DNA methylation analysis using long-read sequencing, methods and application

Brynja D. Sigurpálsdóttir

Dissertation submitted to the School of Technology, Department of Engineering, at Reykjavík University in partial fulfillment of the requirements for the degree of Doctor of Philosophy

2025

Thesis committee:

Bjarni V. Halldórsson, Supervisor
Associate professor, Reykjavík University, Iceland

Hans T. Björnsson, Committee member
Professor, University of Iceland, Iceland

Ólafur A. Stefánsson, Committee member
Project leader, deCODE genetics, Iceland

Ólafur E. Sigurjónsson, Committee member
Dean of the School of Technology, Reykjavík University, Iceland

Kasper D. Hansen, Examiner
Professor, John Hopkins, USA

ISBN 978-9935-9655-9-2 electronic version

ORCID 0000-0001-5227-2372

Table of contents

Abstract 2

List of publications 4

Acknowledgements 5

Introduction 6

Background 7

Summary of papers 11

Paper I 12

Paper II 40

Paper III 71

Discussion 87

References 90

Appendix I: Paper I – Supplementary materials 93

Appendix II: Paper II - Supplementary materials 113

Appendix III: Paper III - Supplementary materials 129

Abstract

Beyond the primary genetic code, DNA carries a second layer of information in the form of epigenetic modification, predominantly DNA methylation, which do not alter the DNA sequence itself but instead alter how genetic information is interpreted and expressed[1]. DNA methylation is a dynamic process that can change in response to aging[2] and environmental exposures[3], such as smoking[4,5]. Importantly, altered DNA methylation patterns have been associated with a wide range of diseases[6], including cardiovascular disorders[7] and cancer risk[7,8]. Characterization of the role of DNA methylation is requisite on accurate genome-wide detection of methylation, however traditional approaches provide limited coverage of the genome and cannot accurately resolve parent-of-origin specific patterns. Long-read sequencing (LRS) overcomes these limitations and offers new opportunities to study DNA methylation.

This thesis makes three main contributions. First, we establish LRS as highly reliable for DNA methylation detection and introduce filtering strategies to ensure high-quality data. Second, we reveal that sequencing variants drive much of the correlation of methylation with gene expression. Third, we map age associated methylation changes across the genome and reveal parent-of-origin specific changes of imprinting fidelity. Together, these studies demonstrate the utility of LRS for large scale methylation analysis and highlight its potential to uncover novel biological insights, refine our understanding of epigenetic regulation and inform future translational applications.

Key words: DNA methylation, long-read sequencing, nanopore sequencing, gene regulation, methylation age

Útdráttur

Erfðarófið samanstendur af kjarnsýrum, en geymir einnig annað lag af upplýsingum í formi utangenamerkja, einkum DNA metýlunar, sem talið er að hafi áhrif á það hvernig erfðaefnið er lesið og tjáð. DNA metýlun er breytileg og breytist meðal annars með aldri, og ýmsum umhverfisáhrifum, s.s. lifnaðarhætti og reykingum. Sýnt hefur verið fram á að breytingar í DNA metýlun tengjast fjölmörgum sjúkdómum, þar á meðal hjarta- og æðasjúkdómum sem og ýmsum tegundum krabbameina. Til að rannsaka hlutverk DNA metýlunar er nauðsynlegt að kortleggja hana yfir allt erfðamengið, en hefðbundnar aðferðir geta ekki greint nema hluta erfðamengisins og geta ekki aðgreint metýleringu á móður- og föðurlitningum. Með þriðju kynslóðar raðgreiningar (e. *long-read sequencing*) er hægt að greina langar raðir af DNA og þar með yfirstíga ýmsar hindranir sem voru til staðar. Það gerir okkur þar með kleift að svara nýjum spurningum varðandi hlutverk DNA metýlunar.

Í þessari rannsókn, sýnum við fram á að mælingar á DNA metýlun með þriðju kynslóðar raðgreiningu eru almennt áræðanlegar og við kynnum nýjar aðferðir við gagnaúrvinnslu sem bæta gæði mælinganna. Við sýnum að erfðabreytileiki hefur mikil áhrif á fylgni milli metýleringar og tjáningu gena. Að lokum kortleggjum við aldurstengdar breytingar í DNA metýlun yfir erfðamengið og sýnum fram á að þær koma einnig fram á stöðum þar sem metýlering ræðst af því hvort um er að ræða móður- eða föðurlitning, þ.e. á svokölluðum genagreyptum svæðum (e. *genomic imprinting*). Með þessu sýnum við fram á að þriðju kynslóðar raðgreining hentar vel til erfðarannsókna á stórum skala sem getur varpað ljósi á hlutverk metýleringar í sjúkdómum sem og öðrum líffræðilegum breytileika mannsins.

Efnisorð: DNA metýlun, þriðju kynslóðar raðgreining, nanopore raðgreining, genatjáning, metýlerunar aldur

List of publications

This thesis consists of two peer-reviewed journal papers and one pending paper submission, found in Appendix I-III.

Paper 1

Sigurpalsdottir, B.D., Stefansson, O.A., Holley, G. *et al.* A comparison of methods for detecting DNA methylation from long-read sequencing of human genomes. *Genome Biol* **25**, 69 (2024). https://doi.org/10.1186/s13059-024-03207-9

Paper 2

Stefansson, O.A., Sigurpalsdottir, B.D., Rognvaldsson, S. *et al.* The correlation between CpG methylation and gene expression is driven by sequence variants. *Nat Genet* **56**, 1624–1631 (2024). https://doi.org/10.1038/s41588-024-01851-2

In this paper I was responsible for creating the set of high-quality CpG sites, phase the data and run the methylation to variant associations.

Paper 3

Sigurpalsdottir, B.D., Holley G., Sverrisson, S.Þ. *et al*. Nanopore sequencing identifies age-associated disruption of imprinting fidelity in the human genome. Manuscript submitted.

Acknowledgements

I would like to express my deepest gratitude to my family for their unwavering support and encouragement throughout the various stages of my education. I am equally grateful to my friends and colleagues at deCODE, whose endless knowledge and support have greatly enriched this work. Without their contributions and encouragement, this thesis would not have been possible.

Introduction

DNA carries a second layer of information in the form of epigenetic modification, which do not alter the DNA sequence itself but instead alter how genetic information is interpreted and expressed[9]. The predominant modification of DNA in humans is the addition of a methyl group to a cytosine preceding a guanine (CpG), commonly referred to as CpG methylation or simply DNA methylation[10]. This modification is heritable through cell division, allowing methylation patterns to be stably maintained as cell proliferate[1]. DNA methylation has long been suspected to have roles in maintaining cellular identity and cell-type specification[11], X chromosome inactivation[12], chromatin and transcriptional regulation[1], and aging[2]. More importantly, abnormal DNA methylation patterns have been found in myriads of human diseases[6].

Accurate and comprehensive detection of DNA methylation patterns is necessary to understand the complex underlying mechanism. Traditional methods, such as arrays[13–16] or oxidative bisulfite sequencing (oxBS)[17–20], provide incomplete coverage of the genome and lack the ability to resolve haplotype specific methylation patterns. With the advancement of LRS, DNA methylation can be detected directly at single-molecule resolution[21]. Extender read lengths of LRS are the key to accurately phase the methylation levels to parental haplotypes[22].

Here, we provide background relevant to my thesis and outline the scope of the three papers on which it is based. This thesis addresses three key research questions: (i) How reliably can LRS detect DNA methylation compared to existing methods, and what strategies ensure high-quality data, (ii) to what extent are correlation between DNA methylation and gene expression driven by sequencing variants, and (iii) how DNA methylation changes across the human lifespan, particularly with respect to parent-of-origin specific patterns. Lastly, we illustrate how this thesis advances the current understanding and methodology in DNA methylation analysis and future directions for the field are outlined.

Background

DNA methylation

DNA methylation is a process which involves the covalent addition of a methyl group to the C5 position of cytosines, predominantly in CpG dinucleotides[10]. This reaction is catalyzed by a family of DNA methyltransferases (DNMTs), which establish de novo methylation patterns during early development and maintain them through subsequent cell divisions[23]. These patterns are highly dynamic during embryogenesis and differentiation[11].

CpGs are unevenly distributed across the genome, often clustering into CpGs islands, which are CpG rich regions on average 1000 bp long[1]. Promoter sequences are frequently embedded within CpG islands, which are typically unmethylated, likely because of TF binding[24]. Because certain TFs recognize motifs containing CpGs, their binding can be directly influenced by CpG methylation[25]. Therefore, CpG methylation can modulate transcriptional regulation although in many cases methylation levels may not be the driving force of gene expression regulation[24,26,27].

One of the clearest functional roles of DNA methylation is in genomic imprinting, a process in which genes are expressed in a parent-of-origin specific manner. Here, methylation selectively silences one parental haplotype, ensuring monoallelic expression[28]. Although, most methylation marks are reset after fertilization, imprinted regions are protected from this reprogramming wave, ensuring parent-specific methylation patterns and expression[29]. Imprinting plays a key role in normal growth, development and metabolism[30,31] and disruption of imprinting controls is associated with disorders, like Prader-Willi and Angelman syndromes[32], as well as with cancer[33].

Technologies for DNA methylation detection

The most used technologies for DNA methylation detection are Illumina methylation arrays[13–16], bisulfite sequencing (BS)[17–20], and LRS[34,35]. BS and methylation arrays both start with bisulfite treatment of DNA, which converts unmethylated cytosines to uracils, while methylated CpGs remain as cytosines and then PCR amplification of the DNA converts uracils to thymine[36]. Methods based on bisulfite treatment of DNA therefore require DNA methylation to be inferred indirectly from the sequenced DNA. In LRS, both nanopore sequencing (NS)[34] and single-molecule real-time sequencing (SMRTS)[35], DNA methylation can be detected directly, without the need for bisulfite conversion or additional sequencing using, enabling broader CpG coverage[21].

<u>Illumina methylation arrays</u>

On methylation arrays the converted DNA is then hybridized to an array containing locus-specific probes designed to target thousands to hundreds of thousands of CpG site across the genome (27k-850k)[13–16]. Bisulfite treated DNA fragments are then bound to locus-specific probes on the array that

distinguish methylated (C) from unmethylated (T) states, using either two probes (Infinium I) or one probe (Infinium II)[14,15]. Next, a single-base extension reaction incorporates a fluorescently labelled nucleotide that matches the methylation state. The intensity of fluorescence from the methylated versus unmethylated probe is measured by a scanner, as done by Illumina short read sequencing (SRS)[14–16]. For each CpG site on the array, a β-value, ranging from 0-1, representing the proportion of DNA molecules that are methylated at that site is reported[37].

BS and oxBS-sequencing

When the bisulfite treated DNA fragments are sequenced directly the technology is known as bisulphite sequencing (BS). In BS, 5-hydroxymethylcytosine (5hmC)[38], another modification found in DNA, is also read as methylated cytosine by sequencing methods that rely on bisulfite treatment and thus cannot be distinguished from 5mC methylation. By adding oxidation step before the bisulfite conversion, 5hmC is converted to 5-formylcytosine, which then is converted to uracil after bisulfite treatment and can therefore be distinguished from 5mC[20,39]. This method is known as oxidative bisulfite sequencing (oxBS)[17]. These treatments negatively influence the quality of DNA samples as they can cause severe DNA degradation and loss of sequence informativity, which makes the alignment more challenging, thereby complicating the sequencing process[13,18–20,40].

Nanopore sequencing

In NS, a single strand of DNA or RNA is passed through a small biological pore (a nanopore) embedded in a protein membrane. An electrical current is applied across the membrane, and as the nucleotides are moved through the nanopore, they disrupt the ionic current. The changes in signal produced by each nucleotide are then translated by machine-learning algorithms to the underlying sequence[21].

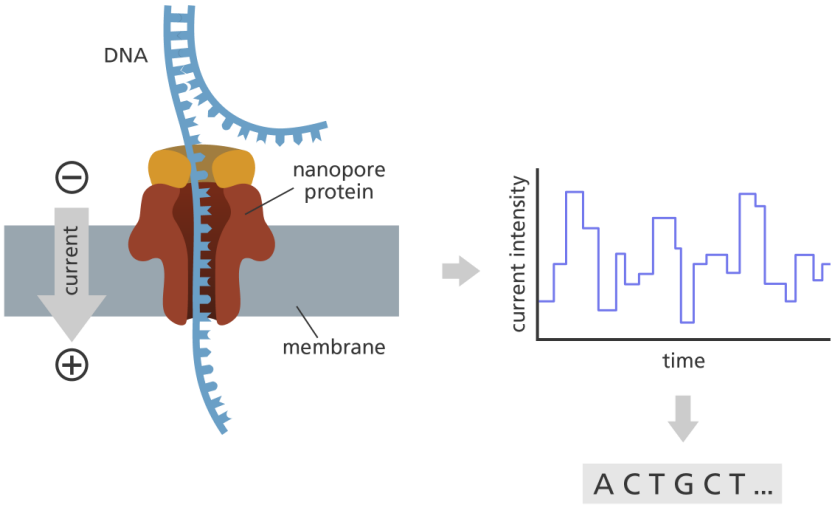


Figure 1. Oxford nanopore technology[41].

Nanopolish

Nanopolish[42] is a methylation detection algorithm for NS that uses a hidden Markov model (HMM) to assign a log-likelihood ratio (LLR) for the presence/absence of cytosine methylation at each CpG site. Measurements where absolute LLRs that do not exceed critical threshold are classified as unambiguous. Then, we calculated the per unit methylation level by the fraction of reads classified as methylated out of all unambiguous reads.

Dorado/Guppy

Dorado[43] is the latest version of the basecalling algorithm from Oxford Nanopore Technologies (ONT), where modified bases can be detected using convolutional neural network (CNN). The algorithm is trained on extended DNA alphabet, including 5mC, 5hmC and other modifications. Likelihood for each CpG being 5mC is reported and if the likelihood does not exceed critical threshold the measurement is classified as unambiguous. As described before, we calculate the per CpG methylation level by the fraction of reads classified as methylated out of all unambiguous reads. Previous versions of the algorithm were referred to as Guppy.

Phased methylation calling

Phasing is the process of determining from which parent genomic locus is inherited. For phasing of the methylation, we use long-range phasing to assign set of variants to parental haplotypes[22]. Heterozygous carriers of phased sequence variants within sequencing read, allowed us to assign CpG methylation calls to maternal and paternal chromosomes. Extended read lengths of LRS are the key to phasing as multiple variants can be present on one read.

DNA methylation aging clocks

As individuals age, changes in methylation levels occurs. These methylation changes form the basis for highly accurate “methylation clocks” that aim to estimate biological age, which is a measure of how old your body seems biologically, based on physiological and/or molecular markers[44–47]. The gold-standard multi-tissue clock from Horvath, was developed from Illumina DNA methylation arrays, encompassing 51 healthy tissue and cell types[45]. The clock consists of 353 CpGs selected using Elastic net model and its test and train median absolute error (medAE) was 3.6 years and 2.9 years, respectively. Similarly, Hannum’s clock was created using 656 samples processed using Illumina Infinium 450k array and consisted of 71 CpGs with overall and test medAE of 3.9 years and 4.9 years, respectively[44]. “Second generation” of methylation clocks aim to go beyond simply estimating chronological age and are trained to capture aspects of biological aging by including clinical biomarkers to measure phenotypic age[46] or incorporate mortality risk factors and time to death data plasma protein levels and smoking history to their training to predict life- and healthspan[47].

While these clocks have enabled valuable insights into age-related phenotypes and risk stratification[44–47], their mechanistic foundation remain poorly understood[48].

Summary of papers

In the first paper of this thesis, we compare methylation detection from 132 oxBS, 50 SMRTS and 7,179 NS whole blood DNA samples from Icelanders. We establish LRS as highly reliable for DNA methylation detection and introduce filtering strategies to ensure high-quality data, increasing confidence in using it for large scale methylation studies. In the second paper, we apply these methods to 7,179 NS whole blood samples to dissect the relationship between gene expression, methylation and genetic variants on haplotype level and reveal that genetic variants drive much of the correlation with gene expression. In the third paper, we use nanopore sequencing to characterize age-associated methylation changes, including parent-of-origin specific effects and imprinting fidelity.

Paper I

Title: A comparison of methods for detecting DNA methylation from long-read sequencing of human genomes

Authors:

Brynja D. Sigurpalsdottir[1,2], Olafur A. Stefansson[1], Guillaume Holley[1], Doruk Beyter[1], Florian Zink[1], Marteinn Þ. Hardarson[1,2], Sverrir Þ. Sverrisson[1], Nina Kristinsdottir[1], Droplaug N. Magnusdottir[1], Olafur Þ. Magnusson[1], Daniel F. Gudbjartsson[1,3], Bjarni V. Halldorsson[1,2*], Kari Stefansson[1,4]

Affiliations:

1. deCODE genetics / Amgen Inc., Sturlugata 8, Reykjavík, Iceland
2. School of Technology, Reykjavík University, Reykjavík, Iceland
3. School of Engineering and Natural Sciences, University of Iceland, Reykjavík, Iceland
4. Faculty of Medicine, School of Health Science, University of Iceland, Reykjavík, Iceland

*Correspondence to: Bjarni V. Halldorsson (Bjarni.Halldorsson@decode.is) and Brynja D. Sigurpalsdottir (Brynja.Sigurpalsdottir@decode.is)

Abstract

Background: Long-read sequencing can enable detection of base modifications, such as CpG methylation, in single molecules of DNA. The most commonly used methods for long-read sequencing are nanopore developed by Oxford Nanopore Technologies (ONT) and single molecule real-time (SMRT) sequencing developed by Pacific Bioscience (PacBio). In this study, we systematically compare the performance of CpG methylation detection from long-read sequencing.

Results: We demonstrate that CpG methylation detection from 7,179 nanopore sequenced DNA samples is highly accurate and consistent with 132 oxidative bisulfite sequenced (oxBS) samples, isolated from the same blood draws. We introduce quality filters for CpGs that further enhance accuracy of CpG methylation detection from nanopore sequenced DNA, while removing at most 30% of CpGs. We evaluate the per-site performance of CpG methylation detection across different genomic features and CpG methylation rates and demonstrate how the latest R10.4 flowcell chemistry and base-calling algorithms improve methylation detection from nanopore sequencing. Additionally, we show how the methylation detection of 50 SMRT-sequenced genomes compares to nanopore sequencing and oxBS.

Conclusions: This study provides the first systematic comparison of CpG methylation detection tools for long-read sequencing methods. We compare two commonly used computational methods for detection of CpG methylation in a large number of nanopore genomes, including samples sequenced using the latest R10.4 nanopore flowcell chemistry and 50 SMRT sequenced samples. We provide insights into the strengths and limitations of each sequencing method as well as recommendations for standardization and evaluation of tools designed for genome scale modified base detection using long-read sequencing.

Background

The predominant modification of DNA in humans is the methylation of a cytosine preceding a guanine (CpG), commonly referred to as either CpG methylation or 5-mCpG (1). Accurate detection of 5-mCpG patterns is necessary to understand the complex regulatory mechanisms underlying gene expression (2), cellular differentiation (3) and imprinting (4). Currently, the most common 5-mCpG detection methods (5,6) do not directly detect base modifications in DNA as they rely on bisulfite conversion of DNA samples followed by either targeted methylation assays or whole genome bisulfite sequencing (WGBS). Array-based methods enable measurement of up to ~900 thousand CpG sites (7), while WGBS has the potential to measure most of the ~30 million CpG sites in the human genome (8).

Bisulfite treatment of DNA converts unmethylated cytosines to uracils, while methylated CpGs remain as cytosines and PCR amplification of the DNA then converts uracils to thymine (9). Methods based on bisulfite treatment of DNA therefore require CpG methylation to be inferred indirectly from the sequenced DNA. Additionally, 5-hydroxymethylcytosine (5hmC), another modification found in DNA (10), is also read as methylated cytosines by sequencing methods that rely on bisulfite treatment and thus cannot be distinguished from 5-mCpG. By adding an oxidation step before the bisulfite conversion, 5hmC is converted to 5-formylcytosine (5fC), which then is converted to uracil after bisulfite treatment and can therefore be distinguished from 5-mCpG (11,12). This method is known as oxidative bisulfite sequencing (oxBS) (6). These treatments negatively influence the quality of DNA samples as they can cause severe DNA degradation, thereby complicating the sequencing process (8).

With the advancement of long-read sequencing, methylation detection can be accomplished directly from the raw sequence data, offering the possibility to perform detection of a wide range of modifications without the need for chemical treatments of the DNA (13). Long-read sequencing technologies have the capability to produce substantially longer reads, at the cost of having higher error rate than previous short-read technologies.

Nanopore sequencing uses a protein nanopore embedded on a synthetic membrane (14). An electrical current is applied across the membrane, leading the negatively charged single-stranded DNA to move through the nanopore. Changes in the electrical current are measured as each DNA molecule disrupts the ion flow in the pore. Importantly, nanopore sequencing has the ability to detect modified bases by distinguishing their electrical current shifts, from those of unmodified bases, measured as they pass through the pore (13,15).

SMRT sequencing uses hairpin adapters to attach to DNA fragments and create a single stranded circular template that can be sequenced continuously. The sample is then loaded to a smart cell containing millions of zero-mode waveguides equipped with fluorescent nucleotides, such that each unique base is labelled with a unique colour of fluorescent. Similar to nanopore sequencing, SMRT

sequencing can distinguish modified bases from unmodified bases by measuring the time it takes to incorporate the next base during DNA synthesis process as modified bases alter the kinetics of this process (13).

Previous studies have extensively evaluated the performance of different tools for methylation detection of nanopore sequencing (16,17). In this study, we present a systematic comparison of 5-mCpG methylation detection tools for nanopore sequencing (ONT) of 7,179 DNA samples, including 22 samples sequenced with the latest nanopore flowcell chemistry, 132 oxBS sequenced samples from the same blood draw, and 50 samples sequenced using SMRT technology (PacBio). By analysing large numbers of genomes, we accurately assess the reliability of CpG methylation predictions from nanopore long-read sequencing and introduce generalized quality filters that can be applied to other cohorts, providing guidance for researchers performing 5-mCpG studies based on long-read sequencing.

Results

Detection of CpG methylation with nanopore sequencing

We sequenced whole blood from 7,179 individuals to an average coverage of 20.6x per sample (median 19.5x, ranging from 10-108x) on 8,906 promethION flowcells from ONT. The same set of samples was used to investigate the correlation between CpG methylation, gene expression and sequence variants (18). CpG methylation detection was performed using Nanopolish (19), which groups CpGs located within 10 bp of each other, referred to here as CpG units. Nanopolish takes reference aligned reads as input and outputs for each read the strand of the reference that was sequenced and for each CpG unit a log-likelihood ratio (LLR) of it being methylated or not. The LLR is then translated to binary values indicating the methylation status of sequenced CpGs. We classified CpG units as “unreliable” when the LLR did not meet our criteria for predicting a CpG unit as either methylated or unmethylated. Here we restrict our analysis to 22,178,458 autosomal CpG units, containing the 27,651,488 CpG sites, detected by Nanopolish in our cohort.

CpG methylation measurements are comparable between nanopore sequencing and oxBS

As a baseline for 5-mCpG rates we used 132 DNA samples sequenced by oxBS in our previous study (20) to an average coverage of 25x (median 24.7x, range 15-41x). For each CpG unit, we calculated the average 5-mCpG rate over all individuals in each dataset separately (7,179 in nanopore and 132 in oxBS) and assessed the performance of Nanopolish by evaluating the Pearson correlation coefficient between average 5-mCpG rates from oxBS data and the corresponding average 5-mCpG rates predicted from Nanopolish, across all CpGs. We refer to this correlation as per CpG average Pearson correlation (APC).

Our analysis revealed a high APC between the 5-mCpG rates in the two datasets (r=0.9594; 95%CI:0.9594-0.9595) and the mean absolute difference (MAD) in the 5-mCpG predictions per CpG was 0.0471 (95%CI:0.0471-0.0472) per CpG.

We measured the overall methylation levels per individual by counting the number of times a methylated status was assigned to a CpG detected in sequences obtained from a given DNA sample to then divide this number by the total number of times we were able to assign a methylation status (unmethylated/methylated) to CpG sites in sequences obtained from that same DNA sample. We find that the overall methylation levels were on average lower in nanopore sequenced samples than in those sequenced by oxBS ($\bar{x}_{\text{Nanopolish}}$=0.767; 95%CI: 0.763-0.770 versus $\bar{x}_{\text{ox-BS}}$=0.773; 95%CI: 0.770-0.775, Wilcoxon rank sum test $p=2\cdot10^{-6}$) (Fig. 1A). As short-read sequences can be more difficult than long-read sequences to align to the reference genome, it is possible that these subtle differences in overall methylation levels between nanopore and oxBS sequenced samples are due to challenges in accurately aligning short-read sequences to the reference genome, which may affect the detectability and thereby measurement of certain CpGs by each of the two methods.

Coverage affects the consistency of CpG methylation measurements in nanopore data

Next, we performed a matched sample-to-sample analysis based on the 132 individuals for which DNA samples were sequenced using both nanopore and oxBS and evaluated the Pearson correlation and MAD. We found that the correlation varied from 0.71 to 0.94 and the MAD from 0.076 to 0.14. The correlation was notably higher and MAD lower for high coverage samples, indicating that sequencing coverage of approximately 12x or more per sample is advisable for accurate methylation detection and sequencing at 20x or greater yields even more accurate results (Fig. 1B, 1C). We then calculated the Pearson correlation for each sample, for all CpG sites with high sequence coverage (greater than 25x) supporting a minimum nanopore sequencing depth of a CpG unit as 20x for obtaining a highly reliable measurement of its 5-mCpG rate (Fig. 1D).

The accuracy of the measured 5-mCpG rate is not affected by different versions of the basecalling algorithm nor changes in the error rate within the range of reported error rate of nanopore sequencing (Additional file 1: Fig. S1, S2, Additional file 2: Tab. S1).

Nanopore data is more consistent in unmethylated and methylated CpG units

To capture the distribution of the methylation predictions, we divided the paired data into four categories based on methylation rates in oxBS: unmethylated (0-0.15), low-methylated (0.15-0.5), intermethylated (0.5-0.85) and methylated (0.85-1). We found that Nanopolish predictions were consistent with oxBS measurements (Fig. 1E, Additional file 2: Tab. S2). We limit our analysis to CpGs with at least 25x coverage in oxBS and consider a prediction made by Nanopolish to be correct if the prediction falls into the same of the four categories as the oxBS. We see that the highest fraction

of correctly predicted CpG units was for unmethylated CpGs (86%), followed by methylated (77%), intermethylated (56%) and low methylated (52%) (Fig. 1F). The lower fraction of correct predictions among low- and intermethylated CpGs may be due to a higher propensity of the methylation in these categories to fall close to the boundaries of these classes and the higher variance of 5-mCpG rates expected for these categories i.e., as the distribution of predicted methylation states is far more uniform for unmethylated and methylated CpGs in comparison to low- and intermethylated CpGs.

Fig. 1. **Nanopore sequencing and oxBS performance in the same DNA samples.** The consistency in 5-mCpG rates measured by nanopore sequencing and oxBS in DNA samples isolated from the same 132 individuals was estimated by: **A** The overall measurement of 5-mCpG rates in each of the 132 DNA samples measured by ONT (red) and oxBS (green), Y-axis is limited to (0.7,0.8). The centre line (solid black) shown in each box represent the median; the box limits represent upper and lower quartile, whiskers represent 1.5x interquartile range. **B** The Pearson *r* correlation coefficient, y-axis and **C** mean of the absolute differences in 5-mCpG rates of each CpG, y-axis, with respect to nanopore sequencing coverage in each sample on the x-axis. Panels D, E, F analyse sites that have >25x coverage in oxBS. **D** CpG coverage underlying the 5-mCpG rates, i.e. the number of sequences that were used to compute the 5-mCpG rate for a given CpG, in nanopore sequenced samples, x-axis, influences the consistency (Pearson *r*), y-axis, with 5-mCpG rates measured with high coverage by oxBS. The y-axis is limited to (0.5,1) **E** CpG rates in nanopore (y-axis) and oxBS (x-axis, binned). The mean is represented with red (ONT) and green (oxBS). **F** Number (y-axis, unit = million CpGs) of correctly classified (blue) by nanopore sequencing in a sample-to-sample comparison. Incorrectly classified CpGs are coloured according to the absolute difference in 5-mCpG rates (colour legend).

Nanopolish methylation prediction quality is affected by CpG unit sequence context

Although the results of nanopore and oxBS are highly correlated, there are regions in the genome where the methylation detection is more difficult due to limitations in the sequencing method, mapping, or the methylation detection algorithms. To evaluate the performance of the methylation detection in nanopore sequenced DNA, we compared the APC of CpG units located inside and outside of regions where we expected difficulties in methylation predictions.

Nanopolish predicts methylation status from reads aligned to the human reference genome (GRCh38) (21), which instigates a risk of error when predicting the methylation status of CpG units located close to sequence variants. We found that CpG units located within 5 bp of a sequence variant had a lower APC (r=0.9219; 95%CI:0.9218-0.9221) than other CpG units (r=0.96560, 95%CI:0.96557-0.96563) (Fig. 2A). This likely is because Nanopolish assumes that aligned sequences are the same as those found in the reference genome. As a result, the electric signal, produced by a short stretch of a DNA sequence containing an unmethylated CpG, but including the alternative allele of a nearby sequence variant, may be similar to the signal produced in the presence of reference allele and a 5-mCpG.

We define dark regions (22) as sequences where ≥90% of the reads have mapping quality <10, coverage <5x on average and base quality <20 in DNA samples analysed on Illumina sequencers. Dark regions often contain large contiguous tandem repeats (e.g., centromeres and telomeres) or larger specific DNA regions that have been duplicated (22), causing the mapping to be unreliable. The APC for CpG units within dark regions was lower (r=0.698; 95%CI:0.697-0.699) than other CpG units (r=0.96320; 95%CI:0.96318-0.96323) (Fig. 2A). This poor correlation in these regions is likely largely attributable to the difficulty in measuring the methylation rates of CpG units that reside within these regions using oxBS, as mapping is generally more reliable in long reads. When the mapping is incorrect, the 5-mCpG rates are predicted from wrong reference sequence leading to incorrect predictions.

We defined abnormal sequencing coverage, as greater than 1.5 times the average coverage or less than 0.5 times the average coverage, and show that these CpG units tend to have lower APC (r=0.7223; 95%CI:0.7218-0.7225) than other (r=0.9646; 95%CI: 0.9645-0.9646) (Fig. 2A, Additional file 1: Fig. S3A, B), likely because of duplicated regions (such as tandem repeats) or mapping errors.

As DNA methylation is in most cases symmetric, meaning that cytosines in CpGs are methylated on both DNA strands (23), and hemi-methylated CpGs, where one strand is methylated while the other is unmethylated, are rare in the genome (24) we investigated strand bias, defined as the difference in the absolute value of the estimated 5-mCpG rates of the forward and reverse strands. We found that the magnitude of strand bias is low in oxBS data, with mean strand bias of 0.026 (Quartiles: 0.0055,0.028) (Additional file 1: Fig. S4). Strand bias was much higher in ONT Nanopolish data (mean=0.095, Quartiles: 0.017,0.11, Wilcoxon rank sum test, $p<2\cdot10^{-16}$), suggesting that strand bias may indicate problematic regions with unreliable methylation predictions. As there is far less strand bias in oxBS, we assume that these are unreliable in nanopore because of methylation detection artefacts. Notably, CpG units with strand bias greater than 0.2 (Additional file 1: Fig. S3C, D) had lower APC (r=0.8279; 95%CI:0.8275-0.8282) than other CpG units (r=0.97411; 95%CI:0.97409-0.97414) (Fig. 2A).

To further investigate the quality of methylation predictions in our nanopore sequenced DNA samples, we examined CpG units with low fraction of reliable reads (FRR), defined as the fraction of reads where the absolute log likelihood ratio exceeds the defined cut-off. CpG units with FRR below 0.5 had a lower APC (r=0.819; 95%CI:0.816-0.820) than other CpG units (r=0.96868; 95%CI:0.96866-0.96871) (Fig. 2A, Additional file 1: Fig. S3E, F).

Consequently, we define problematic CpG units as being within dark regions, within 5 bp distance from a SNP, having coverage ≤0.5 times the average coverage or ≥1.5 times, strand bias ≥0.2 and FRR ≤0.5. These CpGs were removed them from our analysis, resulting in a set of 15,644,462 (70.5%) high-quality CpG units (hq-CpGs), containing 19,685,181 (71.2%) CpG sites in the reference

genomes (hg38). The APC for the hq-CpGs was 0.98582 (95%CI:0.98581-0.98584) compared to 0.9594 (95%CI:0.9594-0.9595) for the complete set and we found lowered MAD (Additional file 2: Tab. S4), between the predictions of hq-CpGs, indicating improved accuracy. The overall 5-mCpG rates were higher among hq-CpGs than among non hq-CpGs (Additional file 2: Tab. S4). Furthermore, correlation coefficients were consistently higher for methylation measurements of hq-CpGs in the same DNA samples analysed by Nanopolish and oxBS (Additional file 1: Fig. S5).

The highest number of CpG units were excluded from the set of hq-CpGs due to their proximity to a sequence variant, followed by high strand bias and low FRR (Fig. 2B). A similar proportion of singletons, defined as CpG units containing one CpG and non-singletons were excluded from the set of high-quality CpG units or 30% and 26%, respectively (Fig. 2C). Notably, a higher proportion of low- (50%) and intermethylated (51%) CpG units were excluded from the set of hq-CpGs than unmethylated (17%) and methylated (19%) (Fig. 2D). Most CpGs (57.7%) are removed from the low- and intermethylated groups because of high strand bias. The hq-CpGs were evenly distributed across the number of CpGs within a unit and chromosomes (Additional file 1: Fig. S6).

Fig. 2 The quality of 5-mCpG rate measurements by DNA sequence attributes. **A** APC estimates (x-axis), for CpG sites located outside (pink) and inside (grey) of DNA sequence attributes, y-axis, and the APC estimates based on all CpGs (vertical black line). **B** The number of CpG units (red) and sites (green), x-axis, found inside of each attribute, y-axis. **C** The proportion of high-quality (dark blue) and non high-quality (light blue) CpG units among singletons and non-singletons, x-axis. **D** The proportion of high-quality and non high-quality CpG units within each methylation state category, x-axis, defined by binning the mean of 5-mCpG rates measured by Nanopolish.

Guppy outperforms Nanopolish per CpG-site in comparison to oxidative bisulfite sequencing data

The recent improvements in algorithms for ONT basecalling have greatly enhanced the accuracy and efficiency of the basecalling. Specifically, a recent version of the basecaller, referred to as Guppy, can now perform CpG methylation detection at the basecalling stage by adding 5-mCpG to the DNA alphabet. We predicted the 5-mCpG rates of CpGs in 304 samples with Guppy (version 6.2.1) and calculated the average rates for each CpG over all individuals. Since Guppy does not group the CpGs like Nanopolish, we assumed the same rates for each CpG within a CpG unit in Nanopolish and compared the rates at the CpG site level.

The methylation calls from Guppy and Nanopolish were highly correlated, with an APC of 0.96558 (95%CI:0.96555-0.96561) for the full set of CpGs. Guppy had higher APC with oxBS data (r=0.97256; 95%CI: 0.97255-0.97259) than Nanopolish (r=0.9594; 95%CI:0.9594-0.9595). The overall 5-mCpG rates were lower for Guppy ($\bar{x}_{Guppy}$=0.7634; 95%CI: 0.7633,0.7635) than oxBS

($\bar{x}_{oxBS}$=0.7756; 95%CI:0.7755-0.7757; p<2·10^{-16} Wilcoxon rank sum test). Interestingly, Guppy had lower mean strand bias ($\bar{x}$=0.064; Quartiles: 0.016,0.077) than Nanopolish ($\bar{x}$=0.095; Quartiles:0.017,0.11; Wilcoxon rank sum test, p<2·10^{-16}), although the strand bias was still higher than in oxBS ($\bar{x}$=0.026; Quartiles:0.0055,0.028; Wilcoxon rank sum test, p<2·10^{-16}).

By applying the same quality filters as specified for Nanopolish, we identified 22,256,402 (80.5%) hq-CpGs. This represents a 9.3% increase compared to the set of hq-CpGs identified using Nanopolish data. This difference is mainly explained by two factors, first this version of Guppy does not report number of reads where the probability of the call was below the threshold and therefore the FRR filter is not applicable and second, Guppy has lower strand bias, leading to more hq-CpGs being retained. The APC between the set of Guppy hq-CpGs and oxBS data was 0.98691 (95%CI: 0.98690-0.98693), compared to 0.97257 (95%CI: 0.97255-0.97259) for the complete set of CpGs (Additional file 2: Tab. S4, S5).

Moreover, we found high correlations between the matched samples for the methylation predictions generated by Nanopolish and Guppy, and Guppy and oxBS (Additional file 1: Fig. S7,S8). The sample-to-sample correlation between the 5-mCpG predictions from Guppy and the corresponding oxBS rates ranged from 0.62-0.90 for the full set of CpGs and increased to 0.65-0.91 for the set of hq-CpGs. For most samples the correlation was higher between Guppy and oxBS than Nanopolish and oxBS (Additional file 1: Fig. S8A). The strand bias and MAD was also lower for Guppy on average per sample (Additional file 1: Fig. S8B, C).

The latest chemistry attains higher accuracy and improved methylation predictions

ONT has made several improvements to its protein nanopore and motor protein, releasing nine versions of the system to date (15). Our dataset consists mainly of samples sequenced on R9.4 flowcells (released in October 2016) and in addition we sequenced 22 samples on 28 R10.4 flowcells (received as an early access) to an average depth of 9.64x. R10.4 flowcells have two sensing regions designed to provide higher consensus accuracy with homopolymers than the R.9.4 flowcells (15).

The R10.4 flowcells have average sequencing error rate (25) of 3.9%, significantly lower than the 8% average sequencing error rate for the R9.4 chemistry. Although there is high APC between 5-mCpG rates measured in all CpGs with the two types of flowcells (r=0.98190, 95%CI: 0.98188-0.98191), the APC between 5-mCpG rates predicted from nanopore data in all CpGs and oxBS data is higher for R10.4 flowcells ($r_{R10.4}$=0.97845; 95%CI:0.97843-0.97846, $r_{R9.4}$=0.97256; 95%CI: 0.97255-0.97259, Additional file 2: Tab. S5). R10.4 flowcells also show lower average strand bias of 0.047 (Quartiles: 0.0097,0.053) over all CpGs in comparison to R9.4 ($\bar{x}$=0.064; Quartiles: 0.016,0.077) (Wilcoxon rank sum test, p<2e-16) indicating improved accuracy (Additional file 2: Tab. S4). Nonetheless, the strand bias observed in R10.4 flowcells is still higher than that observed in oxBS data. Guppy R10.4 further

showed lower MAD between methylation predictions with oxBS than Guppy R9.4 (Additional file 2: Tab. S4).

Applying the same quality filters as before to the R10.4 dataset, we obtain 22,893,522 (82.8%) high-quality autosomal CpGs, with APC of 0.99067 with oxBS (95%CI: 0.99066-0.99068, Additional file 2: Tab. S4, S5). This is a 2.3% increase in the number of hq-CpGs compared to Guppy data sequenced on R9.4 flowcells, and an increase in APC.

CpG methylation measurements are comparable between SMRT-sequencing, nanopore sequencing and oxBS

We SMRT-sequenced whole-blood samples from 50 individuals on 170 flowcells to average sequencing coverage of 28.5x per sample (range 13.6-41.7x), which was higher than for nanopore R9.4 and R10.4 sequencing methods (Additional file 1: Fig. S9A). The average N50, defined as the length of the sequence read at 50th percentile of the total sequence read length, was similar for SMRT and nanopore R9.4 and R10.4 sequencing methods (Additional file 1: Fig. S9B), but the average sequencing error rate was lower for SMRT-sequencing than either of the two nanopore sequencing methods, or 1.12% (range 1.02-1.31%, Additional file 1: Fig. S9C). We used primrose for methylation detection of SMRT-sequenced samples. The methylation detection step is performed by the sequencer after basecalling. The APC between predicted 5-mCpG rates across all 27,527,663 autosomal CpGs from SMRT-sequencing and oxBS data was 0.97010 (95%CI: 0.97008-0.97013) and the MAD was 0.05691 (95%CI:0.05689-0.05694). After applying our quality filters, we identify 22,554,423 (81.9%) hq-CpGs of the autosomal CpGs with APC of 0.979956 (95%CI: 0.97955-0.97579) (Additional file 2: Tab. S4, S5). In summary, the number of hq-CpGs is similar to R10.4, with fewer filters applied and the APC with oxBS is lower than for either the R10.4 or R9.4 nanopore sequencing methods.

Comparison of CpG methylation predictions from nanopore sequencing and SMRT sequencing

In this comparison, we used the 50 SMRT-sequenced samples (average coverage 26.7x) and 50 nanopore sequenced samples analysed using Nanopolish (average coverage 23.4x), 50 nanopore sequenced samples on R9.4 flowcells and methylation called using Guppy (average coverage 22.0x), all of the 22 nanopore sequenced samples on R10.4 flowcells analysed using Guppy (average coverage 9.64x), and 50 DNA samples sequenced by oxBS (average coverage 25.0x) (Additional file 2: Tab. S3).

We averaged the 5-mCpG rates over all samples and compared the APC correlation coefficient between all five methods (SMRT, R9.4-Guppy, R10.4-Guppy, R9.4-Nanopolish and oxBS) and the absolute difference between 5-mCpG rates and oxBS (Table 1A). 26,345,529 autosomal CpGs were detected in all datasets and used for the comparison. The highest APC was seen for Guppy applied to R10.4 and Guppy applied to R9.4. In comparison to oxBS, the highest APC and the lowest MAD was

also seen for Guppy applied to R10.4 (Tab. 1A). We note, however, that some of the differences in APC and MAD observed between methods may be due to differences in age, gender or smoking status of the samples (Additional file 2: Tab. S3).

Sequence variants around or within CpG introduces mapping bias in oxBS, leading to inaccurate methylation measurements and low APC. Therefore, it is less important to filter on CpGs located close to sequence variants for Guppy and PacBio, because low APC is most likely caused by inaccurate measurements in oxBS (Tab. 1B) and higher APC is seen between Guppy R9.4, Guppy R10.4 and PacBio. We note however, that likely all methods benefit from filtering on CpGs where sequence variants are located close to the CpG as all long-read sequencing technologies use the local sequence context and comparison to the reference genome for predicting the methylation status of CpGs. Not filtering on sequence variants would increase the number of hq-CpGs to about 25.1M (90.7%) and 25.8M (93.7%) hq-CpG for Guppy and PacBio with APC 0.98545 (95%CI:0.98544-0.98546) and 0.97561 (95%CI:0.97559-0.97563), respectively.

Table 1 Comparison between methods. **A** APC comparisons are shown below the main diagonal whereas MAD comparisons are shown above the main diagonal. **B** APC comparisons between methods based on all CpGs, or after restricting to those located close to sequence variants or those located within dark regions as indicated.

A Consistency between methods

		Mean absolute difference (MAD)				
		OxBS	**ONT Nanopolish**	**ONT guppy R9.4**	**ONT guppy R10.4**	**PacBio**
APC	OxBS	-	0.05933	0.05730	0.04772	0.05691
	ONT Nanopolish	0.9460	-	0.05040	0.04888	0.05207
	ONT guppy R9.4	0.9594	0.9653	-	0.04655	0.04608
	ONT guppy R10.4	0.9647	0.9637	0.9817	-	0.04363
	PacBio	0.9563	0.9571	0.9705	0.9785	-

B Consistency between methods in sequencing context

	APC									
	NP vs. Guppy R9.4	**NP vs. Guppy R10.4**	**NP vs. PacBio**	**NP vs. oxBS**	**Guppy R9.4 vs. Guppy R10.4**	**Guppy R9.4 vs. PacBio**	**Guppy R9.4 vs. oxBS**	**Guppy R10.4 vs. oxBS**	**Guppy R10.4 vs. PacBio**	**PacBio vs. oxBS**
All CpGs	0.965	0.963	0.957	0.959	0.982	0.971	0.972	0.978	0.979	0.970
CpGs near sequence variants	0.930	0.921	0.913	0.922	0.973	0.957	0.940	0.934	0.962	0.929
CpGs within dark regions	0.841	0.845	0.841	0.698	0.906	0.873	0.716	0.736	0.876	0.691

Distribution of the 5-mCpG rates

5-mCpG rates computed across all individuals in the five subsets of 50 individuals yielded the expected bimodal distribution for all methods (Fig. 3A, 3B). However, we noticed a shift in the distribution of methylated and unmethylated CpG sites away from 1 and 0, for both Guppy applied to R9.4 flowcells and PacBio. PacBio never reaches 0 or 1, while Guppy R9.4 rarely does. Guppy applied to R10.4 flowcells more closely follows the methylation distribution patterns seen in oxBS sequenced samples than R9.4. Additionally, all methods showed a higher number of intermethylated CpGs than oxBS. The distribution for hq-CpGs is similar with slightly lower fraction of low- and intermethylated CpGs for Guppy R10.4 and PacBio (Additional file 1: Fig. S10). Less CpGs are removed due to strand bias and abnormal coverage for Guppy R10.4 and R9.4 compared to Nanopolish. Interestingly, more are removed because of abnormal coverage for PacBio (Additional file 1: Fig. S11).

5-mCpG rates of functional regions

To investigate the influence of biological context on the accuracy of the methylation predictions, we calculated the average 5-mCpG rates in 50 bp intervals relative to the start of the transcription start sites (TSSs) of genes expressed in whole blood. All methylation detection methods closely replicate the methylation patterns observed in oxBS sequenced samples, which demonstrated a lack of methylation within TSSs (Fig. 3C). Notably, PacBio and Guppy R9.4 exhibited higher rates of CpG methylation at TSSs and lower rates away from TSSs, which is consistent with the slight shift in the methylation distributions observed for these two methods (Fig. 3A, B). Guppy applied to R10.4 flowcells, however, more closely follows the TSS methylation levels seen in oxBS (Fig. 3C). Further, Nanopolish has the lowest MAD with oxBS in unmethylated CpG units (Supplementary Fig. S12).

Long-read sequencing calls more CpGs than oxBS

Long-read sequencing provides a significant advantage in the number of CpG sites captured over previous methods. To quantify this, we compared the number of CpGs called per sample by each long-read based method and found that they all called similar number of CpGs. Restricting our analysis to autosomes, all three methylation detection tools for long-reads called similar number of CpGs (Guppy R9.4=27,467,383, Guppy R10.4=27,369,144, PacBio = 26,739,539 CpGs and Nanopolish=26,487,587, within 22,058,476 CpG units). As expected, oxBS called the fewest CpGs, with an average of 26,002,520 CpGs (Fig. 3D). The varying number of CpGs detected in long-read sequencing is most likely because the criteria set by each method to make confident methylation predictions.

Fig. 3 Comparison of CpG methylation detection by method. CpG methylation rates (ranging from 0-1) averaged across individuals yields the expected bimodal distribution seen in oxBS data for **A**

oxBS, Guppy R9.4 and R10.4 and **B** oxBS, PacBio and Nanopore. The units on y-axis are millions (M). **C** CpG methylation rates averaged in 50 bp bins relative to transcription start sites (TSSs) of genes expressed in whole blood. **D** Number of CpGs called by each method. For Nanopolish we count all CpGs within a CpG unit. Note, the y-axis is limited from 24.5 to 27.7 M (millions). The centre line (solid black) shown in each box represent the median; the box limits represent upper and lower quartile, whiskers represent 1.5x interquartile range.

Discussion

ONT and PacBio sequencing technologies both generate long-reads but their underlying differences in chemistry affects the length of the reads, error rate and throughput. Sequence detection algorithms can also affect the error rate. Consequently, each method has their distinct strengths and limitations. ONT excels in generating longer reads than PacBio, but this advantage comes at the expense of higher error rate. Additionally, ONT is more scalable than PacBio resulting in lower sequencing cost per sample. Both of them exceed short-read sequencing in terms of capturing challenging regions, structural variants detection, read phasing accuracy and creating whole chromosome assembly (13).

Long-read sequencing benefits from using native unamplified DNA for sequencing, because the molecules retain base modifications, allowing for their detection. This makes methylation detection more direct and simplifies the process. Several tools have been developed for methylation detection from long read sequencing, such as Nanopolish (26), DeepSignal (27), DeepMod(28), Tombo(29) and Megalodon (30) for nanopore sequencing and ccsmeth (31) and modbamtools(32) for SMRT-sequencing. Extensive benchmark work on methylation detection tools for nanopore sequencing has been done previously (16,17). Liu et al. (17) concluded that Nanopolish and Guppy required the least amount of CPU time and exhibited the lowest peak memory usage, making it feasible for large scale CpG methylation studies. Nanopolish and Guppy, were also among the overall top performers, along with DeepSignal and Megalodon. However, Nanopolish and Guppy detected 4-6% fewer CpGs than DeepSignal and Megalodon, due to more stringent log-likelihood cutoffs (17). These studies however were based on a small number of samples and did not consider the difference in methylation detection between sequencing methods or consider processing strategies to improve those correlations.

By using a large cohort, we can reduce the risk of drawing erroneous conclusions due to random variability and have more robustness to outliers. Our study extends beyond previous studies by showing that the quantification of methylation varies in quality between CpGs. We filter out unreliable CpGs and define a set of hq-CpGs that led to significantly improved accuracy while still providing comprehensive analysis of over 70% of autosomal CpG sites. By tuning the quality attributes on a large cohort, they are more likely to be representative of the broader population and therefore generalized. The most significant improvement in APC was achieved by removing dark regions and regions with abnormal read coverage. APC between all pairs methods was lower for dark

regions, suggesting that these regions are less reliable for all methods. Filtering out CpGs located ≤5 bp from sequence variants is necessary for Nanopolish because of the way the algorithm is designed, but this is not a necessary filtering criteria for other methods. All methods however, likely benefit from filter on CpGs where a sequence variant occurs on either the cytosine or guanine base within the CpG motif itself. We note, that the aforementioned filters may need to be reevaluated depending on each project‘s needs.

We further show that Guppy applied to R10.4 flowcells with updated chemistry, resolves some of the problems seen in the earlier versions of Guppy applied to R9.4 flowcells, such as strand bias and results in larger set of hq-CpGs. We report that for SMRT-sequencing the methylation predictions (12,13) never reach either fully unmethylated, or fully methylated state. Changes to the model, such that the predictions do not regress away from either of these two extremes may be beneficial.

The performance of long-read sequencing technologies relies heavily on the algorithms applied and the samples used as the training dataset. In many cases, the 5-mCpG detection algorithm is trained on fully methylated and fully unmethylated datasets, resulting in these regions being more accurately called than low- and intermethylated sites. For improved methylation predictions, penalized models, i.e. imposing additional cost on the models for making classification mistakes at these regions may improve the methylation detection. Furthermore, expanding the training dataset on more challenging regions and more human DNA sequences, may improve the methylation predictions. Lastly, consensus approaches, based on the combination of predictions from two or more tools, show promising results for improved accuracy but were not investigated in this study (16).

CpG methylation detection from long-read sequencing faces limitation due to the diffusion of the signal around the CpG. Therefore, the algorithms require the use of intervals for the methylation detection and combination of the kinetic information from neighbouring CpGs to increase the confidence in identifying methylated CpGs. Future work could identify problematic *k*-mers and incorporate that information into the training set to improve the detection reliability.

Long-read sequencing has revolutionized our ability to study CpG methylation without the need for chemical treatment of DNA, providing a higher resolution and more accurate picture of CpG methylation diversity. By enabling accurate phasing of the reads (18), long-read sequencing allows for precise characterization of DNA methylation at single base resolution at haplotype level. This has facilitated the exploration of complex patterns of epigenetic modification and the detection of sequences that are infrequently depleted of methylation in the population that would have been missed using traditional array or short-read sequencing.

Conclusions

CpG methylation detection in nanopore sequenced DNA samples is highly accurate, even for samples with high error rate and SMRT sequencing shows similar results. Based on our comparison, we made five key observations. First, coverage of approximately 10x or higher per sample and per CpG is an important factor for accurate methylation detection. Second, we observed strand bias present in the nanopore data that is not seen in oxBS data. The strand bias decreases with lower error rate and more accurate mapping and methylation predictions. Third, the methylation predictions from all methods are highly correlated and consistent with 5-mCpG detection in samples analysed by the well-established oxBS method. They all replicate known 5-mCpG distributions in the human genome, such as the lack of 5-mCpG in promoter sequences. Fourth, we show improved consistency in 5-mCpG by excluding CpGs according to quality parameters identified herein. Between 7-30% of CpGs are filtered, depending on dataset. The lower the error rate, and the more accurate the mapping of the sequenced DNA, the fewer CpGs need to be excluded for further analysis. Fifth, long-read sequencing detects about 3% more CpGs than oxBS. The number of CpGs detected by each long-read method mainly differ due to the criteria defined by each tool to confidently predict 5-mCpG rates. In summary, we have revealed the strengths and limitations of long-read sequencing methods, a crucial step to enable informed decision when selecting the appropriate sequencing technique and data analysis method.

Methods

Nanopore sequencing and analysis

Dataset

In this study we sequence DNA isolated from whole blood samples from 7,179 individuals (3,745 females and 3,434 males) participating in various studies at deCODE genetics. Analysis of structural variants in a subset of 3,622 of these individuals has been described previously (20). The earliest year of birth was 1890 and 1876, for females and males respectively and the latest was 2015 for both genders. All individuals gave informed consent and all personal identifiers were encrypted by an external agent before being imported into the deCODE database.

Sample preparation

DNA from whole blood was extracted using Chemagic method (perkinElmer), an automated procedure that involves the use of M-PVA magnetic beads. Sequencing libraries were generated using the SQK-LSK109 ligation kit from ONT. Sample input varied from 1-5 µg DNA, depending on the exact version of the preparation kit and the flowcell type used for the PromethION sequencing.

Samples were loaded onto PromethION R9.4.1 and R10.4.1 flowcells following ONT standard operating procedures. Sequencing was performed on PromethION devices.

Basecalling

The samples were analysed with two versions of our pipeline, v3 (5761 R9 flowcells) and v4 (3145 R9 flowcells). The main difference between the pipelines is the version of the basecaller. In v3 squiggle data from PromethION was basecalled using Guppy 3.3.0 (3826 flowcells) using either the 'flipflop' or 'hac' model or 3.2.2 (536 flowcells), 3.6.0 (675 flowcells) and 4.0.14 (724 flowcells) using the 'hac' model. In ont_build38_v4 all data was basecalled using guppy 5.0.11, using the 'sup' model (dna_r9.4.1_450bps_sup_prom.cfg). All 7,179 individuals basecalled with guppy had a minimum reference-genome-aligned sequencing coverage of at least 10x at the time of analysis and 3x per flowcell.

Mapping

Basecalled reads were mapped to the human reference genome GRCh38 (21) with minimap2 (33), versions 2.14-r883 (5748 flowcells), 2.17-r941 (13 flowcells) and 2.22-r1105 (3145 flowcells). The aligned reads were sorted using samtools sort (34) and stored in a BAM file.

CpG methylation detection

All R9.4 flowcells were methylation called using Nanopolish (19) versions 0.11.1, 0.11.3 and 0.13.3. Nanopolish uses a hidden Markov model (HMM) to assign a log likelihood ratio for the presence of a cytosine methylation at each CpG site. We interpret values above 1.921 as indication for cytosine methylation and less than -1.921 for unmodified CpG. Nanopolish groups CpGs within 10 bp distance and assigns a methylation status to each such that all CpGs within a group have the same methylation status. For this reason, we refer to CpGs measured by Nanopolish as CpG units. We first detect the methylation on read level and exclude ambiguous methylation predictions ($-1.921 \leq LLR \leq 1.921$). Then we calculated the per unit methylation level by the fraction of reads classified as methylated out of all unambiguous reads.

The LLR threshold is selected based on Wilks' theorem (35), which states that assuming the null hypothesis is true and the sample size approaches infinity, the distribution of the test statistics, $-2\log(\Lambda)$, asymptotically approaches the chi-squared distribution with degrees of freedom equal to the difference in dimensionality. Here, Λ denotes the likelihood ratio. For 1 degree of freedom and p-value of 0.05 the chi-square value is 3.842. Therefore, we choose 1.921 as a threshold.

Additionally, we called CpG methylation in 304 samples on 325 flowcells using Guppy 5.0.11 or 6.2.1, which are versions of the basecalling algorithm that uses extended alphabet, including 5mC. Guppy consists of a convolutional neural network (CNN) trained on fully methylated DNA created by

treating the DNA with CpG methyltransferase M.SssI and fully unmethylated DNA created using PCR amplification. We then used modBam2Bed (www.github.com/epi2me-labs/modbam2bed) script to extract the methylation values from the bam file and calculate the per-site methylation level.

Pipeline v3 vs pipeline v4

The main difference between the two versions is that v3 uses guppy versions 3.3.0, 3.2.2, 3.6.0 and 4.0.14 for the basecalling, resulting in an error rate of 11.53% on average and v4 uses guppy 5.0.11 for the basecalling, resulting in an error rate of 8.06% on average (Additional file 1: Fig. S1). Version v3 is sequenced on older flowcells and hardware, potentially affecting the quality of the sequence reads and the methylation detection.

R10.4 flowcells

Additionally, we sequenced 22 samples on 26 R10.4 flowcells. Basecalling, alignment and CpG methylation detection were performed on the box using Guppy 6.2.7. CpG methyl tags were then copied from unaligned bam file to aligned bam file and analysed the same way as Guppy R9.4 samples.

OxBS-sequencing and analysis

Dataset

All 132 samples were analysed and described by Zink et al (20).

Sample preparation

Samples were prepared using the TrueMethyl® Whole Genome kit (Cambridge Epigenetix) following the manufacturer's recommendations (see URLs). In short, this involved a three-step procedure: (1) genomic DNA (0.2–0.4 µg) was oxidized using a proprietary oxidant (Cambridge Epigenetix). This step was done to convert all 5-hydroxy methylcytosines to their formyl derivatives, 5-formylcytosines; (2) bisulfite treatment of oxidized DNA converted both cytosines and 5-formylcytosines to uracil, leaving the 5-methylcytosines intact; (3) Illumina-compatible oxBS-seq libraries were prepared, using the appropriate primers and sequence adapters.

Sequencing

All sequencing libraries were quality control monitored for size and concentration using a LabChip GX analyser (PerkinElmer). Libraries were first sequenced on a MiSeq system (2 × 25 cycles; Illumina) to evaluate quality (insert size, library diversity, etc.) and then underwent further WGS on either HiSeq 2500 system (2 × 125 cycles; Illumina) or HiSeq X system (2 × 150 cycles; Illumina) with ≥ 20% PhiX spike-in. The method was validated by sequencing four pairs of technical replicates and three pairs of matched biological replicates. Technical replicates were independent library

preparations made from the same oxBS-treated DNA sample. Biological replicates were three pairs of samples from different individuals, matched on age, sex, and library quality parameters.

SMRT sequencing and analysis

Dataset

We sequenced DNA isolated from whole blood samples from 50 individuals (29 females and 21 males) samples to average depth of 26.73x (range 12.74x-39.09x), on 189 flowcells. The earliest year of birth was 1941 and 1946, for females and males respectively and the latest was 1998 for both genders.

PacBio Sample Preparation

Samples were prepared and sequenced using either protocol A) (63 flowcells) or B) (189 flowcells) as described below.

A) HiFi SMRTbell® prep kit 2.0.

Genomic DNA (5 µg) diluted in Elution buffer (EB, 10 mM Tris, pH 8.5) was sheared to a target insert size of 15-20 kb using the MegaRuptor 3 system (Diagenode) with two successive shearing cycles at a speed setting of 31 and 32, respectively. Single stranded overhangs were removed using the DNA prep enzyme master mix by incubating the reaction mixture at 37 °C for 15 min, followed immediately by incubation with the DNA Damage Repair mix v2 at 37 °C for 30 min. End-repair/A-tailing was done by incubating the reaction mix with the End Prep Mix for 10 min at 37 °C, followed by 65 °C for 30 min. Finally, adapter ligation using Overhang adaptor v3, ligation mix, ligation additive and ligation enhancer was done by incubating the reaction mixture at 20 °C for at least 1 hour. The resulting SMRTbell libraries were purified using AMPure® PB beads at a 1.0X volume (beads:sample) and eluted in 15 µL of EB. Damaged SMRTbell templates were removed by nuclease treatment using the SMRTbell Enzyme Clean Up Mix (15 µL sample/55 uL mix) by incubating the reactions at 37 °C for 30 min followed immediately by AMPure® purification as described above. Size selection of the HiFi SMRTbell libraries was performed using the Blue Pippin system (Sage Science). Approximately 1.5 µg of library in a final volume of 30 µL per sample was loaded on each lane of the system followed by 10 µL of loading buffer. Samples were run using the 0.75% DF Marker S1 High-Pass 6-10 kb vs2 Cassette definition file with a run time of 4.5 hours and a selection mode of >10 kb. The collected samples were purified using AMPure® PB beads at a 1.0X volume as described above and eluted in EB in a final volume of 11 µL. Purified SMRTbell libraries were quantified using the dsDNA HS assay kit on the Qubit fluorometer and assessed for sizing using the Fragment

Analyzer 5300 (Agilent). Libraries were stored at -20 °C until further use. All steps in the workflow were performed using wide-bore pipette tips and LoBind (Eppendorf) tubes and/or strips.

B) HiFi SMRTbell® prep kit 3.0.

Genomic DNA (1 µg) was diluted in low TE buffer (10 mM Tris, pH 8.5, 0.1 mM EDTA) and sheared to a target insert length of 15-20 Kb using the MegaRuptor 3 at a shear speed of 31. Samples were purified using 1.0X volume ratio of SMRTbell clean up beads and eluted in 47 µL of low TE buffer. Repair and A-tailing was performed in a mixture of End repair mix and DNA repair mix (RM1) in a reaction volume of 60 µL at 37 °C for 30 min, followed by 5 min at 65 °C. Adapter ligation was done by adding the RM2 mix (SMRTbell adapter, ligation mix and ligation enhancer) to the samples in a final volume of 95 µL and incubating the mixture for 30 min at 20 °C, followed by 1X bead clean-up and elution in 40 µL of EB. Nuclease treatment was done using the RM3 mix by incubating the samples for 15 min at 37 °C. AMPure® PB bead size selection (<5 kb) was performed by pre-diluting the beads to 35% (vol/vol) with EB and using a 3.1X (vol/vol) of diluted beads to each sample. Final elution was done in 15 µL of EB. Quantity and quality of purifed SMRTbell libraries was done as described for method A.

PacBio Sequencing

Run designs were created in the SMRT Link software (v 10 or 11). SMRTbell libraries were bound to Sequel II polymerase 2.2. using either the Binding kit 2.2 or 3.2. Bound pol:DNA complex was purified using SMRTbell clean-up beads, quantified with Qubit and loaded on the Sequel® II sequencing plate 2.0 with on-plate loading concentrations ranging from 30-70 pM, predictive loading enabled and a maximum 2 hour loading time. Samples were sequenced using the SMRT®Cell 8M tray on the Sequel IIe system (HiFi application) with 30 hours movie time per SMRT cell and kinetic data acquisition enabled. Each sample was in general sequenced on 3-5 SMRT® cells depending on HiFi yield.

CpG methylation detection

We use Primrose for methylation detection of SMRT-sequencing. During sequencing the kinetic information, pulse width and duration are stored for each CpG. The 5mC signature the signal is quite diffused and not directly at the site of the modification but primarily few bases downstream. Therefore, SMRT sequencing uses “aggregate on intervals” technique, where the kinetic information is combined for neighbouring CpG sites, increasing the confidence in identifying the methylation at those sites (36). For every CpG in a read a feature vector is produced with the kinetics, pulse width and pulse duration for 16 bp intervals around each site on both strands. This feature vector is then fed into convolutional neural network (CNN) that outputs the probability of methylation for each CpG per read.

The CNN was trained on modified native human DNA (HG002), where fully methylated DNA was generated by treating the DNA with CpG methyltransferase M.SssI, and fully unmethylated DNA was generated using whole genome amplification (WGA). The accuracy increases with the number of passes per read. The methylation probabilities for each CpG per read is stored in a methyl tag in a bam file. We then use RefAlnBam-toModsBed-SAMTags.py script provided by PacBio to calculate the combined methylation per CpG and filter on minimum coverage 4x.

Statistical analysis

Per CpG average Pearson correlation (APC)

We calculate the average 5-mCpG rates per CpG or CpG unit over all individuals in the dataset. Then we evaluate the Pearson correlation coefficient of the per CpG averaged methylation predictions to the corresponding averaged oxBS methylation rates.

Defining a set of high quality CpG units

We assess the APC coefficient for CpGs that fall inside and outside any of the problematic regions, separately. We defined CpGs close to a variant as CpGs within 5 bp of any of 14,476,753 high-quality common variants. We define dark regions from 123 Illumina short-read sequenced samples, as regions where over 90% of the reads have mapping quality less than 10, coverage less than 5x and base quality less than 20 on average. We kept only regions at least 30 bp long. We define high coverage regions as regions that have over 1.5 times the average coverage in the dataset and low coverage regions as having less than 0.5 times the average coverage. We define strand bias as the difference in estimated 5-mCpG rates of forward and reverse strand. We further defined fraction of reliable reads as fraction of reads where the absolute log likelihood ratio exceeds defined cut-off as a fraction of the total number of available reads.

Statistical tests

Statistical tests were performed in R 3.6.0 (37). Correlation and confidence intervals were calculated using the cor.test() function and statistical difference between two distributions was evaluated using the non-parametric wilcox.test() function. Figures were created using ggplot2 (38).

Declarations

Competing interest

All authors are employees of deCODE genetics / Amgen, Inc.

Author's contributions

B.D.S, O.A.S. and B.V.H. designed the experiments. B.D.S. performed all statistical analysis and comparisons. F.Z. analysed the oxBS-seq data. N.K., D.N.M. and Ó.Þ.M. performed the sequencing. S.S., G.H., D.B. and B.D.S. designed the sequencing analysis pipeline. G.H. computed the method for detection of dark regions.

B.D.S. wrote the initial version of the manuscript and B.V.H., O.A.S., G.H., M.Þ.H. and Ó.Þ.M. contributed to the subsequent versions. B.V.H. and K.S. supervised the study. All authors reviewed and approved the final versions.

Availability of data and materials

Access to the raw Icelandic sequence data is available on request from Kári Stefánsson at the premises of deCODE genetics. The data are not publicly available because of Icelandic state law. All analyses in the manuscript can be recreated using the 5-mCpG rates summary statistics, which can be downloaded from our website: www.decode.com/summarydata/ (39).

Acknowledgements

The authors would like to thank our colleagues from deCODE genetics and Amgen Inc. for their helpful feedback. The authors would also want to thank all research participants who provided a biological sample to deCODE genetics and to the Genome in a Bottle Consortium.

Ethics approval and consent to participate

The study was approved by the National Bioethics Committee in Iceland (Approval no. VSN 14-015) and conducted in agreement with instructions issued by the Data Protection Authority in Iceland (PV_2017060950ÞS/--). All participating subjects signed informed consent. The personal identities of the participants and biological samples were encrypted by a third-party system approved and monitored by the Data Protection Authority. In addition, the study confirmed to the principles of the Helsinki declaration.

Additional File 1: Supplementary material

Supplementary notes, figures S1-S12 and data description.

Additional File 2: Supplementary tables

Supplementary Tables S1-S5.

References

1. Luo C, Hajkova P, Ecker JR. Dynamic DNA methylation: In the right place at the right time. Vol. 361, Science. 2018.

2. Kaluscha S, Domcke S, Wirbelauer C, Stadler MB, Durdu S, Burger L, et al. Evidence that direct inhibition of transcription factor binding is the prevailing mode of gene and repeat repression by DNA methylation. Nat Genet. 2022 Dec 1;54(12):1895–906.

3. Borgel J, Guibert S, Li Y, Chiba H, Schübeler D, Sasaki H, et al. Targets and dynamics of promoter DNA methylation during early mouse development. Nat Genet. 2010;42(12).

4. Butz S, Schmolka N, Karemaker ID, Villaseñor R, Schwarz I, Domcke S, et al. DNA sequence and chromatin modifiers cooperate to confer epigenetic bistability at imprinting control regions. Nat Genet. 2022;54(11).

5. Tost J, Gut IG. Analysis of gene-specific DNA methylation patterns by pyrosequencing technology. Methods Mol Biol. 2007;373.

6. Booth MJ, Branco MR, Ficz G, Oxley D, Krueger F, Reik W, et al. Quantitative sequencing of 5-methylcytosine and 5-hydroxymethylcytosine at single-base resolution. Science (1979). 2012 May 18;336(6083):934–7.

7. Noguera-Castells A, García-Prieto CA, Álvarez-Errico D, Esteller M. Validation of the new EPIC DNA methylation microarray (900K EPIC v2) for high-throughput profiling of the human DNA methylome. Epigenetics. 2023;18(1).

8. Wreczycka K, Gosdschan A, Yusuf D, Grüning B, Assenov Y, Akalin A. Strategies for analyzing bisulfite sequencing data. Vol. 261, Journal of Biotechnology. 2017.

9. Frommer M, McDonald LE, Millar DS, Collis CM, Watt F, Grigg GW, et al. A genomic sequencing protocol that yields a positive display of 5- methylcytosine residues in individual DNA strands. Proc Natl Acad Sci U S A. 1992;89(5).

10. Lister R, Mukamel EA, Nery JR, Urich M, Puddifoot CA, Johnson ND, et al. Global epigenomic reconfiguration during mammalian brain development. Science (1979). 2013;341(6146).

11. Skvortsova K, Zotenko E, Luu PL, Gould CM, Nair SS, Clark SJ, et al. Comprehensive evaluation of genome-wide 5-hydroxymethylcytosine profiling approaches in human DNA. Epigenetics Chromatin. 2017;10(1).

12. Booth MJ, Ost TWB, Beraldi D, Bell NM, Branco MR, Reik W, et al. Oxidative bisulfite sequencing of 5-methylcytosine and 5- hydroxymethylcytosine. Nat Protoc. 2013;8(10).

13. Logsdon GA, Vollger MR, Eichler EE. Long-read human genome sequencing and its applications. Vol. 21, Nature Reviews Genetics. 2020.

14. Mazid MA, Ward C, Luo Z, Liu C, Li Y, Lai Y, et al. Rolling back human pluripotent stem cells to an eight-cell embryo-like stage. Nature. 2022;605(7909).

15. Wang Y, Zhao Y, Bollas A, Wang Y, Au KF. Nanopore sequencing technology, bioinformatics and applications. Vol. 39, Nature Biotechnology. 2021.

16. Yuen ZWS, Srivastava A, Daniel R, McNevin D, Jack C, Eyras E. Systematic benchmarking of tools for CpG methylation detection from nanopore sequencing. Nat Commun. 2021;12(1).

17. Liu Y, Rosikiewicz W, Pan Z, Jillette N, Wang P, Taghbalout A, et al. DNA methylation-calling tools for Oxford Nanopore sequencing: a survey and human epigenome-wide evaluation. Genome Biol. 2021;22(1).

18. Stefansson OA, Sigurpalsdottir BD, Rognvaldsson S, Halldorsson GH, Juliusson K, Sveinbjornsson G, et al. The correlation between CpG methylation and gene expression is driven by sequence variants. [Unpublished manuscript].

19. Simpson JT, Workman RE, Zuzarte PC, David M, Dursi LJ, Timp W. Detecting DNA cytosine methylation using nanopore sequencing. Nat Methods. 2017 Feb 20;14(4):407–10.

20. Zink F, Magnusdottir DN, Magnusson OT, Walker NJ, Morris TJ, Sigurdsson A, et al. Insights into imprinting from parent-of-origin phased methylomes and transcriptomes. Nat Genet. 2018;50(11).

21. Schneider VA, Graves-Lindsay T, Howe K, Bouk N, Chen HC, Kitts PA, et al. Evaluation of GRCh38 and de novo haploid genome assemblies demonstrates the enduring quality of the reference assembly. Genome Res. 2017 May;27(5).

22. Ebbert MTW, Jensen TD, Jansen-West K, Sens JP, Reddy JS, Ridge PG, et al. Systematic analysis of dark and camouflaged genes reveals disease-relevant genes hiding in plain sight. Genome Biol. 2019;20(1).

23. Vu TH, Li T, Nguyen D, Nguyen BT, Yao XM, Hu JF, et al. Symmetric and asymmetric DNA methylation in the human IGF2-H19 imprinted region. Genomics. 2000;64(2).

24. Sun S, Li P. HMPL: A pipeline for identifying hemimethylation patterns by comparing two samples. Cancer Inform. 2015;14.

25. Beyter D, Ingimundardottir H, Oddsson A, Eggertsson HP, Bjornsson E, Jonsson H, et al. Long-read sequencing of 3,622 Icelanders provides insight into the role of structural variants in human diseases and other traits. Nat Genet. 2021;53(6).

26. Simpson JT, Workman RE, Zuzarte PC, David M, Dursi LJ, Timp W. Detecting DNA cytosine methylation using nanopore sequencing. Nat Methods. 2017 Feb 20;14(4):407–10.

27. Ni P, Huang N, Zhang Z, Wang DP, Liang F, Miao Y, et al. DeepSignal: Detecting DNA methylation state from Nanopore sequencing reads using deep-learning. Bioinformatics. 2019;35(22).

28. Liu Q, Fang L, Yu G, Wang D, Xiao C Le, Wang K. Detection of DNA base modifications by deep recurrent neural network on Oxford Nanopore sequencing data. Nat Commun. 2019;10(1).

29. Stoiber M, Quick J, Egan R, Eun Lee J, Celniker S, Neely R, et al. De novo Identification of DNA Modifications Enabled by Genome-Guided Nanopore Signal Processing. bioRxiv. 2016;

30. Oxford Nanopore Technologies: Megalodon. https://nanoporetech.github.io/megalodon (2019). Accessed 1 Nov 2023.

31. Ni P, Nie F, Zhong Z, Xu J, Huang N, Zhang J, et al. DNA 5-methylcytosine detection and methylation phasing using PacBio circular consensus sequencing. Nat Commun. 2023;14(1).

32. Razaghi R, Hook PW, Ou S, Schatz MC, Hansen KD, Jain M, et al. Modbamtools: Analysis of single-molecule epigenetic data for long-range profiling, heterogeneity, and clustering. bioRxiv. 2022;

33. Li H. Minimap2: Pairwise alignment for nucleotide sequences. Bioinformatics. 2018;34(18).

34. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009;25(16).

35. Wilks SS. The Large-Sample Distribution of the Likelihood Ratio for Testing Composite Hypotheses. The Annals of Mathematical Statistics. 1938;9(1).

36. Suzuki Y, Korlach J, Turner SW, Tsukahara T, Taniguchi J, Qu W, et al. AgIn: Measuring the landscape of CpG methylation of individual repetitive elements. Bioinformatics. 2016;32(19).

37. R Core Team. R Foundation for Statistical Computing. 2021. R: A language and environment for statistical computing.

38. Wickham H. ggplot2. Wiley Interdiscip Rev Comput Stat. 2011;3(2).

39. Sigurpalsdottir BD, Stefansson OA, Holley G, Beyter D, Zink F, Hardarson MÞ, Sverrisson SÞ, Kristinsdottir NK, Magnusdottir DN, Magnusson OÞ, Gudbjartsson DF, Halldorsson BV, Stefansson K. A comparison of methods for detecting DNA methylation from long-read sequencing of human genomes. Datasets. Zenodo. https://doi.org/10.5281/zenodo.10683994 (2024).

Figures

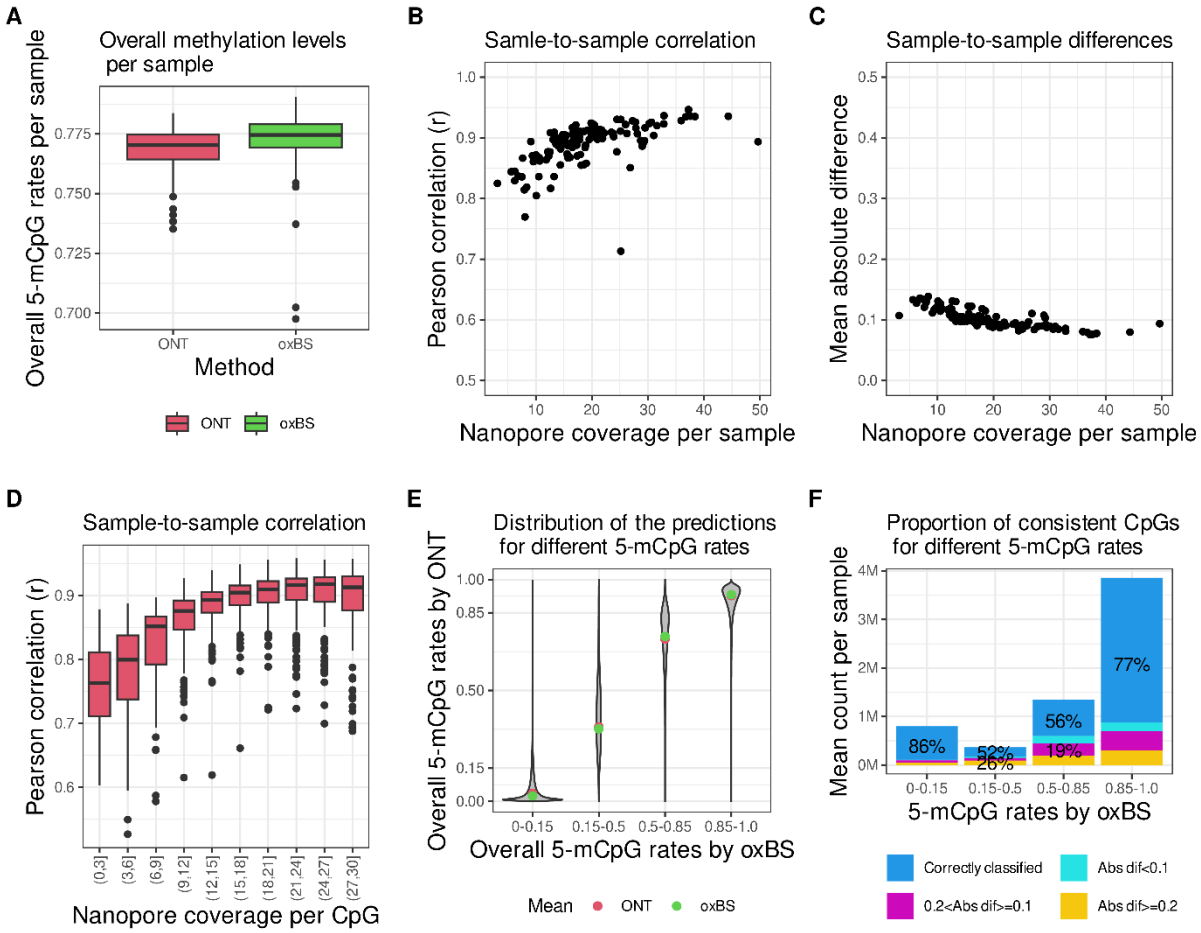


Fig1 Nanopore sequencing and oxBS performance in the same DNA samples. The consistency in 5-mCpG rates measured by nanopore sequencing and oxBS in DNA samples isolated from the same 132 individuals was estimated by: **A** The overall measurement of 5-mCpG rates in each of the 132 DNA samples measured by ONT (red) and oxBS (green), Y-axis is limited to (0.7,0.8). The centre line (solid black) shown in each box represent the median; the box limits represent upper and lower quartile, whiskers represent 1.5x interquartile range. **B** The Pearson *r* correlation coefficient, y-axis and **C** mean of the absolute differences in 5-mCpG rates of each CpG, y-axis, with respect to nanopore sequencing coverage in each sample on the x-axis. Panels D, E, F analyse sites that have >25x coverage in oxBS. **D** CpG coverage underlying the 5-mCpG rates, i.e. the number of sequences that were used to compute the 5-mCpG rate for a given CpG, in nanopore sequenced samples, x-axis, influences the consistency (Pearson *r*), y-axis, with 5-mCpG rates measured with high coverage by oxBS. The y-axis is limited to (0.5,1) **E** CpG rates in nanopore (y-axis) and oxBS (x-axis, binned). The mean is represented with red (ONT) and green (oxBS). **F** Number (y-axis, unit = million CpGs) of correctly classified (blue) by nanopore sequencing in a sample-to-sample comparison. Incorrectly classified CpGs are coloured according to the absolute difference in 5-mCpG rates (colour legend).

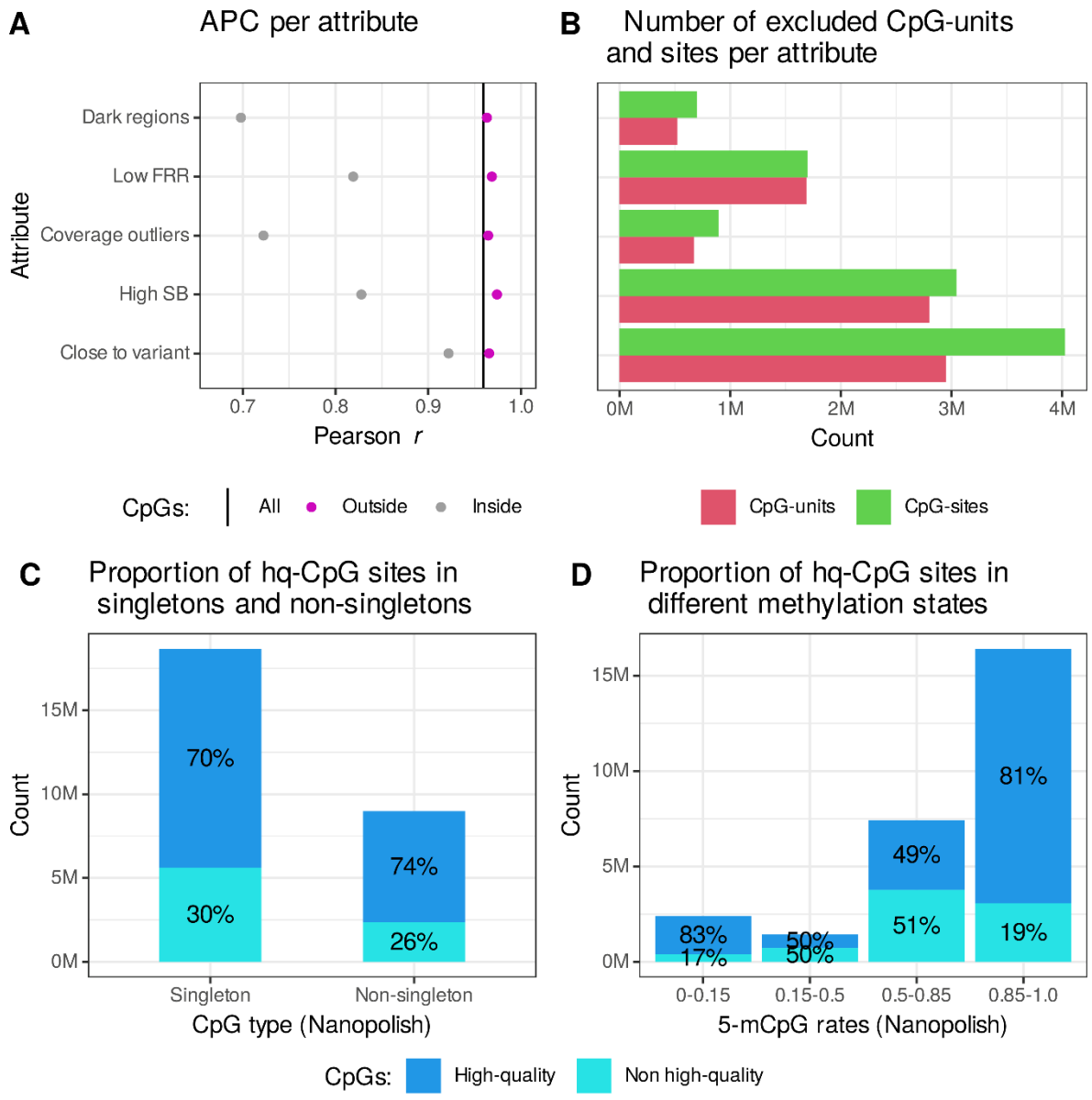


Fig. 2 The quality of 5-mCpG rate measurements by DNA sequence attributes. **A** APC estimates (x-axis), for CpG sites located outside (pink) and inside (grey) of DNA sequence attributes, y-axis, and the APC estimates based on all CpGs (vertical black line). **B** The number of CpG units (red) and sites (green), x-axis, found inside of each attribute, y-axis. **C** The proportion of high-quality (dark blue) and non high-quality (light blue) CpG units among singletons and non-singletons, x-axis. **D** The proportion of high-quality and non high-quality CpG units within each methylation state category, x-axis, defined by binning the mean of 5-mCpG rates measured by Nanopolish.

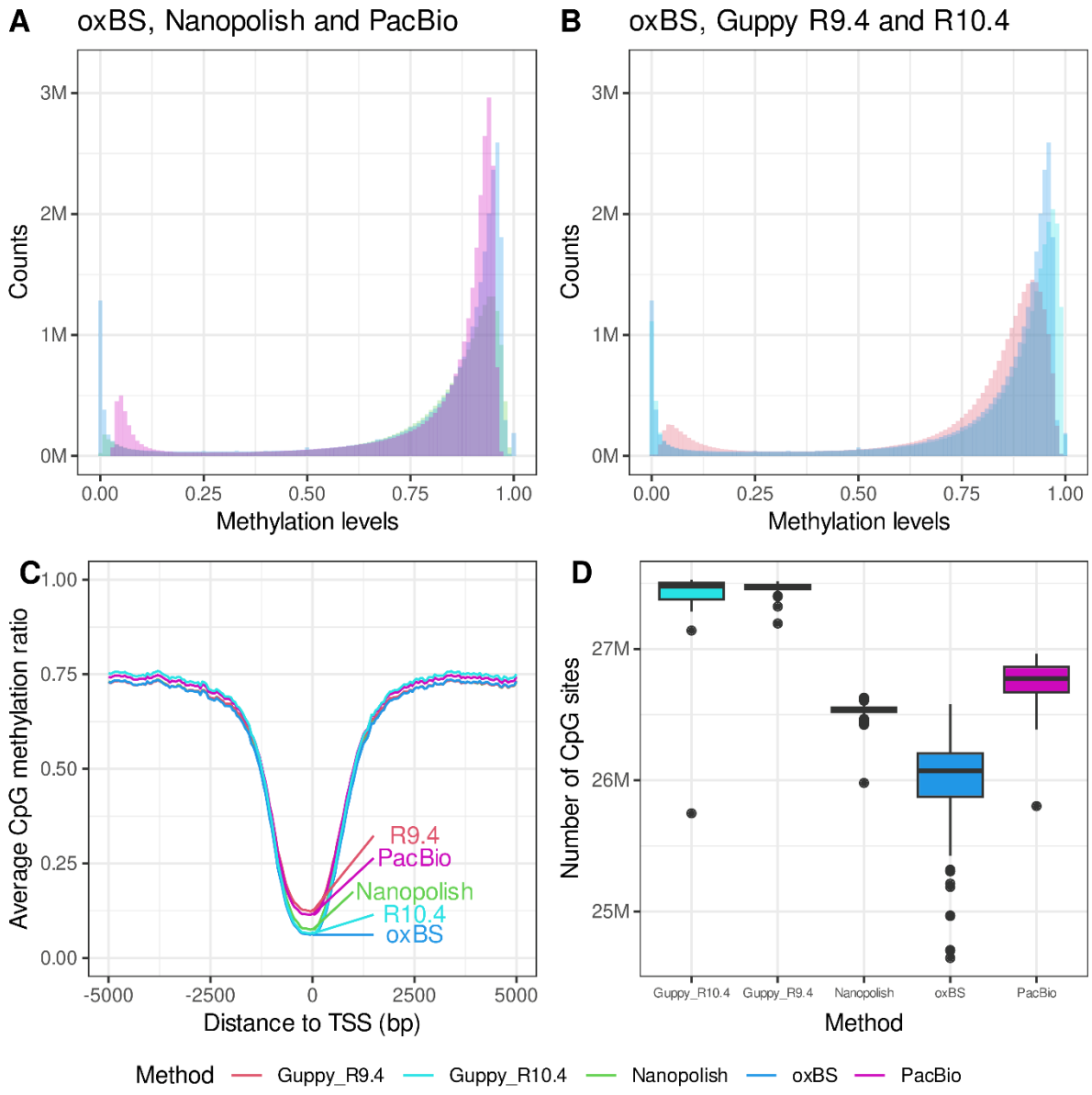


Figure 3. Fig. 3 Comparison of CpG methylation detection by method. CpG methylation rates (ranging from 0-1) averaged across individuals yields the expected bimodal distribution seen in oxBS data for **A** oxBS, Guppy R9.4 and R10.4 and **B** oxBS, PacBio and Nanopore. The units on y-axis are millions (M). **C** CpG methylation rates averaged in 50 bp bins relative to transcription start sites (TSSs) of genes expressed in whole blood. **D** Number of CpGs called by each method. For Nanopolish we count all CpGs within a CpG unit. Note, the y-axis is limited from 24.5 to 27.7 M (millions). The centre line (solid black) shown in each box represent the median; the box limits represent upper and lower quartile, whiskers represent 1.5x interquartile range.

Paper II

ARTICLE

Title:

The correlation between CpG methylation and gene expression is driven by sequence variants

Author list:

Olafur Andri Stefansson[1]*, Brynja Dogg Sigurpalsdottir[1,2], Solvi Rognvaldsson[1], Gisli Hreinn Halldorsson[1,3], Kristinn Juliusson[1], Gardar Sveinbjornsson[1], Bjarni Gunnarsson[1], Doruk Beyter[1], Hakon Jonsson[1], Sigurjon Axel Gudjonsson[1], Thorunn Asta Olafsdottir[1,4], Saedis Saevarsdottir[1,4], Magnus Karl Magnusson[1,4], Sigrun Helga Lund[1,3], Vinicius Tragante[1], Asmundur Oddsson[1], Marteinn Thor Hardarson[1,2], Hannes Petur Eggertsson[1], Reynir L. Gudmundsson[1], Sverrir Sverrisson[1], Michael L. Frigge[1], Florian Zink[1], Hilma Holm[1], Hreinn Stefansson[1], Thorunn Rafnar[1], Ingileif Jonsdottir[1,4], Patrick Sulem[1], Agnar Helgason[1,5], Daniel F. Gudbjartsson[1,3], Bjarni V. Halldorsson[1,2], Unnur Thorsteinsdottir[1,4], Kari Stefansson[1,4]*

Affiliations:

1. deCODE genetics / Amgen Inc., Sturlugata 8, Reykjavik, Iceland
2. School of Technology, Reykjavik University, Reykjavik, Iceland
3. School of Engineering and Natural Sciences, University of Iceland, Reykjavik, Iceland
4. Faculty of Medicine, School of Health Sciences, University of Iceland, Reykjavik, Iceland
5. Department of Anthropology, University of Iceland, Reykjavik, Iceland

*Correspondence to: Olafur Andri Stefansson (olafurs@decode.is), Kari Stefansson (kstefans@decode.is)

Abstract

Gene promoter and enhancer sequences are bound by transcription factors (TFs) and depleted of methylated CpG sites. The absence of methylated CpGs in these sequences typically correlates with increased gene expression, indicating a regulatory role for methylation. We used nanopore sequencing to determine haplotype-specific methylation rates of 15.3 million CpG units in 7,179 whole blood genomes. We identified 189,178 methylation depleted sequences (MDSs) where three or more proximal CpGs were unmethylated on at least one haplotype. 77,789 MDSs (~41%) associated with 80,503 *cis*-acting sequence variants which we termed allele-specific methylation QTLs (ASM-QTLs). RNA sequencing of 896 samples from the same blood draws used to perform nanopore sequencing, showed that the ASM-QTL i.e., DNA sequence variability, drives most of the correlation found between gene expression and CpG methylation. ASM-QTLs were enriched 46.4-fold (95%CI:36.0,58.7) among sequence variants associating with hematological traits, demonstrating that ASM-QTLs are important functional units in the non-coding genome.

Introduction

Cytosines preceding guanines in DNA (CpG dinucleotides) are predominantly modified by the addition of a methyl group in humans and other vertebrates[1]. This modification is commonly known as CpG methylation, or 5-mCpG, where the number five indicates the position of the methyl group on the carbon ring of cytosine. DNA methyltransferases (DNMTs) are responsible for adding methyl groups to CpGs[2]. DNMTs are essential in mice as their deletion results in embryonic death[3], but these embryos are nonetheless able to form all major cell types[4].

DNA regulatory sequences, e.g. promoters and enhancers, are frequently bound by TFs[5,6] and depleted of CpG methylation[7,8]. Promoter sequences are frequently found within so-called CpG islands[9] containing high density of CpGs that are typically unmethylated[10] likely because of TF binding counteracting DNMTs[11,12,13,14,15,16]. DNA sequence variability in TF binding sites represents a plausible mechanism by which variation in CpG methylation arises[16,17,18,19].

There are TFs that bind to CpG-containing motifs in DNA and some (but not all) of these TFs are influenced by CpG methylation in their binding site[20,21]. CpG methylation, by influencing the binding sites of some TFs, may therefore be involved in regulating gene expression[22,23], repeat repression[23] and imprinting[24]. CpG methylation is nonetheless dynamically modified as a consequence of protein (e.g. TF) binding to DNA[8,14,25,26,27]. Hence, CpG methylation is highly malleable by TFs and, for this reason, CpG methylation status is not necessarily the driving force of correlation between CpG methylation and gene expression[28,29,30].

In this study we searched for *cis*-acting influences of sequence variants on CpG methylation and explored the relevance of this DNA sequence variability to correlations observed between CpG methylation and gene expression.

Results

CpG methylation measured by nanopore sequencing

We performed whole genome sequencing using nanopore technology in whole blood samples from 7,179 individuals to at least 10x coverage (mean 20.6x; range: 10-108.3x) on 8.906 PromethION flowcells, each sequenced to at least 3x coverage (mean 16.6x, range 3-39x) (Supplementary Fig. 1). 5-mCpG detection was performed on autosomes using Nanopolish[31]. CpGs were measured as units by Nanopolish if they were located within 10bp of each other and, consequently we referred to them as „CpG units". Most (83.6%) of these CpG units were „singletons", i.e. represented by a single CpG site, but 16.4% of CpG units were represented by two or more CpG sites. The number of CpG units detected by Nanopolish was 22,058,476 which corresponds to 26,487,587 CpG sites in the reference genome (GRCh38).

For each CpG unit and each parental haplotype, we defined the 5-mCpG rate as the number of sequences of that parental haplotype that were methylated at the CpG unit divided by the total number of sequences covering the CpG unit on the same parental haplotype. By comparing 5-mCpG rates, as measured by nanopore sequencing, to those obtained from a subset of 132 DNA samples analyzed with oxidative bisulfite sequencing in our previous study[32], we showed that 15.3 million CpG units were reliably detected by Nanopolish (Supplementary Figs. 2 and 3; Supplementary Notes 1.1 and 1.2), and we confined this study to those units.

5-mCpG rates showed a bimodal distribution (**Fig. 1a**) as expected given previous studies on 5-mC content in DNA and whole methylome sequencing[1].

In this study, we classified CpG units in each individual as unmethylated (5-mCpG rate <0.15) or low-methylated (0.15≤ 5-mCpG rate <0.50). For the individual, the proportion of CpG units classified as low-methylated was on average 4.2% (Quartiles:3.8,4.5%) and the proportion classified as unmethylated was 7.2% (Quartiles:6.5,7.6%).

We further confirmed the well-documented lack of 5-mCpGs in functional regions mapped by Encode[33] and others[34,35] (**Fig. 1b;** Supplementary Fig. 4).

Sequence variants associated with 5-mCpG rates

We imputed genotypes based on whole-genome sequences of 63,460 Icelanders[36] into the 7,179 nanopore sequenced individuals. A total of 34,435,950 high-quality sequence variants with MAF>10^{-}

[4] in the sample set were located 100kb up- or downstream of the CpG units; 23,752,296 SNPs, 5,929,255 indels, 609,536 structural variants[37,38] and 4,144,863 microsatellites[39].

We searched for *cis*-acting sequence variants (100kb up- or downstream of CpG units) associated with 5-mCpG rates over each of the 15.3 million CpG units. Each CpG unit was tested, on average, against ~2,200 sequence variants yielding $3.4 \cdot 10^{10}$ tests, and we used Bonferroni correction to set the threshold for significances at $P < 0.05/3.4 \cdot 10^{10} \sim 10^{-12}$.

A total of 1,625,423 CpG units associated with 1,023,970 sequence variants in a total of 1,669,151 associations; 704,474 SNPs, 205,026 indels, 106,743 microsatellites and 6,727 SVs. 263,403 (25.7%) of these sequence variants associated with more than one CpG unit. Out of the 1,625,423 associated CpG units, 43,728 (~2.7%) were associated with more than one sequence variant.

The majority (73.4%) of sequence variants identified in association with CpG methylation in an external cohort[40] were replicated in our cohort (Supplementary Note 1.3).

5-mCpG depleted sequences

As DNA sequences depleted of 5-mCpGs are indicative of function we searched for haplotypes where 5-mCpG rates were low across closely located CpG units in one of the two haplotypes of at least one individual (**Fig. 1c**). For accuracy, we confined this analysis to methylomes sequenced to an average coverage of >20x (n=2,648 individuals) and CpG units where ≥10 sequences were available for estimating their 5-mCpG haplotype rate. We defined unmethylated haplotypes as those with three or more CpG units each with 5-mCpG rate <0.15, but located no more than 500bp apart. Many of the unmethylated haplotypes found in different individuals had the exact same coordinates, or were found in overlap. We therefore defined clusters of overlapping unmethylated haplotypes. For each cluster, we catalogued the genome coordinates of the most frequently occurring unmethylated haplotype (**Extended Data Fig. 1a**) and removed each of the unmethylated haplotypes that overlapped with these defined coordinates. If any remained, we repeated this procedure until there were no remaining unmethylated haplotypes in the cluster (**Extended Data Fig. 1b**). Low-methylated haplotypes were analogously defined, but with 5-mCpG rate <0.50 and only in sequences where unmethylated haplotypes were not found.

Our algorithm identified 84,924 unmethylated and 104,254 low-methylated haplotypes, hereafter referred to as *methylation depleted sequences* (MDSs).

Collectively, the 189,178 MDSs covered ~83Mb of the genome and consisted of 1.2 million high-quality CpG units. MDSs were, on average, 440bp (Quartiles:153,512bp), and the median number of CpG units that defined each MDS was 3 (Quartiles:3,4).

Sequence variants influence the 5-mCpG rates of MDSs

We searched for *cis*-acting sequence variants (100kb up- or downstream of MDSs) associated with 5-mCpG rates over each of the 189,178 MDSs. On average, each MDS was tested against ~2,400 sequence variants yielding $4.5 \cdot 10^8$ tests, and we set a Bonferroni corrected significance threshold at $P < 0.05/4.5 \cdot 10^8 \sim 10^{-10}$.

A total of 77,789 MDSs associated with 80,503 sequence variants in a total of 86,252 associations; 58,892 SNPs, 11,306 indels, 8,040 microsatellites and 2,265 SVs. 3,760 (4.7%) of these sequence variants associated with more than one MDS.

We refer to these 80,503 sequence variants hereafter as „allele-specific methylation quantitative trait loci“ (ASM-QTL). Out of the 77,789 associated MDSs, 8,513 (~11%) were associated with more than one ASM-QTL. 71,868 (89.3%), out of 80,503 ASM-QTLs, were in strong linkage disequilibrium ($r^2 > 0.80$) to at least one of the ~1 million sequence variants found in association with 5-mCpG rates of individual CpG units.

The median distance from ASM-QTLs to the center of their associated MDS was 3.1kb (Quartiles: 0.2,16kb). Most ASM-QTLs were common: 76,154 with minor allele frequency (MAF) >1% likely because of lack of power to detect associations with rare variants.

For validation, we used whole-blood derived DNA samples previously analyzed by oxidative bisulfite sequencing[40], but restricted to the forty-five individuals that were not included in the larger cohort used for identifying ASM-QTLs. We were able to evaluate 57,273 (out of 86,252) ASM-QTLs for association with 5-mCpG rates of the corresponding MDSs in the independent cohort of 45 individuals. Our results showed that most (89.7%; 95%CI:89.5,90%) of the tested ASM-QTLs were consistent in effect size in the validation cohort, and the effect sizes were strongly correlated (Pearson’s r=0.679; 95%CI:0.675,0.684) (**Extended Data Figure 2**).

Correlations between CpG methylation and mRNA expression

We performed RNA sequencing (polyA) of 896 whole blood samples used for nanopore sequencing to analyze the effect of 5-mCpG on gene expression. The same blood samples were used to measure cellular composition and to isolate DNA and RNA for both nanopore- and RNA sequencing, respectively. RNA sequences were assigned to parental haplotypes based on phase informative alleles in RNA sequence fragments. As there can be alternative TSSs for the same gene, we performed the quantification per mRNA isoform.

We searched for associations between haplotype specific measures of 5-mCpG rates of MDSs and the haplotype specific mRNA isoform expression of genes located 100kb up- or downstream of each MDS. In these analyses, haplotype specific mRNA isoform expression was represented as the

proportion of mRNA sequences expressed from the paternal- and maternal haplotype. 83,963 out of the 189,178 MDSs were located within 100kb from TSSs of 18,923 mRNA isoforms (9,603 genes) expressed in this collection of whole blood samples. On average, we tested each of these MDSs against ~4 mRNA isoforms (Quartiles:2,6) leading to ~380 thousand tests and used Bonferroni correction to set the threshold for significance at $P < 0.05/0.38 \cdot 10^6 \sim 1.3 \cdot 10^{-7}$.

We found 1,103 mRNA isoforms (derived from 773 genes) in association with 957 MDSs in a total of 1,513 associations. The median distance between MDSs and the TSS of the associated mRNA isoform was 23.8kb (Quartiles:9,47kb). Most of the associated MDSs (921; ~96%) did not include the TSS of an associated mRNA isoform (**Extended Data Figure 3a**), but 36 (~4%) contained the TSS of an associated mRNA isoform, also known as promoter methylation (**Extended Data Figure 3b**).

None of these 957 MDSs overlapped with any of the previously known regions where 5-mCpG rates differ between parental chromosomes[41], also known as imprinted regions, which was expected as our models account for parent of origin.

ASM-QTLs correspond to TF binding sites

The frequency of ASM-QTLs was 3.3-fold higher than expected ($P=8 \cdot 10^{-19}$) among sequence variants previously found to influence allele specific binding (ASB) of various proteins to DNA by Chen et al[42]. This same database[42] has six proteins associated with >100 sequence variants: SPI1, CTCF, STAG1, EBF1, POLR2A and POL2RB. ASM-QTLs were more frequently ($P < 0.05/6 \sim 0.008$) found among ASB variants for the three TFs (SP1, CTCF, and EBF1), and the Cohesin Complex subunit STAG1 (**Extended Data Figure 4A**).

ASM-QTLs were also more prevalent than expected among sequence variants located within regulatory elements as defined by the ENCODE project, notably within those bound by the CTCF protein (**Extended Data Figure 4B**).

In accordance with previous studies[16,17,18,19], these results support the notion that sequence variants influence methylation of CpGs through their influences on protein binding to DNA.

ASM-QTLs dominate in correlations between MDSs and mRNA

All of the 957 MDSs that associated with mRNA expression were associated with an ASM-QTL (100%; 95%CI:99.6%,100%). In comparison, a significantly lower proportion, i.e. 40.8% (95%CI:40.6%,41%) of MDSs that did not associate with mRNA expression were associated with an ASM-QTL ($\chi_1^2=148.4$; $P=4 \cdot 10^{-37}$). This correspondence between MDSs associated with mRNA expression and MDSs influenced by an ASM-QTL persists irrespective of the variance in 5-mCpG rates of MDSs (**Fig. 2a**).

For a given MDS found in association with mRNA expression, the fraction of the variance in mRNA expression explained by the associated ASM-QTL tends to be higher (Median=0.24; Quartiles:0.13,0.41) than that explained by the 5-mCpG rate of the MDS (Median=0.097; Quartiles:0.06,0.17) (**Fig. 2b**). Further, the fraction of the variance in mRNA expression explained by the 5-mCpG rate of a given MDS is largely nullified by accounting for the effects of the ASM-QTL associated with the same MDS (Median=0.001; Quartiles:0.0002,0.004) (**Fig. 2b**).

For example, the 5-mCpG rate of the MDS located within the CpG island promoter sequence of *VAMP5* associated with mRNA expression of the main isoform (VAMP5-201) initiated from within that same CpG island (**Extended Data Figure 5a**). The 5-mCpG rate of the MDS explains 23.7% of the variance in mRNA expression of *VAMP5-201*, whereas the ASM-QTL (1bp deletion at chr2:85580659:AT:T), associating with this same MDS, explains 35.9% of the variance in mRNA expression of VAMP5-201. After correcting the 5-mCpG rates of the MDS for the sequence variant (AT>T deletion), the fraction of the variance in mRNA expression explained by the 5-mCpG rate was only 2.5% (**Extended Data Figure 5b**) suggesting that the correlation found between the CpG island promoter methylation and expression of the VAMP5-201 gene is mostly driven by DNA sequence variability (**Extended Data Figure 5c&d**).

The variability in mRNA expression explained by the ASM-QTL genotype status corrected for 5-mCpG rates (ASM-QTL|5-mCpG) (Median=0.09; Quartiles:0.03,0.21) was only moderately lower than that explained by the uncorrected ASM-QTL genotype status (**Fig. 2b**). Similarly, the variability in 5-mCpG rates explained by the ASM-QTL genotype status corrected for mRNA expression (ASM-QTL|mRNA) (Median=0.25; Quartiles:0.11,0.45) was only moderately lower than that explained by the uncorrected ASM-QTL genotype status (**Fig. 2c**).

These results indicate that sequence variants (ASM-QTLs) are responsible for creating most of the variability in CpG methylation that correlates with gene expression.

Modeling the impact of ASM-QTL on methylation and expression

We considered four different models (**Extended Data Figure 6**) to infer the mechanism by which ASM-QTL variants affect CpG methylation and gene expression.

We used the Mendelian Randomization-Steiger test[43] to infer the direction of effect between the 5-mCpG rates of MDSs and mRNA expression; i.e., whether 5-mCpGs are affecting mRNA levels or *vice versa*. Assuming equal measurement error, we found 5-mCpG rates more likely to be affecting mRNA expression for 68% of ASM-QTLs that were nominally associated with both 5-mCpG rates and mRNA expression among individuals with both measurements (P=$3 \cdot 10^{-41}$). The measurement error of mRNA expression would have needed to be 22% greater than that of 5-mCpG rates for us to have observed this result if half of the ASM-QTLs would have supported the conclusion that 5-mCpG

rates are more likely than mRNA expression to be causal. Therefore the Mendelian Randomization-Steiger test provides evidence against Model 4 (**Extended Data Figure 6**) which states that ASM-QTLs affect 5-mCpG rates through their influences on mRNA expression, but provides support for Model 2 (**Extended Data Figure 6**) which states that ASM-QTLs affect mRNA expression through their influences on 5-mCpG rates.

The Mendelian Randomization-Steiger test, however, does not consider horizontal pleiotropy as is present in Models 1 and 3 (**Extended Data Figure 6**). Model 1 states that the ASM-QTL affects the 5-mCpG rate and mRNA expression partially through a common mechanism (e.g. TF binding). Model 3, however, states that the genotype (G) of an ASM-QTL affects the 5-mCpG rate (M) and mRNA expression level (E) through independent mechanisms. Under Model 3, the product of the variance in M and the coefficient from regressing E on M ($\hat{\sigma}_M^2\hat{\beta}_{ME}$) would be expected to be equal to the product of the coefficient from regressing M on G ($\hat{\beta}_{GM}$) and the coefficient from regressing E on G ($\hat{\beta}_{GE}$). However, we found that $\hat{\sigma}_M^2\hat{\beta}_{ME}$ is on average 10% greater (95%CI:9,12%) than $\hat{\beta}_{GM}\hat{\beta}_{GE}$ (**Fig. 2d**), providing evidence against Model 3 (**Extended Data Figure 6**). The differences between $\hat{\beta}_{GM}\hat{\beta}_{GE}$ and $\hat{\sigma}_M^2\hat{\beta}_{ME}$ are nonetheless small which, again, indicates that most of the correlation between 5-mCpG rates and mRNA expression is driven by sequence variants.

Our method of comparing $\hat{\beta}_{GM}\hat{\beta}_{GE}$ and $\hat{\sigma}_M^2\hat{\beta}_{ME}$ is unable to distinguish between the two remaining Models 1 and 2, and the Mendelian Randomization-Steiger test does not consider Models 1 and 3 but provides support for Model 2 over Model 4 (**Extended Data Figure 6**).

Our results are therefore equally consistent with both Models 1 and 2, which we illustrate using a hypothetical example in **Figure 3**.

ASM-QTL enrichment among trait associated sequence variants

We have previously performed genome-wide association studies (GWAS) on a large number of human diseases and other traits in the Icelandic population[36,44,45]. GWAS signals identified in these studies allowed us to quantify the contribution of ASM-QTLs to human phenotypic diversity relative to other types of sequence annotations in the same model. We searched for and identified 5,071 GWAS signals (P <1·10^{-9}) in a selected list of 261 diverse traits; 60 diseases and 201 other traits).

ASM-QTLs were 23.2-fold enriched among GWAS signals (95%CI:18.6,28.4; P <0.0001; **Extended Data Figure 7**; Supplementary Note 1.4). For comparison, in this same model, we also specified an annotation of sequence variants that resided within the MDS coordinates, which yielded 4.2-fold enrichment (95%CI:3.2,5.3; P <0.0001) showing that location alone did not capture all the effects of ASM-QTLs. We noted that ASM-QTLs were more enriched among GWAS signals than any other non-coding annotation such as sequence variants located in DNase hypersensitivity footprints[6] (**Extended Data Figure 7**). Only protein coding variants were more enriched than ASM-QTLs, i.e.

missense (98.7-fold; 95%CI:80.2,118.5; P <0.0001) and loss of function variants (292-fold; 95%CI:178,426; P <0.0001).

The enrichment of ASM-QTLs varied according to the classification of GWAS signals into trait groups (**Fig. 4**). Notably, our ASM-QTLs were 46.4-fold (95%CI:36.0,58.7; P <0.0001) enriched among 2,394 GWAS signals found in association with 56 hematological traits (out of 261 traits) which was higher than the overall 23.2-fold enrichment (P <0.05) (**Fig. 4**). In contrast, ASM-QTLs were 6.6-fold (95%CI:3.9,9.7; P <0.0001) enriched among the remaining 2,677 GWAS signals found in 205 non-hematological traits. This demonstrates that the ASM-QTLs identified here have more effects on hematological traits than on other traits, probably because of tissue-specificity in CpG methylation as our measurements were done using DNA from whole blood samples.

ASM-QTLs found in linkage disequilibrium with sequence variants that associated with gene expression in our in-house RNA-seq data[46], also known as *cis*-expression QTLs, were enriched 69.8-fold (95%CI:50.6,90.8; P <0.0001) among GWAS signals. Furthermore, ASM-QTLs that were not found in linkage disequilibrium with *cis*-expression QTLs were enriched 16.8-fold (95%CI:12.5,21.6; P <0.0001). Hence, ASM-QTLs were enriched among sequence variants relevant to trait diversity irrespective of whether they were also identified as *cis*-expression QTLs. We postulate that the ASM-QTLs that did not appear to influence gene expression, may still be functionally important because their impact on expression may be context-specific, e.g. only relevant after specific stimuli or environmental cues.

The ~1 million sequence variants found in association with 5-mCpG rates of individual CpG units were also enriched among GWAS signals (7.3-fold, 95%CI:6.3,8.3; P <0.0001), which was significantly lower than the corresponding 23.2-fold enrichment for ASM-QTLs associated with 5-mCpG rates of MDSs (P <0.05). These sequence variants were enriched 10.4-fold (95%CI:8.7,12.5) and 4.7-fold (95%CI:3.9-5.7) among GWA signals associated with hematological and non-hematological traits, respectively. Thus, sequence variants associated with 5-mCpG rates of MDSs have greater functional relevance than those associated with individual CpG units, showing that measurements of 5-mCpG rates within MDSs are highly informative of functional activity in DNA and relevant to human trait diversity.

ASM-QTLs correspond to disease associated sequence variants

Previous studies have found that sequence variants associated with human traits overlap those associated with CpG methylation[47,48]. Our results show that 964 and 4,391 ASM-QTLs were in strong linkage disequilibrium (r^2>0.80) with sequence variants associated with 152 diseases and 431 other traits, respectively, based on published GWASs[49].

For an example, on chromosome 2q33.3, rs34329895, associating with type-II diabetes[50], is in strong linkage disequilibrium (r^2=0.97) with the ASM-QTL rs35735821 found in association with promoter methylation of the *PLEKHM3* gene. In this example, the disease-associated variant is not found in linkage disequilibrium to a protein coding variant or any *cis*-expression QTLs. Here, ASM-QTLs offer a valuable complement to other sequence variant annotations in identifying candidate gene targets.

Another example on chromosome 10p15.1, rs12722502, associating with asthma[51], was found in nearly perfect linkage disequilibrium (r^2=0.997) with ASM-QTL rs12722547 which associated with 5-mCpG rates of an MDS residing within a cCRE enhancer element, intronic of *IL2RA*. Neither the ASM-QTL, nor the disease associated sequence variant were in linkage disequilibrium to any protein coding variants. Further, in our RNA-seq datasets, rs12722502 was not found in strong linkage disequilibrium (r2 > 0.8) with any conditionally independent *cis*-expression QTLs in RNA isolated from whole-blood samples from ~17.8 thousand individuals[46]. Whereas neither RNA nor protein coding annotations provided clues to a functional consequence, the ASM-QTL points to a candidate regulatory element, which can then be investigated further in experimental models.

Discussion

In this study, we assigned CpG methylation, gene expression and alleles of sequence variants to parental haplotypes, allowing us to investigate correlations between the three sets of measurements on a haplotype level. We used these data to identify MDSs and found that in instances where their CpG methylation correlated with gene expression, a sequence variant was invariably found in association with the CpG methylation of those same MDSs that explained most of the correlation. Hence, in instances where CpG methylation is found in association with both a sequence variant and gene expression, it is important to be cautious about assuming that the sequence variant influences the gene expression through CpG methylation. Indeed, our results are consistent with a model wherein the correlations found between CpG methylation and gene expression are mostly by-products of variability in TF protein binding to DNA created by sequence variants. In this model TFs, but not the CpG methylation, are responsible for influencing gene expression. Nonetheless, it remains that our results are equally consistent with a model wherein the sequence variant exerts its influences on mRNA expression by affecting CpG methylation. In both models, however, the sequence variant is the primary driver of the correlation between CpG methylation and gene expression.

We show that sequence variants found in association with variation in CpG methylation rates of MDSs have substantial effects on human phenotypic diversity. Previous studies have described haplotype specific influences of sequence variants on CpG methylation[16,17]. However, due to small sample sizes they lacked power to carry out association analyses designed to detect consistent allele-

specific influences across individuals, and as such were not well suited to evaluate the effect of these sequence variants on human phenotype diversity.

Limitations of our study include variability in cell type composition which we accounted for by using a statistical technique, singular value decomposition, while also using information on direct measurements of cell type composition available in a subset of our cohort. We note, however, that both methods are limited in resolution of specific sub-populations of blood cell types and therefore incompletely account for cell type composition. We would also like to note that additional insights may be gained into the relevance of CpG methylation to human diseases, and other traits, by using genetic colocalization methods[52] instead of the linkage disequilibrium approach used here. Also, the mapping of *cis*-expression QTLs is an ongoing effort involving contributions from numerous entities around the world, which therefore complicates our ability to declare whether or not a sequence variant has been identified as a *cis*-expression QTL, as the field is continuously evolving. Finally, as we accounted for parent of origin in our models, our results do not apply to gene imprinting wherein 5-mCpGs may have important regulatory roles[53].

Nanopore sequencing provides accurate detection of CpG methylation in DNA samples and has the benefit of achieving long sequences which facilitate phasing of the sequences to parental chromosomes to yield haplotype resolved methylomes (Supplementary Note 1.5). Measurement of CpG methylation in DNA samples allows for evaluation of protein binding and the effect of sequence variants on protein binding. CpG methylation has many properties that make it more suitable than other chromatin-based assays for functional annotation of the non-coding genome. This includes its chemical stability and its measurement accuracy, which ensures comparability between samples from different individuals. We expect nanopore sequencing of genomes from various cell types and tissues will be instrumental to investigate non-coding sequence variants of functional relevance.

Acknowledgements

The authors would like to thank colleagues at deCODE genetics for helpful discussions and feedback. The authors received no specific funding for this work.

Author contributions statement

Paper was written by OAS and KS with input from UT, BVH, DFG, BDS, SR, KJ, GS, DB, HJ, SAG, MTH, HPE, RLG, SSae, MLF, FZ, AH. Data pipeline was set up and managed by SSv. Analysis and detection of 5-mCpGs in nanopore sequenced DNA was performed by BDS and BVH. Analysis of 5-mCpG rate measurements based on nanopore sequenced DNA was performed by OAS, BDS, SR, BG, DFG and BVH. Phasing of oxBS sequenced DNA was performed by F.Z. and analyzed for validation of nanopore results by OAS. RNA sequencing data was processed by GHH and analyzed by OAS and GHH. Phenotypes were defined by PS, IJ, TR, HS, HH. OAS and SHL selected main traits and OAS,

TAO, SSae, MKM, VT, AO classified GWA signals into trait groups. Analysis of enrichment among GWA signals was performed by OAS, SR and DFG. Study was supervised by OAS and KS. All authors agreed to the final version of the manuscript.

Competing interests statement

All authors are employees of deCODE genetics/Amgen.

Figures

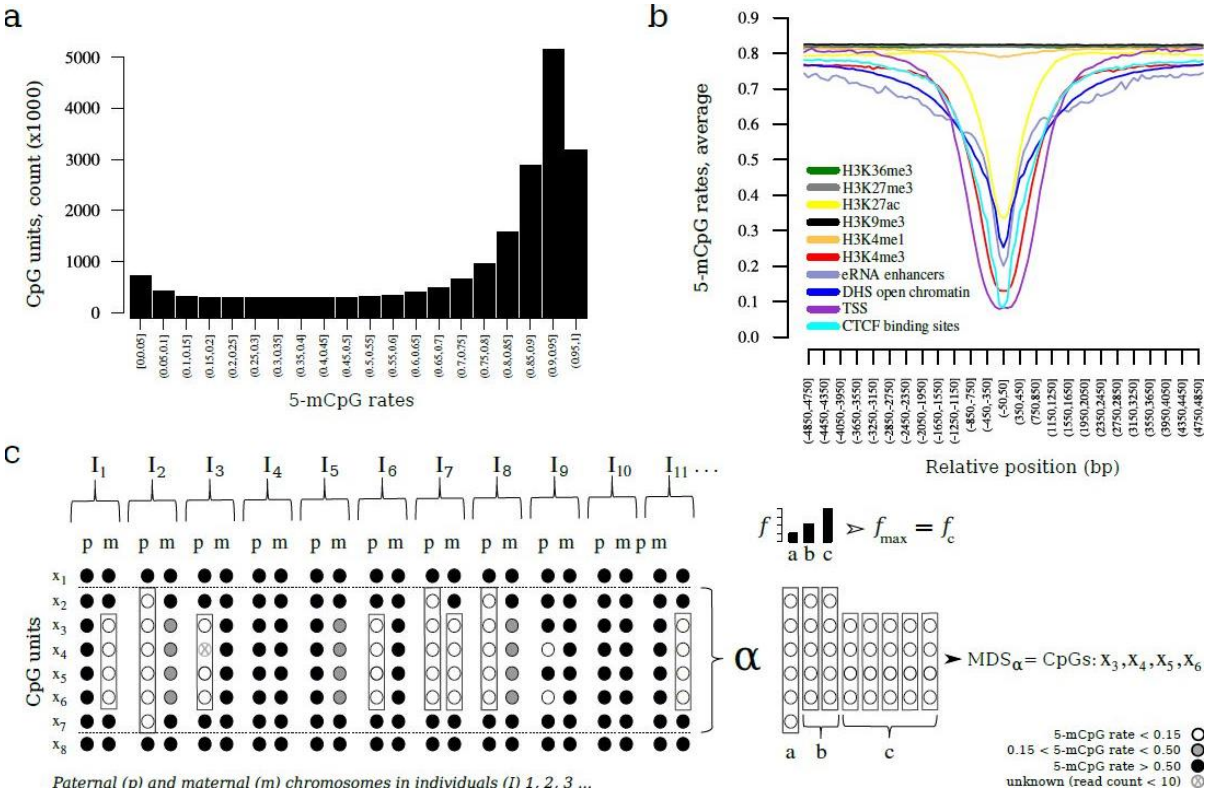


Figure 1: **5-mCpG detected by nanopore sequencing. (a)** 5-mCpG rates computed across individuals yielded the expected bimodal distribution. **(b)** 5-mCpG rates averaged in 100bp bins relative to the mid-position of chromatin makers assayed in relevant cell types, i.e. histone modifications (H3K4me3, H3K27ac, H3K36me3 and H3K9me3) in primary neutrophils, CTCF protein binding sites in primary neutrophils and open chromatin regions (DNase hypersensitive sites; DHS) in CD4+ T-cells obtained from Encode and Roadmap epigenomics project. Additionally, enhancer RNA (eRNA) and main transcription start site (TSS) reference maps for RNA samples isolated from whole blood based on CAGE-seq assays obtained from the Fantom5 project (SlideBase). **(c)** We applied sequence-based phasing to assign 5-mCpG status to paternal (p) or maternal (m) haplotypes in each individual (I). For each CpG unit and each parental haplotype, we computed the 5-mCpG rate and defined unmethylated haplotypes where we found three or more neighbouring CpG units each with 5-mCpG rate <0.15, but located no more than 500bp apart, in at least one haplotype (restricted to the 2,648 individuals sequenced to an average coverage of >20x). Open circles = unmethylated CpG units, grey-filled circles = low methylated CpG units, black-filled circles = higher methylated CpG units. A rectangle is drawn around neighbouring CpG units where such unmethylated haplotypes were detected. The cluster labelled a defines locations containing overlapping unmethylated haplotypes labelled a, b and c. The most frequently occurring unmethylated haplotype (f_{max}) is then nominated as the representative methylation depleted sequence of cluster a (MDS_a) containing CpG units x_3,x_4,x_5,x_6.

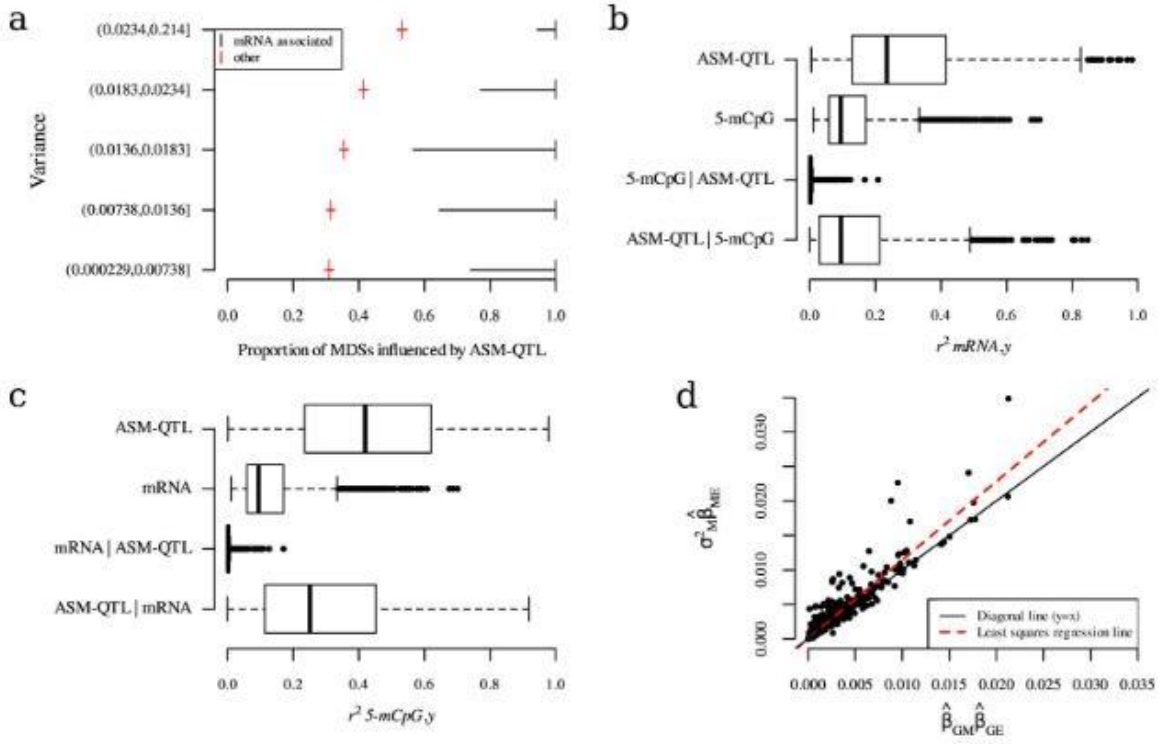


Figure 2: **ASM-QTLs dominate in correlations found between MDSs and mRNA expression**. (**a**) Variance in 5-mCpG rates of MDSs, one MDS from each 100kb segment of the genome (n=25,079 MDSs), binned by quintiles (y-axis; ~5000 MDSs per bin), and plotted against the proportion of these MDSs that have an associated ASM-QTL (x-axis) according to whether the MDS is associated with mRNA expression (black) or not (red). The proportion estimates are shown as tick marks (vertical lines), and their 95% confidence intervals are shown as horizontal lines. **(b)** The fraction of the variance in mRNA expression explained by each of the four variables on the *y*-axis as follows: *y*=′ASM-QTL′ represents the genotype of the ASM-QTL found in association with 5-mCpG rates of MDSs; *y*=′5-mCpG′ represents the 5-mCpG rates of MDSs; *y*=′5-mCpG | ASM-QTL′ represents the 5-mCpG rates of MDSs after correction for the ASM-QTL found in association with that same MDS; *y*=′ASM-QTL | 5-mCpG′ represents the genotype of the ASM-QTL found in association with 5-mCpG rates of MDSs after having corrected the genotype status for the 5-mCpG rates of that same MDS. **(c)** The fraction of the variance in 5-mCpG rates explained by each of the four variables on the *y*-axis where ′mRNA′ represents mRNA expression but otherwise analogous to **(b)**; n=1,513. Note, in **(b)** and **(c)** the center line (solid black) shown in each box represents the median; box limits represent upper and lower quartiles; whiskers represent 1.5x interquartile range. $r^2_{\text{mRNA,5-mCpG}}$ in **(b)** is equivalent to the reversed comparison of $r^2_{\text{5-mCpG,mRNA}}$ in **(c)**. **(d)** The effects of ASM-QTL genotype (G) on CpG methylation ($\hat{\beta}_{GM}$) and mRNA expression ($\hat{\beta}_{GE}$), x-axis, compared to the effects of CpG methylation on mRNA expression ($\hat{\beta}_{ME}$), y-axis. Diagonal line (x=y) is represented as a solid black line, and the least squares regression line is represented as a dashed red line as indicated in the figure legend shown in the bottom right corner. The number of data points in each boxplot in (b) and (c), and in the scatter plot in (d) corresponds to the number of associations

found between methylation and gene expression (n=1,513). Abbreviations: G=ASM-QTL genotype; M=5-mCpG rate; E=mRNA expression.

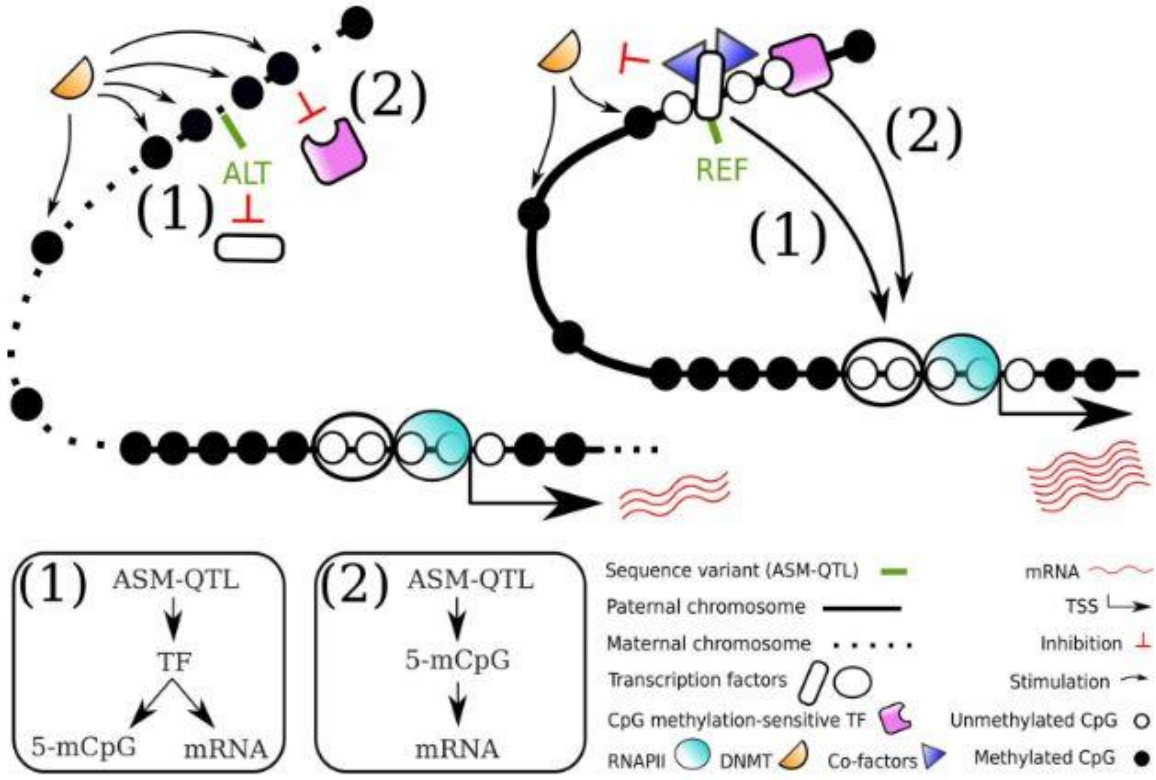


Figure 3: DNA sequence variability affects CpG methylation and gene expression. The figure illustrates the two models consistent with our results, see model diagrams at the bottom left side of the figure. In this hypothetical example, an ASM-QTL gives rise to CpG methylation differences between the maternal (left) and paternal (right) chromosomes of an individual. Under model (1), the ASM-QTL influences TF binding to DNA which, in turn, influences methylation of nearby CpGs, but it is the TF (not methylation) that then results in influences on gene expression. Under model (2), the ASM-QTL influences TF binding to DNA which, again, leads to influences on methylation of nearby CpGs, but here the change in methylation results in influences on gene expression e.g., by enabling binding of a CpG methylation-sensitive TF. Hence, methylation is irrelevant to gene expression in model (1) whereas it is relevant to gene expression in model (2). Importantly, in both models, it is DNA sequence variability that drives the correlation between CpG methylation and gene expression. Abbreviations: MDS=Methylation depleted sequence; REF= Reference allele; ALT=Alternative allele; RNAPII=RNA polymerase II; DNMT=DNA methyltransferase; TSS=Transcription start site.

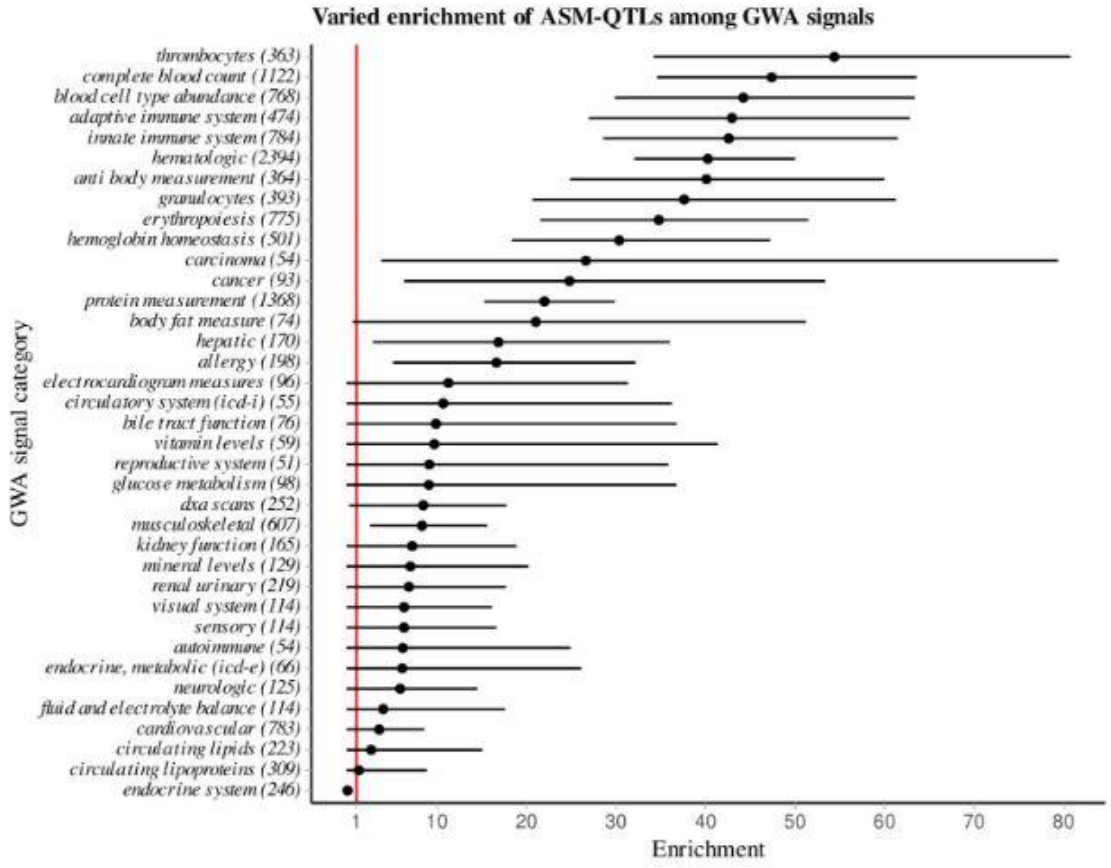


Figure 4: ASM-QTLs are enriched among GWA signals. The enrichment of ASM-QTLs, *x*-axis, among GWA signals varies in magnitude by trait category, *y*-axis. The solid points (black) represent the measure of center, i.e., the enrichment point estimates and the horizontal lines (black) represent their 95%CIs. The number of GWA signals for each trait category is shown within parentheses on the y-axis. Vertical line (red) indicates the point of neutral enrichment on the *x*-axis i.e., where $x=1$. Abbreviations: GWA=Genome-wide association.

References

1. Luo C, Hajkova P, Ecker J.R. Dynamic DNA methylation: In the right place at the right time. *Science* **361**(6409), 1336-1340 (2018).

2. Okano M, Bell D.W., Haber D.A., Li E. DNA methyltransferases Dnmt3a and Dnmt3b are essential for de novo methylation and mammalian development. *Cell* **99**(3), 247-257 (1999).

3. Li E, Bestor T.H., Jaenisch R. Targeted mutation of the DNA methyltransferase gene results in embryonic lethality. *Cell* **69**(6), 915-26 (1992).

4. Clark S.J. et al. Single-cell multi-omics profiling links dynamic DNA methylation to cell fate decisions during mouse early organogenesis. *Genome Biol*. **23**(1):202 (2022).

5. Dynan W.S., Tjian R. The promoter-specific transcription factor Sp1 binds to upstream sequences in the SV40 early promoter. *Cell* **35**(1), 79-87 (1983).

6. Vierstra J. et al. Global reference mapping of human transcription factor footprints. *Nature* **583**(7818), 729-736 (2020).

7. Ziller, M. et al. Charting a dynamic DNA methylation landscape of the human genome. *Nature* **500**, 477–481 (2013).

8. Stadler M.B. et al. DNA-binding factors shape the mouse methylome at distal regulatory regions. *Nature* **480**(7378), 490-495 (2011).

9. Gardiner-Garden M., Frommer M. CpG islands in vertebrate genomes. *J Mol Biol*. **196**(2), 261-82 (1987).

10. Bird, A. CpG-rich islands and the function of DNA methylation. *Nature* **321**, 209–213 (1986).

11. Brandeis, M. et al. Sp1 elements protect a CpG island from de novo methylation. *Nature* **371**, 435–438 (1994).

12. Ooi S.K. et al. DNMT3L connects unmethylated lysine 4 of histone H3 to de novo methylation of DNA. *Nature* **448**(7154), 714-717 (2007).

13. Boulard, M., Edwards, J. & Bestor, T. FBXL10 protects Polycomb-bound genes from hypermethylation. *Nat Genet*. **47**, 479–485 (2015).

14. Krebs, A.R., Dessus-Babus, S., Burger, L., Schubeler, D. High-throughput engineering of a mammalian genome reveals building principles of methylation states at CG rich regions. *eLife* **3**:e04094 (2014).

15. Wachter E. et al. Synthetic CpG islands reveal DNA sequence determinants of chromatin structure. *eLife* **3**:e03397 (2014).

16. Onuchic V. et al. Allele-specific epigenome maps reveal sequence-dependent stochastic switching at regulatory loci. *Science* **361**(6409):eaar3146 (2018).

17. Do C, et al. Mechanisms and Disease Associations of Haplotype-Dependent Allele-Specific DNA Methylation. *Am J Hum Genet*. **98**(5), 934-955 (2016).

18. Lienert F., Wirbelauer C., Som I., Dean A., Mohn F., Schübeler D. Identification of genetic elements that autonomously determine DNA methylation states. *Nat Genet*. **43**(11), 1091-1097 (2011).

19. Bonder MJ. et al. Disease variants alter transcription factor levels and methylation of their binding sites. *Nat Genet*. **49**(1), 131-138 (2017).

20. Yin Y. et al. Impact of cytosine methylation on DNA binding specificities of human transcription factors. *Science* **356**(6337):eaaj2239 (2017).

21. Domcke S. et al. Competition between DNA methylation and transcription factors determines binding of NRF1. *Nature* **528**(7583):575-579 (2015).

22. Borgel J. Targets and dynamics of promoter DNA methylation during early mouse development. *Nat Genet*. **42**(12), 1093-100 (2010).

23. Kaluscha S. et al. Evidence that direct inhibition of transcription factor binding is the prevailing mode of gene and repeat repression by DNA methylation. *Nat Genet*. **54**(12), 1895-1906 (2022).

24. Swain J.L., Stewart T.A., Leder P. Parental legacy determines methylation and expression of an autosomal transgene: a molecular mechanism for parental imprinting. *Cell* **50**(5), 719-727 (1987).

25. Stöger R., Kajimura T.M., Brown W.T., Laird C.D. Epigenetic variation illustrated by DNA methylation patterns of the fragile-X gene FMR1. *Hum Mol Genet*. **6**(11), 1791-1801 (1997).

26. Lin I.G., Tomzynski T.J., Ou Q., Hsieh C.L. Modulation of DNA binding protein affinity directly affects target site demethylation. *Mol Cell Biol*. **20**(7), 2343-2349 (2000).

27. Métivier R. et al. Cyclical DNA methylation of a transcriptionally active promoter. *Nature* **452**(7183), 45-50. (2008).

28. Gutierrez-Arcelus M. et al. Passive and active DNA methylation and the interplay with genetic variation in gene regulation. *eLife* **2**:e00523 (2013).

29. Shang L. et al. meQTL mapping in the GENOA study reveals genetic determinants of DNA methylation in African Americans. *Nat Commun*. **14**(1), 2711 (2023).

30. Pierce BL. Co-occurring expression and methylation QTLs allow detection of common causal variants and shared biological mechanisms. *Nat Commun*. **9**(1), 804 (2018).

31. Simpson, J. et al. Detecting DNA cytosine methylation using nanopore sequencing. *Nat Methods*. **14**, 407–410 (2017).

32. Zink F. et al. Insights into imprinting from parent-of-origin phased methylomes and transcriptomes. *Nat Genet*. **50**(11), 1542-1552 (2018).

33. The ENCODE Project Consortium. Expanded encyclopaedias of DNA elements in the human and mouse genomes. *Nature* **583**, 699–710 (2020).

34. Andersson R. et al. An atlas of active enhancers across human cell types and tissues. *Nature* **507**(7493), 455-461 (2014).

35. FANTOM Consortium and the RIKEN PMI and CLST (DGT). A promoter-level mammalian expression atlas. *Nature* **507**(7493), 462-470 (2014).

36. Gudbjartsson, D. et al. Large-scale whole-genome sequencing of the Icelandic population. *Nat Genet*. **47**, 435–444 (2015).

37. Beyter, D. et al. Long-read sequencing of 3,622 Icelanders provides insight into the role of structural variants in human diseases and other traits. *Nat Genet*. **53**, 779–786 (2021).

38. Eggertsson H.P. et al. GraphTyper2 enables population-scale genotyping of structural variation using pangenome graphs. *Nat Commun*. **10**(1):5402 (2019).

39. Kristmundsdottir S., Eggertsson H.P., Arnadottir G.A., Halldorsson B.V. popSTR2 enables clinical and population-scale genotyping of microsatellites. *Bioinformatics* **36**(7), 2269-2271 (2020).

40. Min, J.L. et al. Genomic and phenotypic insights from an atlas of genetic effects on DNA methylation. *Nat Genet*. **53**, 1311–1321 (2021).

41. Monk D., Mackay D.J.G., Eggermann T., Maher E.R., Riccio A. Genomic imprinting disorders: lessons on how genome, epigenome and environment interact. *Nat Rev Genet.* **20**, 235–248 (2019).

42. Chen, J. et al. A uniform survey of allele-specific binding and expression over 1000-Genomes-Project individuals. *Nat Commun*. **7**, 11101 (2016).

43. Hemani G., Tilling K., Davey Smith G. Orienting the causal relationship between imprecisely measured traits using GWAS summary data. *PLoS Genet*. **13**(11):e1007081 (2017)

44. Sveinbjornsson G. et al. Weighting sequence variants based on their annotation increases power of whole-genome association studies. *Nat Genet*. **48**(3), 314-317 (2016).

45. Halldorsson B.V. et al. Characterizing mutagenic effects of recombination through a sequence-level genetic map. *Science* **363**(6425):eaau1043 (2019).

46. Ferkingstad E. et al. Large-scale integration of the plasma proteome with genetics and disease. *Nat Genet*. **53**(12), 1712-1721 (2021).

47. Oliva M. et al. DNA methylation QTL mapping across diverse human tissues provides molecular links between genetic variation and complex traits. *Nat Genet*. **55**(1), 112-122 (2023).

48. Chen L. et al. Genetic Drivers of Epigenetic and Transcriptional Variation in Human Immune Cells. *Cell* **167**(5), 1398-1414 (2016).

49. Buniello A. et al. The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Res*. **47**(D1), D1005-D1012 (2019).

50. Vujkovic M. et al. Discovery of 318 new risk loci for type 2 diabetes and related vascular outcomes among 1.4 million participants in a multi-ancestry meta-analysis. *Nat Genet*. **52**(7), 680-691 (2020).

51. Han Y. et al. Genome-wide analysis highlights contribution of immune system pathways to the genetic architecture of asthma. *Nat Commun*. **11**(1), 1776 (2020).

52. Wallace C. et al. Dissection of a Complex Disease Susceptibility Region Using a Bayesian Stochastic Search Approach to Fine Mapping. PLoS Genet. 24;11(6):e1005272. (2015).

53. Butz S. et al. DNA sequence and chromatin modifiers cooperate to confer epigenetic bistability at imprinting control regions. *Nat Genet*. **54**(11), 1702-1710 (2022).

Methods

Ethical Statement

This study was approved by the National Bioethics Committee in Iceland (Approval no. VSN 14-015) and conducted in agreement with instructions issued by the Data Protection Authority in Iceland (PV_2017060950ÞS/--). All individuals gave informed consent, and all personal identifiers were encrypted by an external agent before being imported into the deCODE database.

Statistics and reproducibility

In this study we nanopore sequenced DNA isolated from whole blood samples from 7,179 Icelanders (3,434 males, 3,745 females) participating in various studies at deCODE genetics[36,44,45]. The earliest year of birth (YOB) was 1876 and 1890 for males and females respectively and the latest was 2015 for both sexes. The median YOB was 1960 for males and 1958 for females.

No statistical method was used to predetermine sample size. The sample size was determined based on the number of nanopore sequenced DNA samples. We excluded nanopore sequenced DNA samples derived from tissues other than whole blood. We further excluded individuals where we obtained <10x nanopore sequencing coverage after restricting to flowcells sequenced to at least 3x average coverage, as described in the results. No animals were used in the study.

DNA isolation and sequencing

DNA from whole blood was extracted using the Chemagic method (perkinElmer), an automated procedure that involves the use of M-PVA magnetic beads. Quantitation of genomic DNA was performed on the Big Lunatic instrument using software and plates from the manufacturer, and the absorbance ratio for quality. DNA integrity was assessed using the Fragment Analyzer capillary system from AATI, following the manufacturers guidelines.

DNA sequencing libraries were generated using the SQK-LSK109 ligation kit from ONT. Sample input varied from 1 to 5 µg DNA, depending on the exact version of the preparation kit and the flowcell type used for the PromethION sequencing. Nanopore sequencing was performed using PromethION machines, using R.9.4 flowcells, following ONT standard operating procedures. Data acquisition varied from 48 to 60 h per flowcell.

CpG methylation detection by nanopore sequencing

Squiggle data from the sequencers were basecalled using Guppy and mapped to the human reference genome GRCh38 with Minimap2[54]. We then used Nanopolish[31] to detect 5-mCpG from nanopore sequenced DNA. Nanopolish detects methylated cytosines in a CpG context using a Hidden Markov model (HMM) to assign a log-likelihood ratio (LLR) for the presence of a cytosine methylation at each CpG site. We interpret values above +1.921 as indicating support for cytosine methylation and less than -1.921 as support unmethylated CpG. Nanopolish groups CpG sites within 10bp distance and assigns a methylation status to each group such that all CpG sites within a group have the same methylation status. For this reason, we refer to CpG sites measured by Nanopolish as: **CpG units**. 5-mCpG status was assigned as unreliable if the prediction was ambiguous ($-1.921 \leq LLR \leq 1.921$).

We assessed the impact of various attributes of the CpG units on the quality of 5-mCpG detection (**Supplementary Notes 1.1** and **1.2**). First, strand bias and the fraction of reliable reads (FRR) was calculated for each CpG unit by averaging over the whole dataset. We then defined strand bias as the difference in methylation levels between forward and reverse strand, and FRR as the fraction of reliable reads out of all reads. CpG units were removed if the strand bias was ≥0.20 or if the FRR was ≤0.5. Second, we removed CpG units within 5bp of a known SNP locus (MAF>0.001) as the presence of a sequence variant in a pore at the same time as an unmethylated CpG may produce an electric signal similar to the signal of 5-mCpG, and *vice versa*. Third, we removed CpG units of coverage higher than 1.5 times the average as this is evidence of repetitive regions, CpG units of coverage lower than 0.5 times the average as this indicates structural variants or regions where the sequencing or mapping might be problematic. Fourth, we removed CpG units located within so-called dark regions[55] of the genome as these regions contain sequence repeats that cause mapping to

be unreliable. Fifth, we removed CpG units where the fraction of phased sequences was less than 0.3 as this is indicative sequences that are difficult to phase. In total, 15,317,794 CpG units satisfied our criteria for high quality detection of 5-mCpG in nanopore sequenced DNA samples.

We restricted our cohort to DNA samples from whole blood that were nanopore sequenced to at least 10x coverage analyzed on flowcells with at least 3x coverage (**Supplementary Notes 1.1**; **Supplementary Figs. 1** and **2**). The mean N50 (measure of average sequence length) of our nanopore sequence data is 18861bp (median=18376; min=4491, max=50719) (**Supplementary Figure 1**).

We define coverage as the total number of sequenced base pairs divided by $3 \cdot 10^9$, approximately the size of the reference genome.

Assignment of 5-mCpG status to parental haplotypes

DNA sequence variants were called using Graphtyper[56] based on 63,460 whole genome sequenced individuals representing a subset of 173,025 SNP chip-typed individuals from the Icelandic population[36,37,38,39]. Whole genome sequencing was carried out using Illumina sequencing to mean depth of 39.8x (range 20–397.8x). The sequences were then phased to impute haplotypes into the chip-typed individuals.

Whole genome sequences of individuals in our study had been long-range phased and assigned parent of origin[57], enabling us to assign sequences analyzed by Nanopolish to maternal or paternal chromosomes. Parental haplotypes were assigned by examining the phasing status of a set of 8,960,728 high-quality sequence variants[58], using heterozygous carriers of in-read sequence variants, allowing us to assign CpG methylation calls (methylated or unmethylated) to maternal and paternal chromosomes. Sequences overlapping at least 3 heterozygous variants and where at least 70% of the variants where consistent with the phasing of one parent were considered to be phased and used for subsequent analysis.

For each CpG unit and each parental haplotype, we defined the 5-mCpG haplotype rate as the number of sequences of that parental haplotype that are methylated at the CpG unit divided by the number of parental sequences covering the CpG unit.

MDSs (methylation depleted sequences)

Definition: MDSs are sequences depleted of 5-mCpGs across at least three CpG units located within 500bp of each other, on the same haplotype in at least one individual.

We applied measurements of 5-mCpG rates on individual haplotypes to find instances where closely located CpG units are found depleted in methylation on the same haplotype. We confined our search for 5-mCpG depleted haplotypes to nanopore methylomes sequenced to an average coverage of >20x (n=2,648 individuals) and, further, we restricted to CpG units where ≥10 sequences were available for estimating the 5-mCpG rate. For each individual and each of its parental haplotypes, we defined unmethylated haplotypes as the occurrence of three or more CpG units on the same haplotype each of which displaying 5-mCpG rates <0.15, but located no more than 500bp apart from each other. Unmethylated haplotypes found in different individuals often shared the exact same coordinates, i.e. they were defined by the exact same CpG units, or varied slightly from one individual to another. We therefore defined clusters of overlapping unmethylated haplotypes based on CpG position; unmethylated haplotypes sharing one or more CpG units were clustered together. In each such cluster, we restricted to unmethylated haplotypes where the CpG units located immediately up- and downstream were measured, i.e. based on >10 sequences. We then catalogued the genome

coordinates of the most frequently occurring unmethylated haplotype. We then removed unmethylated haplotypes found in overlap to these coordinates, to then determine whether there were any remaining unmethylated haplotypes in that same cluster. If so, we repeated the process of cataloging the most frequently occurring unmethylated haplotype until the cluster was emptied. In instances where there were more than one unmethylated haplotype with the highest frequency we catalogued the coordinates of the longest (bp) haplotype amongst them.

Low-methylated sequences (but not unmethylated) have been shown to be a characteristic feature of distal regulatory elements[8]. We therefore defined low-methylated haplotypes as the occurrence of more than three CpG units on the same haplotype, each with 5-mCpG rate <0.5, but only in sequences where unmethylated haplotypes were not found. We refer to the collection of unmethylated- and low-methylated haplotypes as MDSs.

We then measured the methylation level of each MDS by counting the number of methylated and unmethylated CpG units positioned within each MDS on each parental haplotype in each individual to then compute the 5-mCpG rate: the number of methylated CpG units divided by the number of methylated and unmethylated CpG units. Hence, even though we restrict to a subset of individuals to search for MDSs, we still measure the 5-mCpG rate of MDSs in all of the 7,179 individuals in our cohort.

Allele-specific methylation quantitative trait loci

Definition: ASM-QTLs are alleles of sequence variants that lead to local changes in the methylation status of CpG sites on the same inherited sequence.

To identify ASM-QTLs, we followed a two-phased procedure. In the first phase, we used least squares regression to identify the most likely causal variant. In the second phase, we removed variants that did not associate after we conditioned on cellular composition among the 1,934 of 7,179 (27%) individuals where we had this information.

Phase 1: We defined a total of 34,435,950 high-quality sequence variants[56,36,37,38,39] with minor allele frequency $>10^{-4}$ by filtering on sequence variants with imputation information above 0.9, alternative allele score (AA-score; SNP/indels only) above 0.5, average allele balance of heterozygous and homozygous individuals above 27.5% and 96.5%, respectively, root-mean-square mapping quality of the overlapping reads above 20, between 10% and 90% of the overlapping reads mapped on the forward strand, at least 5 reads supporting the alternative allele in each individual, and at least 30% of the reads supporting the alternative allele in any individual.

These criteria yielded 23,752,296 SNPs[56], 5,929,255 indels[56], 609,536 structural variants[37,38] and 4,144,863 microsatellites[39].

We used multivariate least squares regression (fastLm function in RcppEigen[59] package version 0.3.3.9.4; R version 3.6) to search for sequence variants associated with 5-mCpG rates of each of the ~189 thousand MDSs measured on each of the two haplotypes in the 7,179 individuals in our cohort. We set a Bonferroni corrected significance threshold in accordance with the number of hypothesis tests performed ($P < 0.05/4.5 \cdot 10^8 \sim 10^{-10}$). In each association signal, the most significant variant was selected as the ASM-QTL (representing the likely causal variant); also referred to as the „primary“ association.

As the distribution of 5-mCpG rates cannot be assumed to be normal for each and every MDS, we performed the inverse normal transformation.

For each MDS, we restricted the search to sequence variants located 100kb up- or downstream of the MDS. We included measured haplotypes regardless of the number of sequences used to estimate the 5-mCpG rate of the MDS to perform the test for association with alleles of sequence variants, but we excluded MDSs where less than 100 haplotypes were measured.

Covariates were as follows: age, sex [male, female], parental haplotype [paternal, maternal], and the first five principal components computed on all autosomes apart from the autosome containing the MDS being tested for association, see definition further in the **Covariates** section, below.

In searching for „secondary" associations (i.e. to see if there are more than one ASM-QTLs for each MDS), we confined our search to the major allele of the primary association variant for each MDS, and we then performed the same association analysis accounting for the same covariates and, as before, we used the same *P*-value threshold at $<10^{-10}$ for detecting secondary associations and, as before, we then selected the most significant sequence variant as the index for the secondary ASM-QTL.

Phase 2: Information on cell type composition was available for 1,934 (out of 7,179) individuals measured in the same blood draws as were used to isolate the DNA for nanopore sequencing. We restricted the set of ASM-QTL to those that remained significant at an FDR rate of 0.5% after accounting for cell type composition in the subset of 1,934 individuals. We did this by adding cell count information to the covariates in Phase 1. We included counts of neutrophils, basophils, eosinophils, immature granulocytes, monocytes, white blood cells, and red blood cells in addition to the fraction of nucleated red blood cells. We did not include lymphocyte counts because of their strong correlation with neutrophil counts (Pearson's $r = -0.94$) to avoid co-linearity in the regression.

We then followed the same two-phased procedure to identify sequence variants associated with individual CpG units.

Covariates

We sampled two random subsets of the CpG methylation data: one used for training and the other for testing. Each subset consisted of approximately 1% of the 15.3 million high-quality autosomal CpG units. By using an add one based method we tested the CpG units for association with the following covariates: QC measures (N50, number of ultra long sequences, percentage alignment and percentage error) and sequencing measures (sequencing device, concentration, ratio, source type and storage time).

This association analysis was carried out as follows: First, we ran association of all the covariates to each site in the training subset and ranked the covariates based on their median *P*-value. Second, using the ordering from the first step, we compared the goodness of fit for the model containing n covariates to the model containing $n+1$ covariates for each site in the test set, where $0 \leq n \leq m$, with m denoting the total number of covariates. We continued until adding more covariates no longer yielded a significantly better model for the majority of sites at a nominal *P*-value threshold of < 0.05.

Using these criteria, we chose not to adjust for any of these covariates since the first ordered covariate was only significant for <10% of the sites.

We sampled a random subset of the methylation data and computed principal components (PCs) for each autosome separately by using CpG units from all other autosomes (**Supplementary Figure 5**). We chose to adjust for the first five PCs as they collectively explain approximately 2.84% of the variance in 5-mCpG rates (**Supplementary Figure 6**); Each of the other PCs explain less than 0.12%, and were therefore omitted.

Currently, methods used for predicting cell type counts for array-based measurements of methylation have not been validated for nanopore sequencing data. We therefore accounted for cell type composition by using the first five PCs as they capture a large fraction of the variability in neutrophils and lymphocytes (**Supplementary Figure 5**), while also using direct measures of cell type composition in individuals available for a subset of our cohort, see further in methods sections entitled: **Allele-specific methylation quantitative trait loci** and **RNA sequencing and phasing to parental chromosomes**.

ASM-QTLs validated in an independent cohort

Validation of ASM-QTLs found in association with 5-mCpG rates of MDSs was performed using oxidative bisulfite sequencing performed in our previous study[32] using DNA samples isolated from whole blood samples from forty-five individuals. Note, these forty-five individuals used for validation were not included in our cohort of 7,179 nanopore sequenced individuals and were therefore independent of the study cohort. We performed the same multivariate least squares regression as was performed for nanopore sequenced samples with age, sex and parental haplotype as covariates in both oxBS and nanopore sequencing data. As a measure of consistency between the ASM-QTL effect sizes in the 45 and 7,179 individuals sequenced by oxBS and nanopore sequencing, respectively, we computed:

$$t = \frac{\hat{\beta}_{ox} - \hat{\beta}_{nano}}{\sqrt{s^2_{ox} + s^2_{nano}}}$$

Here, $\hat{\beta}$ represents the effect size estimates for ASM-QTLs and s^2 is the variance of those effect size estimates in oxBS (*ox*) and nanopore (*nano*) sequenced individuals. The proportion of consistent effect sizes was then computed as the number of nominally non-significant differences between the ASM-QTL effect sizes in oxBS and nanopore sequenced individuals ($P>0.05$) divided by the total number of ASM-QTLs that we were able to test for validation in the 45 oxBS sequenced individuals. The *P*-value was based on the *t*-statistic with *n*-1 degrees of freedom where *n* is the number of informative haplotypes in the regression analysis in the oxBS validation cohort.

ASM-QTL in functional annotation maps

We tested whether ASM-QTLs were more likely than expected by chance to be identical, or in high linkage disequilibrium, to sequence variants identified as functionally relevant in other studies[33,42].

ASM-QTLs, or sequence variants in high linkage disequilibrium ($r^2>0.80$) to ASM-QTLs, are often found in close proximity to one another and may therefore be found in the same annotated region. To eliminate such dependencies between observations due to proximity we employed the following procedure: First, we selected the single most significant ASM-QTL (based on the lowest *P*-value). Second, we selected the most significant of the remaining ASM-QTLs that was at least 1Mb away from the already selected ASM-QTL. Third, we repeated step two until no more ASM-QTLs could be found at least 1Mb away from those ASM-QTLs already selected. In this way, we obtained a subset of ASM-QTLs ($n = 1{,}929$), hereafter refered to as the „observed ASM-QTLs“, for use in analysis of enrichment among functional annotation maps of the genome.

The proximity to MDSs and the number of sequence variants found in high linkage disequilibrium ($r^2>0.80$) to the 1,929 observed ASM-QTLs are expected to influence the probability of finding an overlap with functional annotations of the genome. We therefore sampled the same number of sequence variants, i.e. 1,929, from regions located <10kb from the midpoint of any of the 189

thousand MDSs, while also ensuring that the 1,929 sampled variants are matched to the each of the 1,929 observed ASM-QTLs with respect to the number of sequence variants found in high linkage disequilibrium. Additionally, we require that the 1,929 sampled variants are spaced by >1Mb as this was also the requirement for the 1,929 observed ASM-QTLs. We refer to a sampled variant and variants found in high linkage disequilibrium to the sampled variant as a „sampled signal". We then count the number of sampled signals that overlap with sequence variants found within a functional annotation (this count is denoted as z). This procedure is then repeated N=50,000 times. In summary, we are simulating the ASM-QTLs in terms of a) the proximity of ASM-QTLs to MDSs and b) the number of sequence variants in high linkage disequilibrium to the ASM-QTLs.

Let z_i represent the number of sampled signals that were annotated in each i-th set of N samples. The probability that a sampled signal overlaps a functional annotation is then:

$$p = \frac{\sum_i^N z_i}{nN}$$

Here, $n = 1{,}929$ is the number of sampled variants and N=50,000 is the number of sampled sets.

To determine the probability of observing x or more ASM-QTLs in a given annotation, where x is the observed number of ASM-QTLs that overlap with that annotation, we compute $P(X \geq x) = j/N$, where j is the number of times we found x or more of the 1,929 sampled signals in overlap with the annotation in each of the aforementioned N=50,000 sampled sets.

In instances where j=0, we compute the P-value using a binomial approximation.

We define X as a random variable that follows a binomial probability distribution $Bin(n, p)$ representing the number of ASM-QTLs found in a functional annotation. As we have ensured a minimum of 1Mb distance between the observed ASM-QTLs we assume that the observations are largely independent. We then use the probability density function of the binomial distribution for $Bin(n, p)$ to compute $P(X \geq x)$ as the sum of probabilities of finding x or more ASM-QTLs in the functional annotation using the dbinom function in R.

The fold-enrichment is computed as: $\frac{p'}{p}$ where $p' = \frac{x}{n}$

ASM-QTL enrichment among trait associated sequence variants

In this study, a GWA signal is defined as the lead (strongest) association variant for a human trait and sequence variants found in high linkage disequilibrium ($r^2 > 0.80$) to this same lead variant. We used GWA signals (P-value $<10^{-9}$) identified in a diverse set of 261 human phenotypes in the Icelandic population[36,44,45], including 60 diseases and 201 other traits.

We restricted to the ~34.4 million sequence variants that we used to search for ASM-QTLs (methods section entitled: **Allele-specific methylation quantitative trait loci**). We selected the strongest associating variant (the highest χ^2) within each 1Mb interval to make a list of candidate association variants. For each trait, we then retained only sequence variants that are independently associated with the trait i.e. only the strongest association variants are retained from each chromosome if they still associate with the same trait at $P<10^{-9}$ after correcting for other candidate association variants located on the same chromosome.

We defined the enrichment among GWA signals as the fold-change in the proportion of GWA signals among sequence variants of a given annotation category relative to the genome-wide proportion of variants in that category.

We used our previously described model for estimating enrichment of sequence variant annotation among GWA signals[44], but with modifications that were critical for enabling analysis of annotations that are skewed towards common variants as our ASM-QTLs are. In Sveinbjornsson et al[44], the enrichment of an annotation c, E_c, was estimated as $\frac{p_c}{q_c}$, where p_c is the probability of a causal variant being from annotation c and q_c is the probability of a non-causal variant being from annotation c. The derived approximate likelihood over all association signals i, each with marker set M_i was as follows:

$$\mathcal{L}(\boldsymbol{p}) = \prod_i \sum_{m \epsilon M_i} e^{\chi^2_m} p_{c_m} \prod_{m' \neq m} \hat{q}_{c_{m'}}$$

Here, χ^2_m is the test statistic from the GWA for an individual sequence variant m and $c_m$is the annotation of variant m and $\hat{q}_c$ is an estimate of q_c as the proportion of variants coming from annotation c among tested variants. The parameters p_c were inferred by maximizing the above likelihood and the enrichment estimates are then $\hat{E}_c = \frac{\hat{p}_c}{\hat{q}_c}$.

This model works well to estimate enrichment of VEP (Variant Effect Predictor) annotations[60] but has its drawbacks in the current context. ASM-QTLs are identified on the basis of association analyses and as a result their minor allele frequency (MAF) is skewed towards common variants. GWA signals are likewise skewed towards common variants as their detection also depends on statistical power, i.e. the larger the number of individuals the more power there is for detection of associations to sequence variants in the lower end of the MAF spectrum. This skew of ASM-QTLs towards common variants will therefore have the tendency to inflate enrichment estimates among GWA signals, which makes them incomparable to enrichment estimates of other frequency independent annotations in VEP. To alleviate this problem, we modified the model such that the parameter p_c, the probability of a causal variant in an association signal being from annotation c, is allowed to vary with variant frequency. Further, we reparametrized the model under the assumption that the true enrichment of annotation c, E_c, is independent of variant frequency and write for frequency bin f:

$$p_{c,f} = \frac{E_c q_{c,f}}{\sum_{k \epsilon C} E_k q_{k,f}}$$

$q_{c,f}$ is the probability of a non-causal variant coming from annotation c in frequency bin f and C is set of annotations and k is a running index over annotations in C. The denominator of the expression above implicitly ensures that $p_{c,f}$ is between 0 and 1. The modified approximate likelihood, which now is only a function of the enrichments E_c, is then as follows:

$$\mathcal{L}(\boldsymbol{E}) = \prod_i \sum_{m \epsilon M_i} e^{\chi^2_m} \frac{E_{c_m} \hat{q}_{c_{m,f}}}{\sum_{k \epsilon C} E_k \hat{q}_{k,f}} \prod_{m' \neq m} \hat{q}_{c_{m',f}}$$

$\hat{q}_{c,f}$ is an estimate of $q_{c,f}$ as the proportion of variants coming from annotation c in frequency bin f among tested variants. We then estimated the enrichment by maximizing the above likelihood. We assumed $E_c \geq 0$ for all $c \epsilon C$ and to ensure identifiability we set the annotation with the largest number of variants, which in this study are intronic variants, equal to 1. In other words, the largest annotation serves as a baseline and the enrichment estimates of other annotations is relative to it. We selected the frequency bins such that they included approximately the same number of association variants (~1014 GWA signals per each of 5 frequency bins). All sequence variants that belonged to a GWA signal were taken to be in the same frequency bin as the strongest (lead) variant for that signal.

For further details on the derivation of the model see: Sveinbjornsson et al[44]. To validate the model modification, we carried out extensive analyses using random annotations of varying sizes and frequency distributions to show that the method is not sensitive to either frequency distribution or size of the variant annotation.

We employed *Rsolnp* package in R (*solnp* function) to maximize $\mathcal{L}(\boldsymbol{E})$ and obtain the estimates $\hat{E}_c$. In practice, to handle the large numbers resulting from multiplication of the exponential of these test statistics, we input $e^{\chi^2_m}$ in the following form: $e^{\chi^2_m} - e^{max_{m'}(\chi^2_{m'})}$ where the maxium is taken over all χ^2 statistics in GWA signal i.

In the models discussed in the results section we specified twelve annotations of sequence variants based on VEP with the addition of ASM-QTLs and DHS footprints[6] as listed out in Supplementary Table 1.

RNA isolation and sequencing

Total RNA was isolated from PaxGene (Qiagen) blood tubes using the Chemagic Total RNA Kit special (Perkin Elmer). The quality and quantity of the RNA was assessed using either the Agilent BioAnalyzer (RNA 600 Nano kit) or the LabChip GX instrument (Perkin Elmer) using the 96-well RNA kit.

Indexed cDNA libraries were prepared using the TruSeq RNA sample preparation v2 kit from Illumina (96-well plate format). In short, between 0.1-1 µg of total RNA was used for poly-A mRNA capture using oligo-dT attached magnetic beads. cDNA synthesis was done using SuperScript II (Invitrogen) and random hexamer priming. End-repair, 3´-adenylation, ligation of dual indexed adaptors (IDT for Illumina), AMPure XP bead purification and PCR amplification was performed as described by Illumina. Quantity and quality of the resulting cDNA sequencing libraries was assessed using the LabChip GX, followed by standard dilutions to 3 nM. Samples were stored at -20°C in barcoded 96-well trays, with all reagent and sample handling workflows registered in an in-house LIMS.

Further quality assessment was performed by doing pool sequencing (96 samples/pool) on an Illumina MiSeq instrument in order to optimize cluster densities and assess insert size, sample diversity etc.

Samples were pooled, clustered on to flowcells using either Illumina´s cBot and the TruSeq PE cluster kits (4-8 samples/pool/lane), or on NovaSeq S4 flowcells (24 samples/pool/lane) using on-board clustering, respectively. Paired-end sequencing (2x125 cycles) was performed with either HiSeq2500/HiseqX instruments using the TruSeq SBS kits from Illumina or NovaSeq instruments using the S4 flowcells, following the XP workflow.

RNA phasing to parental chromosomes

We aligned RNA-sequences separately to maternal and paternal genome references. The diploid haplotype reference was created for each individual and long-range phased haplotypes[57]. We assigned phasing of each RNA-sequenced fragment to maternal or paternal inheritance based on higher alignment score from STAR-aligner (*v2.5.3a*). We parsed fragments overlapping heterozygous variants to inspect concordant variant alleles and assigned to gene transcript based on genomic location in Ensembl v87 (limiting to BASIC and Support II level transcripts)[62]. We aggregated maternal and paternal fragment counts per gene. We removed measurements for individuals with a high rate of fragment multimapping or inconsistency between fragment alignment phasing and the phase of detected allele in fragment sequence.

We tested for haplotype-specific associations between methylation and expression by fitting least squares regression models wherein 5-mCpG rates measured on each of the two haplotypes of each individual was the outcome variable and the fraction of mRNA expression originating from one of the two haplotypes of each individual was the main predictor with the addition of the following covariates: age of individuals at blood draw, sex [male, female], parent of origin [maternal, paternal], measurements of cell type abundance [quantiative] along with the first five principal components derived from singular value decomposition analyses of methylation across the cohort, see further under the **Covariates** section in methods.

We tested mRNA isoforms if the TSS was located within 100kb of the MDS. We tested MDSs where at least 6 sequences were available for estimating the 5-mCpG rates (**Supplementary Figure 2C**) and, we therefore apply the same criteria on mRNA isoforms. Further, we performed the regression if >100 observations were informative for the outcome, main predictor and covariates.

The same procedure was followed to identify associations between individual CpG units and mRNA isoforms.

Analysis of causality between methylation and expression

We performed Mendelian Randomization-Steiger test[43] to assess whether CpG methylation is more likely to affect mRNA expression or if mRNA expression is more likely to affect CpG methylation using 1,369 ASM-QTL sequence variants associating with both as instruments. Mendelian Randomization-Steiger test involves comparing the correlation of genotypes (G) with CpG methylation (M) (ρ_{GM}) to their correlation with expression (E) (ρ_{GE}). If more sequence variants correlate more strongly with CpG methylation the conclusion is that CpG methylation is more likely to causally affect expression rather than the *vice versa*. CpG methylation and mRNA expression are both measured imprecisely which will reduce their correlation with the sequence variants. The reliability of the result from Mendelian Randomization-Steiger test is estimated by the R statistic[43]. We also assessed how relatively more imprecise the mRNA expression measurements needed to have been for us to erroneously observe the ASM-QTL variants associating more strongly with CpG methylation than they did with mRNA expression by finding the minimum value r such that $r\rho_{GE}$ is less than ρ_{GM} half the time.

Dissecting the effects of methylation on expression

Let G denote the allele count of a sequence variant standardized to have mean 0 and variance 1, M denote methylation levels, and E denote expression levels. Let also assume the model that sequence variants affect methylation and expression through distinct mechanisms and that methylation and expression are normally distributed given G, then:

$$M|G \sim \mathcal{N}(\beta_{GM}G, \sigma_M^2)$$

$$E|G \sim \mathcal{N}(\beta_{GE}G, \sigma_E^2)$$

Under this model we can calculate the expectation of expression given methylation:

$$\mathrm{E}(E|M) = \mathrm{E}(\mathrm{E}(E|G,M)|M) = \mathrm{E}(\mathrm{E}(E|G)|M) = \mathrm{E}(\beta_{GE}G|M) = \beta_{GE}\mathrm{E}(G|M) = \frac{\beta_{GM}\beta_{GE}}{\sigma_M^2 + \beta_M^2}M$$

$$= \frac{\beta_{GM}\beta_{GE}}{\mathrm{Var}(M)}M$$

This tells us that the expectation of the regression coefficient we get when regressing expression on methylation, β_{ME}, is $\frac{\beta_{GM}\beta_{GE}}{\mathrm{Var}(M)}$, or equivalently that $\mathrm{Var}(M)\beta_{ME} = \beta_{GM}\beta_{GE}$

If we observe methylation or expression with random normally distributed noise, then this identity is not affected.

If methylation affects expression or if the sequence variants affect methylation and expression through a common mechanism, e.g. TF binding, then the effect of methylation on expression will be greater than predicted by the effects of the sequence variant on methylation and expression:

$$\mathrm{Var}(M)\beta_{ME} \geq \beta_{GM}\beta_{GE}$$

Data availability

ASM-QTL summary statistics are available upon request from our website (www.decode.com/summarydata/), and can be used without restrictions, by clicking "Summary data" for this article. The legend for the data files can be found in the Supplementary information file, or by clicking "Read me file" on our website. Nanopore whole-genome and RNA sequencing data are not publicly available because of Icelandic state law. However, sequence variants identified in the Icelandic population using whole-genome sequencing have been deposited at the European Variant Archive under accession PRJEB15197 (www.ebi.ac.uk/ena/browser/view/PRJEB15197).

Data from the following publicly available databases were used in the study:

GWAS Catalog (Trait-associated sequence variants, all studies v1.0.3): https://www.ebi.ac.uk/gwas/

GTEx project (Cis-acting expression QTLs, bulk tissue): https://gtexportal.org/home/

eQTLGen (Cis-acting expression QTLs, phase I): https://www.eqtlgen.org/index.html

Ensembl v.87 (Transcript isoforms): https://www.ensembl.org/index.html

VEP (Sequence variant annotations): https://www.ensembl.org/info/docs/tools/vep/index.html

NCBI reference genome assembly (DNA sequence): https://www.ncbi.nlm.nih.gov/

ENCODE ChIP-seq data (ENCSR072QBN, ENCSR254XTB, ENCSR267YXV, ENCSR437MHW, ENCSR586POT, ENCSR393SYU, ENCSR167JFX, ENCSR785YRL): https://www.encodeproject.org/

ENCODE cCRE data (Candidate cis-regulatory elements, V3): https://screen.encodeproject.org

Fantom5 project (eRNA, CAGE-seq for UBERON:0000178): https://fantom.gsc.riken.jp/5/

AlleleDB (ASB, allele-specific binding): http://alleledb.gersteinlab.org/

LOFTEE (High quality annotation for loss of function sequence variants): https://github.com/konradjk/loftee

Code availability

DNA samples were analyzed with two versions of our pipeline, v3 (5761 R9 flowcells) and v4 (3145 R9 flowcells). The main difference between the pipelines is the version of the basecaller. In v3

squiggle data from PromethION was basecalled using Guppy 3.3.0 (https://github.com/nanoporetech/remora), 3826 flowcells, using either the 'flipflop' or 'hac' model or 3.2.2 (536 flowcells), 3.6.0 (675 flowcells) and 4.0.14 (724 flowcells) using the 'hac' model. In v4, all data was basecalled using guppy 5.0.11, using the 'sup' model (dna_r9.4.1_450bps_sup_prom.cfg). All 7,179 individuals basecalled with guppy had a minimum reference-genome-aligned sequencing coverage of at least 10x at the time of analysis. Basecalled reads were mapped to the human reference genome GRCh38 with minimap2 (https://github.com/lh3/minimap2), versions 2.14-r883 (5748 flowcells), 2.17-r941 (13 flowcells) and 2.22-r1105 (3145 flowcells). The aligned reads were sorted using samtools sort and stored in a BAM file. Nanopolish v0.11.0 and v0.13.3 were used to detect CpG methylation in nanopore sequences. Data analysis was performed in R (v3.6.0; https://www.r-project.org/), and the codes were deposited on our Github page (https://github.com/DecodeGenetics/) under the repository name: Stefanssonetal-Nature-Genetics-2024[63].

Methods-only references

54. Li H. Minimap2: pairwise alignment for nucleotide sequences, *Bioinformatics*, **34**(18):3094-3100.

55. Ebbert MTW. et al. Systematic analysis of dark and camouflaged genes reveals disease-relevant genes hiding in plain sight. *Genome Biol*. **20**(1):97 (2019).

56. Eggertsson HP. et al. Graphtyper enables population-scale genotyping using pangenome graphs. *Nat Genet*. **49**(11):1654-1660 (2017).

57. Kong A. et al. Parental origin of sequence variants associated with complex diseases. *Nature*. **462**(7275):868-74 (2009).

58. Halldorsson BV. et al. The rate of meiotic gene conversion varies by sex and age. *Nat Genet*. **48**(11):1377-1384. (2016).

59. Bates D, Eddelbuettel. Fast and elegant numerical linear algebra using the RcppEigen package. *Journal of Statistical Software*. **52**(5), 1–24 (2014).

60. McLaren W. et al. The Ensembl Variant Effect Predictor. *Genome Biol*. **17**(1):122 (2016).

61. Karczewski KJ. et al. The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature*. **581**(7809):434-443 (2020).

62. Cunningham F. et al. Ensembl 2022. *Nucleic Acids Res*. **50**(D1):D988-D995 (2022).

63. DOI: 10.5281/zenodo.12103852.

64. Sigurpalsdottir. et al. A comparison of methods for detecting DNA methylation from long-read sequencing of human genomes. *Genome Biol*. **25**(1):69 (2024).

Paper III

Nanopore sequencing identifies age-associated disruption of imprinting fidelity in the human genome

Authors: Brynja Sigurpalsdottir[1,2], Guillaume Holley[1], Sverrir Þ. Sverrisson[1], Droplaug N. Magnusdottir[1], Pall I. Olafsson[1], Arnaldur Gylfason[1], Olafur Þ. Magnusson[1], Gisli Masson[1], Olafur A. Stefansson[1], Bjarni V. Halldorsson[1,2]

Affiliations

1. Amgen deCODE genetics, Sturlugata 8, Reykjavik, Iceland
2. School of Technology, Reykjavik University, Menntavegur 1, Reykjavik, Iceland

Abstract

Aging is accompanied by widespread DNA methylation changes[1], yet their full genomic scope and parent-of-origin dynamics remain poorly understood. Here, we applied nanopore long-read sequencing to 7,284 whole blood samples enabling methylation measurements of 17,959,684 million high-quality CpG units[2]. Over 20% of the measured CpGs undergo age-associated changes, predominantly hypomethylation. From these data, we constructed a methylation aging clock from 1,373 CpGs, that outperforms gold-standard methylation clocks[3,4]. Importantly, phasing the methylation to parental haplotypes enabled systematic analysis of age effects in parent-of-origin specific context, uncovering 702 CpGs with parent-of-origin specific age-association, most of which were located at imprinted regions. At the DIRAS3 imprinted locus, we detected age-dependent hypermethylation on the active paternal allele, indicative of erosion of imprinting fidelity. Together, these findings establish nanopore sequencing as a powerful tool to map both genome-wide and parent-of-origin specific signatures of methylation aging, revealing imprinting instability as a hallmark of human aging, with potential relevance to disease susceptibility.

Main

DNA methylation, primarily occurring at cytosines within CG dinucleotides (CpG sites), has long been suspected to have roles in maintaining cellular identity and cell-type specification[5], X chromosome inactivation[6] and chromatin and transcriptional regulation[7]. As individuals age, changes in methylation levels occurs. These methylation changes form the basis for highly accurate "methylation clocks" that aim to estimate biological age[3,4,8,9]. While these clocks have enabled valuable insights into age-related phenotypes and risk stratification[8,9], their mechanistic foundation remain poorly understood[10].

A key limitation of previous studies[3,4,7–9,11,12] lies in their reliance on array-based technologies, which cannot resolve methylation at the level of individual haplotypes. Additionally array-based methods only interrogate fraction of the methylome ~30 million CpG sites found in the human genome. As a result, little is known about how aging affects allele-specific methylation patterns or the genome-wide changes with age. This is especially relevant for imprinted regions, where methylation is specified in a parent-of-origin specific manner[12,13]. While methylation within these regions has traditionally been considered stable[14], recent studies suggest that imprinting fidelity may deteriorate at some imprinted loci with age[11], raising important questions about the long-term stability of these imprinted regions.

To map the landscape of age-associated changes in methylation, we analysed whole blood samples from 7,284 Icelanders (3,497 males, 3,787 females), using Oxford nanopore long-read sequencing[2,15]. Here, 17,959,684 million autosomal CpG units, including 22,909,253 CpG sites, satisfied our criteria for high-quality methylation measurements (Supplementary Table 1)[2].

3,739,875 CpG units (20.8%) were significantly associated with chronological age after covariate adjustment and Bonferroni correction ($p<2.8\cdot10^{-9}=0.05/18\cdot10^{6}$, Supplementary Table 2). We found that the mean methylation of age-associated CpGs (μ=0.74) is lower than that of CpGs without age association (μ=0.80, Welch two sample t-test, $p<2.2\cdot10^{-16}$, Supplementary Fig. 1A). The vast majority, or 92.4%, exhibited progressive hypomethylation (negative effect size), whereas 7.6% showed hypermethylation (positive effect size) with age (Fig. 1A). Age-associated hypomethylation mostly occurs in highly methylated CpGs, while age-associate hypermethylation spanned a wider range of methylation levels (Supplementary Fig. 1B). These results are consistent with loss of methylation fidelity with age (Supplementary Fig. 1C)[1].

Age-associated CpGs showed distinct patterns of enrichment across genomic regions depending on whether they exhibited gain or loss of methylation (Fig. 1B). Consistent with previous studies[3,16,17], CpGs that showed hypermethylation with age were enriched in repressed polycomb regions (ReprPC), flanking TSS, bivalent enhancers, CpG islands and coding exons (Fig. 1B, upper panel). Most of the age-associated CpGs however, showed hypomethylation with age, and these CpGs were more

prevalent in introns, intergenic and quiescent regions (Fig. 1B, lower panel). We further demonstrated that age associated CpGs exhibit higher residual variance compared to non-age associated CpGs, indicating more unexplained variance in their methylation levels (Supplementary Fig. 2). Taken together, these patterns suggest that most of the observed age-associated methylation changes occur within non-functional, less stable genomic regions[4,18].

We developed a methylation clock based on 1,373 CpG units (Supplementary Table 3), achieving median absolute error (medAE) of 2.22 years and 2.43 years, in training and test set respectively. Our clock outperforms two cold standard methylation clocks: Horvath's multi tissue methylation clock (train medAE=2.9 years, test medAE = 3.6 years)[3] and Hannum's methylation clock (overall medAE=3.9 years, test medAE=4.9 years)[4]. Model performance remained robust even after CpGs identified as predictive in the model, i.e. those with non-zero coefficient, were iteratively removed and the model retrained on the remaining CpGs, indicating redundancy among predictive sites (Supplementary material Table 2). 398 (29.0%) and 975 (71.0%) of the CpG units included in the clock were hyper- and hypomethylated, respectively. This distribution of hyper- and hypomethylated CpGs contrast sharply with that observed in the age-associated CpG set, where hypermethylated CpGs where 4.95 times more frequent (chi-square test, $p<1.37\cdot10^{-195}$).

1,274,315 CpG units (7.1%) showed age-associated methylation changes on one (54.2%) or both (45.8%) of the two parental haplotypes after Bonferroni correction ($p<2.8\cdot10^{-9}$, Fig. 1A, Supplementary Table 4). Of these CpGs, 99.7% associated with age regardless of parent-of-origin status. Effect sizes across haplotypes were highly correlated (Pearson's $r = 0.94$, Supplementary Fig. 3), reflecting symmetric aging effects at most loci. Interestingly, 702 CpGs across the genome (Supplementary Table 5), showed significant difference in effect size (z-test, $p<3.9\cdot10^{-8} = 0.05/1{,}274{,}315$). 300 CpG units (42.7%) were not amongst the ~3.7 million age-associated CpG-units, identified using unphased data, suggesting that these age effects cannot be identified without using parent-of-origin assigned methylation. Differences in effect sizes were correlated with differences in methylation levels between the two parental haplotypes (*Pearson's r* = 0.2, Fig. 1C), suggesting that differences in effect sizes mostly emerge nearby or within imprinted regions. 664 CpG units (94.5%) out of 702 that exhibited parent-of-origin specific age association and difference in effect sizes, were within 100kb of previously reported imprinted regions[13,19–21] (Supplementary Fig. 4). For comparison, these imprinted regions including 100kb windows up- and downstream cover about 1.4% of the genome.

Of the 702 CpGs, 78.1% showed paternal-specific age association and 21.9% maternal-specific (Fig 1A). Notably, 535 CpGs (76.2%) associated with the unmethylated or non-imprinted allele, which is consistent with a commonly observed pattern in imprinted regions where paternal alleles are typically unmethylated and maternal alleles methylated[22]. As for unphased associations, we typically see

patterns of hypermethylation on the unmethylated haplotype and hypomethylation on the methylated haplotype, indicative of partial loss of imprinting with age (Fig. 2A,B). Together, these findings support a progressive erosion of imprinting fidelity with age in imprinted regions.

Among all imprinted regions, a region found within exon 2 of *DIRAS3* (DIRAS3:Ex2-DMR)[13], stood out as the strongest example of age-related parent-of-origin specific methylation changes (Fig. 2A,B). *DIRAS3* is a maternally imprinted tumour suppressor gene, that is normally expressed exclusively from the paternal allele[23]. As individuals age, we observed progressive methylation on the paternal allele (Fig. 2C, left panel), while the maternal allele remained methylated (Fig 2C, right panel), suggesting that progressive loss of *DIRAS3* occurs due to age-related processes. Notably, these changes occur during adolescent development and early adulthood. Downregulation of *DIRAS3* has been reported in many cancers, including ovary, breast, brain, prostate and thyroid[23–25]. We postulate that age-driven paternal hypermethylation at this locus may contribute to cancer susceptibility later in life.

Our observations at *DIRAS3* suggest that age-dependent methylation changes can destabilize imprinted loci. Building on this example, our study highlights more broadly how long-read sequencing can uncover the full landscape of age-dependent methylation changes, from genome-wide patterns to haplotype specific age effects. By interrogating nearly 18 million hq-CpGs across the genome, we identified widespread age-related methylation changes, developed a highly accurate methylation aging clock, and demonstrated that long-read sequencing can capture methylation changes with greater resolution than previous methods. A key advantage of our work is the ability to phase methylation data to parental haplotypes, which enabled systematic analysis of parent-of-origin specific dynamics. This revealed that age-dependent changes are not only widespread but also asymmetric across haplotypes, with clear disruption of imprinting stability at loci such as *DIRAS3*.

Together, these findings establish long read nanopore sequencing as a powerful framework for studying the biology of human aging. The ability to detect both genome-wide and haplotype-specific methylation changes associated with age opens new opportunities for understanding how age-related methylation dynamics influence disease risk, and for developing biomarkers and therapeutic strategies aimed at preserving methylation stability across the lifespan.

Figures

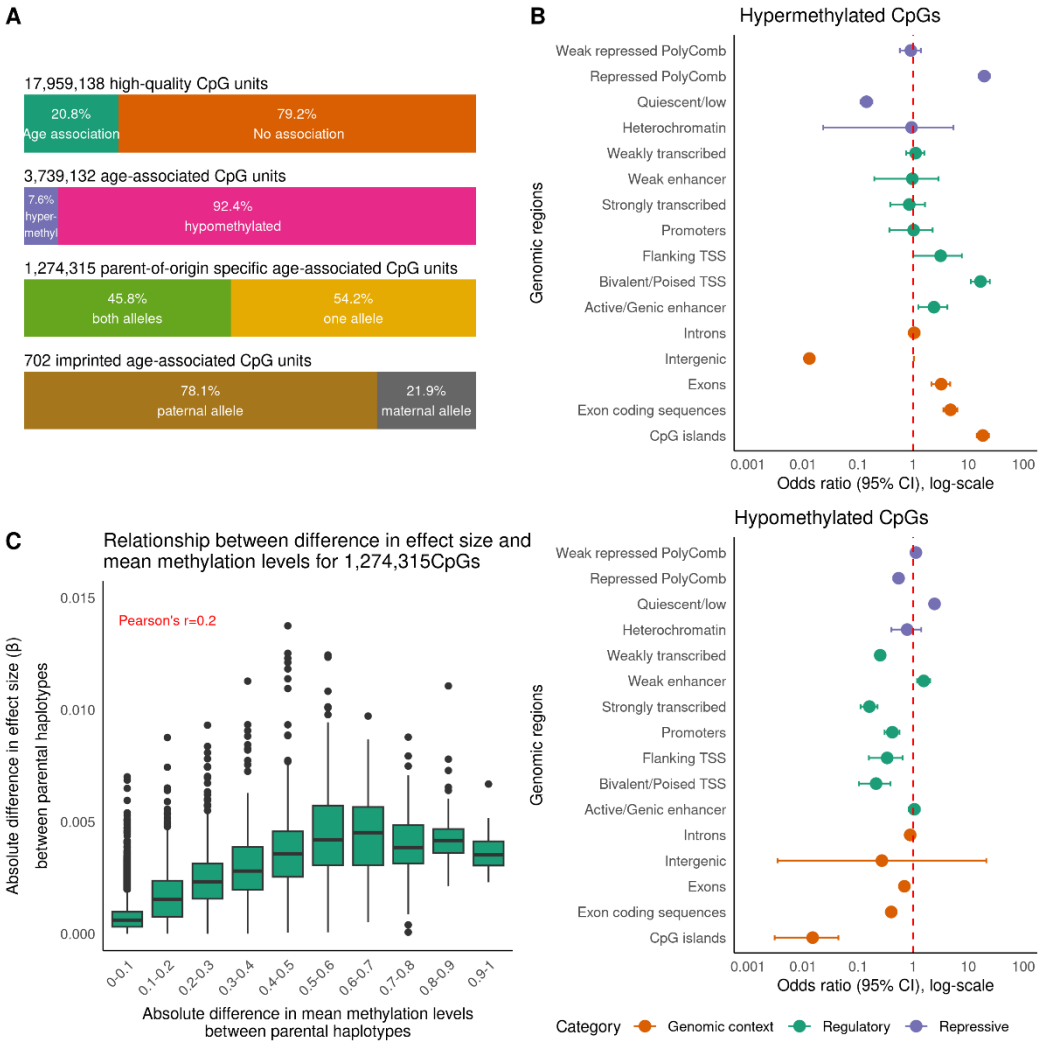


Fig. 1 Age-associated DNA methylation changes across the genome with age.

a. Overview of genome-wide age association methylation patterns. Among 17.96 million CpG units, 20.8% associate with age, of which 7.6% and 92.4% showed hyper- and hypomethylation, respectively. 1,274,315 parent-of-origin-specific associations, with 46.4% of associations reaching significance thresholds on both alleles and 53.6% on one allele. 702 show parent-of-origin specific associations with imprinted methylation, thereof 80.2% associate with the paternal allele and 19.8% with the maternal allele. **b.** Enrichment and depletion of age-associated CpGs across genomic regulatory regions. Odds ratios and 95% confidence intervals shown for each region class. Dashed red line indicates odds ratio = 1 (no enrichment). Shown for hypermethylated age-associated CpGs (upper panel) and hypomethylated age-associated CpGs (lower panel), separately. **c.** Boxplot showing the relationship between differences in mean methylation levels between parental haplotypes (x-axis) and the absolute difference in effect size (β) between haplotypes (y-axis) for 1,274,315 CpG units with parent-of-origin specific age-association.

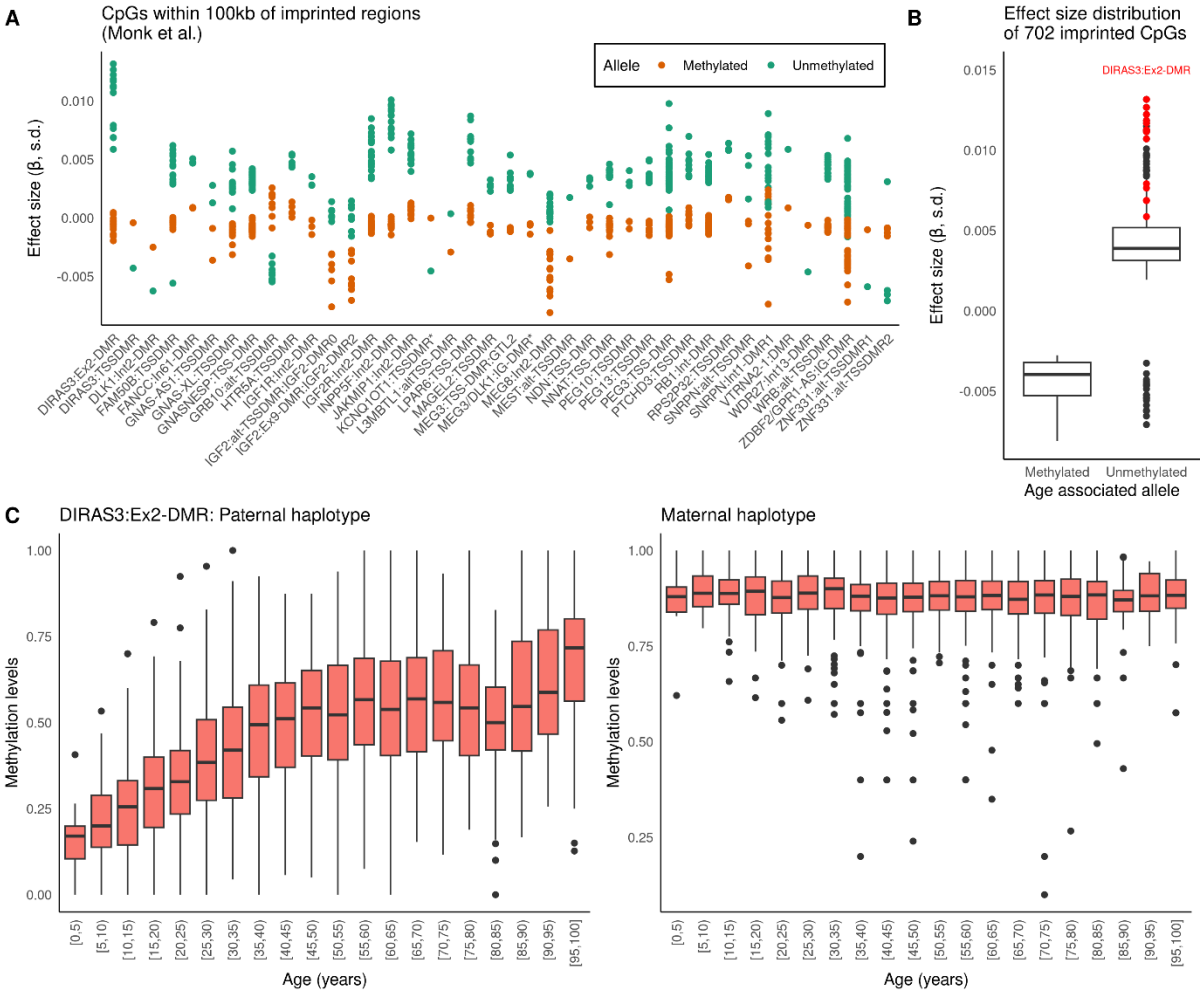


Fig. 2 Haplotype-specific methylation associated with age within imprinted regions.

a. Distribution of effect sizes for CpGs located within 100kb of 40 imprinted regions listed out in Monk et al.[13] (x-axis). Y-axis is expressed in standard deviation (s.d.) units. **b.** Box plot showing effect size for methylated and unmethylated allele. CpGs within the imprinted region at exon 2 of *DIRAS3* shown in red. Y-axis is expressed in s.d. units. **c.** Boxplot showing methylation levels for CpGs with 5x coverage or more (y-axis) of the paternal (left panel) and maternal haplotype (right panel) across age, binned by 5 years (x-axis) in DIRAS3:Ex2:DMR.

References

1. Wang, K. *et al.* Epigenetic regulation of aging: implications for interventions of aging and diseases. *Signal Transduction and Targeted Therapy* vol. 7 Preprint at https://doi.org/10.1038/s41392-022-01211-8 (2022).
2. Sigurpalsdottir, B. D. *et al.* A comparison of methods for detecting DNA methylation from long-read sequencing of human genomes. *Genome Biol* **25**, (2024).
3. Horvath, S. DNA methylation age of human tissues and cell types. *Genome Biol* **14**, R115 (2013).
4. Hannum, G. *et al.* Genome-wide Methylation Profiles Reveal Quantitative Views of Human Aging Rates. *Mol Cell* **49**, (2013).
5. Ehrlich, M. *et al.* Amount and distribution of 5-methylcytosine in human DNA from different types of tissues or cells. *Nucleic Acids Res* **10**, (1982).
6. Li, E., Beard, C. & Jaenisch, R. Role for DNA methylation in genomic imprinting. *Nature* **366**, (1993).
7. Deaton, A. M. & Bird, A. CpG islands and the regulation of transcription. *Genes Dev* **25**, (2011).
8. Levine, M. E. *et al.* An epigenetic biomarker of aging for lifespan and healthspan. *Aging* **10**, 573–591 (2018).
9. Lu, A. T. *et al.* DNA methylation GrimAge strongly predicts lifespan and healthspan. *Aging* **11**, (2019).
10. Bell, C. G. *et al.* DNA methylation aging clocks: Challenges and recommendations. *Genome Biology* vol. 20 Preprint at https://doi.org/10.1186/s13059-019-1824-y (2019).
11. Mancino, S. *et al.* Exploring the Stability of Genomic Imprinting and X-Chromosome Inactivation in the Aged Brain. *Aging Biology* **2**, 20240030 (2024).
12. Li, Y. & Sasaki, H. Genomic imprinting in mammals: Its life cycle, molecular mechanisms and reprogramming. *Cell Research* vol. 21 Preprint at https://doi.org/10.1038/cr.2011.15 (2011).
13. Monk, D., Mackay, D. J. G., Eggermann, T., Maher, E. R. & Riccio, A. Genomic imprinting disorders: lessons on how genome, epigenome and environment interact. *Nature Reviews Genetics* vol. 20 Preprint at https://doi.org/10.1038/s41576-018-0092-0 (2019).
14. Woodfine, K., Huddleston, J. E. & Murrell, A. Quantitative analysis of DNA methylation at all human imprinted regions reveals preservation of epigenetic stability in adult somatic tissue. *Epigenetics Chromatin* **4**, (2011).
15. Stefansson, O. A. *et al.* The correlation between CpG methylation and gene expression is driven by sequence variants. *Nat Genet* **56**, 1624–1631 (2024).
16. Lu, A. T. *et al.* Universal DNA methylation age across mammalian tissues. *Nat Aging* **3**, (2023).
17. Field, A. E. *et al.* DNA Methylation Clocks in Aging: Categories, Causes, and Consequences. *Molecular Cell* vol. 71 Preprint at https://doi.org/10.1016/j.molcel.2018.08.008 (2018).
18. Maegawa, S. *et al.* Widespread and tissue specific age-related DNA methylation changes in mice. *Genome Res* **20**, (2010).
19. Joshi, R. S. S. *et al.* DNA Methylation Profiling of Uniparental Disomy Subjects Provides a Map of Parental Epigenetic Bias in the Human Genome. *Am J Hum Genet* **99**, (2016).

20. Court, F. *et al.* Genome-wide parent-of-origin DNA methylation analysis reveals the intricacies of human imprinting and suggests a germline methylation-independent mechanism of establishment. *Genome Res* **24**, (2014).

21. Zink, F. *et al.* Insights into imprinting from parent-of-origin phased methylomes and transcriptomes. *Nat Genet* **50**, (2018).

22. Moore, G. E. *et al.* The role and interaction of imprinted genes in human fetal growth. *Philosophical Transactions of the Royal Society B: Biological Sciences* vol. 370 Preprint at https://doi.org/10.1098/rstb.2014.0074 (2015).

23. Bildik, G., Liang, X., Sutton, M. N., Bast, R. C. & Lu, Z. DIRAS3: An Imprinted Tumor Suppressor Gene that Regulates RAS and PI3K-driven Cancer Growth, Motility, Autophagy, and Tumor Dormancy. *Molecular Cancer Therapeutics* vol. 21 Preprint at https://doi.org/10.1158/1535-7163.MCT-21-0331 (2022).

24. Yu, Y. *et al.* Biochemistry and Biology of ARHI (DIRAS3), an Imprinted Tumor Suppressor Gene Whose Expression Is Lost in Ovarian and Breast Cancers. *Methods in Enzymology* vol. 407 Preprint at https://doi.org/10.1016/S0076-6879(05)07037-0 (2006).

25. Soboska, K. *et al.* Expression of RASSF1A, DIRAS3, and AKAP9 Genes in Thyroid Lesions: Implications for Differential Diagnosis and Prognosis of Thyroid Carcinomas. *Int J Mol Sci* **25**, (2024).

26. Taylor, D. P. *et al.* HerediGene Population Study IT infrastructure: A model to support genomic research recruitment and precision public health. *AMIA Annu Symp Proc* **2023**, (2023).

27. Schneider, V. A. *et al.* Evaluation of GRCh38 and de novo haploid genome assemblies demonstrates the enduring quality of the reference assembly. *Genome Res* **27**, (2017).

28. Li, H. Minimap2: Pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, (2018).

29. Simpson, J. T. *et al.* Detecting DNA cytosine methylation using nanopore sequencing. *Nat Methods* **14**, 407–410 (2017).

30. Boix, C. A., James, B. T., Park, Y. P., Meuleman, W. & Kellis, M. Regulatory genomic circuitry of human disease loci by integrative epigenomics. *Nature* **590**, (2021).

31. Frankish, A. *et al.* GENCODE 2021. *Nucleic Acids Res* **49**, (2021).

32. Eggertsson, H. P. *et al.* Graphtyper enables population-scale genotyping using pangenome graphs. *Nat Genet* **49**, (2017).

33. Kong, A. *et al.* Detection of sharing by descent, long-range phasing and haplotype imputation. *Nat Genet* **40**, (2008).

Methods

Dataset

We nanopore sequenced DNA isolated from whole-blood samples from 7,284 Icelanders (3,497 male, 3,787 female) participating in various studies at deCODE genetics. Participants were born between 1876 and 2014, with median birth year of 1960 for males and 1958 for females. 778 of the individuals (369 males, 409 females) were deceased at the time of the study. All individuals gave informed consent, and all personal identifiers were encrypted by an external agent before being imported into deCODE's database. For validation we used 545 samples from independent cohort from Intermountain USA (223 males, 322 females) born between 1933 and 2002. The participants were recruited by HerediGene: Population study, a large-scale collaboration between Intermountain Healthcare, deCODE genetics and Amgen, Inc[26].

We calculate chronological age of individuals as years from date of birth to date of blood draw (sampling date). Only 31.4% of samples had sampling date registered. For the remaining samples, we calculated the age as years from the date of registration (when sample was added to the database), which was usually within few days of sampling date (median = 22 days, mean = 310 days). We found 5 individuals with age over 110 years (Supplementary Fig. 5). Those individuals were all registered in March or April 2016. To ensure high-quality training data, we exclude all 53 samples registered in these months from our training set. Additionally, we removed 37 samples with year of death before sampling date. These samples were included in the test set.

Sample processing and sequencing

As previously described[2,15], DNA from whole blood was extracted using the Chemagic method (PerkinElmer), an automated procedure that involves the use of M-PVA magnetic beads. Sequencing libraries were generated using the SQK-LSK109 ligation kit from ONT. Sample input varied from 1 to 5 µg DNA, depending on the exact version of the preparation kit and the flowcell type used for the PromethION sequencing. Samples were loaded onto PromethION R9.4.1 flowcells following ONT standard operating procedures. Sequencing was performed on PromethION devices. Basecalling was performed using guppyrh-5.0.11 and reads were aligned to reference genome GRCh3 [27] 8 using minimap 2 (v. 2.22-r1105-dirty)[28]. For quality control, flowcells with error rate greater than or equal to 15%, with sequencing coverage less than or equal to 3x, n50 less than or equal to 7000bp and mean methylation levels lower than or equal to 0.5 were removed.

DNA methylation detection with Nanopolish

As previously described[2,15], all flowcells were methylation called using Nanopolish [29], version 0.13.3. In short, Nanopolish assigns a log-likelihood ratio for each CpG in a read. We interpret values above 1.921 as 5mC methylation and lower than -1.921 as unmodified C. We exclude ambiguous

methylation predictions (-1.921 < LLR < 1.921). Nanopolish groups CpGs within 10bp distance and assigns the same methylation status to all CpGs within the group. We calculated strand difference and fraction of non-ambiguous reads out of all reads and use these measures to define set of high-quality CpGs[2] We include CpGs within dark regions and CpGs within 5 base pairs of common SNPs, instead we set the methylation levels within 5bp of common variant, as missing values for carriers (Supplementary Fig. 6). Finally we compute methylation levels as $n_{methylated}/(n_{unmethylated} + n_{methylated})$, where $n_{unmethylated}$ and $n_{unmethylated}$ are the number of reads deemed methylated and unmethylated, respectively.

Principal components

To account for potential batch effects and other latent variables, we computed principal components (PCs) based on a random subset of 200,000 high-quality CpG sites (hq-CpGs). PCs were derived separately for each autosome using leave-one-chromosome-out approach to avoid circularity. PCs were computed using age and gender adjusted methylation data, for each flowcell using the *PCA* function *sklearn.decomposition* in python. To calculate variance explained we calculate PCs over all autosomes. The top 10 PCs, selected based on the proportion of variance explained (Supplementary Fig. 7), were included as covariates in all regression models.

These components capture major sources of variation, including blood cell composition, sampling quality, and sequencing batch effects[15]. Finally, for each sample, we derived a sample-level PC profile by averaging across all associated flowcells, which yields similar results as per flowcell (Supplementary Fig. 8). These averaged PCs were then used as covariates.

Age association

To assess the relationship between DNA methylation and chronological age, we applied a linear regression model, implemented with *lm* in *R*. For each CpG site, we rank first normal transformed methylation levels (CpG_{rank_normal}) and modelled as a function of age, with additional covariate adjustment for gender, the first ten PCs to account for technical and biological variability and genotypes of strongest cis-acting ASM-QTL ($genotype_{best\_marker}$).

The model was specified as:

$$CpG_{rank\_normal} \sim age + gender + PC_1 + \cdots + PC_{10} + genotype_{bestMarker}$$

We assessed statistical significance using p-values from linear regression and reported effect sizes as regression coefficients. Positive effect size indicates hypermethylation while negative indicates hypomethylation. To control for multiple testing, we applied Bonferroni correction by adjusting the significance threshold based on the total number of tests performed. To validate that our results are

not an artifact of our data transformations we also calculate association without covariates and without standardizing methylation values and compared the p-values and direction of effect sizes.

We calculate residual variance in methylation (or root mean squared error, RMSE) for each CpG with the following formula.

$$RMSE = \sqrt{\frac{\sum (y_i - \tilde{y}_i)^2}{n}}$$

$(y_i - \tilde{y}_i)^2$ are the squared residuals and n is total number of samples.

Evaluating robustness of CpG-Age associations

To evaluate whether the observed associations between DNA methylation at CpG sites and chronological age could have arisen by chance, we performed a permutation test. Specifically, we tested the null hypothesis that methylation is not associated with age by randomly permuting the age variable across individuals and recalculating the association statistics for each CpG site. For each of the 17,959,684 CpGs, we repeated the age association analysis 3 times with the age values randomly permuted across individuals in each iteration, thereby breaking any real association while preserving the distribution of both variables. In each permutation, we recalculate the p-values from the linear models and counted the number of CpGs with Bonferroni significant association (Supplementary Table 6), yielding an empirical null distribution of the number of significant CpGs under the assumptions of no association. No age associated CpGs in the permutation test reached Bonferroni significance (Supplementary Table 6). Approximately 0.3% of CpGs with differing effect sizes between the parental haplotypes did not pass Bonferroni correction, corresponding to a false positive rate (FPR) of 0.3% for the imprinted associations.

Genomic enrichment analysis

All genomic regions were defined from the GRCh38 reference genome[27]. We used annotations from encode ChromHMM (18-state model)[30] and selected BSS01380_NEUTROPHIL cell line for all analysis. For validation we did the same analysis for the BSS00178_CD14_MONOCYTE cell line (Supplementary Fig. 9), yielding similar results. For genomic context (coding exons, exons, intergenic regions and introns) we used GENCODE (v48)[31]. We split the genome into 100kb windows. From each window we select one CpG and check whether the CpG is age associated and whether it was hyper- (effect size≥0) or hypomethylated (effect size<0). We then check whether the CpG is within the genomic region and create two contingency tables, for hyper- and hypomethylated counts separately. We used Fisher's exact test to calculate odds-ratios and p-values from the contingency tables.

Methylation phasing

As described in our previous study[15], DNA sequence variants were called using Graphtyper[32] based on 63,460 Illumina whole-genome sequenced individuals representing a subset of 173,025 SNP chip-typed individuals from the Icelandic population. The sequenced individuals were then phased to impute haplotypes and assign parent of origin into long-range phased chip-typed individuals[33].

The long-range phasing and assigned parent of origin, enabled us to assign nanopore sequences to maternal or paternal chromosomes[15]. Parental haplotypes were assigned by examining the phasing status of a set of 8,960,728 high-quality sequence variants, using heterozygous carriers of in-read sequence variants, allowing us to assign CpG methylation calls (methylated or unmethylated) to maternal and paternal chromosomes. Reads overlapping at least three heterozygous variants and where at least 70% of the variants where consistent with the phasing of one parent were phased and used for subsequent analysis.

For each CpG and each parental haplotype, we defined the haplotype methylation levels as the number of sequences of that parental haplotype that are methylated at the CpG divided by the total number of parental sequences with reliable methylation calls covering the CpG.

Construction and validation of a methylation aging clock

We use all 7,284 samples for either testing or training, split into 80% training (5,828 samples) and 20% testing (1,625 samples). We ensure high quality training set by excluding samples with registration date in March or April 2016 and with year of birth after sampling date, as this was an indication of error in registration. We further limit our training set to individuals between 18-100 years old. Combined, these criteria excluded 477 samples from the training set, but these samples were used for testing, the remaining 1,148 samples were randomly selected into the testing set.

We used *createDataPartition()* in *R*, stratified by age group and gender, to ensure balanced representation, to split the remaining 6,807 samples randomly into test and training. The training set consisted of 3,030 females and 2,798 males, while the test set consisted of 680 females and 945 males.

We first perform feature selection of CpGs using linear regression between methylation levels in CpGs and chronological age and select 1 million CpGs with lowest p-values. We excluded CpG sites with more than 80% missing values and for the remaining CpGs we mean imputed missing values per CpG, using *SimpleImputer(strategy="mean")* from the *scikit-learn* library in *python*. Subsequentially, all features were standardize using *StandardScaler*, also from *scikit-learn*.

We implemented a Lasso regression model with 5-fold cross-validation, using *GridSearchCV()* from *scikit-learn*, to tune the regularization parameter. We evaluated the performance using the negative mean squared error metric. We trained the final model on the full training dataset using the optimal alpha and included features (CpGs) selected by the Lasso model and evaluated the model performance

on both the remaining samples (test) and an independent validation cohort from Intermountain. Metrics reported include root mean squared error (RMSE), mean absolute error (MAE) and median absolute error (MedAE).

Methods only references

26. Taylor, D. P. *et al.* HerediGene Population Study IT infrastructure: A model to support genomic research recruitment and precision public health. *AMIA Annu Symp Proc* 2023, (2023).

27. Schneider, V. A. *et al.* Evaluation of GRCh38 and de novo haploid genome assemblies demonstrates the enduring quality of the reference assembly. *Genome Res* 27, (2017).

28. Li, H. Minimap2: Pairwise alignment for nucleotide sequences. *Bioinformatics* 34, (2018).

29. Simpson, J. T. *et al.* Detecting DNA cytosine methylation using nanopore sequencing. *Nat Methods* 14, 407–410 (2017).

30. Boix, C. A., James, B. T., Park, Y. P., Meuleman, W. & Kellis, M. Regulatory genomic circuitry of human disease loci by integrative epigenomics. *Nature* 590, (2021).

31. Frankish, A. *et al.* GENCODE 2021. *Nucleic Acids Res* 49, (2021).

32. Eggertsson, H. P. *et al.* Graphtyper enables population-scale genotyping using pangenome graphs. *Nat Genet* 49, (2017).

33. Kong, A. *et al.* Detection of sharing by descent, long-range phasing and haplotype imputation. *Nat Genet* 40, (2008).

Data availability

The sequence data cannot be made publicly available because Icelandic law and the regulations of the Icelandic Data Protection Authority prohibit the release of individual-level and personally identifying data. Data access can be granted only at the facilities of deCODE genetics in Iceland, subject to Icelandic law regarding data usage. Summary statistics are available upon request from our website www.decode.is/summarydata/ and can be used without restrictions by clicking “Summary data” for this article.

Code availability

Publicly available software used, and methods developed at Amgen deCODE genetics are described in Methods.

Author contributions

B.D.S, O.A.S and B.V.H. designed the study. B.D.S and B.V.H. implemented the software. B.D.S, O.A.S and B.V.H. analyzed the data and interpreted the results. G.H., and S.Þ.S. implemented the sequencing pipeline. P.I.O. and A.G. performed the long-range phasing of the variants, supervised by G.M. O.Þ.M. and D.N.S. performed the nanopore sequencing. B.D.S. wrote the initial version of the manuscript and B.V.H. and O.A.S. contributed to the subsequent versions. All authors contributed to the final version of the manuscript.

Acknowledgements

We thank colleagues at Amgen deCODE genetics for helpful discussions and feedback. The authors received no specific funding for this work.

Corresponding authors

Correspondence to Brynja D. Sigurpalsdottir or Bjarni V. Halldorsson.

Ethics declarations

Ethical statement

The study was approved by the National Bioethics Committee in Iceland (Approval no. VSN 14–015) and conducted in agreement with instructions issued by the Data Protection Authority in Iceland (PV_2017060950ÞS/–). All participating subjects signed informed consent. The personal identities of the participants and biological samples were encrypted by a third-party system approved and monitored by the Data Protection Authority.

Competing interest

All authors are employees of Amgen/deCODE genetics.

Discussion

Conclusions

A major contribution of this thesis to the field is the establishment of LRS, particularly NS, as a reliable platform for large-scale DNA methylation studies. Through comparison of existing methods, we demonstrated that LRS achieves both broad CpG coverage and high accuracy, while overcoming limitation of SRS, such as poor resolution in repeat regions. Importantly, we introduced filtering strategies that enhance the quality and interpretability of LRS methylation data. These methodological advances provide the technical basis for addressing deeper biological questions and set the stage for large-scale applications of LRS methylation data.

Beyond technical contributions, we clarified the complex interplay between genetic variation, DNA methylation and gene expression. By phasing the data into parental haplotypes and integrating with gene expression profiles, we showed that much of the correlation traditionally observed between methylation and gene expression is in fact driven by sequence variants. This finding highlights the need for caution when interpreting methylation-expression analysis, as genetic variation must be considered one of the key drivers of the methylation landscape. These insights redefine our understanding of the functional relevance of methylation and highlight the value of haplotype level analysis in disentangling genetic from methylation effect.

Last part of this thesis was to determine the extent to which DNA methylation sites across the genome are influenced by age. By applying NS to thousands of individuals across the lifespan, we found that roughly 20% of CpGs undergo age-associated methylation changes, most of which reflect progressive hypomethylation in non-coding regions. Importantly, assigning the methylation to parental haplotypes revealed that some age effects are parent-of-origin specific, uncovering erosion of imprinting fidelity at loci, including tumour suppressor *DIRAS3*. These findings highlight imprinting instability as hallmark of aging and point to possible mechanisms linking age-associations with disease susceptibility.

Future work

Benchmarking of technological improvements

The Icelandic cohort was sequenced using R9.4 flowcells from Oxford nanopore technologies (ONT), with methylation detection performed using Nanopolish. In future studies, sequencing will be carried out using new flowcell chemistry (R10.4) and methylation will be called using Dorado. Benchmarking will be essential to assess the performance of Dorado compared to existing methylation callers. However, we suspect that similar filtering strategies as proposed in this thesis for Nanopolish and Guppy, can be applied to Dorado data, to ensure robust and high-quality methylation data.

Expanding population diversity

A key priority will be extending the sequencing efforts beyond the Icelandic population, including samples from large-scale cohorts such as the UK biobank. This expansion will substantially increase statistical power, enable replication of findings across populations, facilitate the discovery of population-specific effects and increase diversity in our methylation dataset. Furthermore, sequencing rare disease cases, that remain unsolved by SRS, may reveal pathogenic variants located in genomic regions inaccessible to short-read methods. We could create diagnostic patterns for specific methylation disorders, thereby advancing both diagnostic yield and our understanding of disease mechanisms.

Genetic variation regulating DNA methylation

Building on our work in defining cis-acting that influence methylation (ASM-QTLs), we aim to extend our analysis to identify trans-meQTLs and to evaluate their contribution to complex traits. Genes located near trans-meQTLs may highlight regulators of DNA methylation and provide novel insights into epigenetic control. The ability to phase the methylation data will allow us to investigate trans-meQTLs associated with imprinted methylation, offering a unique opportunity to study the interaction between imprinting and distal regulatory variants.

DNA methylation and aging

Although methylation changes have been shown to occur extensively across the human lifespan, the full extent to which DNA methylation captures the biology of aging, still remains poorly understood. Future work will involve developing and applying novel statistical measures to quantify variance in DNA methylation. Grouping CpGs into co-regulated networks based on variance or shared age-associated effects could uncover higher-order patterns that are not evident when loci are studied individually. Longitudinal studies, in which multiple measures of DNA methylation are collected repeatedly across the lifespan, will be particularly valuable in disentangling stable age-associated

changes from stochastic variation. Additionally, combining methylation age, proteomics age, telomere length measurements could shed light on some of the underlying mechanism of methylation aging.

Multi-omics approaches

Integrating DNA methylation with other molecular layers, such as proteomics, represents an important next step. Multi-omics integration will make it possible to investigate how DNA methylation interacts with molecular phenotypes and with environmental exposures. For example, protein- and expression QTLs (pQTLS, eQTLS) could be analyzed jointly with meQTLs to reveal coordinated regulatory networks. Such integrative approaches can both improve our understanding of the role of methylation in disease and enhance the discovery of biomarkers and potential therapeutic targets. Future work can move beyond single-layer associations to uncover the complex interactions between genetics, methylation and the environment. Together, these efforts will not only refine methodological approaches and broaden our understanding of DNA methylation, but also pave the way for work in translational applications where methylation can inform disease risk prediction, personalized medicine and targeted therapeutic strategies.

Connecting DNA methylation to phenotypes

Building on the methylation-expression and methylation-genetic association analyses presented in this thesis, we plan to expand the investigation to link methylation variation with human traits and diseases. Moreover, we will explore machine learning approaches to predict complex traits, health outcomes and lifestyle factors, such as smoking behavior, from methylation profiles and assess the power of methylation signatures relative to existing biomarkers.

References

1. Deaton, A. M. & Bird, A. CpG islands and the regulation of transcription. *Genes Dev* 25, (2011).
2. Wilson, V. L. & Jones, P. A. DNA methylation decreases in aging but not in immortal cells. *Science (1979)* 220, (1983).
3. Christensen, B. C. *et al.* Aging and environmental exposures alter tissue-specific DNA methylation dependent upon CPG island context. *PLoS Genet* 5, (2009).
4. Lee, K. W. K. & Pausova, Z. Cigarette smoking and DNA methylation. *Frontiers in Genetics* vol. 4 Preprint at https://doi.org/10.3389/fgene.2013.00132 (2013).
5. Zeilinger, S. *et al.* Tobacco Smoking Leads to Extensive Genome-Wide Changes in DNA Methylation. *PLoS One* 8, (2013).
6. Portela, A. & Esteller, M. Epigenetic modifications and human disease. *Nature Biotechnology* vol. 28 Preprint at https://doi.org/10.1038/nbt.1685 (2010).
7. Wang, C. *et al.* DNA methylation-based biomarkers of age acceleration and all-cause death, myocardial infarction, stroke, and cancer in two cohorts: The NAS, and KORA F4. *EBioMedicine* 63, 103151 (2021).
8. Hao, X. *et al.* DNA methylation markers for diagnosis and prognosis of common cancers. *Proc Natl Acad Sci U S A* 114, (2017).
9. Moore, L. D., Le, T. & Fan, G. DNA methylation and its basic function. *Neuropsychopharmacology* vol. 38 Preprint at https://doi.org/10.1038/npp.2012.112 (2013).
10. Luo, C., Hajkova, P. & Ecker, J. R. Dynamic DNA methylation: In the right place at the right time. *Science* vol. 361 Preprint at https://doi.org/10.1126/science.aat6806 (2018).
11. Ehrlich, M. *et al.* Amount and distribution of 5-methylcytosine in human DNA from different types of tissues or cells. *Nucleic Acids Res* 10, (1982).
12. Li, E., Beard, C. & Jaenisch, R. Role for DNA methylation in genomic imprinting. *Nature* 366, (1993).
13. Heiss, J. A. *et al.* Battle of epigenetic proportions: comparing Illumina's EPIC methylation microarrays and TruSeq targeted bisulfite sequencing. *Epigenetics* 15, (2020).
14. Sandoval, J. *et al.* Validation of a DNA methylation microarray for 450,000 CpG sites in the human genome. *Epigenetics* 6, (2011).
15. Moran, S., Arribas, C. & Esteller, M. Validation of a DNA methylation microarray for 850,000 CpG sites of the human genome enriched in enhancer sequences. *Epigenomics* 8, (2016).
16. Noguera-Castells, A., García-Prieto, C. A., Álvarez-Errico, D. & Esteller, M. Validation of the new EPIC DNA methylation microarray (900K EPIC v2) for high-throughput profiling of the human DNA methylome. *Epigenetics* 18, (2023).
17. Booth, M. J. *et al.* Quantitative sequencing of 5-methylcytosine and 5-hydroxymethylcytosine at single-base resolution. *Science (1979)* 336, 934–937 (2012).
18. Li, Y. & Tollefsbol, T. O. DNA methylation detection: Bisulfite genomic sequencing analysis. *Methods in Molecular Biology* 791, 11–21 (2011).

19. Grunau, C., Clark, S. J. & Rosenthal, A. Bisulfite genomic sequencing: systematic investigation of critical experimental parameters. *Nucleic Acids Res* 29, (2001).

20. Booth, M. J. *et al.* Oxidative bisulfite sequencing of 5-methylcytosine and 5-hydroxymethylcytosine. *Nat Protoc* 8, (2013).

21. Logsdon, G. A., Vollger, M. R. & Eichler, E. E. Long-read human genome sequencing and its applications. *Nature Reviews Genetics* vol. 21 Preprint at https://doi.org/10.1038/s41576-020-0236-x (2020).

22. Kong, A. *et al.* Detection of sharing by descent, long-range phasing and haplotype imputation. *Nat Genet* 40, (2008).

23. Okano, M., Bell, D. W., Haber, D. A. & Li, E. DNA methyltransferases Dnmt3a and Dnmt3b are essential for de novo methylation and mammalian development. *Cell* 99, (1999).

24. Krebs, A. R., Dessus-Babus, S., Burger, L. & Schübeler, D. High-throughput engineering of a mammalian genome reveals building principles of methylation states at CG rich regions. *Elife* 3, (2014).

25. Yin, Y. *et al.* Impact of cytosine methylation on DNA binding specificities of human transcription factors. *Science (1979)* 356, (2017).

26. Stadler, M. B. *et al.* DNA-binding factors shape the mouse methylome at distal regulatory regions. *Nature* 480, (2011).

27. Stefansson, O. A. *et al.* The correlation between CpG methylation and gene expression is driven by sequence variants. *Nat Genet* 56, 1624–1631 (2024).

28. Surani, M. A. H., Barton, S. C. & Norris, M. L. Development of reconstituted mouse eggs suggests imprinting of the genome during gametogenesis. *Nature* 308, (1984).

29. Li, Y. & Sasaki, H. Genomic imprinting in mammals: Its life cycle, molecular mechanisms and reprogramming. *Cell Research* vol. 21 Preprint at https://doi.org/10.1038/cr.2011.15 (2011).

30. Zink, F. *et al.* Insights into imprinting from parent-of-origin phased methylomes and transcriptomes. *Nat Genet* 50, (2018).

31. Kong, A. *et al.* Parental origin of sequence variants associated with complex diseases. *Nature* 462, (2009).

32. Do, C. *et al.* Mechanisms and Disease Associations of Haplotype-Dependent Allele-Specific DNA Methylation. *Am J Hum Genet* 98, (2016).

33. Monk, D., Mackay, D. J. G., Eggermann, T., Maher, E. R. & Riccio, A. Genomic imprinting disorders: lessons on how genome, epigenome and environment interact. *Nature Reviews Genetics* vol. 20 Preprint at https://doi.org/10.1038/s41576-018-0092-0 (2019).

34. Deamer, D., Akeson, M. & Branton, D. Three decades of nanopore sequencing. *Nature Biotechnology* vol. 34 Preprint at https://doi.org/10.1038/nbt.3423 (2016).

35. Flusberg, B. A. *et al.* Direct detection of DNA methylation during single-molecule, real-time sequencing. *Nat Methods* 7, 461–465 (2010).

36. Frommer, M. *et al.* A genomic sequencing protocol that yields a positive display of 5-methylcytosine residues in individual DNA strands. *Proc Natl Acad Sci U S A* 89, (1992).

37. Du, P. *et al.* Comparison of Beta-value and M-value methods for quantifying methylation levels by microarray analysis. *BMC Bioinformatics* 11, (2010).

38. Lister, R. *et al.* Global epigenomic reconfiguration during mammalian brain development. *Science (1979)* 341, (2013).

39. Skvortsova, K. *et al.* Comprehensive evaluation of genome-wide 5-hydroxymethylcytosine profiling approaches in human DNA. *Epigenetics Chromatin* 10, (2017).

40. Wreczycka, K. *et al.* Strategies for analyzing bisulfite sequencing data. *Journal of Biotechnology* vol. 261 Preprint at https://doi.org/10.1016/j.jbiotec.2017.08.007 (2017).

41. YourGenome/Wellcome Connecting Science. What is Oxford Nanopore Technology (ONT) sequencing? https://www.yourgenome.org/theme/what-is-oxford-nanopore-technology-ont-sequencing/.

42. Simpson, J. T. *et al.* Detecting DNA cytosine methylation using nanopore sequencing. *Nat Methods* 14, 407–410 (2017).

43. Oxford Nanopore Technologies. Dorado. Preprint at https://github.com/nanoporetech/dorado.

44. Hannum, G. *et al.* Genome-wide Methylation Profiles Reveal Quantitative Views of Human Aging Rates. *Mol Cell* 49, (2013).

45. Horvath, S. DNA methylation age of human tissues and cell types. *Genome Biol* 14, R115 (2013).

46. Levine, M. E. *et al.* An epigenetic biomarker of aging for lifespan and healthspan. *Aging* 10, 573–591 (2018).

47. Lu, A. T. *et al.* DNA methylation GrimAge strongly predicts lifespan and healthspan. *Aging* 11, (2019).

48. Bell, C. G. *et al.* DNA methylation aging clocks: Challenges and recommendations. *Genome Biology* vol. 20 Preprint at https://doi.org/10.1186/s13059-019-1824-y (2019).

Appendix I: Paper I – Supplementary materials

Supplementary material

Additional file 1: Supplementary Material

This document includes:

1. **Supplementary Notes**
2. **Description of the contents in Data S1, S2, S3, S4, S5 and S6**
3. **References cited in Supplementary Materials**

1. Supplementary Notes

1.1 Effect of different pipeline versions on the methylation calling

All samples were analyzed with two version of our pipelines, referred to as v3 and v4. The main difference is the version of guppy, the basecalling algorithm, and flowcell chemistry, resulting in lower error rate for v4 (Fig. S1) (1). We show that the errror rate improves for each version of guppy (Fig. S2A). The sequencing coverage also increases which lowers the error rate (Fig. S2B) but is not affected by mean N50.

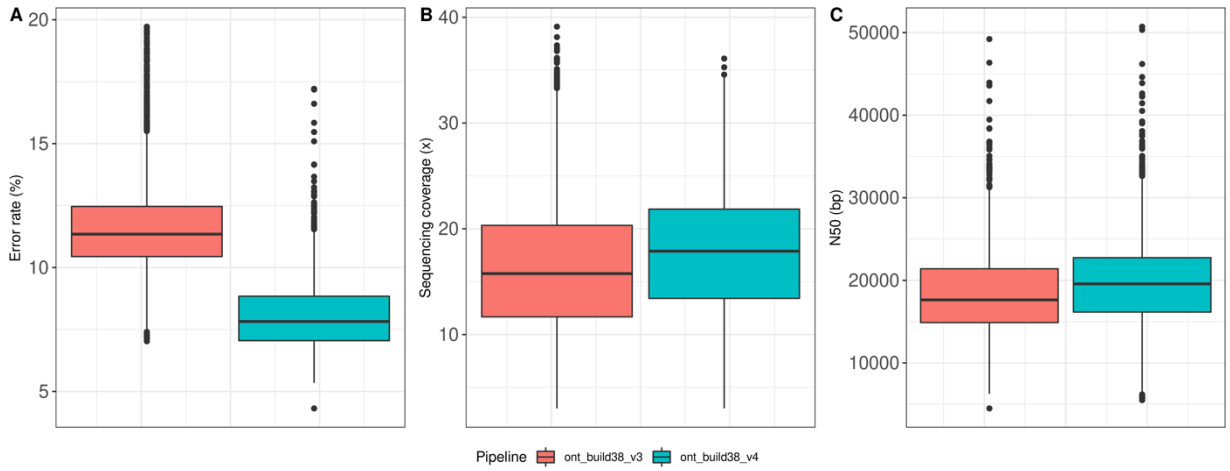


Fig. S1 Statistics for different versions of the pipeline, computed for 50 high coverage samples. **A** Box plot showing the computed error rate for each version of the pipeline. **B** Box plot showing the sequencing coverage for each version of the pipeline. **C** Box plot showing the computed N50 for each version of the pipeline. The centre line (solid black) shown in each box represent the median; the box limits represent upper and lower quartile, whiskers represent 1.5x interquartile range.

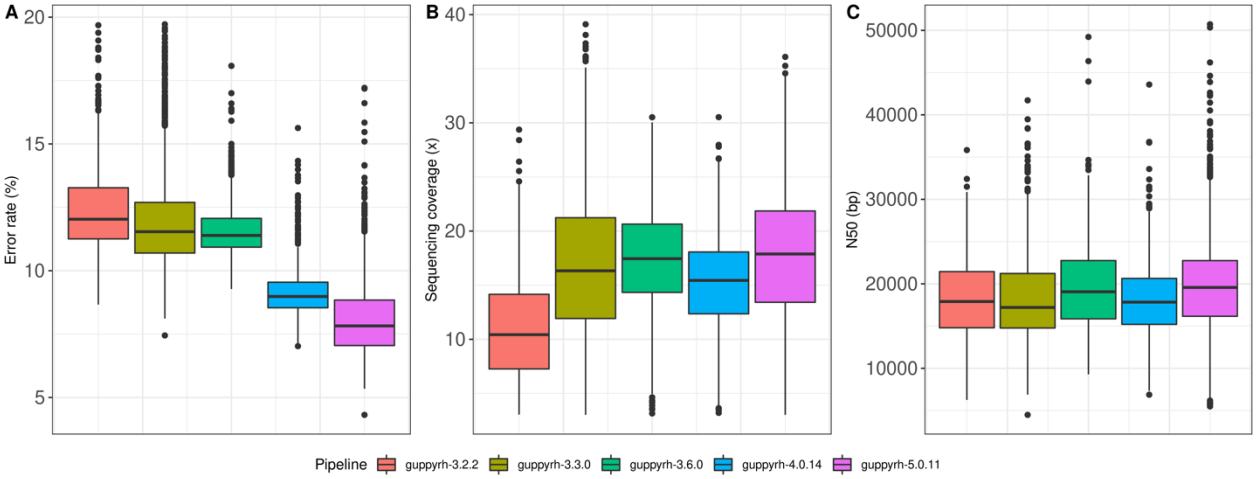


Fig. S2 Statistics for different versions of Guppy, computed for 50 high coverage samples. **A** Box plot showing the computed error rate for each version of guppy. **B** Box plot showing the sequencing coverage for each version of guppy. **C** Box plot showing the computed N50 for each version of guppy. The centre line (solid black) shown in each box represent the median; the box limits represent upper and lower quartile, whiskers represent 1.5x interquartile range.

1.2 Nanopore data is more consistent in unmethylated and methylated CpG-units

CpGs were categorized based on 5-mCpG rates in oxBS, where sequencing coverage in oxBS is high (>25x). We calculated the mean methylation levels within correctly classified CpGs conditioned on the methylation levels measured on oxBS, separately for ONT and oxBS. We further calculated the mean absalute difference and standard deviation between the two measurements. Results are shown in Table S2.

1.3 Nanopolish methylation prediction quality depends on the CpG-unit sequence context

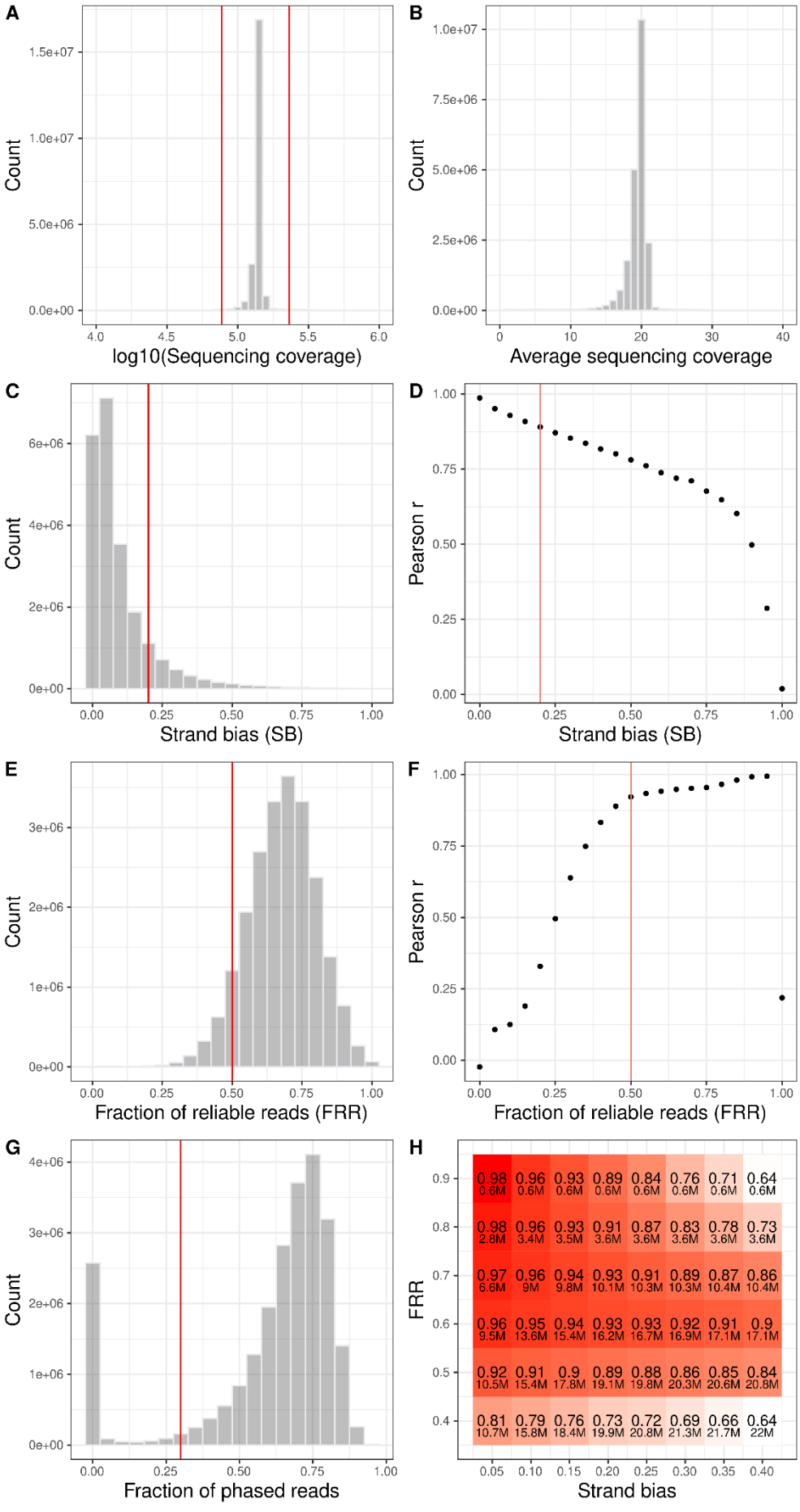

A
Count
log10(Sequencing coverage)
B
Count
Average sequencing coverage
C
Count
Strand bias (SB)
D
Pearson r
Strand bias (SB)
E
Count
Fraction of reliable reads (FRR)
F
Pearson r
Fraction of reliable reads (FRR)
G
Count
Fraction of phased reads
H
FRR
Strand bias
0.98 0.6M 0.96 0.6M 0.93 0.6M 0.89 0.6M 0.84 0.6M 0.76 0.6M 0.71 0.6M 0.64 0.6M
0.98 2.8M 0.96 3.4M 0.93 3.5M 0.91 3.6M 0.87 3.6M 0.83 3.6M 0.78 3.6M 0.73 3.6M
0.97 6.6M 0.96 9M 0.94 9.8M 0.93 10.1M 0.91 10.3M 0.89 10.3M 0.87 10.4M 0.86 10.4M
0.96 9.5M 0.95 13.6M 0.94 15.4M 0.93 16.2M 0.93 16.7M 0.92 16.9M 0.91 17.1M 0.9 17.1M
0.92 10.5M 0.91 15.4M 0.9 17.8M 0.89 19.1M 0.88 19.8M 0.86 20.3M 0.85 20.6M 0.84 20.8M
0.81 10.7M 0.79 15.8M 0.76 18.4M 0.73 19.9M 0.72 20.8M 0.69 21.3M 0.66 21.7M 0.64 22M

Fig. S3 Selection of high-quality CpGs. **A** Histogram of the sequencing coverage distribution of the population, plotted on log10-scale. For better demonstration of the majority of the data we only show values between 4-6 on x-axis. **B** Histogram of the average coverage distribution (coverage divided by number of samples), shown for the range of 10-40x. **C** Histogram of the strand bias in the data. **D** Scatter plot of the correlation coefficient for each cut-off value for strand bias. **E** Histogram of the fraction of reliable reads. **F** Scatter plot of the correlation coefficient for each cut-off value for fraction of reliable reads. **G** Histogram of the fraction of phased reads. Red lines represent the selected cut-off values. **H** Heatmap of the correlation coefficient and number of hq-CpG based on each cutoff.

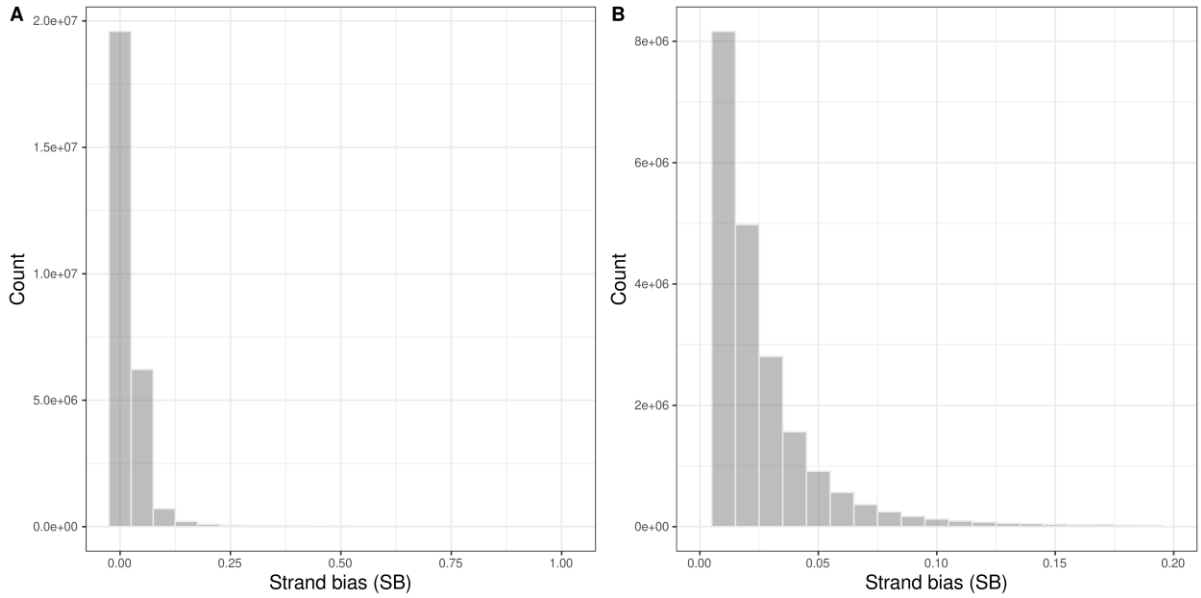


Fig. S4 Strand bias in oxBS data. **A** Histogram for strand bias in oxBS data. **B** Same plot, shown only for strand bias, x-axis, between 0-0.2. The mean strand bias in oxBS data was 0.025.

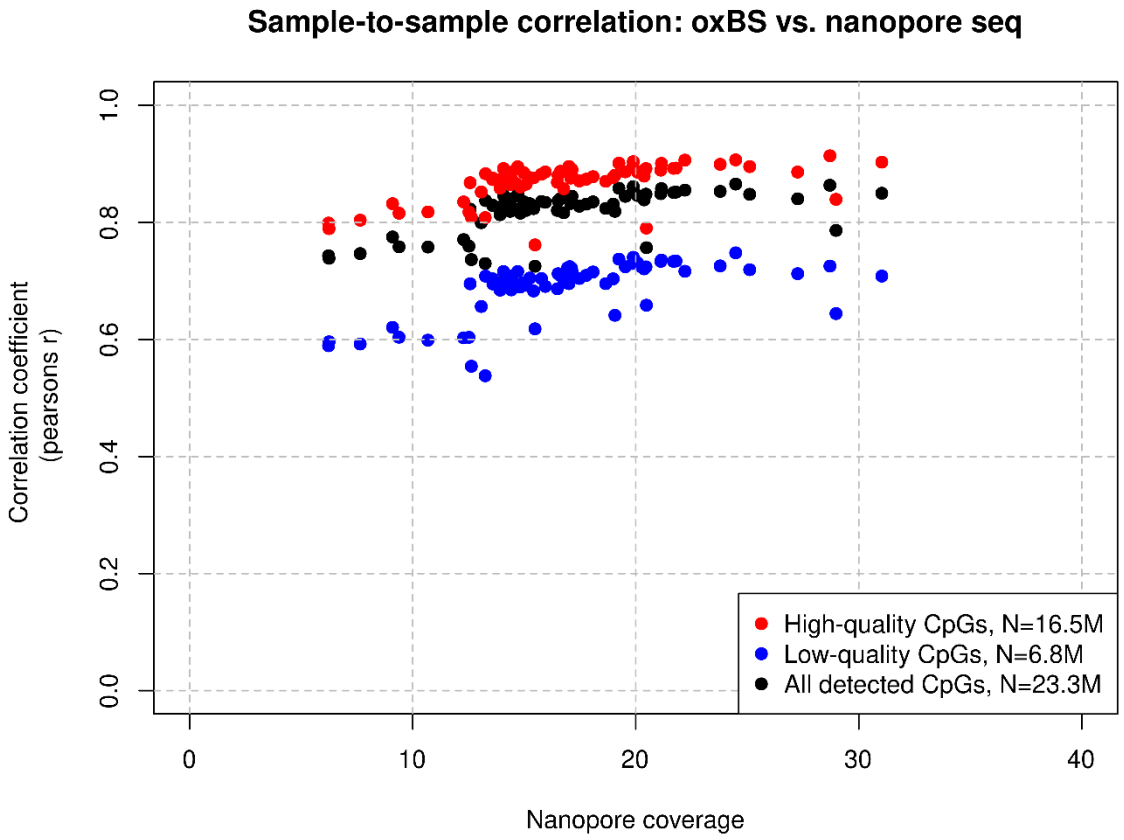


Fig. S5 Sample-to-sample correlation for samples sequenced using oxBS and nanopore sequencing 132 samples isolated from the same whole blood were analyzed using both oxidative bisulfite sequencing and whole genome nanopre sequencing. Shown are correlation coefficients, y-axis, for each sample-to-sample comparison; i.e. 5-mCpG rates measured by nanopore sequencing compared to 5-mCpG rates measured by oxBS in the same individual are plotted on the y-axis. The average coverage of the nanopore sequenced DNA (upper panel) and oxBS sequenced DNA (lower panel) are plotted on the x-axis. Correlation coefficients were calculated on the basis of all CpGs detected (blue colour) in sequenced reads from both sequencing methods, but also based only on CpGs that meet our quality critera for percent error and strand bias (black) and those that do not meet this criteria (red). The 16.5M CpGs categorized as "high quality" (black) consistently show higher sample-to-sample correlation coefficients, but the correlation is clearly affected by average coverage; i.e. nanopore sequenced samples with average coverage < 10 show lower correlation coefficients across all categories of CpGs.

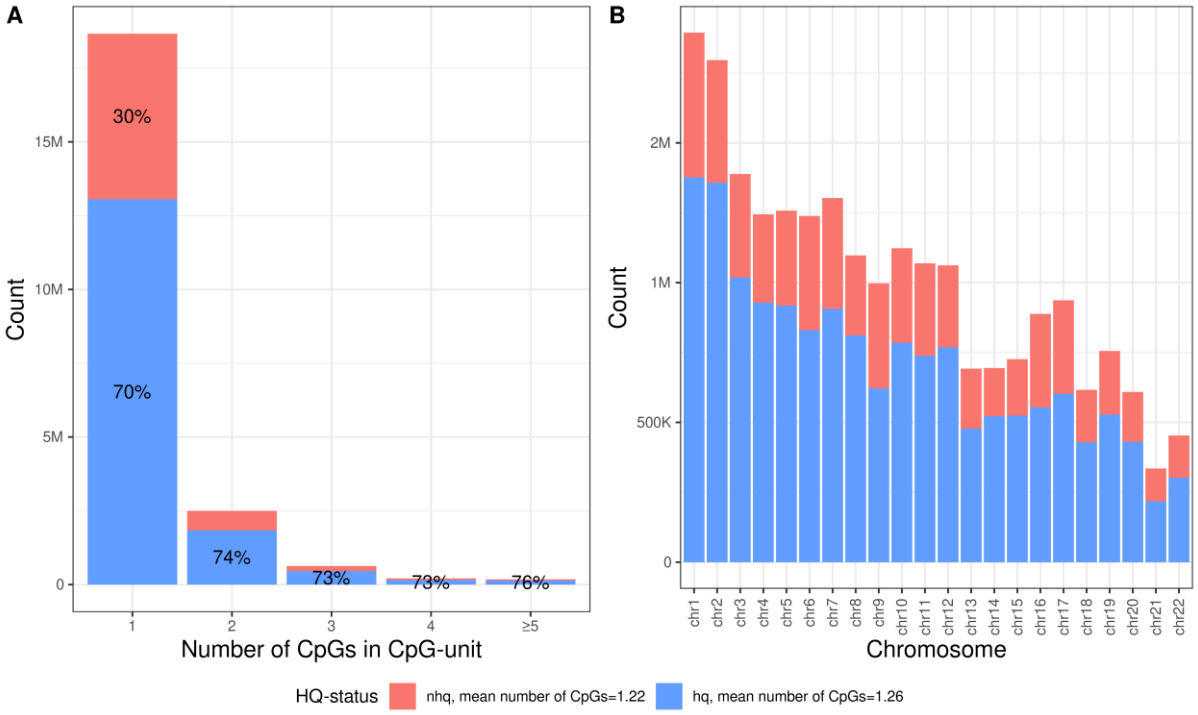


Fig. S6 Distribution of high-quality CpG units**. A** Majority of the CpG units have less than 2 CpGs, for the full set and the hq-CpG units. The average number of CpGs within unit was 1.26 for hq-CpGs and 1.22 for other CpGs. **B** The distribution of hq-CpGs per chromosome is consistent with total number of CpGs per chromosome.

1.4 Guppy outperforms Nanopolish per CpG site in comparison to oxBS

We selected 42 samples, nanopore sequenced on 55 recent flowcells (average coverage 16.6x, average error rate 7.9%), for which we also have oxBS data. We performed additional methylation detection using Guppy. To evaluate the per-site correlation of methylation levels between the nanopore data and oxBS data, we select CpGs with high coverage in both datasets, defined as having on average more than 10 reads covering the CpG per flowcell for oxBS data and as having on average more than 10 reads covering the CpG, exceeding the cutoff threshold per flowcell for nanopore data. Using these high coverage CpGs, we create 10 benchmarking datasets with 0-10%, 10-20%, …, and 90-100% methylation levels in oxBS. We then sampled one CpG per 100kb in autosomes from each group, ending with 254,566 CpGs to compare (not all subsets per bin contained a high coverage CpG).

We examined the APC coefficient between the methylation levels predicted by the two tools and oxBS and found that Guppy has a higher correlation with oxBS than Nanopolish and has lower RMSE values (Fig. S7A). Nanopolish and Guppy agree with the expected methylation level per site for all groups and especially for intermethylated CpGs. Nanopolish is in slightly better agreement with oxBS

for low methylated regions compared to Guppy (0-20%, m_{Guppy}=0.180, $m_{Nanopolish}$=0.151 and m_{oxBS}=0.0905).

We then calculated the proportion of sites with concordant prediction within each window. We define concordant prediction as predictions where the methylation levels fall into the expected prediction window based on oxBS data. Nanopolish has more sites correctly predicted within expected window for unmethylated (0-10%, 10-20% and 20-30%) and highly methylated CpGs (90-100%) while Guppy has higher proportion correctly predicted in low- and intermethylated CpGs (30-90%) (Fig. S8B). Consistently, Guppy has higher number of CpGs within expected range for intermethylated CpGs (40-90%) while Nanopolish has higher number for unmethylated (0-20%) and highly methylated CpGs (90-100%). The performance is similar for low methylated CpGs (20-40%) (Fig. S7B).

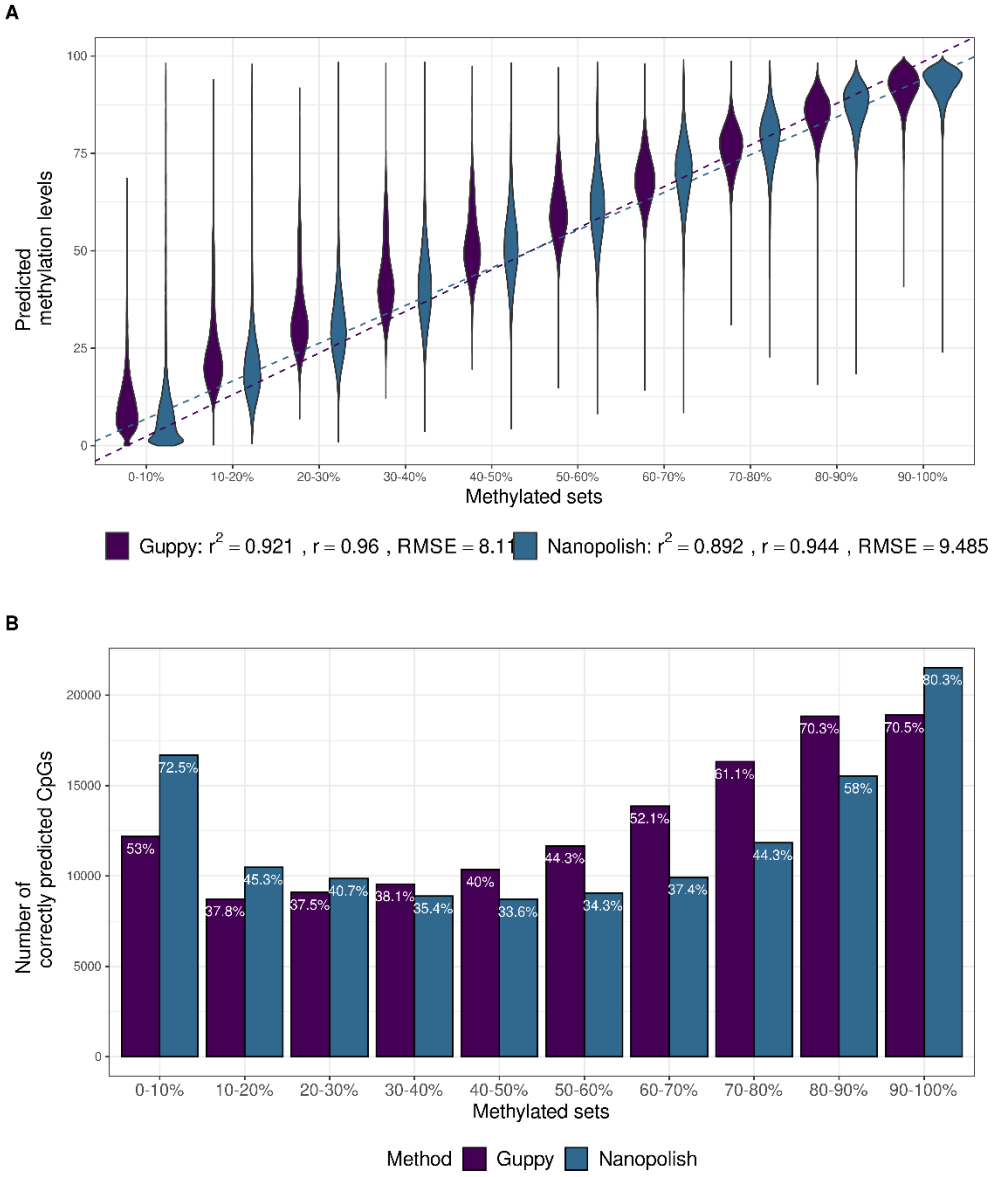
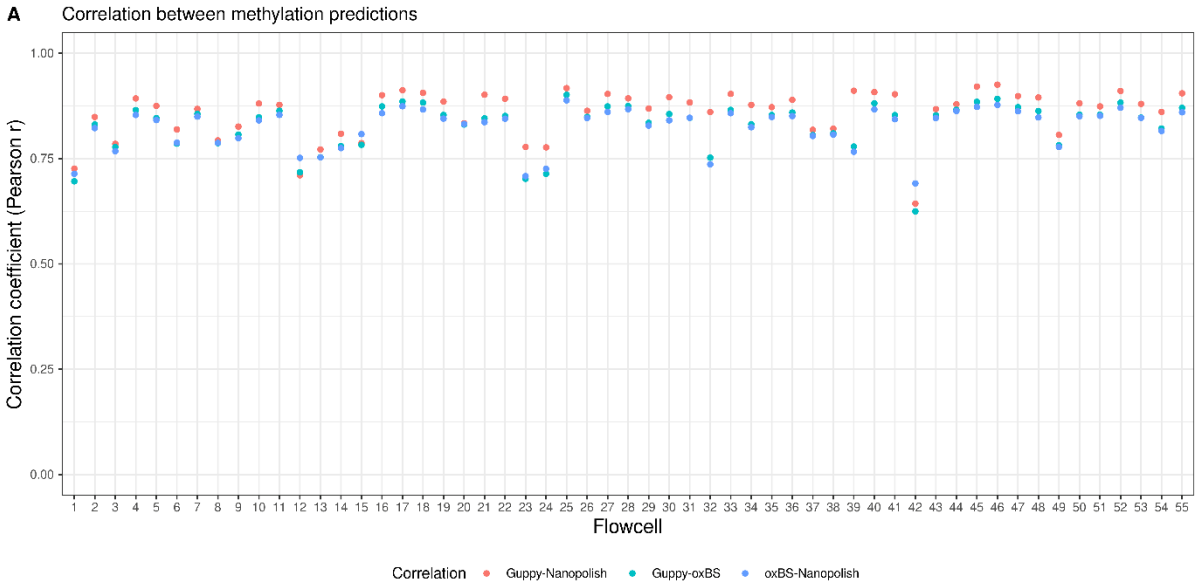
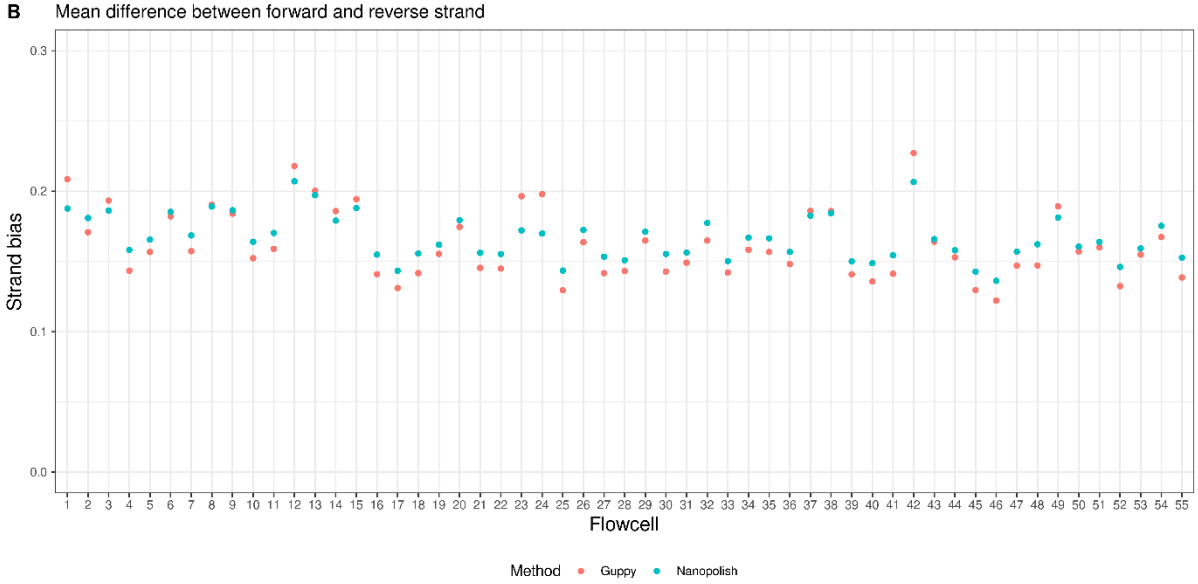


Fig. S7 Comparison of CpG methylation predictions from Guppy and Nanopolish with oxBS within each methylation ratio group. **A** Violin plots showing the predicted methylation levels (y-axis) for each control set with a a given proportion of methylated reads (x-axis). The Pearson correlation coefficient, coefficient of determination and RMSE is given for both methods. **B** The number of congrugent methylation predictions (y-axis) and the proportion, shown for each methylation window (x-axis).

For each sample we calculated the Pearson correlation coefficeint over all CpGs between the predicted levels by Guppy or Nanopolish and the corresponding oxBS data. The per sample correlation ranged from 0.64 to 0.93, with average 0.85 (Fig. S8A). There was a similar average strand bias per sample in the two methods, with 0.162 on average per sample for Guppy and 0.167 for Nanopolish (Fig. S8B) and similar mean absolute difference between the two methods and oxBS.


A
Correlation between methylation predictions
Correlation coefficient (Pearson r)
1.00
0.75
0.50
0.25
0.00
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47 48 49 50 51 52 53 54 55
Flowcell
Correlation
Guppy-Nanopolish
Guppy-oxBS
oxBS-Nanopolish


B
Mean difference between forward and reverse strand
Strand bias
0.3
0.2
0.1
0.0
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47 48 49 50 51 52 53 54 55
Flowcell
Method
Guppy
Nanopolish

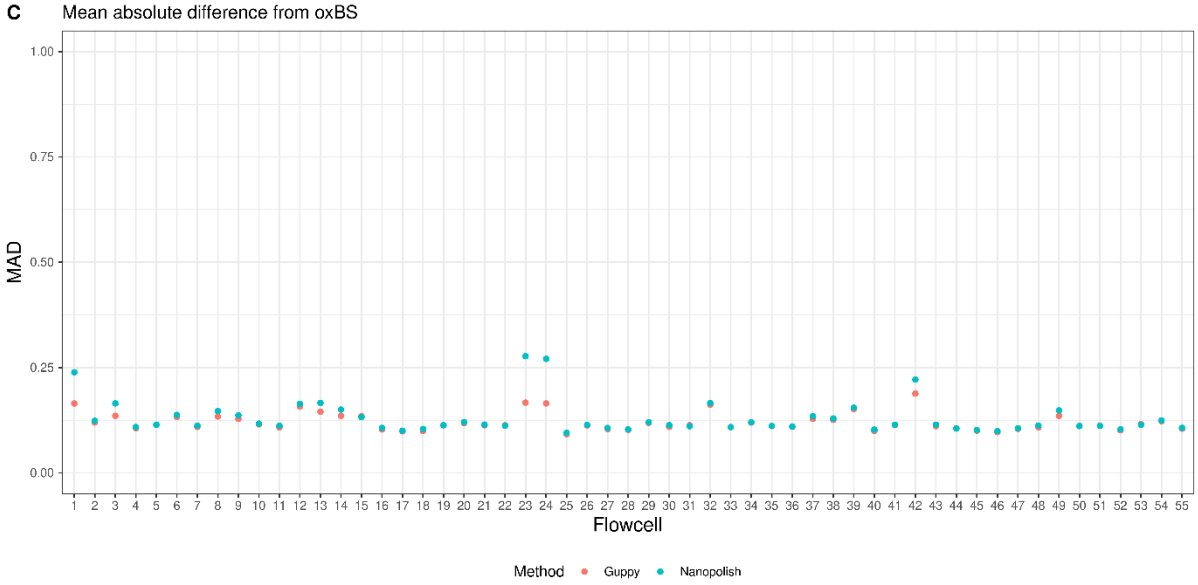

C
Mean absolute difference from oxBS
MAD
1.00
0.75
0.50
0.25
0.00
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47 48 49 50 51 52 53 54 55
Flowcell
Method
Guppy
Nanopolish

Fig. S8 Comparison of CpG methylation predictions from Guppy and Nanopolish with oxBS. **A** Correlation coefficient (y-axis) per flowcell (x-axis). **B** Mean strand bias (y-axis) per flowcell (x-axis). **C** Mean absalute difference in 5-mCpG between Guppy and oxBS and Nanopolish and oxBS (y-axis) per sample (x-axis).

1.5 Comparison of CpG methylation predictions from nanopore sequencing and SMRT sequencing

Sequencing statistics for nanaopore and SMRT sequencing

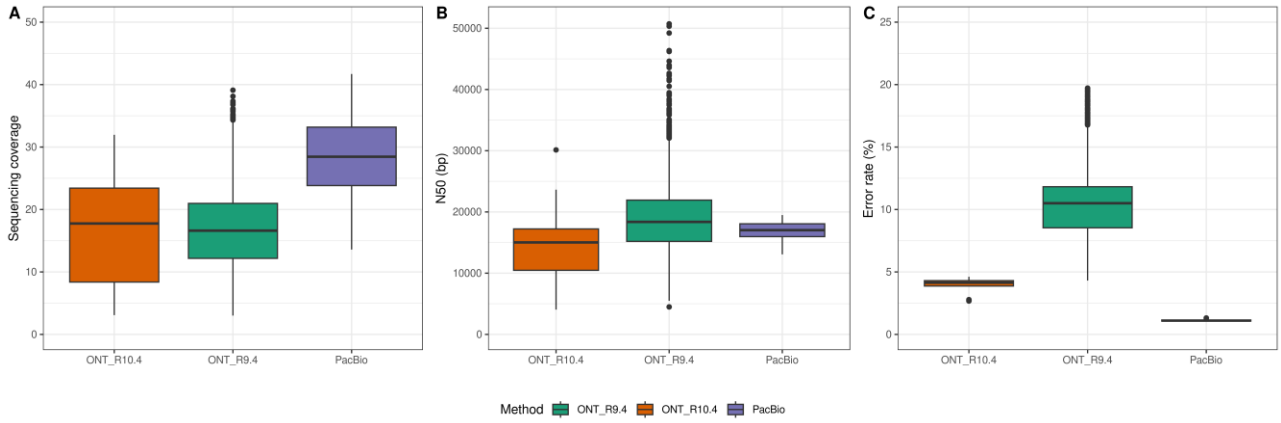


Figure S9 Comparison of sequencing statistics for long-read sequencing techniques. A Boxplot showing the sequencing coverage (y-axis) for each method (x-axis). **B** Box plot showing the N50 (y-axis) distribution for each method (x-axis). **C** Box plot showing the percentage error rate (y-axis) for each method (x-axis). The centre line (solid black) shown in each box represent the median; the box limits represent upper and lower quartile, whiskers represent 1.5x interquartile range.

Subset description

Description of the random subset created is shown in Table S3. We note, that some of the differences in APC and MAD observed between methods may be due to differences in age, gender or smoking status of the samples. The effect of these attributes involve relatively few and specific CpGs, considering the >15 million measured CpGs, and the effects tend to be subtle. For this reason, we assume that these attributes will not have a large effect when comparing all CpGs based on the APC parameter.

Table S3 Description of the random subsets, showing gender and year of birth distribution. Smoking status was derived from ever smokers versus never smokers phenotype created inhouse.

Method	Gender	Number of samples	Earliest YOB	Latest YOB	Median YOB	Smokers

oxBS	Females	26	1923	1989	1952	16
oxBS	Males	24	1931	1986	1951	14
Nanopolish	Females	28	2005	1914	1955	7
Nanopolish	Males	22	2001	1902	1972	6
Guppy_R9.4	Females	25	1991	1895	1952	12
Guppy_R9.4	Males	25	1998	1921	1955	6
Guppy_R10.4	Females	9	1992	1948	1979	NA*
Guppy_R10.4	Males	11	2010	1975	1987	NA*
PacBio	Females	29	1998	1941	1950	9
PacBio	Males	21	1998	1946	1949	9

*None of the participants had smoking status available.

Set of hq-CpGs

First, we assess if restricting to the set of hq-CpGs selected using Nanopolish methylation detection would benefit the Guppy data. Guppy benefit from restricting to the set of high-quality CpGs and we end up with similar correlation between Nanopolish and oxBS and Guppy and oxBS, however the Nanopolish data gains the most from the filters.

Next we applied the same filters for all methods. Not all filters are applicable for all methods, fraction of reliable reads is only filtered for Nanopolish and strand bias is not measured in PacBio as the reading strand information is not known. Fraction of reliable reads filters could be introduced for PacBio and Guppy, by slightly altering the summary scripts, but preliminary result suggest the benefit is not as great as for Nanopolish data. PacBio had higher APC in high and low coverage CpGs and more CpGs filtered out, indicating that this filter may need to be adjusted for PacBio data. Furthermore, PacBio showed the lowest improvement in APC of all methods, suggesting that other filters may be more beneficial for improving PacBio performance. In contrast, Nanopolish data required most filters, which explains why the number of hq-CpGs is lower in this dataset.

We found that applying quality filters improved the APC coefficient between long-read sequencing data and oxBS data. Nanopore data retained more CpGs in datasets sequenced using R10.4 flowcells and with more recent version of Guppy. Guppy applied to R10.4 flowcell and PacBio show higher fraction of high quality in all groups than Guppy applied to R9.4 (Fig. S10). After filtering, all datasets showed higher APC coefficient with oxBS data, lower MAD and higher methylation levels, indicating that more low- or unmethylated CpGs were removed. Guppy called on R10.4 flowcells had the highest number of hq-CpGs followed by PacBio and also highest APC with oxBS data and lowest MAD. R9.4 flowcells methylation called with Guppy and PacBio performed similarly (Additional File 2: Table S4, S5).

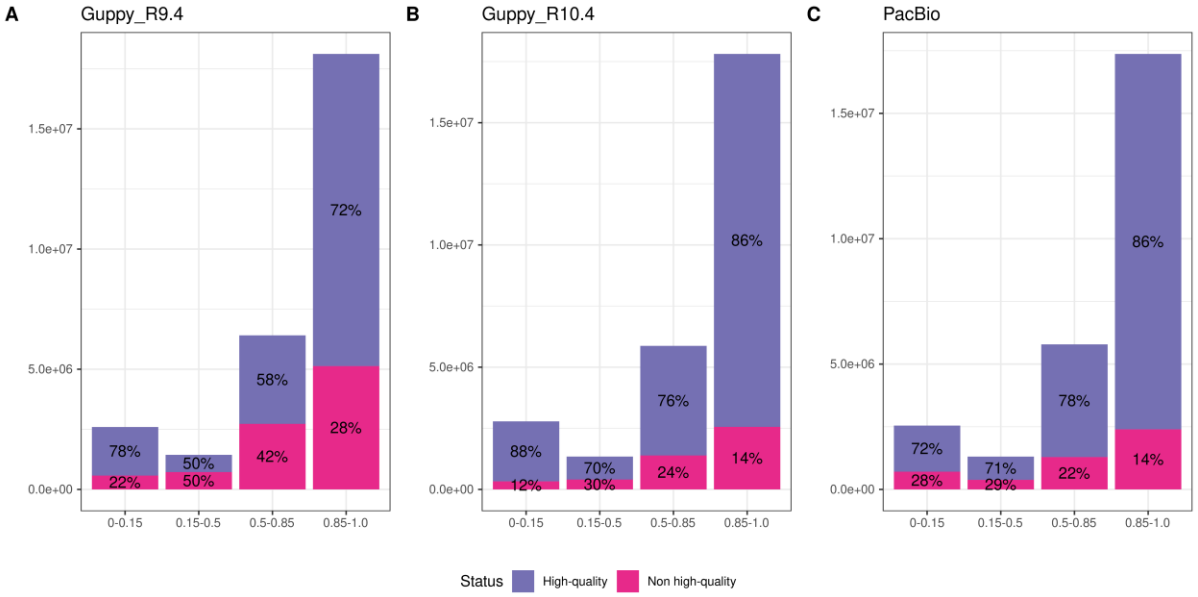


Fig. S10 Fraction of hq-CpGs within each range in 5-mCpG rates, shown for **A** Guppy applied to R9.4 flowcells, **B** Guppy applied to R10.4 flowcells and **C** PacBio.

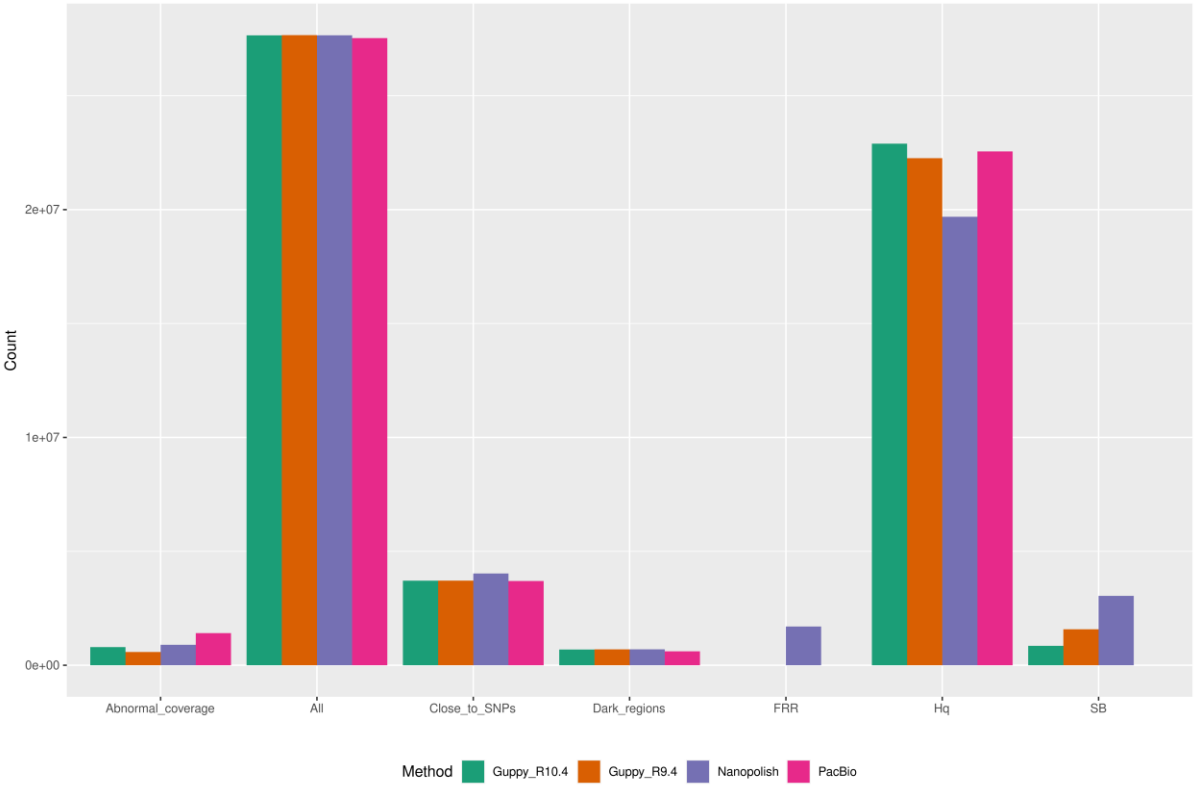


Fig. S11 Number of CpGs (y-axis) removed from each group (x-axis) shown for Guppy R9.4 in orange, Guppy R10.4 in green, Nanopolish in purple and PacBio in pink. Strand bias is not measured in PacBio and FRR is only measured for Nanopolish.

Per-site accuracy of 5-mCpG predictions in different genomic and methylation context

To investigate the variation in methylation predictions among the different methods, we categorized the oxBS data into four bins based on the methylation rates and computed the MAD in 5m-CpG rates between all methods and oxBS. We further looked into the MAD for CpG islands, shelves and shores, separately.

In terms of the full set of CpGs and methylated CpGs all methods demonstrated similar performance with regards to MAD. All methods display a higher MAD in low- and intermethylated CpGs and all methods tend to overestimate the 5-mCpG rates compared to oxBS data, Guppy and PacBio more than Nanopolish, which supports the shift in bimodal distribution (Fig. 3., Additional File 1: Fig. S12). Nanopolish exhibits the lowest MAD in 5-mCpGs when compared to oxBS in unmethylated CpGs and CpG islands. On the other hand, PacBio has the lowest MAD in intermethylated CpGs. Adding more training data from inter- and lowmethylated CpGs might benefit the methylation detection of all tools.

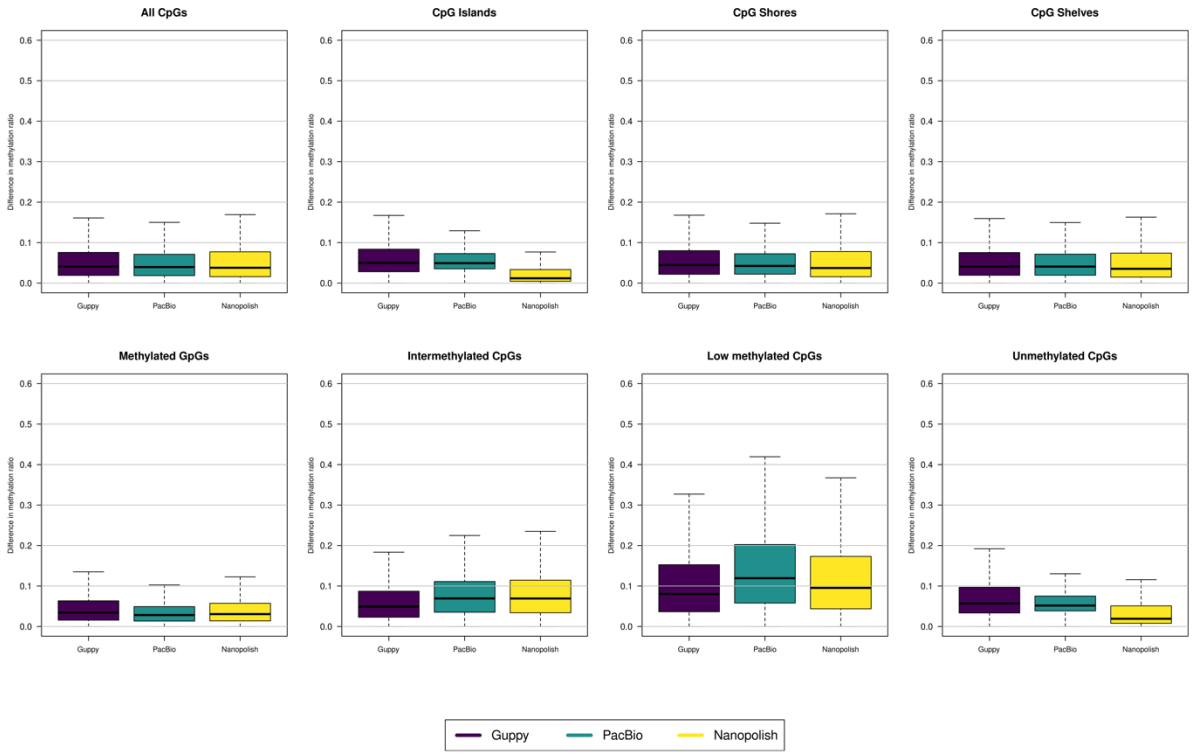


Fig. S12 Mean absolute difference in 5-mCpG rates between the long-read datasets and oxBS data**.** Box plot showing MAD in 5-mCpG rates between all datasets shown separately for different genomic contexts: CpG islands, CpG shores and CpG shelves and for different methylation rates: methylated CpGs, intermethylated CpGs, low methylated CpGs and unmethylated CpGs. The centre line (solid black) shown in each box represent the median; the box limits represent upper and lower quartile, whiskers represent 1.5x interquartile range.

2. Supplementary data description

Data S1: Summary statistics for CpG units averaged over 7,179 samples sequenced using R9.4 flowcell from ONT and methylation called using Nanopolish.

Chrom: *The chromosome of the CpG unit*

start: *The nucleotide starting position of the CpG unit, 0 based*

end: *The nucleotide ending position of the CpG unit*

n_reliable: *The number of realiable reads behind the CpG unit. Reliable is defined as having the absalute log likelihood ratio larger than 1.921*

n_total: *Total number of reads behind the CpG unit*

ratio: *The methylation ratio, calculated as number of methylated reliable reads out of all reliable reads.*

strand_bias: *Difference between the methylation ratio of forward and reverse strand*

FRR: *Fraction of reliable reads, calculated as n_reliable/n_total*

dark_region: *Boolean column indication of whether the CpG unit is within dark region. 1 indicates the CpG unit is within a dark region, 0 indicates it is not*

SNP: *Boolean column indicating whether the CpG unit is within 5bp of SNP*

phased_frac: *Fraction of phased reads*

hq: *Boolean column Indicating whether the CpG unit is high-quality*

hq_phased: *Boolean column indicating whether the CpG unit is high quality with phasing status included*

highSB: *Boolean column indicating whether the CpG unit failes on strand bias test and has strand bias greater than 0.2*

lowFRR: *Boolean column indicating if the CpG unit fails on FRR test and has FRR less than 0.5.*

highCov: *Boolean column indicating if the CpG unit fails on coverage test and has coverage greater than 1.5 times the average coverage*

lowCov: *Boolean column indicating if the CpG unit fails on coverage test and has coverage less than 0.5 times the average coverage*

lowPF: *Boolean column indicating if the CpG unit fails on phasing fraction and has PF less than 0.3*

Data S2: Summary statistics for CpGs averaged over 132 samples sequenced using oxBS.

Chrom: *The chromosome of the CpG*

Pos: *The nucleotide starting position of the CpG, 0 based*

N_total: *Total number of reads behind the CpG*

ratio: *The methylation ratio, calculated as number of methylated reads out of all reads*

Strand_bias: *Difference between the methylation ratio of forward and reverse strand*

Data S3: Summary statistics for CpGs averaged over 304 samples sequenced using R9.4 flowcell from ONT and methylation called using Guppy.

Chrom: *The chromosome of CpG*

startx: *The nucleotide starting position of the CpG, 0 based*

end: *The nucleotide ending position of the CpG*

n_total: *Total number of reads behind the CpG*

ratio: *The methylation ratio, calculated as number of methylated reads out of all reads.*

strand_bias: *Difference between the methylation ratio of forward and reverse strand*

dark_region: *Boolean column indicating if the CpG is within dark region. 1 indicates the CpG unit is within a dark region, 0 indicates it is not*

SNP: *Boolean column indicating if the CpG is within 5bp of SNP*

highSB: *Boolean column indicating if the CpG failes on strand bias test and has strand bias greater than 0.2*

highCov: *Boolean column indicating if the CpG fails on coverage test and has coverage greater than 1.5 times the average coverage*

lowCov: *Boolean column indicating if the CpG fails on coverage test and has coverage less than 0.5 times the average coverage*

hq: *Boolean column indicating if the CpG is high-quality*

Data S4: Summary statistics for CpGs, averaged over 22 samples sequenced using R10.4 flowcell from ONT and methylation called using Guppy.

Chrom: *The chromosome of CpG*

startx: *The nucleotide starting position of the CpG, 0 based*

end: *The nucleotide ending position of the CpG*

n_total: *Total number of reads behind the CpG*

ratio: *The methylation ratio, calculated as number of methylated reads out of all reads.*

strand_bias: *Difference between the methylation ratio of forward and reverse strand*

dark_region: *Boolean column indicating if the CpG is within dark region. 1 indicates the CpG unit is within a dark region, 0 indicates it is not*

SNP: *Boolean column indicating if the CpG is within 5bp of SNP*

highSB: *Boolean column indicating if the CpG failes on strand bias test and has strand bias greater than 0.2*

highCov: *Boolean column indicating if the CpG fails on coverage test and has coverage greater than 1.5 times the average coverage*

lowCov: *Boolean column indicating if the CpG fails on coverage test and has coverage less than 0.5 times the average coverage*

hq: *Boolean column indicating if the CpG is high-quality*

Data S5: Summary statistics for CpGs, averaged over 50 samples sequenced using SMRT-sequencing and methylation called using primrose.

Chrom: *The chromosome of CpG*

Pos: *The nucleotide position of the CpG, 0 based*

n_total: *Total number of reads behind the CpG*

ratio: *The methylation ratio, calculated as number of methylated reads out of all reads.*

dark_region: *Boolean column indicating if the CpG is within dark region. 1 indicates the CpG unit is within a dark region, 0 indicates it is not*

SNP: *Boolean column indicating if the CpG is within 5bp of SNP*

highCov: *Boolean column indicating if the CpG fails on coverage test and has coverage greater than 1.5 times the average coverage*

lowCov: *Boolean column indicating if the CpG fails on coverage test and has coverage less than 0.5 times the average coverage*

hq: *Boolean column indicating if the CpG is high-quality*

Data S6: Summary statistics for CpGs, 132 samples sequenced using both oxBS and ONT, filtered on 25x per CpG in oxBS or greater.

Chrom: *The chromosome of CpG*

pos_ont: *The nucleotide start position of the CpG in ONT data, 0 based*

end_ont: *The nucleotide end position of the CpG in ONT data, 0 based*

ratio_ont: *The methylation ratio, calculated as number of methylated reads out of all reads in ONT.*

pos_oxBS: *The nucleotide start position of the CpG in oxBS data, 0 based*

ratio_oxBS: *The methylation ratio, calculated as number of methylated reads out of all reads in oxBS.* 3. References

1. Beyter D, Ingimundardottir H, Oddsson A, Eggertsson HP, Bjornsson E, Jonsson H, et al. Long-read sequencing of 3,622 Icelanders provides insight into the role of structural variants in human diseases and other traits. Nat Genet. 2021;53(6).

Additional file 2: Supplementary Tables

Table S1 The mean absolute difference between Nanopolish methylation predictions and the corresponding oxBS methylation levels.

	Mean absolute difference	**Standard deviation**	**Pearson r**
Pipeline v3 (AER=12.3%)	0.0494	0.690	0.955
Pipeline v4 (AER=7.94%)	0.0486	0.0694	0.954

Table S2 Mean methylation levels for oxBS and nanopore within correctly classified CpGs, mean absalute difference and standard deviation between nanopore methylation detection and oxBS, conditioned on methylation levels measured in oxBS.

Average methylation in oxBS	**Mean methylation levels (oxBS)**	**Mean methylation levels (ONT)**	**Mean absolute difference**	**Standard deviation**
Unmethylated (0-0.15)	0.0362	0.019	0.0226	0.0247
Low methylated (0.15-0.5)	0.336	0.307	0.0673	0.0560
Intermethylated (0.5-0.85)	0.732	0.731	0.0588	0.0526
Methylated (0.85-1)	0.927	0.936	0.0266	0.0228

Table S3 Description of the random subsets, showing gender and year of birth distribution. Smoking status was derived from ever smokers versus never smokers phenotype created inhouse.

Method	**Gender**	**Number of samples**	**Earliest YOB**	**Latest YOB**	**Median YOB**	**Smokers**
oxBS	Females	26	1923	1989	1952	16
oxBS	Males	24	1931	1986	1951	14
Nanopolish	Females	28	2005	1914	1955	7
Nanopolish	Males	22	2001	1902	1972	6
Guppy_R9.4	Females	25	1991	1895	1952	12
Guppy_R9.4	Males	25	1998	1921	1955	6
Guppy_R10.4	Females	9	1992	1948	1979	NA*
Guppy_R10.4	Males	11	2010	1975	1987	NA*
PacBio	Females	29	1998	1941	1950	9
PacBio	Males	21	1998	1946	1949	9

*None of the participants had smoking status available.

Table S4 Statistics for the full set of CpGs compared to hq and not hq-CpGs. Mean methylation, APC coefficient, MAD, strand bias and number of CpGs shown for each method. Strand bias is not shown for PacBio as both strands are read simultaneously.

		Mean methyl levels	APC	Mean absolute difference	Strand bias	nCpGs
Nanopolish	**All CpGs**	0.793	0.959	0.0471	0.0946	27,651,488
	hq-CpGs	0.817	0.986	0.0314	0.0561	19,685,181
	not hq-CpGs	0.735	0.882	0.0866	0.187	7,966,307
Guppy R9.4	**All CpGS**	0.763	0.973	0.0465	0.0644	27,659,182
	Hq-CpGs	0.769	0.987	0.0383	0.0468	22,256,402
	Not hq-CpGs	0.738	0.907	0.0814	0.138	5,402,780
Guppy R10.4	**All CpGs**	0.790	0.978	0.0339	0.0467	27,648,754
	Hq-CpGs	0.792	0.991	0.0291	0.0339	22,893,522
	Not hq-CpGs	0.781	0.901	0.0689	0.109	4,755,232
PacBio	**All CpGs**	0.782	0.970	0.0437	NA*	27,527,663
	Hq-CpGs	0.797	0.980	0.0380	NA*	22,554,423
	Not hq-CpGs	0.710	0.938	0.0705	NA*	4,973,240

*Not reported for PacBio

Table S5 Correlation analysis for CpGs excluded from the high-quality set of CpGs between oxBS methylation detection and Nanopolish, Guppy and PacBio. For each group of CpGs we provide the number of CpG sites and pearson correlation coefficient, and additionally number of CpG-units for Nanopolish.

	ONT Nanopolish			**ONT Guppy R9.4**		**ONT Guppy R10.4**		**PacBio**	
	Pearson r	Number of CpGs-units	Number of CpGs	Pearson r	Number of CpGs	Pearson r	Number of CpGs	Pearson r	Number of CpGs
All CpGs	0.959	22,178,458	27,651,488	0.972	27,659,182	0.978	27,648,754	0.970	27,527,663
CpGs near sequence variants	0.922	2,949,801	4,026,250	0.940	3,713,303	0.934	3,711,812	0.929	3,696,786
CpGs within dark regions	0.698	522,262	698,627	0.716	695,535	0.736	688,246	0.691	611,068
CpGs with high/low coverage	0.722	673,474	896,984	0.702	581,583	0.752	796,000	0.918	1,409,536
CpGs with SB>0.2	0.828	2,800,184	3,044,175	0.854	1,577,810	0.626	850,227	NA	NA
CpGs with FRR<0.5	0.819	1,688,756	1,698,511	NA	NA	NA	NA	NA	NA
Hq CpGs	0.986	15,644,462	19,685,181	0.987	22,256,402	0.991	22,893,522	0.980	22,554,423

Appendix II: Paper II - Supplementary materials

Table of contents:

1. Supplementary Notes

2. Legends for Extended Data Figures

3. Legends for Data files

4. Supplementary Figures

5. Supplementary Tables

6. List of Source Data

Other supplementary materials for this manuscript include the following summary data which can be downloaded from our website (www.decode.com/summarydata/**):**

Data-S1.tab†: Summary of sequence variants associated with 5-mCpG rates of CpG units

Data-S2.tab†: Coordinates of MDSs

Data-S3.tab†: Summary of sequence variants associated with 5-mCpG rates of MDSs

Data-S4.tab†: Summary of MDSs associated with expression of mRNA isoforms

Data-S5.tab†: Summary of CpG units associated with expression of mRNA isoforms

* Corresponding authors. E-mails: Olafur.Stefansson@decode.is (O.A.S.); kstefans@decode.is (K.S.)

† Coordinates correspond to those of the 1-based hg38 (human reference genome GRCh38).

1. Supplementary Notes

1.1. Nanopore sequencing and oxBS performed in the same DNA samples

We performed a comparison between nanopore sequenced DNA samples isolated from 132 individuals that had previously been analysed by oxidative whole-genome bisulfite sequencing (oxBS)[32]. We showed that the average coverage of nanopore sequenced DNA samples is indicative of how consistent the 5-mCpG rates are with CpG methylation measured by oxBS (**Supplementary Figure 2A&B**). We apply this result in o ur study by excluding individuals whose DNA samples have been sequenced to less than 10x average coverage.

We binned each CpG according to the number of sequences used to compute its 5-mCpG rate in each sample analyzed by nanopore sequencing. Based on the Pearson's correlation coef ficient we then determined how consistent the 5-mCpG rates in each such bin were to 5-mCpG rates measured with high coverage (>25 sequences for each CpG site) by oxBS (**Supplementa ry Figure 2C**). The results indicated that, for each DNA sample, the 5-mCpG rate was more reliably estimated when more than 9 sequences were used in the estimate.

For each CpG-unit, we calculated the average 5-mCpG rate across DNA samples from individuals measured by nanopore sequencing (n=7,179) and oxBS (n=132) to the n compute the Pearson's correlation coefficient across averaged 5-mCpG rates. We refer to this correlation as per CpG average Pearson correlation (PCAPC). By looki ng at all CpGs detected by both methods, we found strong correlation i.e., PCAPC=0.9594 (95%CI:0.9594,0.9595).

We identified five different attributes of CpG units (see methods section entitled: **CpG methylation analysis by nanopore sequencing**) that negatively affected the PCAPC (**Supplementary Figure 3A**), and we removed those CpG units from the study (**Supplementary Figure 3B**).

1.2. Long range phasing of sequences detected by Nanopolish and oxBS

Phasing of the sequences is important for resolving heterogeneity in the data as phasing enables us to quantify 5-mCpG rates of CpG sites separately for each of the two parental chromosomes. For arrays such as HumanMethylation BeadChip developed by Illumina, phasing is not possible and phasing of short sequences from whole-genome bisulfite sequencing is also problematic as, due to the short lengths of sequenced DNA, it is less likely that the sequences contain a known sequence variant that is heterozygous in the individual. Commonly used long-read phasing tools include Margin (github.com/UCSC-nanopore-cgl/margin) and LongPhase. We used a different approach, as we have nearly complete parent of origi n information for our genotype data[56], we could assign each sequence to a parental haplotype based on agreement with the haplotypes (see methods section entitled: **Assignment of 5-mCpG status to parental haplotypes**).

Short-read sequencing reads were assigned to parental haplotypes by examining the phasing status of 17,649,586 sequence variants and one heterozygous variant needed per sequence[32].

7.6% of sequences detected by oxBS were successfully phased whereas 39.5% of sequences detected by nanopore sequencing were successfully phased. The main limitation for the long-read phasing is the set of heterozygous SNPs per individual, as 97.6% of the nanopore sequences covering at least 1 heterozygous SNP, were phased.

On average, 21.87% of the CpGs were phased for oxBS per sample, but 81.05% for nanopore sequencing.

Additionally, we found that 8.61% of aligned bases were phased for oxBS, but 71.20% for nanopore sequencing.

The phasing is affected by sequence length which explains the fact that despite applying stricter conditions on the phasing for long-read sequencing we were still able to phase a higher fraction of sequences and more CpGs on average per individuals than for short-reads. Average sequence length (N50) for phased sequences was 15,798 bp for nanopore sequencing, while average length for sequences that were not phased was 9,732 bp.

When working with phased data, the set of high-quality CpGs should be refined to exclude CpGs where the phasing is often impossible. We calculated the fraction of phased sequences out of all sequences covering SNPs that we are able to phase for each CpG and suggest excluding CpGs where the fraction of phased sequences is less than 0.3 as performed in this study (see methods section entitled: **CpG methylation analysis by nanopore sequencing**).

1.3. Replication analysis: GoDMC study

The GoDMC[40] study is currently the largest study on the effect of sequence variants on 5-mCpG although testing less than 450 thousand CpGs (*versus* 15.3 million CpG units measured in our study). 152,794 of the sequence variants identified as *cis*-methylation QTLs by GoDMC were mapped to our high-quality set of ~34.4 million sequence variants according to chromosome position and alleles. Note, in line with our analysis, we restricted to GoDMC sequence variants located within 100kb of the affected CpG. 112,193 out of the 152,794 (73.4%) sequence variants identified as *cis*-methylation QTLs by GoDMC were in high LD (r2>0.80) to those identified in our study in association with 5-mCpG rates of individual CpG units.

To determine whether the observed overlap with GoDMC sequence variants is higher than expected by chance alone, we binned the ~1 million 5-mCpG associated sequence variants identified in our study, hereafter referred to as the observed sequence variants, into seven MAF bins (0-0.01,>0.01-0.05,>0.05-0.1,>0.1-0.2,>0.2-0.3,>0.3-0.4,>0.4-0.5). We counted the number of observed sequence variants in each MAF bin and sampled the same number of sequence variants from the 34.4 million sequence variants (located within 100kb from each of the 15.3 million measured CpG units). Hence, we obtained ~1 million sampled sequence variants matched to the observed sequence variants with respect to MAF. We found that approximately 24.4% of the ~1 million sampled sequence variants were in high LD to at least one sequence variant identified by GoDMC as a *cis*-methylation QTL. The 73.4% observed overlap with our findings is significantly higher than the 24.4% expected overlap (Binomial test, P <0.05).

Further, we tested the GoDMC *cis*-methylation QTLs in our cohort using the same sequence variants and CpG positions as reported by GoDMC. The effect sizes observed in our cohort were consistent with those reported by GoDMC (Pearson's r =0.736, 95%CI:0.734,0.739).

We note, however, that the effect sizes in our cohort tended to be lower than those reported by GoDMC; the regression effect on GoDMC effect sizes for deCODE effect sizes was

βˆ

$\hat{\beta}$

=1.57 (97%CI:1.56,1.58). 5-
mCpG rates measured using >12 sequences were highly consistent with rates measured by bisulfite se quencing (Supplementary Figure 2C). By restricting to haplotypes where >12 sequences were availab le for measuring the 5-
mCpG rate we demonstrated that both the coverage/read depth of the CpG unit and the number of obs ervations, i.e. the number of haplotypes measured with
>12x coverage/read depth, were likely important contributors to the observed bias towards lower effec t sizes in our cohort (**Supplementary Figure 7**).

1.4. Enrichment analysis: ASM-QTLs stratified by TSS proximity

We find that 38.4% (38.0,38.7%) of ASM-QTLs (i.e., 30,889 out of 80,503) reside within 25kb from the TSS of a protein coding gene. 53.1% (95%CI:52.4,53.8%) of GWA signals reside within 25kb from the TSS of a protein coding gene. Hence, in comparison to index variants of GWA signals, the index variants of ASM-QTLs are less frequently found near TSSs (<25kb).

We therefore sought to explore the question of whether proximity to protein coding TSSs influences t he enrichment of ASM-QTLs among GWA signals.

To this end we binned sequence variants (other than missense and loss of function) according to distance to the nearest TSS as follows: >25kb (distal), 10-25kb, 5-10kb, <5kb. We then determined the enrichment of these non-coding sequence variants depending on whether or not they are ASM-QTLs.

We found that ASM-QTLs located <5kb from protein coding TSSs were approximately 70-fold (95%CI:45.8,96.4) enriched among GWA signals (**Supplementary Figure 8**), whereas other non-coding sequence variants located <5kb from protein coding TSSs were less enriched i.e., 5.5-fold (95%CI:4.8,6.2). Hence, although the 70-fold enrichment for ASM-QTLs among GWA signals seems impressive, there is already this 5.5-fold enrichment given the TSS proximity (<5kb from a protein coding TSS).

Note, in this model, non-coding sequence variants located >25kb from a protein coding TSSs were used as the baseline annotation (neutral enrichment). The corresponding category of ASM-QTLs, i.e. those located >25kb from protein coding TSSs, were enriched by 19.3-fold (95%CI:13.3,26) among GWA signals (**Supplementary Figure 8**). This estimate is similar to the 23.2-fold enrichment of ASM-QTLs among GWA signals that we computed in a model without considering TSS proximity (**Extended Data Figure 7**).

Hence, we show here that non-coding sequence variants located close (<5kb) to protein coding TSSs are enriched among GWA signa ls. ASM-QTLs, however, are far more enriched among GWA signals regardless of whether or not they are located close to protein c oding TSSs.

1.5. Summary of main limitations and strengths of 5-mCpG detection by nanopore sequencing

Currently, the main **limitations** of using nanopore sequencing for 5-mCpG detection are as follows: 1) sequencing coverage in the DNA sample as low coverage is indicative of unreliable measurements of 5-mCpG rates in the sample. 2) The presence of

a sequence variant near or within a CpG site is problematic as this may cause erroneous 5-mCpG detection by some algorithms, like Nanopolish. 3) The observation of strand bias in 5-mCpG detection at a CpG site is fairly frequent and we demonstrated that this is indicative of erroneo us 5-mCpG detection.

The main **strengths** are as follows:
1) Nanopore sequencing delivers longer sequences than other sequencing methods, including whole g enome bisulfite sequencing. This property of nanopore sequencing considerably enhances the efficien cy of phasing the sequences. Phasing is important for resolving the heterogenity in the measurements of 5-mCpG rates, a task that cannot be accomplished with array-based methods. 2)
The large number of CpG sites detected by nanopore sequencing allows us to access regions in the ge nome that are not accessible using array-
based methods, thereby opening avenues for making novel discoveries.

In our previous work[64] we compared, in more detail, different algorithms and long read sequencing technologies for detecting 5-mCpG status in DNA samples. We note that Remora (of Guppy) exhibits less sensitivity to sequence variation in close proximity to the CpG site thereby enabling high-quality measurements of more CpGs than can be reliably measured by Nanopolish.

2. Legends for Extended Data Figures

Extended Data Figure 1: **Methylation depleted sequences**. Lines on the y-axis represent unmethylated haplotypes found in different individuals that overlap in position,
with chromosomal position on the x axis. (a) For each such "cluster of overlapping unmethylated haplotypes", as shown here, we look for the most frequently occurring unmethylated haplotype, shown in red. This "dominant" form is then taken as the representative for the cluster of unmethylated haplotypes, which then enables comparison in 5-mCpG haplotype rates over the same coordinates (those of the representative) across all individuals in the cohort. (b) An example cluster where two (or more) representatives were found which we identify, and measure separately as distinct entities.

Extended Data Figure 2: **Validation of ASM-QTL influences in oxBS-seq data**. ASM-QTLs validated using oxBS in independent DNA samples derived from whole blood of forty-five individuals. Note, these forty-five individuals were not included in the larger cohort of 7,179 nanopore sequenced individuals. The effect size of each ASM-QTL on the 5-mCpG haplotype rates of MDSs in oxBS sequenced individuals, y-axis, are plotted against the effect size of the same ASM-QTLs on 5-mCpG haplotype rates of the corresponding MDS in nanopore sequenced samples, x-axis. We binned the validation tests according to the number of haplotypes (n) that were available for validation testing in oxBS individuals as indicated on the top of each figure. A least squares regression line, where the effect sizes in the oxBS validation cohort are used as the outcome and those in nanopore sequenced cohort as the predictor, is shown in each figure, blue color, along with a diagonal line of identity, red color, as we expect the regression coefficient to be equal to one (and an intercept of zero) if the effect sizes are identical in the oxBS and nanopore sequenced individuals. The regression coefficients and their 95% confidence intervals are shown at the top left-hand corner of each figure. Note that as the number of informative haplotypes (n) in the oxBS validation cohort

increases, the regression coefficient approaches that of a diagonal line indicating that the effect sizes become more similar as the number of haplotypes available for testing (n) in the oxBS cohort increases.

Extended Data Figure 3: **MDSs associated with gene expression.** Distance, x-axis, from MDS midpoints to the TSS locations of the associated mRNA transcript isoforms *versus* the effect size on expression, y-axis. MDS midpoints are represented as black dots, and the MDSs are represented as ranges i.e., horizontal lines. (a) Distal associations i.e., where the MDS does not contain the TSS of the associated mRNA. (b) Promoter associations i.e., where the MDS contains the TSS of the associated mRNA, where the MDS is represented by a line (left to right) indicating the position relative to the TSS of the associated mRNA. (a) and (b) TSS is represented as the origin of the x-axis (x=0), shown as a vertical dashed line, red. Note, we modify the sign of the distance such that negative distances reflect upstream positions of the MDS midpoint relative to the TSS of the associated mRNA. Hence, the upstream positions indicate that the MDS midpoint does not intersect with the direction of mRNA transcription. In contrast, downstream positions indicate that the MDS midpoint does intersect with the direction of mRNA transcription.

Extended Data Figure 4: Functional attributes of ASM-QTLs. ASM-QTLs were more likely than expected by chance, to be found in high LD (r^2 >0.80) to A) sequence variants that influence the DNA binding affinity to transcription factors SPI1, CTCF and EBF1, and cohesin protein STAG1 and B) sequence variants located within regulatory sequences defined by ENCODE[33], i.e., candidate cis-regulatory elements (*version 3*). Note, sequence variants that influence ASBs were identified in Chen et al[42]. The *P*-values shown, unadjusted for multiple comparison, were computed using our permutation-based method described under methods section „ASM-QTL in functional annotation maps“.

Extended Data Figure 5: ASM-QTLs dominate in correlations between MDSs and mRNA. (**a)** Example of an association found between expression of VAMP5-201 (ENST00000306384), y-axis, and 5-mCpG haplotype rates, x-axis, of an MDS (chr2:85584168-85584976) located within the CpG island promoter of the TSS; i.e., an example of CpG island promoter methylation. (**b)** The fraction of variance in VAMP5-201 mRNA isoform expression explained by (*top)* ASM-QTL found in association with the MDS *(middle)* CpG methylation of the MDS and *(bottom)* CpG me thylation of the MDS after correction for the ASM-QTL. This same association between MDS at chr2:85584168-85584976 and mRNA expression of VAMP5-201 is then shown separately on (**c)** the reference- and (**d)** the alternative alleles of ASM-QTL chr2:85580659:IG.0:0, respectively. Least squares regression line is shown within each of three scatter plots, in blue color, and the estimated regression coefficient and their 95% confidence intervals are shown at the bottom, along with the corresponding two-sided *P*-values.

Extended Data Figure 6: Modeling the impact of ASM-QTL on methylation and expression. Model diagrams (leftmost column) describing the four different hypotheses on the mechanism by which an ASM-QTL genotype (G) influences CpG methylation (M) and gene expression (E). The outcomes of two different tests (MR-Steiger and our method of comparing

βˆGMβˆGE

$\hat{\beta}_{GM}\hat{\beta}_{GE}$

and

σˆ2MβˆME

$\hat{\sigma}_M^2\hat{\beta}_{ME}$

) are shown; "v" shape indicates that the model is supported by the test, whereas "x" indicates that the model is not supported by the test and empty cells indicates that the test is unable to consider whether or not the model is supported. For clarity, each model is labelled with a number which are then used as reference to each diagram in the main text and in **Figure 3**.

Extended Data Figure 7: Varied contribution of sequence variants to human trait variability. Enrichment, x-axis, of ASM-QTLs and other sequence variant annotations, y-axis, among GWA signals found in various human diseases and other traits. The solid points (black) represent the measure of center, i.e., the enrichment point estimates and the horizontal lines (black) represent their 95%CIs. The number of sequence variants in each annotation is shown within parentheses on the y-axis. Shown are the enrichment point estimates, points, and their 95% confidence intervals, horizontal lines, for each sequence variant annotation, and the point of neutral enrichment on the x-axis as vertical line, blue. UTR=Untranslated Region, DHS=DNase hypersensitivity site, MDSs=Methylation depleted sequences.

3. Legends for Data files

Data files can be downloaded from our website: www.decode.com/summarydata/

Data-S1.tab: Sequence variants associated with 5-mCpG rates of individual CpG units; GRCh38 coordinates, 1-based.

Chrom*: The chromosome of sequence variants and CpG units*

SeqVariant_start*: The nucleotide starting position of the sequence variant*

SeqVariant _end*: The nucleotide ending position of the sequence variant*

SeqVariant _name*: The inhouse name of the sequence variant*

SeqVariant _rsname*: The rs-identifier of the sequence variant*

SeqVariant _ref*: The reference allele of the sequence variant*

SeqVariant _alt*: The alternative allele of the sequence variant*

SeqVariant _minorallele*: The minor allele of the sequence variant*

SeqVariant _MAF*: The minor allele frequency of the sequence variant*

SeqVariant _LD_count*: The number of sequence variants found in high LD (r^2>0.8) to the sequence variant*

SeqVariant _vartype*: sequence variant type (SNV=single nucleotide variant, Indel, Micosatellite and SV=Structural variant)*

SeqVariant _rank*:*
The rank of the sequence variant: primary=the most significantly associated variant, secondary= the most significantly associated variant on the major allele of the primary

CpG_start*: The nucleotide starting position of the CpG unit*

CpG_end*: The nucleotide ending position of the CpG unit*

CpG_ref_methylrate*: The average 5-mCpG rate of the CpG unit on reference alleles*

CpG_alt_methylrate*: The average 5-mCpG rate of the CpG unit on alternative alleles*

SeqVariant _5mCpG_effectsize*: The effect size (SD units) of the sequence variant on 5-mCpG rates of the CpG umit*

SeqVariant _5mCpG_95CI*: The 95%CI of the sequence variant effect size*

n_haplotypes*: The number of informative haplotypes used to identify the sequence variant*

GWAS_disease*: Comma separated list*
of disesases (or disorders) found in association to a sequence variants in high LD (r^2>0.8) to the 5-mCpG-associated sequence variant

GWAS_disease_markers*: Comma separate list of disease-associated sequence variants in high LD (r2>0.8) to the 5-mCpG-associated sequence variant. Format: disease-associated variant name (disease-associated variant effect size | r^2 of disease-associated variant to 5-mCpG-*
associated sequence variant), comma separated list in the same order as „GWAS_disease“ column

GWAS_other.traits*: Similar to „GWAS_disease_markers“ columnt except for traits not classified as e ither disease or disorder*

GWAS_other.traits_markers*: Similar to „GWAS_disease_markers“ except for traits not classified as either disease or disorder, comma separated list in the same order as „GWAS_other.traits“ column*

Data-S2.tab: The coordinates of the methylation depleted sequences (MDSs); GRCh38 coordinates, 1-based.

Chrom*:The chromosome of the MDS*

Start*:The nucleotide starting position of the MDS*

End*:The nucleotide ending position of the MDS*

MDS*:Coordinate-identifier of the MDS*

Class*:The type of MDS*

Data-S3.tab: Sequence variants associated with 5-mCpG rates of MDSs; GRCh38 coordinates, 1-based.

Chrom*: The chromosome of sequence variants and MDSs*

SeqVariant_start*: The nucleotide starting position of the sequence variant*

SeqVariant _end*: The nucleotide ending position of the sequence variant*

SeqVariant _name*: The inhouse name of the sequence variant*

SeqVariant _rsname*: The rs-identifier of the sequence variant*

SeqVariant _ref*: The reference allele of the sequence variant*

SeqVariant _alt*: The alternative allele of the sequence variant*

SeqVariant _minorallele*: The minor allele of the sequence variant*

SeqVariant _MAF*: The minor allele frequency of the sequence variant*

SeqVariant _LD_count*: The number of sequence variants found in high LD (r^2>0.8) to the sequence variant*

SeqVariant _vartype*: sequence variant type (SNV=single nucleotide variant, Indel, Micosatellite and SV=Structural variant)*

SeqVariant _rank*:*
The rank of the sequence variant: primary=the most significantly associated variant, secondary= the most significantly associated variant on the major allele of the primary

MDS_start*: The nucleotide starting position of the MDS*

MDS_end*: The nucleotide ending position of the MDS*

MDS_ref_methylrate*: The average 5-mCpG rate of the MDS on reference alleles*

MDS_alt_methylrate*: The average 5-mCpG rate of the MDS on alternative alleles*

SeqVariant _5mCpG_effectsize*: The effect size (SD units) of the sequence variant on 5-mCpG rates of the MDS*

SeqVariant _5mCpG_95CI*: The 95%CI of the sequence variant effect size*

n_haplotypes*: The number of informative haplotypes used to identify the sequence variant*

GWAS_disease*: Comma separated list of disesases (or disorders) found in association to a sequence variants in high LD (r^2>0.8) to the MDS-associated sequence variant*

GWAS_disease_markers*: Comma separate list of disease-associated sequence variants in high LD (r2>0.8) to the MDS-associated sequence variant. Format: disease-associated variant name (disease-associated variant effect size | r^2 of disease-associated variant to MDS-associated sequence variant), comma separated list in the same order as „GWAS_disease" column*

GWAS_other.traits*: Similar to „GWAS_disease_markers" columnt except for traits not classified as e ither disease or disorder*

GWAS_other.traits_markers*: Similar to „GWAS_disease_markers" except for traits not classified as either disease or disorder, comma separated list in the same order as „GWAS_other.traits" column*

Data-S4.tab: 5-mCpG rates of MDSs found in association with expression of mRNA isoform; GRCh38 coordinates, 1-based.

Chrom*: The chromosome of the MDS*

Start*: The nucleotide starting position of the MDS*

End*: The nucleotide ending position of the MDS*

MDS*: Coordinate-identifier of the MDS*

mRNA_isoform*: Ensembl identifier of the mRNA isoform*

Gene*: The gene identifier of the mRNA isoform*

effectsize*: The effect size (SD units) of the 5-mCpG rate of the MDS on expression level of the mRNA isoform*

p_value*: P-value for the MDS in the association between 5-mCpG rates and mRNA isoform expression*

Data-S5.tab: 5-mCpG rates of CpG units found in association with expression of mRNA isoform; GRCh38 coordinates, 1-based.

Chrom*: The chromosome of the CpG unit*

Start*: The nucleotide starting position of the CpG unit*

End*: The nucleotide ending position of the CpG unit*

CpG*: Coordinate-identifier of the CpG unit*

mRNA_isoform*: Ensembl identifier of the mRNA isoform*

Gene*: The gene identifier of the mRNA isoform*

effectsize*: The effect size (SD units) of the 5-mCpG rate of the CpG unit on expression level of the mRNA isoform*

p_value*: P-value for the CpG unit in the association between 5-mCpG rates and mRNA isoform expression*

4. Supplementary Figures

Supplementary Figure 1: Shown are histograms for N50, upper panel, and the average coverage for each DNA sample, lower panel.

Supplementary Figure 2: Nanopore sequencing and oxBS performed in the same DNA samples. The consistency in 5-mCpG rates measured by nanopore sequencing and oxBS in DNA samples isolated from the same 132 individuals was estimated by (A) Pearsons r correlation coefficient, y-axis, and (B) mean of the absolute differences in 5-mCpG rates of each CpG site, y-axis, with respect to nanopore sequencing coverage in each sample on the x-axis. Note, the y-axes in A and B are limited to enhance readability of the data. (C) CpG coverage underlying the 5-mCpG rates, i.e., the number of sequences that were used to compute the 5-mCpG rate for a given CpG, in the nanopore sequenced samples, x-axis, influences the consistency (Pearsons r), y-axis, with 5-mCpG rates of CpGs measured with high coverage (>25 sequences per CpG) by oxBS. The center line (solid black) shown in each box represents the median; box limits represent upper and lower quartiles; vertical lines represent the range spanning 1.5x quartiles. Note, the y-axis is limited (0.5, 1). Nanopore sequencing and oxBS performed in the same DNA samples.

Supplementary Figure 3: The quality of 5-mCpG rate measurements varies by different attributes of the sequenced DNA. (A) Shown are PCAPC estimates, x-axis, for CpG sites located outside (orange colour) and inside (brown colour) of different attributes of the sequenced DNA, y-axis, and the PCAPC estimate based on all CpGs (vertical line, black colour). (B) The number of CpG-units (red) and CpG-sites (blue), x-axis, found inside of each attribute, y-axis. Attributes on y-axis are as follows (top to bottom): „Dark regions“ contain sequence repeats that cause mapping to be unreliable, „Low FRR“ (Fraction of Reliable Reads ≤0.5) is indicative of problematic sequences, „Coverage outliers“ (defined where the average is <0.5 or >1.5) are indicative of repetitive regions, „high SB“ (Strand Bias ≥0.20) is a measure of strand bias in 5-

mCpG detection based on the absolute difference between methylation ratio on forward and reverse st rand, „Close to variant“ indicates the presence of a known SNP locus (MAF>0.001) within 5bp of measured CpGs.

Supplementary Figure 4: Depletion of 5-mCpGs in functional regions. (A) Genomic attributes defined in the well-studied B-lymphoblastoid GM12878 cell line, i.e. regions marked by histone modifications (ChIP-seq) along with transcriptional start sites (RAMPAGE assay) and CpG islands. 5-mCpG measured by WGBS in GM12878 (Encode project), y-axis, and measures of 5-mCpG by nanopore sequencing of whole blood samples, x-axis. (B) TF protein binding sites defined in B-lymphoblastoid GM12878 cell line have similar 5-mCpG analyzed by WGBS in GM12878, y-axis, as those measured by nanopore sequencing in whole blood samples, x-axis. CBX3 is also known as HP1-γ (heterochromatin protein 1 homolog gamma). In both (A) and (B), we show the means of 5-mCpG rates of the inner core region (-50 to +50bp relative to midpoint) and 95%CI for each genomic attribute defined in GM12878; i.e. measures of mean methylation within the same region coordinates for GM12878 (WGBS), y-axis, and our whole-blood samples (nanopore sequencing), x-axis. Additionally, a diagonal line (y=x) is drawn in both (A) and (B). 50 thousand „random_peaks“ were selected from randomly selected locations in the genome shown on the figure as genome-wide reference points. A trend towards overall lower 5-mCpG in B-lymphoblastoid GM12878 cells is a well-known feature of methylomes in cultured cell lines. Abbreviations: DHS=DNase hypersensitivity sites, TSS=Transcription start sites.

Supplementary Figure 5: Pearson´s correlation coefficient between PCs and covariates. The su m of r^2 for each covariate (y-axis) across PCs 1 to 10 are shown on the right-hand side of the heatmap and, within parentheses, the sum across PCs 1 to 5. The correlations were computed for each covariate against each of the ten PCs; here , using the PCs computed from a random sample of CpG units (hq) located on autosomes except for chromsome 1. Abbreviations: WBC=White blood cells, RBC=Red blood cells, PCTERR=Percent error, PCTALN=Percent alignment, NULTRA=Number of ultra long sequences, NRBC=Fraction of nucleated red blood cells, NE=Neutrophils, MO=Monocytes, LY=Lymphocytes, IG=Immature granulocytes, HGB=Hemoglobin, EO=Eosinophils, BA=Basophils.

Supplementary Figure 6: Percent of the variance in 5-mCpG rates explained by each of the first ten PCs. The first five PCs, vertical line (red), explain ~ 2.84% of the variance and each of the other PCs (i.e. the sixth, seventh etc.) explain less than 0.12%. The PCs were computed using randomly sampled autosomal CpG units (hq). PC=Principal compone nt.

Supplementary Figure 7: GoDMC *cis*-methylation QTLs tested for association in deCODE cohort after restricting to haplotypes with at least 12x coverage/read depth for measuring the 5-mCpG rate. The effect size (SD units) of sequence variants in the GoDMC study, y-axis, plotted against the effect size (SD units) of the same sequence variants on the same CpGs in our cohort, x-axis. The number of observations, i.e. the number of haplotypes with >12x reads (based on Suppleme ntary Figure 2C), that were available to test for associations between a *cis*-meQTL (GoDMC) sequence variant and CpG unit methylation in our cohort is shown at the top of eac h figure as follows: (A) >100 haplotypes (107,046 *cis*-meQTLs), (B) >1000 haplotypes (30,332 *cis*-meQTLs), (C) >1500 haplotypes (9,807 *cis*-meQTLs) and (D) >2000 haplotypes (1,661 *cis*-meQTLs). The regression coefficient on GoDMC effect sizes, y-axis, for deCODE effect sizes, x-axis, is shown in blue colour at the top-left of each figure. The least squares regression line (blue) and diagonal line (red) are shown.

Supplementary Figure 8: The enrichment of ASM-QTLs among GWA signals varies according to distance from protein coding TSSs. The enrichment, x-axis, of each ASM-QTLs category among GWA signals found in various human diseases and other traits relative to other sequence variant annotations, x-axis. The solid points (black) represent the measure of center, i.e., the enrichment point estimates and the horizontal lines (black) represent their 95%CIs. The number of sequence variants in each annotation is shown within parentheses on the y-axis. Shown are the enrichment point estimates, points, and their 95% confidence intervals, horizontal lines, for each sequence variant annotation, and the point of neutral enrichment on the x-axis as vertical line, blue. TSS=Transcription Start Site.

5. Supplementary Tables

Supplementary Table 1: Annotations of **s**equence variants used in the analysis of enrichment among trait associated sequence variants.

Index	Name	Merged from	Annotation description
1	missense_variant		*Variant resulting in substitution of an amino acid in the protein product of a gene.*
2	synonymous_variant		*Variant that does not affect the amino acid sequence of the protein product.*
3	splice_region_variant		*Variant located in region of splice site, 1-3 bases into the exon or 3-8 bases into the intron.*
4	3_prime_UTR_variant		*Variant located in the untranslated region of the 3´end of a gene.*
5	5_prime_UTR_variant		*Variant located in the untranslated region of the 5´end of a gene.*
6	DHS footprints, hematopoietic cells		*DHS footprints defined in hematopoietic cell types, defined by Vierstra et al*[6]

7	DHS footprints, non-hematopoietic cells		*DHS footprints defined in other, non-hematopoietic cell types, defined by Vierstra et al*[6]
8	intron_variant		*Variant located in an intron of a gene.*
9	loss_of_function	frameshift, stop gained, start lost, splice_donor_variant, splice_acceptor_variant	*Variant predicted to have loss of function consequences to a gene with high confidence according to loftee*[61]
10	ASM-QTL		*Alleles of sequence variants found in association with changes in 5-mCpG rates of MDSs on the same haplotype (defined in methods section entitled:* ***Allele-specific methylation quantitative trait loci)****.*
11	Methyl-depleted sequences, MDSs		*Variants located within DNA sequences depleted of 5-mCpGs; MDSs were defined as described in methods section entitled:* ***MDSs (methylation depleted sequences)****.*
12	Other	Merged from multiple other VEP annotations, but not those from 1 to 11.	*Variant that does not belong in any of the categories from 1 to 11.*

6. List of Source Data

6.1. Source data files for the main figures

Figure 1: source-data-fig1.xlsx

Figure 2: source-data-fig2.xlsx

Figure 3: Not applicable.

Figure 4: source-data-fig4.xlsx

6.2. Source data files for extended data figures

Extended Data Figure 1: Not applicable.

Extended Data Figure 2: source-data-EDFig2.xlsx

Extended Data Figure 3: source-data-EDFig3.xlsx

Extended Data Figure 4: Not applicable.

Extended Data Figure 5: source-data-EDFig5.xlsx

Extended Data Figure 6: Not applicable.

Extended Data Figure 7: source-data-EDFig7.xlsx

Appendix III: Paper III - Supplementary materials

Supplementary Notes 1

Construction and validation of a methylation aging clock

The methylation aging clock selected 1,373 CpGs using Lasso regression. The model achieved a median absolute error (medAE) of 2.34 years (Supplementary Notes Fig 1A-B, Table 1, training medAE = 2.22 years, test medAE = 2.43), outperforming Horvath's benchmark DNAmAge clock (medAE = 2.7 years)[45]. The model showed decreased performance in individuals under 18 years old, likely due to their exclusion from training (Supplementary Notes Fig. 1A,B). The model retained high accuracy in an independent validation cohort (medAE = 2.15 years, Supplementary Notes Fig. 1C). Model performance remained robust even when top-ranked CpGs were iteratively removed and the model retrained 4 times, indicating redundancy among predictive sites (Supplementary Notes Table 2). Interestingly, adjusting the data for the first 10 principal components before training resulted in higher error (Supplementary Notes Table 1).

Phased methylation clocks

Following the same steps, we created a phased methylation aging clock. The phased clock selected 3,001 CpGs by Lasso regression. The model achieved a training medAE 3.70 years and test medAE 3.86 years (Supplementary Notes Table 1). We further trained separately a clock for paternal and maternal haplotypes. Both parent-specific clocks had higher error than clocks trained on unphased data. Out of the two parent-of-origin specific clocks, the medAE for paternal-specific methylation clock was slightly lower than for the maternal-specific, independent of CpG set size ($medAE_{maternal}$=12.05 > $medAE_{paternal}$= 11.81 years, Wilcoxon rank sum test p-value = 0.019, Supplementary Notes Fig. 2, Table 3).

Our results suggest that phased clocks are less accurate than traditional methylation aging clocks. Because phasing reduces the number of reads supporting each CpG by approximately half, the accuracy of methylation estimates per CpG is diminished. Given that age-associated effects on methylation are usually small, this reduction in coverage likely contributes to the higher error observed in phased methylation clocks.

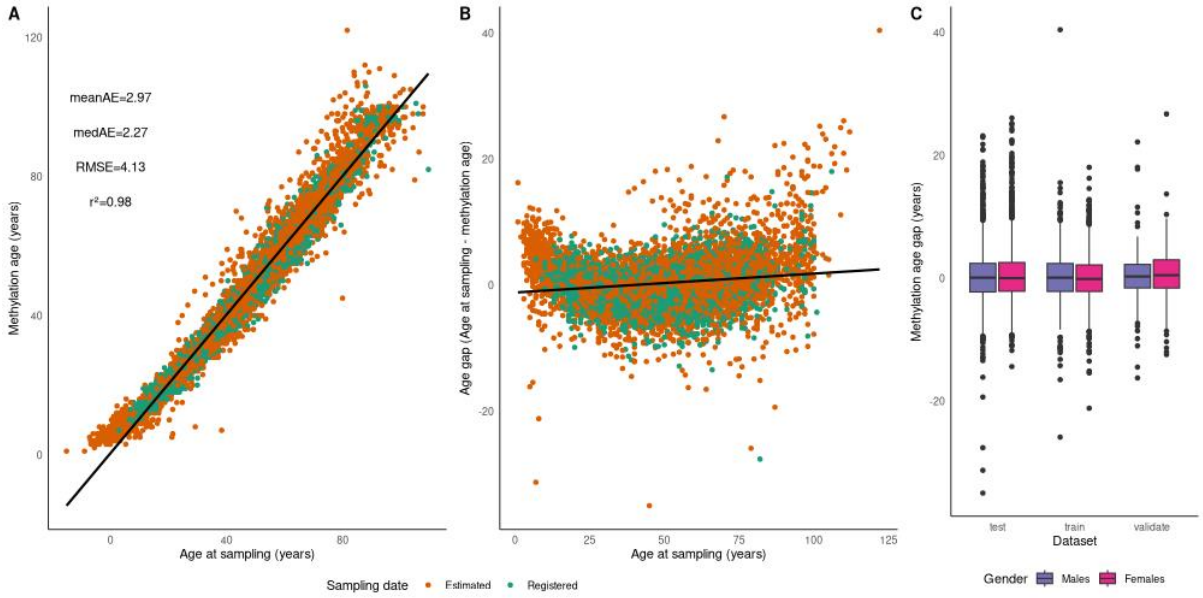
Supplementary Notes 2

Validation of age associations

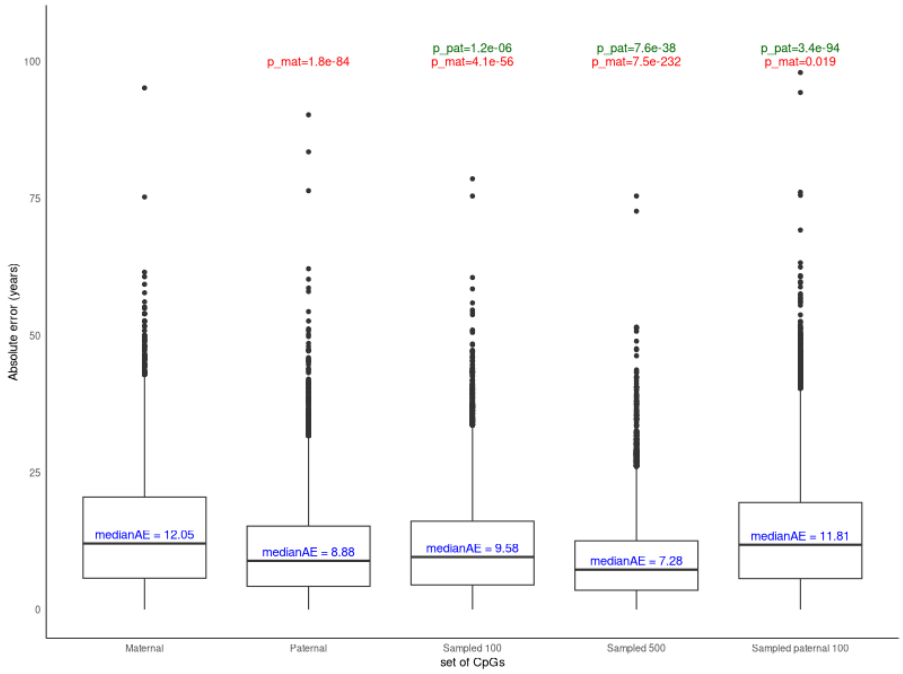
We found that 3,736,010 out of 3,739,875 (99.9%) age-associations are replicated without covariates adjustment and standardization of methylation levels ($p<0.05$), thereof 83.4% were Bonferroni significant. Correlation coefficient (Pearson's r) between effect size was 0.95, suggesting that adjusting for covariates and the rank normal transformation are not driving the associations or the size of the effect. Further, permutation test confirmed that our test statistic is well calibrated (Supplementary Table 6). In independent cohort of 545 samples from Intermountain, we found 710,836 out of 3,739,875 (19%) age-association are replicated ($p<0.05$) and 3,106,669 (83%) have the same direction of effect size and Pearson's correlation coefficient between the two effect sizes was 0.58.

Similarly, for the 1,274,315 parent-of-origin specific age associations were replicated for a model without covariate adjustment and without standardization of methylation levels ($p<0.05$), there of 87.9% were Bonferroni significant. Moreover, all 702 asymmetric age-associated CpGs are nominally significant without standardization, there of 85.0% were Bonferoni significant. As we cannot assign haplotypes to parent-of-origin in our independent cohort, we did not attempt to replicate the parent-of-origin specific age association.

Supplementary notes figures



Supplementary Notes Fig. 1 Properties of DNAmAge clock generated from ONT data. **(A)** Relationship between predicted methylation age (years, y-axis) and age at sampling (years, x-axis). Mean absolute error (AE), median AE, RMSE and R-squared shown on figure. Best line shown in black. **(B)** Relationship between predicted age (y-axis) and age at sampling (x-axis). The colours represent whether the date of the sample draw was estimated from run date (orange) or accurately registered (green). **(C)** Distribution of methylation age gap (y-axis) across test, train and validation set (x-axis).



Supplementary Notes Fig. 2 Boxplot showing the absolute error (y-axis) distribution for clocks trained on different set of CpGs (x-axis). First two are trained on haplotype-specific methylation associated with age on maternal haplotype versus paternal haplotype, next two are trained on a subset of 100 and 500 CpGs that were strongly associated with age and the last one is subset of 100 CpGs from the paternal dataset. Median absolute error in each set is written in blue. P-values from Wilcoxon rank sum test between maternal distribution and other set of CpGs is written in red (p_mat). P-values from Wilcoxon rank sum test between paternal distribution and other set of CpGs is written in green (p_pat).

Supplementary Notes Tables

Supplementary Notes Table 1. Summary of the performance of different epigenetic age models.

	Info	Mean abs error	Median abs error (MedAE)	RMSE	Number of non-zero features	Test error (MedAE)	Validation error (MedAE)
500k top age associating CpGs	Linear regression to find top 1M age associating CpGs, with effect size >15. 500,000 CpGs selected at random	2.46	1.89	3.33	1476	3.01	3.92
500k top age associating CpGs + 10PCs	Linear regression to find top 1M age associating CpGs, with effect size >15. 500,000 CpGs selected at random and adjusted for first 10 principal components	4.95	5.51	6.68	3214	-	-
Top 500k age associated parent-of-origin specific CpGs	Linear regression to find top 1M parent-of-origin specific age associating CpGs and effect size > 1.5. 500,000 CpGs selected at random	4.92	3.70	6.89	3,001	3.86	-
Maternal-specific age association	133 CpGs with maternal-specific age association	13.40	11.50	16.70	115	12.56	-
Paternal-specific age association	539 CpGs with paternal-specific age association	9.85	8.02	12.50	316	9.63	-

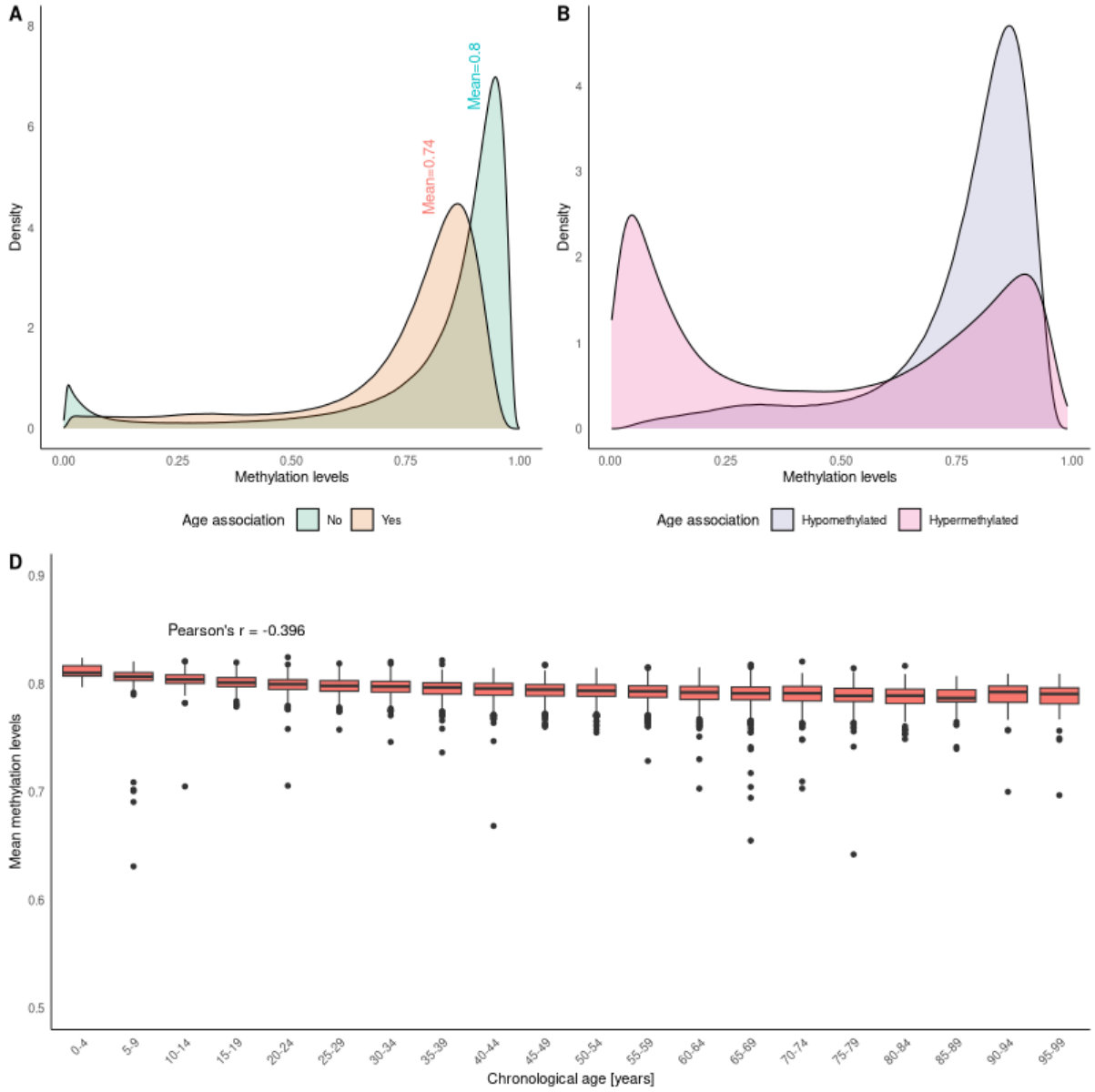
Supplementary Notes table 2. Results for training 5 clocks by iteratively removing selected CpGs each round.

	Train median abs error (medAE)	**Train RMSE**	**Number of non-zero features**	**Test medAE**
Round 1	2.22	3.71	1,373	2.41
Round 2	2.04	3.83	1,053	3.32
Round 3	1.90	3.69	2,090	4.09
Round 4	1.99	3.59	2,036	3.77
Round 5	2.00	3.80	2,036	3.96

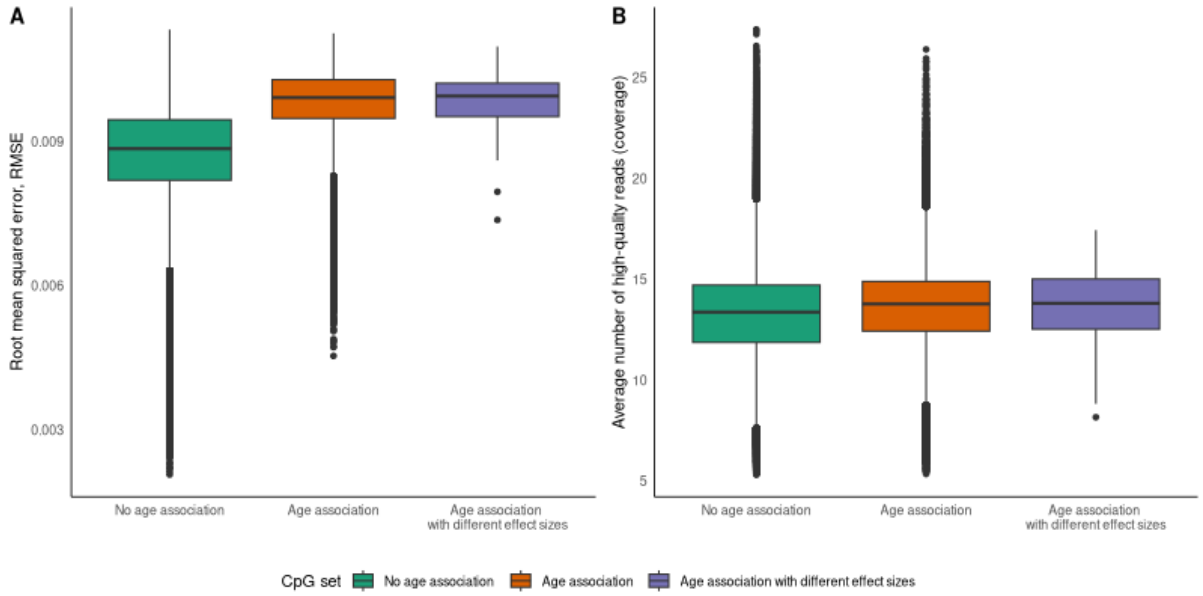
Supplementary Notes Table 3. Comparison of performance of two clocks trained on haplotype-specific methylation associated with age on maternal haplotype versus paternal haplotype.

Clock	**Mean abs train error**	**Median abs train error (MedAE)**	**Train RMSE**	**Number of non-zero features**	**Test error (medAE)**
Maternal	13.40	11.50	16.70	115	12.56
Paternal	9.85	8.02	12.50	316	9.63
Sampled 100	11.10	9.39	13.93	82	11.50
Sampled 500	8.21	6.86	10.34	289	7.61
Sampled paternal 100	13.42	11.59	16.60	83	12.76

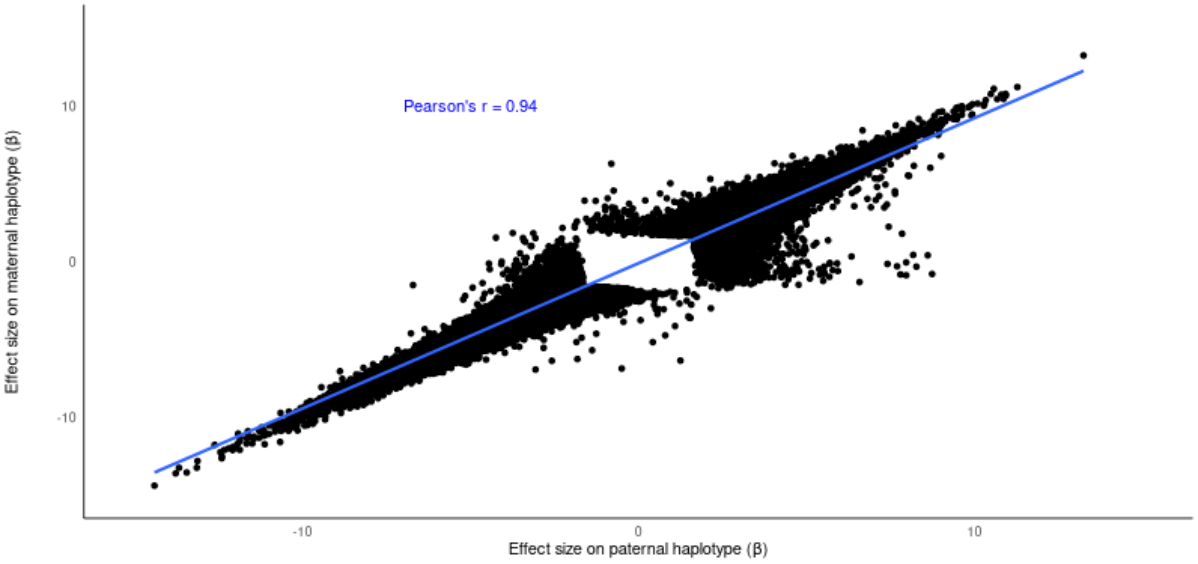
Supplementary Figures



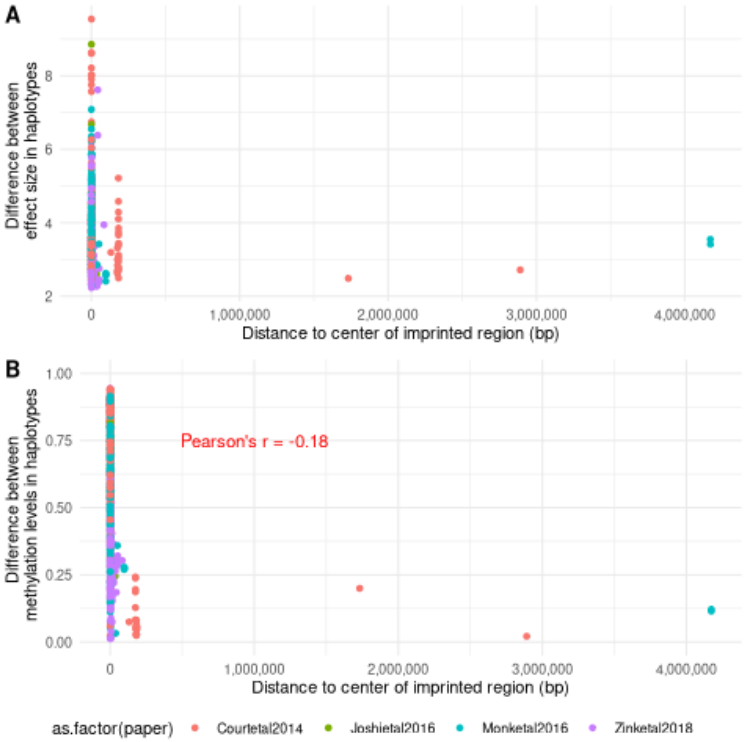
Supplementary Fig. 1. Genome-wide changes in methylation with age. **(A)** Density plot showing the distribution of methylation levels at CpG sites with (orange) and without (green) Bonferroni significant ($p<2.8 \cdot 10^{-9}$) age associations. Mean methylation levels of each distribution are written on top. **(C)** Density distribution of methylation levels at CpG sites that are loosing methylation (hypermethylation, pink) compared to CpGs that are gaining methylation (hypermethylation, purple). **(D)** Boxplot showing the relationship between chronological age, binned by 5 years (x-axis) and mean methylation levels (y-axis).



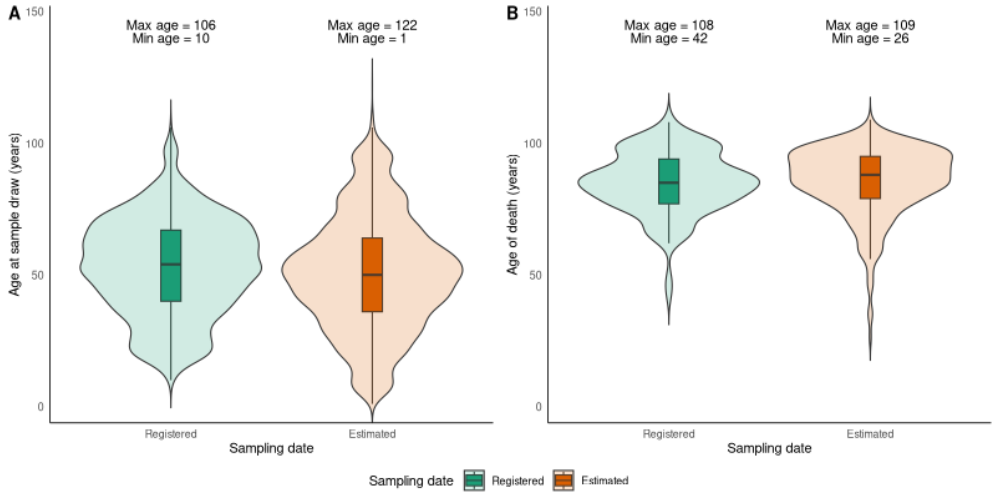
Supplementary Fig. 2 Distribution of residual variance of methylation for three sets of CpGs calculated over haplotypes. **(A)** Boxplot showing the distribution of root mean squared residuals from linear regression between age and methylation levels per CpG, shown for non-age associated, age associated CpGs with no differece between haplotypes and age associated CpGs with different estimate sizes, calculated over haplotypes. **(B)** Boxplot showing the distribution of number of high-quality reads on average behind the same three groups of CpGs.



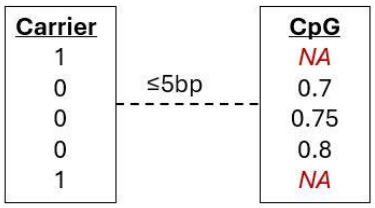
Supplementary Fig. 3. Comparison of 1,274,315 CpGs that exhibit age associations on at least one haplotype. Effect sizes for maternal haplotype are shown on y-axis and paternal haplotypes on x-axis. Blue line shows the best-fit line, with Pearson's r = 0.94.



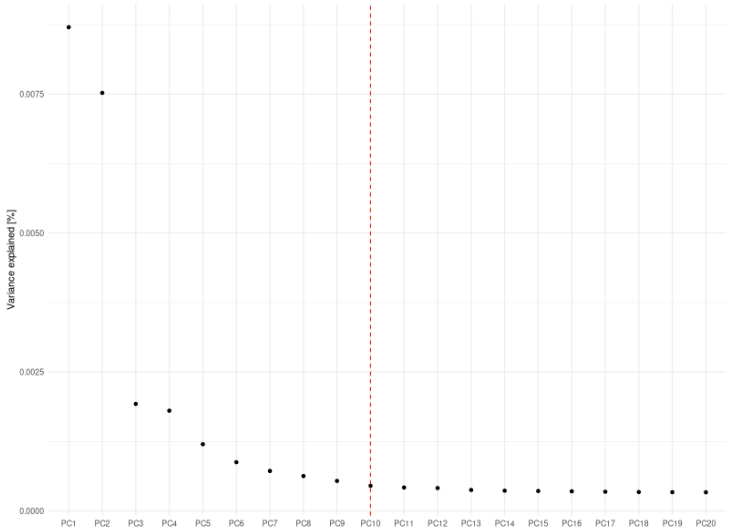
Supplementary Fig. 4 Haplotype-specific methylation age association has stronger effect size near imprinted regions. **(A)** Difference in effect sizes of maternal- and paternal haplotypes (y-axis) with respect to proximity to imprinted region (x-axis). **(B)** Difference in methylation levels of maternal- and paternal haplotypes (y-axis) with respect to proximity to imprinted region (x-axis). Pearson correlation coefficient between variable on x- and y-axis is shown in red.



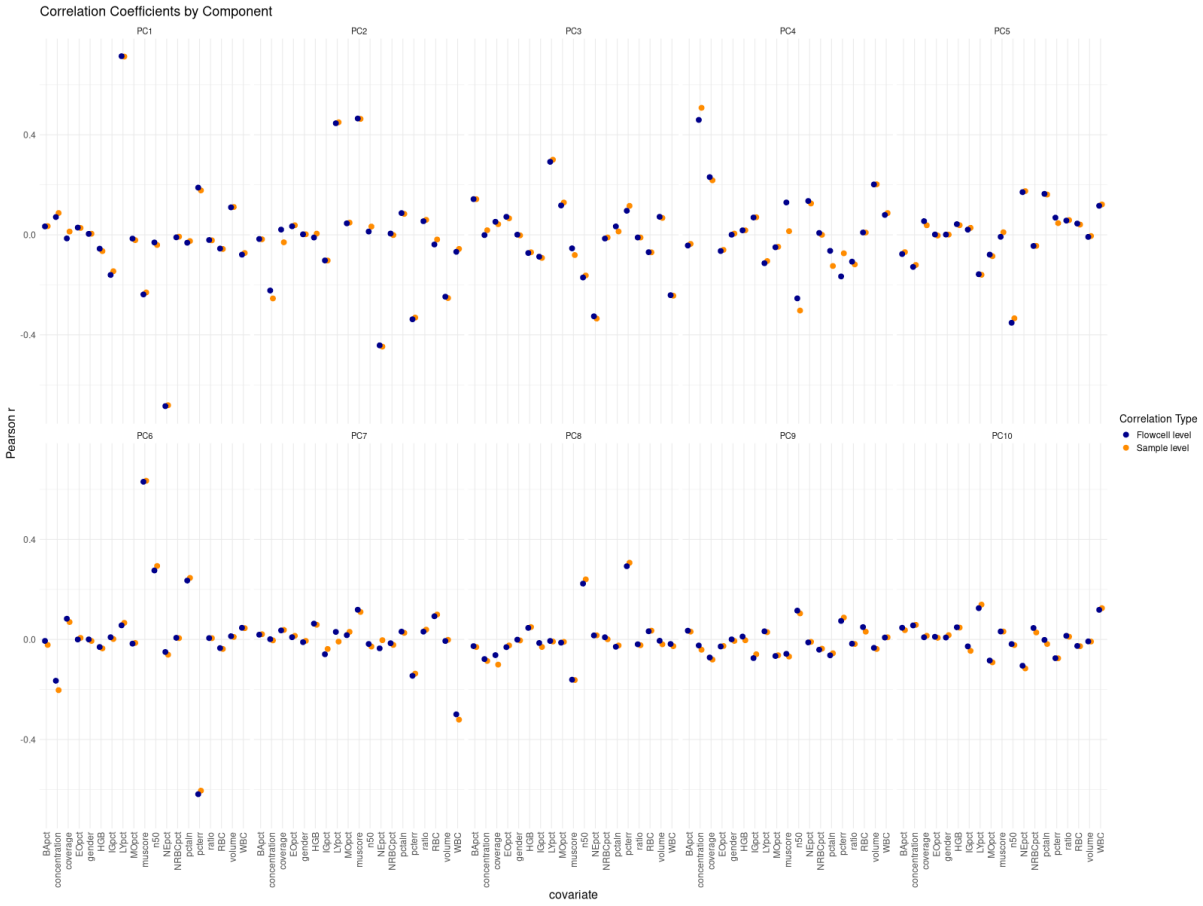
Supplementary Fig. 5. Violin plot with boxplot showing the distribution of **(A)** Age at sampling in years (x-axis) and **(B)** age at death in years (x-axis). Orange represents the estimated date of sample draw while Green represents the accurately registered date of sampling (y-axis).



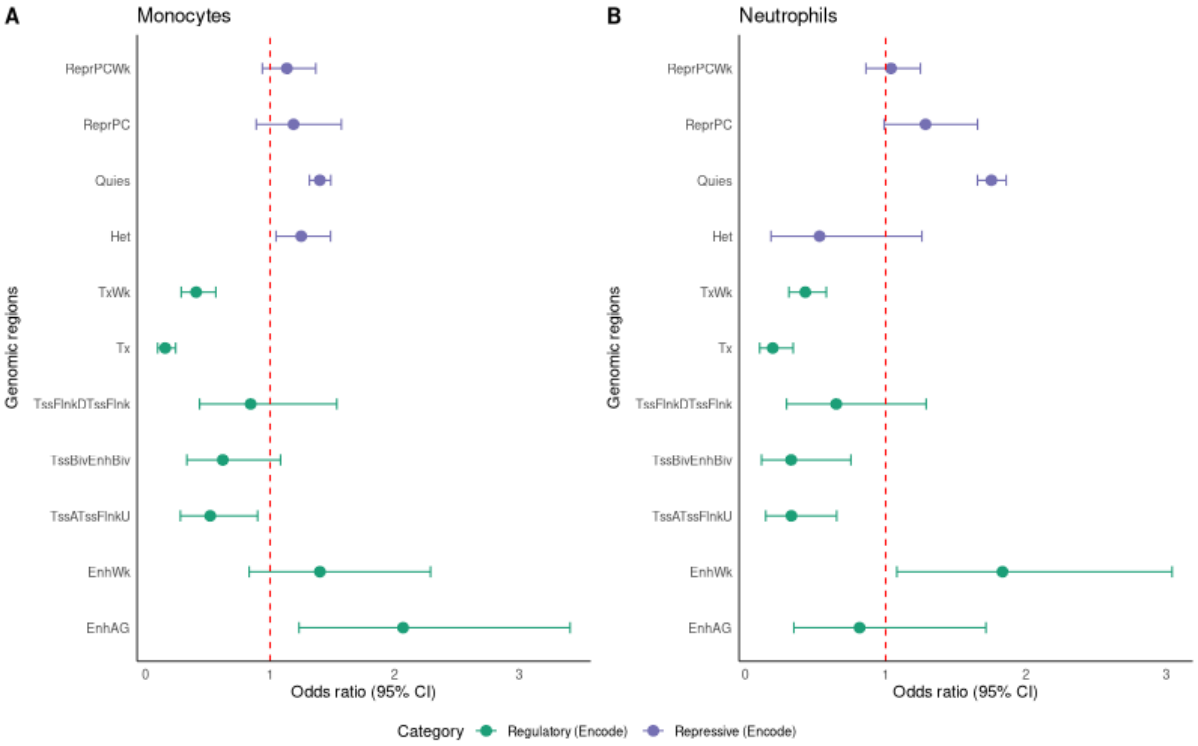
Supplementary Figure 6: Schema showing how we deal with common variants near CpGs. "Carrier" box shows the genotype of 5 individuals and "CpG" box shows the corresponding methylation levels.



Supplementary Fig. 7. Variance explained in percentages (y-axis) for each principal component (1-20, x-axis). Red dashed line represents the cutoff of 10 PCs that we selected.



Supplementary Fig. 8. Pearson correlation (y-axis) shown for each of the covariates (x-axis) for each PC separately. Blue represents the correlation on flowcell level and orange on sample level, where the values have been averaged over multiple flowcells.



Supplementary Fig. 9. Genomic enrichment in regulatory (green) and repressive (purple) regions selected from **(A)** monocytes and **(B)** neutrophils. Enrichment and depletion of age-associated CpGs across genomic regulatory regions (y-axis). Odds ratios (OR) and 95% confidence intervals (x-axis) shown for each genomic region. Dashed red line indicates OR=1 or no enrichment.

Supplementary tables

Supplementary Table 1 – decode summary statistics

Summary statistics of all 22,178,471 autosomal CpGs showing whether they are high-quality and if not, why they failed.

Chrom: chromosome

Start: start coordinate

End: end coordinate

Ratio: Average methylation levels

Passed_strandBias: binary column indicating whether the CpG passed strand bias filter

Passed_fractionReliable: binary column indicating whether the CpG passed fraction of reliable reads filter

Passed_coverage: binary column indicating whether the CpG passed coverage filter

Supplementary Table 2 – deCODE summary statistics

Summary statistics from associations for all 3,739,876autosomal CpGs that have Bonferroni significant age-association.

Chrom: chromosome

Start: start coordinate

End: end coordinate

Intercept: intercept calculated from linear regression model

Effect_size: effect size calculated from linear regression model

Std_err: standard error calculated from linear regression model

T_val: standard error calculated from linear regression model

P_val: p-value calculated from linear regression model

Supplementary Table 3 – deCODE summary statistics

Summary statistics from methylation aging clock, showing intercept and coefficients for 1,373 CpGs used in the model.

Attribute: Name of the attribute in model, either intercept or CpG name on the format chrom_start_end

Coefficient: Value of coefficient or intercept.

Supplementary Table 4 – deCODE summary statistics

Summary statistics for all 1,274,315 autosomal CpGs that have Bonferroni significant age-association on one of the parental haplotypes.

Chrom: chromosome

Start: start coordinate

End: end coordinate

Mean_M: mean methylation levels of maternal haplotype

Mean_P: mean methylation levels of paternal haplotype

Coef_M: effect size for maternal haplotype calculated from linear regression model

Coef_P: effect size for paternal haplotype calculated from linear regression model

P_val_M: p-value for maternal haplotype calculated from linear regression model

P_val_P: p-value for paternal haplotype calculated from linear regression mode

Stderr_M: standard error for maternal haplotype calculated from linear regression model

Stderr_P: standard error for paternal haplotype calculated from linear regression model

Supplementary Table 5 – deCODE summary statistics

Summary statistics for 702 CpGs with significant difference between the effect sizes calculated for maternal- and paternal haplotypes.

Chrom: chromosome

Start: start coordinate

End: end coordinate

Mean_M: mean methylation levels of maternal haplotype

Mean_P: mean methylation levels of paternal haplotype

Coef_M: effect size for maternal haplotype calculated from linear regression model

Coef_P: effect size for paternal haplotype calculated from linear regression model

P_val_M: p-value for maternal haplotype calculated from linear regression model

P_val_P: p-value for paternal haplotype calculated from linear regression mode

Stderr_M: standard error for maternal haplotype calculated from linear regression model

Stderr_P: standard error for paternal haplotype calculated from linear regression model

Cpgs: CpG name

Z: z-value for effect size comparison

P: p-value for effect size comparison

Hap: Significant haplotype

Unmet: unmethylated haplotype

Allele: Methylation status of allele

Met_dif: Absolute difference between mean_M and mean_P

Est_dif: Absolute difference between coef_M and coef_P

Supplementary Table 6. Number of Bonferroni- ($p<2.8\cdot10^{-9}$) and nominally significant ($p<0.05$) CpGs in original analysis and three rounds of permuted analysis for **(A)** combined haplotypes model, and **(B)** parent-of-origin specific model.

A) Combined haplotypes model

	Bonferroni significant CpGs, $p<2.8\cdot10^{-9}$ (%)	**p-value<$1\cdot10^{-6}$ (%)**	**p-value<0.001 (%)**	**Nominally significant, p<0.05 (%)**
Overall	3,739,875 (20.8%)	4,972,493 (27.7%)	7,390,066 (41.1%)	10,368,881 (57.7%)
Permutation 1	0 (0%)	23 (0.0001%)	17,649 (0.1%)	896,087 (5.0%)
Permutation 2	0 (0%)	13 (0.00007%)	17,983 (0.1%)	895,223 (5.0%)
Permutation 3	0 (0%)	16 (0.00009%)	17,976 (0.1%)	897,699 (5.0%)

B) Parent-of-origin specific model

	Bonferroni significant CpGs, $p<2.8\cdot10^{-9}$ (%)	**p-value<$1\cdot10^{-6}$ (%)**	**p-value<0.001 (%)**	**Nominally significant, p<0.05 (%)**	**Bonferroni significant, $p<2.8\cdot10^{-9}$ Effect size difference, $p<3.9\cdot10^{-8}$ (%)**
Overall	1,274,315 (3.5%)	2,729,364 (7.6%)	5,877,004 (16.3%)	10,173,092 (28.3%)	702 (0.05%)
Permutation 1	5 (0.00001%)	15 (0.00004%)	35,492 (0.01%)	1,731,361 (4.8%)	2 (0.3%)
Permutation 2	9 (0.00003%)	15 (0.00004%)	35,434 (0.01%)	1,728,349 (4.8%)	2 (0.3%)
Permutation 3	8 (0.00002%)	17 (0.00005%)	35,465 (0.01%)	1,729,939 (4.8%)	3 (0.4%)