Supplementary Information

MeLoDe: detecting DNA methylation from low-depth long-read sequencing data

Cao *et al.*

Supplementary Note (sections 1-3)
Supplementary Figures 1-25
Supplementary Tables 1-7
Supplementary References

Supplementary Note

Section 1: Optimization of the read-level model

MeLoDe's read-level model was developed to predict CpG site 5mC probabilities from PacBio sequencing data, including both HiFi reads and non-HiFi reads. The model's architecture was first established through a hyperparameter search. Subsequently, we systematically analyzed factors that could potentially influence model's performance to ensure robust model training, as detailed below. For the training data, read-level models were trained and validated on a 90%/10% split of CpGs from M.SssI-treated and WGA DNA from five HPRC-derived lymphoblastoid cell lines (HG02080, HG01109, HG02723, HG005, and HG002), and generalization was assessed on an independent cell line not included in training.

1.1. Nucleotide window size

To determine the optimal nucleotide window size for the read-level model, centered on a CpG site, using HiFi reads, we generated training datasets with window spans of ±4 bp, ±8 bp, ±10 bp, ±20 bp, ±30 bp, and ±50 bp from a cell line sample of HG02080. These training datasets were split into 90% for training; and 10% for validation and model selection. To evaluate the selected model's generalization, we subsequently tested the model on an independent cell line sample of HG01109. In this part of the study, to minimize confounding effects from kinetic signal interactions, we restricted our analysis to windows containing only one CpG site. This yielded approximately 20 million training CpGs for ±50 bp windows, a sample size deemed adequate for robust read-model training (see Section 1.3). From these regions, we created ±4 bp, ±8 bp, ±10 bp, ±20 bp, and ±30 bp windows, thus isolating nucleotide window size as the sole experimental variable.

As shown in **Supplementary Fig. 1a**, in the validation subset, a ±10 bp window achieved optimal prediction accuracy (AUC=0.95; Accuracy=0.86), with no significant improvement observed for larger windows. Independent sample testing confirmed that this model maintained a similar AUC of 0.95 and accuracy of 0.86 at whole-genome scale, indicating that IPD and PW signals carrying information about 5mCpG occurrence are localized within ±10 bp of the CpG site. This result is consistent with a previous study[1], which demonstrated that the addition of a 5mC group to a CpG dinucleotide creates a kinetic footprint at positions 0, 2, 3, and 6 bases downstream of the CpG site as DNA polymerase moves along the template. Additionally, the ±10 bp window guarantees that the majority of CpG sites are singular within their designated regions. Considering the above results and constraints, we selected a 21-nt nucleotide window for the read-level model.

Furthermore, as expected, the read-level model at double-strand resolution exhibited higher accuracy than the single-strand resolution model, AUC: 0.96 vs 0.90 (**Supplementary Fig. 1b**). In a real-world situation, nearly all methylated CpG sites occur symmetrically on both strands[2]. Therefore, in MeLoDe, we assumed symmetric 5mCpG occurrence for the read-level model, and features from both the forward and reverse strands of a DNA template are combined as input. Nevertheless, we developed single-strand resolution models to enable the detection of hemimethylation (where only one strand is methylated) and to accommodate non-HiFi-Open reads (where only one strand is sequenced).

1.2. Model hyperparameter tuning

The CNN architecture of the read-level model was determined by searching a pre-defined parameter space using the *BayesianOptimization* function, which uses the Gaussian process to guide the selection of hyperparameter combinations to test. The cell line sample of HG002 was used as a training sample, and the training data was split into 90% for training; and 10% for validation. From 50 experimental tests, we selected the top-performing model with a smaller number of parameters, which contained 134, 593 parameters and achieved 0.86 accuracy (**Supplementary Table 2**).

1.3. Training dataset sample size

To evaluate the effect of training dataset size, we trained read-level models using varying numbers of unique CpGs (300K, 500K, 1M, 5M, 10M, 15M, 20M, and 30M) from cell line samples of HG02080, HG02723, and HG01109 and subsequently assessed model performance on an independent cell line sample of HG005 used in this study. Since multiple sequencing reads can be aligned to a single CpG site, we further evaluated the following two scenarios:

In the first scenario, we investigated whether including multiple reads per CpG site in the training data would improve model performance, although the sequence context is nearly identical across reads mapped to the same site (except for sequencing errors). We generated training data from 300K unique CpG sites (from multiple individuals), varying the coverage depth to 1, 2, 3, 5, 10, and 15 reads per site. As shown in **Supplementary Fig. 3a**, model accuracy improved progressively with increasing read depth. This result demonstrates that, given a fixed number of unique CpGs, especially when there are limited unique CpGs (say, below 5M, see below for details), including more reads improved the read-level model's performance.

In the second scenario, only one sequencing read is chosen to represent a CpG site, which allows us to determine the minimum number of unique CpG sites needed for read-level model training.

As shown in **Supplementary Fig. 3b**, model performance improved with increasing number of unique CpG sites, eventually converging at an AUC of 0.96 and an accuracy of 0.89. The performance curve revealed a transition point at ~10M CpG sites. Below this threshold, models exhibited substantial performance variation and suboptimal AUC and accuracy values, indicating insufficient training. This trend was consistent when using CpGs from multiple individuals; we observed that models trained on data from multiple individuals achieved higher accuracy than those trained on data from single individual when the number of unique CpG sites was fewer than 5 million.

In conclusion, both the read depth and the number of unique CpG sites contribute to read-level model performance. Therefore, in MeLoDe's implementation, the read-level models were trained using all available sequencing reads covering genome-wide CpG sites from multiple individuals ($n = 5$).

1.4. Influence of training sample ancestry

To investigate potential ancestry-related biases, we trained and tested read-level models using representative samples from diverse populations: European (HG002), African (HG02723), admixed American (HG01109), Southeast Asian (HG02080), and East Asian (HG005 [Chinese_1] and Ind02 [Chinese_2]). Results demonstrated that read-level models trained on a single ancestry achieved consistently high accuracy across all the other tested populations (**Supplementary Fig. 4**). This indicates that training sample ancestry does not significantly influence model performance, suggesting the model's broad applicability across diverse genetic backgrounds.

1.5. Read-level model tuning for read types

We classified sequencing reads from PacBio LRS into HiFi reads, non-HiFi-C reads, non-HiFi-O reads (**Fig. 1**). Using different read type sources, we trained variations of read-level model, followed by testing to find the best model for each read type. These models are distinguished by their training read types and source data, as specified below:

- Read-level (HiFi; HiFi): trained by HiFi data using features from HiFi consensus file;
- Read-level (HiFi; Subread): trained by HiFi data using features summarized from subreads;
- Read-level (non-HiFi; Subread): trained by non-HiFi reads using features summarized from subreads;
- Read-level (Mix; Subread): trained by mixture of both HiFi and non-HiFi reads using features summarized from subreads.

For this analysis, we performed four replications using cell line samples, ensuring no overlap between training and evaluation data, for example, training on HG02080 and testing on HG01109. The resulting comparison of the read-level model is presented in **Supplementary Fig. 5**. For HiFi data, models trained on HiFi reads showed higher prediction accuracy than those using subreads. For non-HiFi data (for which features could only be summarized from subreads), models trained on a combination of HiFi data and non-HiFi data demonstrated greater robustness.

Section 2: Optimization of the site-level model

MeLoDe's site-level model was developed to predict 5mC methylation frequency by jointly modelling a target CpG site and its surrounding CpG sites. Each CpG site is encoded as binned 5mC probabilities from its aligned reads. The model's architecture was first established through a hyperparameter search. Subsequently, factors potentially influencing model performance were profiled and characterized using naïve wild-type samples, as described below.

2.1. Surrounding CpGs included

To determine the optimal number of surrounding CpGs to be included in the site-level model, centered on a CpG site, using HiFi read data, we generated training datasets with ±5, ±10, ±15, ±25, and ±50 CpGs included in a study window, from the cell line sample HG02723. These training datasets were split into 90% for training; and 10% for validation, and model selection. To showcase the selected model's generalization, we subsequently tested the model on an independent cell line sample of HG005.

As shown in **Supplementary Fig. 2a**, including surrounding ±5 CpGs was sufficient to achieve a prediction accuracy that reached a plateau. Using more surrounding CpGs did not significantly contribute to higher prediction accuracy, as measured by concordance correlation coefficient (CCC), Pearson correlation coefficient (PCC), and mean absolute error (MAE). Therefore, we set the ±5 CpGs as the default for the site-level modeling. Independent sample (HG005) testing showed that this model reached a CCC of 0.91, a PCC of 0.91 and an MAE of 0.099.

2.2. Model hyperparameters tuning

The architecture of the site-level model was determined by searching a pre-defined parameter space using the *BayesianOptimization* function, which uses the Gaussian process to guide the selection of hyperparameter combinations to evaluate. The cell line sample of HG002 was used as a training sample, and training data were split into 90% for training; and 10% for validation. From 50 experimental tests, the best model was selected as default in the software implementation (**Supplementary Table 3**).

2.3. Training dataset sample size

To determine the number of focal CpGs required for model training, we trained site-level models using training datasets with 100K, 500K, 1M, 5M, 10M and 15M focal CpGs randomly selected from the whole genome, followed by testing these models on independent samples. Using cell line data, a total of four replications were conducted, for example, training on HG02080 and testing on HG01109. As shown in **Supplementary Fig 2b**, we found that at least 5M focal CpGs were needed to generate a stable and saturated site-level model, and models trained with fewer CpG sites than this number were likely to suffer from instability and reduced prediction accuracy, as measured by the CCC, PCC and MAE.

2.4. Comparison of different machine learning methods

We additionally developed the MeLoDe using a Transformer architecture. Specifically, we examined the performance of the read-level model and the site-level model when built upon either a Convolutional Neural Network (C) or a Transformer (T). A total of four combinations were then tested, namely "CNN + CNN (C+C)", "CNN + Transformer (C+T)", "Transformer + CNN (T+C)" and "Transformer + Transformer (T+T)". As shown in **Supplementary Fig. 7**, the selection of CNN or Transformer did not significantly influence the prediction accuracy. Therefore, we employed the "CNN + CNN (C+C)" combination as the default setting, where both the read-level model and the site-level model were constructed under the CNN framework.

2.5 Site-level model tuning for CpG site types

One issue that warrants consideration is whether to build a single site-level model using all read types to jointly predict H-sites, C-sites, and O-sites, or to model each of the three categories separately using a subset of aligned reads prioritized by read-level accuracy. To address this, we trained site-level model variations to determine the optimal read-type combination for each category. As listed below, four model variations were trained for each site type (12 models total). During training, we down-sampled three high-depth untreated DNA samples to generate datasets ranging from 3× to 15× template coverage. Two of the down-sampled datasets were used for training and validation (one blood sample, and one cell-line sample), while the third served as an independent test set (blood sample).

- S_1_HiFi: Uses surrounding CpGs with ≥1× HiFi read coverage; CpG encoding is based on HiFi-derived 5mCpG probabilities.
- S_4_HiFi: Uses surrounding CpGs with ≥4× HiFi coverage; CpG encoding is based on HiFi-derived probabilities.

- S_4_ZMW(nC): Uses surrounding CpGs with ≥4× combined HiFi and non-HiFi-C coverage; CpG encoding is based on probabilities from these two read types.
- S_4_ZMW(nC & nO): Uses surrounding CpGs with ≥4× total coverage from all three read types; CpG encoding is based on probabilities from all read types.

Model testing results demonstrated that for H-sites, models using HiFi reads alone achieved the highest accuracy, while further inclusion of non-HiFi reads in the model reduced the accuracy by 1.3–2.0% (**Supplementary Fig. 8a**), as measured by the concordance correlation coefficient (CCC), a metric evaluating both correlation and mean offset between predictions and ground truth. Conversely, for C-sites and O-sites, models incorporating both HiFi and non-HiFi reads achieved higher prediction accuracy than those using HiFi reads alone (**Supplementary Fig. 8a**). The reason for this improvement is that, for C-sites and O-sites, incorporating non-HiFi data enables the site-level model to leverage closer neighboring CpGs with strong correlations to the focal one, whereas with HiFi reads alone the usable neighboring CpGs are farther away and less informative (**Supplementary Fig. 8b**). In conclusion, for H-sites, the site-level model does not incorporate non-HiFi reads, whereas for C-sites and O-sites, non-HiFi reads are included to improve prediction accuracy.

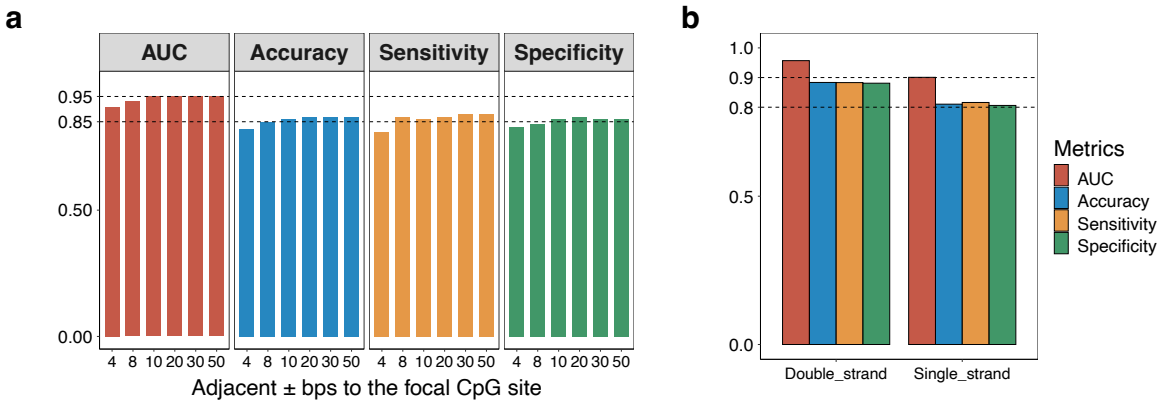
2.6 Training using population-scale data

In cohort-level applications, such as MWAS and meQTL analyses, the most informative CpG sites are those exhibiting high inter-individual variability. However, our initial site-level level model was trained on data from only a few individuals; consequently, 97% of the CpG sites in the training data showed little variability within the 1KCP sub-cohort ($n = 110$). Because the training process was dominated by these conserved sites, the model was not optimized for variable ones. To address this, we stratified CpG sites as either 'variable' (vGpGs) or 'conserved' (cCpGs) based on their inter-individual variance.

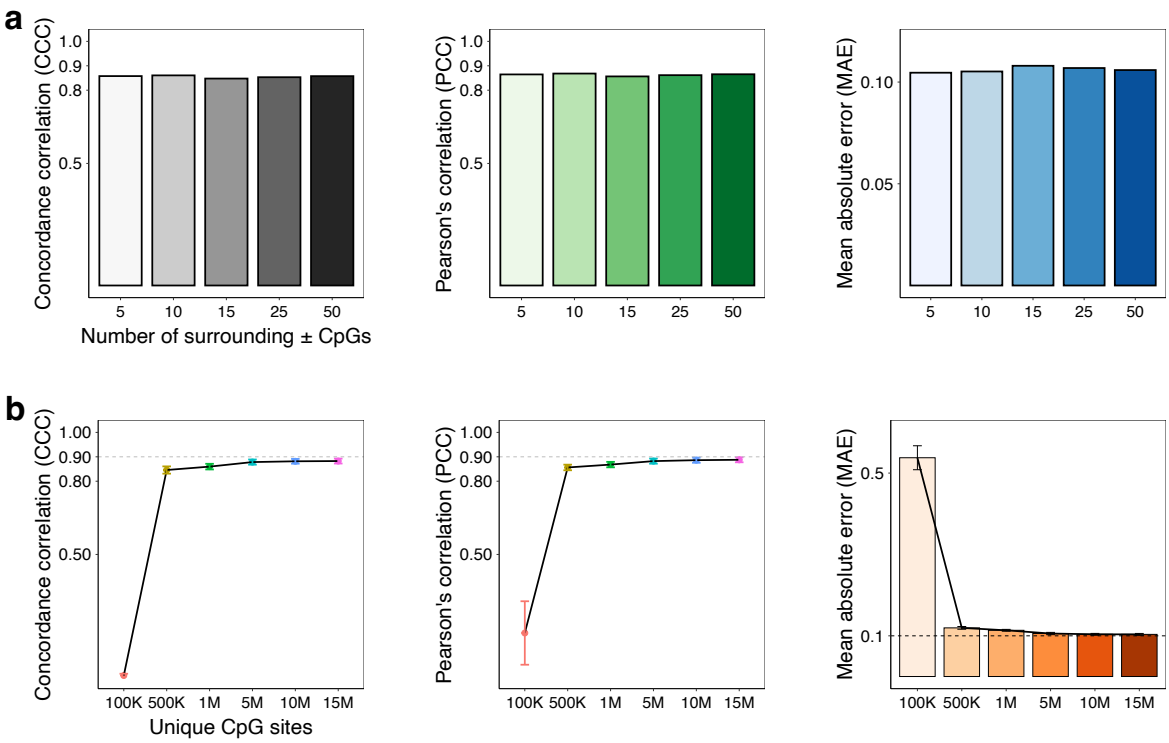
Specifically, we calculated inter-individual methylation variance from the 1KCP sub-cohort WGBS data ($n = 110$). After examining the variance distribution, we set a threshold of ~2.70 (corresponding to two standard deviations above the mean variance of 0.70) to classify sites. Using the focal GpG sites as indices, we then created training datasets consisting of only vCpGs, only cCpGs, or mixtures of the two in various ratios (**Supplementary Table 4**). To minimize the risk of overfitting, we implemented a leave-chromosome-out validation strategy, e.g., training on chromosomes 2–22 and evaluating on the held-out chromosome 1. As shown in **Supplementary Fig. 9**, a model trained on a 1:1 ratio of variable and conserved sites achieved the best overall performance.

Section 3: Haplotype resolved prediction

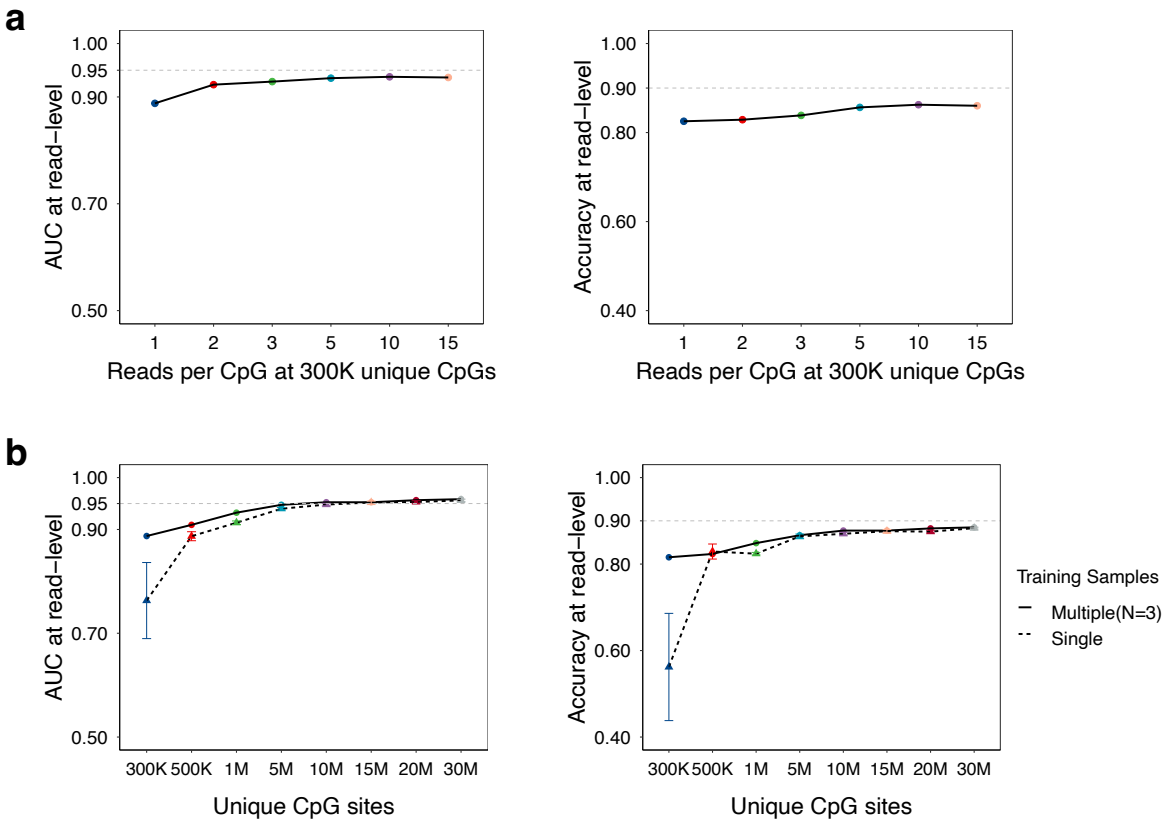
We created haplotype-resolved predictions following the steps described below: First, the DeepVariant[3] software was used to determine genetic variants from the alignment BAM file of the 30× HiFi reads data. Then, the genetic variants were phased using the whatshap *phase* command[4], with the long read BAM file as a supporting input file. In the output phased VCF file, variants located on the same haplotype were marked with the same phase set identifier (*PS* tag). Finally, the whatshap *haplotag* command was used to tag the BAM file. In the haplotype tagged BAM file, reads were labelled with "*HP:i:1*" or "*HP:i:2*" depending on which haplotype they belonged to. Additionally, a *PS* tag was added to describe in which haplotype block the read was located in.



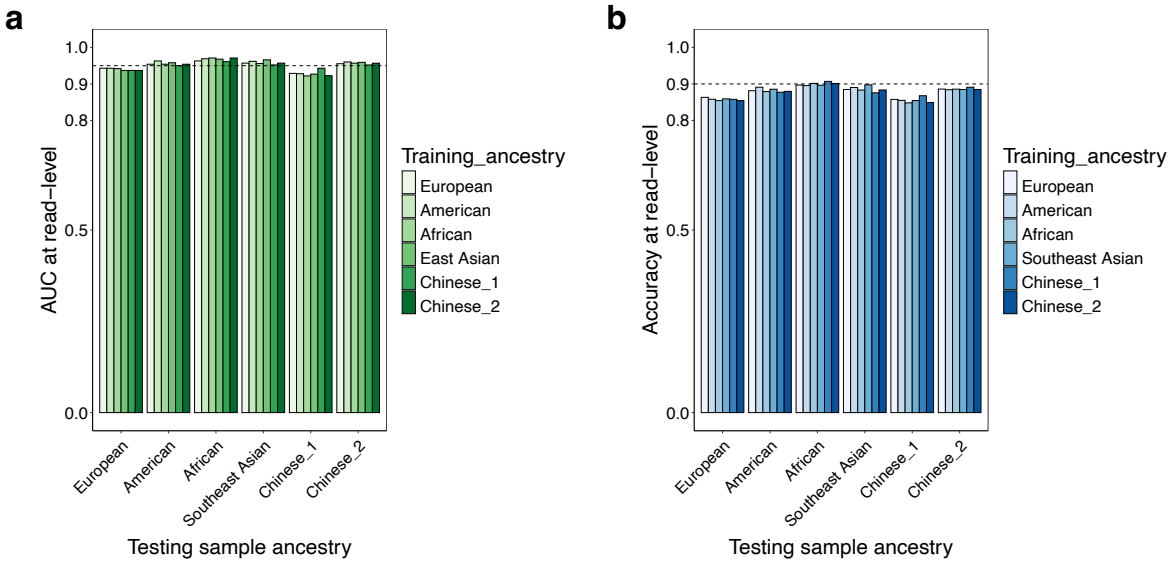
Supplementary Figure 1. Read-level model structure. a, Number of adjacent base pairs included in the read-level model. **b**, Read-level model in double-strand and single-strand resolution. "Double_strand" denotes models utilizing features from both forward and reverse strands of a DNA template, and "Single_strand" denotes models using features from only one strand.



Supplementary Figure 2. Influential factors for the site-level model. a, Number of surrounding CpG sites included. **b,** Number of unique CpG sites in the training dataset.



Supplementary Figure 3. Influence of training dataset size on read-level models. a, Influence of reads depth. Given 300K unique CpG sites, models are trained using 1, 3, 5, 10 and 15 reads per CpG site. **b,** Influence of unique training CpG sites. For each CpG site only one read is selected to represent the CpG-dependent sequence context.

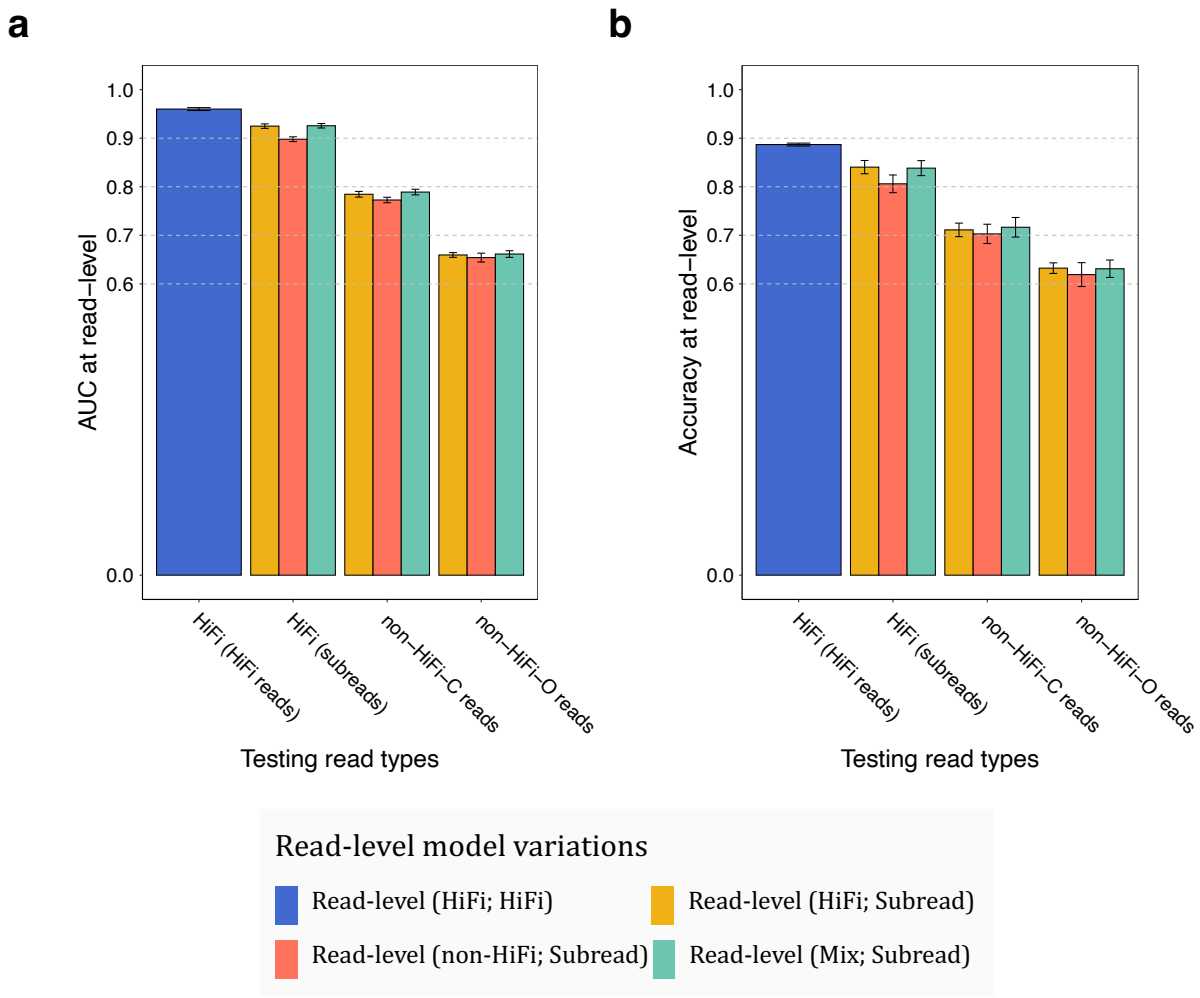


252

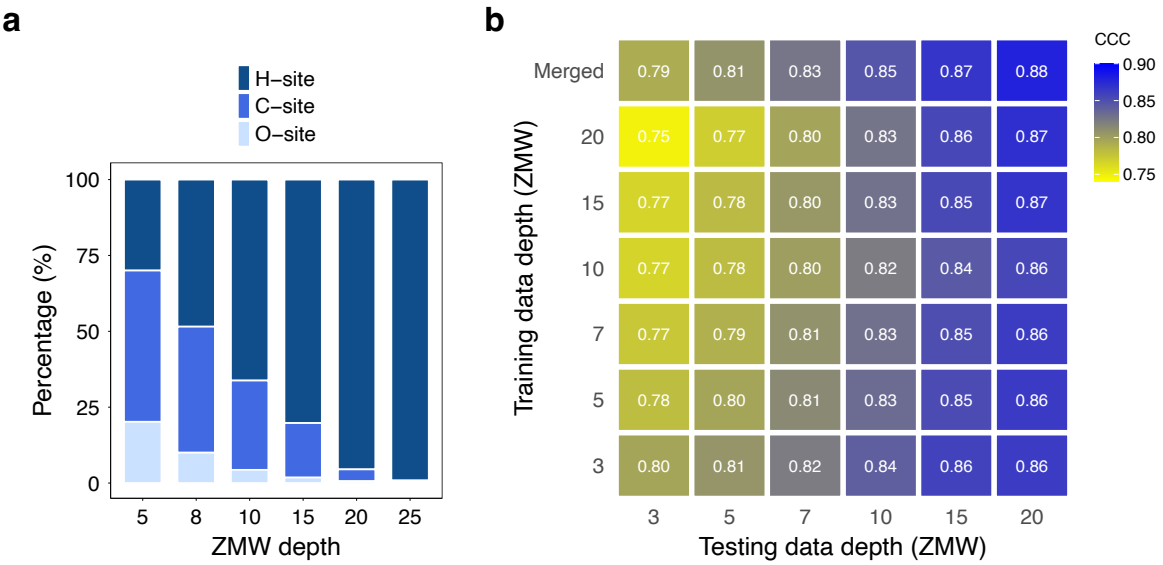
253 **Supplementary Figure 4. Influence of training sample ancestry on read-level models. a,**

254 Prediction AUC. **b,** Prediction accuracy.

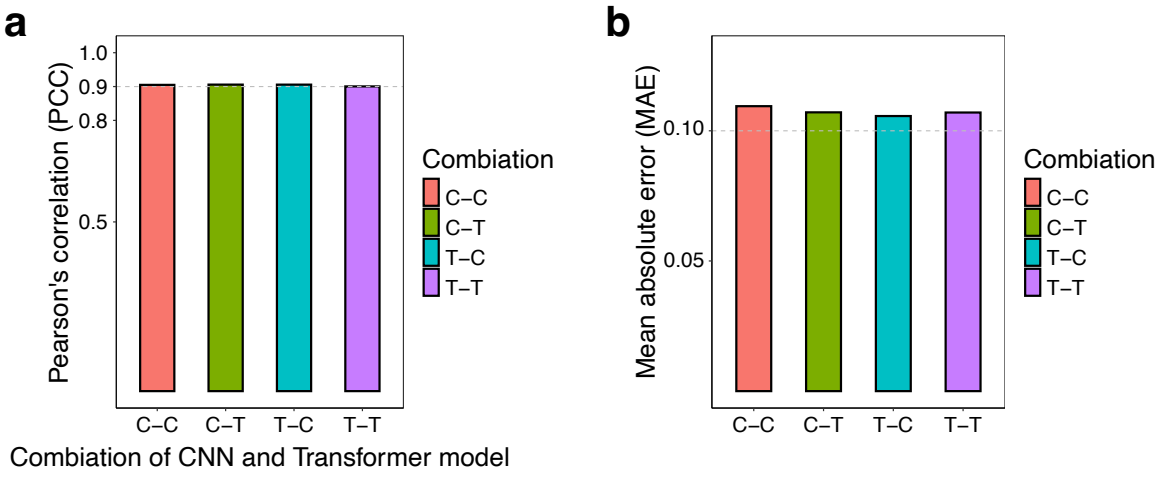
255



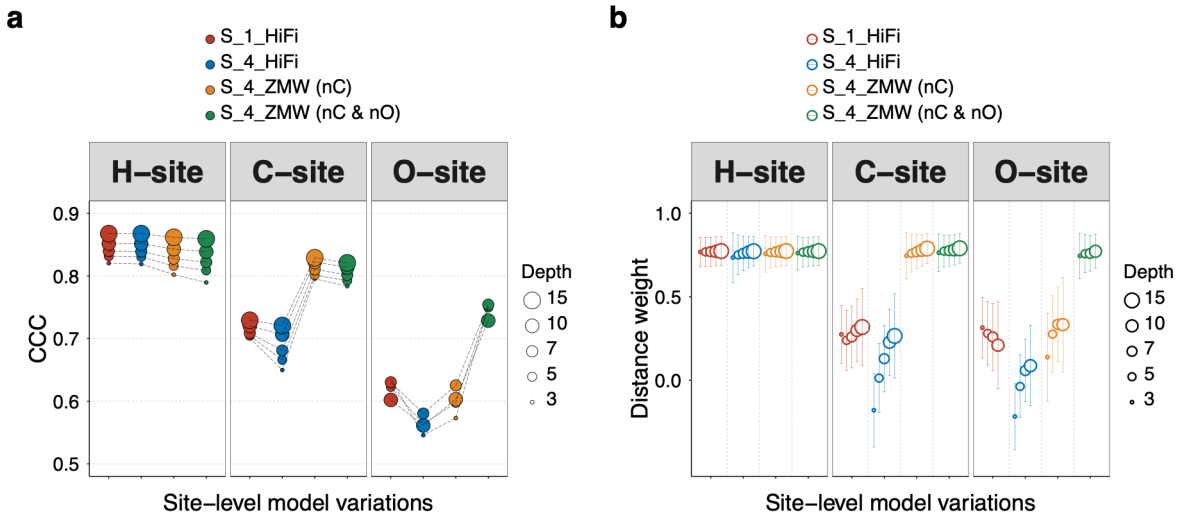
Supplementary Figure 5. Read-level model tuning. a, Prediction AUC. **b,** Prediction accuracy. Read-level model variations are distinguished in terms of the training read types and source data origin, as described below: Read-level (HiFi; HiFi): trained by HiFi data using features from HiFi reads file; Read-level (HiFi; Subread): trained by HiFi data using features summarized from subreads; Read-level (non-HiFi; Subread): trained by non-HiFi reads using features summarized from subreads; Read-level (Mix; Subread): trained with a mixture of both HiFi and non-HiFi reads using features summarized from subreads.



Supplementary Figure 6. CpG sites classification and impact of training data depth on the site-level model performance. a, Proportions of CpG classes. CpG classes are defined as follows (depth threshold at 4×): H-site: ≥4× coverage from HiFi reads; C-site: HiFi<4× & (HiFi + non-HiFi-C) ≥4×; O-site: HiFi<4× & (HiFi + non-HiFi-C<4× & HiFi + non-HiFi-C + non-HiFi-O) ≥4×. In the calculation, the denominators are CpG sites with ≥4× DNA template coverage. **b,** Site-level model across depths of coverage. Concordance correlation coefficients (CCC) values are plotted. The x-axis shows the testing data depth measured in ZMW molecules, and the y-axis shows the training data depth. *Merged* refers to data from individual sequencing depths (3×–20×) combined into a single unified training dataset.



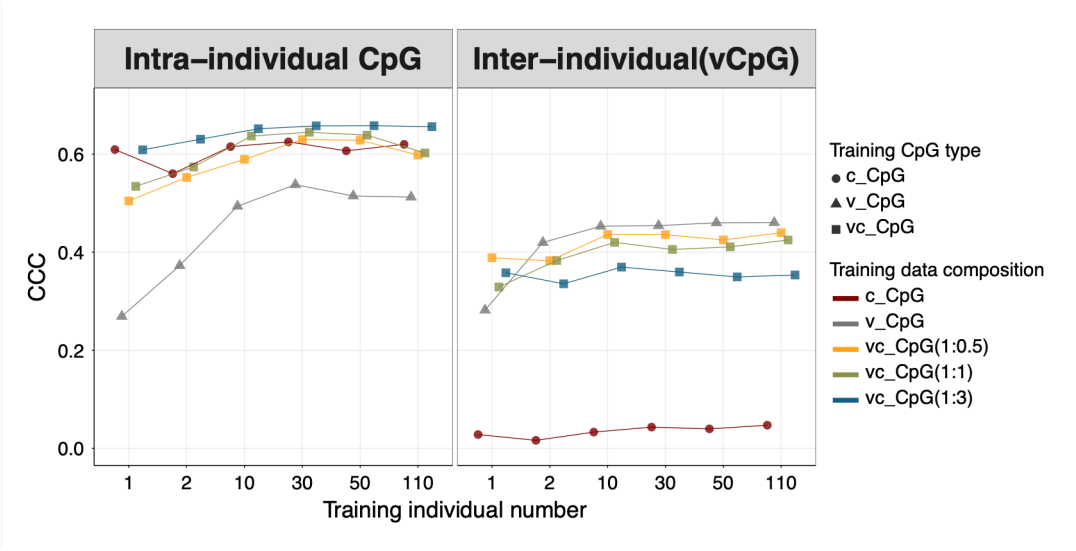
Supplementary Figure 7. Comparison of CNN and Transformer architecture for the read-level model and site-level models. CNN and Transformer machine learning methods were tested for the read-level model and the site-level model. "C-C" denotes the use of CNN for both the read-level model and site-level models. Similarly, "C-T", "T-C", and "T-T" denote the use of CNN-Transformer, Transformer-CNN, and Transformer-Transformer combinations, respectively. Pearson's correlation coefficient (PCC) and mean absolute error (MAE) are shown in panel **a** and **b**, respectively.



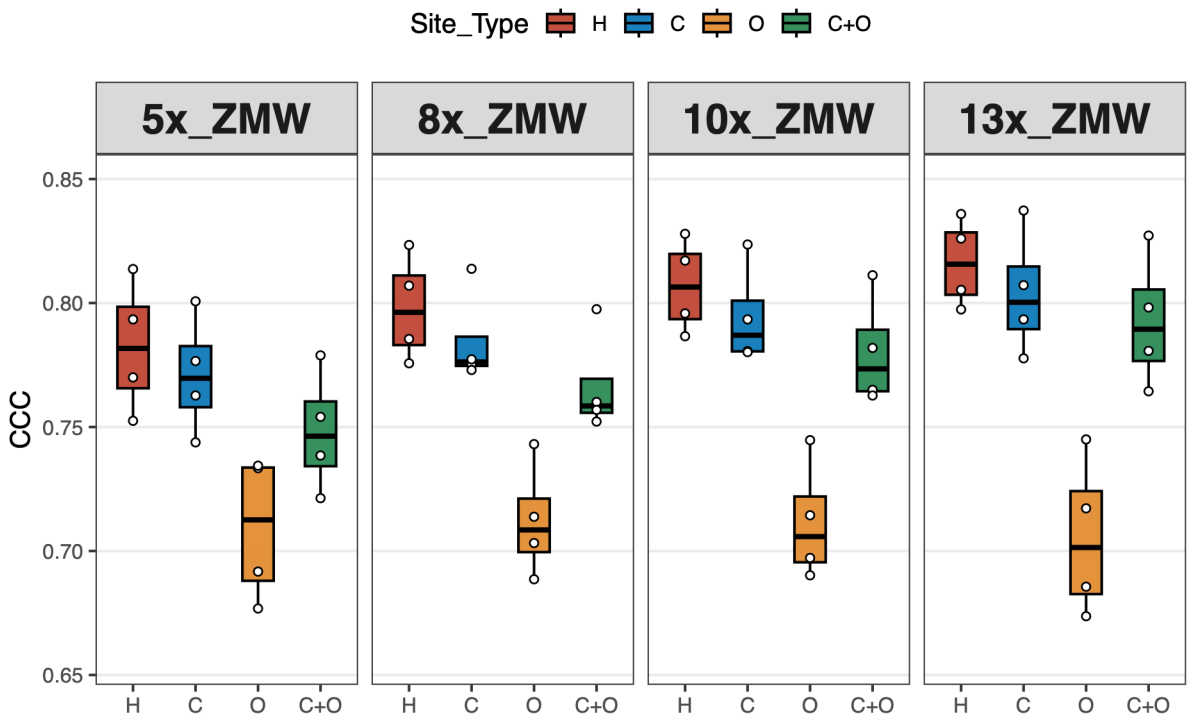
285

286 **Supplementary Figure 8. MeLoDe site-level model training and tuning. a,** Comparison of
287 site-level model variations. Prediction accuracy, measured by the concordance correlation
288 coefficient (CCC), is shown for models using different data sources to select and encode the
289 surrounding CpG sites. The models are: S_1_HiFi (≥1× HiFi coverage), S_4_HiFi (≥4× HiFi
290 coverage), S_4_ZMW(nC) (≥4× combined HiFi and non-HiFi-C coverage), and S_4_ZMW(nC & nO)
291 (≥4× coverage from all read types). **b,** Distance weight from surrounding CpG sites to the focal
292 site across different model variations. In panels a and b, results for O-sites at 15× depth are
293 censored due to the low number of qualifying sites at higher sequencing depths.

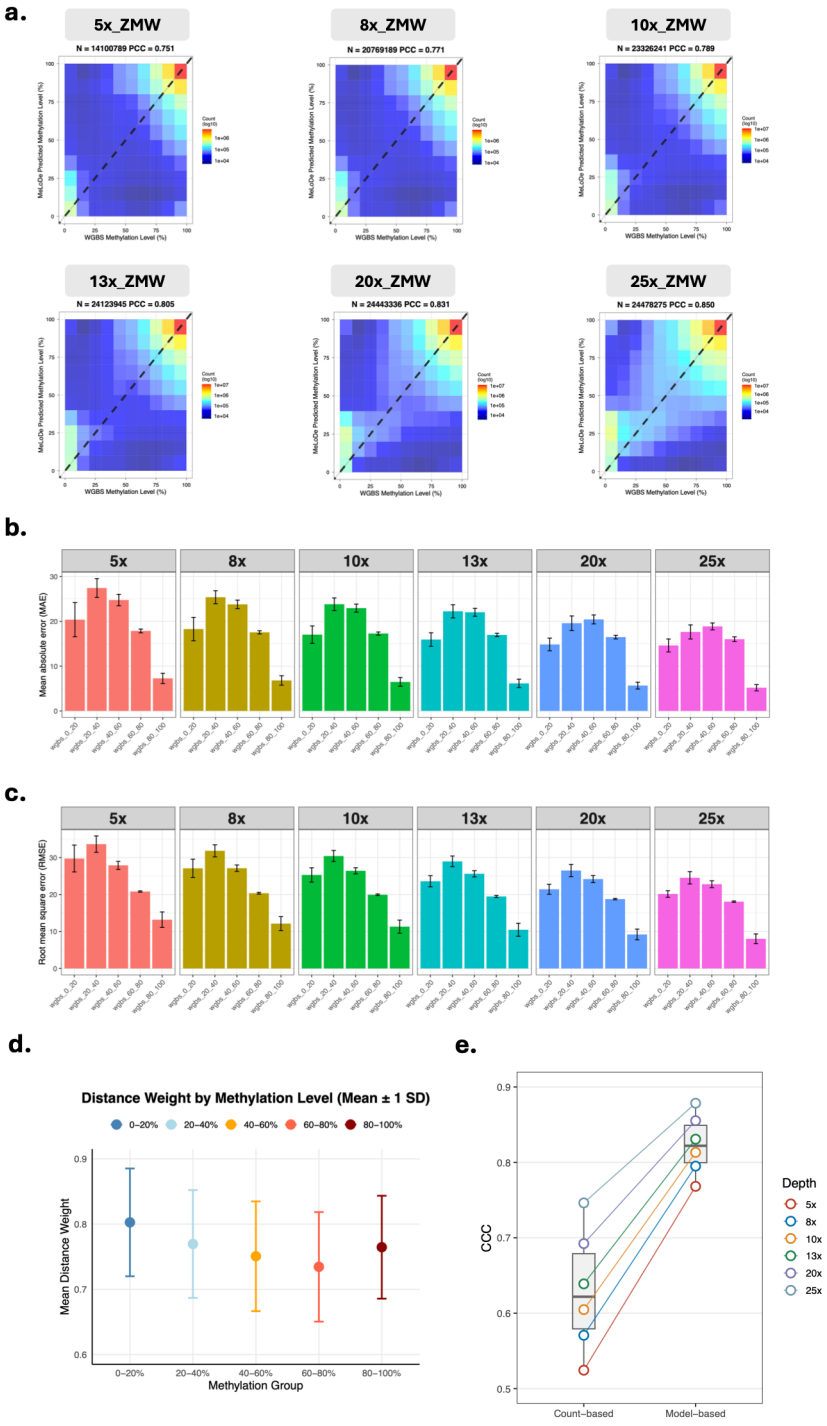
294



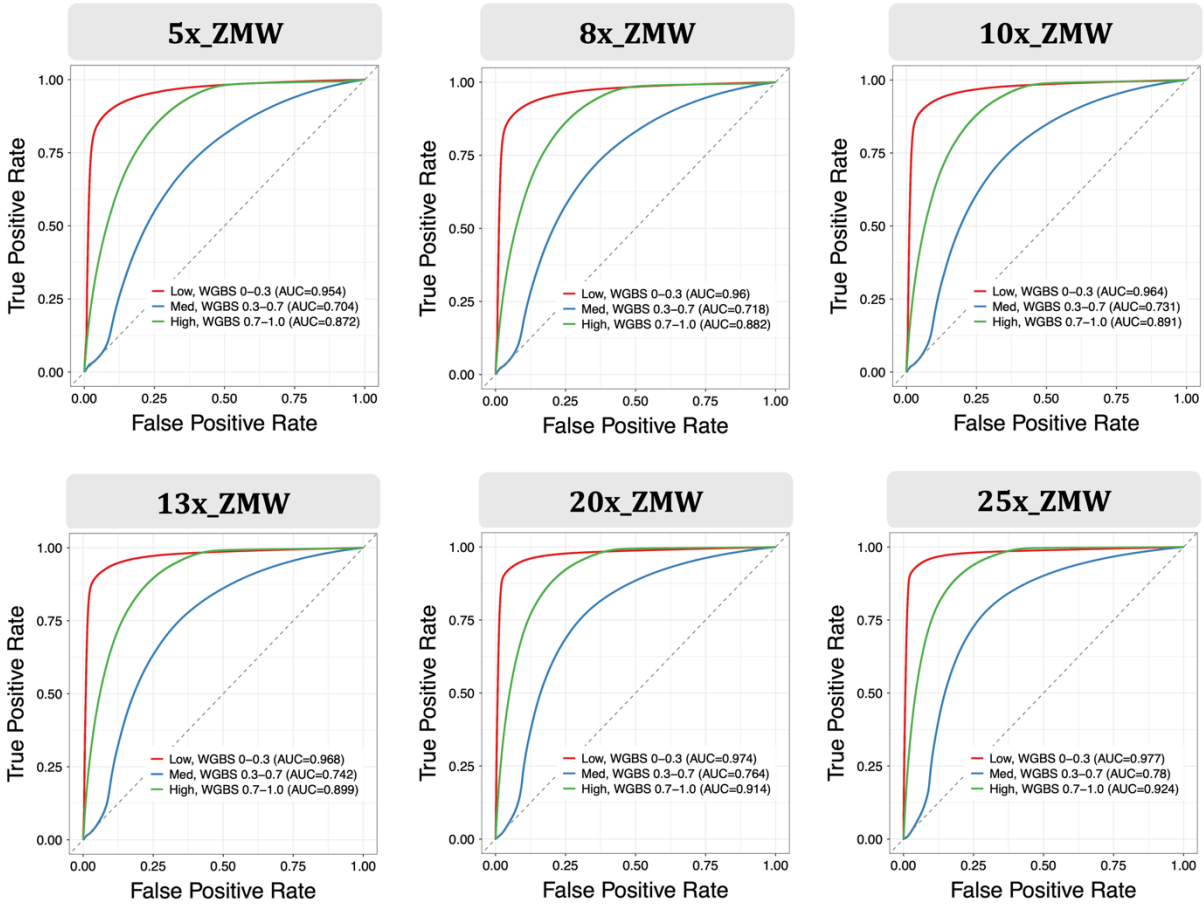
Supplementary Figure 9. MeLoDe model training on population-scale data. Optimizing the training data composition for the site-level model. Models were trained on datasets with varying proportions of variable (vCpGs) and conserved (cCpGs) sites, and tested on independent held-out datasets. The classifications of vCpGs and cCpGs are based on inter-individual methylation frequency variance calculated from WGBS data.



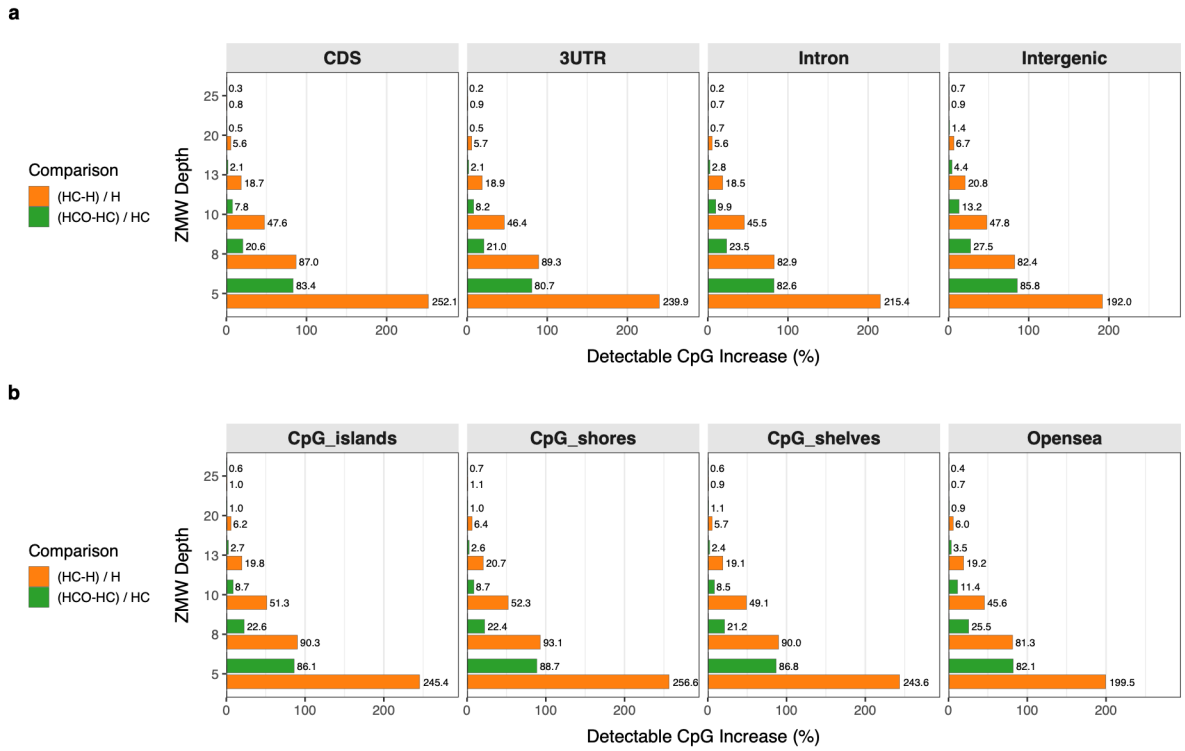
Supplementary Figure 10. Prediction accuracy of MeLoDe for different CpG sites. Site classes are defined using a 4× coverage threshold: H-sites (HiFi coverage ≥4×), C-sites (HiFi <4× and HiFi + non-HiFi-C ≥4×), and O-sites (HiFi + non-HiFi-C <4× and total coverage ≥4×).



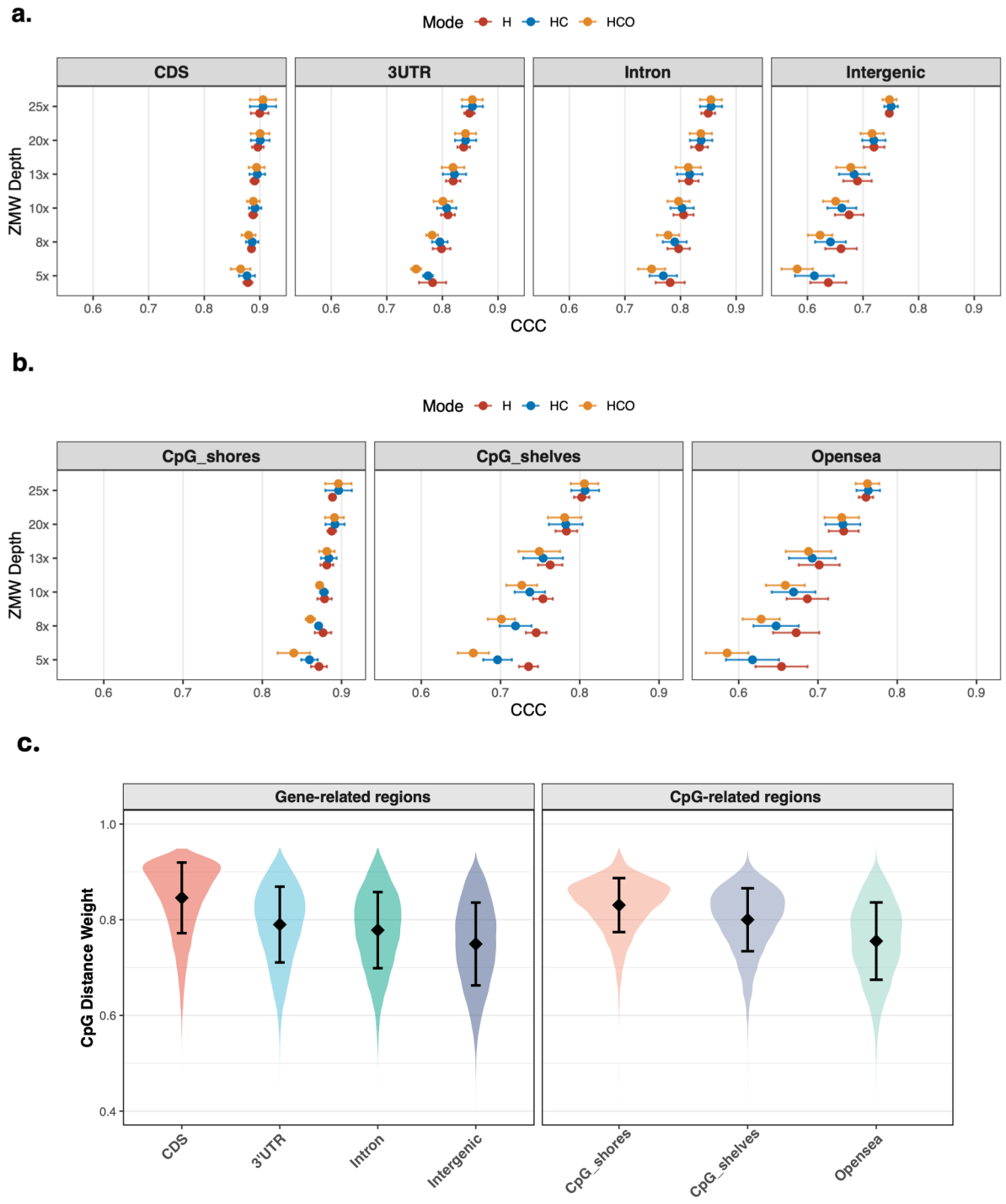
Supplementary Figure 11. MeLoDe-HCO performance. **a**, Pearson correlation coefficient (PCC) with WGBS. **b**, Mean absolute error (MAE). **c**, Root mean square error (RMSE). **d**, Focal CpG distance weight stratified by WGBS-based methylation bins. In panel **a**, predicted values largely follow the diagonal line across the 0–1 methylation spectrum (PCC = 0.75–0.85), indicating strong agreement with WGBS. Elevated MAE (panel **b**) and RMSE (panel **c**) are observed at intermediate methylation levels, coinciding with reduced distance weights in this range (panel **d**). **e**, Model-based and count-based prediction.



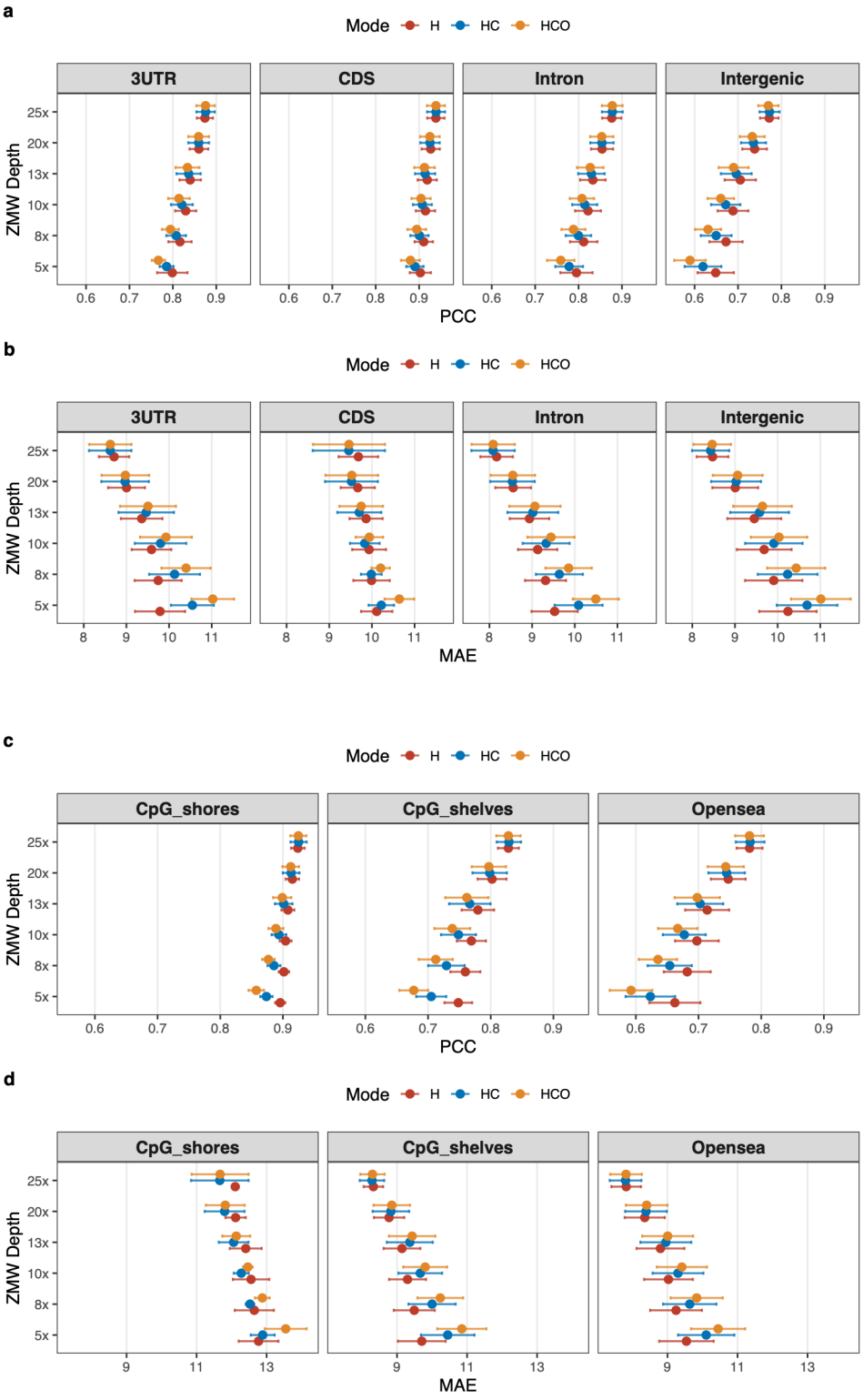
Supplementary Figure 12. Classification accuracy for MeLoDe methylation calls. CpG sites are stratified into three groups by WGBS methylation level: low (0–0.3, red), intermediate (0.3–0.7, blue), and high (0.7–1.0, green). One-vs-rest AUC is calculated for each group using MeLoDe-predicted values. The low- and high-methylation groups achieve high AUC (0.87–0.98) and together comprise 92% of sites, whereas the intermediate group (8% of sites) shows lower AUC (0.70–0.78).



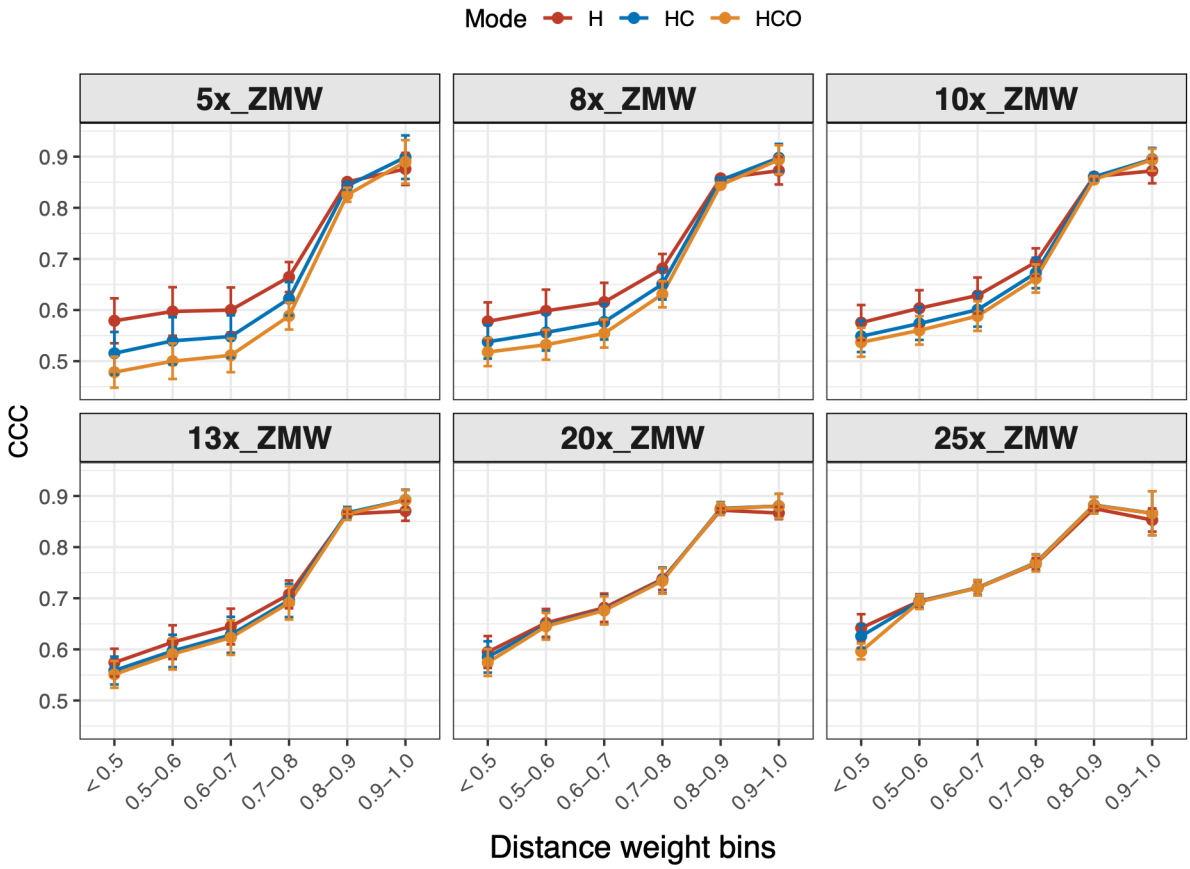
Supplementary Figure 13. Percentage increase of detectable CpG sites across genomic contexts. Detectable CpG increase (%) is defined as the relative gain in detectable CpGs between MeLoDe modes, such as (HCO − HC)/HC × 100% for MeLoDe-HCO over MeLoDe-HC, plotted against ZMW depth. a, CpGs located in gene-related regions (CDS, 3′ UTR, intron, and intergenic). b, CpGs located in CpG-related features (CpG islands, CpG shores, CpG shelves, and open sea). Orange bars, (HC − H)/H; green bars, (HCO − HC)/HC.



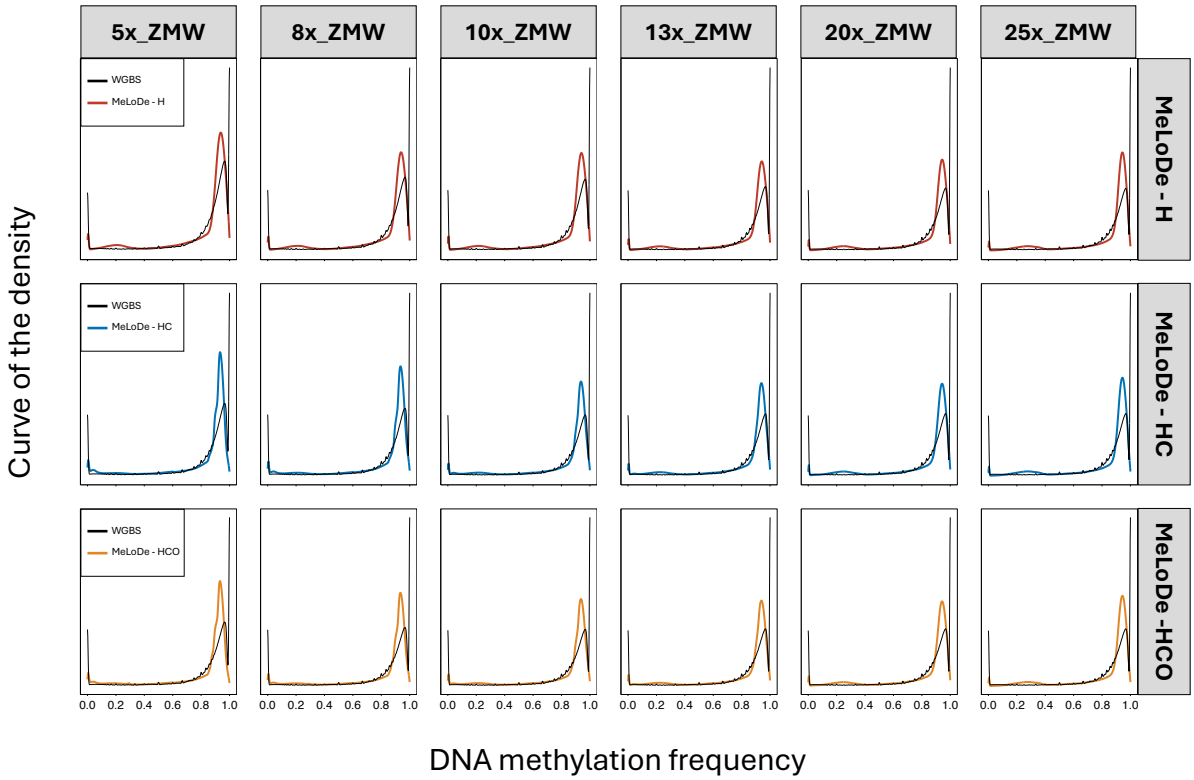
Supplementary Figure 14. MeLoDe performance across genomic contexts (in CCC). **a**, CCCs for CpGs in gene-related regions (CDS, 3′ UTR, intron, and intergenic). **b**, CCCs for CpGs in CpG-related regions (CpG shores, CpG shelves, and open sea). GC-rich CpG islands are not included because of their reduced read depth in the ground-truth WGBS data. **c**, Distance weight across genomic contexts.



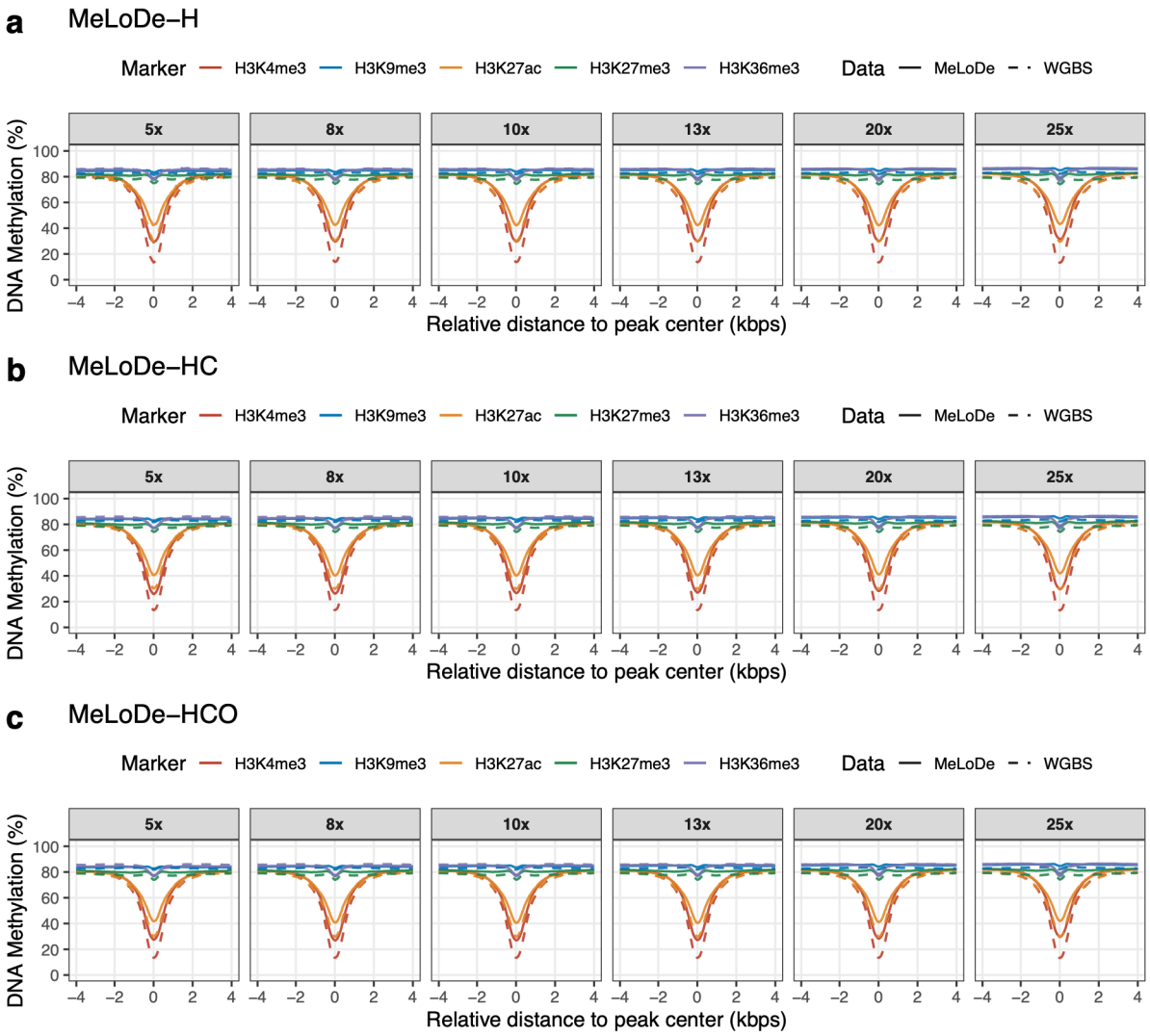
Supplementary Figure 15. MeLoDe performance across genomic contexts (in PCC and MAE). (**a**, **b**) PCC and MAE for CpGs in gene-related regions (CDS, 3′ UTR, intron, and intergenic). (**c**, **d**) PCC and MAE for CpGs in CpG-related regions (CpG shores, CpG shelves, and open sea). GC-rich CpG islands are not included because of their reduced read depth in the ground-truth WGBS data.



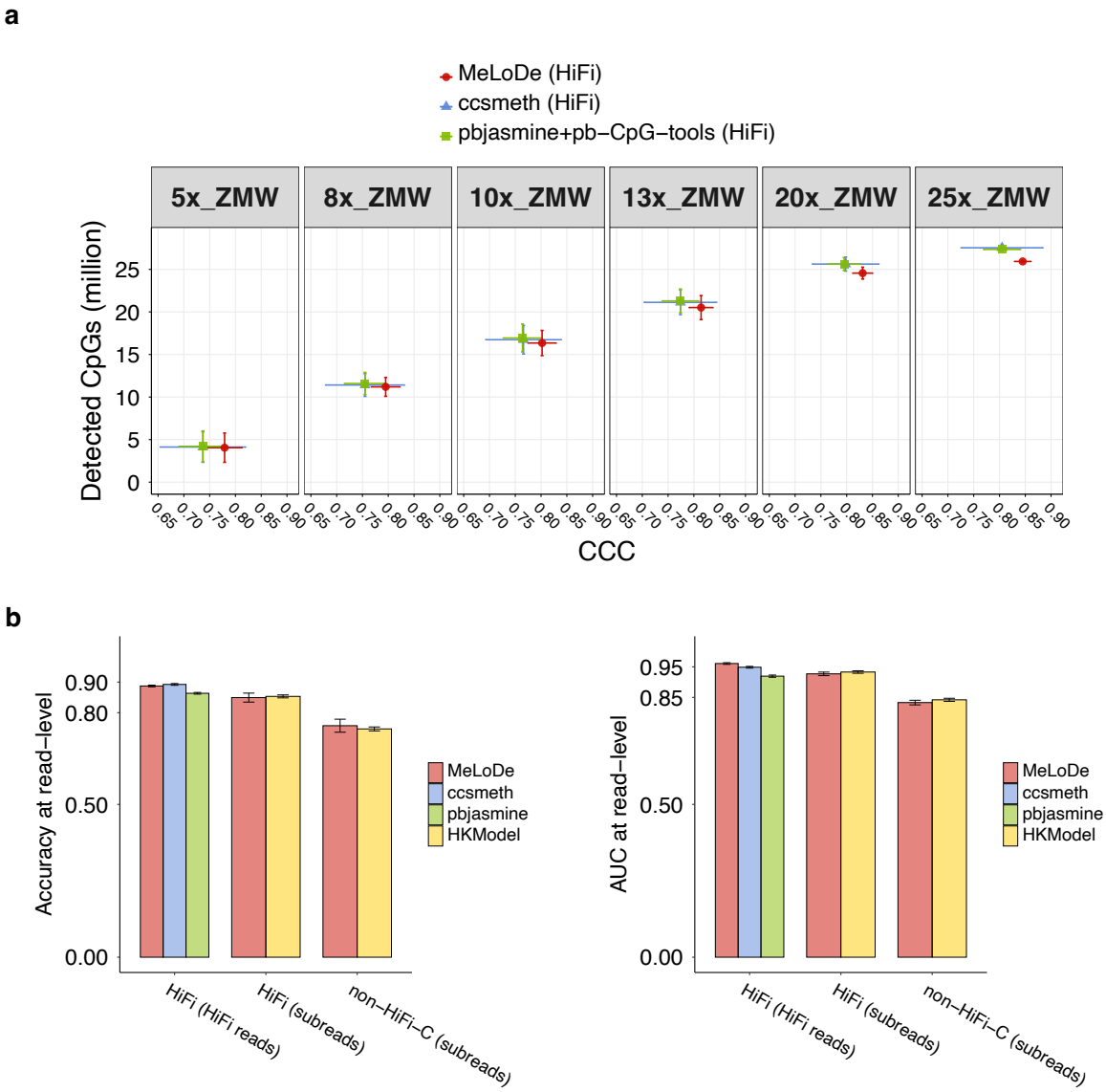
Supplementary Figure 16. Prediction accuracy of MeLoDe stratified by distance weight. The x-axis represents binned distance weight intervals, where higher distance weights denote closer genomic proximity. The y-axis shows prediction accuracy as measured by the concordance correlation coefficient (CCC). Accuracy increases with proximity, with an inflection point at distance weights of 0.7–0.8 (CCC ~0.7; roughly 625–1,836 bp). Approximately ~80% of human CpG sites have a distance weight ≥0.7.



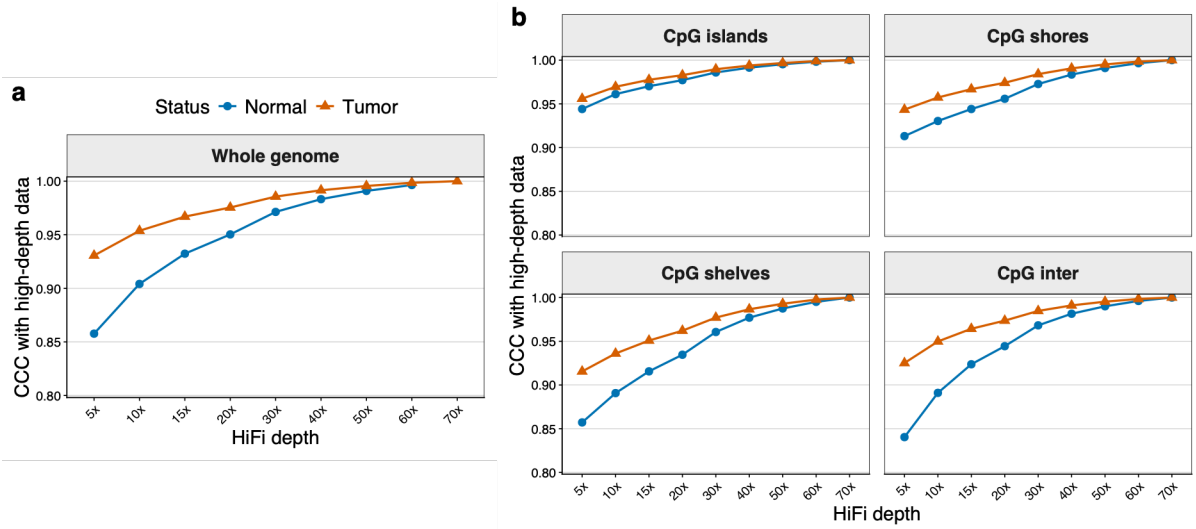
Supplementary Figure 17. Bimodal distribution of DNA methylation frequencies predicted by MeLoDe.



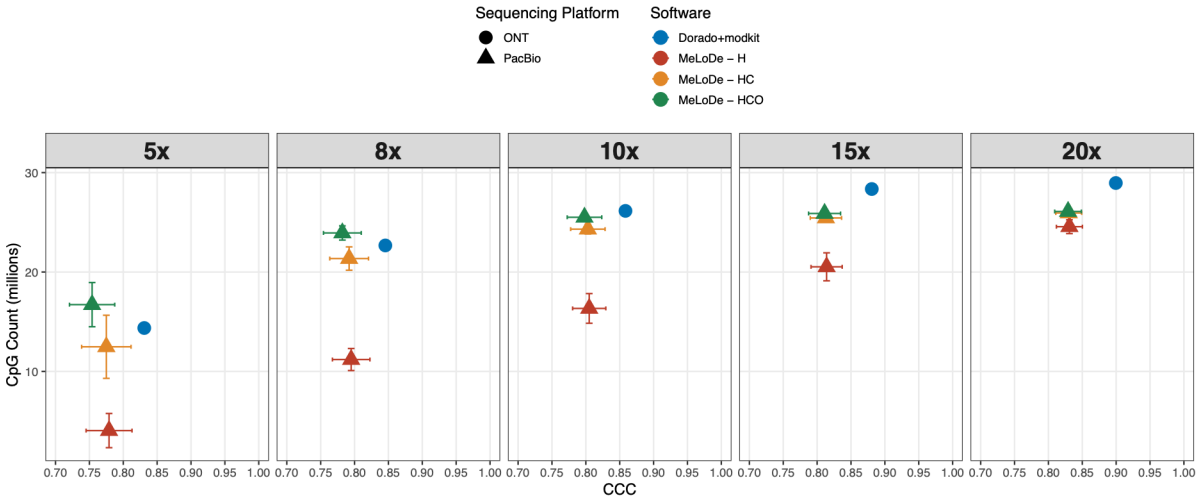
Supplementary Figure 18. DNA methylation at regulatory elements by MeLoDe. The x-axis shows the relative distance to the center of regulatory elements as defined by the ENCODE project[5]. Shown in panel **a**, **b** and **c** are predictions from MeLoDe-H, MeLoDe-HC and MeLoDe-HCO, respectively.



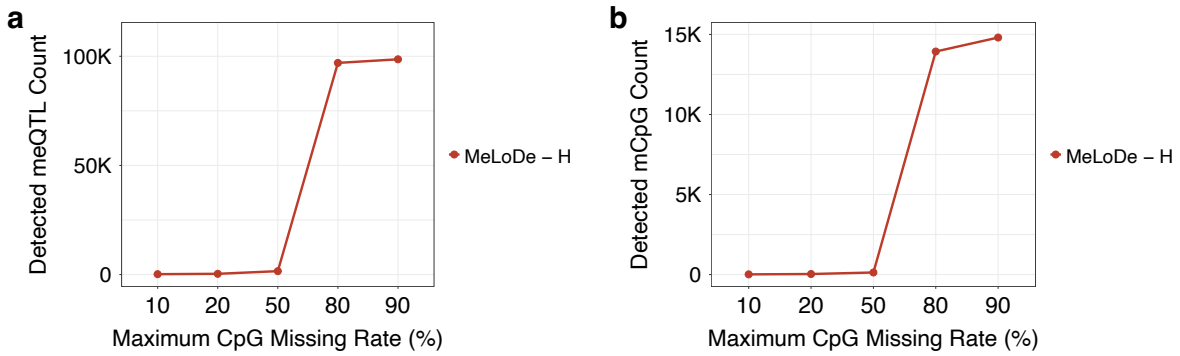
Supplementary Figure 19. Benchmarking of MeLoDe on HiFi reads. a, Site-level comparison. CCC (x-axis): Concordance correlation coefficients between predicted and ground-truth methylation frequency. Y-axis: Number of CpG sites detected. HiFi reads were generated from each ZMW. **b**, Read-level comparison. The x-axis represents testing data with the input data format used in parentheses: HiFi (HiFi reads): HiFi data using HiFi reads; HiFi (subreads): HiFi data using subreads; non-HiFi-C (subreads): non-HiFi-Closed reads using subreads. Non-HiFi-Open reads, which can only be processed by MeLoDe, are not included in the benchmarking.



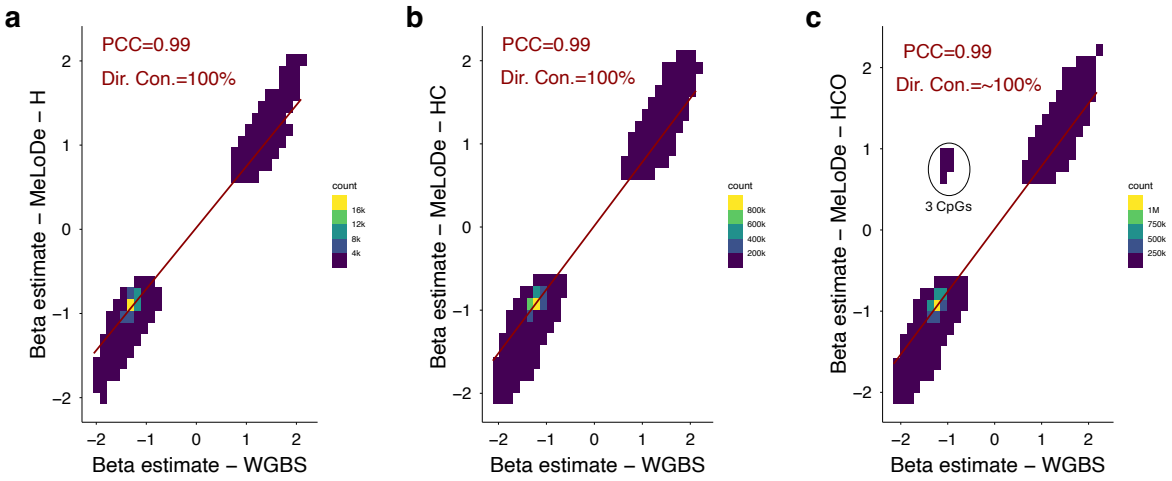
Supplementary Figure 20. Accuracy of HiFi-only methylation calling as a function of HiFi sequencing depth in the COLO829 cancer dataset. CpGs from the whole genome and CpGs overlapping with CpG-related features, including CpG islands, CpG shores, CpG shelves and CpG inter (open sea) are shown in panel **a** and **b**, respectively. Publicly available cell line datasets of COLO829, including tumor and normal HiFi reads, are used in this figure. Lacking matched WGBS data, methylation frequencies from full-depth HiFi data (~70× HiFi) serve as the reference; down-sampled depths span 5×–70× in 5× or 10× increments.



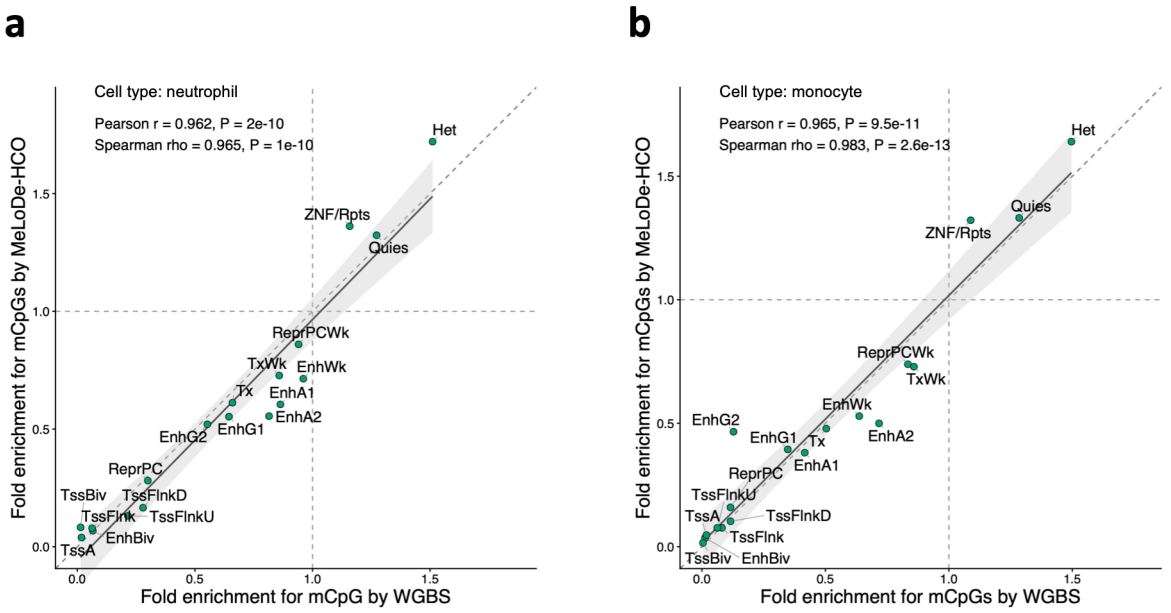
Supplementary Figure 21. Comparison of ONT-based DNA methylation prediction with PacBio-based MeLoDe prediction. Public ONT data (FAST5 and FASTQ) from the HG02080 cell line were processed using Dorado with the dna_r9.4.1_e8_hac@v3.3 model for read-level 5mC methylation calling, followed by modkit for CpG site-level prediction. ONT reads were down-sampled to match the sequencing depth of the PacBio dataset (ZMW-equivalent) for comparison. The x-axis shows the CCCs, and the y-axis shows the number of detected CpG sites with methylation estimates.



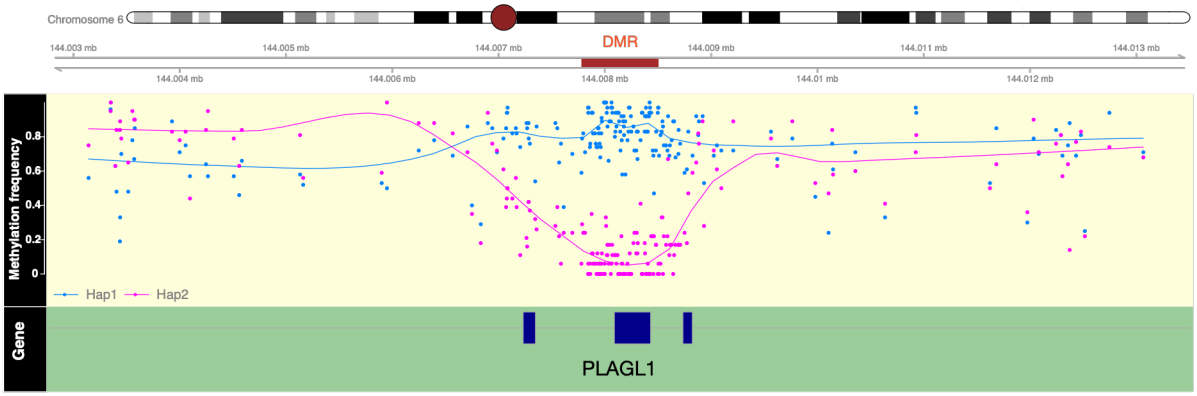
Supplementary Figure 22. Detected meQTL and mCpG using HiFi reads only. CpG methylation frequency was predicted by MeLoDe–H. The allowed missingness rate is shown on the x-axis. Results for meQTLs and mCpGs are shown in panels **a** and **b**, respectively.



Supplementary Figure 23. Comparison of meQTL effects based on MeLoDe calls with those based on WGBS calls. Person's Correlation Coefficient (PCC) and concordance of effect direction (Dir. Con.) for significant meQTLs effect size derived from MeLoDe-CpGs (y-axis) versus WGBS-CpGs (x-axis). Panels **a**, **b**, **c** denote CpGs detected by MeLoDe - H, MeLoDe - HC and MeLoDe - HCO, respectively.



Supplementary Figure 24. Regulatory element enrichment of mCpGs derived from WGBS and MeLoDe. Candidate *cis*-regulatory elements (cCREs) are tested for neutrophils in panel **a** (ENCODE Accession: ENCSR042TMK) and for monocytes in panel **b** (ENCODE Accession: ENCSR108EPF). These two cell types account for 57% and 38% of the cellular composition of our blood samples, respectively. Full names of these cCREs are listed as follows: EnhA1 (strong active enhancer), EnhA2 (weak or moderate active enhancer), EnhBiv (bivalent/poised enhancer), EnhG1 (active enhancer overlapping gene bodies), EnhG2 (weak genic enhancer), EnhWk (weak or putative enhancer with low signal), TssA (active promoter), TssBiv (bivalent/poised promoter), TssFlnk (flanking region upstream of TSS), TssFlnkD (downstream flanking region of TSS), TssFlnkU (upstream flanking region of TSS), Tx (actively transcribed gene body), TxWk (weakly transcribed region), ReprPC (Polycomb-repressed region), ReprPCWk (weakly Polycomb-repressed region), Het (constitutive heterochromatin), Quies (quiescent/silent region), and ZNF/Rpts (zinc finger genes and repetitive regions).



Supplementary Figure 25. Differentially methylated region between two haplotypes. Each dot represents a CpG site, and the methylation frequency (y-axis) was predicted by MeLoDe - H in haplotype resolution for HG005 using HiFi reads (count-based prediction). A differentially methylated region (DMR) between haplotypes is labelled.

Supplementary Table 1. Summary of read-level model and site-level model

Models	Analysis unit	Input features	Measurement window	Training samples	Training labels	Model output	Evaluation metric	Interpretation
Read-level	CpGs captured by single read	IPD, PW, sequence context	Focal CpG ± 10 bps	fully-methylated and fully-unmethylated DNA samples	0 or 1	[0,1]	AUC of the ROC curve	Probability of 5mC at the read level
Site-level	CpG sites on genome	Read-level prediction probabilities	Focal_CpG ± 5 surrounding CpGs	Untreated DNA samples, with CpG methylation levels quantified by WGBS	[0,1]	[0,1]	Concordance correlation coefficient (CCC)	Methylation level of the CpG site on genome level

Supplementary Table 2. Read-level model hyperparameter tuning (top10)

rank	accuracy	conv_layers (count)	conv_filters (count)	kernel_size	dropout_rate	dense_layers (count)	dense_0 (units)	dense_1 (units)	dense_2 (units)	dense_3 (units)	dense_4 (units)
1	0.86	5	256	5	0.2	5	128	128	32	128	32
2*	0.86	5	256	5	0.2	2	128	32	0	0	0
3	0.86	5	256	5	0.2	5	32	128	128	128	128
4	0.86	5	256	5	0.4	2	32	32	0	0	0
5	0.86	5	256	5	0.2	5	128	32	128	128	128
6	0.86	5	256	5	0.2	5	128	128	32	128	128
7	0.86	5	256	5	0.2	5	32	32	128	128	32
8	0.86	5	256	5	0.2	5	32	32	128	128	32
9	0.86	5	256	5	0.2	5	32	32	32	128	128
10	0.86	5	256	5	0.2	2	128	32	32	128	32

Note: conv_layers (count): number of Conv1D layers; conv_filters (count): number of filters in each Conv1D layer; kernel_size: kernel size; dropout_rate: dropout rate; dense_layers (count): number of dense layers; dense (units): number of units in a dense layer. The selected read-level model in MeLoDe is marked with an asterisk.

Supplementary Table 3. Site-level model hyperparameter tuning (top10)

rank	mean absolute error	conv_layers (count)	conv_filters	kernel_size	dropout_rate	dense_layers (count)	dense_0 (units)	dense_1 (units)	dense_2 (units)
1	0.074	2	96	4	0.3	3	30	10	10
2	0.074	2	96	3	0.3	3	30	10	10
3	0.075	2	64	2	0.3	2	30	20	NA
4	0.075	2	96	3	0.3	3	30	10	10
5	0.075	2	96	3	0.4	3	30	10	10
6	0.075	2	96	3	0.4	3	30	10	10
7	0.076	2	96	3	0.4	3	30	10	10
8	0.076	3	32	3	0.3	3	20	20	10
9	0.076	2	64	2	0.3	2	30	20	NA
10	0.076	3	32	3	0.3	3	20	20	10

Note: conv_layers (count): number of Conv1D layers; conv_filters (count): number of filters in each Conv1D layer; kernel_size: kernel size; dropout_rate: dropout rate; dense_layers (count): number of dense layers; dense (units): number of units in a dense layer.

Supplementary Table 4. Site-level model training data composition for cCpG and vCpG study

Individual	vCpG (million)	cCpG (million)	vCpG : cCpG ratio		
			1:0.5	1:1	1:3
1	0.34	12.3	0.34 : 0.17	0.34 : 0.34	0.34 : 1.01
2	0.69	26.97	0.69 : 0.35	0.69 : 0.69	0.69 : 2.07
10	3.43	129.38	3.43 : 1.72	3.43 : 3.43	3.43: 10.30
30	10.44	366.37	10.44 : 5.22	10.44 : 10.44	10.44 : 31.33
50	17.46	594.94	17.46 : 8.73	17.46 : 17.46	17.46 : 52.39
110	38.41	1267.83	38.41 : 19.21	38.41 : 38.41	38.41 : 115.23

Note: Training CpGs were composed solely of vCpGs, solely of cCpGs, or mixtures of cCpGs and vCpGs at different ratios. Listed numbers are the CpG site counts in millions.

Supplementary Table 5. MeLoDe computational efficiency

	Data	Software	Throughput (reads per second per CPU core)
Read-level prediction *	HiFi read	MeLoDe	7.8 ± 2.0
	HiFi read	ccsmeth	0.5 ± 0.1
	HiFi read	pbjasmine	71.2 ± 15.5

	Data	Software	Time (seconds per 1M sites, given 10× depth)
Site-level prediction **	HiFi read	MeLoDe	100.0 ± 7.9
	HiFi read	ccsmeth	723.0 ± 41.8
	HiFi read	pb-CpG-tools	94.8 ± 8.4

Note: Hardware specifications of the high-performance computing (HPC) cluster systems used in this study were as follows: two CPU node types were available—(1) 2 × AMD EPYC 7532 (2.4 GHz, 32 cores) and (2) 2 × Intel Xeon 8358 (2.6 GHz, 32 cores), both with 512 GB RAM (3200 MHz).

* From unaligned HiFi reads with kinetic signal tags of IPD(*fi*, *ri*) and PW(*fp*, *rp*), stored in BAM format, to methylation-annotated HiFi reads in BAM format, adding *MM* tags to specify modified base positions, and *ML* tags to quantify the statistical confidence score for methylation calls.

** From aligned methylation-annotated HiFi reads in BAM format (with *MM*/*ML* tags), to genome-wide CpG site prediction. Read alignment, performed using the third-party software, is not included in the runtime calculation.

Supplementary Table 6. Genetic (meQTL) validation of CpG sites recovered from non-HiFi reads

Sites	*cis*-meQTL	mCpG	Missingness	Tested CpGs	CpGs with ≥1 meQTL (%)
H-sites	99,598	15,299	81.0±11.7%	12,609,305	0.12
C-sites + O-sites	1,901,686	132,846	33.8±19.1%	26,010,635	0.51

Note: *cis*-meQTL mapping was performed separately for the pooled C- and O-sites and for H-sites in the low-depth 1KCP cohort. Significant *cis*-meQTLs: nominal $P < 5\times10^{-8}$; mCpG: CpG sites with at least one *cis*-meQTL; missingness: inter-individual missing rata; tested CpGs: CpGs with inter-individual missingness less than 90%.

Supplementary Table 7. Summary of MeLoDe's modes

Mode	Reads included	CpG sites recovered	Completeness	Accuracy	Recommended use case
MeLoDe - H	- HiFi	- H-sites	+++	+++++	High-depth data; accuracy-critical applications
MeLoDe - HC	- HiFi - non-HiFi-C	- H-sites - C-sites	++++	++++	Low-depth data; balanced completeness and accuracy
MeLoDe -HCO	- HiFi - non-HiFi-C - non-HiFi-O	- H-sites - C-sites - O-sites	+++++	+++	Low-depth cohort/meQTL designs where completeness is limiting (default)

Note: non-HiFi-Closed (non-HiFi-C) denotes reads with both DNA strands sequenced, while non-HiFi-Open (non-HiFi-O) denotes reads with only one strand sequenced. Site categories are defined based on a coverage threshold (say, 4×) as follows: H-sites (HiFi coverage ≥4×), C-sites (HiFi <4× and HiFi + non-HiFi-C ≥4×), and O-sites (HiFi + non-HiFi-C <4× and total coverage ≥4×).

Supplementary References

1. Flusberg BA, Webster DR, Lee JH, Travers KJ, Olivares EC, Clark TA, et al. Direct detection of DNA methylation during single-molecule, real-time sequencing. Nature Methods. 2010; 7: 461-5.

2. Lister R, Pelizzola M, Dowen RH, Hawkins RD, Hon G, Tonti-Filippini J, et al. Human DNA methylomes at base resolution show widespread epigenomic differences. Nature. 2009; 462: 315-22.

3. Poplin R, Chang P-C, Alexander D, Schwartz S, Colthurst T, Ku A, et al. A universal SNP and small-indel variant caller using deep neural networks. Nature Biotechnology. 2018; 36: 983-7.
4. Marcel M, Murray P, Shilpa G, Sarah OF, Nadia P, Gunnar WK, et al. WhatsHap: fast and accurate read-based phasing. bioRxiv. 2016: 085050.
5. Abascal F, Acosta R, Addleman NJ, Adrian J, Afzal V, Ai R, et al. Expanded encyclopaedias of DNA elements in the human and mouse genomes. Nature. 2020; 583: 699-710.