Methylation reference datasets from Quartet DNA materials for benchmarking epigenome sequencing

Guo, X. *et al*.

Table of Contents

Abbreviations **2**

Supplementary Figures **3**

Supplementary Fig. 1. Overview of methylation sequencing data analysis workflow. 3

Supplementary Fig. 2. Cytosine detection counts across sequencing depth thresholds. 4

Supplementary Fig. 3. Strand concordance across sequencing depth thresholds. 5

Supplementary Fig. 4. Methylation frequency distribution across different batches. 6

Supplementary Fig. 5. PCA analysis of different protocols and pipelines. 7

Supplementary Fig. 6. Intra-sample reproducibility assessment at concordant CpG sites. 9

Supplementary Fig. 7. Inter-sample reproducibility assessment at concordant CpG sites. 10

Supplementary Fig. 8. Comparison of unique CpG sites covered and missed across different samples and batches. 11

Supplementary Fig. 9. Workflow for identifying annotation types of CpG sites in reference datasets. 13

Supplementary Fig. 10. Annotation distribution across different sample pairs based on reference datasets. 14

Supplementary Fig. 11. Schematic overview of 850K for validating reference datasets. 16

Supplementary Fig. 12. Performance assessment based on reference datasets across different regions. 17

Supplementary Fig. 13. Performance assessment based on reference datasets across different pipes. 18

Supplementary Tables **20**

Supplementary Table 1. Performance metrics overview. 20

Supplementary Table 2. Orthogonal validation of methylation reference datasets using 850K 21

Supplementary Table 3. Sensitivity analysis of the reference dataset robustness for sample D6 across 24 exclusion regimes 22

Abbreviations

5mC: DNA 5-methylcytosine

850K: Illumina Infinium Methylation EPIC (850K) array

bp: Base pair

CpG: Cytosine-phosphate-guanine

CHH: Cytosine-hydrogen-hydrogen (non-CpG methylation context)

CGI: CpG island

EMseq: Enzymatic Methyl-Seq

gDNA: Genomic DNA

IQR: Interquartile Range

MAD: Median Absolute Deviation

MAE: Mean Absolute Error

ONT: Oxford Nanopore Technologies

oxBS-Seq: Oxidative bisulfite sequencing

PacBio: Pacific Biosciences

PC: Principal Component

PCA: Principal Component Analysis

PCC: Pearson Correlation Coefficient

Q30: Quality score threshold (Phred score ≥30)

RMSE: Root Mean Square Error

SNP: Single-Nucleotide Polymorphism

SNR: Signal-to-Noise Ratio

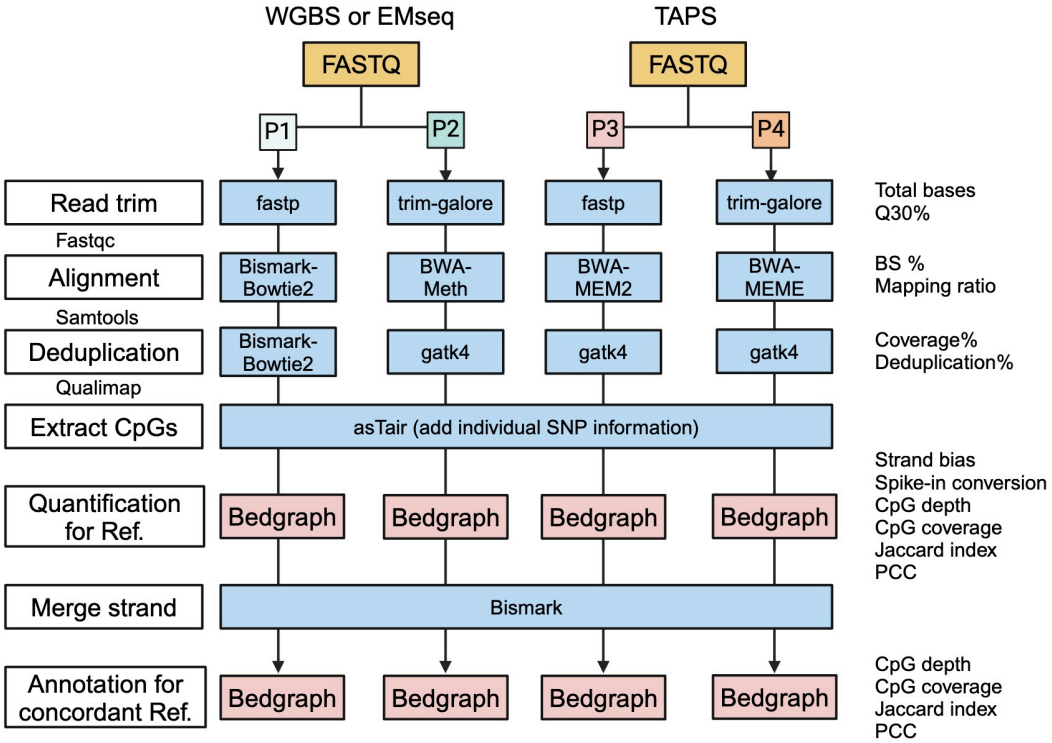
TAB-seq: Ten-eleven translocation-assisted bisulfite sequencing

TAPS: Ten-eleven translocation-assisted pyridine borane sequencing

WGBS: Whole-Genome Bisulfite Sequencing

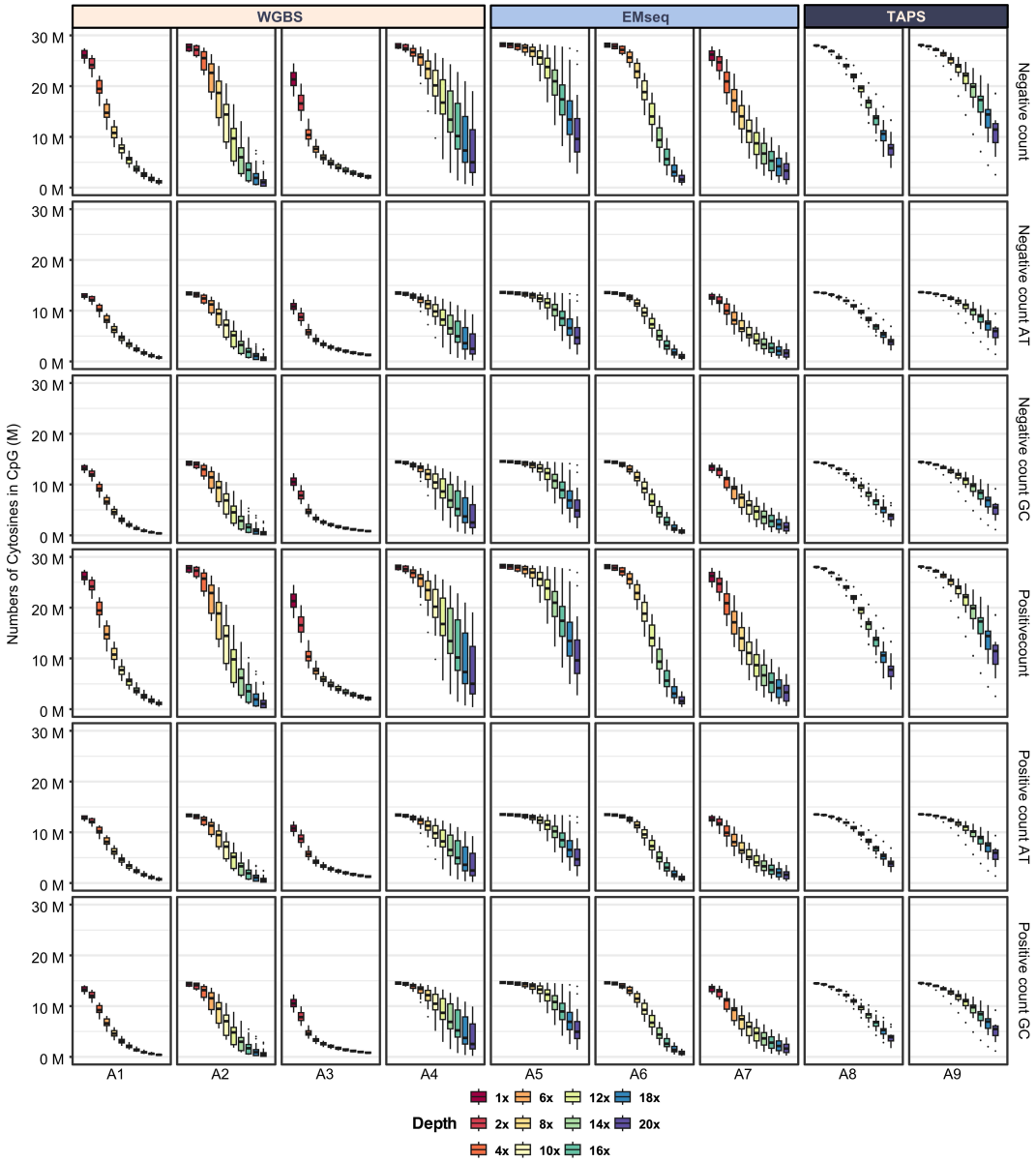
Supplementary Figures

Supplementary Fig. 1. Overview of methylation sequencing data analysis workflow.



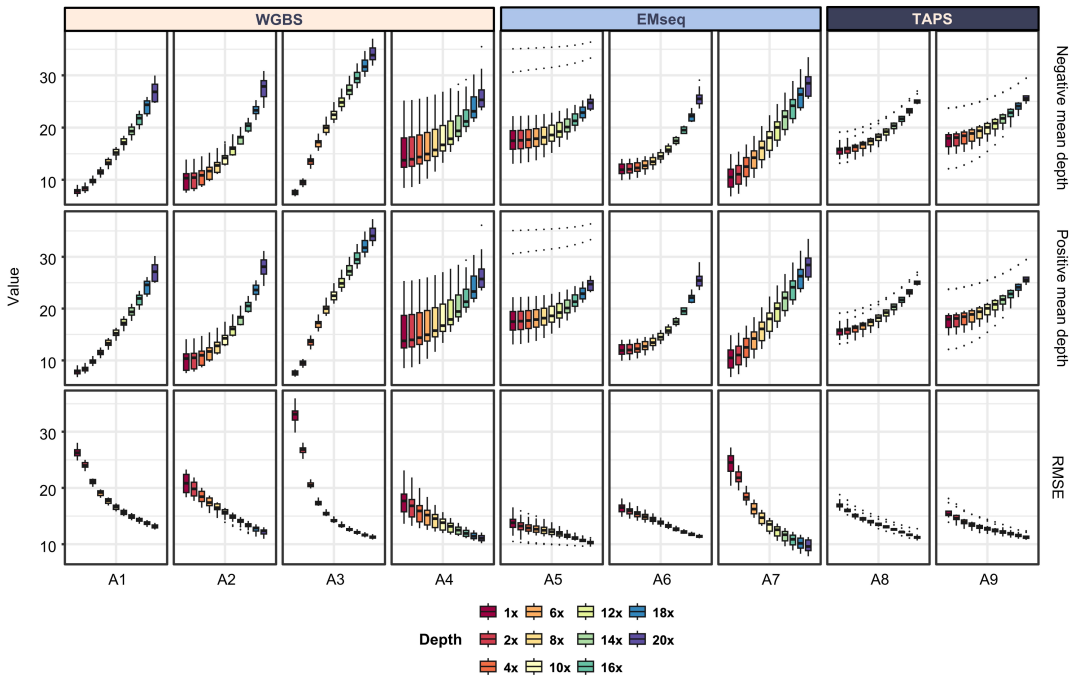
Supplementary Fig. 1. Overview of methylation sequencing data analysis workflow. This figure illustrates a detailed flowchart outlining the steps in processing FASTQ files for WGBS or EMseq (P1, P2) and TAPS (P3, P4) datasets. The flowchart highlights the key steps, including read trimming, alignment, deduplication, CpG extraction, quantification, strand merging, and annotation of reference datasets. The associated evaluation metrics for each step are also shown on the right side. Created in BioRender. Xi, H. (2025) https://BioRender.com/fbst4bu

Supplementary Fig. 2. Cytosine detection counts across sequencing depth thresholds.



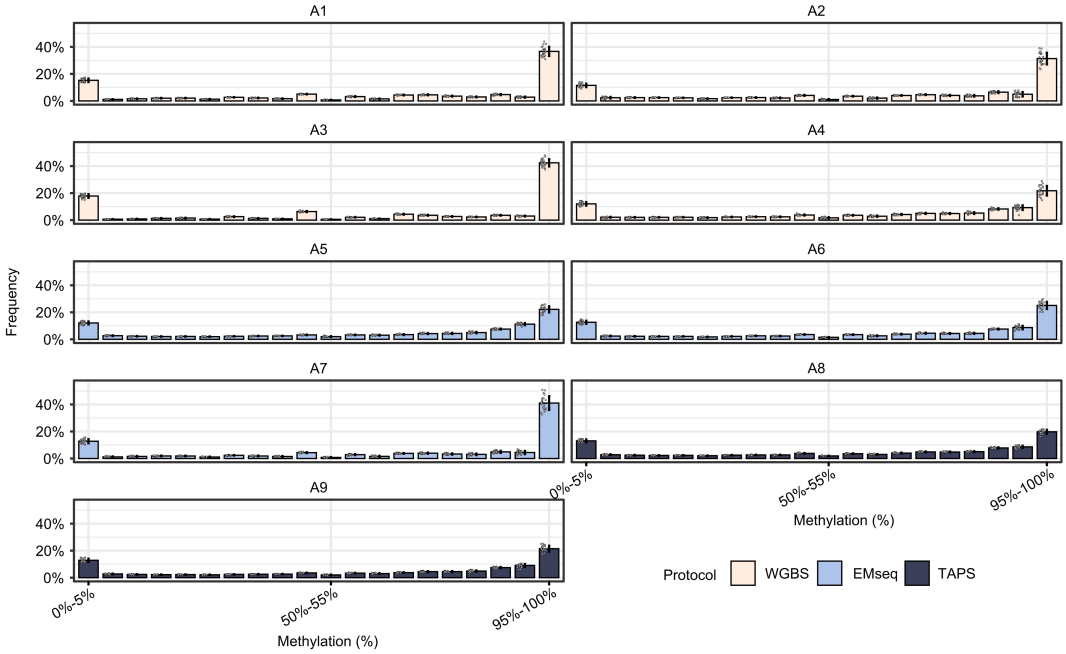
Supplementary Fig. 2. Cytosine detection counts across sequencing depth thresholds. This box plot illustrates the impact of cytosine sequencing depth on detection counts (millions) for positive and negative strands across nine batches (A1-A9), with box colors indicating the minimum cytosine sequencing depth threshold from 1× to 20×. Each row represents a strand-context combination: total cytosines (rows 1 and 4), CGA/CGT contexts (rows 2 and 5), and CGG/CGC contexts (rows 3 and 6). Source data are provided as a Source Data file.

Supplementary Fig. 3. Strand concordance across sequencing depth thresholds.



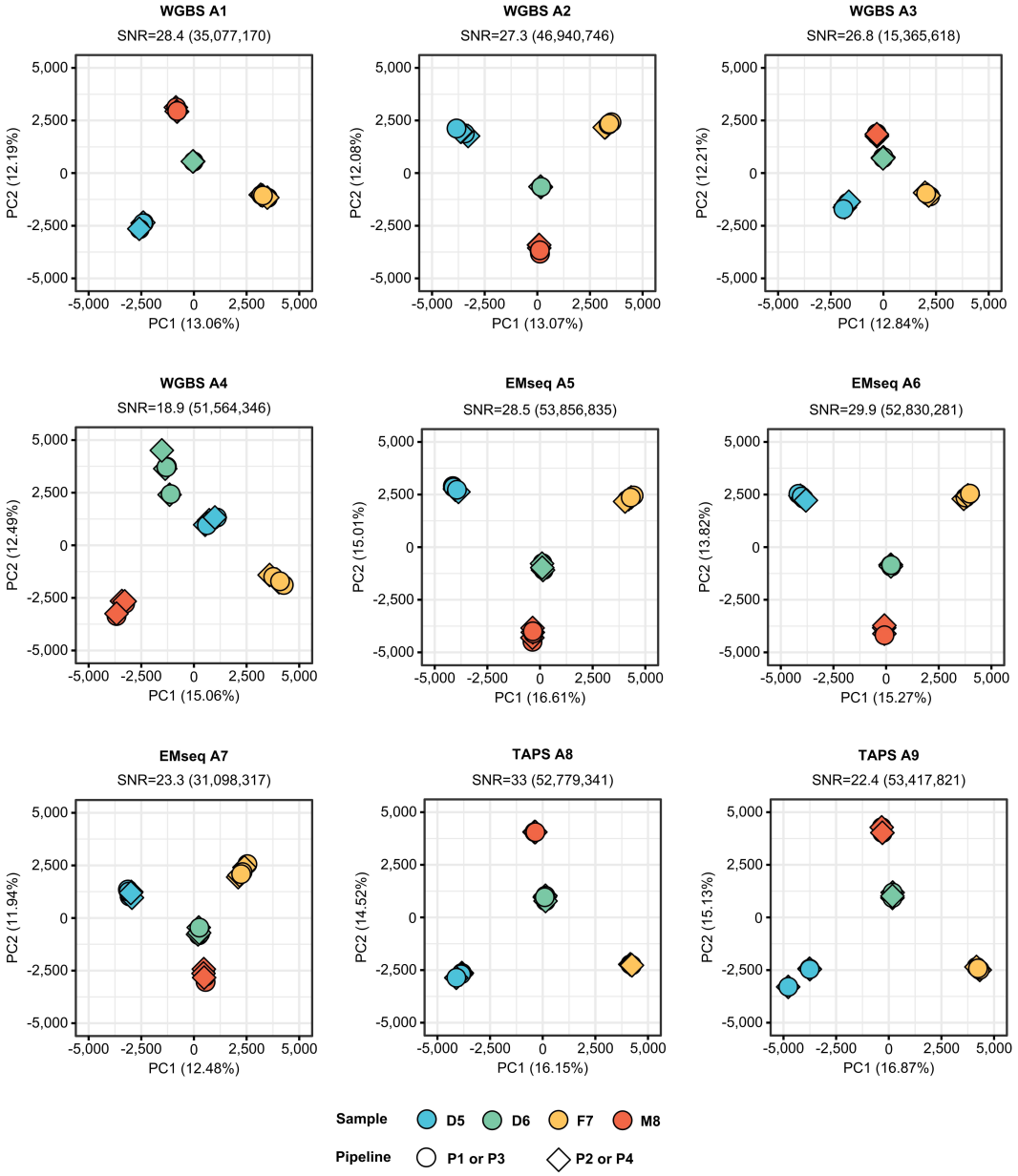
Supplementary Fig. 3. Strand concordance across sequencing depth thresholds. This box plot illustrates the impact of sequencing depth on cytosine concordance for positive and negative strands across nine batches (A1-A9), with box colors indicating the minimum sequencing depth threshold from 1× to 20×. The y-axis panels display (from top to bottom) the average cytosine sequencing depth on negative strands, average depth on positive strands, and RMSE of overlapping cytosines between strands. Source data are provided as a Source Data file.

Supplementary Fig. 4. Methylation frequency distribution across different batches.



Supplementary Fig. 4. Methylation frequency distribution across different batches. Bar plots represent the frequency of varying methylation levels, categorized into intervals ranging from 0% to 100%, with each interval spanning 5%. The y-axis shows the frequency of samples within each methylation interval, while the x-axis indicates the methylation level ranges. Each bar represents the mean frequency ± standard deviation (SD) for each interval, calculated from 24 measurements (4 sample types × 3 replicates × 2 pipelines). The plot is presented in nine grids, corresponding to one batch from A1 to A9, using WGBS, EMseq, or TAPS protocols, respectively. Source data are provided as a Source Data file.

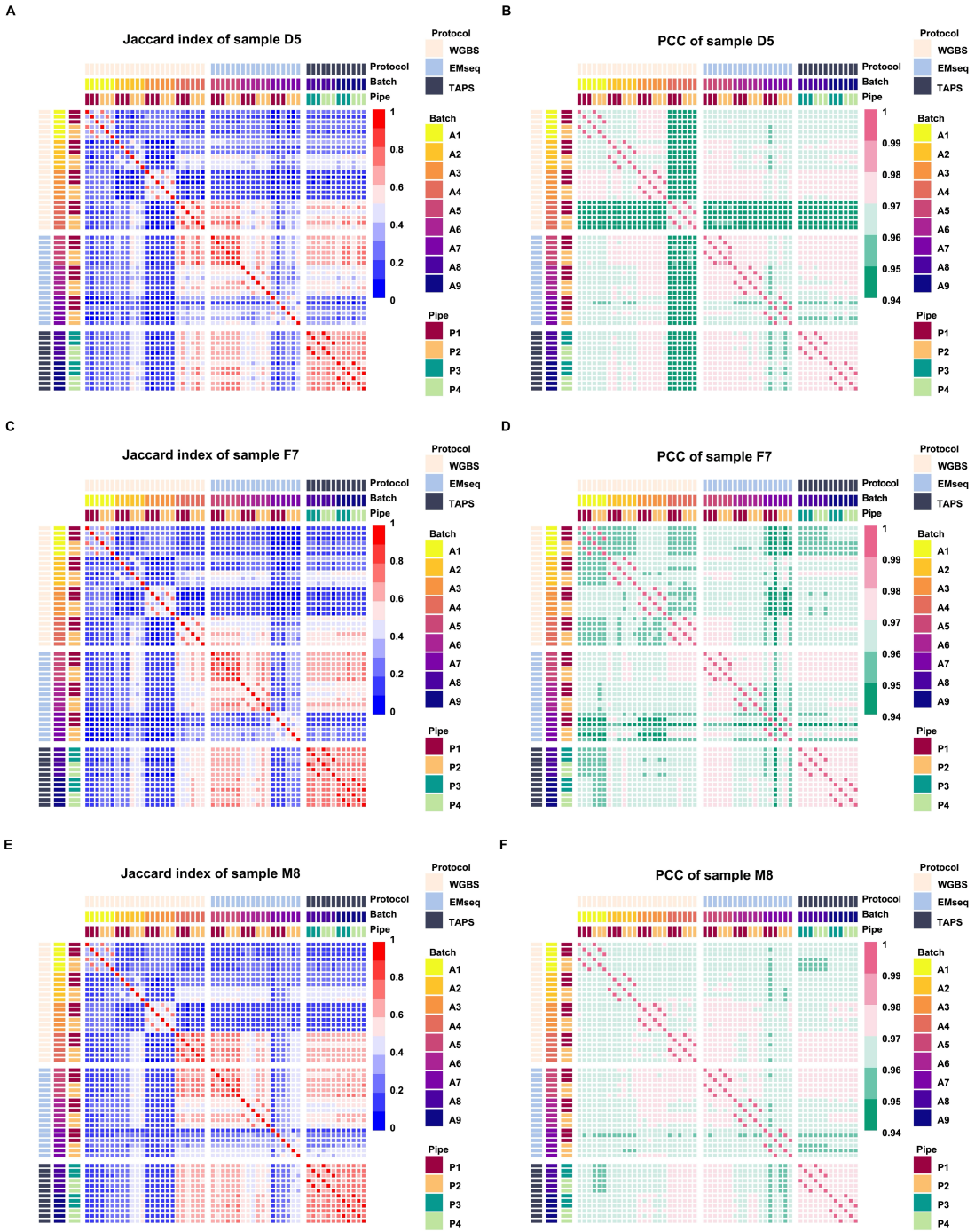
Supplementary Fig. 5. PCA analysis of different protocols and pipelines.



Supplementary Fig. 5. PCA analysis of different protocols and pipelines. Principal component analysis (PCA) was performed on data from four different samples (D5, D6, F7, and M8) and two analysis pipelines (P1/P3 and P2/P4). SNR values and the number of sites used for plotting are indicated in the panel titles. Samples are represented by different colors: cyan for D5, green for D6, orange for F7, and red for M8. Different pipeline choices are denoted by shape: circles for P1/P3 and diamonds for P2/P4. The axes represent the first two principal components (PC1 and PC2), with the percentage of variance explained by each principal component noted on the axis labels. The plot is

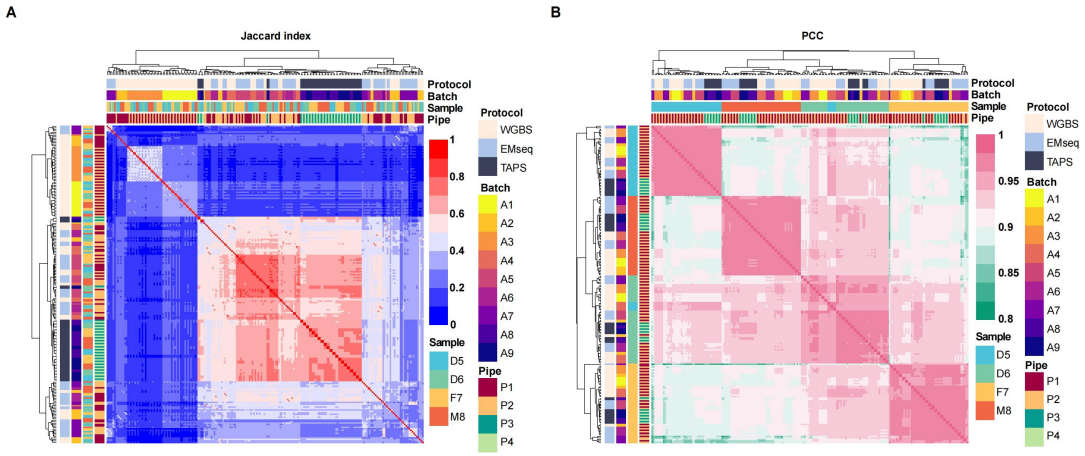
112 presented in a 3 × 3 grid of panels, corresponding to one batch from A1 to A9, using
113 WGBS, EMseq, or TAPS protocols, respectively. Source data are provided as a Source
114 Data file.
115

Supplementary Fig. 6. Intra-sample reproducibility assessment at concordant CpG sites.



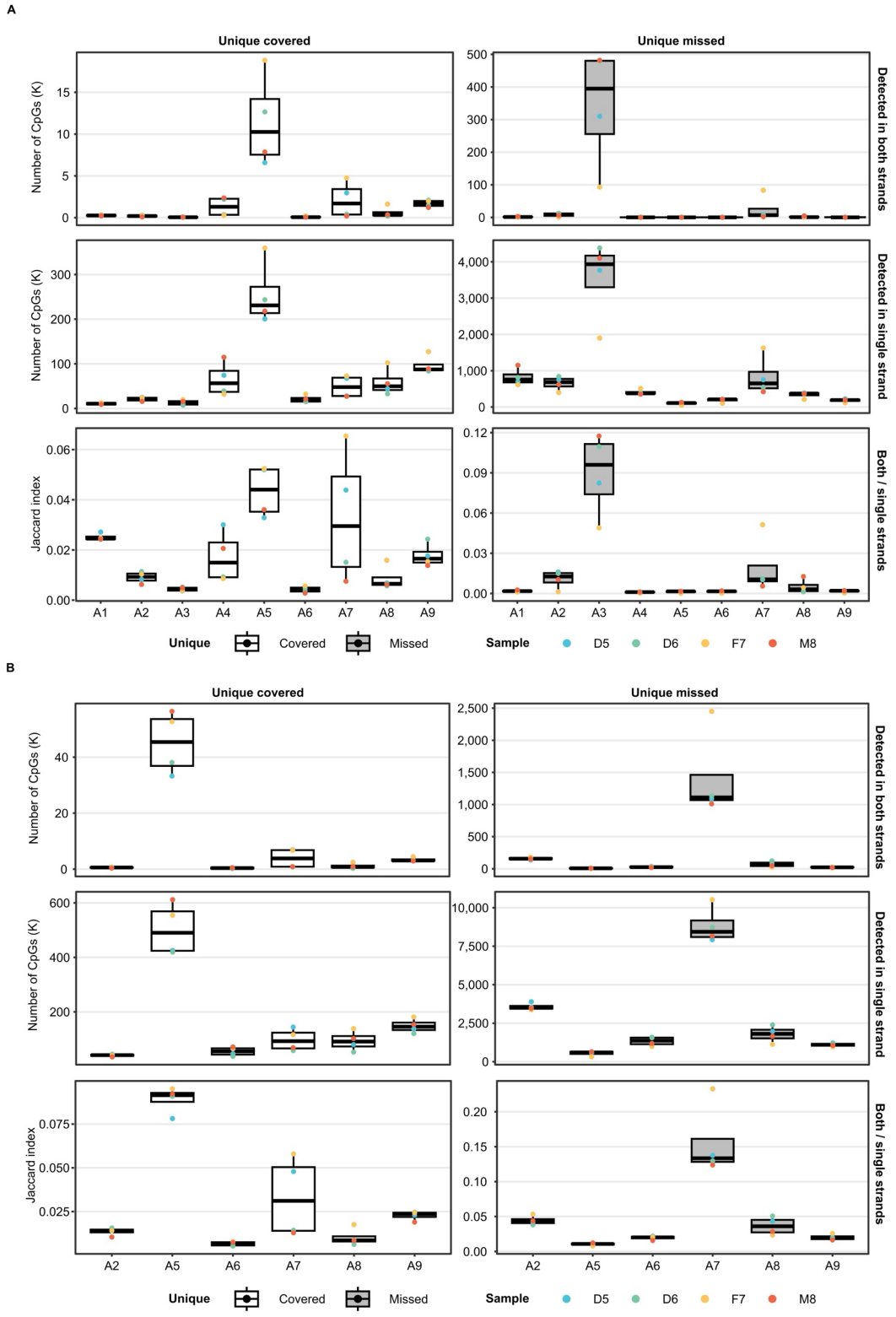
Supplementary Fig. 6. Intra-sample reproducibility assessment at concordant CpG sites. Jaccard index of ≥ 20× concordant CpG sites for samples D5 (**A**), F7 (**C**), and M8 (**E**). Pearson coefficient correlation (PCC) of ≥ 20× concordant CpG sites for samples D5 (**B**), F7 (**D**), and M8 (**F**). Source data are provided as a Source Data file.

Supplementary Fig. 7. Inter-sample reproducibility assessment at concordant CpG sites.



Supplementary Fig. 7. Inter-sample reproducibility assessment at concordant CpG sites. **A** Jaccard index of ≥ 20× concordant CpG sites for Quartet samples. **B** Pearson coefficient correlation (PCC) of ≥ 20× concordant CpG sites for Quartet samples. Source data are provided as a Source Data file.

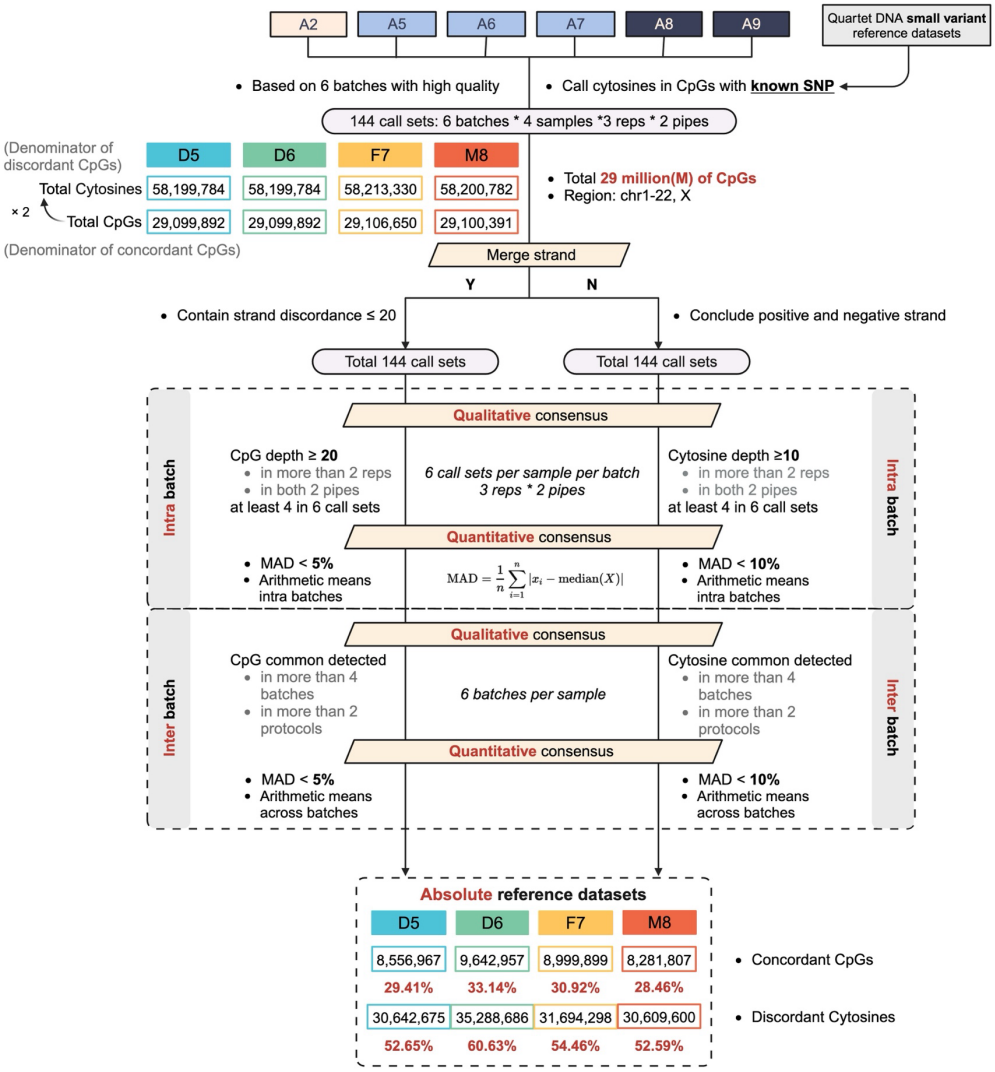
Supplementary Fig. 8. Comparison of unique CpG sites covered and missed across different samples and batches.



Supplementary Fig. 8. Comparison of unique CpG sites covered and missed across different samples and batches. **A** The box plots display the number of unique CpG sites

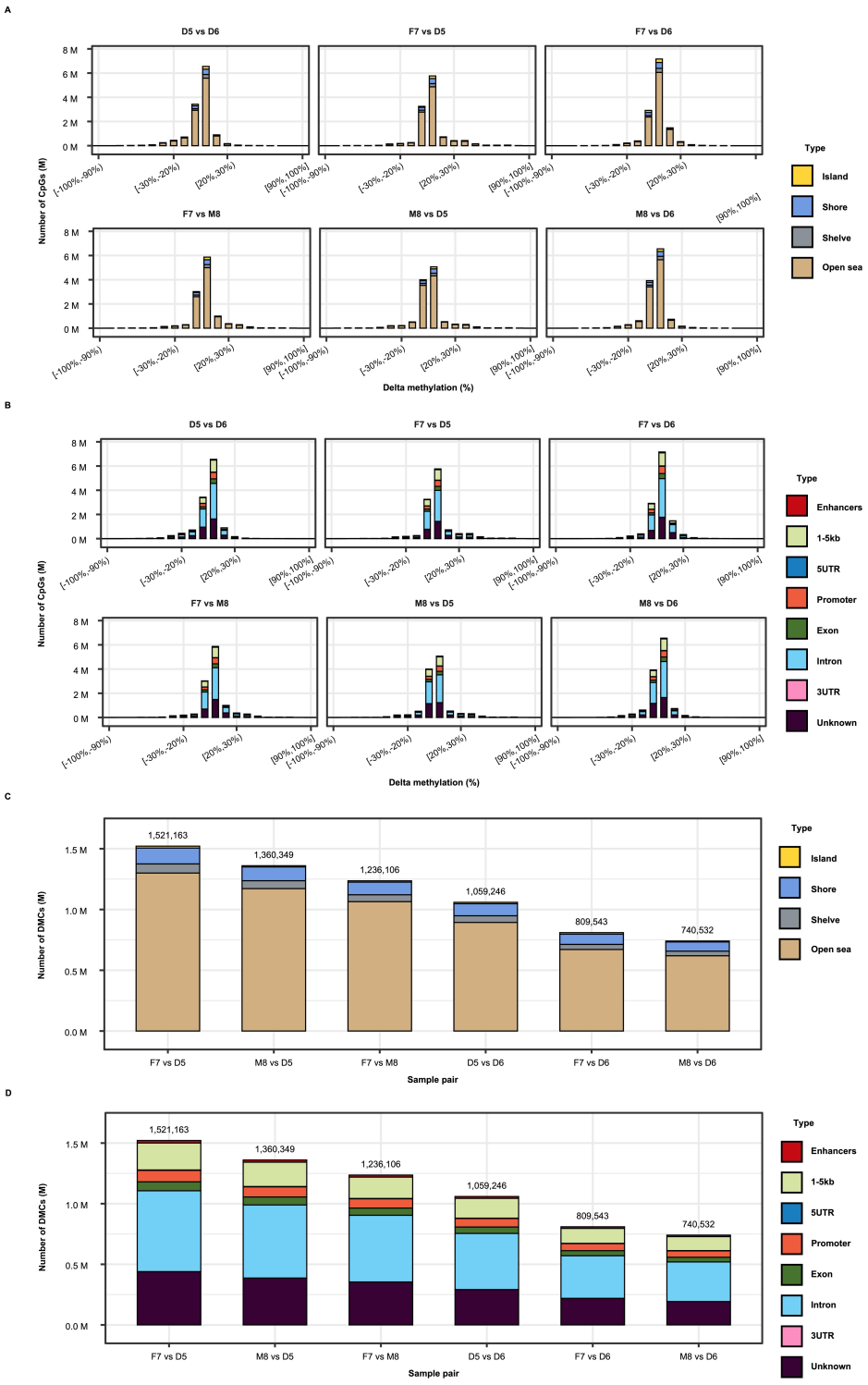
covered and missed in each of the nine batches (A1-A9) across four samples (D5, D6, F7, M8). The left three panels show unique CpG sites covered, while the right three panels show those missed. Plot are divided into three categories: (Top) sites detected in both strands, (Middle) sites detected in a single strand, and (Bottom) the combined detection across both and single strands. Points represent individual data points for each sample (D5 in cyan, D6 in green, F7 in yellow, M8 in red). The box color indicates the detection (covered) or missing status (missed) of the CpG sites, with the y-axis representing the number of sites (in thousands) or the Jaccard index. **B** The comparison of unique CpG sites covered and missed across four samples (D5, D6, F7, M8) for the 6 batches (A2, A5, A6, A7, A8, A9) used to construct the reference datasets. The plots are categorized in the same manner as in panel A. Source data are provided as a Source Data file.

Supplementary Fig. 9. Workflow for identifying annotation types of CpG sites in reference datasets.



Supplementary Fig. 9. Workflow for identifying annotation types of CpG sites in reference datasets. Annotation for reference datasets was conducted according to the following steps: (1) merging CpG sites with a difference of less than 20% on the positive and negative strands; (2) identifying intra-batch qualification consistency CpGs; (3) identifying inter-batch quantification consistency CpGs; (4) calculating methylation levels based on reliable CpGs that were expressed consensus. Created in BioRender. Xi, H. (2025) https://BioRender.com/fbst4bu

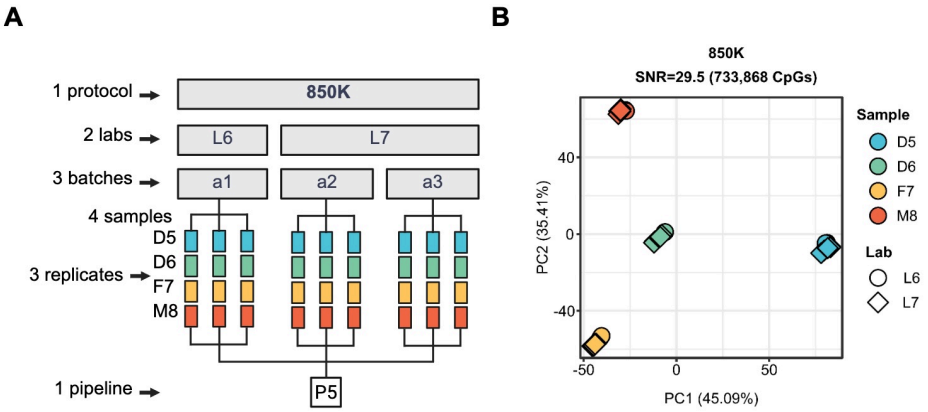
Supplementary Fig. 10. Annotation distribution across different sample pairs based on reference datasets.



Supplementary Fig. 10. Annotation distribution across different sample pairs based on reference datasets. **A**, **B** Distribution of CpG sites across six sample pairs, stratified by delta CpG methylation (%). The x-axis shows the delta methylation levels between each

sample pair; the y-axis shows the number of CpG sites at each delta methylation levels. **A** colored CpGs by CpG island (CGI) context; **B** colored CpGs by genic features. **C, D** Total number of differentially methylated CpG sites (DMCs) identified in each sample pair. The DMC threshold is defined as an absolute methylation difference ≥ 20%. The x-axis indicates the six sample pairs; the y-axis indicates the number of DMCs. **C** colored DMCs by CGI context; **D** colored DMCs by genic features. Source data are provided as a Source Data file.

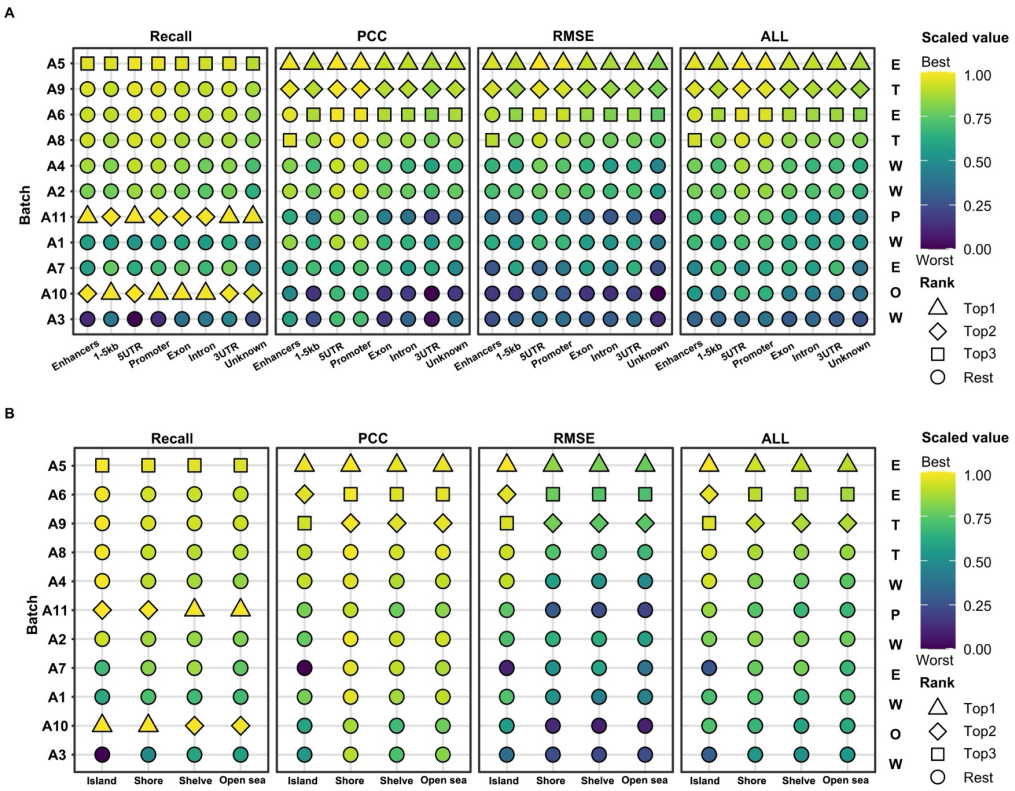
Supplementary Fig. 11. Schematic overview of 850K for validating reference datasets.



Supplementary Fig. 11. Schematic overview of 850K for validating reference datasets. **A** Scheme for generating the 850K dataset. Three technical replicates of each Quartet DNA reference material were processed in three batches using 850K arrays across two labs. A unified analytical pipeline identified 733,868 high-confidence CpG sites common to all batches. Created in BioRender. Xi, H. (2025) https://BioRender.com/fbst4bu

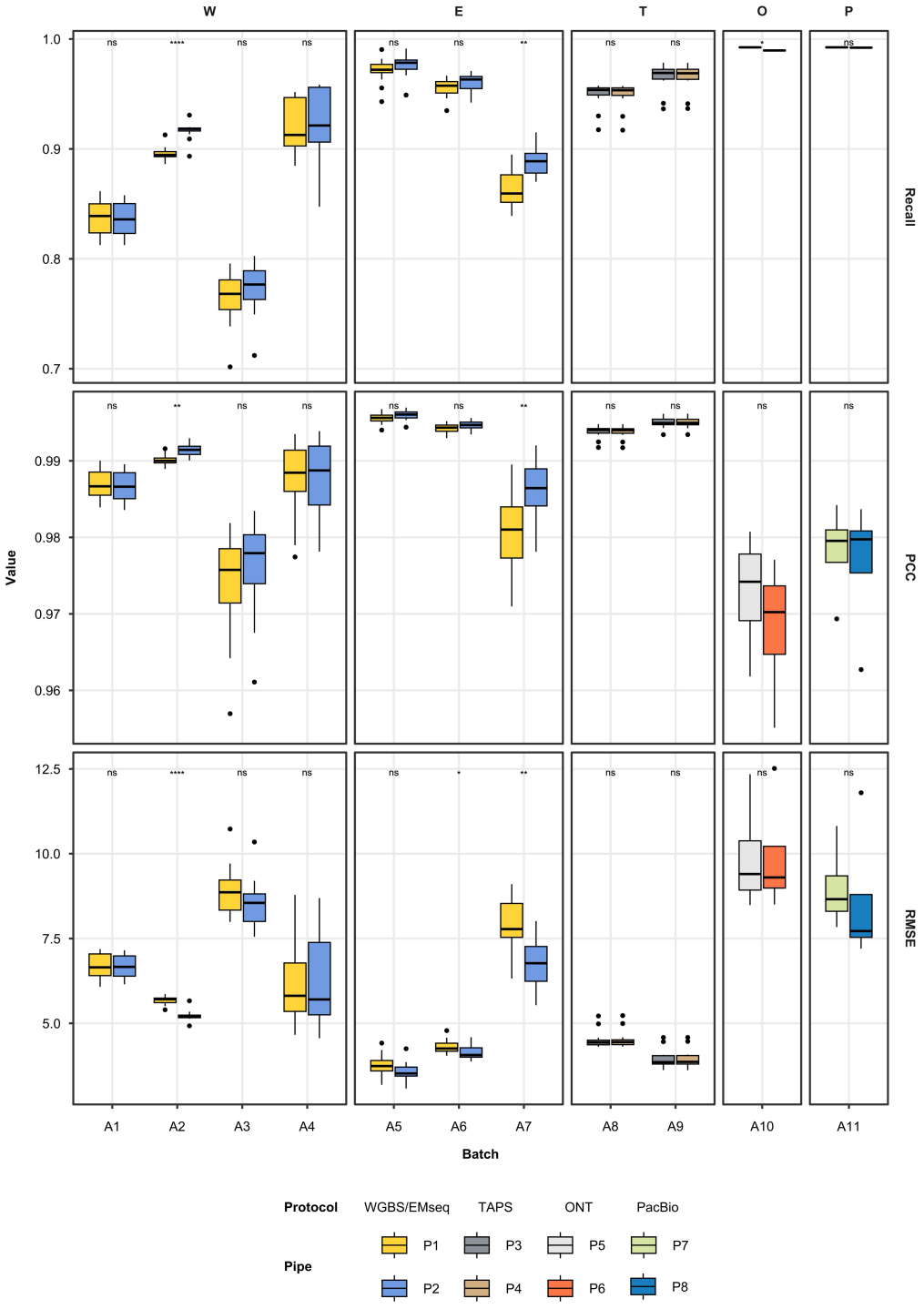
B PCA and SNR of 850K array data. 850K analyses incorporated 733,868 CpGs, inherently treated as concordant due to single-strand profiling. Source data are provided as a Source Data file.

Supplementary Fig. 12. Performance assessment based on reference datasets across different regions.



Supplementary Fig. 12. Performance assessment based on reference datasets across different regions. Evaluation scores under **A**, genic features, and **B**, CGI context, across three reference-dependent metrics, with the top three performers highlighted for each metric. ALL scores represent the arithmetic mean of scaled (0–1) Recall and PCC values, and inversely scaled (0–1) RMSE. W- WGBS, E- EMseq, T- TAPS, O- ONT, P- PacBio. All metrics were assessed at single-CpG resolution, with short-read analyses restricted to strand-concordant sites (inter-strand methylation difference ≤ 20%), while long-read data inherently merged both strands into duplex CpG calls due to the inability of PacBio to resolve strand-specific methylation. Source data are provided as a Source Data file.

Supplementary Fig. 13. Performance assessment based on reference datasets across different pipes.



Supplementary Fig. 13. Performance assessment based on reference datasets across different regions. Pipe performance evaluation across 11 batches using reference-dependent metrics: Recall, Pearson Correlation Coefficient (PCC) and Root Mean Square Error (RMSE). W- WGBS, E- EMseq, T- TAPS, O- ONT, P- PacBio. All

metrics were assessed at single-CpG resolution, with short-read analyses restricted to strand-concordant sites (inter-strand methylation difference ≤ 20%), while long-read data inherently merged both strands into duplex CpG calls due to the inability of PacBio to resolve strand-specific methylation. Source data are provided as a Source Data file.

Supplementary Tables

Supplementary Table 1. Performance metrics overview.

Metrics	WGBS A1	WGBS A2	WGBS A3	WGBS A4	EMseq A5	EMseq A6	EMseq A7	TAPS A8	TAPS A9
Total bases (G)	96.75±6.96	103.68±6.36	107.31±9.58	117.54±33.84	155.87±50.44	110.64±7.6	111.08±10.75	108.47±9.82	127.65±16.15
Total reads (M)	645.02±46.42	691.6±43.22	713.47±63.74	783.6±225.59	1037.75±336.26	735.58±50.51	740.54±71.68	753.24±68.18	886.47±112.17
Q30%	88.77±0.81	89.88±1.13	91.77±0.2	88.98±0.73	90.05±0.57	91.84±0.34	94.09±0.57	94.27±0.54	94.38±0.42
Duplication%	11.38±1.31	12.75±2.67	20.41±1.80	1.95±0.50	8.78±1.90	12.61±1.54	22.74±11.64	2.87±0.39	3.14±1.29
GC%	20.26±0.07	23.78±0.72	20.72±0.12	22.19±0.26	21.75±0.17	21.36±0.15	23.82±1.58	41.06±0.92	40.67±0.42
Insert size (bp)	238±8.28	247.58±14.17	268±0.00	263.92±0.29	268±0.00	268±0.00	233.75±12.53	161±0.00	168.92±27.42
Spike-in conversion%	93.91±13.51	97.61±0.05	99.78±0.04	99.88±0.05	NA	99.97±0.06	97.11±0.28	91.2±0.63	96.49±1.05
CHH methylation%	0.43±0.06	3.19±0.14	0.3±0.00	0.2±0.00	0.1±0.00	0.12±0.04	3.42±0.34	0.32±0.04	0.3±0.00
Coverage% 20×	0.19±0.03	0.38±0.08	0.13±0.01	0.64±0.13	0.81±0.05	0.58±0.05	0.32±0.07	0.66±0.04	0.71±0.07
CpG depth	15.54±1.03	20.6±2.45	13.03±0.79	31.55±8.87	37.63±9.61	24.3±1.39	21.57±3.33	32.24±3.29	35.95±5.64
Jaccard	0.44±0.01	0.34±0.06	0.47±0.01	0.62±0.10	0.78±0.03	0.58±0.03	0.41±0.04	0.74±0.02	0.77±0.03
PCC	0.97±0.00	0.97±0.00	0.98±0.00	0.98±0.00	0.98±0.00	0.97±0.00	0.97±0.00	0.98±0.00	0.98±0.00
SNR	28.4	27.3	26.8	18.9	28.5	29.9	23.3	33.0	22.4

Supplementary Table 1. Performance metrics overview. This table presents the performance metrics for each batch, with values reported as mean ± standard deviation (SD) for all indicators, except for the last row, which shows single values for each batch. Greener cells indicate better performance in the respective batch (for some metrics, higher values correspond to better performance, and for others, the reverse is true). Non-directional metrics (GC% and insert size) are filled with light grey, and the numerical value does not directly reflect the performance. Source data are provided as a Source Data file.

Supplementary Table 2. Orthogonal validation of methylation reference datasets using 850K.

Metrics	D5	D6	F7	M8
Total CpGs on 850K	733,868	733,868	733,868	733,868
Overlapping CpGs with reference	442,014	521,554	483,431	436,324
CpGs within reference ± 3SD	362,887	435,276	396,165	357,759
Detection rate (%)	60.23%	71.07%	65.87%	59.46%
Validation rate (%)	82.10%	83.46%	81.95%	81.99%
Pearson correlation (PCC)	0.99	0.99	0.99	0.99
Spearman correlation (SCC)	0.95	0.96	0.95	0.96
Coefficient of determination (R^2)	0.98	0.97	0.98	0.98
Mean absolute error (MAE, %)	5.14%	5.49%	5.01%	5.20%
Root mean square error (RMSE, %)	6.97%	7.34%	6.78%	6.99%

Supplementary Table 2. Orthogonal validation of methylation reference datasets using 850K. Total CpGs on 850K indicate the number of high-confidence probes retained after quality filtering (733,868 each sample). Overlapping CpGs with reference represent shared sites between array data and reference datasets, with the detection rate calculated as the percentage of array CpGs overlapping the reference. The validation rate reflects the proportion of overlapping CpGs where array methylation levels (0-100% scale) fell within ±3 standard deviations of reference values. PCC, SCC, R^2, MAE, and RSME metrics were calculated exclusively across all overlapping CpG sites shared between the reference datasets and 850K. Source data are provided as a Source Data file.

Supplementary Table 3. Sensitivity analysis of the reference dataset robustness for sample D6 across 24 exclusion regimes.

Removed	Length	Intersection	Union	Jaccard	PCC	SCC	MAE	RMSE
A2_all_D6	19063274	13131539	19142125	0.6860	0.9997	0.9977	0.5367%	0.8876%
A2_rep1_D6	15408679	12859403	15759666	0.8160	0.9997	0.9979	0.4662%	0.8196%
A2_rep2_D6	15013488	12847470	15376408	0.8355	0.9997	0.9979	0.4667%	0.8206%
A2_rep3_D6	14835646	12827386	15218650	0.8429	0.9997	0.9978	0.4786%	0.8319%
A5_all_D6	13977313	13011871	14175832	0.9179	0.9997	0.9976	0.5923%	0.9136%
A5_rep1_D6	13337929	12724818	13823501	0.9205	0.9999	0.9986	0.3240%	0.6006%
A5_rep2_D6	12863252	12483532	13590110	0.9186	0.9998	0.9985	0.3325%	0.6188%
A5_rep3_D6	13324267	12657475	13877182	0.9121	0.9998	0.9985	0.3494%	0.6286%
A6_all_D6	19271863	13157888	19324365	0.6809	0.9998	0.9984	0.3813%	0.7321%
A6_rep1_D6	14330913	13015602	14525701	0.8960	0.9998	0.9986	0.3625%	0.7290%
A6_rep2_D6	14712621	13025700	14897311	0.8744	0.9998	0.9985	0.3975%	0.7717%
A6_rep3_D6	14109027	13031575	14287842	0.9121	0.9998	0.9988	0.3255%	0.6858%
A7_all_D6	17681395	13084437	17807348	0.7348	0.9996	0.9974	0.6064%	0.9487%
A7_rep1_D6	15023962	12706856	15527496	0.8183	0.9998	0.9982	0.4018%	0.7141%
A7_rep2_D6	14711383	12674967	15246806	0.8313	0.9998	0.9981	0.4232%	0.7396%
A7_rep3_D6	14606029	12688993	15127426	0.8388	0.9998	0.9981	0.4239%	0.7386%
A8_all_D6	15408896	13068940	15550346	0.8404	0.9996	0.9966	0.6768%	1.0299%
A8_rep1_D6	13578673	12463116	14325947	0.8700	0.9998	0.9973	0.4108%	0.7225%
A8_rep2_D6	13523144	12429181	14304353	0.8689	0.9998	0.9972	0.4252%	0.7381%
A8_rep3_D6	13062915	12318863	13954442	0.8828	0.9998	0.9972	0.4270%	0.7462%
A9_all_D6	14585856	13035534	14760712	0.8831	0.9996	0.9972	0.6184%	0.9572%
A9_rep1_D6	13165823	12427265	13948948	0.8909	0.9998	0.9979	0.3708%	0.6597%
A9_rep2_D6	12860093	12343394	13727089	0.8992	0.9998	0.9980	0.3631%	0.6555%
A9_rep3_D6	13341186	12564669	13986907	0.8983	0.9998	0.9982	0.3558%	0.6418%

Supplementary Table 3. Sensitivity analysis of the reference dataset robustness for sample D6 across 24 exclusion regimes. The table shows the impact of excluding individual batches or technical replicates on reference dataset construction. Removed specifies excluded data subsets (e.g., rep1-3: removal of single libraries per batch; all: removal of entire batches). Length indicates retained CpG sites in perturbed references, Intersection and Union quantify overlaps with the original D6 reference dataset (22,078,295 CpG sites), and Jaccard measures their similarity. PCC and SCC assess methylation value correlations, while MAE and RMSE evaluate quantitative deviations. Source data are provided as a Source Data file.