Australian National University

Targeted nanopore sequencing for forensic SNP genotyping and CpG methylation profiling

Zaka Wing-Sze Yuen

BAppScFor (Hons)

The John Curtin School of Medical Research

This thesis submitted for the degree of Doctor of Philosophy of
The Australian National University

Declaration

I, Zaka Wing-Sze Yuen, certify that:

- This thesis is composed of my original work since the commencement of my higher degree by research candidature, except where work which has formed part of jointly authored publications has been included.
- This thesis contains no material previously published or written by other persons, except where due reference has been made in the text.
- This thesis does not include any material that has been submitted for the award of any other degree or diploma in my name, in any university or other tertiary institution.

Zaka Wing-Sze Yuen

March 2023

Acknowledgements

This PhD journey was long and challenging, and there were many ups and downs in these four years.

First, I would like to thank my supervisors, Prof. Eduardo Eyras and Cameron Jack. I have known Cam before studying at ANU because he was one of my supervisors for my honours project at UC. Because of Cam, I started to code and learn more about bioinformatics, and I found myself interested in it and wanted to pursue a PhD/career in bioinformatics. After I finished my honours in 2019, Cam put me in contact with Eduardo who was about to join JCSMR at that time. Eduardo kindly offered to be my primary supervisor; since then, he has continuously encouraged and guided me. I feel lucky and proud to be a part of Eyras lab, and I am so grateful to have Eduardo and Cam as my supervisors. They have pushed me to succeed and were supportive at those inevitable times when I was lost. Working under your supervision has been absolutely enjoyable, and I have learned and grown a lot. Thanks, Eduardo and Cam!

Next, I would like to thank Prof. Dennis McNevin and Dr. Runa Daniel for contributing to my PhD project. I sincerely appreciate how you two have continuously encouraged and guided me from my honours to my PhD. You have always been so friendly and supportive, and I was happy to see you in person finally at the ANZFSS conference.

I am also grateful beyond words to my dearest "family" in JCSMR: Marina, Adrià, Emiliana, Alex, Attila, Akanksha, Pablo, Tim, Lex, Ann. Thank you all for the support, encouragement, love, laughter, and sweet and unforgettable memories during the journey. You guys always have my back and help lift my spirits whenever I feel down. Special thanks Akanksha for helping me with the consensus model on my first paper.

I gratefully acknowledge the assistance of Carolina from Biomolecular Resource Facility (BRF) and Niccy from Ecogenomics and Bioinformatics Lab (EBL) in library preparation and nanopore sequencing. Thanks also to Nadine for sample preparation, as well as to Simon, Soma and Maurice for the help with the blood samples.

Thanks Marcin and Benjamin for giving me the opportunity to be a demonstrator for your course, which allows me to gain work experience in teaching.

Many thanks to my friends outside JCSMR in Australia: Fanfan, Anna, Kim, for always being there when I needed them.

Thanks to my partner, Sarah for all her love and support. I could not have come this far without her. No words can express my gratitude for everything you have done for me.

Lastly, I wanted to sincerely acknowledge my family and friends from my hometown – Hong Kong, so I wrote an acknowledgment for them in my mother tongue, Cantonese.

以下落嚟係我特別想鳴謝嘅人排名不分先後：

多謝爹哋媽咪養育之恩。佢哋係非常開明嘅父母，從來都唔俾壓力我，比我自由發展

所長，成為我刻苦嘅嘅留學生捱嘅動力。今日我學有所成，成為澳洲國立大學嘅博士

畢業生，成為科學家，佢哋功不可沒。

我特別要多謝我姑姐，佢係我從小到大心目中嘅偶像，無可取替嘅榜樣。非常感謝姑姐無償供書教學，提供無限資源栽培我，我今日終於可以好似姑姐一樣，成為一個獨立自主有學識嘅人。

我好多謝我細妹美美，佢係我最好嘅聆聽者，亦係我心中最成熟最生性最痛錫我嘅細妹。希望以後你都會做我第一個聆聽者成為我最大嘅後盾，多謝你美美！

最後，好多謝我嘅好朋友！Fatlum同Chankei成日聽我呻，我嘅事姐系佢地嘅事，我成日都喺最迷惘嘅時候，你地就鞭策我；Lily永遠都好快覆我信息同聽我電話，成日都買嘢送俾我，你買俾我嘅袋仲用到而家;Micmic成日去旅行寄啲得意嘢俾我好warm，次次翻HK你都陪我行街食好野；Carmen,細細同Waiying每次都送我機陪住系我身邊，其實好遺憾我哋冇去嘅台灣畢業旅行，希望大家40歲前我哋四個真係可以去到！

This thesis is dedicated with love and the deepest gratitude

To my parents;

To my aunt.

I hope I have made you proud.

Abstract

Forensic DNA phenotyping can provide intelligence leads by inferring externally visible characteristics, biogeographical ancestry, and the chronological age of an unidentified subject of an investigation. Current analyses are based on the detection of informative single nucleotide polymorphisms (SNPs) and DNA methylation markers, using a broad spectrum of laboratory techniques and detection technologies. Currently, these SNP and DNA methylation markers must be detected in separate assays using specific primers for every single marker. Oxford Nanopore Technologies (ONT) sequencing platforms offer the advantage of detecting genetic and epigenetic markers simultaneously using biochemical-free targeted approaches, without the need for conventional laboratory techniques such as polymerase chain reaction (PCR). While base modifications can be revealed by the current signals from nanopore sequencing, the prediction accuracy of current nanopore-based methylation calling tools is still unclear, and no systematic comparison of these tools has been performed to date. The aim of this study is to evaluate nanopore sequencing applied to the field of forensic DNA phenotyping, specifically SNP genotyping and DNA methylation profiling.

Firstly, a pipeline called METEORE is presented that provides a standardized workflow for methylation detection using six different nanopore-based methylation calling tools. The per-site and per-read prediction performance of these tools was benchmarked and different

strategies to improve prediction accuracy are described. Among six individual tools, Megalodon achieved the best performance overall.

Two PCR-free targeted nanopore sequencing approaches – Cas9 enrichment and adaptive sampling were then validated. Cas9-based targeted nanopore sequencing (nCATS) involved sequencing of ten forensically relevant regions and evaluate the six tested tools further using public bisulfite sequencing dataset. Adaptive sampling involves in-silico enrichment of regions of interest (ROI) during sequencing, recovering key methylation information from the targeted loci in a cost-effective, simple, and direct way. Although adaptive sampling resulted in lower coverage than the biochemical methods, it was nevertheless accurate for SNP genotyping and methylation profiling. This study represents a proof of concept for the potential of nanopore sequencing for forensic applications. Nanopore sequencing is expected to become the new gold standard for methylation detection as it not only correlates with sequencing of bisulfite-treated DNA, but it also detects methylation without the need for high coverage.

Despite its great potential, the technology requires key improvements for its implementation in routine forensic workflows, especially to work with low-input or mixed DNA samples. The thesis concludes with a discussion of the trends in ongoing improvements to both bioinformatic tools and library preparation methods, which are rapidly making new techniques and new opportunities available for forensic operation.

List of abbreviations

A	Adenine
AISNP	Ancestry-informative SNP
API	Application programming interface
APM	Age prediction model
ASIC	Application-specific integrated circuit
AUC	Area under the ROC Curve
BGA	Biogeographical ancestry
C	Cytosine
CE	Capillary electrophoresis
CGI	CpG island
CNN	Convolution neural network
CPU	Central processing unit
CRISPR	Clustered regularly interspaced short palindromic repeats
crRNA	CRISPR RNA
dATP	Deoxyadenosine triphosphate
ddNTP	Dideoxynucleotide triphosphate
DMR	Differentially methylated region
DNA	Deoxyribonucleic acid
dsDNA	double-stranded DNA
DVI	Disaster victim identification
ECDF	Empirical cumulative distribution function

ENA	European Nucleotide Archive
ENCODE	Encyclopedia of DNA Elements
EVC	externally visible characteristics
FDP	Forensic DNA phenotyping
FDR	False discovery rate
FM index	Ferragina-Manzini index
FN	False negative
FNR	False negative rate
FP	False positive
FPR	False positive rate
G	Guanine
GIAB	Genome in a Bottle Consortium
gRNA	Guide RNA
GPU	Graphics processing unit
HMM	Hidden Markov model
IGS	Intergenic spacer
IGV	Integrative Genomics Viewer
LLR	Log-likelihood ratio
LOESS	Locally weighted scatterplot smoothing
MAE	Mean absolute error
MBD	Methyl-CpG-binding domain
MPS	Massively parallel sequencing
MRE	Methylation-sensitive restriction enzymes
ONT	Oxford Nanopore Technologies
PAM	Protospacer-adjacent motif
PCR	Polymerase chain reaction
PISNP	Phenotype-informative SNP
PR	Precision-recall
REG	Regression
RF	Random forest
RMSE	Root mean square error

RNA	Ribonucleic acid
RNN	Recurrent neural network
RNP	Ribonucleoprotein
ROC	Receiver operating characteristic
ROI	Regions of interest
RRBS	Reduced-representation bisulfite sequencing
SBE	Single base extension
SFM	Short fragment mode
SNP	Single nucleotide polymorphism
STR	Short tandem repeat
T	Thymine
TFS	Thermo Fisher Scientific
TN	True negative
TP	True positive
TPR	True positive rate
tracrRNA	Transactivating CRISPR RNA
TSV	Tab-separated values
U	Uracil
USB	Universal serial bus
WGBS	Whole genome bisulfite sequencing

List of publications

The following publication has derived from the work contained in this thesis:

Yuen ZWS, Srivastava A, Daniel R, McNevin D, Jack C, Eyras E (2021). Systematic benchmarking of tools for CpG methylation detection from nanopore sequencing. Nature Communications, 12(1), 3438. doi:10.1038/s41467-021-23778-6

List of figures

List of tables

Table of Contents

CHAPTER I: INTRODUCTION

1 DNA methylation

DNA can be modified through the addition of chemical groups to nucleotides. These modifications play a fundamental role in the regulation of genome stability and gene expression during mammalian development, disease progression and aging (Greenberg & Bourc'his, 2019; Jones *et al.*, 2015; Jones, 2012; Kader & Ghai, 2015; Raiber *et al.*, 2017). Of more than 17 possible DNA modifications, 5-methylcytosine (5mC), also known as the fifth base of DNA, is the most frequently observed in relation to gene regulation (Raiber *et al.*, 2017). 5mC is formed by the addition of a methyl group ($-CH_3$) to the 5^{th} carbon of the cytosine (C) ring and predominantly occurs in the CG context (CpG). CpG islands (CGIs) are CpG-rich regions found in many promoters known to control gene expression. Gene silencing is often associated with highly methylated CGIs (Deaton & Bird, 2011). Conversely, highly expressed genes are often associated with unmethylated or slightly methylated CGIs (Deaton & Bird, 2011). Many other DNA modifications have been found in bacterial and eukaryotic genomes (Raiber *et al.*, 2017), such as 5-hydroxymethylcytosine (5hmC), 5-formyluridine (5fU) and N6-methyladenine (6mA), but 5mC in the CpG context is the most studied and most common among all these DNA modifications nevertheless (Kumar *et al.*, 2018) and is the main focus of this thesis.

1.1 Experimental methods for genome-wide profiling of DNA methylation

Key advances in the understanding of the function of 5mC have been made possible through the development of dedicated genome-wide profiling techniques. Commonly used methods include restriction enzyme digestion, affinity enrichment, or bisulfite conversion, followed by microarray hybridisation or short-read sequencing (Figure 1) (Yong *et al.*, 2016).

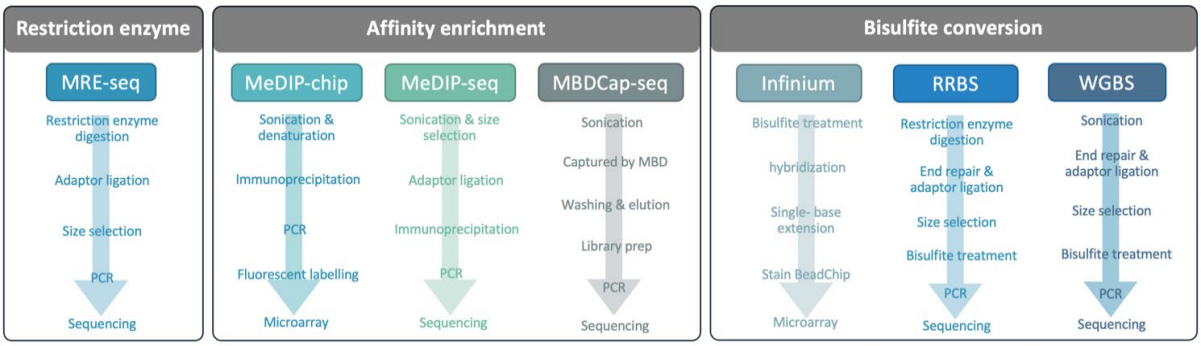


Figure 1. Overview of three major experimental method for genome-wide DNA methylation profiling methods.
(from left to right) restriction enzyme digestion-based, affinity enrichment-based and bisulfite conversion-based. MRE-seq: Methylation-sensitive restriction enzyme sequencing. MeDIP-chip: Methylated DNA immunoprecipitation microarray. MeDIP-seq: Methylated DNA immunoprecipitation sequencing. MBDCap-seq: Methyl-CpG binding domain-based capture and sequencing. RRBS: Reduced representation bisulfite sequencing. WGBS: Whole genome bisulfite sequencing. PCR: polymerase chain reaction.

Methylation-sensitive restriction enzymes (MREs) are used for restriction enzyme-based methods to cut unmethylated target sequences while leaving the methylated DNA intact. Hence, DNA methylation levels can be predicted from the enzymatic digestion patterns coupled with high-throughput sequencing technologies (MRE-seq). Although the MRE digestion methods are cost-effective, they are limited by restriction enzyme recognition sites and prone to false-positive results due to incomplete cleavage (Yong *et al.*, 2016).

Alternatively, affinity enrichment-based methods utilise either methyl-CpG-binding domain (MBD) proteins or methylation-specific antibodies to enrich methylated DNA regions. The

former involves the capture of methylated DNA sequences with the MBD proteins and subsequent sequencing of eluted DNA (MBDCap-seq) (Brinkman *et al.*, 2010). The latter uses a method called methylated DNA immunoprecipitation (MeDIP) (Weber *et al.*, 2005), which utilises an antibody to immunoprecipitate DNA with methylated CpG sites, and the purified methylated DNA can be subsequently detected by either microarray (MeDIP-chip) or high-throughput sequencing (MeDIP-seq). These methods are cost-effective and require only minimal DNA sample input amounts (as low as 1 ng) or even fragmented DNA as starting material (Brinkman *et al.*, 2010; Weber *et al.*, 2005; Yong *et al.*, 2016; Zhao *et al.*, 2014). However, the major limitations are low sensitivity in regions with high CpG density and lack of single-base resolution (Yong *et al.*, 2016).

Bisulfite genomic sequencing invented by Frommer *et al.* (1992) represents a major advance in DNA methylation detection because it provides a qualitative and quantitative approach to determine 5mC at single base resolution. This method consists of a bisulfite conversion step that treats denatured DNA with sodium bisulfite, which converts unmodified cytosine (C) to uracil (U) while leaving 5mC unchanged, providing a means to discriminate between C and 5mC. A subsequent amplification by polymerase chain reaction (PCR) is carried out to replace uracil with thymine. The bisulfite-treated DNA can then be subjected to microarray hybridisation or sequencing for detection.

In terms of detection technology, Illumina's Infinium HumanMethylation450 (450K) BeadChip microarray and whole-genome bisulfite sequencing (WGBS) are the most commonly used bisulfite conversion-based method to study genome-wide DNA methylation (Jones, 2012; Yong *et al.*, 2016). The former covers more than 450,000 CpG sites and produces data storing fluorescence intensities that quantify the relative abundance of methylated and unmethylated sites, and the latter is based on high-throughput short-read sequencing and

requires downstream analysis of the sequencing data, including alignment and methylation calling. Although both can provide single-base resolution, WGBS is considered the gold standard and has advantages over the methylation array as it can determine absolute methylation levels of nearly all CpG sites in the sample, regardless of CpG density, and provide sequence context information (i.e. CG, CHG and CHH) (Singer, 2019; Yong *et al.*, 2016), computational analysis of WGBS data is challenging and consists of multiple steps, as described in the following section.

Other bisulfite sequencing approaches, such as reduced-representation bisulfite sequencing (RRBS) and pyrosequencing, are also available. RRBS involves a size selection step using site specific CpG restriction enzymes such as MspI prior to bisulfite conversion, enabling profiling of CpG-rich regions of a genome like CGIs at a lower cost (Meissner *et al.*, 2005; Singer, 2019; Yong *et al.*, 2016). On the other hand, pyrosequencing allows locus-specific DNA methylation profiling and provides higher coverage for the target loci (Singer, 2019).

1.2 Computational methods for detecting DNA methylation

WGBS is widely used for genome-wide quantification of DNA methylation at single-base resolution. The entire workflow for WGBS analysis involves multiple steps including preprocessing, alignment of reads to the reference genome, and methylation calling, with the use of different computational tools, as described below:

The first step in WGBS data analysis is preprocessing, which involves adapter removal and read filtering to eliminate adapter bases and low-quality bases. The quality of the raw sequence data is first assessed with tools such as FastQC (Andrews, 2010), and then the reads are subjected to read quality trimming using tools such as Trimmomatic (Bolger *et al.*, 2014).

Next, the filtered reads are aligned to the reference genome. This step poses a big challenge due to decreased sequence complexity and non-complementarity in the Watson and Crick strands resulting from the bisulfite conversion of unmethylated cytosine to thymine without affecting the other three canonical bases or methylated cytosine (Xi & Li, 2009). Here, two groups of algorithms can be employed to map bisulfite-converted reads to the reference: wild-card and three-letter aligners (Bock, 2012). Wild-card aligners such as BSMAP (Xi & Li, 2009) substitute Cs with Ys in the reference genome and thus both Cs and Ts in the sequencing reads can be aligned to Ys in the reference, while other wild-card aligners such as Last (Frith *et al.*, 2012) use a modified score matrix that allows mismatches between Ts in the read and Cs in the reference (Bock, 2012; Tsuji & Weng, 2015). Alternatively, three-letter aligners such as BS-Seeker2 (Guo *et al.*, 2013) and Bismark (Krueger & Andrews, 2011) convert all Cs into Ts on both sequencing reads and reference genome in order to perform subsequent alignment of three nucleotides (A, G and T) using conventional short-read aligners such as Bowtie (Langmead *et al.*, 2009) and BWA (Bock, 2012; Li & Durbin, 2010; Tsuji & Weng, 2015).

After sequence alignment, PCR duplicates are discarded to prevent false positive signals in downstream analysis. Methylation calling is then carried out using the corresponding bisulfite aligners as mentioned before. Following methylation calling, a quality control still needs to be performed to minimise any methylation biases rising from inadequate adapter trimming, incomplete C to T conversion or DNA degradation. This can usually be done by adding spike-in sequences with a known sequence containing unmethylated Cs to calculate the conversion rate. The alignment data can also be visualized with tools such as UCSC genome browser (Fujita *et al.*, 2010) and Integrative Genomics Viewer (IGV) (Robinson *et al.*, 2011; Thorvaldsdóttir *et al.*, 2013). Apart from data visualization, the average methylation level and genome-wide methylation level can be calculated and plotted, and differentially methylated regions (DMRs) can be identified.

2 Forensic DNA phenotyping

Short tandem repeat (STR) analysis has been the gold standard for human identification in forensic analysis including criminal casework and paternity analysis because STRs are highly polymorphic genetic markers (Parson *et al.*, 2016). STRs, also commonly known as microsatellites, are repetitive nucleotide sequences of 2–10 base pairs (bps). Due to their high discrimination power, STRs can be used to differentiate between individuals by comparing the STR profile of the DNA sample left at a crime scene with either the DNA from a suspect or the STR profiles from a reference DNA database (Budowle *et al.*, 2017; Montesanto *et al.*, 2020). The analysis method is based on multiplex PCR to amplify the STR loci using primers flanking the repeat region, followed by capillary electrophoresis (CE) with an internal size standard to separate DNA fragments by their length (Budowle *et al.*, 2017; Parson *et al.*, 2016). The alleles for each STR locus are then determined by comparing the DNA fragments lengths with an "allelic ladder" which is a standard mixture of fragments representing most known STR alleles. The STR profile of an individual is presented in an electropherogram which can be converted to a text-based table.

Single nucleotide polymorphisms (SNPs) are alternative genetic markers that can not only supplement STRs for human identification but also be applied to forensic DNA phenotyping (FDP) (Mehta *et al.*, 2017). SNP markers have some potential advantages over STRs including lower mutation rates and shorter amplicon lengths that are more applicable to degraded DNA

samples. Comparative DNA profiling is deemed inefficient when there are no or many suspects or no matching profiles within a reference DNA database (Daniel *et al.*, 2015). To overcome this, FDP can provide intelligence leads to assist in an investigation and subsequently in identifing the unknown perpetrator by using ancestry-informative SNPs (AISNPs), phenotype-informative SNPs (PISNPs) and DNA methylation markers to infer respectively the biogeographical ancestry (BGA), externally visible characteristics (EVCs) and the chronological age of a donor from the DNA sample found at crime scenes (Mehta *et al.*, 2017).

2.1 SNaPshot assays for SNP genotyping

SNaPshot®, a single base extension (SBE) multiplex assay coupled with CE-based detection, has been widely used to genotype SNPs in forensic DNA analysis (Daniel *et al.*, 2015; Mehta *et al.*, 2017). After an initial multiplex PCR, SBE primers hybridise to the single-stranded PCR product and are extended with a fluorescently labelled dideoxynucleotide triphosphate (ddNTP) chain terminator complementary to the nucleotide at the SNP locus of interest, followed by the detection of the extended primers by CE (Børsting & Morling, 2015). Various AISNP and PISNP SNaPshot assays have been developed, for instance, the SNP*for*ID 34-plex for differentiation of European, Asian and African populations (Phillips *et al.*, 2007), Pacifiplex for differentiation of Oceanian populations (Santos *et al.*, 2016) and HIrisPlex for hair and eye colour prediction (Walsh *et al.*, 2013). While a single SNaPshot assay (i.e. PCR-SBE-CE) can only genotype 30-40 SNPs, massively parallel sequencing (MPS) offers potential as an alternative SNP genotyping method that sequences a large pool of amplicons from the existing SNaPshot multiplexes and detects a large number of SNP loci with single base pair (bp) resolution in a single run (Budowle *et al.*, 2017; Daniel *et al.*, 2015; Mehta *et al.*, 2016). It has been demonstrated that AISNP and PISNP amplicons from SNaPshot assays can be combined

and genotyped on medium-throughput benchtop MPS platforms such as the Ion GeneStudio S5™ (Thermo Fisher Scientific) (TFS) (Phillips *et al.*, 2019), MiSeq (Verogen/Illumina) (Mehta *et al.*, 2016) and MiSeq FGx (Verogen/Illumina) (Xavier *et al.*, 2022).

MPS may have the potential to replace CE and become the standard sequencing technology for forensic genotyping of SNPs and STRs. One of the advantages of MPS is that it can detect every nucleotide in a sequence and therefore enables the analysis of variants with single base pair resolution (Børsting & Morling, 2015; Daniel *et al.*, 2015; Phillips *et al.*, 2019). It also allows the investigation of a larger number of loci in a single assay, thereby decreasing the usage of evidential DNA samples, especially those with low quantity or quality (Budowle *et al.*, 2017). However, MPS has not been considered as a replacement for CE in short to medium term. One of the primary factors behind the lack of widespread implementation of MPS by forensic laboratories is its higher cost relative to CE (Plesivkova *et al.*, 2019). Currently, it is an additional genotyping method which is mostly used to genotype forensically relevant SNPs while STR genotyping using MPS is validated.

Semiconductor sequencing and sequencing by reversible termination are employed in the Ion Torrent™ and Verogen/Illumina platforms, respectively. Semiconductor sequencing resembles pyrosequencing, but instead of using the optical detection of fluorescently labelled nucleotides, it detects proton release during nucleotide incorporation using an ion sensor (Van Dijk *et al.*, 2014). On the other hand, Illumina sequencing (sequencing by synthesis) utilises fluorescently labeled nucleotides with reversible 3' terminators where each nucleotide incorporation is detected by a camera, followed by the removal of the terminator and the fluorescent dyes, allowing the next nucleotide to be incorporated (Børsting & Morling, 2015; Van Dijk *et al.*, 2014).

2.2 Forensic DNA methylation profiling

DNA methylation has been identified as one of the most promising biomarkers for age prediction and body fluid identification in forensic applications (Dias *et al.*, 2020; Freire-Aradas *et al.*, 2020; Jones *et al.*, 2015; Lee *et al.*, 2022; Montesanto *et al.*, 2020; Woźniak *et al.*, 2021). It offers potential as part of FDP to provide intelligence by inferring the age of a person to aid an investigation. Age prediction helps characterize unknown perpetrators and re-prioritise a suspect pool in addition to forensically validated inference of eye and hair colour and BGA (Vidaki & Kayser, 2018; Woźniak *et al.*, 2021; Xavier *et al.*, 2022). Some EVCs such as hair loss, hair greying, and facial wrinkles are age-dependent and have been the focus of current research efforts (Vidaki & Kayser, 2018; Woźniak *et al.*, 2021).

Numerous age prediction models (APMs) have been developed to determine the biological age of individuals (as a proxy for chronological age) by DNA methylation and are based on either hundreds of CpG sites using genome-wide BeadChip microarray technology (Hannum *et al.*, 2013; Horvath, 2013; Levine *et al.*, 2018) or small sets of biomarkers containing age-associated CpG sites using a broad spectrum of technologies with a targeted approach such as Sanger sequencing (Dias *et al.*, 2020), pyrosequencing (Bekaert *et al.*, 2015; Feng *et al.*, 2018; Fleckhaus & Schneider, 2020; Li *et al.*, 2020; Park *et al.*, 2016; Thong *et al.*, 2017; Weidner *et al.*, 2014; Zbieć-Piekarska *et al.*, 2015a; Zbieć-Piekarska *et al.*, 2015b), SNaPshot (Hong *et al.*, 2017; Jung *et al.*, 2019; Lee *et al.*, 2022), MPS (Aliferi *et al.*, 2018; Naue *et al.*, 2017; Naue *et al.*, 2018; Vidaki & Kayser, 2018; Woźniak *et al.*, 2021) and EpiTYPER mass spectrometry (Freire-Aradas *et al.*, 2018; Freire-Aradas *et al.*, 2016). Most of these studies incorporate only a few CpG sites into the model and work with 20 ng DNA input for a robust methylation quantification. These specific CpG sites are consistently related to age across individuals and used to predict age.

3 Limitations of the current technologies

SNaPshot (SBE-PCR-CE) is favoured in the forensic workflow as it not only genotypes SNPs from low input amounts of DNA but also generates DNA methylation profiles at specific CpG sites from the bisulfite-converted DNA (Daniel *et al.*, 2015; Lee *et al.*, 2022; Phillips *et al.*, 2009). However, the conventional SNaPshot system only allows up to 30-40 SNPs per assay, and the ability to provide accurate BGA and EVC intelligence to forensic investigators requires the analysis of more SNPs than can be genotyped from a single SNaPshot assay (Daniel *et al.*, 2015; Xavier *et al.*, 2020). On the other hand, short-read sequencing can be used to simultaneously genotype hundreds of SNPs from multiple SNaPshot assays (SBE-PCR-MPS). MPS relies on PCR to amplify the target loci defined by the primers before sequencing, which requires specialized reagents and consists of multiple steps in the sequencing workflow, but most limiting for DNA methylation profiling, base modifications (e.g. DNA methylation) are lost during amplification.

Furthermore, most DNA methylation profiling methods mentioned above are based on bisulfite conversion of the genomic DNA (gDNA). The significant drawbacks of bisulfite conversion-based methods are substantial DNA degradation, incomplete conversion, reduced sequence variability and the necessary coupling with lengthy protocols (Freire-Aradas *et al.*, 2020; Grunau *et al.*, 2001; Yong *et al.*, 2016) . The conversion of 5mC to uracil is also very sensitive

to the reaction conditions (Ehrich *et al.*, 2007). More importantly, DNA methylation quantification is influenced by the amount of bisulfite-converted DNA used for PCR and the systematic errors associated with the types of genetic analyser used in CE analysis from the SNaPshot assays (Freire-Aradas *et al.*, 2020; Lee *et al.*, 2022). While short-read sequencing-based methylation profiling methods such as WGBS and RRBS have been very effective at mapping 5mC sites at the genome-scale, they still presents various disadvantages including high mapping uncertainty in repetitive regions and amplification bias (Ni *et al.*, 2021).

Last but not least, although many AISNPs, PISNPs and CpG sites have been identified and applied for ancestry analysis, prediction of human eye/hair/skin colour and age prediction/body fluid identification, respectively, these forensic markers are detected separately in each FDP assay, especially separating SNP markers from DNA methylation markers, which increases the number of experiments to be performed.

4 Nanopore sequencing

4.1 Overview of nanopore sequencing

Oxford Nanopore Technologies (ONT) has developed a nanopore-based sequencing platform for long-read DNA/RNA sequencing. Protein nanopores are embedded within an electrically resistant membrane, and an ionic current is passed through the nanopores by applying a voltage across the membrane. When a single-stranded nucleic acid passes through the nanopore, the nucleotides within the pore affect the electrical resistance, which causes disruptions in the electrical current. The small consumable flow cell used in nanopore sequencing contains an arrayed sensor chip for measuring the current and an Application-Specific Integrated Circuit (ASIC) for signal processing. The resulting current signals known as "squiggles' are recorded as raw signal data by the nanopore and can be translated to nucleotide sequence by a computational process called basecalling (Wick *et al.*, 2019). It is a computationally intensive step and requires GPUs.

This long-read sequencing technology can produce reads of varying lengths, from a few hundred bases to over 10 kilobase pairs (kb), depending on the fragment length of the input sample, whereas short-read sequencing technology such as sequencing by synthesis (Illumina) is limited to hundreds of bases (Amarasinghe *et al.*, 2020). While the basecalling accuracy has

improved dramatically over the past few years (Amarasinghe *et al.*, 2020; Wick *et al.*, 2019) nanopore sequencing is still error-prone when sequencing homopolymer regions (Amarasinghe *et al.*, 2020; Cornelis *et al.*, 2017; Tytgat *et al.*, 2020; Tytgat *et al.*, 2022). As translocation of homopolymers does not change the k-mers of nucleotides residing in the pore, this event generates no change in the signal, making resolving homopolymers difficult (Amarasinghe *et al.*, 2020).

Guppy has been utilized for numerous years and integrated into the MinKNOW software for real-time basecalling. Recently, ONT has developed new basecalling tools, namely Dorado (Oxford Nanopore Technologies, 2023b) and Bonito (Oxford Nanopore Technologies, 2023a), which are anticipated to replace Guppy in the near future. Currently, Dorado exists as a standalone tool but will soon be incorporated into MinKNOW. On the other hand, Bonito is an open-source research basecaller that is actively being developed for research purposes.

A recent study by Dittforth *et al.* (2023) compared the basecalling performance of Guppy and Dorado on 20 different Amazon Elastic Compute Cloud (Amazon EC2) instance types equipped with GPU accelerators. The study's findings indicate that Dorado surpasses Guppy in terms of basecalling and methylation calling, demonstrating superior performance. The results of the study revealed that Dorado demonstrates superior performance over Guppy in both basecalling and methylation calling.

4.2 The MinION nanopore sequencer

ONT has released several sequencers in different sizes and formats, such as the portable pocket-sized MinION and the benchtop PromethION for large-scale projects. The MinION is a lightweight, cheap, and USB-powered sequencer that runs on a desktop or laptop computer. Taking advantage of its portability and real-time sequencing ability, reads generated from the sequencer can be basecalled in real time, aligned and analysed as soon as the run starts, thus it

has been applied in field research and clinical diagnosis (Burton *et al.*, 2020; Faria *et al.*, 2016; Quick *et al.*, 2017; Quick *et al.*, 2016; Sim & Chapman, 2019; Truelove *et al.*, 2019), showing that sequencing is not constrained to a traditional laboratory setting. For instance, the MinION was used in a mobile lab to conduct genomic surveillance of Zika virus (Faria *et al.*, 2016; Quick *et al.*, 2017), and it was used to conduct the first ever DNA sequencing in space (Burton *et al.*, 2020). The MinION flowcell has 512 pores for sequencing, and the throughput per flow cell is 15–30 Gb based on the manufacturer's data. MinKNOW is the software on the host computer for controlling the MinION device, collecting sequencing data in real time and processing the raw signal into reads.

4.3 Modified base detection from nanopore sequencing data

Nanopore long-read sequencing technologies have made also possible methylation profiling at a single-nucleotide and even single-cell resolution. By sequencing native DNA molecules with nanopore technology, modified bases such as 5mC can be detected without bisulfite treatment of the DNA sample and PCR amplification (Laszlo *et al.*, 2013; Simpson *et al.*, 2017). Since changes in electrical current signals depend on the sequence of the nucleotides within the nanopore, modified bases affect the current signals differently than the equivalent unmodified bases, and hence base modifications can be revealed by their unique signal shapes (Amarasinghe *et al.*, 2020; Laszlo *et al.*, 2013; Rand *et al.*, 2017; Simpson *et al.*, 2017). Still, signals are dependent on sequence context and different copies of the same molecule present considerable signal variation. Thus, it is necessary to apply computational models to interpret the signals and predict the site-specific methylation status.

There are various computational tools available for the detection of modified bases from nanopore reads. Some of the most commonly used tools are: Tombo, Nanopolish, DeepMod,

DeepSignal, Megaldon, and Guppy. Tombo implements an alternative base detection approach specific for CpG detection to detect targeted bases by statistically comparing signals against canonical and alternate 5mC at CG motifs (Stoiber *et al.*, 2017). Nanopolish detects CpG methylation with a hidden Markov model (HMM) (Simpson *et al.*, 2017). f5c, an optimised version of Nanopolish, allows faster processing in event alignment and methylation calling steps with the use of graphics processing unit (GPU) (Gamaarachchi *et al.*, 2020). DeepMod, DeepSignal, and DeepSignal-plant use neural networks to predict the methylation states (Liu *et al.*, 2019a; Ni *et al.*, 2021; Ni *et al.*, 2019). Megalodon (Oxford Nanopore Technologies, 2020b), which is a tool developed by ONT, extracts modified base calls from raw nanopore signals by anchoring the basecalling output produced by Guppy to a reference genome, using the modified base model included in Remora (Oxford Nanopore Technologies, 2022).

5 Potential of Nanopore sequencing for forensic applications

There are already some potential field-based DNA profiling systems such as RapidHIT ID (Applied Bioystems), ParaDNA (LGC Forensics) and ANDE Rapid DNA (ANDE) (Watherston *et al.*, 2022). Although these field-based systems are fast and the size of RapidHIT ID and ParaDNA is smaller than that of benchtop sequencers like the MiSeq FGx (Illumina/Verogen) and the Ion S5 (TFS), they are still heavy, not very portable and are very expensive per sample. The MinION, on the other hand, is small, palm-sized, weighs 87 g only and has a low entry cost (US $1000 from the starter pack including a MinION, two flow cells and one library preparation kit).

Given its portability, it has the potential for in-field forensic DNA analyses. For example, it could aid disaster victim identification (DVI) in remote locations (Bowman *et al.*, 2022; Watherston *et al.*, 2022) and provide DNA intelligence in covert operations (Gargis *et al.*, 2019), thereby reducing the turnaround time for forensic investigations. ONT has also developed the VolTRAX system for automated library preparation and MinIT device that is a companion to the MinION for real-time basecalling using a GPU. Otherwise, the MinION can be connected to a laptop computer with a GPU to support real-time basecalling. With the ongoing improvement to bioinformatic tools and future developments for more automated and

simplified library preparation, it offers more opportunities to forensic scientists for field-based analyses.

5.1 Targeted Nanopore sequencing

It is possible to sequence different types of forensic genetic and epigenetic markers, as well as a high number of forensic markers, at the same time with nanopore technology, generating data from the same sequencing run for forensic SNP/STR analysis and forensic DNA methylation profiling. The commonly used targeted genotyping approach employing PCR may not be compatible with nanopore sequencing because amplicons generated from PCR generally range from 75 to 1000 bp, plus amplification erases base modifications, which limits nanopore technology's ability to obtain long reads and detect methylation after PCR. With PCR-free nanopore targeted approaches, conventional laboratory techniques for methylation profiling including PCR and bisulfite treatment, are no longer required, which can reduce the cost of analysis, simplify the experimental protocols and minimize analysis time.

5.1.1 CRISPR/Cas9-based enrichment

One of the strategies to perform targeted nanopore sequencing uses clustered regularly interspaced short palindromic repeats (CRISPR), which is adapted from a naturally occurring genome editing system in bacteria. The genome editing properties of this complex have been adapted for targeted sequencing. The CRISPR/Cas family of proteins consists of DNA/RNA-cleaving enzymes that can be programmed to cleave specific sequences using RNA oligos. Cas9 is an endonuclease that cuts double-stranded DNA (dsDNA). It requires a CRISPR RNA (crRNA) to bind and cleave DNA at sites that are identical or highly similar to the crRNA

sequence. crRNA is a 36 nt long RNA that recognises a 20 mer target site called the protospacer, and the protospacer-adjacent motif (PAM) which is immediately next to the 3' end of the target on the opposite strand. The crRNA must be combined with the transactivating crRNA (tracrRNA) to form a guide RNA (gRNA) duplex, followed by the formation of a functional ribonucleoprotein (RNP) complex with Cas9. The complex melts dsDNA and searches for PAM sites. Once the crRNA successfully basepairs with the target sequence, Cas9 will then cut both strands of the target sequence three bp upstream of the PAM site.

One of the PCR-free targeted methods using nanopore sequencing is to utilise Cas9 to cut dsDNA and preferentially ligate nanopore sequencing adapters to the cleaved ends (Gilpatrick *et al.*, 2020) (Figure 2). In this Cas-mediated enrichment approach, DNA ends are first dephosphorylated to prevent subsequent off-target ligation, then the Cas9-gRNA complex is added to the sample to bind and cleave the target regions, introducing newly exposed ends with 5' phosphates. After dA-tailing, sequencing adapters are primarily ligated to the target DNA ends at Cas9 cleavage sites which are both 5' phosphorylated and 3' dA-tailed. Thus, non-target DNA is depleted by dephosphorylating pre-existing DNA ends while target DNA is enriched by selectively ligating adapters to the Cas9 cleavage sides. The native target DNA strands are then sequenced through the nanopores and therefore base modifications and fragment length can be preserved.

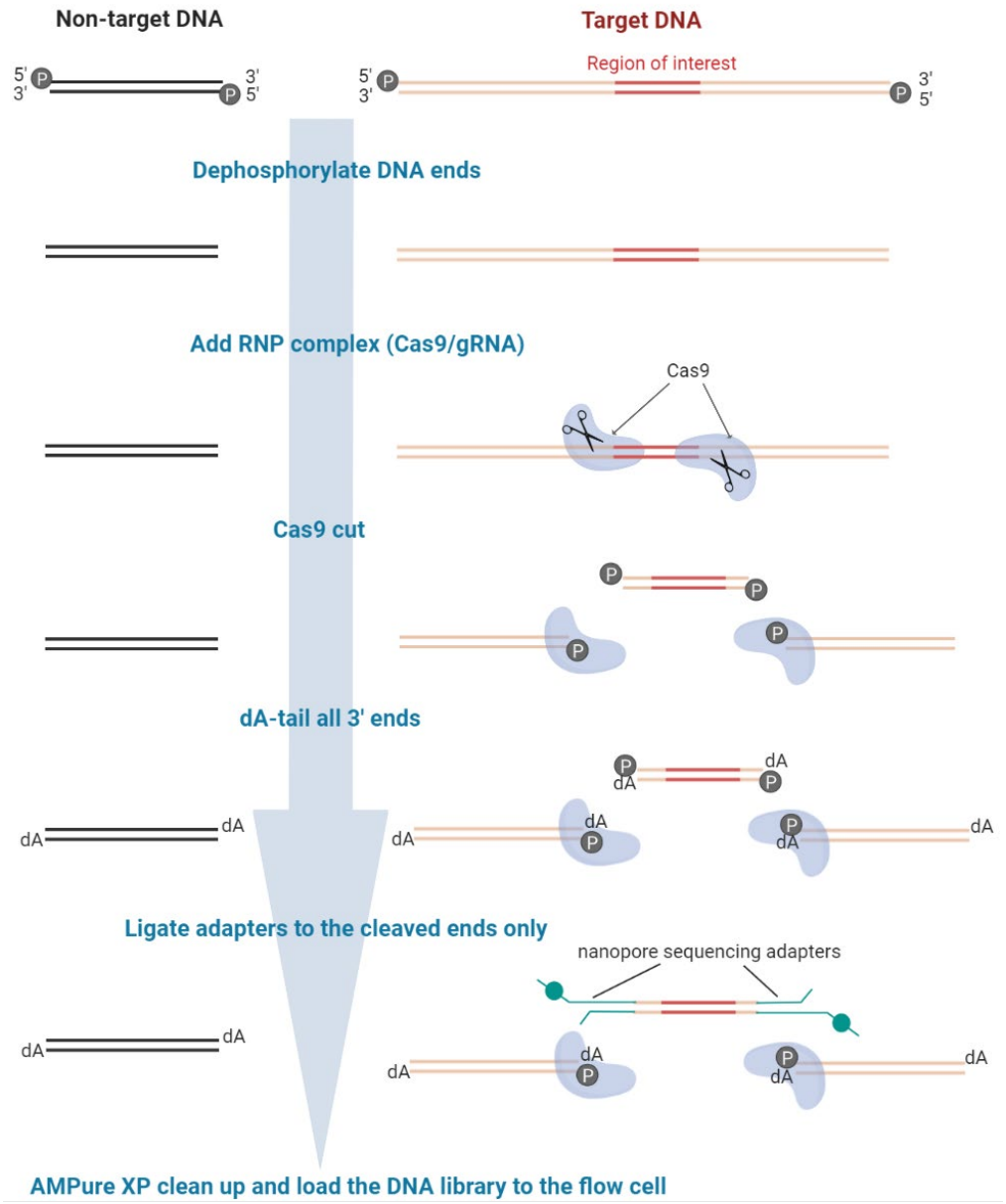


Figure 2. Workflow of the Cas9 enrichment and library preparation for nanopore sequencing.
The DNA ends of the sample are first dephosphorylated to prevent adapters from ligating to non-target DNA. The ribonucleoprotein (RNP) complex, which is formed by the combination of guide RNA (gRNA) and Cas9 protein, is then added to the sample. Cas9 enzyme then cleaves the region of interest (ROI) such that the cleaved DNA ends with ligatable 5' phosphates. After that, all 3' ends are dA-tailed and nanopore sequencing adapters are preferentially ligated to the cleaved ends, followed by clean-up using AMPure XP beads. The resulting DNA library is loaded to the flow cell. Created with BioRender.com.

5.1.2 Adaptive sampling

An alternative PCR-free targeted nanopore sequencing approach is adaptive sampling, also known as ReadUntil, where the nanopore sequencer can identify the target sequences and

selectively reject non-target strands from the pore during sequencing (Kovaka *et al.*, 2021; Payne *et al.*, 2021; Stevanovski *et al.*, 2022) (Figure 3). It is purely a computational method to selectively sequence individual molecules by reversing the voltage across the pores for approximately 0.1 seconds to reject specific strands. Without additional laboratory techniques, users can simply perform the standard ONT library preparation and load the library onto the flow cell. Adaptive sampling can be activated in MinKNOW and requires a fasta/bed file specifying the target regions. Target reads are determined during sequencing itself by real-time basecalling and rapid mapping of at least the first 400 bp of the sequenced reads to the provided input target reference (Marquet *et al.*, 2022) (Figure 3).

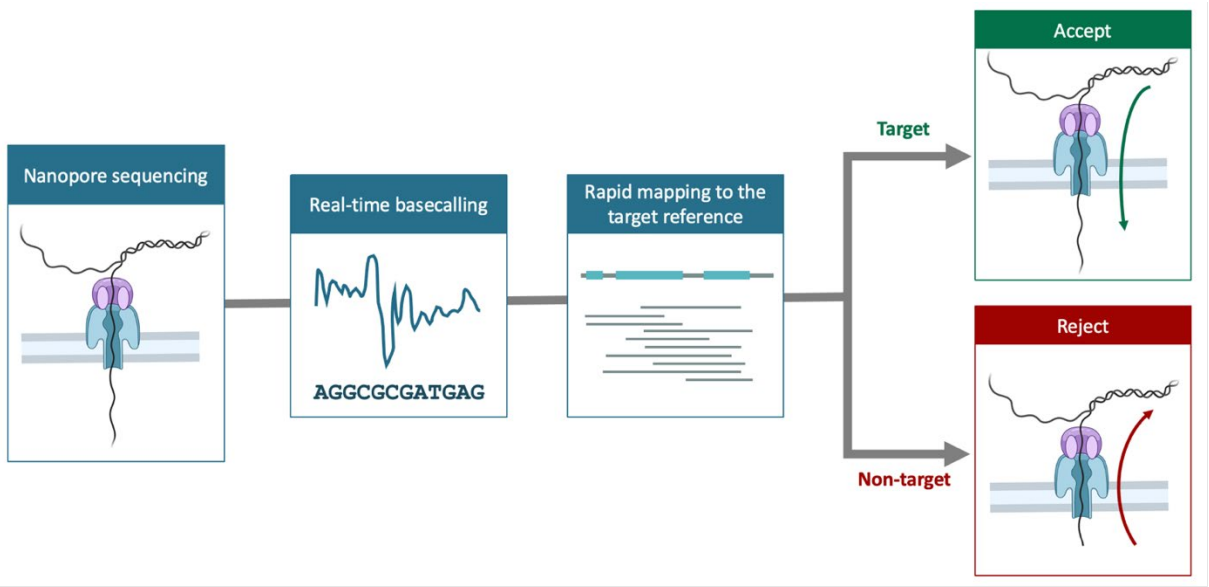


Figure 3. Overview of adaptive sampling.
Adaptive sampling is an in-silico enrichment during nanopore sequencing. A DNA library is prepared using the standard nanopore library preparation kit (LSK109 or LSK110) without a PCR workflow before sequencing on a MinION flow cell. An adaptive sampling option is present in Oxford Nanopore's MinKNOW software in which a target reference file in fasta format specifying target sequences is required for this functionality. Target selection is performed during sequencing by real-time basecalling and rapid mapping of at least the first 400 bp of the sequenced reads to the provided target reference. If the target region is found, sequencing continues. Otherwise, the read is rejected, and the pore will sequence the next strand. Created with BioRender.com.

Alternatively, ONT provides the Read Until application programming interface (API) as a tool for researchers to integrate their own application with the MinKNOW server to obtain raw

signal data in real-time (Oxford Nanopore Technologies, 2020c). The data can then be handled and analysed in any way fit for purpose, and requests can be sent back to the server to eject the read from the specified pore by reversing the polarity of the voltage. Two open-source computational tools have been developed – Readfish (Payne *et al.*, 2021) and UNCALLED (Kovaka *et al.*, 2021), which utilise the ReadUntil API for selective rejection of reads from pores in real time. The former one uses GPU for live basecalling and selectively rejects off-target reads (Payne *et al.*, 2021), and the latter one maps the signals to the reference without the basecalling step by utilising a probabilistic model to consider k-mers that are representable by the signal and then eliminating the candidates based on the reference genome encoded within a Ferragina-Manzini (FM) index (Kovaka *et al.*, 2021).

CHAPTER II: AIMS

1. Systematic benchmarking of tools for DNA methylation detection from nanopore sequencing

Although there are multiple tools that enable target detection from Nanopore sequencing, a lack of systematic benchmarking poses a significant challenge for users in assessing the reliability of their predictions. Although Nanopore provides an opportunity to detect methylation at the single-molecule level, its accuracy in this context remains largely unknown. This precludes the development of reliable and cost-effective applications in medical, forensic, and environmental samples. It is thus necessary to thoroughly characterize the strengths and limitations of the different available tools and establish different strategies for the accurate detection of 5mC in DNA.

A systematic benchmarking of nanopore methylation tools has been performed using controlled methylation mixtures built from *E. coli* methylated and unmethylated control samples. The performance of these tools is compared at both read and genome levels. To facilitate the overall methylation analysis, Snakemake pipelines were developed for each tool to enable the systematic application of the analyses to other datasets and ensure consistent inputs and outputs. Various strategies are also introduced to improve the prediction accuracy, including the combination of predictions from different tools, the use of different score cutoffs and the removal of reads with uncertain methylation state.

Research objectives:

- Evaluate the state-of-the-art computational tools for DNA methylation detection from nanopore sequencing using methylation control samples
- Develop a pipeline to facilitate the test and comparison methylation analysis tools
- Develop strategies to improve the methylation prediction accuracy

2. Validation of nanopore Cas9-targeted sequencing (nCATS)

Targeted sequencing allows sequencing regions of interest (ROIs) with high coverage in a cost-effective manner. Existing assays usually depend on PCR, resulting in loss of base modifications and short amplicon lengths. However, these limitations are overcome by nanopore sequencing which can be coupled with PCR-free targeting methods and call SNPs and methylated bases. In order to enrich and sequence regions of forensic relevance, targeted nanopore sequencing was performed in two different ways, Cas9-based enrichment and adaptive sampling.

The benchmarking work described in Aim 1, was extended to the resulting nanopore Cas9-targeted sequencing (nCATS) data from the GM12878 human cell line for a systematic evaluation of different methylation detection tools. The methylation analyses were then performed using those previously developed pipelines.

To compare DNA methylation levels between nanopore and an established technology, a publicly available whole genome bisulfite sequencing dataset from the Encyclopedia of DNA

Elements (ENCODE) generated by the HiSeq 2500 platform (Illumina) was included in the downstream analysis.

Research objectives:

- Evaluate the use of nCATS for methylation profiling and SNP calling
- Compare DNA methylation levels between bisulfite sequencing (Illumina) and nanopore sequencing (ONT)
- Provide evidence for sequence context biases associated with DNA methylation detection from different tested tools

3. Validation of nanopore adaptive sampling

Taking advantage of nanopore sequencing, the applicability of the unique targeted nanopore sequencing approach - adaptive sampling to target specific DNA-based forensic markers for DNA analysis was scrutinised. Adaptive sampling is an in-silico enrichment approach for ROIs and the depletion of unwanted reads during sequencing. Adaptive sampling was used to enrich and sequence forensic STRs, SNPs, and methylation markers simultaneously. Similar to the nCATS approach, it can be used to generate information-rich data in a single sequencing run without the use of PCR, meaning that base modifications are retained. Additionally, many target loci spanning more than 10kb can be sequenced in single reads.

This technology was tested on known SNPs for EVCs, BGA and DNA methylation markers for age, using methylation control samples, Genome in a Bottle Consortium (GIAB) reference DNA sample HG002 and DNA samples from ten blood donors aged 25 to 76 years.

Research objectives:

- Evaluate the use of adaptive sampling for methylation profiling and SNP calling
- Validate DNA methylation biomarkers from the developed assays for age prediction

CHAPTER III: METHODS

1 METEORE: a pipeline for MEthylation deTEction with nanopORE sequencing

1.1 Methylation control datasets

Positive and negative control sequence reads from E. coli (K12 ER2925) were leveraged from an earlier study (Simpson *et al.*, 2017), for which DNA was amplified by PCR (unmethylated, negative control) and half of the PCR-amplified DNA was subjected to CpG methyltransferase (M.SssI) treatment to methylate cytosines at CpG sites (methylated, positive control). It was reported that M.SssI has an efficiency of 95% (Simpson *et al.*, 2017). This would not affect the benchmarking with the fully unmethylated set (0% methylated), but it might affect the other mixture datasets. Accordingly, to avoid potential biases in our benchmarking analysis, methylation results were gathered in bins of percentage methylation range, with steps of 10%. For instance, the bin of the fully methylated sites included sites of 90-100% methylation. After alignment of all PASS reads for both controls using Minimap2 (Li, 2018) and removing secondary and supplementary reads, 110,795 reads of unmethylated control and 69,453 reads of methylated control remained.

From the 346,793 CpG sites in the *E. coli* reference genome (NC_000913.3), a universally observed set of 100 random sites was selected, each with a single CpG in a 20nt window

(NNNNNNNNNCGNNNNNNNNN, with no CG in the region with Ns) that had aligned sequencing reads from both positive and negative control datasets. Using the filtered PASS reads covering these 100 selected CpG sites from the control datasets (Supplementary Data 1), 11 benchmarking datasets were created with specific mixtures of methylated and unmethylated reads, namely, containing 0%, 10%, …, 90%, 100% of methylated reads. These mixture groups were used for the initial benchmarking of the tools to predict different methylation mixtures per genomic site (mixture dataset 1) (

Table 1a) (Supplementary Data 3). Each set contained approximately 2,400 reads (Table 1a).

For independent validation of the developed model, mixture dataset 2 was developed from a different set of 50 CG sites (single CG in a 20nt window) (Supplementary Data 2), and 6,803 different reads from the remaining fully methylated or fully methylated reads not used in the dataset above. Another 11 benchmarking datasets with specific mixtures of methylated and unmethylated reads was created from this, namely, containing 0%, 10%, 20%, …, 90%, and 100% of methylated reads (mixture dataset 2) (Table 1b) (Supplementary Data 4). Mixture dataset 2 was used to obtain the correlation and error measurements of the different methods, with the expected percentage methylations using the default thresholds suggested by each tool and alternative thresholds obtained from dataset 1. This dataset was also used to test the combined model.

Table 1. Methylation control mixtures.
Descriptions of the mixture datasets 1 (a) and 2 (b) used for the benchmarking of different methylation proportions built from fully unmethylated and fully methylated reads.

a

Set name	% methylated	Total reads	Unmethylated reads (PCR)	Methylated reads (PCR+M.SssI)
m0	0	2390	2390	0
m10	10	2437	2182	255
m20	20	2431	1946	485
m30	30	2434	1714	720
m40	40	2410	1458	952
m50	50	2432	1231	1201
m60	60	2414	981	1433
m70	70	2420	739	1681
m80	80	2410	498	1912
m90	90	2399	256	2143
m100	100	2225	0	2225

b

Set name	% methylated	Total reads	Unmethylated reads (PCR)	Methylated reads (PCR+M.SssI)
m0	0	3420	3420	0
m10	10	3426	3081	345
m20	20	3413	2745	668
m30	30	3423	2410	1013
m40	40	3415	2068	1347
m50	50	3434	1736	1698
m60	60	3403	1377	2026
m70	70	3406	1039	2367
m80	80	3398	690	2708
m90	90	3396	354	3042
m100	100	3383	0	3383

1.2 CpG methylation detection with six different tools

We developed METEORE, a programmatic, standardized workflow for six tools (Table 2): Nanopolish (Simpson *et al.*, 2017), Megalodon (Oxford Nanopore Technologies, 2020b), DeepSignal (Ni *et al.*, 2019), Guppy (Oxford Nanopore Technologies, 2020a), Tombo (Stoiber *et al.*, 2017) and DeepMod (Liu *et al.*, 2019a). Snakemake pipelines to run these tools are available on GitHub (Yuen *et al.*, 2021). The following tools were excluded from the selection process for this study: NanoMod (Liu *et al.*, 2019b), as it only allowed detection of methylation

differences using a control sample and a test sample; SignalAlign (Rand *et al.*, 2017), as its repository had not been updated for over four years; and mCaller (McIntyre *et al.*, 2019), because it only had been trained for 6mA, but not for 5mC. Methylation level was collected for individual CpG sites in both strands and considered per site and per read or summarized (methylation frequency) per site.

Table 2. Tools for 5-methylcytocine (5mC) detection in Nanopore sequencing.
These six tools are included in METEORE for systematic benchmarking. HMM = Hidden Markov model; LLR = Log-likelihood ratio; BiLSTM = Bidirectional Long Short-Term Memory; RNN = Recurrent neural network; BiRNN = Bidirectional recurrent neural network; LSTM = Long Short-Term Memory

Tools	FAST5 input	Trained model	Algorithm
Nanopolish	Multi-read	*E.coli*	HMM
Tombo	Single-read	*E.coli*	LLR
DeepSignal	Single-read	*E.coli*	BiLSTM+Inception
Guppy	Multi-read	*H. sapiens* & *E.coli*	RNN
Megalodon	Multi-read	*H. sapiens* & *E.coli*	RNN
DeepMod	Single-read	*E.coli*	BiRNN+LSTM

1.2.1 Nanopolish

Nanopolish (v0.13.2) assigns a log-likelihood ratio to each individual CpG site or to a group of nearby CpG sites that share the same methylation level in each site within the group. Nanopolish uses reads with a minimum mapping quality score of 20. A positive log-likelihood ratio value indicates evidence of methylation. To include all the predictions per read and per site, such CpG groups were split up into the constituent sites with the same log-likelihood ratio using a Python script incorporated in our Snakemake pipeline (https://github.com/comprna/METEORE) (Yuen *et al.*, 2021). The output file was further processed in R (v3.6.3). Default cutoffs were considered, as originally suggested by Simpson *et al.* (2017), i.e., log-likelihood > 2.5 for methylated sites and < −2.5 for unmethylated sites. Although the default log-likelihood ratio threshold has changed to 2.0 in v0.12 of Nanopolish, no major differences were observed between both thresholds (Figure 4). For the benchmarking

analysis a standardised cutoff with log-likelihood ratio of 2.5 was used for sites in individual reads. The methylation frequency was then calculated for each site as the number of mapped reads predicted as methylated, divided by the number of total mapped reads.

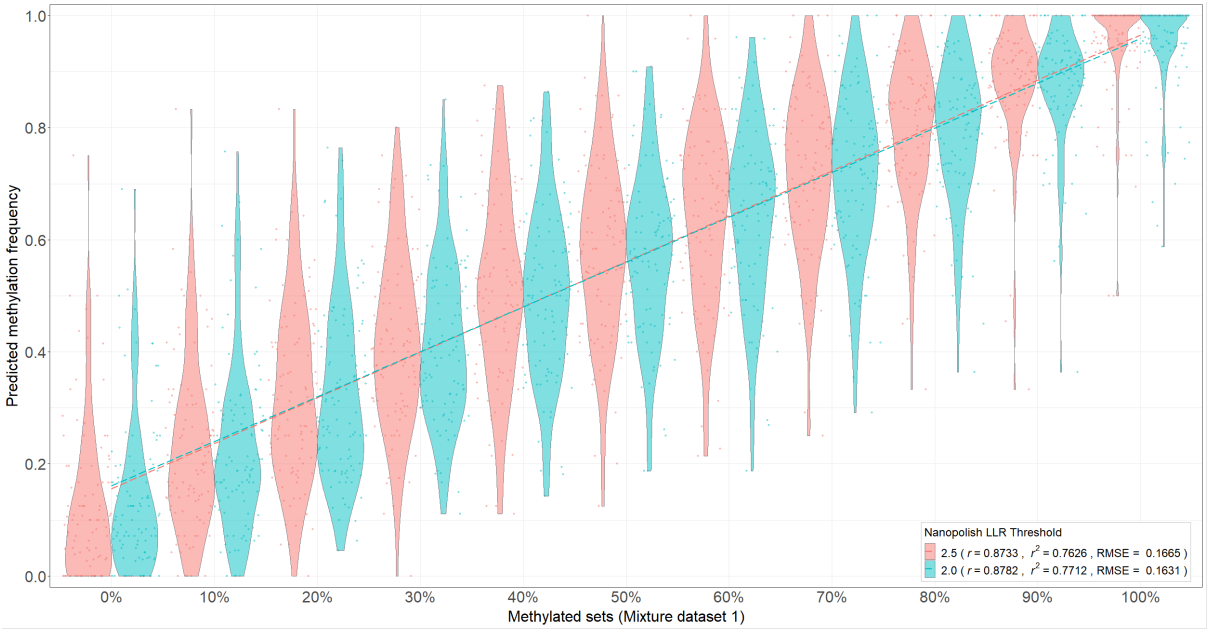


Figure 4. Comparison of methylation frequencies using log-likelihood ratio (LLR) thresholds for Nanopolish. Two different LLR thresholds – 2.5 and 2.0 were used to make a per-site methylation call on control mixture dataset 1.

1.2.2 Tombo

Tombo (v1.5.1) re-squiggles all raw nanopore reads before modified base detection. It then implements three approaches to detect nucleotide modifications: 1) the alternative base detection approach, which computes a statistic by scaling log-likelihood ratios to identify targeted bases where the signals match the expected level for a non-canonical base; 2) the de novo approach, which performs a hypothesis test by statistically comparing signals to an in-silico reference; and 3) the sample comparison approach. This latter approach provides two different ways for modified base detection, one uses a canonical model adjusted by a control set of reads to identify deviations between expected and observed levels, while the other one compares signal levels from two sets of reads at each reference position. Of the three

approaches, the alternative model specific to CpG was used to be able to predict CpG methylation in individual samples and reads. This model compares the signal levels against expected canonical C and alternate 5mC bases at CG motifs, producing the per-read binary statistics in HDF5 format, where positive values indicate canonical bases and negative values modified bases. For the initial analysis, the default cutoffs of (-1.5, 2.5) were used, where scores below -1.5 were considered to be methylated, above 2.5 unmethylated, and scores between these thresholds did not contribute to the per-site methylation. Tombo outputs four individual wiggle files (WIG format), one per strand, reporting the read coverage level and the methylation scores. These files were converted to the same tab-separated values (TSV) format used by other tools with a Python script included in our Snakemake pipeline. For the benchmarking analysis per site and per read, the per-read binary statistics given by Tombo were used as the output score.

1.2.3 DeepSignal

DeepSignal (v0.1.7) uses the re-squiggling algorithm from Tombo in the first step. It then predicts the methylation state of the targeted cytosine at CpG motifs and outputs the probabilities of being methylated, P(m), and unmethylated, P(u), for each cytosine in each read. For the initial analysis, a base was considered as methylated if the methylated probability was greater than the unmethylated probability, P(m)>P(u), and unmethylated otherwise. Additionally, a Python script was added to our Snakemake pipeline to calculate the methylation frequency at each CpG site. For the benchmarking analyses per site and per read, the score calculated as the log2 ratio of the probabilities, i.e., log2(P(m)/P(u)) was taken as the result.

1.2.4 Guppy

Guppy (v3.6.0) filters the Nanopore reads based on the read alignment and implements a neural network model that basecalls 5mC at CG sites as a fifth base along with the four canonical DNA bases. It is only available to members of the Nanopore community (https://community.nanoporetech.com). The configuration file named dna_r9.4.1_450bps_modbases_dam-dcm-cpg_hac.cfg was used to run running Guppy's modified basecalling model. This produced the reads supporting the modifications in fast5 format. The scripts provided at https://github.com/kpalin/gcf52ref were then used to convert the guppy methylation calls from fast5 files to reference anchored files, similar to Nanopolish. These scripts were incorporated into the Snakemake pipeline (https://github.com/comprna/METEORE) (Yuen *et al.*, 2021).

1.2.5 Megalodon

Megalodon (v2.2.4) implements a neural network model that uses Guppy (v4.0.11) to re-basecall all reads and then identifies 5mC by anchoring the basecalling output to the reference, assigning a score for the candidate modified base and performing a calibration for the conversion of the raw scores to estimated empirical probabilities. Megalodon requires Guppy to be installed and the path to Guppy basecalling executable server to be set. The most recent basecalling model in Rerio for Megalodon (Oxford Nanopore Technologies, 2020d) (res_dna_r941_min_modbases_5mC_CpG_v001.cfg) was used. Megalodon produces per-read modified base log probability and canonical base log probability at each mapped CpG site. A default threshold of 0.75 was used as a minimum score for both modified and canonical base probabilities to include a modified basecall in the final aggregated output in the bedMethyl format file containing per-site coverage and methylation percentage. An additional script was

built to convert the Megalodon's output into a per-site result file in an augmented BED format same as other tested tools (https://github.com/comprna/METEORE). For the benchmarking analysis per site and per read, a log-likelihood score calculated by subtracting the natural log-probability of the modified base, log(M) and the natural log-probability of the canonical base, log(C), resulting in log(M/C) was used as the result.

1.2.6 DeepMod

DeepMod takes single-read fast5 files as input and uses reads with a mapping quality score greater than 10. It then outputs a methylation prediction summary per site at genome level for each strand in a BED format containing the coverage, the number of methylated reads and methylation percentage. As DeepMod did not provide any information for individual reads, it could not be used for the per-read benchmarking analyses.

1.3 Processing of per-site methylation calls

The raw output of each of the tools contained methylation information for each aggregated CpG site on both strands. That is, each mapped CpG site on the positive strand of the human reference genome had a counterpart CpG site mapped on the negative strand. To perform the per-site benchmarking, the positive strand coordinate system was used, so the mapped sites that were on the negative strand were lifted to positive coordinates by subtracting 1 from the coordinate position. If there were predictions made on both strands for the same site, the mean methylation frequency for that site was recorded. If there was no per-site information only for one of the strands, then the same prediction from that strand was noted and lifted to the positive strand if necessary.

1.4 METEORE consensus models

METEORE was created to provide a consensus prediction using the scores from two or more methods. Both types of supervised learning approaches, classification and regression, have been implemented in METEORE to test combinations of tools. For the classification model, we used a random forest (RF) classifier (Breiman, 2001) from the Python sklearn library (Pedregosa *et al.*, 2011). The prediction scores from each individual method were scaled to the range of [0,1] using min-max scaling (Breiman, 2001). The Receiver operating characteristic (ROC) and precision-recall (PR) curves were built from a 10-fold cross-validation on the prediction scores from the input methods. The reads were randomly selected during cross-fold validation.

For the METEORE implementation, we used the parameters max_dep=3 and n_estimator=10. The default parameters of RF from sklearn (n_estimator = 100 and max_dep = None) were also tested. The METEORE RF model tested in these analyses was trained on the entire mixture dataset 1 with the scores from Megalodon and DeepSignal and using parameters max_dep=3 and n_estimator=10.

The multiple linear regression-based approach used sklearn's RidgeCV linear regression (REG) model, and min-max scaling as in the RF model. The initial ROC and PR curves were produced with the built-in 5-fold cross validation in RidgeCV. The METEORE REG model tested in these analyses was trained on the entire mixture dataset 1 with the scores from Megalodon and DeepSignal. After prediction, resulting scores were classified as unmethylated if they were less than 0.5 and methylated if greater. METEORE REG required reads with prediction scores from all provided tool inputs for using multiple linear regression to model

expected methylation, leading to a slight (~10%) decrease in observed reads in each test. The scripts to train and run METEORE are available at https://github.com/comprna/METEORE (Yuen *et al.*, 2021).

1.5 Sequence motif analysis and visualization

7-mers of targeted CpG sites from *E.coli* mixture datasets were collected and all sites detected in the CGIs from the nCATS datasets. Sequence logos were generated from 7-mer data in FASTA format using WebLogo (Crooks *et al.*, 2004), and the k-mer contexts of CpG sites using the high-coverage nCATS data from Gilpatrick *et al.* (2020) was also explored. The sequence motifs were grouped by the absolute difference between the methylation frequencies produced by each tool and the WGBS values for each CpG site in an 8-mer context (`NNNCGNNN`), where a site with a score of 0 was labelled as "low discrepancy", and with a score greater than 0.5 was labelled as "high discrepancy". The pLogo Web tool (O'Shea *et al.*, 2013) was used to visualize the sequence motifs and assess the significance of the differences in the frequency of residues between the "low discrepancy" k-mers (foreground dataset) and the "high discrepancy" k-mers (background dataset).

2 Cas9-targeted Nanopore sequencing (nCATS)

2.1 DNA sample used in the nCATS experiment

The DNA of a reference human genome for NA12878 from Utah with Northern and Western European ancestry was sequenced. This individual has been sequenced on a wide variety of platforms and has well-validated variation call datasets from 1000 Genomes as well as bisulfite sequencing data from the ENCODE projects, which enabled to benchmark the sequencing results.

2.2 Guide RNA design and RNP complex assembly

We used the Cas9-targeted Nanopore sequencing (nCATS) protocol (Gilpatrick *et al.*, 2020) to target ten regions of the human genome. This PCR-free protocol uses Cas9 to cut double-stranded DNAs (dsDNAs) at specific sites and then preferentially ligates sequencing adapters to the cleaved ends for enrichment. The cleaved target DNA strands with adapters attached are then sequenced. We designed ten pairs of RNA CRISPR guides (crRNAs) for ten forensically relevant regions (Table 3).

Table 3. Ten forensically relevant regions used for the nCATS protocol.
The table provides the coordinates (GRCh38) of the ten regions used to sequence native DNA with the nCATS protocol. The table also indicates whether the region contains an ancestry-informative SNP (aiSNP), a phenotypic-informative SNP (piSNP), or a short tandem repeat (STR).

Region	Size (nt)	Target locus	Type of variants	Associated gene(s)
chr1:159199780-159212236	12456	rs2814778	aiSNP	*CADM3*, *ACKR1*
chr2:1480363-1494141	13778	TPOX	STR	*TPO*
chr4:99314722-99323024	8302	rs1229984	aiSNP	*ADH1B*
chr5:33944710-33959555	14845	rs16891982	piSNP	*SLC45A2*
chr6:392228-401463	9235	rs12203592	piSNP	*IPF4*
chr11:89258999-89295942	36943	rs1393350	piSNP	*TYR*
chr14:92303402-92323757	20355	rs12896399	piSNP	*SLC24A4*
chr15:27983280-27993166	9886	rs1800407	piSNP	*OCA2*
chr15:281127012-8130250	17549	rs12913832	piSNP	*HERC2*
chr21:43627562-43644088	16526	PentaD	STR	*HSF2BP*

An initial panel of candidate crRNAs was designed using the freely available tool CHOPCHOP (Labun *et al.*, 2019) as recommended in the ONT protocol. For each target region, about three to five best crRNAs were selected based on the cleavage location, crRNA efficiency, and the number of predicted mismatches using CHOPCHOP. The crRNAs were then evaluated by the IDT's design checker (Integrated DNA Technologies, 2019) to select for high on-target performance and low off-target activity.

Ten pairs of guide RNA were used to enrich ten human regions ranging from 8-36kb. Here, one gRNA was used on either side of each target region (Table 4). Each RNA oligo including crRNAs (IDT, custom-designed) and tracrRNA (IDT, 11-05-01-12) was resuspended in IDTE buffer pH 7.5 (IDT, 11-01-02-02) to a final concentration of 100uM. crRNAs were then pooled to make an equimolar crRNA mix by combining equal volumes of each crRNA. In order to form gRNA duplexes, the crRNA pool and tracrRNA were then combined in equimolar concentrations with duplex buffer (IDT, 11-05-01-12), followed by denaturation for 5 min at 95 °C, then allowing to cool to room temperature. RNP complexes were created by assembling

the following components: gRNA duplexes, CutSmart buffer (NEB, B7204), nuclease-free water and Cas9 endonuclease (IDT, 1081058).

Table 4. Guide RNA (gRNA) panel used for the nCATs protocol.
The table describe the ten pairs of gRNAs used to target the ten regions from Table 3. To enrich for each target region, two gRNAs were used to make a cut on each side, one upstream of the region of interest (ROI) targeting the positive strand, and the other one downstream targeting the negative strand.

Target	Guide RNA sequence	PAM	Cleaved site
rs12913832	CTTGTTCTCAATCCAACGAG	CGG	chr15:28112701(+)
	GATCAGATGACCATGTTCGA	AGG	chr15:28130250(-)
rs1800407	GTAGAGCTCTAACTAAGTGG	AGG	chr15:27983280(+)
	TATCCAATCCTGCTGACCAG	TGG	chr15:27993166(-)
rs12896399	GCTGGAACGCCCCATCAACA	CGG	chr14:92303402(+)
	GAGTGCAATCAGTGGCCGAG	CGG	chr14:92323757(-)
rs16891982	TGTGATCACCACGACGACAA	CGG	chr5:33944710(+)
	GAGTGCAACGAGGAACTAAG	AGG	chr5:33959555(-)
rs1393350	TCCTTGCTGCACGAATCAGT	GGG	ch11:89258999(+)
	GCTGGATGTGTTATAGACGC	TGG	chr11:89295942(-)
rs12203592	TAAGGGGCCCAAGCTCACGG	CGG	chr6:392228(+)
	ACGTGGTCAGCTCCTTCACG	AGG	chr6:401463(-)
TPOX	CGTATTTGAAAGATCCACGG	TGG	chr2:1480363(+)
	CTTACGTAAGAGTTGAATGG	TGG	chr2:1494141(-)
Penta D	CGGTACCTATCCCAGAACTA	TGG	chr21:43627562(+)
	TAACACGTAGATCATTCACT	TGG	chr21:43644088(-)
rs2814778	CCTACCACGCCATCATCGGT	GGG	chr1:159199780(+)
	GCAATTGTCTTTCAGTGCGT	TGG	chr1:159212236(-)
rs1229984	ACCATCTGCTAACACGTATG	AGG	chr4:99314772(+)
	GCGTTAACATATCTCCACAA	GGG	chr4:99323024(-)

2.3 Library Preparation

Genomic DNA (gDNA) from the GM12878 human cell line was obtained from the Coriell Institute (coriell.org) (cat. No. NA12878). The purity of the purchased gDNA was measured with the Nanodrop spectrophotometer (TFS) at the 260/280nm and 260/230 nm values. Dephosphorylation of 5' ends of DNA was performed as follows. A 5ug amount of input DNA was dephosphorylated to prevent downstream adapter ligation using Quick dephosphorylation

Kit (NEB, M0508). DNAs were resuspended in the CutSmart buffer and dephosphorylated with Quick CIP enzyme in a PCR tube for 10 min at 37 °C, followed by heating for 2 min at 80 °C for CIP enzyme inactivation. Cas9 Cleavage and dA-tailing was performed as follows. A 100 mM aliquot of deoxyadenosine triphosphate (dATP) (NEB, N0440S) was first diluted to 10 mM. After allowing the dephosphorylated DNA sample to return to room temperature, the pre-assembled RNP complexes, 10 mM dATP and Taq DNA Polymerase with Standard Taq Buffer (NEB, M0273) were added to the PCR tube containing the sample for the in vitro digestion reaction and subsequent dA-tailing. The sample was then incubated at 37 °C for 15 min, and then at 72 °C for 5 min.

By dephosphorylating pre-existing DNA ends prior to Cas9 cleavage, sequencing adapters and ligation buffer from the Oxford Nanopore Ligation Sequencing Kit (ONT, LSK109) were preferentially ligated to the cleaved DNA ends at Cas9 cleavage sites using T4 Ligase from the NEBNext Quick Ligation Module (NEB, E6056) for 10 min at room temperature. The sample was cleaned up to remove excess adapters using the Agencourt AMPure XP beads (Beckman Coulter, A63881), washing twice on a magnetic rack with the long-fragment buffer (ONT, LSK109) before eluting in 14 µl of elution buffer (ONT, LSK109). A 1 µl aliquot of the final library was quantified using the Qubit dsDNA Broad Range assay kit (TFS). Starting with 5 µg of input DNA, 1 µg of DNA was recovered after library preparation. The library was stored on ice until ready to load.

2.4 Sequencing and data processing

Before loading, the flowcell was primed with a solution consisting of flush buffer (ONT, LSK109) and flush tether (ONT, LSK109). The sequencing library was prepared by adding sequencing buffer (ONT, LSK109) and loading beads (ONT, LSK109) into the DNA library,

and then loaded into the flowcell. The sample was run on a MinION flow cell (FLO-MIN106, R9.4.1 pore) using the MinION sequencer for 19 hr, operated using the MinKNOW software. Live basecalling was carried out during the experiment using Guppy's fast basecalling model in MinKNOW. The resulted FASTQ files were immediately aligned to the human reference genome (GRCh38/hg38) using minimap2 (Li, 2018), followed by visualization with IGV (Robinson *et al.*, 2011; Thorvaldsdóttir *et al.*, 2013) (Thorvaldsdóttir *et al.*, 2013) to confirm the generation of on-target sequencing reads.

Post-run basecalling was performed using Guppy (v.3.2.4) high-accuracy model to generate the final set of sequencing reads with higher read accuracy than the fast model and recognition of modified bases from the electrical signal data. Reads were aligned to the human reference genome (GRCh38/hg38) using minimap2. Using Samtools (v1.9) (Danecek *et al.*, 2021), we collected aligned reads within the enriched regions as on-target reads. Those outside the targeted regions were considered off-target reads and subsequently discarded. Coverage plots for target regions were generated using the Snakemake (Köster & Rahmann, 2012) workflow developed by Oxford Nanopore Technologies (Oxford Nanopore Technologies, 2019).

2.5 Comparison with bisulfite sequencing data

We used METEORE to detect modifications using raw nanopore signals from the fast5 files, which generates per-site predictions in an augmented BED format for each nanopore methylation tool included in our METEORE pipelines. We then compared the resulting predictions with the published WGBS data for NA12878 (ENCODE accessions: ENCFF279HCL, ENCFF835NTC) using two different approaches. In the first approach, we compared every CpG site on both strands for Nanopore and WGBS data, preserving the strand information. In the second approach, we used the positive strand coordinate system by lifting

all sites from the negative strand to be on the positive strand, i.e., site position – 1. For the sites that had methylation evidence from both strands, we combined the sites by averaging the methylation frequencies and adding up the coverage. Otherwise, we preserved the methylation frequency if the site only had methylation prediction from one strand.

To obtain high confidence methylation calls from WGBS data for validation, the resulting individual or combined CpG sites were processed in the following way. CpG sites with zero coverage from both WGBS replicates were discarded. Furthermore, we calculated the difference in methylation frequency between both WGBS replicates and considered the 0.1 and 0.9 quantiles of the distribution of differences. A CpG site was kept if the difference for that site was between those 0.1 and 0.9 quantiles, otherwise it was removed. We finally calculated the number of sites, Pearson and Spearman correlations and coefficient of determination between methylation frequencies calculated from WGBS and those calculated from Nanopore reads by each of the tested methods using `cor.test()` and `lm()` functions from R (R Core Team, 2020). We also subsampled the reads for the CpG sites that were covered with at least 1, 5, 10, 15, 20, etc. reads and calculated Pearson correlation at different levels of read coverage for further evaluation.

3 Adaptive sampling

3.1 DNA samples

There were three different types of gDNA samples used in our adaptive sampling runs. Firstly, we ran adaptive sampling on the GIAB reference DNA sample HG002. This reference DNA sample was used to test the assay's performance with our custom designed target panel.

Secondly, we performed two other adaptive sampling runs on two methylation control samples from EpigenDX (https://www.epigendx.com/d/products/methylation-controls). These methylation controls consisted in DNA derived from human whole blood, enzymatically treated resulting in greater than 85% methylation in the high methylated control and less than 5% methylation in the low methylated control.

Lastly, we tested adaptive sampling on human blood samples. This study was approved by the ethics committees at the Australian National University (ANU) (ethics protocol no. ETH.1.16.01/ETH.01.15.015). Blood samples were collected from the participants aged 25−76 years at the Centre for Personalised Immunology, ANU (Table 5). DNA was then extracted from blood using the Monarch HMW DNA Extraction Kit for Cells & Blood (NEB, T3050) according to the manufacturer's protocol.

Table 5. Details of ten human blood samples.

Sample no.	Age	Gender
1	25	F
2	26	F
3	28	F
4	31	F
5	42	F
6	65	M
7	67	M
8	72	M
9	75	M
10	76	M

3.2 Quality control

The purity of the gDNA samples was measured with the NanoDrop based on A260:A280 and A260:A230 ratios and concentration in ng/μL. Quantification of gDNA concentration (in ng/μL) was determined using the Qubit dsDNA Broad Range Assay Kit (TFS). The Agilent Femto Pulse system was also used to assess the fragment size of the samples, and thus determine suitability for shearing and size selection in the next step.

3.3 Shearing and size selection

Based on the initial QC results, samples were needle-sheared 5 to 15 times with 26 and/or 29 gauge needles, followed by confirmation on the Agilent Femto Pulse system. The samples were then electrophoretically size-selected using the BluePippin system (Sage science) with 0.75% Agarose gel cassettes and marker S1, enriching for fragments that were at least 20kb long. After that, samples were recovered and cleaned up with the AMPure XP beads and quantified on Qubit with dsDNA Broad Range reagents to assess suitability for subsequent library preparation. The fragment sizes of the samples ranged from 20 to 50 kb.

3.4 Library preparation and sequencing with adaptive sampling

Standard Nanopore sequencing libraries were prepared from genomic DNA methylation controls purchased from EpigenDx, using SQK-LSK110 gDNA by ligation sequencing kit (ONT) following the manufacturer's instructions. Each resulting library was then loaded into a MinION flow cell (FLO-MIN106D, R9.4.1 pore) and sequenced on the MinION sequencer which plugged into a Type-A USB 3.0 port on the computer workstation with the NVIDIA GeForce RTX 3070 GPU card. MinKNOW version v20.10.3 and Guppy version v4.2.2 were installed on the computer connected to the MinION.

An adaptive sampling run was set up on the MinKNOW software, where the adaptive sampling was set to “enrich” and the sequencing parameters were set to use the fast basecalling model of Guppy. Target regions were defined as a multi-fasta file including forensically relevant STR and SNP loci, DNA methylation markers associated with age, and three copies of a tandem repeat of the rDNA unit (KY962518.1) containing two units in three possible orientations with respect to the direction of transcription: 1) the same orientation, 2) convergent or 3) divergent. Additional information indicating the targeted regions can be found in Supplementary Data 5. Most samples were completed within 48 hours, with nuclease flushing of the flow cell to increase throughput and additional loading to the washed flow cell. Based on sequencing output, samples that produced only a total of ~2 million reads required additional library preparation and ran for another ~36 hours, at which point the health of the flow cells was diminished, to generate sufficient data.

3.5 Methylation analysis

Nanopore reads were analysed with Megalodon (v2.5.0) coupled with Guppy (v6.1.1) to perform basecalling, alignment to the modified GRCh38/hg38 human genome assembly containing three additional tandem copies of the rDNA canonical unit (KY962518.1) and methylation calling using Remora model (v1.1.1) (Oxford Nanopore Technologies, 2022). The modified reference file used in the analysis is provided in Supplementary Data 6. The outputs from Megalodon were converted into per-site result files in an augmented BED format using the script in METEORE (https://github.com/comprna/METEORE). The percentage of methylated reads was calculated per site in the methylation control samples. For individual rDNA-containing nanopore reads, sites were considered methylated if the per-read modified base score was above 0.75, and unmethylated otherwise.

3.6 Variant calling

The following tools were used to call SNPs with nanopore data: BCFtools (v1.15.1), GATK (v4.2.6), Nanopolish (v0.13.3), Longshot (v0.4.5) and PEPPER-Margin-DeepVariant (version r0.8). For BCFtools, samtools was first used to calculate the genotype likelihoods, then the resulting likelihoods were piped to BCFtools to perform SNP calling and generate the final VCF file (Danecek *et al.*, 2021). GATK is a popular method to call variants from short reads. GATK's haplotypeCaller in a single-sample mode was used to call variants using the FASTQ file containing raw unmapped sequence to generate the VCF file, following the pipeline as shown in Broad Institute (2023). Longshot utilises a pair-HMM for a small local window around candidate sites to call SNPs on long-read data (Edge & Bansal, 2019), whereas Nanopolish utilises a signal level HMM to detect variants. Prior to calling methylation or

variants or aligning events by Nanopolish, the reads were indexed with f5c (Gamaarachchi *et al.*, 2020). PEPPER-Margin-DeepVariant is a haplotype-aware variant calling pipeline specially designed for detecting variants in a reference genome with long-reads, and it implements several methods including a recurrent neural network (RNN), a convolution neural network (CNN) and an HMM to produce variant calls (Shafin *et al.*, 2021). Recall was calculated as correctly called SNPs (true positives) out of all true SNPs (true positives + false negatives). Precision was calculated as correctly called SNPs (true positives) out of all predicted positive SNPs (true positives + false positives). The F1 score was included as a measure of overall test accuracy, calculated as the harmonic mean of precision and recall.

CHAPTER IV: RESULTS

1 Systematic benchmarking of tools for DNA methylation detection from nanopore sequencing using the METEORE pipeline

A standardized workflow was developed to obtain 5mC calls at CpG sites from Nanopolish (Simpson *et al.*, 2017), Megalodon (Oxford Nanopore Technologies, 2020b), DeepSignal (Ni *et al.*, 2019), Guppy (Oxford Nanopore Technologies, 2020a), Tombo (Stoiber *et al.*, 2017) and DeepMod (Liu *et al.*, 2019a) (Figure 5), using Snakemake pipelines (Köster & Rahmann, 2012) (https://github.com/comprna/METEORE) (Yuen *et al.*, 2021). This workflow ensures consistent inputs and outputs for all tools and facilitates the integration and interpretation of DNA methylation calls. Nanopolish detects CpG methylation with a HMM, while Megalodon, DeepSignal and DeepMod use neural networks, and Tombo applies a statistical test to identify DNA modifications. Both Tombo and DeepSignal re-squiggle the raw signals before detection. Differently from the other tools, Guppy directly basecalls 5mC using an extended alphabet. Megalodon anchors the Guppy basecalling output to the reference and assigns a score for the candidate modified base. All tools output per-site methylation calls on both strands at genome level.

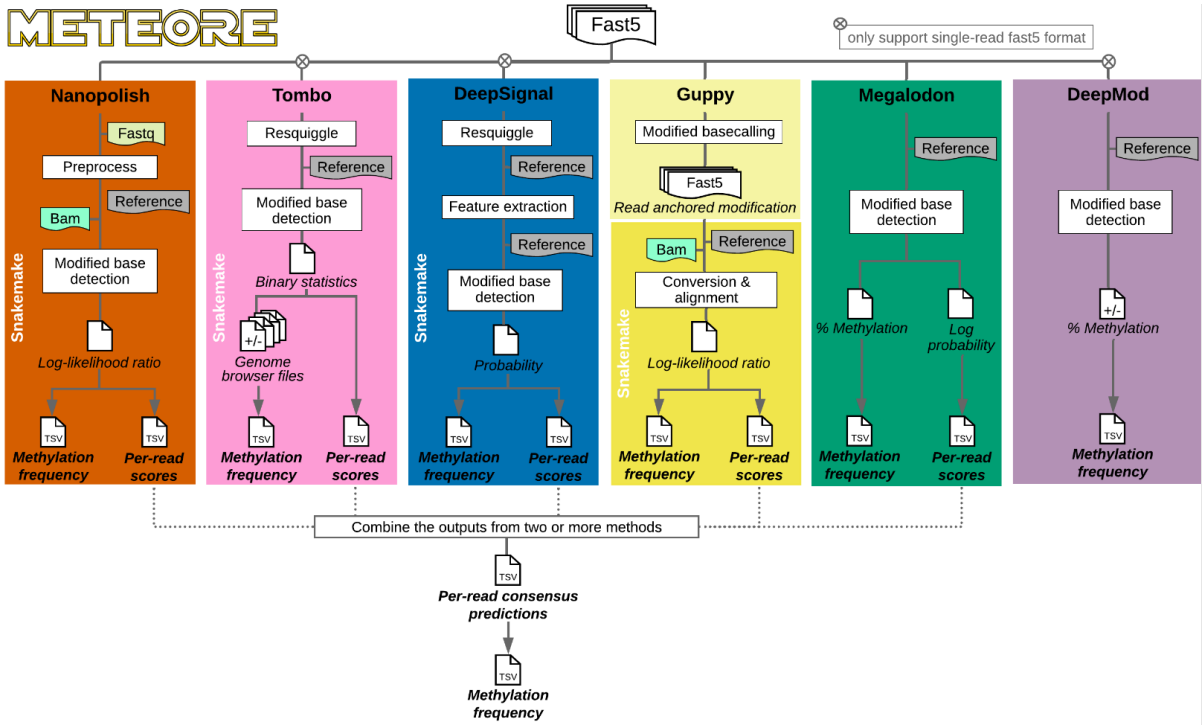


Figure 5. Analysis pipeline for 5mC detection at CpG sites from Nanopore sequencing.
The diagram describes the approach used to test the tools Nanopolish, Tombo, DeepSignal, Megalodon, Guppy and DeepMod. SnakeMake pipelines and command lines used are available at https://github.com/comprna/METEORE (Yuen *et al.*, 2021). A SnakeMake pipeline was not developed for Megalodon and DeepMod as they can be run with a single command. Excluding DeepMod, all tools produce predictions per individual read and per CG site. Additionally, all tools predict the methylation frequency at each genome site from fast5 input files. Methods that currently only accept single-read fast5 format are indicated.

1.1 Performance of CpG methylation detection on methylation controls

We first evaluated the six tools on methylation control datasets built from PCR-amplified DNA (negative control) and M.SssI-treated DNA (positive control) (Simpson *et al.*, 2017). From the 346,793 CpG sites in the E. coli reference genome, we selected 100 arbitrary sites containing a single CpG site with a 20nt window on either side with no CGs (Supplementary Data 1), and with a minimum read coverage of 50x (median coverage 85x), from both positive (methylated) and negative (unmethylated) control datasets. Using these reads, we created 11 benchmarking datasets with specific mixtures of methylated and unmethylated reads, namely, containing 0%,

10%, …, 90%, and 100% of methylated reads, each set with approximately 2,400 reads (mixture dataset 1) (Supplementary Data 3). There are no strong differences in the sequence context around the selected CpG sites (Figure 6).

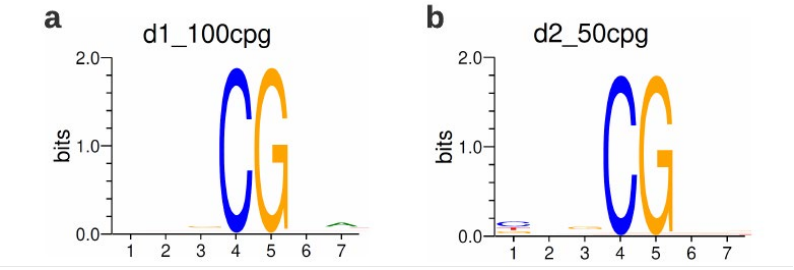


Figure 6. Sequence contexts of the CpG sites in datasets 1 and 2.
We show the sequence logos with the information content in bits (y axis) for the positions surrounding the CpG sites used in mixture datasets 1 (a) and 2 (b). The two datasets do not show strong differences in the sequence context around the selected sites.

We examined the per-site methylation frequency predicted by each tool across these 100 sites using the default cutoff for each method (Supplementary Data 7). All tools, except Tombo, achieved Pearson correlations above 0.8 (p-value < 2.2e-16 for all tools) (Figure 7a). The highest correlation and lowest root mean square error (RMSE) values were attained by Megalodon, followed by DeepMod and DeepSignal.

Despite the good correlation values, most tools showed high dispersion and low agreement with the expected percentage methylation per site. Guppy had the highest RMSE values and systematically underpredicted the per-site methylation (Figure 7a). Nanopolish and Tombo showed high dispersion and systematically overpredicted. To further assess how the dispersion affected the per-site accuracy, we calculated the proportion of sites predicted outside a 10% window around the expected value for each percentage methylation subset. Guppy had most sites predicted outside expected windows (Figure 7b). Nanopolish and Tombo had the lowest proportion of sites predicted outside the m90 and m100 windows, but the highest proportions

at low methylation frequency (Figure 7b). In contrast, Megalodon had most sites predicted within the expected windows in low methylation subsets (Figure 7b).

We further assessed the classification of fully unmethylated and fully methylated sites according to specific thresholds (Table 6) (Supplementary Data 8 and 9). For 0% methylation, Guppy and Megalodon correctly recovered all sites as unmethylated if a methylation proportion <0.1 was considered unmethylated, whereas other methods predicted correctly fewer unmethylated sites (Figure 7c). In the 100% methylated set, using a methylation proportion >0.8, only Nanopolish, Tombo, and Megalodon correctly predicted most of the sites as fully methylated (Figure 7d). In contrast, Guppy failed to predict any sites at this cutoff (Figure 7d). These differences in the accuracy at fully methylated and fully unmethylated sites, as well as the general high dispersion observed at intermediate methylation levels, were the impetus to identify alternative strategies to achieve higher accuracies.

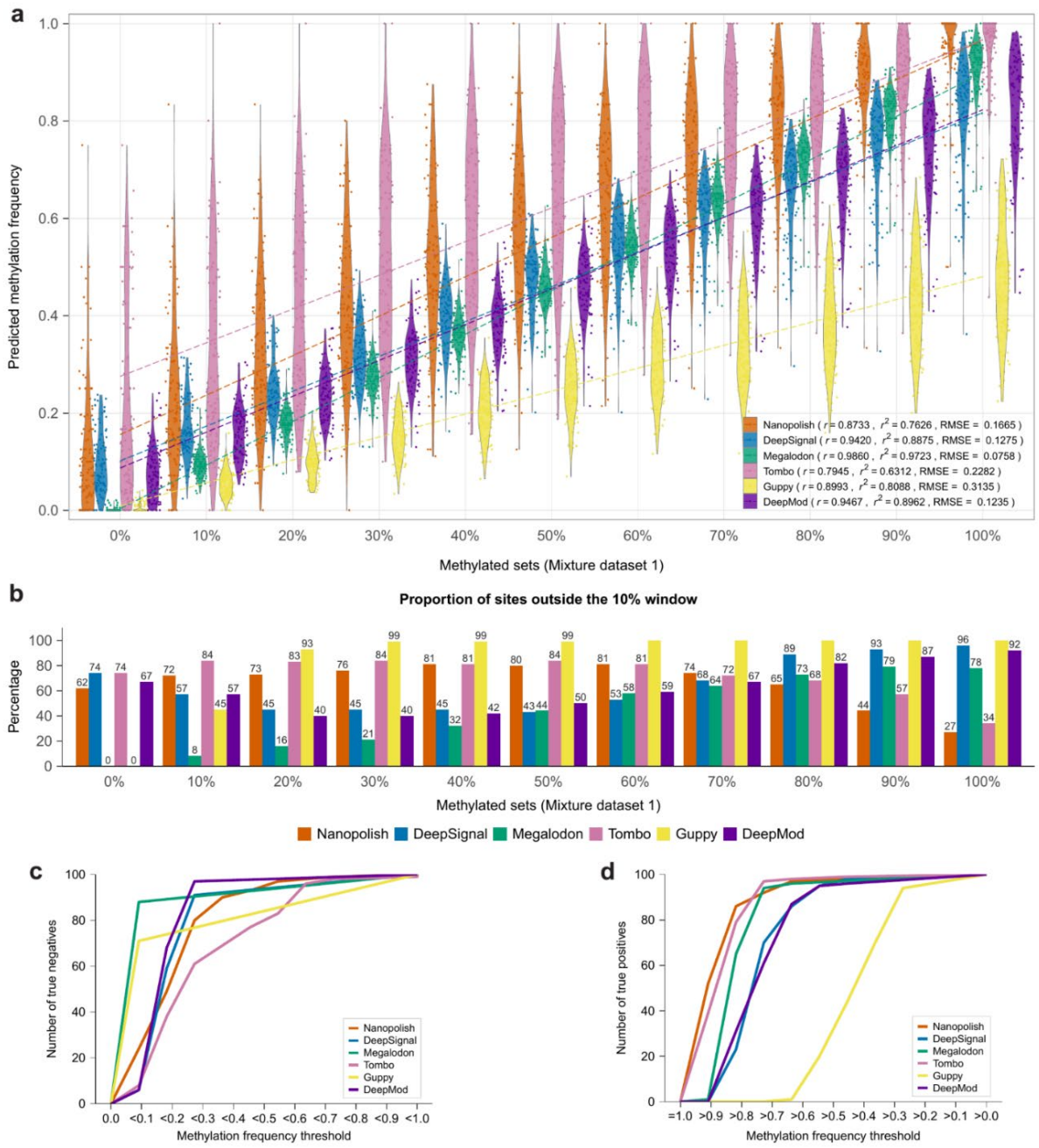


Figure 7. Accuracy analysis per CpG site on control mixture dataset 1.

(a) Violin plots showing the predicted methylation frequencies (y axis) for each control mixture set with a given proportion of methylated reads (x axis). The Pearson's correlation (r), coefficient of determination (r2) and root mean square error (RMSE) are given for each tool. **(b**) For each method we indicate the proportion of sites predicted outside a 10% window around the expected methylation proportion, i.e. each predicted site in the m% dataset was classified as "outside" if its predicted percentage methylation was outside the interval [(m-5)%, (m+5)%] for intermediate methylation values, or outside the intervals [0,5%] or [95%,100%] for the fully unmethylated or fully methylated sets, respectively. The percentage is indicated on top of each bar, except for 100%. **(c)** Empirical cumulative distribution function (ECDF) plot showing the number of true negatives (y axis) for each tool according to different thresholds for the predicted methylation frequency below which a site was called unmethylated (x axis), using the dataset of 100 fully unmethylated sites. **(d)** ECDF plot showing the number of true positives (y axis) for each tool according to different thresholds for the predicted methylation frequency above which a site was called fully methylated (x axis), using the dataset of 100 fully methylated sites.

Table 6. Per-site performance.

The table shows the accuracies in fully methylated or fully unmethylated CpG sites for the six tested tools using two different methylation frequency thresholds to classify methylated and unmethylated sites. For each pair of thresholds (a,b), we defined a site to be unmethylated if the predicted methylation frequency was <a, and methylated if the predicted methylation frequency was >b, for (a,b) = (0.1,0.9), (0.2,0.8), (0.3,0.7), (0.4,0.6), and (0.5,0.5).

Unmethylated if freq < 0.1, methylated if freq > 0.9					
	Accuracy	Specificity	Precision	Recall	Error rate
Nanopolish	0.675	0.490	0.628	0.860	0.325
DeepSignal	0.410	0.590	0.359	0.230	0.590
Megalodon	0.825	1.000	1.000	0.650	0.175
Tombo	0.588	0.384	0.564	0.790	0.412
Guppy	0.500	1.000	NA	0.000	0.500
DeepMod	0.495	0.680	0.492	0.310	0.505

Unmethylated if freq < 0.2, methylated if freq > 0.8					
	Accuracy	Specificity	Precision	Recall	Error rate
Nanopolish	0.860	0.800	0.821	0.920	0.140
DeepSignal	0.805	0.910	0.886	0.700	0.195
Megalodon	0.970	1.000	1.000	0.940	0.030
Tombo	0.794	0.616	0.719	0.970	0.206
Guppy	0.500	1.000	NA	0.000	0.500
DeepMod	0.790	0.970	0.953	0.610	0.210

Unmethylated if freq < 0.3, methylated if freq > 0.7					
	Accuracy	Specificity	Precision	Recall	Error rate
Nanopolish	0.935	0.900	0.907	0.970	0.065
DeepSignal	0.930	1.000	1.000	0.860	0.070
Megalodon	0.980	1.000	1.000	0.960	0.020
Tombo	0.839	0.697	0.766	0.980	0.161
Guppy	0.505	1.000	1.000	0.010	0.495
DeepMod	0.935	1.000	1.000	0.870	0.065

Unmethylated if freq < 0.4, methylated if freq > 0.6					
	Accuracy	Specificity	Precision	Recall	Error rate
Nanopolish	0.955	0.930	0.933	0.980	0.045
DeepSignal	0.975	1.000	1.000	0.950	0.025
Megalodon	1.000	1.000	1.000	1.000	0.000
Tombo	0.884	0.778	0.818	0.990	0.116
Guppy	0.600	1.000	1.000	0.200	0.400
DeepMod	0.975	1.000	1.000	0.950	0.025

Table 6 (continued).

	Unmethylated if freq < 0.5, methylated if freq > 0.5				
	Accuracy	Specificity	Precision	Recall	Error rate
Nanopolish	0.975	0.970	0.970	0.980	0.025
DeepSignal	0.990	1.000	1.000	0.980	0.010
Megalodon	1.000	1.000	1.000	1.000	0.000
Tombo	0.915	0.838	0.861	0.990	0.085
Guppy	0.720	1.000	1.000	0.440	0.280
DeepMod	0.980	1.000	1.000	0.960	0.020

1.2 Methylation prediction accuracy in individual molecules

Nanopore sequencing provides the opportunity to detect nucleotides and their modifications in individual molecules. The accuracy of the tools for identifying 5mC sites in individual reads was therefore expolored. The per-read performance of each tool was evaluated across their range of prediction scores at each site in individual reads (Supplementary Data 10 and 11), excluding DeepMod which only provides the percentage of methylation per site, so it could not be included in this analysis. All tested tools achieved areas under the receiver operating characteristic (ROC) curve (AUC) (Figure 8a) and areas under the precision-recall (PR) curve AUC_{PR} above 0.8 (Figure 8b). Megalodon showed the highest AUC and AUC_{PR} values, followed by DeepSignal and Nanopolish (Figure 8b). Guppy had decreased precision at high recall values (Figure 8b). The differing accuracies of methods across different conditions suggested that a consensus approach may capture the advantages of the methods and compensate for their potential deficiencies.

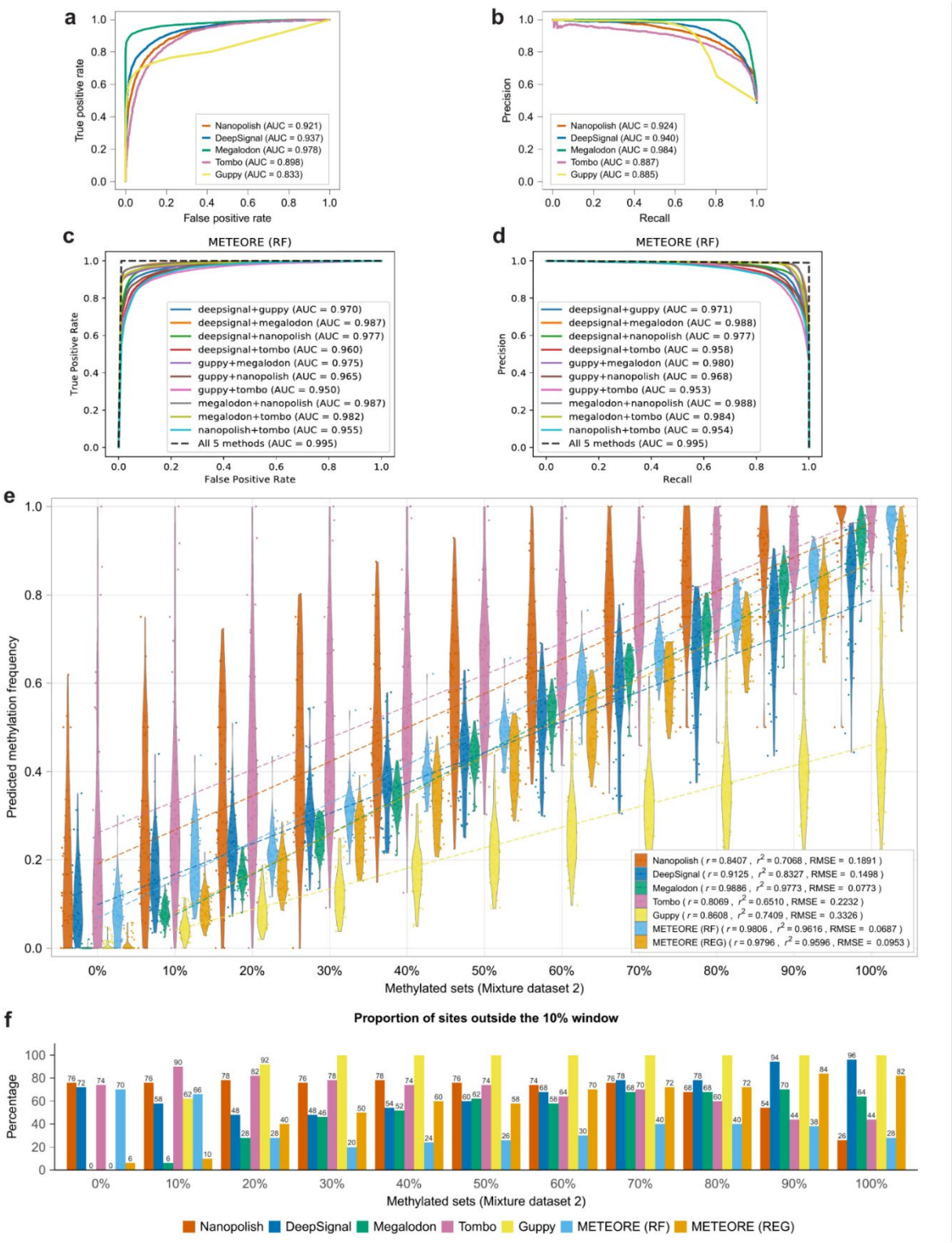


Figure 8. Model accuracy at the individual read level and per-site accuracy analysis on control mixture dataset 2.

(a) Receiver operating characteristic (ROC) curves showing the false positive rate (x axis) and true positive rate (y axis) for the predictions at individual read levels for the 5 methods tested, using reads from 0% and 100% methylated sets. **(b)** Precision-recall (PR) curves showing the recall (x axis) and precision (y axis) for the predictions at individual read levels for the 5 methods tested, using reads from 0% and 100% methylated sets. **(c)** ROC curves for METEORE for the random forest (RF) model (parameters: max_depth=3 and n_estimator=10) combining two methods, as well as combining the 5 methods. The curves were built from the average of a 10-

fold cross validation with mixture dataset 1. Similar plots for another RF model using default parameters and a regression (REG) model are shown in Fig. 3.7. **(d)** PR curves for the same models as in (c). **(e)** Violin plots showing the predicted methylation frequencies (y axis) for each control mixture set with a given proportion of methylated reads (x axis) from the mixture dataset 2 for the five tested tools plus METEORE combining Megalodon and DeepSignal using a random forest (RF) or a regression (REG) model. The Pearson's correlation (r) and coefficient of determination (r2) are given for each tool. **(f)** We indicate the proportion of sites predicted outside a window around the expected methylation proportion, i.e. a site in the m% dataset was "outside" if the predicted percentage methylation was outside the interval [(m−5)%,(m+5)%] for intermediate methylation sets, or outside the intervals [0,5%] or [95%,100%] for the fully unmethylated or fully methylated sets, respectively. The percentage is indicated on top of each bar, except for 100%.

1.3 Combinations of predictive methods improve accuracy in individual molecules

We also developed a consensus approach in METEORE (Yuen et al., 2021), that combines the predictions from two or more tools. The consensus was implemented using two different models, a random forest (RF) and a multiple linear regression (REG). The combination of two methods using either of these two models provided overall an increase in accuracy compared to individual methods (Figure 8c, Figure 8d & Figure 9). The five individual tools were then compared with METEORE (RF and REG) for the combination of Megalodon and DeepSignal, on a different collection of methylation mixtures (mixture dataset 2) (Table 1b) (Supplementary Data 2 and 4). The predictions were performed on individual reads and then summarized per CG site to compare with the expected percentage methylation (Supplementary Data 12). METEORE RF combining Megalodon and DeepSignal achieved lower RMSE compared with the individual tools (Figure 8e). METEORE REG performed similarly to Megalodon, and improved accuracy over the other tools (Figure 8e). METEORE RF also showed an improvement in the proportion of sites predicted within the expected 10% window at intermediate and high methylation sets (Figure 8f).

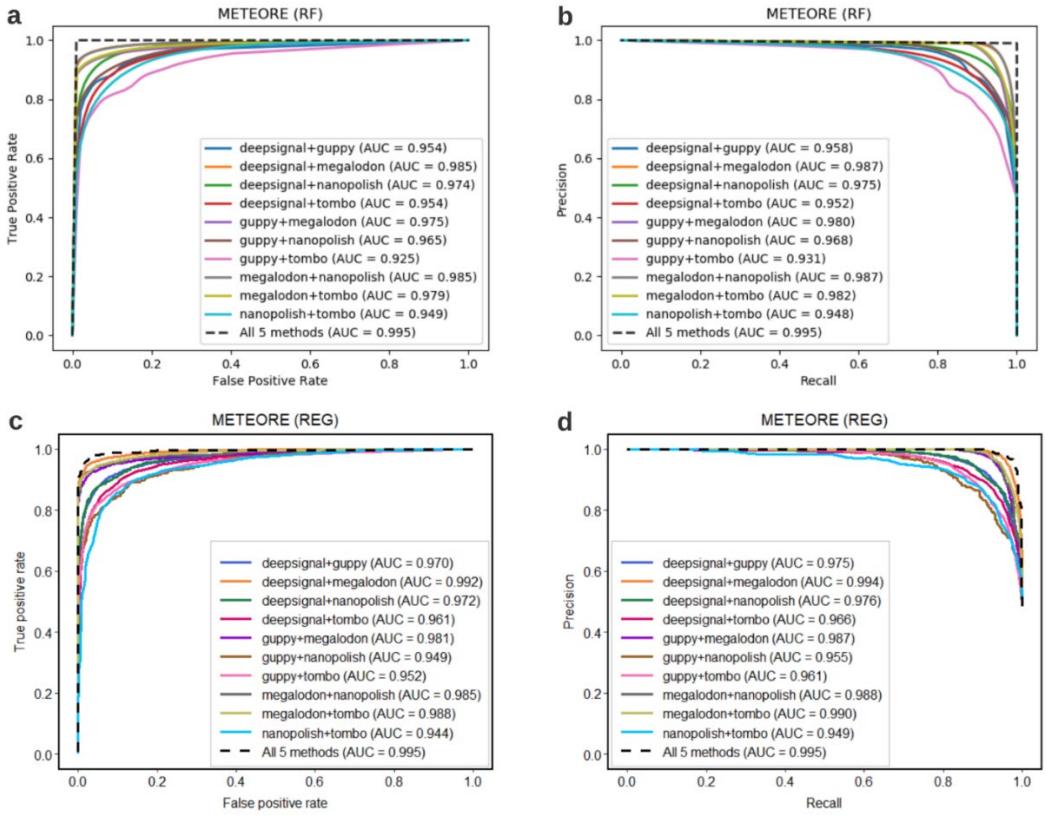


Figure 9. Model accuracy for METEORE at the individual read level.
(a) Receiver operating characteristic (ROC) curves showing the false positive rate (x axis) and true positive rate (y axis) for the predictions of the METEORE random forest (RF) model combining two methods with default parameters (n_estimator = 100 and max_depth = None) using the sites of the mixture dataset 1 at individual read level. The curves were built from the average of a 10-fold cross validation. **(b)** Precision-recall (PR) curves showing the recall (x axis) and precision (y axis) for the RF model using the sites of the mixture dataset 1 at individual read level, also built from 10-fold cross validation. **(c)** ROC curves for the METEORE regression (REG) model combining two methods at individual read level. Curves were built from a 5-fold cross validation using the sites of mixture dataset 1. **(d)** PR curves for the METEORE REG model using the sites of the mixture dataset 1 at individual read level, built from a 5-fold cross validation.

To measure the runtime and memory consumption of each tool, we tested each METEORE pipeline on the m50 set (2,432 reads and a total of 15,564,827 bases) from the mixture dataset 1 (Table 7). We recorded the real time (wall clock time) from start to finish of the pipeline/command(s) for Nanopolish, DeepSignal, Tombo, Megalodon, Guppy and DeepMod, which took a fast5 directory as an input and output a prediction at genome level, i.e., methylation frequency for each site. For the METEORE consensus model, we only recorded the time needed for running a single Python script to make the per-read consensus predictions,

which was independent of the two methods being combined. Here we used METEORE combining DeepSignal and Megalodon using a RF (parameters: max_depth=3 and n_estimator=10) and a REG model, where both models took per-read prediction outputs generated from the selected tools. To estimate the overall central processing unit (CPU) time and memory used for the METEORE consensus model, any two tools can be considered, plus the overhead of an additional step of METEORE.

Table 7. Runtime and memory usage for each tested tool.
We ran each tested tool on a computer with 12 central processing unit (CPU) processors (Intel Core i7 (8th Gen) 8700 @ 3.2 GHz). Guppy can be run on CPUs or graphics processing units (GPUs), but on a GPU the basecalling speed increase significantly. Megalodon requires Guppy for the basecalling step and uses GPU-enabled Guppy by default. If Megalodon is used with Guppy (CPU), the --guppy-timeout argument should be specified (here we used 200 seconds) to allow sufficient time for calling a read during CPU basecalling. Additionally, we ran Megalodon and Guppy on another computer with a GPU (GeForce RTX 2080 Ti) and 32 CPU processors (AMD Ryzen threadripper 2950x). Real time per CPU = real (wall-clock) time taken x no. of CPUs. Bases per second = total no. of bases / (real time per CPU x 60).

	No. of CPUs used	Real time per CPU (min)	Peak memory (GB)	Bases per second
Nanopolish	2	10.2	0.2	25433
DeepSignal	9	334.8	23.7	775
Tombo	9	30.4	23.6	8533
Megalodon	11	3258.7	1.4	80
Megalodon (GPU)	1 GPU	3.8 per GPU	2.4	69710
Guppy	11	1494.6	3.3	174
Guppy (GPU)	1 GPU	7.0 per GPU	0.6	37596
DeepMod	8	176.0	2.5	1474
METEORE (RF)	1	0.1	0.3	-
METEORE (REG)	1	1.8	9.5	-

Since the tested tools generally show a high dispersion and present common patterns across the increasing methylation control sets in both mixture datasets, we considered the most reliable sites with either 0% or 100% methylation using WGBS data for validation (Figure 10a & Figure 10b). The tested tools still presented dispersion as observed before (Figure 10c & Figure 10d),

indicating there is still a major limitation in the capacity to correctly recover the methylation levels from individual reads.

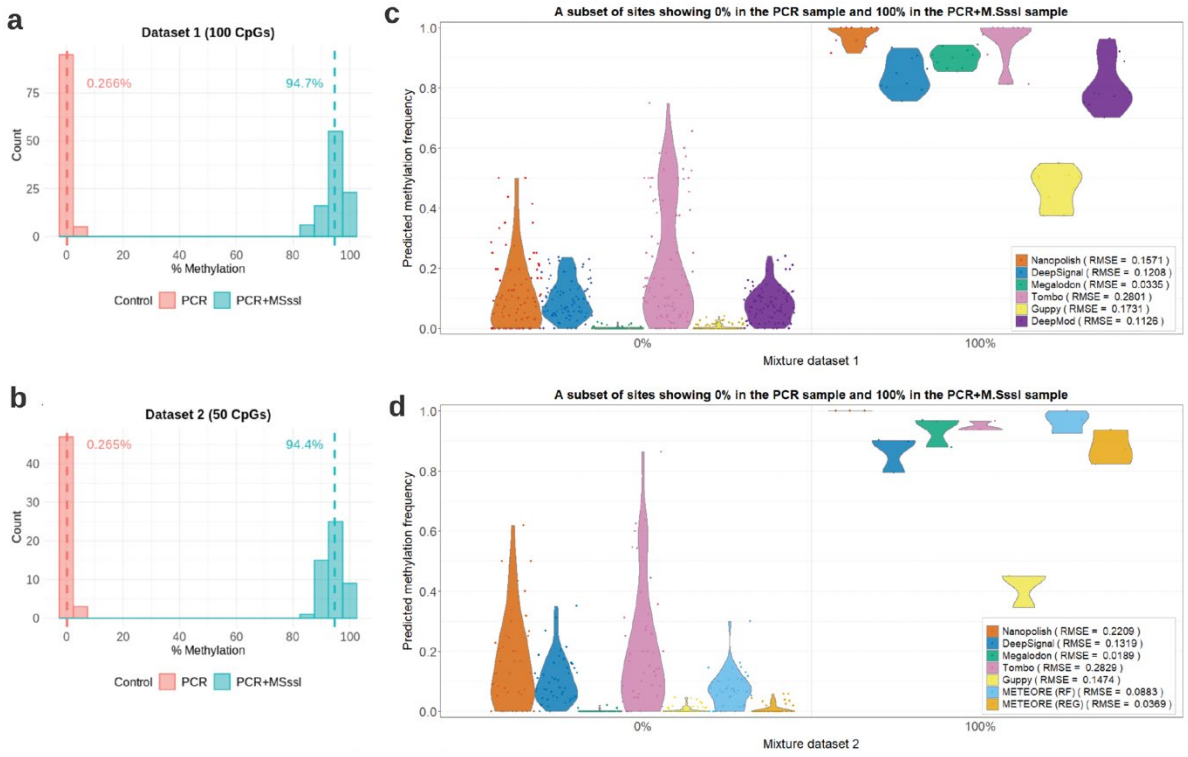


Figure 10. Confirmation of the methylation levels of control samples using whole genome bisulfite sequencing (WGBS) data.

(a) Distribution plot showing the methylation percentage for the 100 target sites in the mixture dataset 1 in the positive (PCR+M.SssI) (mean methylation = 94.7%) and negative (PCR) (mean methylation = 0.266%) controls. **(b)** Distribution plot showing methylation percentage for the 50 target sites in the mixture dataset 2 in the positive (mean = 94.4%) and negative (mean = 0.265%) controls. **(c)** Violin plots showing the distribution of predicted methylation frequencies (y axis) in the sites from dataset 1 that in (a) showed exactly 0% methylation in the negative control and exactly 100% methylation in the positive control, using WGBS. The root mean square error (RMSE) is given for each tool. **(d)** Violin plots showing the distribution of predicted methylation frequencies (y axis) in the sites from the mixture dataset 2 that in (b) showed exactly 0% methylation in the negative control and exactly 100% methylation in the positive control, using WGBS. The root mean square error (RMSE) is given for each tool.

1.4 Varying single score cutoffs improve the accuracy of methylation predictions

At this stage, the tools were applied with their default score cutoffs. We reasoned that it might be possible to identify different cutoffs for the scores at the individual read level to improve the accuracy of the predictions of methylation frequency per site. To do this, we considered the distribution of scores (Figure 11a) and several accuracy metrics using individual reads from the mixture dataset 1 (Figure 11b).

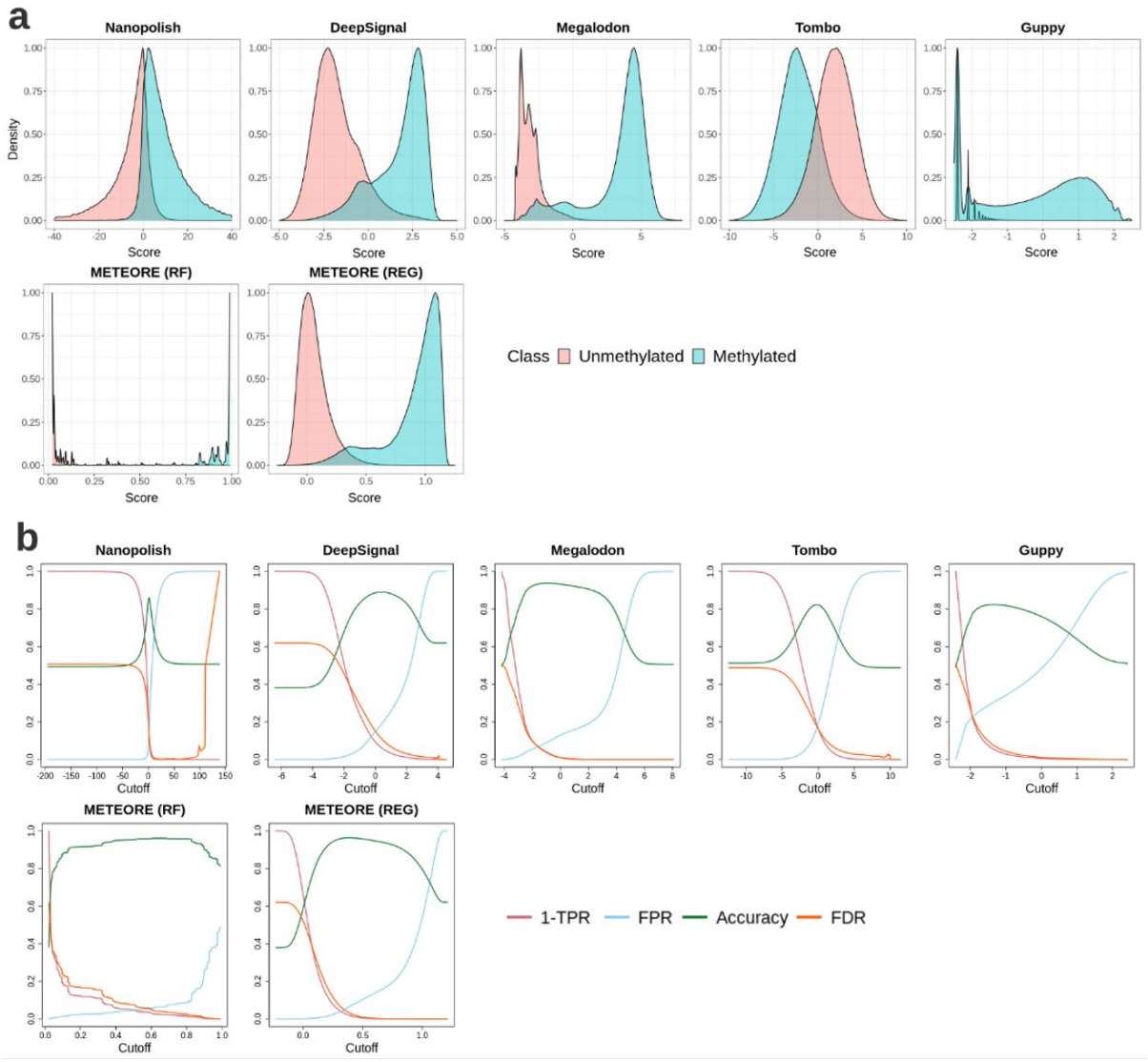


Figure 11. Score distributions and accuracy metrics.

Figure 11. (continued)
(a) Score (x axis) distribution for methylated and unmethylated sites in individual reads for each tested tool on the mixture dataset 1, including METEORE. We used METEORE combining the predictions of Megalodon and DeepSignal with a random forest (RF) (parameters: max_dep=3 and n_estimator=10) and with a regression model (REG). **(b)** Distribution of various accuracy metrics (y axis) according to the score (x axis) for each method shown in (a). We show the false positive rate (FPR), false discovery rate (FDR), 1 - true positive rate (TPR), and Accuracy curves as a function of the single score (x axis) cutoff for each tool, where FPR = FP / (FP+TN), TPR = TP / (TP+FN), FDR = FP / (TP+FP), accuracy = (TP + TN) / (TP + TN + FP + FN); and TP = true positives, FP = false positives, TN = true negatives, and FN = false negatives.

We then determined for each method the score that corresponded to the maximum of TPR – FPR (TPR = true positive rate and FPR = false positive rate) (Table 8). Applying these scores to separate methylated and unmethylated sites per-read on the mixture dataset 2, all tools had a lower RMSE value with respect to the default cutoffs (Figure 12a). In terms of correlation, all tools except Nanopolish and Megalodon improved. Selecting the score corresponding to the minimum of $FPR^2 + (1\text{-}TPR)^2$ led to similar cutoffs (Table 8) and results (Figure 12b). Applying these cutoffs to METEORE (RF and REG) combining Megalodon and DeepSignal improved upon the default cutoffs and achieved a lower RMSE compared with the individual tools (Figure 8e, Figure 12a & Figure 12b). Nevertheless, all tools still showed a high proportion of predicted sites outside the expected 10% window. In particular, Nanopolish and Megalodon had more sites predicted outside the expected windows at high and low methylation sets, respectively (Figure 12c & Figure 12d). In contrast, METEORE RF and REG had fewer sites predicted outside the windows at low methylation and high methylation respectively (Figure 12c & Figure 12d).

Table 8. Single score cutoffs.

Cutoffs obtained by maximising the value of TPR−FPR or by minimizing the value of $(FPR-0)^2 + (TPR-1)^2$. In both optimization we used all reads from the mixture dataset 1. These cutoffs were applied to the per-read data generated by each tool. For all these tools except for Tombo, if a read with a score above the cutoff, we consider it as methylated, and unmethylated for the scores below the cutoff. For Tombo, a read is considered methylated if its score is below the cutoff, and unmethylated for a score above the cutoff. All values were rounded up to 2 decimal places. For METEORE REG, both strategies led to exactly the same cutoffs after rounding up.

		Nanopolish	DeepSignal	Megalodon	Tombo	Guppy	METEORE (RF)	METEORE (REG)
Maximum of (TPR – FPR)	Cutoff	1.03	-0.05	-0.98	-0.13	-1.35	0.62	0.33
	TPR	0.87	0.86	0.91	0.83	0.69	0.94	0.95
	FPR	0.15	0.10	0.04	0.18	0.05	0.03	0.03
Minimum of $(FPR-0)^2 + (TPR-1)^2$	Cutoff	1.04	-0.19	-1.36	-0.11	-1.70	0.55	0.33
	TPR	0.87	0.87	0.92	0.82	0.73	0.95	0.96
	FPR	0.1	0.11	0.05	0.18	0.12	0.04	0.04

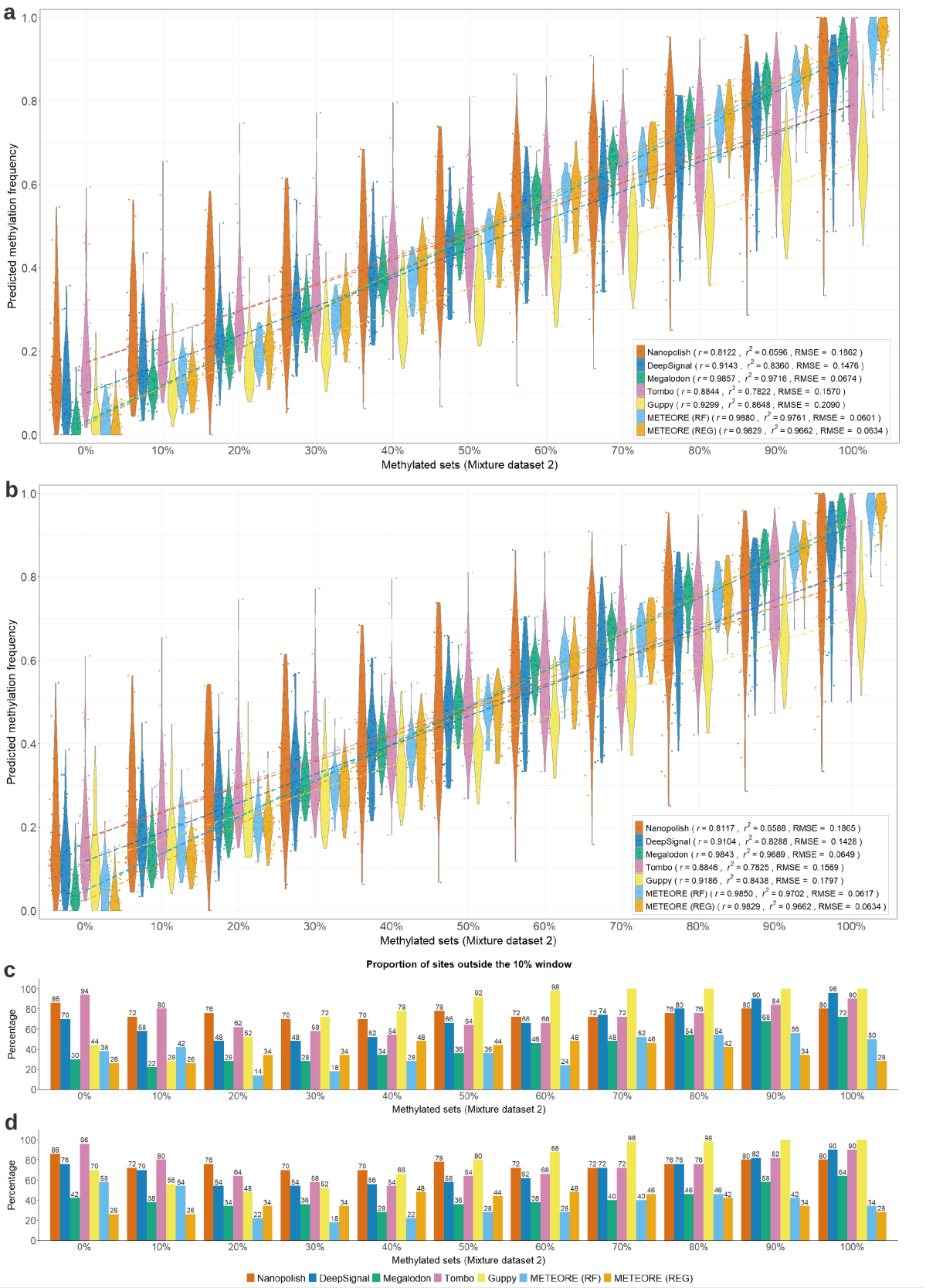


Figure 12. Accuracy analysis of five individual tested tools using a single score cutoff.

(a-b) Violin plots showing the predicted methylation frequencies (y axis) for each control mixture set with a given proportion of methylated reads (x axis) from the mixture dataset 2 for the five tested tools plus METEORE combining Megalodon and DeepSignal using random forest (RF) and regression (REG) models with the single

threshold obtained by **(a)** the maximum value of (TPR-FPR) or **(b)** the minimum value of $FPR^2 + (1-TPR)^2$. Score thresholds are given in Table 8. The Pearson's correlation (r), coefficient of determination (r^2) and the root mean square error (RMSE) are given for each tool. **(c-d)** Barplot showing the proportion of sites predicted outside a 10% window around the expected methylation proportion for each method with the single threshold obtained by **(c)** the maximum value of (TPR-FPR) or **(d)** the minimum value of $FPR^2 + (1-TPR)^2$. Each predicted site in the m% dataset was classified as "outside" if its predicted percentage methylation was outside the interval [(m-5)%, (m+5)%] for intermediate methylation values, or outside the intervals [0,5%] or [95%,100%] for the fully unmethylated or fully methylated sets, respectively. TPR = true positive rate, FPR = false positive rate.

1.5 Discarding reads of uncertain methylation prediction state improves accuracy

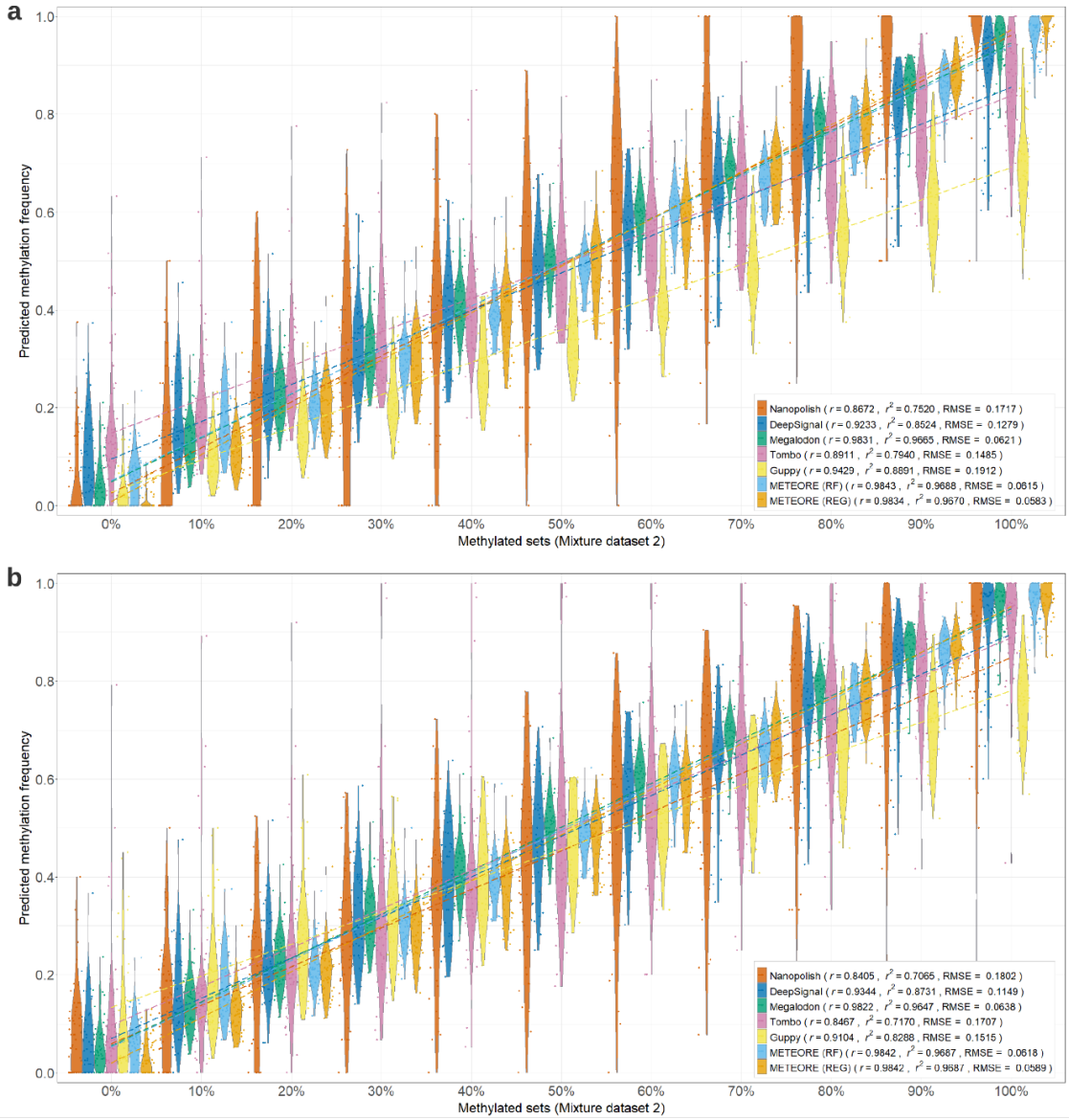
We considered the alternative strategy of discarding sites in reads with uncertain methylation state. This is used by default in Nanopolish and Tombo, which define a double cutoff (higher and lower than the point of indecision). To test this strategy, we used the distribution of scores (Figure 11) and removed the 10% of sites in individual reads that were closest to the score at which the FPR and 1-TPR curves crossed (Table 9).

Table 9. Double score cutoffs.

Cutoffs obtained by removing 10% of reads around the intersection point of the FPR curve and 1-TPR curve or removing the cases that fall between the score at 1- TPR = 0.05 and FPR = 0.05. In the first optimization we used all reads from the mixture dataset 1. In the second optimization we used the 0% and 100% methylation sets from dataset 1. For each double cutoff (a,b), all sites in reads with score < a are considered unmethylated, with score > b are considered methylated, and all cases between these values are discarded. For Tombo the score scale has the opposite orientation, i.e., a read is considered methylated if its score is < a, and unmethylated for a score > b. METEORE (RF) and METEORE (REG) are combination of DeepSignal and Megalodon, which use a random forest (parameters: max_depth=3 and n_estimator=10) and a regression model, respectively. In the REG model, filtering is symmetric by the ranking of scores about the tipping point, i.e., 5% of reads with scores lower than the tipping point and 5% of reads with scores higher than the tipping point.

	Remove **10%** of reads around the cross point of FPR and 1-TPR curves	Remove the reads that fall between the score at **1- TPR=0.05** and **FPR=0.05**
Nanopolish	(-3.58 ,5.80)	(-0.65, 3.53)
DeepSignal	(-0.60, 0.09)	(-1.07, 0.60)
Megalodon	(-2.07, -1.20)	(-2.14, -1.29)
Tombo	(-0.54,0.34)	(-1.82, 1.66)
Guppy	(-1.93, -0.91)	(-2.41, -1.31)
METEORE (RF)	(0.47,0.56)	(0.48,0.55)
METEORE (REG)	(0.20,0.46)	(0.29, 0.34)

Using this approach, all methods except Megalodon achieved higher correlation and lower RMSE values compared with the default cutoffs (Figure 13a). We also assessed the scores at which FPR=0.05 and 1-TPR=0.05 and removed all predictions between these two values (Table 9). This led to improved performance of all methods with respect to default cutoffs, except for Megalodon and Nanopolish (Figure 13b). Additionally, we observed that Nanopolish used much fewer reads compared to the other methods (Figure 13c).



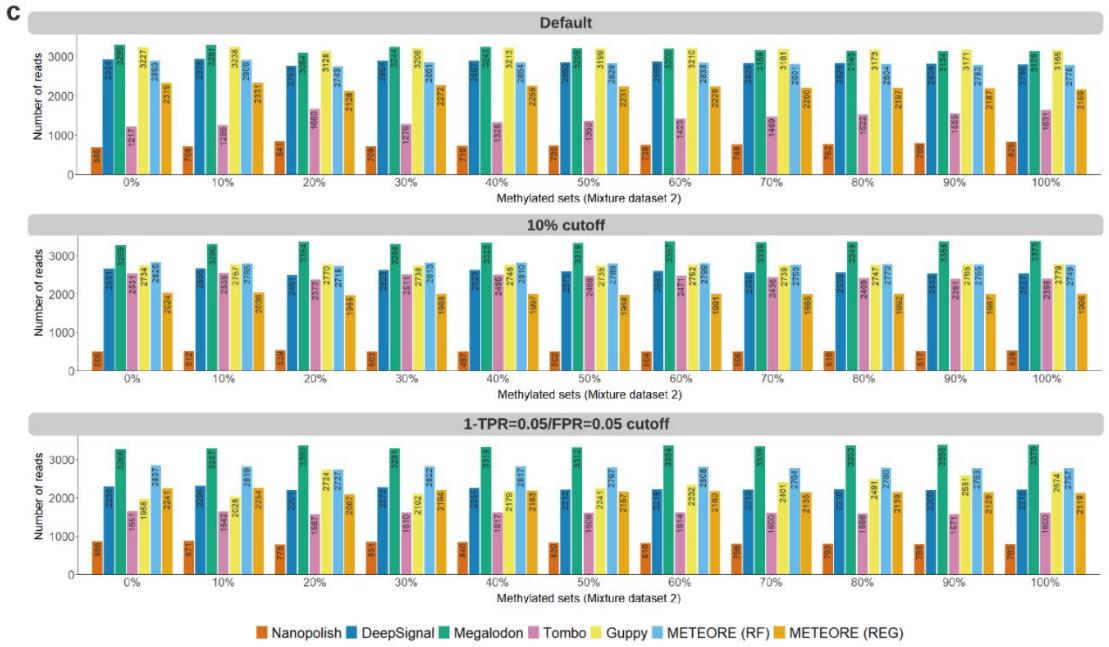


Figure 13. Accuracy analysis using a double cutoff and discarding reads.

(a) Violin plots showing the predicted methylation frequencies (y axis) for each control mixture set with a given proportion of methylated reads (x axis) from the mixture dataset 2 for the five tested tools plus METEORE combining Megalodon and DeepSignal using random forest (RF) and regression (REG) models, after discarding 10% of the reads with a score closest to the value corresponding to the intersection between FPR and 1−TPR. The Pearson's correlation (r), coefficient of determination (r2) and the root mean square error (RMSE) are given for each tool. **(b)** Similar plot as (a) but considering the scores at which FPR=0.05 and 1-TPR=0.05 and removing all sites in reads with a score between these two values. Cutoffs are given in Table 9. **(c)** Total number of reads reported by each method in different approaches: default setting of each method (top), the use of a double cutoff to remove 20% of the reads (middle) and the use of a double cutoff to remove reads with the scores between the cutoff values at FPR=0.05 and 1−TPR=0.05 (bottom). The number of reads is shown inside each bar.

2 Evaluation of DNA methylation with nanopore Cas9-targeted sequencing (nCATS)

Commonly used methods for targeted sequencing include PCR and CRISPR/Cas9 enrichment, which has only become widely available in the last three years (Fiol *et al.*, 2022; Gilpatrick *et al.*, 2020; Goldsmith *et al.*, 2021; López-Girona *et al.*, 2020). PCR becomes problematic when amplifying long DNA fragments (> 5kb), which limits the ability to generate long reads with nanopore sequencing (Kovaka *et al.*, 2021). ONT has introduced short fragment mode (SFM) that integrated into the MinKNOW following the release of the MinKNOW v22.03 on 23 March 2022. This unlocks the ability to sequence short fragments down to 20 bases long, but PCR-based methods still cannot identify base modifications without additional processing such as bisulfite conversion.

In this chapter, a CRISPR/Cas9 approach was used for targeted enrichment of ten forensically relevant regions followed by nanopore sequencing (nCATS) in order to sequence native DNA molecules harboring base modifications of a human lymphoblastoid cell line (GM12878) with a MinION flowcell.

2.1 Evaluation of the nCATS method

The ability of the nCATS method to capture the specified ROI was tested by targeting ten forensic loci (*rs2814778, TPOX, rs1229984, rs16891982, rs12203592, rs1393350, rs12896399, rs1800407, rs12913832, PentaD*) (Table 3) in our panel and used one gRNA on either side of each locus (Table 4) in the GM12878 cell line. These ten regions were successfully enriched and sequenced but with variable coverage, likely due to on-target cutting efficiency of different gRNAs (Figure 14a). We also found that off-targeted reads were distributed randomly across the genome with 1× median coverage. This may be explained by the adapter ligation at random cut sites.

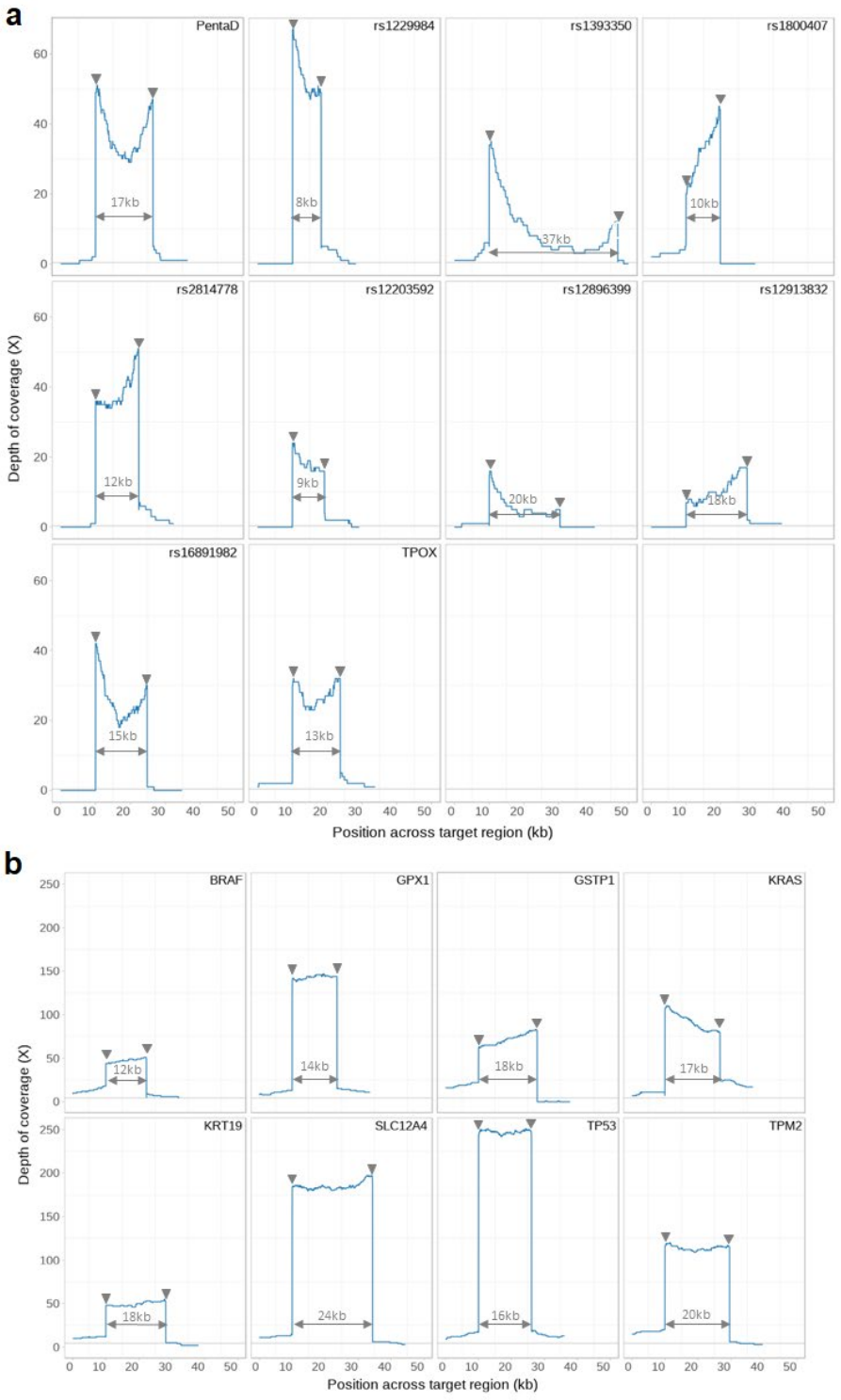


Figure 14. Coverage plots from the nCATS experiments.

(a) The ten regions of forensic relevance targeted with the Cas9-targeted Nanopore sequencing (nCATS) protocol. The details of these ten regions are given in Table 3. **(b)** Eight regions targeted with the nCATS protocol from Gilpatrick *et al.* (2020). The target locus is shown in the top right corner of each panel. For each of target regions, we show the number of reads (y axis) aligning at each position along the region (x axis). The boundaries and length of each region are also indicated. For the coverage, reads mapped in forward and reverse were considered.

2.2 Nanopore recapitulates bisulfite sequencing data at sites of low and high methylation

We applied METEORE pipeline to analyse the CpG methylation patterns using the reads corresponding to the targeted regions and then compared the results with existing WGBS data for NA12878 from the ENCODE project (Supplementary Data 13), which is one of the most used techniques to study CpG methylation at a genome-wide scale (Dunham *et al.*, 2012). Combining the methylation predictions from both on CpG sites showed an improved correlation for all tools compared with using the methylation prediction independently for each strand (Figure 15). Using the combined methylation from both strands, all tools showed a positive correlation with the WGBS signals (Table 10). METEORE (REG) combining Megalodon and DeepSignal achieved the highest Pearson correlation and lowest RMSE values (Table 10).

Table 10. Comparison of the Cas9-targeted nanopore sequencing (nCATS) data with whole genome bisulfite sequencing (WGBS) Illumina data.
(a) nCATS data from our study, targeting ten forensically relevant regions. **(b)** nCATS data from Gilpatrick *et al.* (2020). For each method we give the number of sites (*N*), the Pearson's correlation (*r*), coefficient of determination (r^2), the Spearman's rank correlation (ρ), and the root mean square error (RMSE) for the comparison of the percentage methylation predicted from Nanopore with the percentage methylation calculated from WGBS Illumina data. We show the results for the six tested tools and METEORE combining Megalodon and DeepSignal using a random forest (RF) (parameters: max_depth=3 and n_estimator=10) or a regression (REG) model.

a

	Our nCATS data				
	N	*r*	r^2	ρ	RMSE
Nanopolish	1704	0.8652	0.7485	0.8326	0.2248
DeepSignal	1731	0.9177	0.8423	0.8765	0.1708
Megalodon	1723	0.9117	0.8312	0.8801	0.1772
Tombo	1661	0.7765	0.6030	0.7537	0.2996
Guppy	1738	0.8513	0.7246	0.8316	0.2334
DeepMod	1739	0.7401	0.5477	0.7264	0.2874
METEORE (RF)	1723	0.9174	0.8416	0.8862	0.1829
METEORE (REG)	1723	0.9262	0.8579	0.8885	0.1607

Table 10 (continued).

b

nCATS data from Gilpatrick *et al.* (2020)					
	N	*r*	r^2	*ρ*	RMSE
Nanopolish	2171	0.9463	0.8954	0.8106	0.1490
DeepSignal	2171	0.9651	0.9315	0.8147	0.1252
Megalodon	2171	0.9665	0.9340	0.8201	0.1184
Tombo	2171	0.8694	0.7559	0.7326	0.2390
Guppy	2171	0.9377	0.8792	0.8004	0.2121
DeepMod	2171	0.8479	0.7190	0.7589	0.2585
METEORE (RF)	2171	0.9641	0.9294	0.8259	0.1307
METEORE (REG)	2171	0.9689	0.9387	0.8253	0.1155

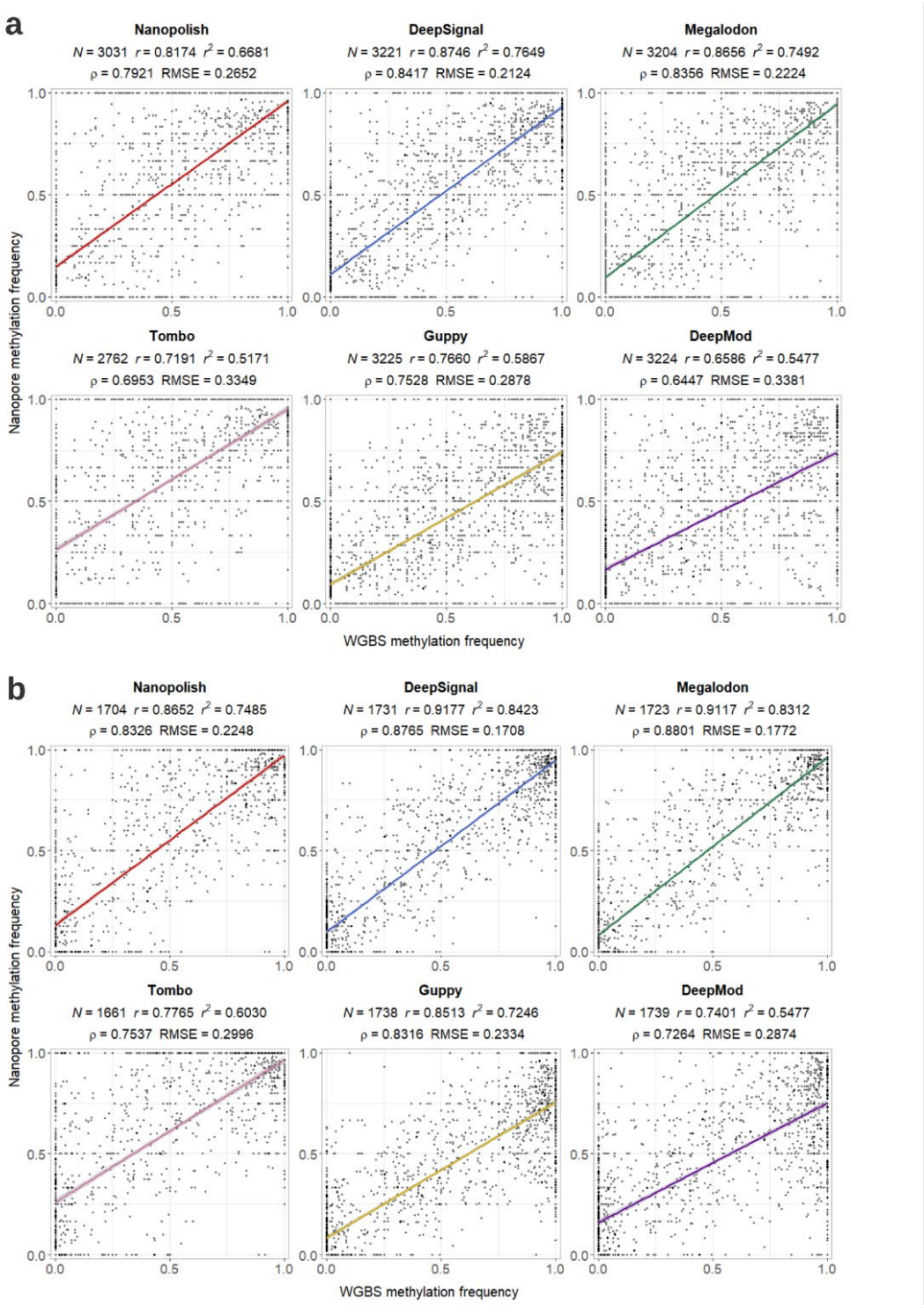


Figure 15. Scatter plots with regression lines added showing per-site performance of methylation detection for each tool.

For each tool, we show the methylation fraction predicted by each tool (y axis) and the fraction calculated from WGBS (x axis), using either **(a)** individual predictions on both strands or **(b)** combined predictions from both strands. The number of sites (*N*), the Pearson's correlation (*r*), coefficient of determination (r^2), the Spearman's rank correlation (ρ), and the root mean square error (RMSE) are provided for each tested tool. The plots include the correlation bands.

To compare the spread of methylation predictions, we categorized the WGBS data into three bins of increasing methylation frequency (Figure 16a). Tombo showed the largest dispersion of values at low (0.0-0.3) and intermediate (0.3-0.7) methylation, Guppy and DeepMod underpredicted at high methylation (0.7-1.0), and all methods overpredicted at intermediate methylation (Figure 16a).

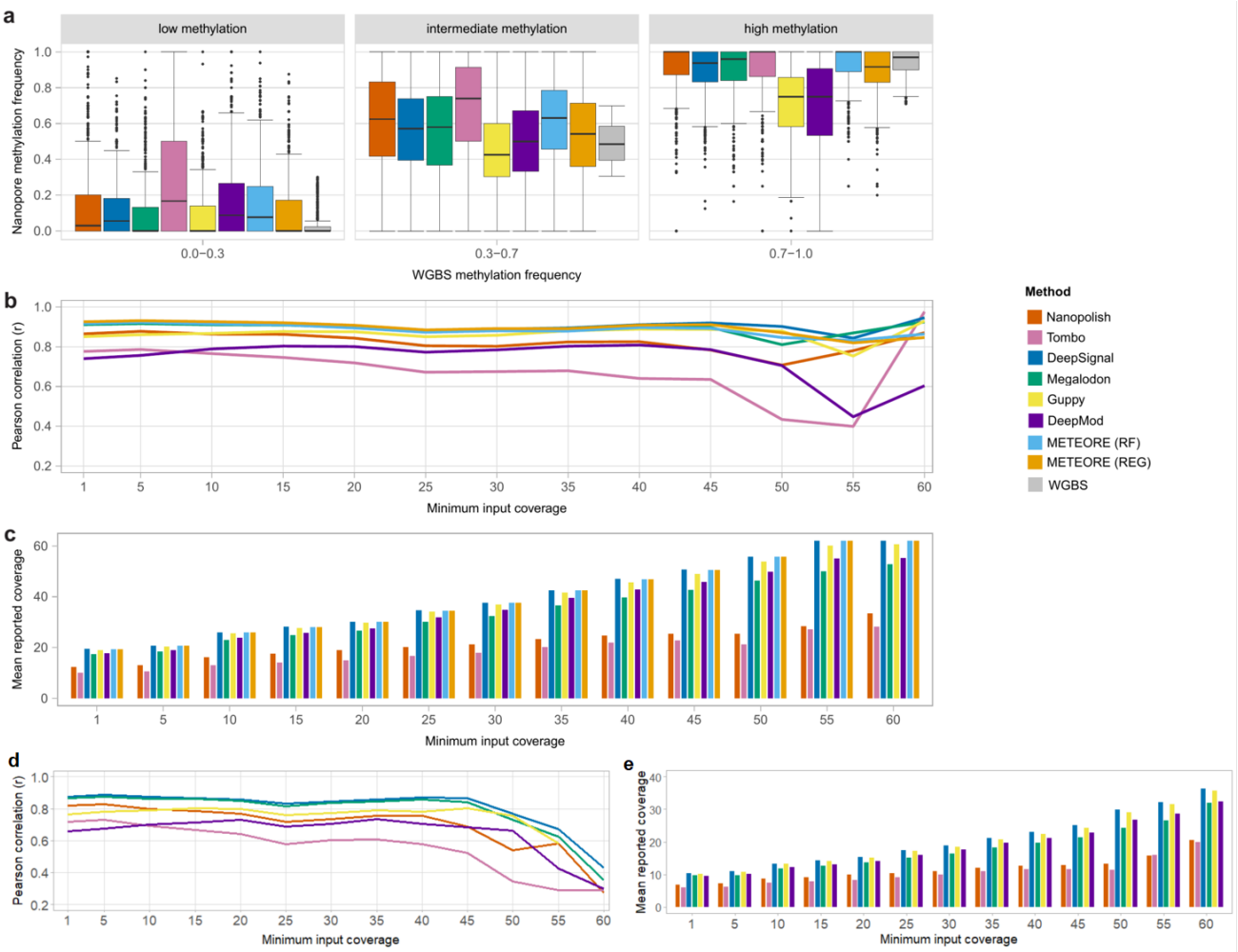


Figure 16. Summary of per-site performance comparing methylation predictions from nanopore with whole genome bisulfite sequencing (WGBS).

(a-c) Using combined predictions from both strands by averaging the methylation frequencies and adding up the coverage. **(d-e)** Using predictions independently on each strand. **(a)** Distribution of methylation calls (n = 1743) from Nanopore (y axis) across three WGBS methylation bins: no or low methylation (0.0–0.3) (n = 793), intermediate methylation (0.3–0.7) (n = 264), and high or full methylation (0.7–1.0) (n = 686). In the boxplots, the lower and upper boundaries of the box are the first and third quartiles of the data, respectively, with the median indicated by a thick black line. The lower and upper whiskers extend to 1.5 times the interquartile range. The outliers are represented by the black dots. **(b)** Pearson's correlation (*r*) (y axis) of each methylation calling tool with WGBS across different levels of minimal input coverage, i.e., minimum number of Nanopore reads

considered per site as reported from the BAM file (x axis). **(c)** Mean reported coverage (using the coverage reported by each tool for each site and then combining coverage from both strands) at each value of minimum input coverage in (b). METEORE (RF) is the combination of Megalodon and DeepSignal using a random forest (parameters: max_depth=3 and n_estimator=10), and METEORE (REG) is the combination of Megalodon and DeepSignal using a regression model. **(d)** Pearson's correlation (*r*) (y axis) of each methylation calling tool with WGBS across different levels of minimal input coverage, i.e., minimum number of Nanopore reads considered per site as reported from the BAM file (x axis). **(e)** Mean reported coverage (using the coverage reported by each tool for each site) considered at each value of minimum input coverage in (d). METEORE is not included since it performs predictions only combining both strands.

The same analyses on similar experimental datasets from eight other regions from Gilpatrick *et al.* (2020) (Figure 14b) (Supplementary Data 14) showed similar trends in terms of methylation predictions (Figure 17 & Figure 18). For this dataset, all methods achieved higher correlations and lower RMSE values with WGBS data (Table 10b), possibly because these regions contained more CpG sites with high or low methylation.

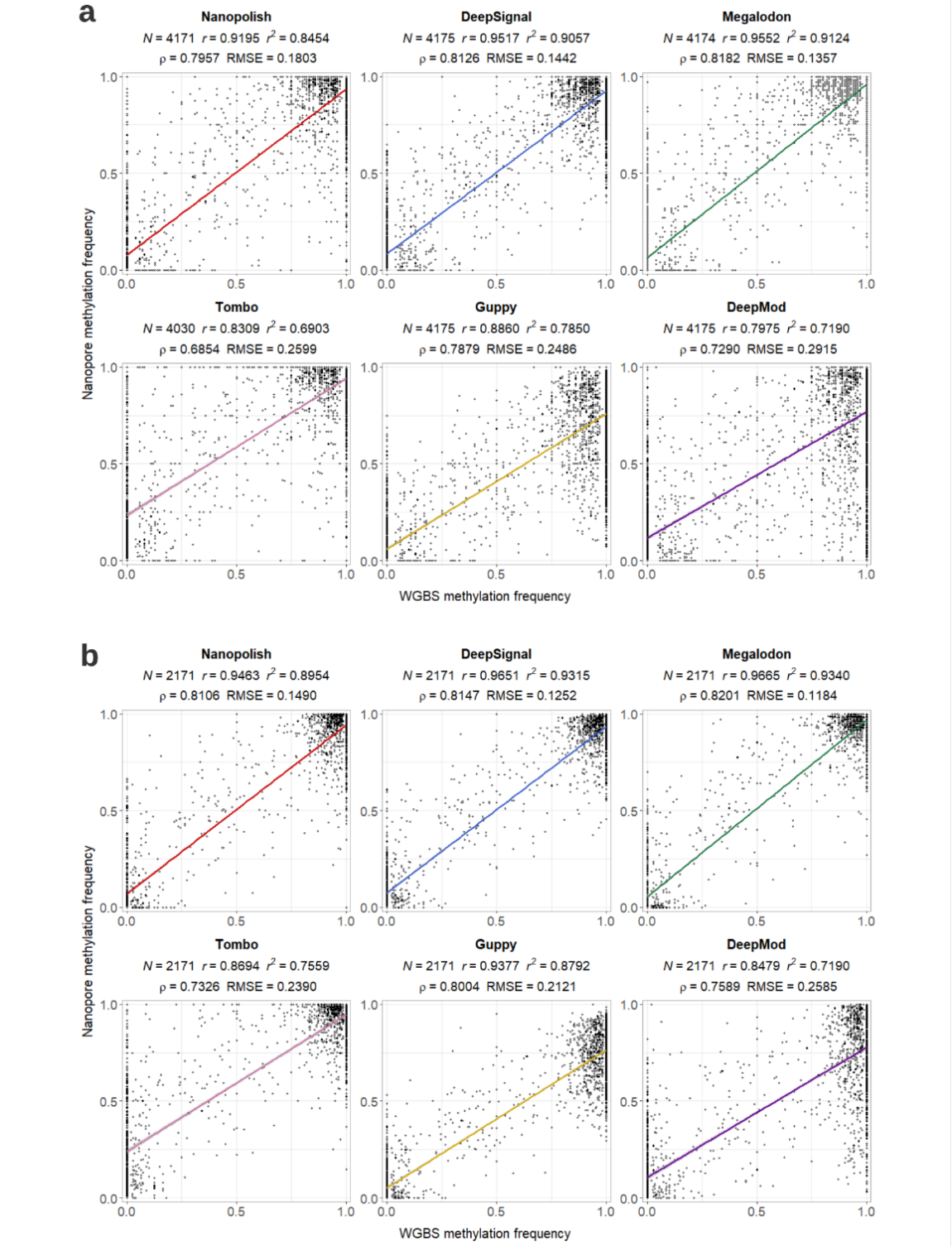


Figure 17. Scatter plots with regression lines added showing per-site performance of methylation detection for each tool, using Cas9-targeted nanopore sequencing (nCATS) data from Gilpatrick *et al.* (2020).
For each tool, we show the methylation fraction predicted by each tool (y axis) and the fraction calculated from WGBS (x axis), using either **(a)** individual predictions on both strands or **(b)** combined predictions from both strands. The number of sites (*N*), the Pearson's correlation (*r*), coefficient of determination (r^2), the Spearman's rank correlation (ρ), and the root mean square error (RMSE) are provided for each tested tool. The plots include the correlation bands.

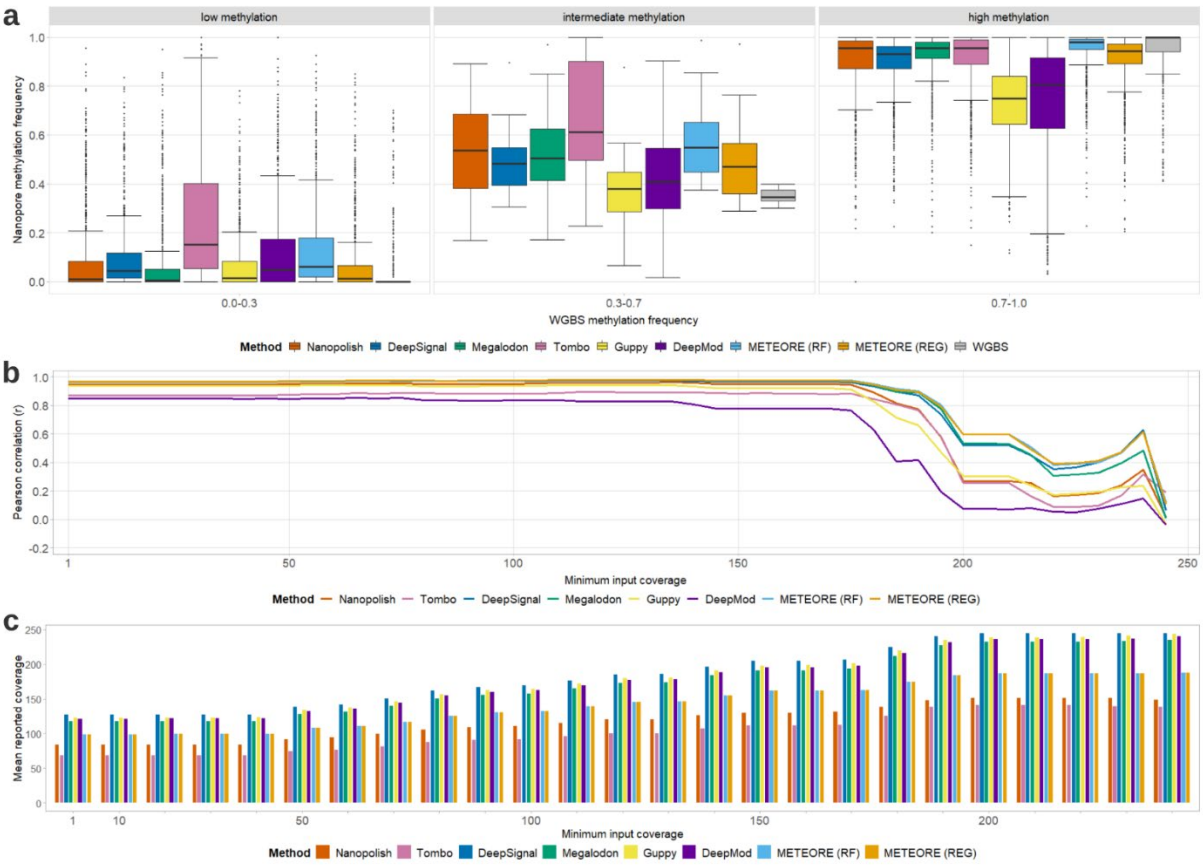


Figure 18. Summary of per-site performance comparing methylation predictions from nanopore with whole genome bisulfite sequencing (WGBS), using Cas9-targeted nanopore sequencing (nCATS) data from Gilpatrick *et al.* (2020) and combined predictions from both strands by averaging the methylation frequencies and adding up the coverage.

(a) Distribution of methylation calls (n = 1743) from Nanopore (y-axis) across three WGBS methylation bins: no or low methylation (0.0−0.3) (n = 793), intermediate methylation (0.3−0.7) (n = 264), and high or full methylation (0.7−1.0) (n = 686). In the boxplots, the lower and upper boundaries of the box are the first and third quartiles of the data, respectively, with the median indicated by a thick black line. The lower and upper whiskers extend to 1.5 times the interquartile range. The black dots represent the outliers. **(b)** Pearson's correlation (*r*) (y axis) oy-axis methylation calling tool with WGBS across different levels of minimal input coverage, i.e., minimum the number of Nanopore reads considered per site as reported from the BAM file (x axis). **(c)** Mean reported coverage at each value of minimum input coverage in (b). METEORE (RF) is the combination of Megalodon and DeepSignal using a random forest (parameters: max_depth=3 and n_estimator=10), and METEORE (REG) is the combination of Megalodon and DeepSignal using a regression model.

2.3 Methylation accuracy is stable at low coverage

Overall, correlations between the Nanopore methylation data and WGBS were approximately constant for all tools across different levels of coverage (Figure 16b & Figure 16d). Megalodon, DeepSignal and METEORE had correlations above 0.9, whereas Tombo had lower correlations across all coverage levels, which agreed with our findings that Tombo was less accurate at separating fully methylated from fully unmethylated sites (Figure 16b & Figure 16d). Moreover, both Nanopolish and Tombo generally reported much lower mean coverage of reads compared with other tools, consistent with their application of double cutoffs to discard reads (Figure 16c & Figure 16e).

The same analyses with data from Gilpatrick *et al.* (2020) also showed stable correlations with WGBS at most levels of coverage for all tools, with a marked drop at high minimum coverage (Figure 18b & Figure 18c). The general drop in accuracy at high minimum input coverage was due to the reduced number of sites available. To compensate for this, we subsampled reads to varying coverage levels (5x, 10x, 20x, 50x) using the same sites across all tools, but with different coverage for each level, and found that the correlation increased slightly with increased coverage (Table 11).

Table 11. Pearson correlations (*r*) of each methylation calling tool with WGBS at different coverage.
We subsampled 5x, 10x, 20x, 50x read coverage for 941 CpG sites, using Cas9-targeted Nanopore sequencing (nCATS) data from (Gilpatrick *et al.*, 2020).

	5x	10x	20x	50x
Nanopolish	0.7456	0.7994	0.8095	0.8145
DeepSignal	0.8065	0.8361	0.8438	0.8458
Megalodon	0.8247	0.8400	0.8514	0.8513
Guppy	0.7021	0.7469	0.7595	0.7696
Tombo	0.6245	0.6934	0.7105	0.7142
METEORE (RF)	0.8335	0.8378	0.8511	0.8509
METEORE (REG)	0.8326	0.8461	0.8551	0.8565

2.4 Nanopore recovers the methylation patterns along genomic regions

We compared the methylation profiles from WGBS and Nanopore along our ten tested regions. In general, all tools were consistent with the overall WGBS pattern independently of coverage and CG content (Figure 19). All tools recapitulated the known pattern of hypomethylation at CGIs (Chen *et al.*, 2011) (Figure 19). However, there were local inconsistencies with some of the tools. For instance, in the region spanning the first and second introns of *IPF4* (`chr6:392228-401463`), Guppy and DeepMod underpredicted the methylation frequency (Figure 19). In the region spanning the genes *ACKR1* and *CADM3* (`chr1:159199780-159212236`), Tombo overpredicted the methylation frequency, and Guppy failed to predict an increase in methylation described by WGBS and the other tools (Figure 19). At the *TPO* locus, the intermediate methylation at a CGI described by WGBS was recovered by all tools, except Tombo, Nanopolish and METEORE (RF) where they overpredicted (Figure 19). Using the eight different regions from Gilpatrick *et al.* (2020) we found similar results (Figure 20). For instance, Guppy and DeepMod underpredicted relative to WGBS in the region of *GPX1* promoter (`chr3: 49352525-49366169`), whereas Tombo overpredicted at a CGI in the same region (Figure 20).

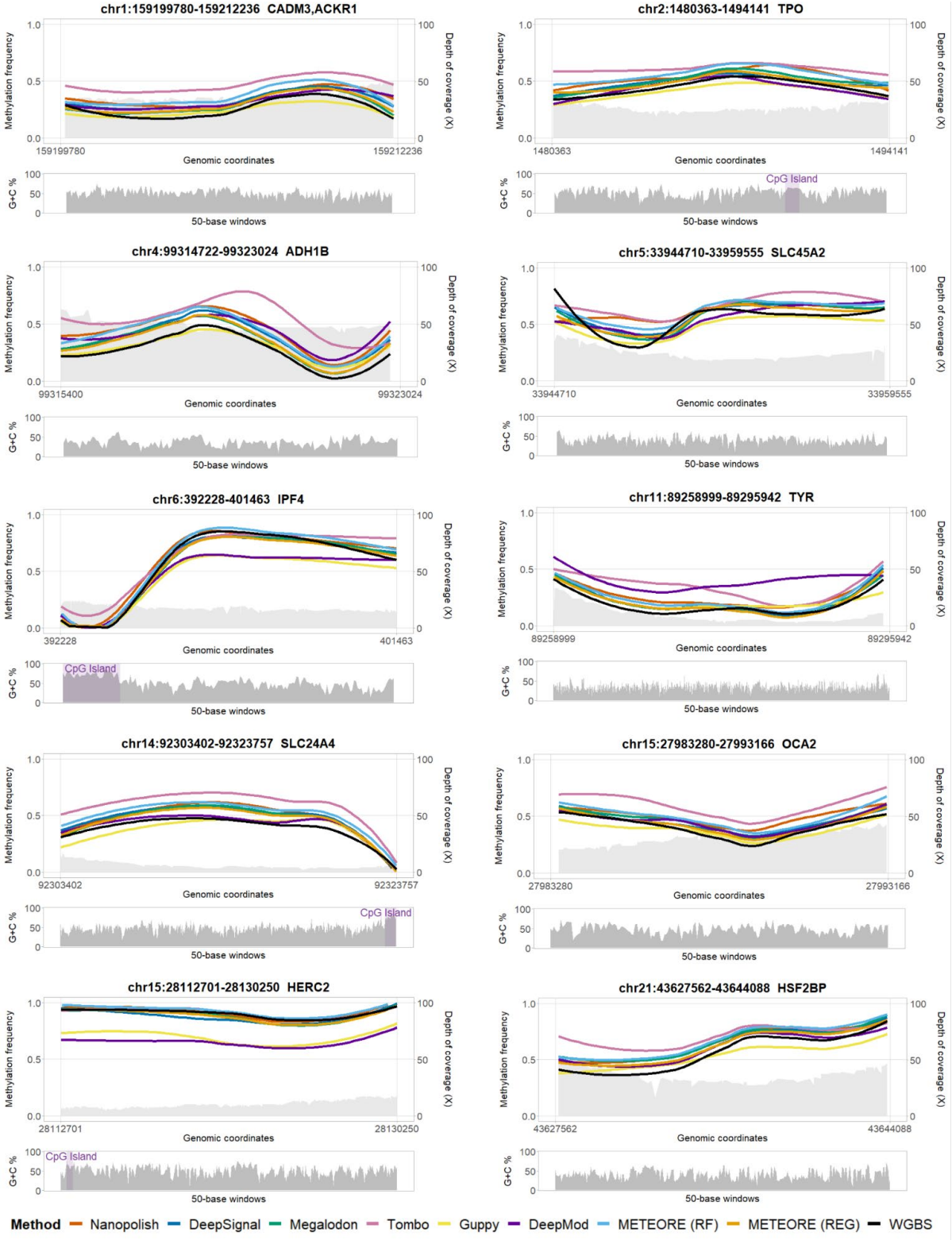


Figure 19. Locally Weighted Scatterplot Smoothing (LOESS) smoothing line plots of methylation calls for each method across our ten targeted regions using Cas9-targeted nanopore sequencing (nCATS).

We show METEORE with the combination of Megalodon and DeepSignal using either a random forest model (RF) or a regression model (REG). The plots include the Nanopore coverage, shown as a light grey area. Below each plot, we include the GC-content of the region.

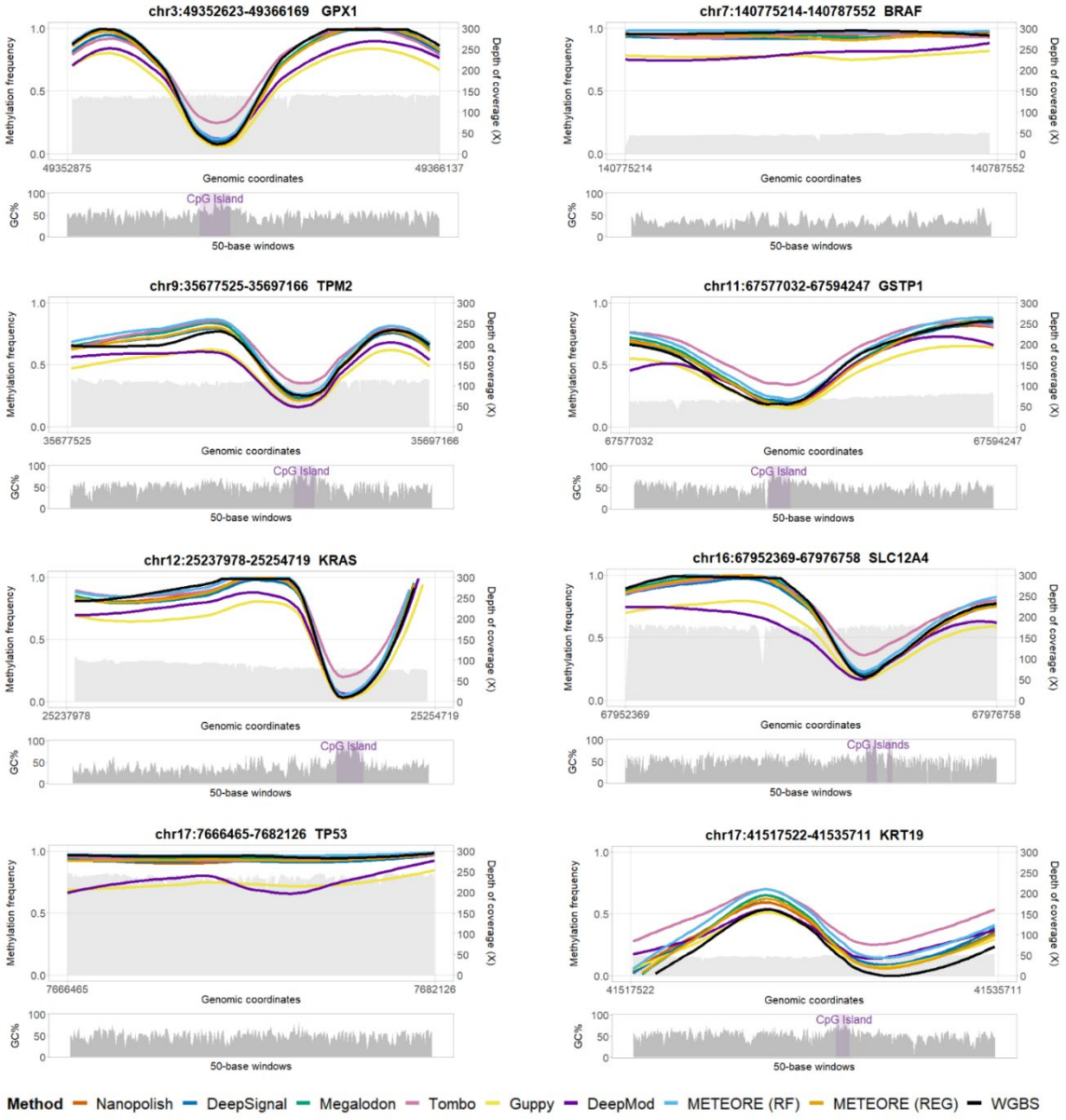


Figure 20. Locally Weighted Scatterplot Smoothing (LOESS) smoothing line plots of methylation calls for each method across the eight targeted regions using Cas9-targeted nanopore sequencing (nCATS) from Gilpatrick *et al.* (2020).

We show METEORE with the combination of Megalodon and DeepSignal using either a random forest model (RF) or a regression model (REG). The plots include the Nanopore coverage, shown as a light grey area. Below each plot, we include the GC-content of the region.

Next, we applied those thresholds derived before from the individual read analysis (Results section 1.4 and 1.5) for different methods on these two independent nCATS data. Using a single score cutoff based on the maximum value of TPR-FPR (Table 8), we observed slight

improvements in correlation and RMSE values for DeepSignal, Tombo and METEORE (RF) in both nCATS datasets (Table 12).

Table 12. Comparison of the Cas9-targeted nanopore sequencing (nCATS) data with whole genome bisulfite sequencing (WGBS) Illumina data, using single score thresholds obtained by the maximum value of (TPR-FPR). **(a)** Compare WGBS with the nCATS data from our study. **(b)** Compare WGBS with the nCATS data from Gilpatrick et al. (2020). We used the score cutoffs that maximized TPR-FPR in the mixture dataset 1 (Table 8). For each method we provide the number of sites (*N*), the Pearson's correlation (*r*), coefficient of determination (r^2), the Spearman's rank correlation (ρ), and the root mean square error (RMSE) for the comparison of the percentage methylation predicted from nanopore with the percentage methylation calculated from WGBS data. We show the results for five tested tools and METEORE combining DeepSignal and Megalodon using a random forest (RF) (parameters: max_depth=3 and n_estimator=10) or a regression (REG) model.

a

Our nCATS data					
	N	*r*	r^2	ρ	RMSE
Nanopolish	1724	0.8648	0.7478	0.8362	0.2171
DeepSignal	1731	0.9196	0.8456	0.8785	0.1693
Megalodon	1723	0.9040	0.8172	0.8753	0.1938
Tombo	1734	0.8037	0.6460	0.7871	0.2551
Guppy	1738	0.7706	0.5938	0.7741	0.2978
METEORE (RF)	1723	0.9217	0.8496	0.8878	0.1736
METEORE (REG)	1723	0.9164	0.8397	0.8871	0.1829

b

nCATS data from Gilpatrick et al. (2020)					
	N	*r*	r^2	ρ	RMSE
Nanopolish	2171	0.9265	0.8584	0.7943	0.2061
DeepSignal	2171	0.9654	0.9319	0.8149	0.1246
Megalodon	2171	0.9602	0.9219	0.8195	0.1372
Tombo	2171	0.8936	0.7986	0.7802	0.2343
Guppy	2171	0.9034	0.8162	0.7957	0.2148
METEORE (RF)	2171	0.9668	0.9348	0.8262	0.1221
METEORE (REG)	2171	0.9646	0.9305	0.8263	0.1271

Using the strategy of a double cutoff based on the removal of the 10% of reads that were closest to the score at which the FPR and 1-TPR curves crossed (Table 9), the concordance with WGBS improved for DeepSignal, Tombo and METEORE (RF) in both datasets, and for Nanopolish only in our experimental dataset (Table 13). The results showed signs of improvement in some methods such as DeepSignal, Tombo and METEORE (RF) by using an

alternative cutoff, whereas Nanopolish, Megalodon, Guppy and METEORE (REG) showed no improvement in accuracy. This might be due to the scores provided not being suitable for the prediction task (Figure 11), or to the high variability of the Nanopore signal across different sequence contexts. This prompted us to study the potential sequence biases associated with correct and incorrect predictions.

Table 13. Comparison of the Cas9-targeted nanopore sequencing (nCATS) data with whole genome bisulfite sequencing (WGBS) Illumina data, using the double cutoff obtained by discarding 10% of reads.
(a) Compare WGBS with the nCATS data from our study. **(b)** Compare WGBS with the nCATS data from Gilpatrick *et al.* (2020). For each site, we removed the 10% of reads with scores closest to the cross point of the FPR and 1-TPR curves in the mixture dataset 1 (Table 9). For each method, we provide the number of sites (*N*), the Pearson's correlation (*r*), the coefficient of determination (r^2), the Spearman's rank correlation (ρ), and the root mean square error (RMSE) for the comparison of the percentage methylation predicted from nanopore with the percentage methylation calculated from WGBS data. We show the results for five tested tools and METEORE combining DeepSignal and Megalodon using a random forest (RF) (parameters: max_depth=3 and n_estimator=10) or a regression (REG) model.

a

Our nCATS data					
	N	*r*	r^2	ρ	RMSE
Nanopolish	1621	0.8726	0.7614	0.8327	0.2165
DeepSignal	1731	0.9220	0.8500	0.8805	0.1698
Megalodon	1723	0.8786	0.7720	0.8643	0.2336
Tombo	1733	0.8150	0.6642	0.7924	0.2468
Guppy	1733	0.7345	0.5394	0.7275	0.3243
METEORE (RF)	1723	0.9185	0.8437	0.8874	0.1812
METEORE (REG)	1722	0.9167	0.8404	0.8917	0.1866

b

nCATS data from Gilpatrick et al. (2020)					
	N	*r*	r^2	ρ	RMSE
Nanopolish	2168	0.9376	0.8792	0.8128	0.1641
DeepSignal	2171	0.9677	0.9364	0.8176	0.1187
Megalodon	2171	0.9492	0.9010	0.8128	0.1632
Tombo	2171	0.9013	0.8123	0.7809	0.2177
Guppy	2171	0.8867	0.7862	0.7811	0.2248
METEORE (RF)	2171	0.9652	0.9317	0.8274	0.1276
METEORE (REG)	2171	0.9653	0.9318	0.8310	0.1253

2.5 Sequence context biases in the capability of methylation prediction

To explore whether the sequence context affects methylation prediction, we calculated the absolute differences between methylation calls from Nanopore reads by each tool and the methylation calculated from WGBS for each CpG site in an 8-mer context (`NNNCGNNN`). We observed that despite the variability between methods, there is a subset of k-mers where all methods have low or no discrepancy with WGBS, whereas for other k-mers all methods show a high discrepancy (Figure 21). Looking at the highest and lowest 40 8-mers according to the average discrepancy across methods, we observed the highest discordance between Nanopore and WGBS in an AT-rich context (Figure 22a). We grouped the k-mers according to the GC content to further investigate these biases. Nanopolish, DeepSignal, Megalodon, and METEORE showed similar agreement with WGBS across all GC-content, whereas Guppy and DeepMod underperformed compared to other tools, especially at low GC-content, and Tombo showed higher dispersion at intermediate GC content (Figure 21b).

To investigate the specific biases for each tool, we grouped the k-mers according to the low (=0) or high (>0.5) absolute difference with WGBS. By assessing the differences in the sequences of 'high' and 'low' k-mers for each tool, we observed that Nanopolish, Megalodon and METEORE showed no sequence biases (Figure 22). In contrast, DeepSignal showed a bias for high discrepancy for a T at the 1st position and an A at the 8th position of the 8-mer, Tombo and Guppy were affected by the residues at the 1st (A for Tombo and T for Guppy) and ,6th (T for Tombo and G for Guppy) positions, whereas DeepMod had many k-mers with a high discrepancy that with a significant bias for A and T (Figure 22).

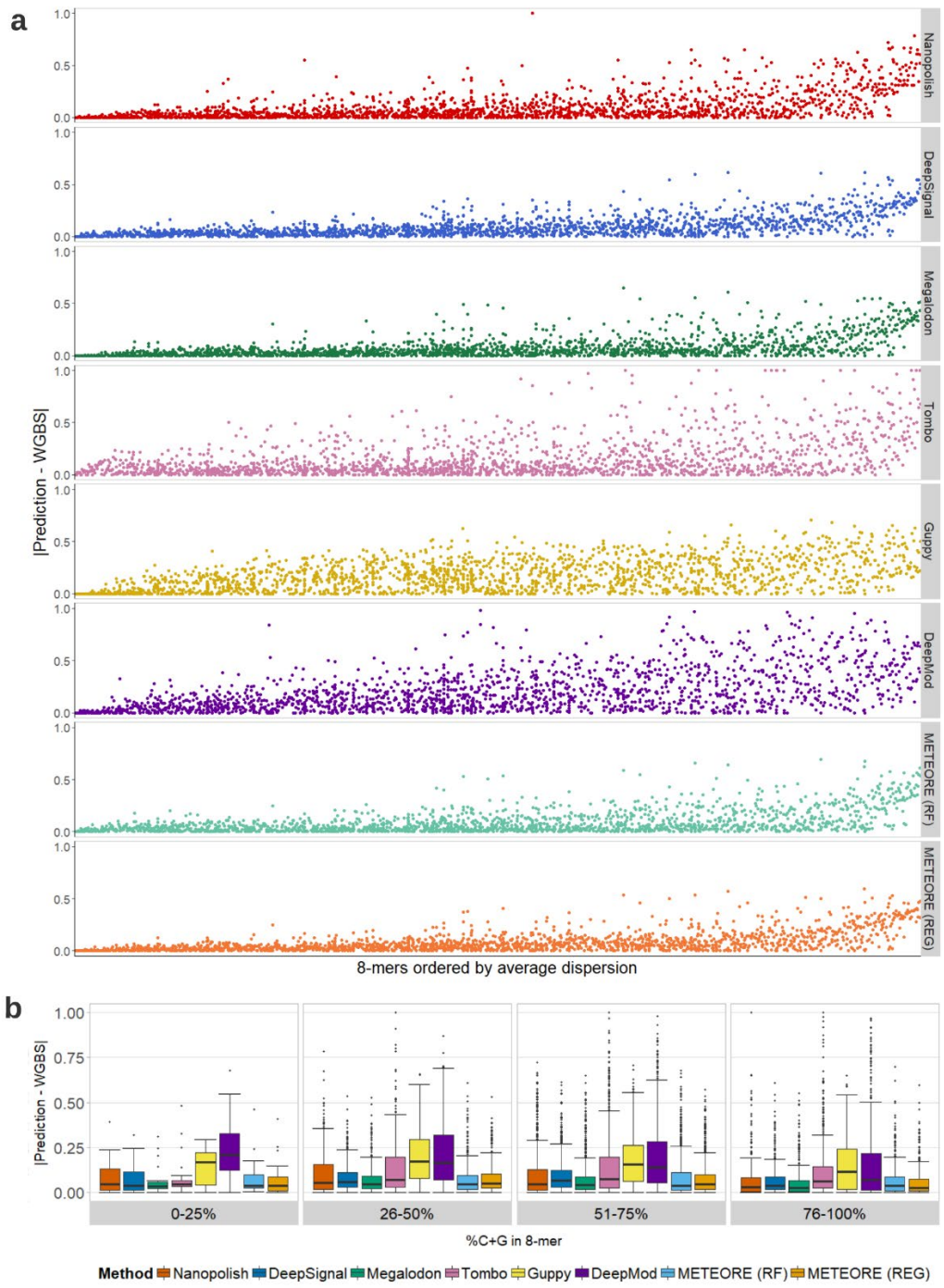


Figure 21. Sequence context analysis.

(a) Absolute difference between methylation frequencies from Nanopore-based methylation detection tools and whole genome bisulfite sequencing (WGBS) data for all 8-mers with a CpG in the middle (NNNCGNNN) (x axis) ordered from left to right in ascending order according to the average value of the absolute difference (y axis) across all tools. **(b)** Same absolute difference of methylation frequencies in 8-mers (n = 1911) from (a) (y axis) but stratified by the percentage of C and G residues in the 8-mers (0-25%: n = 20, 26-50%: n = 317, 51-75%: n = 1093 and 76-100%: n = 481), separated by tool. In the boxplots, the lower and upper boundaries of the box are the first and third quartiles of the data, respectively, with the median indicated by a thick black line. The lower and upper whiskers extend to 1.5 times the interquartile range. The outliers are represented by the black dots.

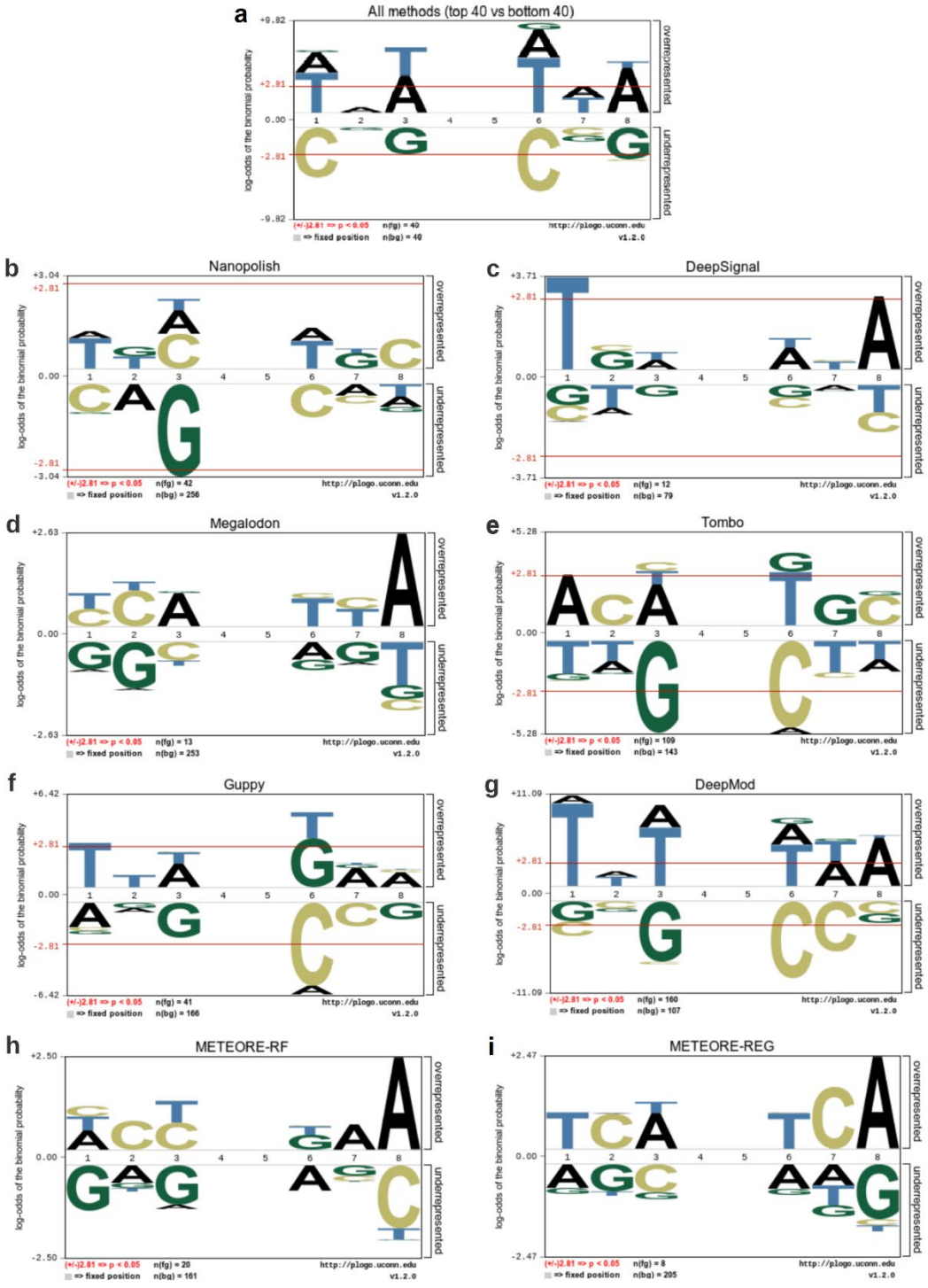


Figure 22. Sequence context associated to sites of high and low discrepancy with whole genome bisulfite sequencing (WGBS).

Here we show the pLogo plots for **(a)** the top/bottom 40 8-mers from the ranking in Figure 21a, **(b)** Nanopolish, **(c)** DeepSignal, **(d)** Megalodon, **(e)** Tombo, **(f)** Guppy, **(g)** DeepMod, **(h)** METEORE with random forest model (RF) and **(i)** METEORE with regression model (REG). **(a)** The bottom 40 8-mers were considered the foreground (fg) dataset, as shown in the upper panel, whereas the top 40 8-mers where considered the background (bg) dataset, as shown in the lower panel. The following residues from the foreground dataset are considered as statistically significant, i.e. p-value < 0.05, as calculated by pLogo: T at 1st, 3rd and 6th positions (*p*-values = 0.002, 0.025 and 3.943e-5 respectively) and A at 3rd, 6th and 8th positions (p-values = 0.003, 0.025 and 4.147e-4 respectively). **(b-**

i) For each method, CG-containing 8-mers (NNNCGNNN) were labelled as "low discrepancy" (fg dataset, as shown in the upper panel of each plot) if the discrepancy with WGBS (in absolute value) was >0.5, or as "high discrepancy" (bg dataset, as shown in the lower panel of each plot) if the discrepancy with WGBS was exactly 0. The pLogo Web tool uses binomial test to determine the statistical significance of the residues in the k-mers. One-tailed significance was determined by computing the area under the desired side (fg and bg datasets) of a probability distribution. Residues are scaled proportional to the log odds of the significance of foreground dataset (overrepresentation) versus the significance of background dataset (underrepresentation) (y axes) and stacked according to the statistical significance, with the most significant residues positioned closest to the x axis. The significance level (α) and the number of foreground and background sequences used, i.e., n(fg) and n(bg) values, are given at the bottom of each logo. The red horizontal lines correspond to α = 0.05 following Bonferroni correction. The following residues from the foreground dataset are considered as statistically significant, i.e., *p*-value < 0.05, as calculated by pLogo: T at 1st position (*p*-value = 0.006) and A at 8th position (*p*-value = 0.035) for DeepSignal; T at 1st position (*p*-value = 0.024) and G at 6th position (*p*-value = 0.013) for Guppy; T at 1st, 3rd and 6th position (*p*-values = 2.484e-9, 2.072e-6 and 2.540e-4 respectively) and A at 7th and 8th position (*p*-values = 0.026 and 4.110e-5 respectively) for DeepMod.

2.6 Increased coverage improves the ability to call SNPs

SNP analysis was performed in addition to methylation analysis by analysing on-target reads from the nCATS experiment, and variants were detected using BCFtools (Danecek *et al.*, 2021). with two different calling models: multiallelic and consensus. For validation, we used the variant calls from the 1000 Genome Project datasets on the GRCh38 reference assembly as ground truth for SNPs. There is a total of 4310 SNPs across the ten target loci from the 1000 Genome Project. All 4310 SNPs were detected with the consensus model (recall: 99.54%, *F*1 score: 99.54%) (Figure 23). On the other hands, the multiallelic model produced fewer false positive variant calls than the consensus model but missed 25 SNPs (*F*1 score: 99.34%) (Figure 23).

We further explored how the increased coverage from the nCATS protocol would affect the ability to call variants from nanopore data, so we subsampled the variant calls stratified by coverage. The higher the coverage, the better the recall and the precision, as higher sequence coverage makes it easier to generate a consensus sequence. We found that a good correlation between coverage and the performance of calling SNPs from nanopore data using both

BCFtools calling models. A minimum coverage of 35x is required to achieve 100% recall and precision (Figure 23).

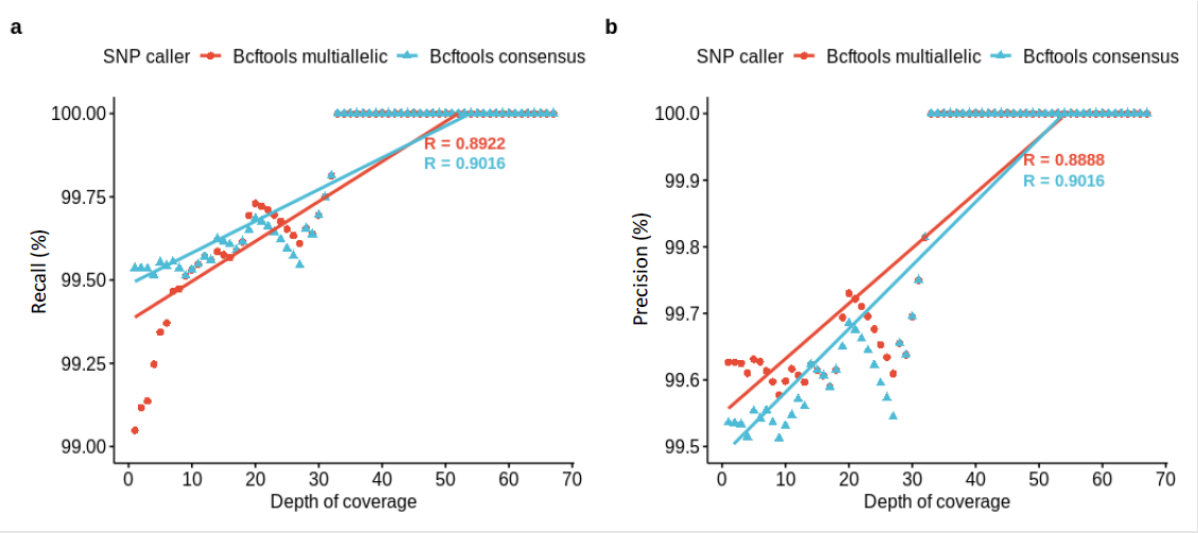


Figure 23. SNP calling performance stratified by coverage.
We called SNPs using BCFtools with two calling models: multiallelic and consensus. **(a)** Recall = true positives / all SNPs from the 1000 Genome project. **(b)** Precision = true positives / all predicted SNPs

.

3 DNA methylation profiling and SNP genotyping of forensically relevant loci with nanopore adaptive sampling

The Cas9-targeted Nanopore sequencing (nCATS) protocol allows the generation of high coverage sequencing data and the detection of SNPs and methylation. However, it requires specialized reagents such as Cas9 and gRNA, and the time required to design gRNAs for every target locus can be a limitation, with at least two sets of gRNA pairs used to improve the reliability of site coverage (Gilpatrick *et al.*, 2020). nCATS is restricted to the number of target regions in a single panel (panels of up to ~100 crRNAs are considered to exhibit acceptable cut kinetics) and the size of the regions (up to 50 kb long). Adaptive sampling, also known as ReadUntil, represents a promising alternative to nCATS for targeted nanopore sequencing. ONT sequencing devices can be programmed to recognize target DNA molecules through real-time basecalling, rapid mapping, and selectively ejecting reads that are identified as off-target during a sequencing experiment. This method allows computational enrichment of the target regions and sequencing of native DNA molecules harbouring base modifications without requiring any further processing beyond standard nanopore library preparation. In this section, we demonstrate that adaptive sampling can be used to profile DNA methylation in human samples without complex sample processing.

3.1 Evaluate the use of adaptive sampling with control samples

Adaptive sampling is a novel way of sequencing ROI but is largely untested for forensic studies. To evaluate the efficacy of this technology, we targeted 165 AISNPs from the TFS Precision ID Ancestry panel, 24 PISNPs from the HIrisPlex system, 36 STR loci commonly in forensic DNA profiling, 10 genes commonly included in forensic age prediction models, 307 ageing-related human genes from the Ageing Gene Database (Tacutu *et al.*, 2018), 393 orthologous genes showing conserved time-dependent behavior across species (Wang *et al.*, 2020), 31 genes associated with body fluid identification, 3 tandem copies of the rDNA unit and some innate immune response genes. The entire locus for each gene was targeted, including 20 kb of flanking sequencing in either direction. The entire panel covered about 113 Mb in total, targeting approximately 3.65% of the human reference human genome (hg38).

To examine the target enrichment performance, adaptive sampling was performed on the GIAB reference sample HG002. Accepted reads show a bimodal distribution with two peaks, where one matches our desired target size of about 40kb, and the other peak indicates the presence of short reads around 500 bases long (Figure 24a). Conversely, rejected reads show a sharp and narrow peak around 500 bases. The presence of short, rejected reads in our sequencing data is because those reads are ejected from the pores after sequencing around 500 bases. It is important to note that the nanopore sequencer still identifies DNA fragments that are not of interest. While these are ejected, a short fragment is still sequenced. There are also reads with the size ranges from 100 to about 2,500 bases, that the software cannot decide whether they belong to the target or not, so these reads get sequenced before any decision is made.

After the experiment, all nanopore data were re-basecalled using Guppy with the HAC basecalling model, followed by aligning the reads to the human reference genome (hg38). Off-

target reads produce a sharp peak of short reads, with N50 of 500 bps compared to on-target reads with N50 of 28.6 kb (Figure 24b). Moreover, a median 6-fold enrichment in sequencing depth within target regions was granted by adaptive sampling (Figure 24c). This shows that DNA fragments outside the target regions were successfully rejected while the DNA fragments within the target regions were fully sequenced (Figure 24b & c).

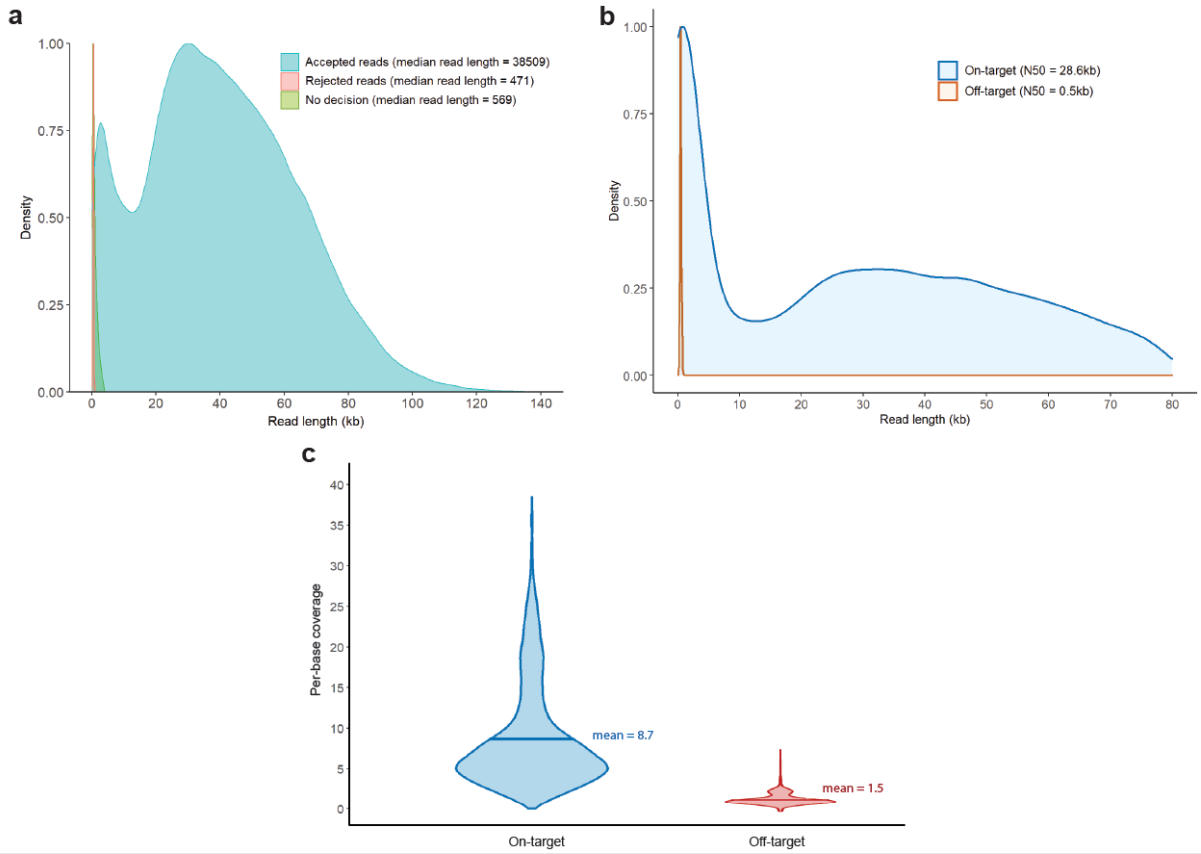


Figure 24. Targeted nanopore sequencing with adaptive sampling on the reference DNA sample HG002. (a**)** Density plot showing read length distribution for different decisions made by the MinKNOW software. **(b)** Density plot showing read length distribution for on-target (blue; N50 = 28.6 kb) versus off-target (orange; N50 = 0.5 kb) alignments. **(c)** Violin plots showing per-base coverage distributions within on-target regions (blue) and within randomly selected off-target regions (red). A horizontal line is added to each violin plot to show the mean coverage of 8.7 and 1.5.

3.2 Confirm the methylation status of control samples with adaptive sampling

Prior to profiling methylation on a human sample, control samples with high and low methylation were sequenced to determine if we could confirm the methylation status of the control samples using nanopore adaptive sampling. The high and low methylation samples serve as the +'ve and –'ve controls respectively in these experiments to assess technology accuracy. The control samples are derived from whole blood and treated enzymatically, resulting in greater than 85% methylation in the high methylation standard and less than 5% methylation in the low methylation standard, according to the manufacturer (EpigenDX, 2018). Megalodon was selected to detect methylation after it demonstrated the highest performance in our benchmarking studies (see Chapter IV - 1 & 2).

We first assessed the sequence context of CpG sites detected from our data. For both samples, bases around the CpG sites were evenly distributed at each position in the 8-mers, indicating no sequence biases in methylation predictions with nanopore sequencing (Figure 25a & b). The overall methylation level was consistent with the expected methylation level from each control sample, where 90% of CpG sites are fully methylated in the positive control and ~95% of sites are fully unmethylated in the negative control (Figure 25c).

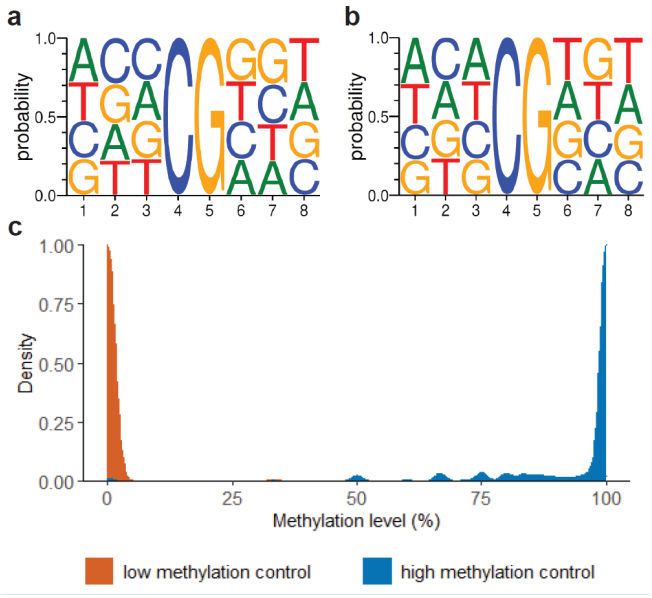


Figure 25. Methylation control samples.
Sequence motifs in DNA sequences in 8-mer context with a CG in the middle (NNNCGNNN) for **(a)** high methylation control and **(b)** low methylation control. **(c)** Density plot showing methylation level distribution for both controls. For the low methylation control, the mean and median methylation are 3.4 and 0 respectively (N = 113,606). For the high methylation control, the mean and median methylation are 94.3 and 100 respectively (N = 207,599). The expected methylation level claimed by EpigenDX for high and low methylation controls are >85 and <5 respectively.

To further examine if CpG methylation can be accurately detected using nanopore adaptive sampling, we included in the target regions three copies of a tandem repeat of the rDNA unit, each containing two units in three different orientations. We could then take advantage of the adaptive sampling technology to enrich and sequence individual rDNA copies as well as check the methylation status using the same methylation control samples. First, we assessed the per-read modified and canonical base scores for both controls. A majority of the CpG sites in rDNA-containing reads showed a per-read modified base score of 0.9 or higher and canonical base score of 0.2 or lower in the high-methylation control sample (Figure 26a & b). In contrast, most of these sites showed a per-read modified base score lower than 0.25 and canonical base score higher than 0.75 in the low-methylation control sample (Figure 26c & d). Additionally, the percentage methylation quantification along the whole rDNA sequence was consistent with the expected methylation levels in both control samples (Figure 26e). Thus, while the rDNA is

composed of high GC content, especially in the 18S and 28S rRNA genes and the intergenic spacer (IGS), nanopore can accurately identify the methylation levels in the control samples.

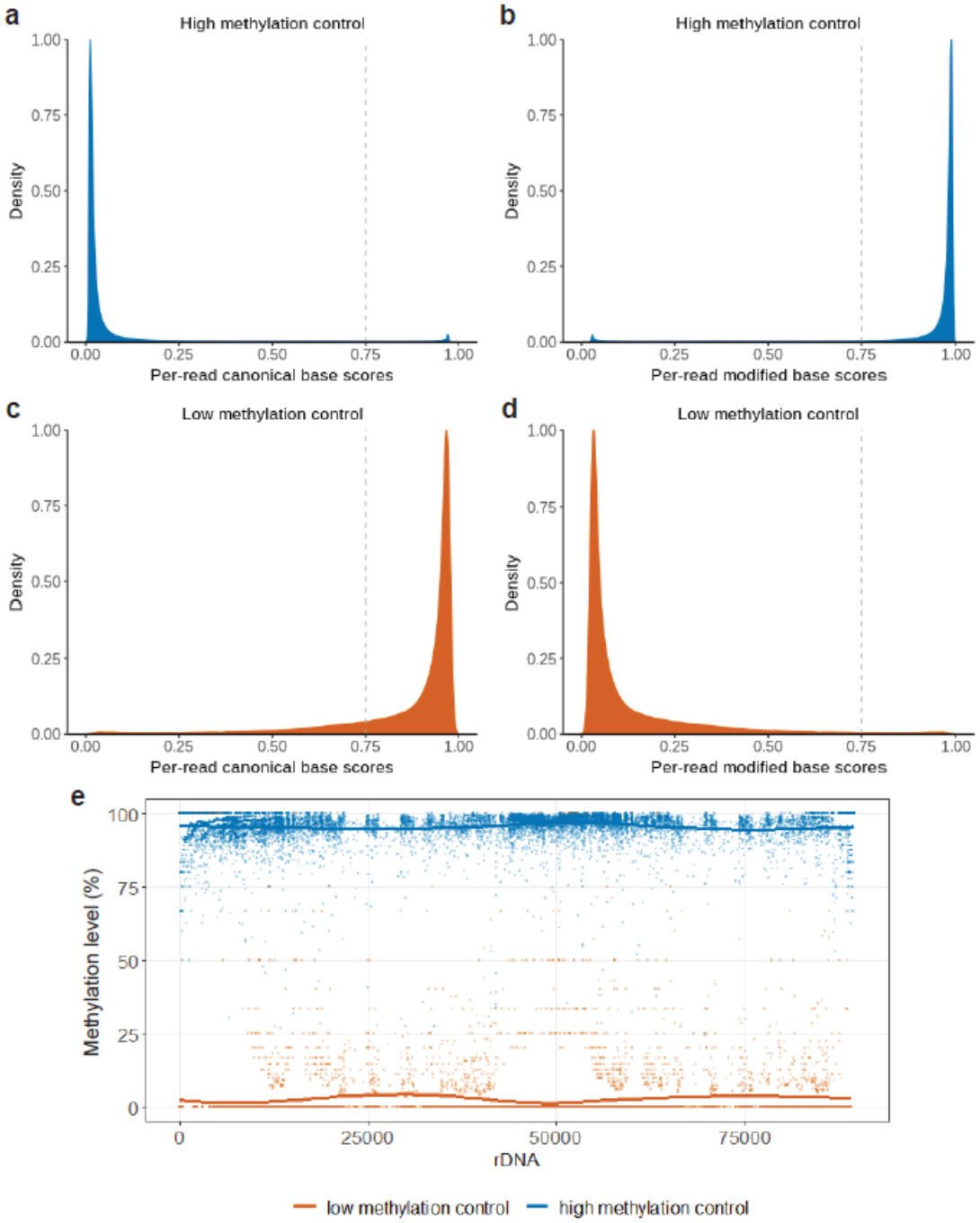


Figure 26. Methylation profiling of ribosomal DNA (rDNA) using nanopore sequencing data that was generated from adaptive sampling.

(a-b) The per-read **(a)** canonical and **(b)** modified base scores for high methylation control. **(c-d)** The per-read **(c)** canonical and **(d)** modified base scores for low methylation control. A dashed line is added in each plot to indicate a threshold of 0.75 which is the threshold used in Megalodon by default to consider a canonical/modified base. **(e)** Methylation levels across the target rDNA sequence for both controls.

3.3 SNP genotyping performance

The scope of forensic studies using nanopore sequencing technology has been limited because of its high error rate compared to short-read technologies, as the basecaller cannot distinguish between some k-mers and failing to discriminate signal events in homopolymeric regions. Using the data generated with adaptive sampling, we validated SNP calls using the reference DNA sample HG002 and compared the resulting SNP calls to the Illumina data from the GIAB project (GRCh38 v4.2.1) as the ground truth, where a total of 6,216 annotated SNPs were benchmarked across our target regions. We then calculated the numbers of predicted SNPs, correct and incorrect SNP calls, followed by various evaluation metrics, including recall, precision and F1 score.

In the previous chapter, we only benchmarked the variant-calling performance of BCFtools on our nCATS datasest. Here we compared the SNP calling performance with five existing variant callers: 1) BCFtools, which is used along with Samtools to generate genotype likelihoods from the alignment data (Danecek *et al.*, 2021); 2) GATK, which is designed for calling variants in short Illumina reads (Van der Auwera & O'Connor, 2020); 3) Nanopolish, which utilises an HMM to interrogate the raw signals from the fast5 file and the alignment data from the bam file (Simpson *et al.*, 2017); 4) Longshot, which calls variants in long reads using a pair-HMM for a short window covering the SNP locus (Edge & Bansal, 2019) ; and 5) PEPPER-Margin-DeepVariant, which is nanopore-based variant calling pipeline using neural networks and HMM (Shafin *et al.*, 2021). SNPs were called on the HG002 sample using the variant calling tool as mentioned above, with 15× mean target coverage across the cohort (Figure 27). PEPPER-Margin-DeepVariant and Longshot outperformed the other tested callers, with the highest F1 score achieved by PEPPER-Margin-DeepVariant, closely followed by Longshot (Figure 27). The significance of this is that even though it yields less coverage (~15×) than the

biochemical method, nanopore adaptive sampling achieves accurate SNP detection, predicting >98% of the known SNPs with >94% recall using nanopore-specific variant calling tools such as PEPPER-Margin-DeepVariant and Longshot. (Figure 27). Nanopolish, however, exhibited reduced performance in our study (Figure 27).

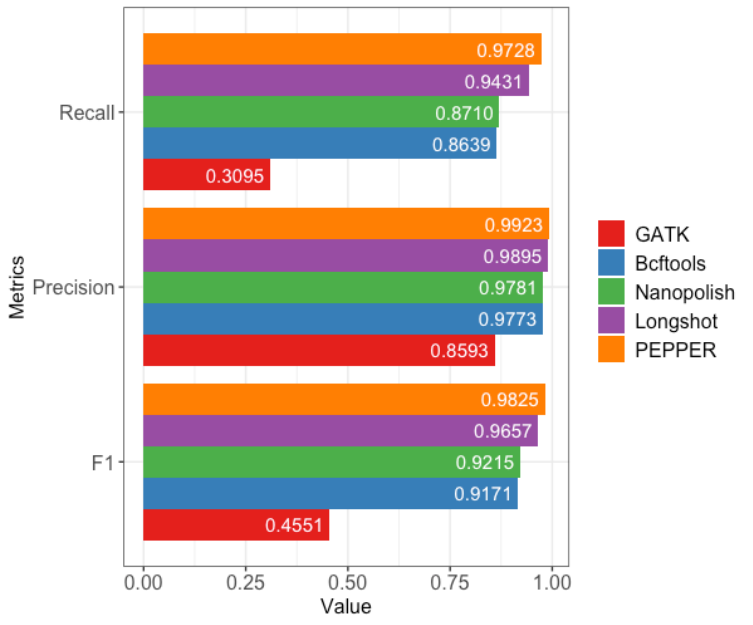


Figure 27. Variant calling performance of different methods on HG002 sample.
We benchmarked single nucleotide polymorphisms (SNPs) over the enriched regions of our panel without ribosomal DNA and sex chromosomes, which have a total of 6,216 annotated SNPs in the Genome in a bottle (GIAB) dataset for HG002. Recall = true positives / (true positives + false negatives); Precision = true positives / (true positives + false positives); F1-score = (2 × precision × recall) / (precision + recall)

3.4 Validation of age-associated methylation markers with adaptive sampling

Up until now, many forensic age prediction models (APMs) were based on the detection of DNA methylation levels using a few individual CpG sites that are known to be associated with age. For example, the gene *ELOVL2* has been identified as a promising marker for age estimation, but there are only a couple of CpG sites in this gene that have been commonly

included in APMs. Also, APMs have been developed using a broad spectrum of technologies with a targeted approach. However, the utilization of Nanopore sequencing in developing such models for humans is still lacking, except for the recent development of the first epigenetic clock for cattle using nanopore sequencing (Hayes *et al.*, 2021). Furthermore, the existing APMs have not been validated using Nanopore sequencing. We used nanopore adaptive sampling to target the genes associated with age, where we investigated all CpG sites found in the gene rather than a few single CpG sites.

Ten DNA samples from blood between 25 and 76 years old were analysed using targeted nanopore sequencing with an adaptive sampling approach for all CpG sites from eight genes commonly included in forensic age prediction models (*ELOVL2*, *FHL2*, *MIR29B2CHG*, *CCDC102B*, *ASPA*, *KLF14*, *TRIM59* and *PDE4C*), 307 ageing-related human genes from the Ageing Gene Database (Tacutu *et al.*, 2018), 393 orthologous genes showing conserved time-dependent behaviour across species (Wang *et al.*, 2020). Simple linear regression analyses were performed between chronological age and methylation levels of each CpG site (Supplementary data 16). New markers can then be identified through this analysis.

We first tested whether age markers used in previous APMs could be validated in our experiments (Table 14). Simple linear regression between chronological age and methylation levels of each CpG site revealed that these common sites have a weak correlation between methylation and age ($r^2 < 0.5$), except for `chr6:11044634`, `chr6:11044642`, `chr6:11044644`, `chr6:11044647`, `chr6:11044655`, `chr1:207823681`, `chr1:207823715`, `chr2:105399282`. It is possible that the number of samples available is not sufficient to recover the age-predictive nature of those sites through simple regression. Increasing the number of samples could improve the linear regression model and further increase the R-squared and reduce RMSE and mean absolute error (MAE).

It has been demonstrated that the ELOVL2 gene is an excellent predictor of chronological age in blood (Jung *et al.*, 2019; Zbieć-Piekarska *et al.*, 2015a), and multiple CpG sites (chr6:11044628, chr6:11044631, chr6:11044634, chr6:11044640, chr6:11044642, chr6:11044644, chr6:11044647, chr6:11044655, chr6:11044661) in this gene have been analysed and used in different APMs. We found that these CpG sites correlate with age showing differences between CpG sites (r^2 varies between 0.1089 and 0.7444) (Table 14). Again, a limited number of samples could cause such varied performance. Nonetheless, the results show that not just a single CpG site from an age-associated gene but rather multiple sites that are close to each other could be informative for age prediction. In this regard, an alternative site could be used to add information or to compensate for a site that cannot be used due to the absence of reads.

Current techniques for age-related DNA methylation analysis use either a genome-wide or a gene-specific approach, depending on the number of CpG sites analysed. For instance, Horvath (2013) introduced an epigenetic clock using a total of 353 CpG sites across a broad spectrum of human tissues and cell types. On the other hand, the established forensic APMs to date are based on a small number of CpG sites, mainly in blood samples, but also in other tissues including saliva and semen. DNA methylation detection in these studies was done either by pyrosequencing, SNaPshot, EpiTYPER, or MPS. A shared characteristic of these methylation detection technologies is their dependency on bisulfite conversion which is a complex workflow. The ability of nanopore adaptive sampling to simultaneously detect modified bases without the use of bisulfite conversion while targeting a large number of loci makes it ideal for validating a high number of CpG sites and detecting methylation in a simpler and less expensive approach.

Table 14. Performance of DNA methylation markers commonly used in the age prediction model (APM). These CpG sites are commonly used and included in forensic APMs and located at eight known genes (*ELOVL2*, *FHL2*, *MIR29B2CHG*, *CCDC102B*, *KLF14*, *TRIM59* and *PDE4C*), with the exact chromosomal location (GRCh38/hg38) provided. Simple linear regression was performed between the chronological age of the individuals and methylation levels for each individual CpG site. The adj r^2 = adjusted R-squared; MAE = mean absolute error; RMSE = root mean square error.

Age markers	Associated genes	Adj r^2	*p*-value	MAE	RMSE	Included in APM(s)
chr6:11044628	ELOVL2	0.3392	0.0452	11.8537	16.0843	Bekaert *et al.* (2015); Feng *et al.* (2018); Freire-Aradas *et al.* (2018); Jung *et al.* (2019); Lee *et al.* (2022)
chr6:11044631	ELOVL2	0.4302	0.0235	12.0617	14.9357	Feng *et al.* (2018); Freire-Aradas *et al.* (2018)
chr6:11044634	ELOVL2	0.7032	0.0015	8.6720	10.7798	Bekaert *et al.* (2015); Feng *et al.* (2018); Fleckhaus and Schneider (2020); Freire-Aradas *et al.* (2018); Woźniak *et al.* (2021); Zbieć-Piekarska *et al.* (2015a)
chr6:11044640	ELOVL2	0.1089	0.1854	16.8062	18.6776	Bekaert *et al.* (2015); Dias *et al.* (2020); Woźniak *et al.* (2021); Zbieć-Piekarska *et al.* (2015a)
chr6:11044642	ELOVL2	0.5867	0.0059	11.0218	12.7204	Bekaert *et al.* (2015); Fleckhaus and Schneider (2020); Zbieć-Piekarska *et al.* (2015a)
chr6:11044644	ELOVL2	0.5717	0.0069	11.558	12.9485	Bekaert *et al.* (2015); Naue *et al.* (2017); Naue *et al.* (2018); Zbieć-Piekarska *et al.* (2015a)
chr6:11044647	ELOVL2	0.7444	0.0008	7.0109	10.0026	Bekaert *et al.* (2015); Zbieć-Piekarska *et al.* (2015a)
chr6:11044655	ELOVL2	0.7393	0.0009	7.8912	10.1027	Bekaert *et al.* (2015); Zbieć-Piekarska *et al.* (2015a)
chr6:11044661	ELOVL2	0.3392	0.0452	11.8537	16.0843	Fleckhaus and Schneider (2020); Zbieć-Piekarska *et al.* (2015a)
chr1:207823681	MIR29B2CHG	0.8369	0.0001	6.4979	7.9894	Dias *et al.* (2020); Jung *et al.* (2019); Lee *et al.* (2022); Woźniak *et al.* (2021); Zbieć-Piekarska *et al.* (2015b)
chr1:207823715	MIR29B2CHG	0.7065	0.0014	9.6897	10.7199	Feng *et al.* (2018); Fleckhaus and Schneider (2020)
chr2:105399282	FHL2	0.7279	0.0010	8.6374	10.3214	Fleckhaus and Schneider (2020); Freire-Aradas *et al.* (2018); Freire-Aradas *et al.* (2016); Jung *et al.* (2019); Lee *et al.* (2022); Woźniak *et al.* (2021)
chr2:105399288	FHL2	0.4993	0.0134	9.9815	14.0004	Dias *et al.* (2020); Zbieć-Piekarska *et al.* (2015b)
chr18:68722210	CCDC102B	0.2149	0.0997	16.1506	17.5308	Feng *et al.* (2018); Freire-Aradas *et al.* (2018)
chr7:130734355	KLF14	0.1777	0.1245	16.7061	17.9418	Fleckhaus and Schneider (2020); Freire-Aradas *et al.* (2016); Jung *et al.* (2019); Lee *et al.* (2022); Zbieć-Piekarska *et al.* (2015b)
chr3:160450189	TRIM59	0.4741	0.0166	11.9448	14.3482	Feng *et al.* (2018); Freire-Aradas *et al.* (2016); Jung *et al.* (2019); Lee *et al.* (2022); Naue *et al.* (2017); Naue *et al.* (2018)

Table 14 (continued).

Age markers	Associated genes	Adj r^2	*p*-value	MAE	RMSE	Included in APM(s)
chr3:160450199	TRIM59	0.3363	0.0461	12.1928	16.1194	Feng *et al.* (2018); Zbieć-Piekarska *et al.* (2015b)
chr19:18233127	PDE4C	-0.0692	0.5362	19.4047	20.4587	Feng *et al.* (2018); Fleckhaus and Schneider (2020); Freire-Aradas *et al.* (2018); Woźniak *et al.* (2021)
chr19:18233131	PDE4C	-0.048	0.4652	19.1477	20.2548	Feng *et al.* (2018); Fleckhaus and Schneider (2020); Freire-Aradas *et al.* (2018)
chr19:18233133	PDE4C	-0.0859	0.6058	19.6377	20.6176	Feng *et al.* (2018); Fleckhaus and Schneider (2020); Freire-Aradas *et al.* (2018)

CHAPTER V: DISCUSSION & CONCLUSIONS

1.1 The seven tested analytic tools exhibit different performances of methylation status prediction

Our systematic benchmarking of DNA methylation prediction from Nanopore sequencing using individual reads, controlled methylation mixtures, Cas9-targeted sequencing, and WGBS, indicated that no single method generates correct predictions across all ranges of methylation frequency. Extreme cases were Guppy, which correctly identified unmethylated sites but failed at fully methylated sites, as well as Nanopolish and Tombo, which were able to recover fully methylated sites but had a high rate of false positives at unmethylated sites. Moreover, the predictions per-site generally showed a high dispersion and did not generally agree with the expected methylation frequency. A possible reason for this low agreement is that the *M.SssI* treatment used to generate the CpG-methylated control only has 95% efficiency (Simpson *et al.*, 2017). Although we used 10% incremental windows to accommodate for this potential variability in our tests, the dispersion occurred across all methods and methylation levels. Additionally, when we focused on the most reliable sites for either 0% or 100% methylation (Simpson *et al.*, 2017), the tested tools still presented dispersion as observed before (Figure 7). This indicates that there is still a major limitation in the capacity to correctly recover the methylation levels from individual reads.

The observed low agreements motivated us to propose a consensus approach, METEORE, aiming to ameliorate the deficiencies from some methods and incorporate the advantages from others. The combination of two methods improved the overall accuracy at the levels of individual reads and per-site methylation frequency. Our analyses thus suggested that it is generally advantageous to run at least two tools to obtain an accurate picture of the DNA methylation patterns. While combining multiple tools improved the accuracy even further, it might be impractical for routine analyses, due to the running times for some methods without

GPU support (Table 6). Megalodon provided overall the best performance but relies on GPUs to be time efficient. In contrast, Nanopolish and Tombo were the fastest on CPUs but showed the largest dispersions and false-positive rates. The combination of Megalodon and DeepSignal struck a good balance for accuracy and running times but would require a GPU for efficiency. On the other hand, on a CPU the combination of Nanopolish and DeepSignal can achieve an accuracy comparable to Megalodon and is time efficient (Table 6).

Additionally, we found that by reassessing the score cutoffs for individual reads, the per-site methylation predictions could be improved in some methods. Furthermore, there was an advantage in removing sites with uncertain methylation status. This strategy improved the accuracy of some methods. However, it had a large impact on the number of reads for Nanopolish and Tombo, which resulted in more reads discarded than the rest of the methods.

Nanopore sequencing offers several advantages over bisulfite treatment followed by sequencing (either Sanger or MPS) for detecting methylated CpGs, including higher accuracy, speed, throughput, and cost-effectiveness. Our study shows that nanopore sequencing detected methylation without the use of bisulfite treatment and PCR, while it still resembled the methylation pattern for the WGBS data. This eliminates sequencing errors caused by bisulfite treatment and PCR-related biases. The analysis workflow for nanopore sequencing is also simpler than that of WGBS, as it does not involve preprocessing, alignment of non-complementary reads due to bisulfite conversion, removal of PCR duplicates. In addition, nanopore sequencing can sequence longer reads, resulting in higher throughput, and it requires fewer reagents and processing steps, making it more cost-effective than bisulfite sequencing.

In the rapidly evolving field of nanopore sequencing, newer methods have been developed since the completion of our benchmarking study. Regrettably, the newly introduced ONT basecaller Dorado (Oxford Nanopore Technologies, 2023b) and Bonito (Oxford Nanopore Technologies, 2023a) were not included in our benchmarking analysis. Furthermore, the Remora model has undergone updates, with a wider range of models now available, specifically designed for the new sequencing kit and chemistry (Oxford Nanopore Technologies, 2022). Despite this, our study demonstrates the feasibility of detecting bases and even modified bases through the analysis of ionic signals in nanopore sequencing. However, we highly recommend utilizing the most up-to-date tools and models to ensure optimal accuracy. With ongoing advancements in the field, we are confident that the accuracy of nanopore sequencing will continue to improve over time.

1.2 Good concordance of nCATS data with the published bisulfite sequencing

We tested nCATS, which uses the CRISPR-Cas9 system to cleave DNA with Cas9 to ligate adapters for nanopore sequencing. Our study showed that while the nCATS approach is highly specific in targeting specific genomic regions, it still occasionally cleaves unintended DNA sequences, leading to off-target effects and potential errors in the sequencing results. Nevertheless, it possesses several other advantages that could make it well-suited for forensic applications, including: 1) real-time sequencing, since the data can be generated and analysed in real-time, allowing rapid access to the sequencing results; 2) no amplification bias, as PCR is not required for nCATS and PCR can reduce sequencing quality, especially in regions with high or low GC content; 3) flexibility, since the user can design the crRNA sequences to target different ROI, which makes it a flexible and versatile tool for targeted sequencing; and 4) portability - in our experiments we used the MINION device that is palm-sized and portable, which makes it possible to perform sequencing in the field.

The comparison with WGBS datasets using two independent nCATS experiments demonstrated consistency for both individual reads and control mixtures. Although the analytic tools recovered the overall WGBS patterns across different genomic regions, and independently of GC content and coverage, there were some remarkable variations. In particular, Nanopolish tended to overpredict methylation, DeepMod and Guppy tended to underpredict methylation, and Tombo and Guppy revealed local discrepancies with the other tools. Megalodon, DeepSignal, and the consensus of these two methods using METEORE were most consistent with the WGBS data.

Furthermore, we observed a limited impact of coverage on accuracy. This suggests a strong potential for the development of sensitive diagnostic and forensic tests without relying on high

coverage. Moreover, the use of our improved strategies with single or double cutoffs in some methods such as DeepSignal and Tombo or with METEORE consensus predictions will further facilitate accurate analyses using individual reads.

Besides, we observed that the highest discrepancy with WGBS occurred at CpG sites within AT-rich 8-mers (Figure 22a). This bias was particularly strong for DeepMod and DeepSignal, whereas there were no significant sequence biases for Nanopolish, Megalodon, and METEORE related to the discrepancy with WGBS. Although these discrepancies could be due to the detection differences between Nanopore and WGBS, they could also point to the disparate types of models and training regimes used by the tools tested.

Lastly, we demonstrate that the nCATS protocol can be used to query SNPs in ROIs using commonly used BCFtools. Despite not being designed for SNP detection in long reads, BCFtools can accurately call all SNPs with a minimum coverage of 35×. With continuous improvement in basecaller algorithms and the development of new SNP detection algorithms, we expect to see further improvement in the accuracy of SNP genotyping.

In summary, we highlighted the strengths and weakness of state-of-the-art methods to predict DNA methylation from Nanopore sequencing and provided various strategies to improve the prediction accuracy. We expect METEORE and the pipelines provided will facilitate the accurate analysis of genome-wide methylation patterns both per site and in individual molecules in multiple biological contexts.

1.3 Evaluate nanopore adaptive sampling and age-correlated DNA methylation markers

A useful expansion of FDP beyond the current use of SNP analysis, which are used to infer EVCs and BGA, is to incorporate forensic epigenomic profiling as a supplementary tool to gather intelligence in criminal investigations (Vidaki & Kayser, 2018). This approach would allow for the prediction of age, using DNA samples collected from crime scenes. By combining epigenetic profiling of age and genomic profiling of appearance and ancestry, pools of suspects in police investigations can be narrowed down further, thereby saving valuable police resources (Vidaki & Kayser, 2018).

The advancement of epigenetic technology has greatly facilitated scientific progress in identifying and comprehending both established and new epigenetic modifications. An instance of this is the widespread use of hybridisation-based microarray technology (Figure 1), such as the Illumina 850 K array, however, this only covers approximately 4% of the human epigenome (Vidaki & Kayser, 2018). Alternatively, various sequencing technologies have been developed that can be categorised into two strategies: 1) to unravel the epigenome-wide DNA methylation status, using methods such as MRE-seq, RRBS and WGBS; 2) to target specific epigenetic regions using SBE-PCR-CE and SBE-PCR-MPS.

However, all these microarray and sequencing platforms are restricted to CpG methylation only and require the development of separate assays for different markers. The use of novel technologies that provide a PCR-free and long-read sequencing approach, currently provided by ONT, will be crucial in the forensic field to minimize wet lab preparation including PCR, library preparation and clean-up, increase the number of markers used in a single assay and broaden the scope of epigenetic marks. Nanopore sequencing can offer the potential to combine genomic and epigenomic information. Despite its higher sequencing error rate compared to

other sequencing methods, and the fact that providing less coverage than PCR-based methods, the results of this study showed that nanopore sequencing still delivers accurate SNP genotyping (with >15× coverage) and methylation profiling.

We evaluated the performance of several new methods, such as Nanopolish, Longshot, and PEPPER-Margin-DeepVariant, to call variants using long-read sequencing technology. Our findings showed that nanopore-based tools outperformed short-read-based methods, as demonstrated in Figure 27, which was in line with the results of a previous study (Shafin *et al.*, 2021). To enable fast turnaround time from sample to genotyping results and serve as a useful tool for forensic biologists, an analysis pipeline for SNP genotyping should be developed using the portable MinION and adaptive sampling technique, along with existing nanopore-based computational tools for downstream analysis.

While most of age prediction studies are based on targeted methylation detection via SNaPshot, EpiTYPER system and MPS, this study addresses the question of whether a palm-sized MinION sequencer could provide CpG methylation quantification with single base resolution, but also allow for a more straightforward and less costly strategy with similar accuracy to WGBS that could be applied in a forensic setting.

With nanopore adaptive sampling, targeted sequencing is no longer done by experimental routines but via a computational approach. Simultaneous detection of SNPs and DNA methylation can now be achieved during the sequencing run without the need for time-consuming and error-inducing steps such as PCR and bisulfite conversion. Since shorter reads (<500 bp) are sequenced before ejection, resulting in short, off-target reads in the sequencing data, adaptive sampling is found to be more effective with longer reads (Figure 24a & b).

Although the dataset only included ten samples for simple linear regression analyses, we still found that the majority of CpG sites utilised in APMs exhibited a significant association with age in blood. In contrast, a relatively weaker association between PDE4C and age was found in our study. We also identified sites strongly associated with age (Supplementary data 16), but a more detailed analysis of the DNA methylation-age correlation for these particular sites should be carried out using a larger set of samples. If future studies further confirm the effectiveness of nanopore adaptive sampling technologies in detecting DNA methylation for forensic purposes, it could lead to a transformative shift in targeted forensic epigenetic profiling.

1.4 Limitations

- For now, nanopore sequencing still necessitates relatively large quantities of initial DNA (at least 1 µg, as suggested by ONT), which restricts its usage in forensics. However, this requirement is expected to decrease over time because of new chemistry. A sensitivity study should be carried out to determine whether nanopore sequencing can reliably detect low-abundance DNA samples or DNA from a mixture of different sources.
- The nCATS method has some limitations, including: 1) off-target effects due to cleavage of unintended DNA sequences; and 2) uneven sequencing depth across different regions, as the effectiveness of Cas9 targeting depends on the location of the target region and the quality of the gRNA sequences.
- The blood samples utilised in this study were either from young (~25 years old) or advanced-age (~70 years old) donors, and lacked an adequate number of samples from the middle-aged group (30-40 years old). Collecting sufficient samples from various age groups can be a challenging task.
- When predicting the age of an unknown blood sample in forensic investigations, the disease status is typically unknown. Therefore, it is crucial to construct a strong APM that includes age markers not prone to displaying distinct methylation profiles as a result of disease status.

1.5 Conclusions

- A systematic benchmarking of six computational tools for 5mC detection from Nanopore sequencing was performed. The detection capabilities of the six tools were established at the single-molecule level and in discerning different stoichiometries, at different levels of coverage and GC-content.
- Nanopore sequencing was able to recover the overall methylation patterns found in the WGBS data. The tested tools present a tradeoff between true positives and false positives and a high dispersion with respect to the expected methylation frequency values.
- A consensus-based approach using two or more methods, as implemented in METEORE, improves the accuracy of nanopore-based methylation detection.
- Snakemake pipelines such as the one implemented in METEORE enhance reproducibility and enable the systematic and accurate application of nanopore-based methylation detection to multiple datasets.
- Targeted nanopore sequencing using either Cas9-guided adapter ligation or adaptive sampling can be a reliable method for SNP genotyping and methylation profiling.
- The MinION sequencer not only offers CpG methylation quantification with single base resolution, but also provides a simpler and more cost-effective approach with accuracy comparable to the WGBS method, which could be employed for forensic purposes.
- Nanopore sequencing technology shows great promise to simultaneously detect all kinds of forensic markers from the same sample by sequencing unamplified DNA, without the need for chemical or enzymatical conversions or PCR. This would generate information-rich data from a single sequencing run to provide intelligence, including BGA, EVCs and age.

1.6 Future research

The validation of simultaneous SNP and STR genotyping using nCATS and adaptive sampling is an essential future research direction for advancing forensic DNA analysis. Adaptive sampling may prove to be a more efficient approach compared to nCATS as it eliminates the need for crRNAs and Cas9 to identify target sequences, reducing both time and cost. As our panel already included 37 STR loci using adaptive sampling, further research could focus on a follow-up evaluation of STR genotyping by comparing nanopore sequencing of these loci with Illumina sequencing, which could be a significant leap forward in forensic genotyping using nanopore sequencing data.

Future research projects could also attempt to use age-specific DNA methylation patterns to build an accurate model for age prediction using nanopore sequencing data from blood samples. To improve the accuracy of such a model, a larger number of samples should be included. This will help reduce impact of random variation, increase the representativeness of the samples (e.g. covering different age ranges, or ethnic/ancestral backgrounds), improving the precision of model estimates.

Among the CpG sites of the targeted age-associated genes in our study, a set of CpG sites could be selected based on the results of the simple regression analysis, followed by a stepwise regression for variable selection to identify the CpG sites that could contribute the most to age estimation modelling. Then a multiple regression analysis could be performed with these sites to develop a DNA methylation-based age prediction model using nanopore sequencing.

To implement an APM using nanopore sequencing for forensic casework, it is necessary to allocate appropriate resources and develop expertise in new sequencing technologies and DNA analysis. This will ensure that the new method's performance is thoroughly tested, and new

data interpretation guidelines are developed. Nanopore sequencing offers several advantages for forensic applications, including the ability to detect methylation without PCR and bisulfite conversion, which eliminates PCR-related amplification biases and potential bisulfite conversion errors linked to conversion efficiency and specificity, as well as DNA degradation. A validation of the new technique should include testing of its sensitivity, reproducibility, and applicability to real samples.

Taking nanopore sequencing to further steps in the field would allow for the detection of biosecurity threats such as the illegal trade of tree/roots and the presence of pathogens, which could be detected on-site by the MinION device. It is important to continue to explore the potential of this technology and develop new applications that can benefit forensic investigations.

Lastly, while our study has predominantly concentrated on 5mC, given its prevalence as the most common form of DNA methylation in humans, it is crucial to explore other modifications and their respective detection methods in future research. Modifications such as 5hmC, and 6mA also play pivotal roles in gene expression regulation, yet they have received comparatively less attention due to the scarcity of reliable methods and reference datasets. Exploring these less-studied modifications offers a promising path for future research, where the development of robust methodologies and the establishment of comprehensive reference datasets would greatly enhance our understanding of epigenetic regulation.

CHAPTER VI: BIBLIOGRAPHY

Aliferi, A., Ballard, D., Gallidabino, M. D., *et al.* (2018). DNA methylation-based age prediction using massively parallel sequencing data and multiple machine learning models. *Forensic Science International: Genetics, 37*, 215-226.

Amarasinghe, S. L., Su, S., Dong, X., *et al.* (2020). Opportunities and challenges in long-read sequencing data analysis. *Genome Biology, 21*(1), 30. doi:10.1186/s13059-020-1935-5

Andrews, S. (2010). FastQC: a quality control tool for high throughput sequence data. In: Babraham Bioinformatics, Babraham Institute, Cambridge, United Kingdom.

Bekaert, B., Kamalandua, A., Zapico, S. C., *et al.* (2015). Improved age determination of blood and teeth samples using a selected set of DNA methylation markers. *Epigenetics, 10*(10), 922-930.

Bock, C. (2012). Analysing and interpreting DNA methylation data. *Nature Reviews Genetics, 13*(10), 705-719. doi:10.1038/nrg3273

Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics, 30*(15), 2114-2120.

Børsting, C., & Morling, N. (2015). Next generation sequencing and its applications in forensic genetics. *Forensic Science International: Genetics, 18*, 78-89. doi:https://doi.org/10.1016/j.fsigen.2015.02.002

Bowman, Z., Daniel, R., Gerostamoulos, D., *et al.* (2022). Rapid DNA from a disaster victim identification perspective: Is it a game changer? *Forensic Science International: Genetics*, 102684.

Breiman, L. (2001). Random forests. *Machine learning, 45*(1), 5-32.

Brinkman, A. B., Simmer, F., Ma, K., *et al.* (2010). Whole-genome DNA methylation profiling using MethylCap-seq. *Methods, 52*(3), 232-236.

Broad Institute. (2023). GATK Best Practices Workflows - Germline short variant discoverty (SNPs+Indels). Retrieved from https://gatk.broadinstitute.org/hc/en-us/articles/360035535932-Germline-short-variant-discovery-SNPs-Indels-

Budowle, B., Schmedes, S. E., & Wendt, F. R. (2017). Increasing the reach of forensic genetics with massively parallel sequencing. *Forensic Science, Medicine and Pathology, 13*(3), 342-349. doi:10.1007/s12024-017-9882-5

Burton, A. S., Stahl, S. E., John, K. K., *et al.* (2020). Off Earth Identification of Bacterial Populations Using 16S rDNA Nanopore Sequencing. *Genes, 11*(1), 76. Retrieved from https://www.mdpi.com/2073-4425/11/1/76

Chen, P.-Y., Feng, S., Joo, J. W. J., *et al.* (2011). A comparative analysis of DNA methylation across human embryonic stem cell lines. *Genome Biology, 12*(7), R62. doi:10.1186/gb-2011-12-7-r62

Cornelis, S., Gansemans, Y., Deleye, L., *et al.* (2017). Forensic SNP genotyping using nanopore MinION sequencing. *Scientific reports, 7*(1), 1-5.

Crooks, G. E., Hon, G., Chandonia, J.-M., *et al.* (2004). WebLogo: a sequence logo generator. *Genome research, 14*(6), 1188-1190.

Danecek, P., Bonfield, J. K., Liddle, J., *et al.* (2021). Twelve years of SAMtools and BCFtools. *GigaScience, 10*(2). doi:10.1093/gigascience/giab008

Daniel, R., Santos, C., Phillips, C., *et al.* (2015). A SNaPshot of next generation sequencing for forensic SNP analysis. *Forensic Science International: Genetics, 14*, 50-60. doi:https://doi.org/10.1016/j.fsigen.2014.08.013

Deaton, A. M., & Bird, A. (2011). CpG islands and the regulation of transcription. *Genes & development, 25*(10), 1010-1022.

Dias, H. C., Cunha, E., Real, F. C., *et al.* (2020). Age prediction in living: Forensic epigenetic age estimation based on blood samples. *Legal Medicine, 47*, 101763.

Dittforth, S., Ozturk, D., & Mueller, M. (2023). Benchmarking the Oxford Nanopore Technologies basecallers on AWS. Retrieved from https://aws.amazon.com/blogs/hpc/benchmarking-the-oxford-nanopore-technologies-basecallers-on-aws/

Dunham, I., Kundaje, A., Aldred, S. F., *et al.* (2012). An integrated encyclopedia of DNA elements in the human genome. *Nature, 489*(7414), 57-74. doi:10.1038/nature11247

Edge, P., & Bansal, V. (2019). Longshot enables accurate variant calling in diploid genomes from single-molecule long read sequencing. *Nature Communications, 10*(1), 1-10.

Ehrich, M., Zoll, S., Sur, S., *et al.* (2007). A new method for accurate assessment of DNA quality after bisulfite treatment. *Nucleic Acids Research, 35*(5), e29.

EpigenDX. (2018). Methylation Controls. Retrieved from https://www.epigendx.com/d/products/methylation-controls

Faria, N. R., Sabino, E. C., Nunes, M. R. T., *et al.* (2016). Mobile real-time surveillance of Zika virus in Brazil. *Genome Medicine, 8*(1), 97. doi:10.1186/s13073-016-0356-2

Feng, L., Peng, F., Li, S., *et al.* (2018). Systematic feature selection improves accuracy of methylation-based forensic age estimation in Han Chinese males. *Forensic Science International: Genetics, 35*, 38-45.

Fiol, A., Jurado-Ruiz, F., López-Girona, E., *et al.* (2022). An efficient CRISPR-Cas9 enrichment sequencing strategy for characterizing complex and highly duplicated genomic regions. A case study in the Prunus salicina LG3-MYB10 genes cluster. *Plant Methods, 18*(1), 105. doi:10.1186/s13007-022-00937-4

Fleckhaus, J., & Schneider, P. M. (2020). Novel multiplex strategy for DNA methylation-based age prediction from small amounts of DNA via Pyrosequencing. *Forensic Science International: Genetics, 44*, 102189.

Freire-Aradas, A., Phillips, C., Girón-Santamaría, L., *et al.* (2018). Tracking age-correlated DNA methylation markers in the young. *Forensic Science International: Genetics, 36*, 50-59.

Freire-Aradas, A., Phillips, C., Mosquera-Miguel, A., *et al.* (2016). Development of a methylation marker set for forensic age estimation using analysis of public methylation data and the Agena Bioscience EpiTYPER system. *Forensic Science International: Genetics, 24*, 65-74.

Freire-Aradas, A., Pośpiech, E., Aliferi, A., *et al.* (2020). A comparison of forensic age prediction models using data from four DNA methylation technologies. *Frontiers in genetics*, 932.

Frith, M. C., Mori, R., & Asai, K. (2012). A mostly traditional approach improves alignment of bisulfite-converted DNA. *Nucleic Acids Research, 40*(13), e100-e100.

Frommer, M., McDonald, L. E., Millar, D. S., *et al.* (1992). A genomic sequencing protocol that yields a positive display of 5-methylcytosine residues in individual DNA strands. *Proceedings of the National Academy of Sciences, 89*(5), 1827-1831.

Fujita, P. A., Rhead, B., Zweig, A. S., *et al.* (2010). The UCSC genome browser database: update 2011. *Nucleic Acids Research, 39*(suppl_1), D876-D882.

Gamaarachchi, H., Lam, C. W., Jayatilaka, G., *et al.* (2020). GPU accelerated adaptive banded event alignment for rapid comparative nanopore signal analysis. *BMC bioinformatics, 21*(1), 1-13.

Gargis, A. S., Cherney, B., Conley, A. B., *et al.* (2019). Rapid Detection of Genetic Engineering, Structural Variation, and Antimicrobial Resistance Markers in Bacterial Biothreat Pathogens by Nanopore Sequencing. *Scientific reports, 9*(1), 13501. doi:10.1038/s41598-019-49700-1

Gilpatrick, T., Lee, I., Graham, J. E., *et al.* (2020). Targeted nanopore sequencing with Cas9-guided adapter ligation. *Nature Biotechnology*. doi:10.1038/s41587-020-0407-5

Goldsmith, C., Cohen, D., Dubois, A., *et al.* (2021). Cas9-targeted nanopore sequencing reveals epigenetic heterogeneity after de novo assembly of native full-length hepatitis B virus genomes. *Microbial Genomics, 7*(5). doi:https://doi.org/10.1099/mgen.0.000507

Greenberg, M. V. C., & Bourc'his, D. (2019). The diverse roles of DNA methylation in mammalian development and disease. *Nature Reviews Molecular Cell Biology, 20*(10), 590-607. doi:10.1038/s41580-019-0159-6

Grunau, C., Clark, S., & Rosenthal, A. (2001). Bisulfite genomic sequencing: systematic investigation of critical experimental parameters. *Nucleic Acids Research, 29*(13), e65-e65.

Guo, W., Fiziev, P., Yan, W., *et al.* (2013). BS-Seeker2: a versatile aligning pipeline for bisulfite sequencing data. *BMC Genomics, 14*(1), 1-8.

Hannum, G., Guinney, J., Zhao, L., *et al.* (2013). Genome-wide methylation profiles reveal quantitative views of human aging rates. *Molecular cell, 49*(2), 359-367.

Hayes, B. J., Nguyen, L. T., Forutan, M., *et al.* (2021). An Epigenetic Aging Clock for Cattle Using Portable Sequencing Technology. *Frontiers in genetics, 12*. doi:10.3389/fgene.2021.760450

Hong, S. R., Jung, S.-E., Lee, E. H., *et al.* (2017). DNA methylation-based age prediction from saliva: High age predictability by combination of 7 CpG markers. *Forensic Science International: Genetics, 29*, 118-125.

Horvath, S. (2013). DNA methylation age of human tissues and cell types. *Genome Biology, 14*(10), 1-20.

Integrated DNA Technologies. (2019). CRISPR-Cas9 Guide RNA Design Checker. Retrieved from https://sg.idtdna.com/site/order/designtool/index/CRISPR_SEQUENCE

Jones, M. J., Goodman, S. J., & Kobor, M. S. (2015). DNA methylation and healthy human aging. *Aging cell, 14*(6), 924-932.

Jones, P. A. (2012). Functions of DNA methylation: islands, start sites, gene bodies and beyond. *Nature Reviews Genetics, 13*(7), 484-492. doi:10.1038/nrg3230

Jung, S.-E., Lim, S. M., Hong, S. R., *et al.* (2019). DNA methylation of the ELOVL2, FHL2, KLF14, C1orf132/MIR29B2C, and TRIM59 genes for age prediction from blood, saliva, and buccal swab samples. *Forensic Science International: Genetics, 38*, 1-8.

Kader, F., & Ghai, M. (2015). DNA methylation and application in forensic sciences. *Forensic science international, 249*, 255-265. doi:10.1016/j.forsciint.2015.01.037

Köster, J., & Rahmann, S. (2012). Snakemake—a scalable bioinformatics workflow engine. *Bioinformatics, 28*(19), 2520-2522. doi:10.1093/bioinformatics/bts480

Kovaka, S., Fan, Y., Ni, B., *et al.* (2021). Targeted nanopore sequencing by real-time mapping of raw electrical signal with UNCALLED. *Nature Biotechnology, 39*(4), 431-441.

Krueger, F., & Andrews, S. R. (2011). Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications. *Bioinformatics, 27*(11), 1571-1572.

Kumar, S., Chinnusamy, V., & Mohapatra, T. (2018). Epigenetics of Modified DNA Bases: 5-Methylcytosine and Beyond. *Frontiers in genetics, 9*, 640-640. doi:10.3389/fgene.2018.00640

Labun, K., Montague, T. G., Krause, M., *et al.* (2019). CHOPCHOP v3: expanding the CRISPR web toolbox beyond genome editing. *Nucleic Acids Research, 47*(W1), W171-W174. doi:10.1093/nar/gkz365

Langmead, B., Trapnell, C., Pop, M., *et al.* (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biology, 10*(3), 1-10.

Laszlo, A. H., Derrington, I. M., Brinkerhoff, H., *et al.* (2013). Detection and mapping of 5-methylcytosine and 5-hydroxymethylcytosine with nanopore MspA. *Proceedings of the National Academy of Sciences, 110*(47), 18904-18909. doi:10.1073/pnas.1310240110

Lee, J. E., Lee, J. M., Naue, J., *et al.* (2022). A collaborative exercise on DNA methylation-based age prediction and body fluid typing. *Forensic Science International: Genetics, 57*, 102656.

Levine, M. E., Lu, A. T., Quach, A.*, et al.* (2018). An epigenetic biomarker of aging for lifespan and healthspan. *Aging (Albany NY), 10*(4), 573.

Li, H. (2018). Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics, 34*(18), 3094-3100. doi:10.1093/bioinformatics/bty191

Li, H., & Durbin, R. (2010). Fast and accurate long-read alignment with Burrows–Wheeler transform. *Bioinformatics, 26*(5), 589-595.

Li, L., Song, F., Lang, M.*, et al.* (2020). Methylation-based age prediction using pyrosequencing platform from seminal stains in han chinese males. *Journal of forensic sciences, 65*(2), 610-619.

Liu, Q., Fang, L., Yu, G.*, et al.* (2019a). Detection of DNA base modifications by deep recurrent neural network on Oxford Nanopore sequencing data. *Nature Communications, 10*(1), 2449. doi:10.1038/s41467-019-10168-2

Liu, Q., Georgieva, D. C., Egli, D.*, et al.* (2019b). NanoMod: a computational tool to detect DNA modifications using Nanopore long-read sequencing data. *BMC Genomics, 20*(1), 78. doi:10.1186/s12864-018-5372-8

López-Girona, E., Davy, M. W., Albert, N. W.*, et al.* (2020). CRISPR-Cas9 enrichment and long read sequencing for fine mapping in plants. *Plant Methods, 16*(1), 121. doi:10.1186/s13007-020-00661-x

Marquet, M., Zöllkau, J., Pastuschek, J.*, et al.* (2022). Evaluation of microbiome enrichment and host DNA depletion in human vaginal samples using Oxford Nanopore's adaptive sequencing. *Scientific reports, 12*(1), 4000. doi:10.1038/s41598-022-08003-8

McIntyre, A. B. R., Alexander, N., Grigorev, K.*, et al.* (2019). Single-molecule sequencing detection of N6-methyladenine in microbial reference materials. *Nature Communications, 10*(1), 579. doi:10.1038/s41467-019-08289-9

Mehta, B., Daniel, R., Phillips, C.*, et al.* (2016). Massively parallel sequencing of customised forensically informative SNP panels on the MiSeq. *Electrophoresis, 37*(21), 2832-2840.

Mehta, B., Daniel, R., Phillips, C.*, et al.* (2017). Forensically relevant SNaPshot® assays for human DNA SNP analysis: a review. *International Journal of Legal Medicine, 131*(1), 21-37. doi:10.1007/s00414-016-1490-5

Meissner, A., Gnirke, A., Bell, G. W.*, et al.* (2005). Reduced representation bisulfite sequencing for comparative high-resolution DNA methylation analysis. *Nucleic Acids Research, 33*(18), 5868-5877.

Montesanto, A., D'Aquila, P., Lagani, V.*, et al.* (2020). A new robust epigenetic model for forensic age prediction. *Journal of forensic sciences, 65*(5), 1424-1431.

Naue, J., Hoefsloot, H. C., Mook, O. R.*, et al.* (2017). Chronological age prediction based on DNA methylation: massive parallel sequencing and random forest regression. *Forensic Science International: Genetics, 31*, 19-28.

Naue, J., Sänger, T., Hoefsloot, H. C., *et al.* (2018). Proof of concept study of age-dependent DNA methylation markers across different tissues by massive parallel sequencing. *Forensic Science International: Genetics, 36*, 152-159.

Ni, P., Huang, N., Nie, F., *et al.* (2021). Genome-wide detection of cytosine methylations in plant from Nanopore data using deep learning. *Nature Communications, 12*(1), 1-11.

Ni, P., Huang, N., Zhang, Z., *et al.* (2019). DeepSignal: detecting DNA methylation state from Nanopore sequencing reads using deep-learning. *Bioinformatics, 35*(22), 4586-4595.

O'Shea, J. P., Chou, M. F., Quader, S. A., *et al.* (2013). pLogo: a probabilistic approach to visualizing sequence motifs. *Nature Methods, 10*(12), 1211-1212. doi:10.1038/nmeth.2646

Oxford Nanopore Technologies. (2019). Evaluation of read-mapping characteristics from a Cas-mediated PCR-free enrichment. Retrieved from https://community.nanoporetech.com/knowledge/bioinformatics/evaluation-of-read-mapping/tutorial

Oxford Nanopore Technologies. (2020a). Oxford Nanopore Technologies GitHub. Retrieved from https://github.com/nanoporetech

Oxford Nanopore Technologies. (2020b). Oxford Nanopore Technologies GitHub - Megalodon Retrieved from https://github.com/nanoporetech/megalodon

Oxford Nanopore Technologies. (2020c). Read Until API. Retrieved from https://github.com/nanoporetech/read_until_api

Oxford Nanopore Technologies. (2020d). Rerio GitHub page. Retrieved from https://github.com/nanoporetech/rerio

Oxford Nanopore Technologies. (2022). Remora GitHub page. Retrieved from https://github.com/nanoporetech/remora

Oxford Nanopore Technologies. (2023a, 1 Jun 2023). Oxford Nanopore Technologies GitHub - Bonito. Retrieved from https://github.com/nanoporetech/bonito

Oxford Nanopore Technologies. (2023b, 27 June 2023). Oxford Nanopore Technologies GitHub - Dorado. Retrieved from https://github.com/nanoporetech/dorado

Park, J.-L., Kim, J. H., Seo, E., *et al.* (2016). Identification and evaluation of age-correlated DNA methylation markers for forensic use. *Forensic Science International: Genetics, 23*, 64-70.

Parson, W., Ballard, D., Budowle, B., *et al.* (2016). Massively parallel sequencing of forensic STRs: considerations of the DNA commission of the International Society for Forensic Genetics (ISFG) on minimal nomenclature requirements. *Forensic Science International: Genetics, 22*, 54-63.

Payne, A., Holmes, N., Clarke, T., *et al.* (2021). Readfish enables targeted nanopore sequencing of gigabase-sized genomes. *Nature Biotechnology, 39*(4), 442-450.

Pedregosa, F., Varoquaux, G., Gramfort, A., *et al.* (2011). Scikit-learn: Machine learning in Python. *the Journal of machine Learning research, 12*, 2825-2830.

Phillips, C., McNevin, D., Kidd, K., *et al.* (2019). MAPlex-A massively parallel sequencing ancestry analysis multiplex for Asia-Pacific populations. *Forensic Science International: Genetics, 42*, 213-226.

Phillips, C., Prieto, L., Fondevila, M., *et al.* (2009). Ancestry analysis in the 11-M Madrid bomb attack investigation. *PLOS ONE, 4*(8), e6583.

Phillips, C., Salas, A., Sanchez, J., *et al.* (2007). Inferring ancestral origin using a single multiplex assay of ancestry-informative marker SNPs. *Forensic Science International: Genetics, 1*(3-4), 273-280.

Plesivkova, D., Richards, R., & Harbison, S. (2019). A review of the potential of the MinION™ single-molecule sequencing system for forensic applications. *Wiley Interdisciplinary Reviews: Forensic Science, 1*(1), e1323.

Quick, J., Grubaugh, N. D., Pullan, S. T., *et al.* (2017). Multiplex PCR method for MinION and Illumina sequencing of Zika and other virus genomes directly from clinical samples. *Nature Protocols, 12*(6), 1261-1276. doi:10.1038/nprot.2017.066

Quick, J., Loman, N. J., Duraffour, S., *et al.* (2016). Real-time, portable genome sequencing for Ebola surveillance. *Nature, 530*(7589), 228-232. doi:10.1038/nature16996

R Core Team. (2020). R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from https://www.R-project.org/

Raiber, E.-A., Hardisty, R., van Delft, P., *et al.* (2017). Mapping and elucidating the function of modified bases in DNA. *Nature Reviews Chemistry, 1*(9), 0069. doi:10.1038/s41570-017-0069

Rand, A. C., Jain, M., Eizenga, J. M., *et al.* (2017). Mapping DNA methylation with high-throughput nanopore sequencing. *Nature Methods, 14*(4), 411-413. doi:10.1038/nmeth.4189

Robinson, J. T., Thorvaldsdóttir, H., Winckler, W., *et al.* (2011). Integrative genomics viewer. *Nature Biotechnology, 29*(1), 24-26. doi:10.1038/nbt.1754

Santos, C., Phillips, C., Fondevila, M., *et al.* (2016). Pacifiplex: an ancestry-informative SNP panel centred on Australia and the Pacific region. *Forensic Science International: Genetics, 20*, 71-80.

Shafin, K., Pesout, T., Chang, P.-C., *et al.* (2021). Haplotype-aware variant calling with PEPPER-Margin-DeepVariant enables high accuracy in nanopore long-reads. *Nature Methods, 18*(11), 1322-1332.

Sim, J., & Chapman, B. (2019). In-field whole genome sequencing using the MinION nanopore sequencer to detect the presence of high-prized military targets. *Australian Journal of Forensic Sciences, 51*(sup1), S86-S90. doi:10.1080/00450618.2019.1568562

Simpson, J. T., Workman, R. E., Zuzarte, P. C.*, et al.* (2017). Detecting DNA cytosine methylation using nanopore sequencing. *Nature Methods, 14*(4), 407-410. doi:10.1038/nmeth.4184

Singer, B. D. (2019). A practical guide to the measurement and analysis of DNA methylation. *American journal of respiratory cell and molecular biology, 61*(4), 417-428.

Stevanovski, I., Chintalaphani, S. R., Gamaarachchi, H.*, et al.* (2022). Comprehensive genetic diagnosis of tandem repeat expansion disorders with programmable targeted nanopore sequencing. *Science Advances, 8*(9), eabm5386. doi:doi:10.1126/sciadv.abm5386

Stoiber, M., Quick, J., Egan, R.*, et al.* (2017). De novo Identification of DNA Modifications Enabled by Genome-Guided Nanopore Signal Processing. *bioRxiv*, 094672. doi:10.1101/094672

Tacutu, R., Thornton, D., Johnson, E.*, et al.* (2018). Human ageing genomic resources: new and updated databases. *Nucleic Acids Research, 46*(D1), D1083-D1090.

Thong, Z., Chan, X. L. S., Tan, J. Y. Y.*, et al.* (2017). Evaluation of DNA methylation-based age prediction on blood. *Forensic Science International: Genetics Supplement Series, 6*, e249-e251.

Thorvaldsdóttir, H., Robinson, J. T., & Mesirov, J. P. (2013). Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Briefings in Bioinformatics, 14*(2), 178-192.

Truelove, N. K., Andruszkiewicz, E. A., & Block, B. A. (2019). A rapid environmental DNA method for detecting white sharks in the open ocean. *Methods in Ecology and Evolution, 10*(8), 1128-1135. doi:10.1111/2041-210x.13201

Tsuji, J., & Weng, Z. (2015). Evaluation of preprocessing, mapping and postprocessing algorithms for analysing whole genome bisulfite sequencing data. *Briefings in Bioinformatics, 17*(6), 938-952. doi:10.1093/bib/bbv103

Tytgat, O., Gansemans, Y., Weymaere, J.*, et al.* (2020). Nanopore sequencing of a forensic STR multiplex reveals loci suitable for single-contributor STR profiling. *Genes, 11*(4), 381.

Tytgat, O., Škevin, S., Deforce, D.*, et al.* (2022). Nanopore sequencing of a forensic combined STR and SNP multiplex. *Forensic Science International: Genetics, 56*, 102621.

Van der Auwera, G. A., & O'Connor, B. D. (2020). *Genomics in the cloud: using Docker, GATK, and WDL in Terra*: O'Reilly Media.

Van Dijk, E. L., Auger, H., Jaszczyszyn, Y.*, et al.* (2014). Ten years of next-generation sequencing technology. *Trends in genetics, 30*(9), 418-426.

Vidaki, A., & Kayser, M. (2018). Recent progress, methods and perspectives in forensic epigenetics. *Forensic Science International: Genetics, 37*, 180-195.

Walsh, S., Liu, F., Wollstein, A.*, et al.* (2013). The HIrisPlex system for simultaneous prediction of hair and eye colour from DNA. *Forensic Science International: Genetics, 7*(1), 98-115.

Wang, T., Ma, J., Hogan, A. N.*, et al.* (2020). Quantitative translation of dog-to-human aging by conserved remodeling of the DNA methylome. *Cell systems, 11*(2), 176-185. e176.

Watherston, J., Watson, J., Bruce, D.*, et al.* (2022). An in-field evaluation of rapid DNA instruments for disaster victim identification. *International Journal of Legal Medicine, 136*(2), 493-499. doi:10.1007/s00414-021-02748-z

Weber, M., Davies, J. J., Wittig, D.*, et al.* (2005). Chromosome-wide and promoter-specific analyses identify sites of differential DNA methylation in normal and transformed human cells. *Nature genetics, 37*(8), 853-862.

Weidner, C. I., Lin, Q., Koch, C. M.*, et al.* (2014). Aging of blood can be tracked by DNA methylation changes at just three CpG sites. *Genome Biology, 15*(2), 1-12.

Wick, R. R., Judd, L. M., & Holt, K. E. (2019). Performance of neural network basecalling tools for Oxford Nanopore sequencing. *Genome Biology, 20*(1), 129. doi:10.1186/s13059-019-1727-y

Woźniak, A., Heidegger, A., Piniewska-Róg, D.*, et al.* (2021). Development of the VISAGE enhanced tool and statistical models for epigenetic age estimation in blood, buccal cells and bones. *Aging (Albany NY), 13*(5), 6459.

Xavier, C., de la Puente, M., Mosquera-Miguel, A.*, et al.* (2020). Development and validation of the VISAGE AmpliSeq basic tool to predict appearance and ancestry from DNA. *Forensic Science International: Genetics, 48*, 102336.

Xavier, C., de la Puente, M., Sidstedt, M.*, et al.* (2022). Evaluation of the VISAGE basic tool for appearance and ancestry inference using ForenSeq® chemistry on the MiSeq FGx® system. *Forensic Science International: Genetics, 58*, 102675.

Xi, Y., & Li, W. (2009). BSMAP: whole genome bisulfite sequence MAPping program. *BMC bioinformatics, 10*(1), 1-9.

Yong, W.-S., Hsu, F.-M., & Chen, P.-Y. (2016). Profiling genome-wide DNA methylation. *Epigenetics & Chromatin, 9*(1), 26. doi:10.1186/s13072-016-0075-3

Yuen, Z. W.-S., Srivastava, A., Jack, C.*, et al.* (2021). METEORE: Automatic DNA methylation detection from Nanopore tools and their consensus models.

Zbieć-Piekarska, R., Spólnicka, M., Kupiec, T.*, et al.* (2015a). Examination of DNA methylation status of the ELOVL2 marker may be useful for human age prediction in forensic science. *Forensic Science International: Genetics, 14*, 161-167.

Zbieć-Piekarska, R., Spólnicka, M., Kupiec, T.*, et al.* (2015b). Development of a forensically useful age prediction method based on DNA methylation analysis. *Forensic Science International: Genetics, 17*, 173-179.

Zhao, M.-T., Whyte, J. J., Hopkins, G. M.*, et al.* (2014). Methylated DNA immunoprecipitation and high-throughput sequencing (MeDIP-seq) using low amounts of genomic DNA. *Cellular Reprogramming (Formerly Cloning and Stem Cells), 16*(3), 175-184.

CHAPTER VII: APPENDIX

Supplementary material

Supplementary Data 1 Supplementary_data_1.xlsx

Description: 100 CpG sites used in mixture dataset 1. A tab separated file including the following fields:

1. Chromosome
2. Ending position

Supplementary Data 2 Supplementary_data_2.xlsx

Description: 50 CpG sites used in mixture dataset 2. A tab separated file including the following fields:

1. Chromosome
2. Ending position

Supplementary Data 3 Supplementary_data_3.xlsx

Description: Fast5 files used to generate each subset in mixture dataset 1. The fast5 files were generated by Simpson et al. (2017) and can be downloaded from the European Nucleotide Archive (ENA) under accession PRJEB13021 (ERR1676719 for negative control and ERR1676720 for positive control). A tab separated file including the following fields:

1. Name of the subset (m0, m10, m20…)
2. Fast5 filename

Supplementary Data 4 **Supplementary_data_4.xlsx**

Description: Fast5 files used to generate each subset in mixture dataset 2. The fast5 files was generated by Simpson et al. (2017) and can be downloaded from the European Nucleotide Archive (ENA) under accession PRJEB13021 (ERR1676719 for negative control and ERR1676720 for positive control). The reads used in mixture dataset 2 are different from mixture dataset 1. A tab separated file including the following fields:

1. Name of the subset (m0, m10, m20…)
2. Fast5 filename

Supplementary Data 5 **Supplementary_data_5.bed**

Description: A bed file containing target regions used in adaptive sampling, which includes the following columns:

1. Chromosome
2. Starting position
3. Ending position
4. Target ID (can be gene name, SNP ID, STR name)

Supplementary Data 6 **Supplementary_data_6.fa**

Description: A modified reference file in FASTA format was used for alignment in the adaptive sampling analyses. It corresponds to the GRCh38/hg38 human genome assembly and includes three additional copies of a tandem repeat of the rDNA unit (KY962518.1), each containing two units arranged in three orientations.

Supplementary Data 7 **Supplementary_data_7.xlsx**

Description: Methylation frequencies of CpG sites calculated by each tested tool for each subset in control mixture dataset 1. An Excel file including the following fields:

1. Chromosome
2. Ending position
3. % Methylated sets
4. Expected methylation frequency
5. Predicted methylation frequency
6. Method
7. Left bound of each 10% methylated bin

8. Right bound of each 10% methylated bin
9. Predicted [inside/outside] a 10% window around the expected methylation proportion

Supplementary Data 8 Supplementary_data_8.xlsx

Description: The number of true negatives for each tested tool according to different thresholds for the predicted methylation frequency below which a site was called unmethylated, using 0% methylated set from mixture dataset 1. An Excel file including the following fields:

1. Threshold
2. Method
3. Number of true negatives

Supplementary Data 9 Supplementary_data_9.xlsx

Description: The number of true positives for each tested tool according to different thresholds for the predicted methylation frequency above which a site was called methylated, using 100% methylated set from mixture dataset 1. An Excel file including the following fields:

1. Threshold
2. Method
3. Number of true positives

Supplementary Data 10 Supplementary_data_10.xlsx

Description: Per-read performance of a binary classifier at all possible classification thresholds for five tested tools measured by receiver operating characteristic (ROC) curve, using reads from 0% methylated set from mixture dataset 1. An Excel file including the following fields:

1. Threshold
2. Specificity, i.e. TN/(TN +FP) = 1−FPR
3. Sensitivity, i.e. TP/(TP/FN) = 1−FNR = TPR
4. FPR
5. Method

, where TP = true positives, FP = false positives, TN = true negatives, FN = false negatives, TPR = true positive rate, FNR = false negative rate, and FPR = false positive rate.

Supplementary Data 11 Supplementary_data_11.xlsx

Description: Per-read performance of a binary classifier at all possible classification thresholds for five tested tools measure by precision-recall (PR) curve, using reads from 0% methylated set from mixture dataset 1. An Excel file including the following fields:

1. Threshold
2. Recall, i.e. TP/(TP/FN) = 1−FNR = TPR
3. Precision, i.e. TP/(TP+FP) = 1−FDR
4. Method

, where TP = true positives, FP = false positives, TN = true negatives, FN = false negatives, TPR = true positive rate, FNR = false negative rate, and FDR = false discovery rate.

Supplementary Data 12 Supplementary_data_12.xlsx

Description: Methylation frequencies of CpG sites calculated by each tested tool for each subset in control mixture dataset 2. An Excel file including the following fields:

1. Chromosome
2. Ending position
3. % Methylated sets
4. Expected methylation frequency
5. Predicted methylation frequency
6. Method
7. Left bound of each 10% methylated bin
8. Right bound of each 10% methylated bin
9. Predicted [inside/outside] a 10% window around the expected methylation proportion

Supplementary Data 13 Supplementary_data_13.xlsx

Description: Methylation frequencies and coverage of CpG sites reported by each tested tool across the 10 targeted regions from our Cas9-targeted Nanopore sequencing (nCATS) data. An Excel file including the following fields:

1. Chromosome
2. Ending position
3. Methylation frequency reported by Nanopolish
4. Coverage reported by Nanopolish
5. Methylation frequency reported by DeepSignal
6. Coverage reported by DeepSignal
7. Methylation frequency reported by Megalodon
8. Coverage reported by Megalodon
9. Methylation frequency reported by Tombo
10. Coverage reported by Tombo
11. Methylation frequency reported by Guppy
12. Coverage reported by Guppy
13. Methylation frequency reported by DeepMod
14. Coverage reported by DeepMod
15. Methylation frequency reported by METEORE (RF)
16. Coverage reported by METEORE (RF)
17. Methylation frequency reported by METEORE (REG)
18. Coverage reported by METEORE (REG)
19. Methylation frequency reported by whole genome bisulfite sequencing (WGBS)
20. Depth of coverage reported from the BAM file
21. WGBS methylation bin
22. Target region

Supplementary Data 14 Supplementary_data_14.xlsx

Description: Methylation frequencies and coverage of CpG sites reported by each tested tool across the 8 targeted regions from the Cas9-targeted Nanopore sequencing (nCATS) data generated by Gilpatrick et al. (2020). An Excel file including the following fields:

1. Chromosome
2. Ending position
3. Methylation frequency reported by Nanopolish
4. Coverage reported by Nanopolish
5. Methylation frequency reported by DeepSignal
6. Coverage reported by DeepSignal
7. Methylation frequency reported by Megalodon
8. Coverage reported by Megalodon
9. Methylation frequency reported by Tombo
10. Coverage reported by Tombo
11. Methylation frequency reported by Guppy
12. Coverage reported by Guppy
13. Methylation frequency reported by DeepMod
14. Coverage reported by DeepMod
15. Methylation frequency reported by METEORE (RF)
16. Coverage reported by METEORE (RF)
17. Methylation frequency reported by METEORE (REG)
18. Coverage reported by METEORE (REG)
19. Methylation frequency reported by whole genome bisulfite sequencing (WGBS)
20. Depth of coverage reported from the BAM file
21. WGBS methylation bin
22. Target region

Supplementary Data 15 Supplementary_data_15.xlsx

Description: 8-mer contexts of the CpG sites and absolute difference between methylation frequencies from Nanopore-based methylation detection tools and whole genome bisulfite sequencing (WGBS) data stratified by different CG content in the 8-mers. Cas9-targeted Nanopore sequencing (nCATS) data from Gilpatrick et al. (2020) were used. An Excel file including the following fields:

1. Chromosome
2. Ending position
3. 8-mer with a CpG in the middle (XXXCGXXX)
4. Method

5. Absolute difference of methylation frequencies between Nanopore-based tools and WGBS
6. Percentage of C and G residues in the 8-mers
7. GC content bin

Supplementary Data 16 Supplementart_data_16.xlsx

Description: Simple linear regression between chronological age and methylation levels of each CpG site enriched and sequenced from the nanopore adaptive sampling experiment. An Excel file including the following fields:

1. Chromosomal position of the CpG site (GRCh38/hg38)
2. Associated gene
3. R-squared
4. p-value
5. Mean absolute error (MAE)
6. Root mean square error (RMSE)