Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Benchmarking and Quality Control of Quantitative Accuracy in Next-Generation Sequencing-based DNA Methylation Profiling

Yuanfeng Zhang
National Center for Clinical Laboratories, National Clinical Research Center for Gerontology, Chinese Academy of Medical Sciences

Yu Ma
National Center for Clinical Laboratories, National Clinical Research Center for Gerontology, Chinese Academy of Medical Sciences

Xiaorou Guo
Fudan University

Duo Wang
National Center for Clinical Laboratories, National Clinical Research Center for Gerontology, Chinese Academy of Medical Sciences

Yaqing Liu
Fudan University

Yanxi Han
National Center for Clinical Laboratories, National Clinical Research Center for Gerontology, Chinese Academy of Medical Sciences

Qingwang Chen
Fudan University

Lin Li
National Center for Clinical Laboratories, National Clinical Research Center for Gerontology, Chinese Academy of Medical Sciences

Yingshuo Ma
National Center for Clinical Laboratories, National Clinical Research Center for Gerontology, Chinese Academy of Medical Sciences

Shumeng Duan
Fudan University

Cong Liu
National Center for Clinical Laboratories, National Clinical Research Center for Gerontology, Chinese Academy of Medical Sciences

Hanjing Hou

National Center for Clinical Laboratories, National Clinical Research Center for Gerontology, Chinese Academy of Medical Sciences

Wanwan Hou
Fudan University

Zhuorui Ran
National Center for Clinical Laboratories, National Clinical Research Center for Gerontology, Chinese Academy of Medical Sciences

Ying Yu
Fudan University

Lianhua Dong
National Institute of Metrology

Leming Shi
Fudan University

Jinming Li
National Center for Clinical Laboratories, National Clinical Research Center for Gerontology, Chinese Academy of Medical Sciences

Yuanting Zheng
Fudan University

Rui Zhang
ruizhang@nccl.org.cn

National Center for Clinical Laboratories, National Clinical Research Center for Gerontology, Chinese Academy of Medical Sciences

---

Research Article

Keywords: DNA methylation, epigenetics, sequencing, quality control, standardization

Posted Date: November 11th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11537881/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License. Read Full License

Additional Declarations: No competing interests reported.

Benchmarking and Quality Control of Quantitative Accuracy in Next-Generation Sequencing-based DNA Methylation Profiling

Authors

Yuanfeng Zhang[1,2,3#], Yu Ma[1,2,3#], Xiaorou Guo[4#], Duo Wang[1,2,3], Yaqing Liu[4], Yanxi Han[1,2,3], Qingwang Chen[4], Lin Li[1,2,3], Yingshuo Ma[1,2,3], Shumeng Duan[4], Cong Liu[1,2,3], Hanjing Hou[1,2,3], Wanwan Hou[4], Zhuorui Ran[1,2,3], Ying Yu[4], Lianhua Dong[5], Leming Shi[4,6,7*], Jinming Li[1,2,3*], Yuanting Zheng[4,6,8*], Rui Zhang[1,2,3*]*

[1] National Center for Clinical Laboratories, Beijing Hospital, National Center for Gerontology; National Clinical Research Center for Gerontology; The Key Laboratory of Geriatrics of NHC; Institute of Geriatric Medicine, Chinese Academy of Medical Sciences, 100005, P.R. China

[2] National Center for Clinical Laboratories, Chinese Academy of Medical Sciences & Peking Union Medical College, Beijing 100005, P.R. China

[3] Beijing Engineering Research Center of Laboratory Medicine, Beijing Hospital, Beijing 100005, P.R. China

[4] State Key Laboratory of Genetics and Development of Complex Phenotypes, School of Life Sciences, Human Phenome Institute and Shanghai Cancer Center, Fudan University, Shanghai 200, P.R. China

[5] Center for Advanced Measurement Science, National Institute of Metrology, Beijing 100013, P.R. China

[6] The International Human Phenome Institutes, Shanghai 200438, P.R. China

[7] Shanghai Cancer Center, Fudan University, Shanghai 200438, P.R. China

[8] Department of Pathology, Zhongshan Hospital, Fudan University, Shanghai 200438, P.R. China

# These authors contributed equally: Yuanfeng Zhang, Yu Ma and Xiaorou Guo.

email of all authors:

Yuanfeng Zhang: yuanfengzhang1997@foxmail.com

Yu Ma: 2162657681@qq.com

Xiaorou Guo: xrguo24@m.fudan.edu.cn

Duo Wang: 18801232285@163.com

Yaqing Liu: yaqing.liu@outlook.com

Yanxi Han: hanyanxicici@163.com

Qingwang Chen: qwchen20@fudan.edu.cn

Lin Li: lilin5052@bjhmoh.cn
Yingshuo Ma: mys75153@163.com
Shumeng Duan: dsm@fudan.edu.cn
Cong Liu: liucong0302@163.com
Hanjing Hou: houhanjing0321@163.com
Wanwan Hou: houwanwan@fudan.edu.cn
Zhuorui Ran: 1098778233@qq.com
Ying Yu: ying_yu@fudan.edu.cn
Lianhua Dong: donglh@nim.ac.cn
Leming Shi: lemingshi@fudan.edu.cn
Jinming Li: jmli@nccl.org.cn
Yuanting Zheng: zhengyuanting@fudan.edu.cn
Rui Zhang: ruizhang@nccl.org.cn

Corresponding author’s phone number:
Rui Zhang: 86-10-58115053

Abstract

Biomedical applications of DNA methylation rely on accurate next-generation sequencing (NGS)-based DNA methylation profiling. However, routine quality metrics and limited reference coverage leave it unclear whether NGS assays provide the reliable measurements required for robust downstream inferences. Here we benchmarked 12 real-world NGS assays across four methods, alongside five microarray and nanopore assays as external benchmarks. First, we assessed quantitative performance across three dimensions, including absolute methylation accuracy, biological order preservation, and the magnitude of relative methylation differences. Second, we examined the robustness of methods across varying GC content, methylation level, mappability, and characteristics of tumor genome. As a result, methods showed distinct profiles across quantitative dimensions, and conventional quality metrics standard metrics cannot capture quantitative errors. Specifically, non-targeted NGS assays were dominated by sequencing depth-dependent binomial sampling noise. To address this, we developed MethRidgeQC, a reference-free quality control tool that can classify CpG measurements into four reliability classes. Retaining high-reliability classes reduced RMSE by 51–74% compared with low-reliability measurements, outperforming depth-based filtering by up to 47% while preserving 54–79% of CpGs for downstream analyses. Our study provides guidance on method selection, and actionable, reference-free quality control for various non-targeted NGS-based methods.

Keywords

DNA methylation, epigenetics, sequencing, quality control, standardization

Introduction

DNA methylation is a central quantitative layer in epigenomic regulation [1], cellular state [2], development [3], environmental response [4], aging [5], and cancer [6]. Biological interpretation and translational use of methylation data increasingly depend on high-resolution, genome-wide profiling, because regional signatures, differential-methylation calls, biomarker scores, and cross-platform analyses are built from methylation estimates at individual cytosines [7–9]. Next-generation sequencing (NGS)-based methods that provide single-base readouts across large parts of the genome have therefore become a foundation for DNA methylation datasets and cohorts [10, 11]. The numerical accuracy and reproducibility of reported methylation levels increasingly determine whether datasets can support cross-study biological insights [12, 13], as these applications place different demands on quantitative accuracy. Comparison of absolute methylation levels requires accurate numerical estimates, cell clustering requires preservation of biological ordering, and differential methylation analysis requires faithful estimates of the effect size of the difference. Benchmarking therefore needs to define the contexts in which an NGS assay is quantitatively reliable for downstream analyses. Therefore, the quantitative accuracy of NGS assays should be evaluated across the scenarios and genomic contexts in which the assays are used.

However, prior benchmarking studies mainly used data quality and concordance metrics, including mapping efficiency, coverage, conversion efficiency, cross-method concordance, and quantitative reproducibility [14–21], as proxies for direct accuracy, assuming without direct proof that high quality or reproducibility implies low

quantitative error. These metrics characterize technical properties of NGS assays, but do not establish whether an assay recovers methylation signals with enough fidelity. The BLUEPRINT consortium addressed this issue using a truth set containing 16,435 selected loci, which has been used to assess quantitative deviation across various methods and computational workflows [21, 22]. Yet, these selected sparse loci do not capture the full distribution of sequence composition, methylation states, mappability, and tumor-genome structures encountered by genome-wide methylation sequencing.

Genome-wide benchmarking matters since methylation measurements are shaped by interactions between assay procedures and genomic context. Library preparation and conversion protocols can alter fragment recovery, duplication, CpG recovery, and coverage uniformity. For instance, bisulfite-associated DNA degradation introduces sequence-context bias, whereas bisulfite-free methods show reduced coverage bias in GC-rich regions [14–16, 18]. In parallel, alignment and downstream bioinformatic processing influence the cytosine recovery and methylation estimates, particularly in repetitive, low-mappability, and other complex contexts [19–21]. For tumor-normal comparisons, the extent to which copy number variations (CNV) and structural complexity introduce additional quantitative difficulties also requires direct assessment. To this end, a genome-wide benchmark needs to locate these context-dependent errors and assess the robustness of NGS methylation assays.

Here we benchmark 12 real-world NGS assays across four methods in terms of absolute methylation accuracy, biological order preservation, and magnitude of relative methylation differences, using genome-wide single-base references, a titrated mixture

series, and five microarray and nanopore assays as external benchmarks. We further profile quantitative error across GC content, methylation level, mappability, and tumor-genome characteristics to identify dominant sources. Observing that non-targeted assays share a common, depth-dependent error structure, we develop a reference-free approach for annotate the reliability of CpG measurements and filter them based on assigned classes. Together, these analyses provide guidance for method selection and accuracy-aware quality control of NGS methylation data.

Methods

Assay cohorts

This study employed a real-world multi-center strategy. A primary NGS cohort was constructed to capture the current diversity of NGS methylation sequencing in routine practice, including conversion chemistries, real-world experimental protocols, and heterogeneous versus harmonized multi-center implementations. This cohort included 12 NGS-based methylation sequencing assays from 12 sequencing centers using their routine protocols, grouped by base-conversion principle as four whole-genome bisulfite sequencing (WGBS) assays, four enzymatic methyl sequencing (EM-seq) assays, one reduced representation bisulfite sequencing (RRBS) assay, and three ten-eleven translocation (TET)-assisted pyridine borane sequencing (TAPS) assays. These abbreviations denote conversion chemistry groups, not commercial kits or vendor-specific implementations. WGBS, RRBS, and EM-seq employ a reverse conversion strategy in which unmethylated cytosines are converted, while TAPS uses a direct

conversion strategy, in which methylated cytosines are converted. In addition, three Illumina EPICv2 microarray assays, one Oxford Nanopore-based reduced representation methylation sequencing (RRMS) assay, and one approximately 30× Oxford Nanopore whole-genome sequencing (NP-WGS) dataset were used as external benchmarks to contextualize NGS performance limits, rather than for direct head-to-head ranking.

Reference Materials

The sample panel included 11 genomic DNA (gDNA) samples: HF, D5, D6, F7, M8, T1, T2, T3, T4, BC, and BL. The HF sample was the 1:1 mixture of the completely unmethylated and fully methylated DNA standards (Zymo Research, D5301) derived from the HCT116 cell line (ATCC CCL-247). D5, D6, F7, and M8 were Quartet gDNA reference materials produced from Epstein-Barr virus-immortalized B-lymphoblastoid cell lines by Fudan University and the National Institute of Metrology, China, with certified reference material codes GBW09900 to GBW09903 [23]. T1–T4 were prepared by mixing D5 and D6 gDNA at 4:1, 3:2, 2:3, and 1:4, respectively. BC and BL were prepared from HCC1395 (ATCC CRL-2324) and HCC1395BL (ATCC CRL-2325). Base conversion assays received 0.2% lambda DNA and 0.01% pUC19 DNA spike-ins for all samples except HF. The reference truth was annotated with Genome-in-a-Bottle (GIAB) stratification resource and BisMap/UMap mappability.

Whole genome bisulfite sequencing

WGBS was performed across four centers (assays BS1–BS4) using varied methodologies, with substantial heterogeneity in fragmentation methods, bisulfite-

conversion conditions, library-construction timing, and sequencing platforms.

Fragmentation was achieved via ultrasonication in BS1–BS3, whereas BS4 utilized enzymatic tagmentation integrated into the library preparation kit. Regarding bisulfite conversion, BS1 and BS2 employed high-temperature bisulfite conversion (≥ 150 min at ≥ 64°C) using Zymo EZ DNA Methylation-Gold Kit (CAT: D5006 and D5005, respectively), while BS3 and BS4 utilized rapid low-temperature conversion (60 min at 54°C) with Zymo kits (Zymo Research Inc; CAT: D5031 and D5465). Regarding library construction, BS1 adopted a pre-conversion strategy using the MGIEasy Whole Genome Methylation Sequencing Library Prep Kit (CAT: 940-001527-00), whereas BS2–BS4 employed post-conversion strategies using a self-developed kit, the IDT Accel-NGS Methyl-seq DNA Library Kit (CAT: 1009824), and the Zymo-Seq WGBS Library Kit (CAT: D5465), respectively. The number of PCR amplification cycles varied by center, with BS1, BS2, BS3, and BS4 employing 13, 10, 4, and 6 cycles, respectively. Sequencing was performed on different platforms. BS1 used MGI DNBSEQ-G400RS (PE150), while BS2 and BS3 used the Illumina NovaSeq 6000 (PE150), and BS4 used the Illumina NovaSeq X (PE150).

Enzymatic methylation sequencing

Four EM-seq assays from four sequencing centers were labeled EM1–EM4. Assays EM1 and EM4 employed the Vazyme EpiArt DNA Enzymatic Methylation Kit (CAT: EM301-01, 330 min at 37°C) for enzymatic conversion but differed in downstream processing. EM1 utilized the Vazyme EpiArt DNA Methylation Library Kit for Illumina V3 with a post-conversion strategy, followed by 10 PCR cycles and sequencing on an

Illumina NovaSeq 6000. In contrast, EM4 adopted a similar post-conversion strategy with the Vazyme kit but reduced PCR amplification to 7 cycles and chose the MGI DNBSEQ-T7 platform.

Assays EM2 and EM3 utilized the NEBNext Enzymatic Methyl-seq Conversion Module (CAT: E7125, 37°C) coupled with the MGI 940-001527-00 kit on the DNBSEQ-T7 platform. These two assays primarily diverged in their library construction timing and reaction duration: EM2 adopted a pre-conversion strategy with a 280-min incubation, whereas EM3 employed a post-conversion strategy with a shorter 180-min incubation. Both EM2 and EM3 libraries were amplified for 10 PCR cycles. Across all EM-seq assays, sequencing was performed to generate 150 bp paired-end reads.

Reduced representation bisulfite sequencing

For the only RRBS assay (RR1), gDNA was fragmented using enzymatic digestion. Libraries were constructed using a self-developed kit employing a pre-conversion strategy, followed by bisulfite conversion with the QIAGEN EpiTect Fast DNA Bisulfite Kit (CAT: 59802, 60 min at 60°C). Following magnetic bead purification, libraries were amplified for 14 PCR cycles and sequenced on the MGI DNBSEQ-T7 platform to generate 100 bp paired-end reads.

TET-assisted pyridine borane sequencing

TAPS was performed across three sequencing centers (PS1–PS3) using a standardized workflow. The gDNA was fragmented by ultrasonication. Enzymatic oxidation and subsequent chemical reduction were carried out using the GenePlus TET (mTET2) kit

at 37°C for 80 min. Post-conversion libraries were constructed using the Hieff NGS Ultima Pro DNA Library Prep Kit (CAT: 12197ES), followed by magnetic bead purification and amplification with 9 PCR cycles. All libraries were sequenced on the MGI DNBSEQ-T7 platform to generate 150 bp paired-end reads.

Illumina Infinium MethylationEPIC v2.0

Three centers (labeled MA1–MA3) used the Infinium MethylationEPIC v2.0 BeadChip (935K microarray), adhering uniformly to the manufacturer's official assay (Illumina, Inc.). In brief, genomic DNA was quantified via fluorescence and subjected to bisulfite conversion using the Zymo EZ-96 DNA Methylation-Lightning MagPrep Kit (CAT: D5001, 12 hr at 50°C). Converted DNA was whole-genome amplified, fragmented, and hybridized to the EPIC v2.0 BeadChips using the Infinium MethylationEPIC v2.0 Kit. The libraries were analyzed on the Illumina NextSeq 550/550Dx system using the Infinium HD Assay workflow. Data were generated with consistent scanning parameters across all sites to support comparability of methylation beta-values among the three cohorts.

Reduced representation methylation sequencing

RRMS (RM1) was performed using Oxford Nanopore Technologies to enable native DNA methylation detection without chemical or enzymatic conversion. The gDNA was quantified via fluorescence and fragmented enzymatically. Libraries were prepared using the Oxford Nanopore Native Barcoding Kit 96V14 (CAT: SQK-NBD114.96) following a PCR-free workflow. Sequencing was conducted on the Oxford Nanopore PromethION P2 platform. Adaptive sampling was performed with the default

configuration (https://community.nanoporetech.com/adaptive_sampling_catalogue/beds/7). Methylation calls were derived directly from raw signal deviations compared to unmodified controls using standard nanopore base calling pipelines.

Oxford Nanopore whole genome sequencing

To extend the reference beyond NGS-accessible regions, we used 10 μg DNA from each of the 11 samples (HF, D5, D6, F7, M8, T1–T4, BC, BL) for ~30× whole-genome sequencing on Oxford Nanopore PromethION (Novogene Inc., Beijing) with the R10.4.1 kit, following the manufacturer's protocol.

Sample preparation

For the evaluated assays, comprising four WGBS (BS1–BS4), four EM-seq (EM1–EM4), one RRBS (RR1), and three TAPS (PS1–PS3) assays, we prepared a 22-sample panel containing duplicates of all 11 samples (HF, D5, D6, F7, M8, T1, T2, T3, T4, BC, BL). To ensure unbiased processing, centers were blinded to both the source of and relationships between all anonymized gDNA samples.

All aliquots distributed to the centers originated from the same master pool to minimize batch effects. To accommodate the specific sequencing requirements of each center, we prepared distinct samples accordingly. Specifically, for WGBS, EM-seq, RRBS, TAPS, and EPICv2 assays, each sample was aliquoted at 1 μg and stored in 2 mL cryovials. For non-targeted assays involving a base-conversion step, all samples except HF were spiked with 0.2% lambda DNA and 0.01% pUC19 DNA as internal controls. For the RRMS center, each sample was aliquoted at 10 μg without lambda or pUC19 DNA.

The concentrations of the Quartet reference materials (D5, D6, F7, M8, T1, T2, T3, T4) were 220 ng/μL, while BC and BL were at 110 ng/μL, and HF was at 250 ng/μL. All samples were stored at -80°C.

Bioinformatic processing

NGS data were processed with two comparable chemistry-specific standardized pipelines. Reverse-conversion (WGBS/EM-seq/RRBS) data were trimmed using Trim Galore v0.6.10 (https://www.bioinformatics.babraham.ac.uk/projects/trim_galore/), aligned to GRCh38.p14 with bwa-meth v0.2.7 [24], deduplicated using GATK v4.6.2.0 [25], and quantified by MethylDackel v0.6.1 (https://github.com/dpryan79/MethylDackel). For RRBS data, the deduplication step was omitted following established recommendations. For TAPS, the corresponding tools were fastp v0.24.1 [26], BWA-MEME v1.0.6 [27], GATK v4.6.2.0 [25], and asTair v3.3.3 [28].

For EPICv2 data, the raw microarray files were processed with ChAMP v2.26.0 [29] and minfi v1.36.0 [30]. For Oxford Nanopore data, we converted the fast5 files to pod5 format by pod5-tools v0.3.27. Base calling and alignment were conducted with Dorado v1.0.1 (https://github.com/nanoporetech/dorado) and the dna_r10.4.1_e8.2_400bps_hac@v5.2.0_5mC_5hmC@v1 model. Reads with mean base quality below Q7 were removed. The methylation levels of cytosines were called by Modkit v0.5.0 (https://github.com/nanoporetech/modkit). The ploidy of BC sample was calculated from NP-WGS data with Clair3 v2.0.1 [31] and SAVANA v1.3.7 [32].

Performance metrics

To mitigate the confounding effect of sequencing depth, all metrics for sequencing-based assays were calculated using cytosines with ≥10× depth, unless otherwise specified.

Quantitative accuracy and technical reproducibility were assessed using root mean squared error (RMSE) and Spearman's rank correlation (Spearman's rho; polar-rs v1.14.0) [33], respectively. For each library, RMSE was calculated over cytosines present in both the bioinformatic pipeline output and the truth set:

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(\mathrm{M}_{reported,i} - \mathrm{M}_{truth,i}\right)^2} \quad (1)$$

where $\mathrm{M}_{\mathrm{reported,i}}$ and $\mathrm{M}_{\mathrm{truth,i}}$ are the methylation levels of the i-th cytosine in the pipeline output and reference dataset, respectively.

Binomial-standardized root mean squared error (bRMSE) was used to quantify methylation level-dependent error excess after accounting for binomial sampling variance and was calculated for sequencing-based assays at reported cytosines with methylation levels of 10%–90%:

$$bRMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(\frac{\mathrm{M}_{reported,i} - \mathrm{M}_{truth,i}}{\sqrt{M_{\mathrm{truth},i}(100 - M_{\mathrm{truth},i})/d_i}}\right)^2} \quad (2)$$

At each cytosine, the difference between the reported and reference methylation levels was divided by the expected binomial standard deviation. This standard deviation was computed from reference methylation and read depth. The resulting standardized

differences were squared and averaged across cytosines, and bRMSE was the square root of that mean. Because binomial read sampling alone gives an expected squared standardized difference of 1, a bRMSE value of 1 is the stochastic reference point. Values greater than 1 indicate excess variation beyond that expectation, whereas values less than 1 indicate variation below the nominal binomial expectation.

The preservation of expected methylation ranks across the titration series (D5, T1–T4, and D6) was evaluated using Kendall's rank correlation (Kendall's tau). For each assay, technical replicates of D5, T1, T2, T3, T4, and D6 were first averaged at shared cytosines. Cytosines were retained only when all six samples had reported methylation levels and the absolute D5-D6 methylation difference met the prespecified threshold (≥50% or 25%–50%). For each retained cytosine, all 15 pairwise order comparisons among D5, T1, T2, T3, T4, and D6 were counted:

$$Kendall's\ tau = \frac{1}{N}\sum_{i=1}^{N} s_i\left(\frac{C_i - D_i}{15}\right) \quad (3)$$

where Ci and Di are the numbers of concordant and discordant pairwise comparisons for cytosine i under the decreasing or increasing D5, T1, T2, T3, T4, and D6 reference order, respectively. Tied comparisons contribute to neither count. The sign term si equals 1 when MD5,i ≥ MD6,i and -1 when MD5,i < MD6,i. The reported Kendall's tau is the mean site-level value across retained cytosines, with values closer to 1 indicating stronger preservation of the expected sample ordering.

The pairwise relative ΔMeth was calculated using the normalization method described

in Downs et al. [34], with a designated baseline sample in place of the original normal tissue. The “normal” and “tumor” comparison pairs were BL and BC for the BC-BL contrast, BL and D6 for BL-D6, F7 and M8 for F7-M8, and D5 and D6 for D5-D6. For each cytosine, the relative ΔMeth was defined as follows:

$$r_{s,i} = \begin{cases} \dfrac{M_{s,T,i} - M_{s,B,i}}{100 - M_{s,B,i}}, & M_{s,T,i} > M_{s,B,i}, \\ \dfrac{M_{s,B,i} - M_{s,T,i}}{M_{s,B,i}}, & M_{s,T,i} < M_{s,B,i}, \\ 0, & M_{s,T,i} = M_{s,B,i}, \end{cases} \tag{4}$$

where B and T are the baseline and comparison samples, respectively. M is the methylation level for cytosine *i* on the 0–100 scale. The relative ΔMeth ranges from 0 to 1, where 0 indicates no change and 1 indicates the maximum possible change from the baseline in the observed direction. Effect-size error was summarized as the MAE between the reported and truth-derived relative ΔMeth values across the *N* evaluated cytosines:

$$\mathrm{MAE}_{\text{relative }\Delta\text{Meth}} = \frac{1}{N}\sum_{i=1}^{N}\left|r_{\text{reported},i} - r_{\text{truth},i}\right| \tag{5}$$

where lower values indicate more accurate recovery of relative change magnitude and 0 indicates exact agreement in magnitude at all evaluated cytosines. It should be noted that the MAE reflects the magnitude without assessing the agreement in the direction of change.

For data-quality metrics, incomplete and excessive conversion were assessed at reported cytosines with depth ≥10× in spike-in unmethylated lambda phage DNA, methylated pUC19 vector DNA, and mitochondrial DNA (chrM), which was treated as

unmethylated. The incomplete conversion and excessive conversion are calculated as follows:

$$\textit{incomplete conversion} = \frac{1}{N}\sum_{i=1}^{N}\delta_i \text{, where} \delta_i = \begin{cases} 100 - M_i^{\text{obs}}, & \text{if } d = \text{direct} \\ M_i^{\text{obs}} - 0, & \text{if } d = \text{reverse} \end{cases} \quad (6)$$

$$\textit{excessive conversion} = \frac{1}{N}\sum_{i=1}^{N}\epsilon_i \text{, where} \epsilon_i = \begin{cases} M_i^{\text{obs}} - 0, & \text{if } d = \text{direct} \\ 100 - M_i^{\text{obs}}, & \text{if } d = \text{reverse} \end{cases} \quad (7)$$

For the reverse conversion strategy (WGBS, EM-seq, and RRBS), incomplete conversion was quantified using lambda (except for RR1) and chrM, and excessive conversion using pUC19. For the direct conversion strategy (TAPS), these controls were inverted.

Dynamic post-conversion GC content (dGC%) and native GC% were calculated in the ±75 bp window around each cytosine in the reference datasets. For every cytosine:

$$\text{dGC}_i(\%) = 100 \times \frac{N_{G,i} + \sum_{j=1}^{N_{C,i}} M_{ij}}{L_i} \quad (8)$$

where i is the reference cytosine, L_i is the window size, and $N_{G,i}$ is the number of guanines in this window. M_{ij} is the reference methylation level of cytosine j among the $N_{C,i}$ cytosines in this window.

Generalized linear modeling of methylation-level error profiles

To separate intrinsic binomial sampling variance from protocol-specific bias, we decomposed relative mean squared error (Relative MSE) using the modeled bRMSE curves. Let p denote the reference methylation level. Under fixed sequencing depth assumptions, the theoretical binomial relative MSE baseline is proportional to p(1-p).

We defined the error components as follows:

(1) Binomial relative MSE: $MSE_{bin} = p(1-p)$;

(2) Protocol excess relative MSE: $MSE_{exc} = p(1-p) \times max(\hat{Y}^2_{\text{bRMSE}} - 1, 0)$;

(3) Total relative MSE: $MSE_{tot} = MSE_{bin} + MSE_{exc}$

The term $max(\hat{Y}^2_{\text{bRMSE}} - 1, 0)$ effectively censors the excess error at zero, enforcing that any observed bRMSE ≤1 is attributed entirely to expected binomial sampling. The fraction MSE_{exc}/MSE_{tot} was interpreted as the non-binomial excess fraction, representing the proportion of total relative error attributable to protocol-specific overdispersion from library preparation biases or bioinformatic artifacts.

To capture this error excess pattern while controlling for sample dependencies, we constructed the observational dataset directly from paired reference sample matrices (D5 and D6). The primary advantage of this paired design is that it allows for the direct comparison of error performance between different sample dilutions or types within exactly the same genomic background. This enables the local methylation density and sequence complexity of co-occurring sites to be incorporated into the model as a structural control term, rather than being treated as independent, uncontrolled noise. This design reduces the influence of sample-specific methylation-level distributions.

Based on this paired structure, we employed a generalized linear modeling (GLM) framework (Gaussian family, identity link function) to fit the error profiles. For a given assay, the expected error $\mathbb{E}[Y_i]$ was modeled as:

$$\mathbb{E}[Y_i] = M_0 + M_1 p_i + M_2 {p_i}^2 + \gamma S_i + \delta_1 c_i + \delta_2 c_i^2 \quad (9)$$

where Y_i is the observed bRMSE value. p_i and p_i^2 are the raw quadratic polynomial terms for the target methylation bin midpoint, allowing the model to capture non-linear error penalties across the methylation spectrum. S_i is a fixed indicator covariate for the target sample (D5 or D6), accounting for systematic baseline shifts between two reference dilutions. c_i and c_i^2 are the raw quadratic polynomial terms for the context methylation bin midpoint (the paired counterpart in the same matrix row), acting as a structural control to adjust for the local methylation background.

Selection of k-mer size for mappability

We evaluated BisMap and UMap tracks (for reverse-conversion and other methods, respectively) using k-mers of 24, 36, 50, and 100. We retained tracks with an adjusted partial $R^2 > 0.7$ that ranked in the top two for both R^2 and monotonicity (negative Spearman correlation between mappability-bin midpoints and adjusted RMSE). Final selection between the top two tracks minimized the mean RMSE contrast between the 0–10 and 90–100 mappability bins across Quartet samples.

Evaluation of alignment tools

For reverse-converted chemistry, the assays with the best quantitative accuracy for WGBS and EM-seq were selected as representative assays (BS2 and EM3, respectively). For direct-converted chemistry, PS1 and PS3 were chosen. Two reference samples with markedly distinct genomic backgrounds were selected: D6 (Han Chinese, diploid, B-lymphocyte) and BC (Caucasian, aneuploid, breast cancer cell line). For aligner evaluation, BatMeth2 (commit 107c960) [35], BSMAPz v1.1.3 (https://github.com/zyndagj/bsmapz) [36], Bismark v0.25.1

(https://github.com/FelixKrueger/Bismark) and bwa-meth v0.2.7 [24] were assessed for reverse-converted chemistry. For direct-conversion chemistry, Bowtie2 v2.5.4 [37], BWA-MEME v1.0.6 [27], strobealign v0.17.0 [38], and Whisper v2.0.2 [39] were assessed for direct-converted chemistry. In addition, BISCUIT v1.7.0 [40], gem3-mapper v3.6.1-27 (https://github.com/smarco/gem3-mapper), and minibwa v0.2-r379 [41] were utilized across data from both reverse- and direct-converted chemistries.

Identification of tumor-specific difficulty

To evaluate the impact of the tumor genome, we used the breast cancer cell line HCC1395 (BC) and its matched non-tumor counterpart (BL). The BC genome was stratified into six ploidy groups: loss, diploid-like, 3-copy-like, 4-copy-like, moderate-amplification, and high-amplification.

In the GIAB-stratified root mean square error (RMSE) analysis, we focused on multi-assay NGS methods (WGBS, EM-seq, and TAPS). The RMSE of each assay within a specific GIAB stratum was normalized by the corresponding sample-level global RMSE to calculate a stratum-specific penalty. The non-tumor background common difficulty was defined as the median penalty across five non-tumor samples (BL, D5, D6, F7, and M8). The BC-specific excess penalty was calculated by subtracting this background median from the BC penalty.

To identify specific stress regions in the BC sample, we established a high-stringency RMSE cutoff, defined empirically as the maximum absolute localized BC-BL RMSE difference observed among primary NGS assays in the ploidy analysis. GIAB strata were classified based on the following criteria: (1) Strict BC-specific excess: passed the

BC-specific excess rule (exceeding this empirically derived cutoff in at least two assays per primary NGS group) but not the background common-difficulty rule, with BC exhibiting the highest mean penalty; (2) BC-amplified common difficulty: passed both the BC-specific excess and background common-difficulty rules; (3) Difficulty inversion: candidates of strict BC-specific excess where assays were more accurate on non-tumor samples, yet BC showed a positive penalty. Signals present in two or one primary method groups were labeled as partial NGS recurrent and method-specific, respectively.

Statistical tests

Spearman correlations and Mantel tests were computed using linkET v0.1.0 [42] to evaluate monotonic associations among quality metrics, RMSE, and recall. Blocked comparisons across genomic contexts and ploidy states were performed using Friedman tests from R v4.4.3 [43], followed by Nemenyi post hoc tests from PMCMRplus v1.9.12 [44]. The wilcoxon signed-rank test was conducted using SciPy v1.17.1 [45]. To estimate the impact of assay-specific errors and binomial errors, generalized linear models (GLMs) were built using R v4.4.3. Statistical significance was set at a two-sided $P < 0.05$, with Holm correction applied for multiple comparisons within predefined analysis families.

MethRidgeQC

MethRidgeQC was designed to assess the reliability of individual CpG methylation measurements in non-targeted NGS libraries. The two strands were assessed separately, retaining their original methylation values and read counts. For a cytosine with *m*

methylated and u unmethylated reads, the observed methylation proportion was $p = m/(m + u)$. The counting uncertainty was estimated from the Jeffreys posterior, Beta($m + 0.5$, $u + 0.5$). Counting risk was defined as the posterior probability that methylation of this cytosine differed from p by more than 10%. When both strands had ≥1 read counts, the strand discordance was estimated as the probability that methylation levels drawn independently from their posteriors differed by more than 0.10. The larger estimate was used as an internal risk anchor for training. The counting risk was used solely if there was no available reads on the opposite strand.

The ridge regression model included eight variables: sequencing depth on each strand, observed methylation levels, width of the 95% posterior interval, posterior standard deviation, absolute and standardized differences between strands, and an indicator for unavailable opposite-strand information. Only up to 200,000 CpGs with valid counts were selected per library to avoid overfitting by reproducible sampling with seed 42. Missing values were imputed using median values, and variables were standardized separately within each cross-validation training set. The Ridge regularization parameter α was selected from 1, 10 and 100, using four-fold cross-validation. CpG cytosines were grouped by genomic coordinates into fixed, contiguous, non-overlapping 1 Mb blocks, with all records from each block assigned deterministically to the same fold. Following the one-standard-error rule, we selected the largest α whose fold-averaged validation mean squared error (MSE) against the internal risk anchors did not exceed the minimum mean MSE plus the standard error estimated across folds for the minimum-MSE candidate. When training records occupied only two or three genomic

blocks, two- or three-fold cross-validation was used, respectively; when only one block was available, α was fixed at 100 without cross-validation. The model was then refitted to the full training set and used to score all eligible measurements. Predicted risks were restricted to [0, 1], and were divided into four reliability classes: high (H), moderately high (MH), moderately low (ML), and low (L). Each class required at least 1% of the training measurements from at least two genomic blocks. Measurements with invalid or zero-depth counts were left unclassified (U), as were those in libraries with insufficient training data to define four classes. Class boundaries were determined separately for each library.

Results

Study design

We designed the benchmark around a reference panel containing 11 unique genomic DNA (gDNA) samples and a multi-center assay cohort (Fig.1A). The reference panel included four diploid samples from a Quartet family (D5, D6, F7, and M8), a controlled D5-D6 mixture series spanning the two source samples and four intermediate mixtures (D5, T1-T4, and D6), and a tumor-normal pair consisting of the breast cancer cell line HCC1395 (BC) and its matched normal cell line HCC1395BL (BL). The reference panel was organized to evaluate absolute accuracy, relative signal preservation, and conventional technical performance. The diploid Quartet samples provided genome-wide single-base truth for measuring absolute base-level error by root mean squared error (RMSE). The D5-D6 mixture series (D5, T1-T4 and D6) created controlled,

graded methylation differences among adjacent samples, allowing us to test whether each assay preserved biological ordering relevant to inference. This design complemented absolute accuracy without requiring separate reference definitions for the diverse paradigms and bioinformatic tools used in differential methylation analysis. The BC-BL sample pair was used to measure the ability to detect tumor-normal methylation differences. Conversion efficiency and other conventional metrics were calculated using spike-in controls, mitochondrial DNA (chrM), and technical replicates.

The primary NGS cohort was constructed to capture the current diversity of NGS methylation sequencing in routine practice, including conversion chemistries, real-world experimental protocols, and heterogeneous versus harmonized multi-center implementations. This cohort included 12 NGS methylation sequencing assays from 12 sequencing centers using their routine protocols (Fig.1A, Table S1), grouped by base-conversion principle as whole-genome bisulfite sequencing (WGBS), enzymatic methyl sequencing (EM-seq), reduced representation bisulfite sequencing (RRBS), and ten-eleven translocation (TET)-assisted pyridine borane sequencing (TAPS). These abbreviations denote conversion chemistry groups, not commercial kits or vendor-specific implementations. WGBS, RRBS, and EM-seq employ a reverse conversion strategy in which unmethylated cytosines are converted, while TAPS uses a direct conversion strategy, in which methylated cytosines are converted. To explicitly decouple wet-lab protocol variations from bioinformatic heterogeneity, all NGS data were uniformly processed using two conversion strategy-specific yet highly comparable standardized pipelines. In addition, three Illumina EPICv2 microarray

assays, one Oxford Nanopore-based reduced representation methylation sequencing (RRMS) assay, and one approximately 30× Oxford Nanopore whole-genome sequencing (NP-WGS) dataset were used as external benchmarks to contextualize NGS performance limits, rather than for direct head-to-head ranking.

In subsequent genomic-context error profiling, we used two paired designs to reduce confounding. To evaluate methylation-level effects while reducing confounding by sequence context, we modeled quantitative error using methylation-level and error differences between the D5-D6 twin samples at the same genomic locations (cytosines), because the reference truth sets of these twin cell lines share matched genomic backgrounds. In the tumor-genome analysis, we calculated the excess quantitative error of the BC sample relative to the matched BL sample. We combined this with the error excess in regions of interest relative to global error as dual metrics to improve the specificity and robustness of the analysis.

Quantitative accuracy as the basis of NGS performance across scenarios

To ensure our genome-wide benchmark addresses diverse downstream needs, we evaluated quantitative performance across three distinct dimensions. First, root mean squared error (RMSE) measured absolute quantitative accuracy at individual cytosines, which is the basic input of analyses of differential methylation and genetic imprinting (Fig. 1B). Based on diploid Quartet samples, the most favorable WGBS, EM-seq, and TAPS assays converged within a narrow RMSE interval of 7.32–8.38, providing an approximate empirical performance limit. This interval provided a practical NGS baseline because it was reached across chemistries, wet-lab protocols, and sequencing

centers. As external benchmarks with poorer observed performance than NGS, NP-WGS showed an RMSE of 11.26, while EPICv2 and RRMS ranged from 9.01 to 9.79.

The genomic scope of the truth set also affected how faithfully the RMSE estimates represented genome-wide performance. Based on diploid Quartet samples, we compared RMSE across three matched truth-set scopes: BLUEPRINT-covered CpGs (~1.6×10^4), EPICv2-targeted CpGs (~1.9×10^6), and genome-wide truth (~5.1×10^7) (Fig. S1). The restricted truth sets yielded systematically more optimistic accuracy estimates than the genome-wide truth (Wilcoxon $P < 0.01$). Compared with estimates obtained using the genome-wide truth set, BLUEPRINT-based sample-level RMSE estimates were lower by 13.1%–46.8% for NGS assays, by 27.5%–34.9% for EPICv2, and by up to 73.5% and 19.9% for the Oxford Nanopore-based RM1 and NP1 assays, respectively). Using EPICv2-targeted CpGs reduced this underestimation, but the RMSE estimates for NGS assays remained 5.6%–25.6% lower than the global truth (Wilcoxon $P < 0.01$). These results support the need for a genome-wide single-base reference to characterize genome-wide assay performance, whereas smaller targeted truth sets should be interpreted within covered loci.

We next assessed the assays' ability to preserve the expected methylation order across the six-sample titration series using Kendall's tau (Fig. 1C). This reference-free metric relied on ranking to determine whether the expected biological ordering of samples was preserved during gradual methylation changes instead of absolute methylation levels, thereby tolerating uniform shifts or signal compression. At cytosines with methylation differences of ≥10% between adjacent samples, 66.7% (8/12) of NGS assays had

546 tau >0.867 and produced no more than one discordant sample pair on average. When
547 adjacent differences narrowed to 5%–10%, only 50% (6/12) of NGS assays with
548 tau >0.733 maintained no more than two discordant pairs. By contrast, while the
549 EPICv2 assays exhibited higher genome-wide RMSE values than NGS assays, they
550 maintained Kendall's tau >0.97 under both conditions. This demonstrated an almost
551 complete absence of discordant sample pairs. The Oxford Nanopore–based RM1 and
552 NP1 ranked behind all NGS assays, including the outlier BS4. Notably, the methylation
553 levels and differences for D5 and D6 samples were derived from assay outputs rather
554 than reference truth. These results reveal that microarray methods maintain excellent
555 internal consistency when profiling multiple samples within the targeted scope. NGS
556 methods showed partial deficiencies, while Oxford Nanopore methods were highly
557 prone to internal contradictions.

558 Third, we used the mean absolute error (MAE) of relative methylation difference
559 (ΔMeth) [34] to evaluate how accurately assays measured the sizes of methylation
560 changes (Fig. 1D), as downstream analyses inherently depend on effect sizes rather than
561 mere statistical significance. We analyzed four sample-pair contrasts with distinct
562 biological relationships: BC-BL for tumor versus matched normal, BL-D6 for normal
563 individuals with different ancestry backgrounds, F7-M8 for unrelated normal
564 individuals with the same ancestry background, and D5-D6 for the twin-daughter pair
565 with divergent cellular differentiation trajectories. As a result, the F7-M8 pair yielded
566 the highest MAE for all assays, indicating that the biological relationships did not
567 translate into a corresponding order of measurement difficulty. Among evaluated assays,

RR1 had higher MAE than all non-targeted NGS assays in BC-BL (0.25) and BL-D6 (0.25), despite its top ranking by Kendall's tau. Oxford Nanopore-based assays showed a different relationship between the two metrics. NP1 achieved MAE values that were indistinguishable from the NGS assays across all contrasts, while RRMS had the highest MAE among all assays on three contrasts.

Notably, while these metrics fundamentally capture distinct dimensions of performance, converting them into performance rankings reveals the concordance within NGS assays. Specifically, the sample-level RMSE ranking of NGS assays was highly concordant with their ranking for biological order preservation, achieving a Spearman's rho of 0.98 ($P < 0.01$) when adjacent differences were ≥10%, and 0.93 ($P < 0.01$) when adjacent differences were 5%–10%. Furthermore, the sample-level RMSE ranking of the NGS assays is also highly consistent with their ranking based on the MAE of ΔMeth for the D5-D6 pair (Spearman's rho = 0.77, $P < 0.01$). This underscored the critical importance of single-sample quantitative accuracy of NGS assays for the robust detection of subtle methylation differences. The outlier BS4 was excluded from the following analyses.

Genomic sequence context stratifies quantitative errors

We extended the genome-wide assessment by resolving the performance of NGS assays across genomic contexts, and found that GC content and methylation-level stratification converted global RMSE into context-specific error profiles. The GC content strata from Genome-in-a-Bottle (GIAB) stratification resources [46] revealed lower RMSE in extreme native GC% bins than in moderate bins, indicating that unconverted GC content alone did not fully capture the influence (Fig. 2A). Considering the base-

conversion state, we defined dynamic GC content (dGC%) of ±75 bp flanking sequence for every cytosine (see Methods). At a given native GC%, lower dGC% reflects hypomethylated contexts that become AT-rich after reverse conversion, whereas higher dGC% indicates cytosine retention and increased GC-richness.

Accounting for dGC% separated distinct error landscapes across GC% bins (Figs. 2B, S2, S3). EM-seq and TAPS peaked at moderate native GC (60–80%) with low-to-moderate dGC (20–40%). Meanwhile, WGBS, RRBS, and Oxford Nanopore assays showed higher RMSE in GC-poor/low-dGC contexts (0–40%). While RRBS (RR1) exhibited the narrowest RMSE range (4.0) across strata, it did not universally achieve the minimal absolute errors. In contrast, in GC-rich and high-dGC contexts (both 65%–100%), RRMS (RMSE=1.8) and TAPS (RMSE=2.4 each) achieved the lowest quantitative errors across all assays and GIAB strata. Targeted EPICv2 assays showed a distinct profile, with maximal errors concentrated in extreme GC-rich/high-dGC regions (80–100%).

In the further analysis of the impact of methylation levels, we found that raw RMSE was lower at the methylation extremes and increased at intermediate methylation levels across all sequencing-based assays, producing inverted U-shaped profiles. This pattern is expected for sequencing-based methylation estimates, where binomial sampling makes counting noise depend on methylation level and read depth [47–49]. To evaluate assay-specific effects of methylation level beyond the binomial sampling variance, we introduced binomial-standardized RMSE (bRMSE) (Methods). By normalizing observed errors against the expected stochastic variance given reference methylation

and read depth, bRMSE helped separate assay-specific, non-binomial, methylation level-dependent error excess from baseline counting noise. Specifically, a bRMSE value of 1.0 indicates errors consistent with binomial read sampling alone. A value greater than 1.0 indicates that there is assay-specific error excess stemming from experimental practice or bioinformatic processing beyond the sampling noise.

Modeling bRMSE from the D5/D6 sample pair using a generalized linear model (see Methods) showed that WGBS had the lowest mean bRMSE across methylation-level bins (1.082), followed by TAPS (1.095) and EM-seq (1.106), whereas RRMS, NP-WGS, and RRBS showed higher error excess (1.263, 1.370, and 1.489, respectively). The assay-level curves revealed distinct methylation-dependent profiles (Fig. 2C). Specifically, EM-seq, RRMS, and NP-WGS assays exhibited larger error excess at low methylation levels. In contrast, the error excess of TAPS increased toward high methylation levels, while RR1 maintained consistently high error excess. The performance of WGBS varied considerably across different assays. RMSE and bRMSE decomposition indicated that the higher RMSE of WGBS and TAPS relative to EM-seq was primarily driven by the sequencing depth-associated binomial component. Conversely, more than half of the total error in RR1 was non-binomial, strongly suggesting the presence of systematic bias that cannot be mitigated by increasing sequencing depth.

Mappability differentiates chemistry-specific quantitative sensitivity

Low-mappability regions are often enriched for repetitive and structurally complex genomic elements and represent a persistent challenge in single-base-resolution

epigenomics. While binary GIAB strata showed a reproducible (Wilcoxon $P < 0.01$ for NGS methods) but marginal RMSE increment across NGS assays (0.03–1.56, Fig. 3A), this obscured assay-specific sensitivities. To resolve the continuous impact of mappability on quantitative fidelity, we optimized assay-specific k-mer sizes of BisMap (for reverse conversion chemistries) or UMap tracks (see Methods) [50]. WGBS assays selected K24 (2/3, no candidate for BS1), EM-seq assays split between K24 (1/4) and K100 (3/4), and TAPS assays split between K24 (1/3) and K36 (2/3). By RMSE range, EM-seq exhibited the highest robustness to mappability variations (1.48–2.11), whereas targeted assays were highly susceptible, with EPICv2 showing extreme sensitivity (15.19–23.16) (Fig. 3B).

Quantitative sensitivity to mappability also depended on the aligner used. Under reverse conversion (Fig. 3C), RMSE ranges varied from 1.18 (Bismark) to 2.52 (BSMAPz). Under direct conversion (Fig. 3D), aligners exhibited generally higher and more variable sensitivities, with RMSE ranges spanning from 2.04 (BISCUIT) to 4.42 (Bowtie2). For several tools (e.g., BISCUIT, Bismark/Bowtie2, gem3-mapper, and minibwa), both the selected k-mer size and the RMSE range differed between the two conversion directions. These results indicate that alignment algorithms should be considered when interpreting mappability-associated methylation error.

Tumor genome contexts reveal localized quantitative difficulty

While previous analyses established baseline error profiles in diploid genomes, tumor genomes introduce additional challenges because somatic CNVs and aneuploidy can alter sequencing analysis. We used a matched breast cancer cell line (BC, HCC1395)

and its normal counterpart (BL) as a stress-test model. We stratified the BC genome into six ploidy states, ranging from loss to high-amplification (Fig. 4A). Friedman rank-sum and Nemenyi post hoc tests showed that copy-number-stratified RMSE difference (ΔRMSE, BC minus BL) differed significantly among these states in both genome-wide and GIAB "NotInAllDifficultRegions" analyses (both $P < 0.01$; Fig. 4B). Specifically, non-targeted NGS methods (WGBS, EM-seq, TAPS) exhibited the highest ΔRMSE in loss and diploid-like regions, whereas high- and moderate-amplification regions maintained constrained error inflation. Conversely, this ordering was completely reversed for targeted methods (RRBS and RRMS) (Fig. 4C), revealing opposite quantitative vulnerabilities to tumor aneuploidy. Across all evaluated methods, non-targeted NGS assays had overall ΔRMSE values from 0.47 (EM1) to 1.69 (PS3), comparable with NP-WGS (0.97) and smaller than EPICv2 (ΔRMSE range: 2.40–4.13) and RRMS (5.66). The maximum absolute localized ΔRMSE was smaller for NGS assays (2.33, observed in BS3 within high-amplification regions) than for EPICv2 (4.12), RRMS (4.44), and NP-WGS (3.95).

Ploidy states captured broad error inflation associated with tumor aneuploidy, but this stratification could obscure local genomic features linked to these biases. To further localize these difficulties, we mapped ΔRMSE across 85 functionally and structurally annotated GIAB strata (Fig. 4D). For each assay, stratum RMSE was first normalized by subtracting the corresponding sample-level global RMSE. The "common difficulty" was defined as the median normalized penalty across BL, D5, D6, F7, and M8 samples. The "BC-extra penalty" was defined as the normalized penalty in the BC sample minus

this “common difficulty” background. We applied a practical effect-size threshold of 2.33 RMSE units, corresponding to the maximum absolute localized ΔRMSE observed across non-targeted NGS assays in the preceding ploidy analysis, and required recurrence in at least two assays within each of WGBS, EM-seq, and TAPS.

Seven strata met the strict BC-specific criteria without meeting the common-difficulty criterion. Among them, the most pronounced excess occurred in “SimpleRepeat_triTR_50to149_slop5” (4.75 RMSE units), followed by “population_CNV_FP_regions” (3.84), “gnomAD_InbreedingCoeff_slop1bp_merge1000bp” (3.77), and “SimpleRepeat_diTR_50to149_slop5” (3.65). The remaining three strata, including “gc85_slop50”, “gc75to80_slop50”, and “BadPromoters”, had median BC-extra penalties of 2.68–3.05. Two additional strata were classified as common difficulties that were amplified in BC sample, including “CMRGv1.00_duplicationinKMT2C” and “collapsed_duplication_FP_regions”, with median common difficulties of 19.16 and 2.83, respectively. The KMT2C stratum represented an extreme outlier. Under this classification framework, in the HCC1395 genome, recurrent error excess involved specific repeat, CNV-associated, GC-rich, promoter, and duplication contexts.

Reference-free quality control selects reliable CpG methylation measurements

The landscape of quantitative errors observed in NGS assays highlighted the need to identify reliable CpG measurements with individual samples before downstream analyses. To make such quality control applicable without matched ground truth, we first tested whether routine metrics could serve as proxies for quantitative accuracy

(Table S2). Quantitative reproducibility between technical replicates, measured by Spearman's rank correlation (Spearman's rho), was weakly associated with quantitative accuracy, with $R^2 < 0.15$ across all three benchmarking scopes (Fig. 5A). Although NGS assays showed distinct conversion-efficiency and intermediate-methylation bias patterns (Fig. 5B), including incomplete conversion and underestimation in TAPS assays, Spearman and Mantel tests did not support these metrics as substitutes for sample-level RMSE. This result held under the pre-specified association criterion of both Spearman's rho and Mantel's $r \geq 0.7$ with $P < 0.05$ in both tests, regardless of whether the outlier BS4 was excluded (Table S3). Nevertheless, conventional metrics retained diagnostic value, as the abnormal chrM conversion rate of outlier BS4 (82.5% vs. 97.8%–99.5% for other WGBS) flagged its poor performance (Table S4). These results underscored the need for a quality control approach beyond conventional metrics to ensure quantitative accuracy.

These results showed the limitations of sample-level QC and underscored the need for a practical single-base QC approach. We developed MethRidgeQC, a QC tool that could assess quantitative accuracy without relying on conventional metrics. Non-targeted NGS methods (WGBS, EM-seq, and TAPS) were focused because of their dominant depth-dependent binomial error. We incorporated this error structure together with the strand concordance for CpGs as a QC indicator in our previous work [23]. MethRidgeQC trains a ridge regression model independently for every sequencing library, using internal risk anchor derived from counting uncertainty and strand discordance as surrogate training targets, eliminating the need for matched ground truth.

Each CpG cytosine is then assigned to one of five reliability classes: high reliability (H), moderately high reliability (MH), moderately low reliability (ML), low reliability (L), and unclassified (U).

Based on the genome-wide reference, we found that MethRidgeQC successfully we found that MethRidgeQC distinguished CpG methylation measurements with lower and higher quantitative error across the six original samples (D5, D6, F7, M8, BC, and BL). For practical filtering, we evaluated two selection criteria: retaining H measurements alone or retaining both H and MH measurements (H+MH) (Figs. 6A, S4-S14). Across non-targeted NGS libraries excluding the outlier BS4, H+MH had 51.4%–73.7% lower RMSE than the ML+L, while H alone had 52.6%–72.6% lower RMSE than the remaining classified measurements (MH, ML, and L). The H+MH selection outperformed depth filtering, yielding 26.1%–46.9% lower RMSE than a 5× cutoff and 13.5%–42.6% lower RMSE than a 10× cutoff (Fig. 6B). Within both the 5–10× and ≥10× depth strata, quantitative accuracy generally improved from L to H (Figs. 6C, S4-S14), indicating that the reliability classes provided useful discrimination even at moderate sequencing depth. These accuracy gains were accompanied by substantial data retention: the H and H+MH selections retained 17.2%–48.4% and 54.0%–78.7% of CpG cytosine measurements per library, respectively, with both selections yielding lower RMSE than the corresponding unfiltered libraries. MethRidgeQC also supported filtering of the poorly performing BS4 libraries, where H+MH measurements had 50.7%–68.1% lower RMSE than measurements in the ML and L classes combined. The H+MH selection excluded an average of more than 18.4 million low-reliability CpG

744 cytosine measurements per BS4 library, while the H and H+MH selections retained
745 averages of 12.1 million and 30.3 million measurements, respectively.

746

747 **Discussion**

748 Methylation readouts are shaped by protocol, platform, coverage, CpG capture,
749 normalization, and bioinformatic workflows, rather than by biology alone [14–16, 18–
750 21]. Extending earlier selected-locus benchmarking [22], our genome-wide single-base
751 references allow direct evaluation across broader sequence and tumor genomic contexts.
752 This scope matters because reproducibility and conversion efficiency assess
753 complementary aspects of methylation sequencing quality without establishing
754 numerical accuracy at the loci used downstream [15, 51]. Therefore, routine quality
755 control needs to be complemented by direct quantitative assessment.

756 Method selection for methylation profiling depends on the intended analysis. Previous
757 cross-method benchmarks have reported high concordance alongside differences in
758 CpG recovery, coverage, GC bias, and platform behavior across NGS, long-read
759 sequencing, and microarray-based methods [15, 18, 52, 53]. The performance
760 differences among these methods in our benchmark reveals why such apparent
761 concordance demands cautious interpretation. Furthermore, although absolute
762 methylation accuracy and biological order preservation yielded similar assay rankings,
763 the separation in performance between assays differed substantially across these
764 dimensions. In several differential methylation scenarios, the Spearman correlations fell
765 below 0.7 between accuracy measures. Method selection and cross-platform integration

should therefore consider the analytical priorities and the magnitude of the biological differences under investigation, with EM-seq demonstrating the most robust overall performance among the evaluated methods (Fig. S15).

Separating absolute error from error above the binomial expectation helps identify priorities for assay improvement. Where depth-dependent sampling error predominates, increasing coverage is a plausible route to better precision. By contrast, persistent excess error motivates investigation of specific genomic contexts and wet-lab protocols. For RRBS, this limitation stems from the restriction-enzyme digestion (e.g., MspI) and subsequent size selection, and approaches like expanded enzyme designs and unique molecular identifiers (UMIs) have been shown to improve performance [54, 55]. Oxford Nanopore methylation detection relies on changes in nanopore ionic current signals. Corresponding methylation inference algorithms are sensitive to k-mer and other contexts [56–59].

MethRidgeQC translates the error analyses into base-level quality control by using internal evidence to stratify CpG measurements by reliability, without requiring an external reference. The H+MH strategy offers a practical balance between quantitative reliability and the number of CpGs available for further analysis, while the H-only strategy provides a more stringent option that require greater quantitative accuracy.

Genomic contexts further define where these conclusions apply. The lower error observed in GC-rich regions is conditional on our consensus reference and ≥10× evaluation threshold. It therefore describes accuracy among retained measurements, while earlier studies examined coverage loss during library preparation and conversion

[14, 15, 18, 60]. Likewise, the dependence of quantitative error on chemistry-specific mappability extends alignment-focused evaluations [19, 61, 62], and supports benchmarking across multiple conversion-aware mappability scales, including k-mer scales close to explicit seed or filtration lengths where such models are defined. In thee HCC1395 tumor genome, elevated errors in regions with complex copy representation are consistent with known sequencing and reference-associated difficulties [46]. Population-CNV, excess heterozygosity and collapsed duplication annotations refer to regions where differences in copy representation between the linear reference genome and the sample can complicate read assignment and variant interpretation [63], highlighting the need to use telomere-to-telomere and pangenome references in further epigenetic analyses [64–66].

Several boundaries should guide the interpretation of this study. Our real-world, multi-center design captures routine assay performance and reflects the availability and implementation of different conversion chemistries, but it introduces inherent protocol diversity, preventing attributing observed errors to specific variables, and estimation of the optimized, best-case limits of each chemistry [14–16, 52, 67]. Consequently, this study should not be interpreted as an idealized method comparison. The tumor analysis relies on only one matched pair, and the targeted and nanopore methods are represented by limited assays. Challenging clinical specimens like formalin-fixed paraffin-embedded (FFPE) samples and circulating cell-free DNA (cfDNA) were excluded to control pre-analytical confounding, so performance in fragmented, damaged, or low-input material requires separate validation [17, 68–70]. Besides, MethRidgeQC also

requires broader evaluation across sample types and assay implementations.

Together, our study connects the genome-wide reference with practical quality control measures by genome-wide benchmarking. Rather than asking whether a method is globally acceptable, it identifies where methylation measurements remain numerically reliable for biological inferences and translational applications.

Declarations

Ethics approval and consent to participate

The need for approval was waived since there was no human or animal materials involved in this study. All DNA materials were constructed from commercial DNA standards.

Consent for publication

Not applicable.

Availability of data and materials

The raw sequence and microarray data reported in this paper have been deposited in the Genome Sequence Archive (HRA010124) and OMIX (OMIX012036) at the National Genomics Data Center, China National Center for Bioinformation / Beijing Institute of Genomics, Chinese Academy of Sciences [71, 72]. The annotated reference datasets and analysis scripts supporting the findings of this study are openly available at 10.5281/zenodo.23587832. The Snakemake pipeline for processing NGS raw data is available and maintained at: https://github.com/YuanfengZhang/MethFlow.

Competing interests

The authors declare that they have no competing interests.

Funding

This study was supported in part by Noncommunicable Chronic Diseases-National Science and Technology Major Project (2024ZD0533100/2024ZD0533102), National Key R&D Program of China (2023YFC3402500/2023YFC3402503), and Beijing Natural Science Foundation (L254025).

Authors' contributions

Y. Zhang and YM prepared samples and analyzed the sequencing data. Y. Zhang also performed the statistical analyses. XG analyzed the Illumina microarray data, prepared the ground-truth datasets, and performed the statistical analyses. Y. Zhang, YM, and XG drafted the manuscript and prepared the figures. DW performed the statistical analyses, revised the manuscript, and acquired funding. YL performed the sequencing experiments, prepared the sample ground-truth sets, and revised the manuscript. YH prepared samples and coordinated the multicenter cohorts. QC performed the sequencing experiments and prepared the sample ground-truth sets. LL analyzed the sequencing data. YSM coordinated the multicenter cohorts. SD prepared the sample ground-truth sets. CL, HH, and ZR performed sample validation. WH and YY performed the sequencing experiments. LD and LS conceptualized the study and revised the manuscript. JL conceptualized the study, revised the manuscript, and acquired funding. Y. Zheng conceptualized and designed the study and revised the manuscript. RZ conceptualized and designed the study, supervised the research, revised

the manuscript, and acquired funding. All authors reviewed and approved the final version of the manuscript.

Acknowledgements

The computations in this research were performed using the CFFF platform of Fudan University. The authors are grateful to the Quartet Project team members for their valuable contributions to the design and execution of this project, and to all the sequencing centers participating in this study.

References

1. Loyfer N, Magenheim J, Peretz A, Cann G, Bredno J, Klochendler A, et al. A DNA methylation atlas of normal human cell types. Nature. Nature Publishing Group; 2023;613(7943):355–64.

2. Menezes NA, Peterson KJ, Guo X, Castiglioni V, Kalvisa A, Filimonow K, et al. Cell-context response to germ layer differentiation signals is predetermined by the epigenome in regionalized epiblast populations. Nat Commun. Nature Publishing Group; 2025;16(1):5000.

3. Rosenski J, Sabag O, Marcus E, Loyfer N, Dor Y, Cedar H, et al. The genetic basis for DNA methylation variation across tissues and development. Nat Commun. 2026;17(1):5194.

4. Wang M, Zhao J, Wang Y, Mao Y, Zhao X, Huang P, et al. Genome-wide DNA methylation analysis reveals significant impact of long-term ambient air pollution exposure on biological functions related to mitochondria and immune response. Environ Pollut. Barking, Essex : 1987; 2020;264:114707.

5. Moqri M, Cipriano A, Simpson DJ, Rasouli S, Murty T, de Jong TA, et al. PRC2-AgeIndex as a universal biomarker of aging and rejuvenation. Nat Commun. Nature Publishing Group; 2024;15(1):5956.

6. Heeke S, Gay CM, Estecio MR, Tran H, Morris BB, Zhang B, et al. Tumor- and circulating-free DNA methylation identifies clinically relevant small cell lung cancer subtypes. Cancer Cell. 2024;42(2):225-237.e5.

7. Grand RS, Burger L, Gräwe C, Michael AK, Isbel L, Hess D, et al. BANP opens

chromatin and activates CpG-island-regulated genes. Nature. Nature Publishing Group; 2021;596(7870):133–7.

8. Kremer LPM, Cerrizuela S, El-Sammak H, Al Shukairi ME, Ellinger T, Straub J, et al. DNA methylation controls stemness of astrocytes in health and ischaemia. Nature. Nature Publishing Group; 2024;634(8033):415–23.

9. Liu J, Dai L, Wang Q, Li C, Liu Z, Gong T, et al. Multimodal analysis of cfDNA methylomes for early detecting esophageal squamous cell carcinoma and precancerous lesions. Nat Commun. Nature Publishing Group; 2024;15(1):3700.

10. Batra RN, Lifshitz A, Vidakovic AT, Chin S-F, Sati-Batra A, Sammut S-J, et al. DNA methylation landscapes of 1538 breast cancers reveal a replication-linked clock, epigenomic instability and cis-regulation. Nat Commun. Nature Publishing Group; 2021;12(1):5406.

11. Zhu Z, Zhou Q, Sun Y, Lai F, Wang Z, Hao Z, et al. MethMarkerDB: a comprehensive cancer DNA methylation biomarker database. Nucleic Acids Res. 2024;52(D1):D1380–92.

12. Medina JE, Dracopoli NC, Bach PB, Lau A, Scharpf RB, Meijer GA, et al. Cell-free DNA approaches for cancer early detection and interception. J Immunother Cancer. 2023;11(9):e006013.

13. Pharo H, Vedeld HM, Sjurgard IV, Pinto R, Lind GE. From concept to clinic: a roadmap for DNA methylation biomarkers in liquid biopsies. Oncogene. Nature Publishing Group; 2025;44(49):4814–31.

14. Olova N, Krueger F, Andrews S, Oxley D, Berrens RV, Branco MR, et al. Comparison of whole-genome bisulfite sequencing library preparation strategies identifies sources of biases affecting DNA methylation data. Genome Biol. 2018;19(1):33.

15. Foox J, Nordlund J, Lalancette C, Gong T, Lacey M, Lent S, et al. The SEQC2 epigenomics quality control (EpiQC) study. Genome Biol. 2021;22(1):332.

16. Morrison J, Koeman JM, Johnson BK, Foy KK, Beddows I, Zhou W, et al. Evaluation of whole-genome DNA methylation sequencing library preparation protocols. Epigenetics Chromatin. 2021;14(1):28.

17. Kresse SH, Brandt-Winge S, Pharo H, Flatin BTB, Jeanmougin M, Vedeld HM, et al. Evaluation of commercial kits for isolation and bisulfite conversion of circulating cell-free tumor DNA from blood. Clin Epigenetics. 2023;15(1):151.

18. Guanzon D, Ross JP, Ma C, Berry O, Liew YJ. Comparing methylation levels assayed in GC-rich regions with current and emerging methods. BMC Genomics. 2024;25(1):741.

19. Nunn A, Otto C, Stadler PF, Langenberger D. Comprehensive benchmarking of software for mapping whole genome bisulfite data: from read alignment to DNA methylation analysis. Brief Bioinform. 2021;22(5):bbab021.

20. Gong W, Pan X, Xu D, Ji G, Wang Y, Tian Y, et al. Benchmarking DNA methylation analysis of 14 alignment algorithms for whole genome bisulfite sequencing in mammals. Comput Struct Biotechnol J. 2022;20:4704–16.

21. Lin Y-Y, Breuer K, Weichenhan D, Lafrenz P, Sarnataro A, Wilk A, et al. Pipeline Olympics: continuable benchmarking of computational workflows for DNA methylation sequencing data against an experimental gold standard. Nucleic Acids Res. 2025;53(19):gkaf970.

22. Bock C, Halbritter F, Carmona FJ, Tierling S, Datlinger P, Assenov Y, et al. Quantitative comparison of DNA methylation assays for biomarker development and clinical applications. Nat Biotechnol. Nature Publishing Group; 2016;34(7):726–37.

23. Guo X, Chen Q, Zhang Y, Zhang Y, Liu Y, Duan S, et al. Methylation reference datasets from quartet DNA materials for benchmarking epigenome sequencing. Nat Commun. Nature Publishing Group; 2025;16(1):9202.

24. Pedersen BS, Eyring K, De S, Yang IV, Schwartz DA. Fast and accurate alignment of long bisulfite-seq reads [Internet]. arXiv; 2014 [cited 2026 Apr 16]. https://doi.org/10.48550/arXiv.1401.1129

25. Van der Auwera GA, O'Connor BD. Genomics in the cloud: using Docker, GATK, and WDL in Terra [Internet]. Sebastopol (CA): O'Reilly Media; 2020. https://www.oreilly.com/library/view/genomics-in-the/9781491975183/

26. Chen S. Ultrafast one-pass FASTQ data preprocessing, quality control, and deduplication using fastp. iMeta. 2023;2(2):e107.

27. Jung Y, Han D. BWA-MEME: BWA-MEM emulated with a machine learning approach. Bioinformatics. 2022;38(9):2404–13.

28. Velikova GV. Development and application of computational methods to study DNA modifications [Internet] [http://purl.org/dc/dcmitype/Text]. University of Oxford; 2020 [cited 2026 Apr 16]. https://ora.ox.ac.uk/objects/uuid:9c7e449a-f0c1-439d-82c6-1b64e52d2dd6. Accessed 16 Apr 2026

29. Tian Y, Morris TJ, Webster AP, Yang Z, Beck S, Feber A, et al. ChAMP: updated methylation analysis pipeline for Illumina BeadChips. Bioinformatics. 2017;33(24):3982–4.

30. Aryee MJ, Jaffe AE, Corrada-Bravo H, Ladd-Acosta C, Feinberg AP, Hansen KD, et al. Minfi: a flexible and comprehensive Bioconductor package for the analysis of Infinium DNA methylation microarrays. Bioinformatics. 2014;30(10):1363–9.

31. Zheng Z, Li S, Su J, Leung AW-S, Lam T-W, Luo R. Symphonizing pileup and full-alignment for deep learning-based long-read variant calling. Nat Comput Sci. Nature Publishing Group; 2022;2(12):797–803.

32. Elrick H, Sauer CM, Espejo Valle-Inclan J, Trevers K, Tanguy M, Zumalave S, et al. SAVANA: reliable analysis of somatic structural variants and copy number aberrations using long-read sequencing. Nat Methods. Nature Publishing Group; 2025;22(7):1436–46.

33. Vink R, Gooijer S de, Beedie A, Gorelli ME, Peters O, nameexhaustion, et al. pola-rs/polars: Python Polars 1.14.0 [Internet]. Zenodo; 2024 [cited 2026 Mar 9]. https://doi.org/10.5281/zenodo.14176391

34. Downs BM, Hu J, Park JS, Lei H, Wang T-H, Pisanic TR, et al. An unbiased approach to measure aberrant DNA methylation alterations. Nat Commun. Nature Publishing Group; 2026;17(1):4522.

35. Zhou Q, Lim J-Q, Sung W-K, Li G. An integrated package for bisulfite DNA methylation data analysis with Indel-sensitive mapping. BMC Bioinformatics. 2019;20(1):47.

36. Xi Y, Li W. BSMAP: whole genome bisulfite sequence MAPping program. BMC Bioinformatics. 2009;10(1):232.

37. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nat Methods. Nature Publishing Group; 2012;9(4):357–9.

38. Sahlin K. Strobealign: flexible seed size enables ultra-fast and accurate read alignment. Genome Biol. 2022;23(1):260.

39. Deorowicz S, Debudaj-Grabysz A, Gudyś A, Grabowski S. Whisper: read sorting allows robust mapping of DNA sequencing data. Bioinformatics. 2019;35(12):2043–50.

40. Zhou W, Johnson BK, Morrison J, Beddows I, Eapen J, Katsman E, et al. BISCUIT: an efficient, standards-compliant tool suite for simultaneous genetic and epigenetic inference in bulk and single-cell studies. Nucleic Acids Res. 2024-2-27;52(6):e32.

41. Li H, Homer N. Fast genomic read alignment with minibwa [Internet]. arXiv; 2026 [cited 2026 July 19]. https://doi.org/10.48550/arXiv.2606.15357

42. Houyun H. Hy4m/linkET: Everything is Linkable [Internet]. Hy4mlinkET Everything Linkable. [cited 2026 July 15]. https://github.com/Hy4m/linkET. Accessed 15 July 2026

43. R Core Team. R: A Language and Environment for Statistical Computing [Internet]. Vienna, Austria: R Foundation for Statistical Computing; 2025. https://www.R-

project.org/

44. Pohlert T. PMCMRplus: Calculate Pairwise Multiple Comparisons of Mean Rank Sums Extended [Internet]. 2024 [cited 2026 July 15]. https://cran.r-project.org/web/packages/PMCMRplus/index.html. Accessed 15 July 2026

45. Virtanen P, Gommers R, Oliphant TE, Haberland M, Reddy T, Cournapeau D, et al. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. Nat Methods. 2020;17:261–72.

46. Dwarshuis N, Kalra D, McDaniel J, Sanio P, Alvarez Jerez P, Jadhav B, et al. The GIAB genomic stratifications resource for human reference genomes. Nat Commun. Nature Publishing Group; 2024;15(1):9029.

47. Hansen KD, Langmead B, Irizarry RA. BSmooth: from whole genome bisulfite sequencing reads to differentially methylated regions. Genome Biol. 2012;13(10):R83.

48. Park Y, Figueroa ME, Rozek LS, Sartor MA. MethylSig: a whole genome DNA methylation analysis pipeline. Bioinformatics. 2014;30(17):2414–22.

49. Park Y, Wu H. Differential methylation analysis for BS-seq data under general experimental design. Bioinformatics. 2016;32(10):1446–53.

50. Karimzadeh M, Ernst C, Kundaje A, Hoffman MM. Umap and Bismap: quantifying genome and methylome mappability. Nucleic Acids Res. 2018;46(20):e120.

51. Liang N, Li B, Jia Z, Wang C, Wu P, Zheng T, et al. Ultrasensitive detection of circulating tumour DNA via deep methylation sequencing aided by machine learning. Nat Biomed Eng. Nature Publishing Group; 2021;5(6):586–99.

52. Liu X, Pang Y, Shan J, Wang Y, Zheng Y, Xue Y, et al. Beyond the base pairs: comparative genome-wide DNA methylation profiling across sequencing technologies. Brief Bioinform. 2024;25(5):bbae440.

53. Sigurpalsdottir BD, Stefansson OA, Holley G, Beyter D, Zink F, Hardarson MÞ, et al. A comparison of methods for detecting DNA methylation from long-read sequencing of human genomes. Genome Biol. 2024;25(1):69.

54. Wang J, Xia Y, Li L, Gong D, Yao Y, Luo H, et al. Double restriction-enzyme digestion improves the coverage and accuracy of genome-wide CpG methylation profiling by reduced representation bisulfite sequencing. BMC Genomics. 2013;14(1):11.

55. Wang K, Li X, Dong S, Liang J, Mao F, Zeng C, et al. Q-RRBS: a quantitative reduced representation bisulfite sequencing method for single-cell methylome analyses. Epigenetics. Taylor & Francis; 2015;10(9):775–83.

56. Liu Y, Rosikiewicz W, Pan Z, Jillette N, Wang P, Taghbalout A, et al. DNA methylation-calling tools for Oxford Nanopore sequencing: a survey and human epigenome-wide evaluation. Genome Biol. 2021;22(1):295.

57. Yuen ZW-S, Srivastava A, Daniel R, McNevin D, Jack C, Eyras E. Systematic benchmarking of tools for CpG methylation detection from nanopore sequencing. Nat Commun. Nature Publishing Group; 2021;12(1):3438.

58. Straver R, Vermeulen C, Verity-Legg JR, Pagès-Gallego M, Stoker DGG, van Oudenaarden A, et al. ReQuant: improved base modification calling by k-mer value imputation. Nucleic Acids Res. 2025;53(9):gkaf323.

59. Kulkarni O, Mathew RJ, Jana R, Zaveri L, Ara S, Nagabandi T, et al. Comprehensive benchmarking of tools for nanopore-based detection of DNA methylation. Nat Commun. Nature Publishing Group; 2026; https://doi.org/10.1038/s41467-026-75183-6

60. de Abreu AR, Ibrahim J, Lemonidis V, Mateiu L, Van Camp G, Op de Beeck K. Comparison of current methods for genome-wide DNA methylation profiling. Epigenetics Chromatin. 2025;18(1):57.

61. Kunde-Ramamoorthy G, Coarfa C, Laritsky E, Kessler NJ, Harris RA, Xu M, et al. Comparison and quantitative verification of mapping algorithms for whole-genome bisulfite sequencing. Nucleic Acids Res. 2014;42(6):e43.

62. Tsuji J, Weng Z. Evaluation of preprocessing, mapping and postprocessing algorithms for analyzing whole genome bisulfite sequencing data. Brief Bioinform. 2016;17(6):938–52.

63. Aganezov S, Yan SM, Soto DC, Kirsche M, Zarate S, Avdeyev P, et al. A complete reference genome improves analysis of human genetic variation. Science. American Association for the Advancement of Science; 2022;376(6588):eabl3533.

64. Olson ND, Wagner J, Dwarshuis N, Miga KH, Sedlazeck FJ, Salit M, et al. Variant calling and benchmarking in an era of complete human genome sequences. Nat Rev Genet. Nature Publishing Group; 2023;24(7):464–83.

65. Taylor DJ, Eizenga JM, Li Q, Das A, Jenike KM, Kenny EE, et al. Beyond the Human Genome Project: The Age of Complete Human Genome Sequences and Pangenome References. Annu Rev Genomics Hum Genet. Annual Reviews; 2024;25(Volume 25, 2024):77–104.

66. Folland C, Monahan G, Breen J, Johari M, Patel HR, Ravenscroft G. Rare disease genomics in an era of human pangenomics and telomere-to-telomere genome references. Eur J Hum Genet. Nature Publishing Group; 2026;34(8):1033–46.

67. Zhou L, Ng HK, Drautz-Moses DI, Schuster SC, Beck S, Kim C, et al. Systematic

evaluation of library preparation methods and sequencing platforms for high-throughput whole genome bisulfite sequencing. Sci Rep. 2019;9(1):10383.

68. Ohmomo H, Komaki S, Ono K, Sutoh Y, Hachiya T, Arai E, et al. Evaluation of clinical formalin-fixed paraffin-embedded tissue quality for targeted-bisulfite sequencing. Pathol Int. 2021;71(2):135–40.

69. Chen X, Liu J, Li J, Xie Y, Yu Z, Shen L, et al. Identification of DNA methylation and genetic alteration simultaneously from a single blood biopsy. Genes Genomics. 2023;45(5):627–35.

70. Cheng JC, Swarup N, Morselli M, Huang W-L, Aziz M, Caggiano C, et al. Single-stranded pre-methylated 5mC adapters uncover the methylation profile of plasma ultrashort Single-stranded cell-free DNA. Nucleic Acids Res. 2024;52(11):e50.

71. Zhang S, Chen X, Jin E, Wang A, Chen T, Zhang X, et al. The GSA Family in 2025: A Broadened Sharing Platform for Multi-omics and Multimodal Data. Genomics Proteomics Bioinformatics. 2025;23(4):qzaf072.

72. CNCB-NGDC Members and Partners. Database Resources of the National Genomics Data Center, China National Center for Bioinformation in 2025. Nucleic Acids Res. 2025;53(D1):D30–44.

Figures Legends

Fig. 1. Study design and quantitative performance of evaluated assays. **A**: Reference materials and multicenter assay cohort. **B**: Sample-level quantitative accuracy (RMSE) across assays and Quartet samples. **C**: Kendall's tau for D5-D6 titration series methylation with 10% or 5% adjacent methylation differences. **D**: Mean absolute error (MAE) of relative ΔMeth estimates across four biological contrasts.

Fig. 2. Quantitative accuracy influenced by GC content and methylation state of template DNA across methods. **A**: Impact of GC content on quantitative accuracy across methods. RMSE values are grouped by GIAB-style GC strata, including moderate (25–65% and 30–55%), extreme (<25% or >65% and <30% or >55%), and genome-wide (global) range. Wilcoxon signed-rank tests were performed to compare the RMSE values across strata. **B**: RMSE heatmaps of the joint effects of GC content and methylation levels on assay performance. Both GC% and dGC% are grouped into five bins. For every assay, the cell values indicate the mean RMSE across the duplicates of Quartet samples (n=8). **C**: Assay-specific fitted curves showing binomial-normalized RMSE (bRMSE) as a function of methylation level. A bRMSE value of 1.0 represents the nominal binomial sampling baseline, whereas the values above 1.0 indicate additional quantitative error beyond that expected from binomial sampling alone within the corresponding methylation range. The sizes and colors of dots represent the absolute quantitative error and the relative contribution of non-binomial error across methylation levels, respectively. Higher fractions indicate greater relative contribution of error beyond depth-dependent binomial sampling.

Fig. 3. Mappability sensitivity of methods and aligners. **A**: The RMSE values of all 16 assays on Quartet samples across GIAB LowMappabilityAll and NotInLowMappabilityAll strata, together with global RMSE values. Wilcoxon signed-rank test were performed to compare the RMSE values across strata. **B**: Assay-specific RMSE profiles across mappability bins using the selected BisMap or UMap track for each assay. **C** and **D**: Sample-level RMSE and RMSE range across mappability bins for reverse- and direct-conversion aligners, respectively. Selected BisMap/UMap tracks are filled with color, while the others are represented by hollow circles. Aligners are sorted by global RMSE from top to bottom.

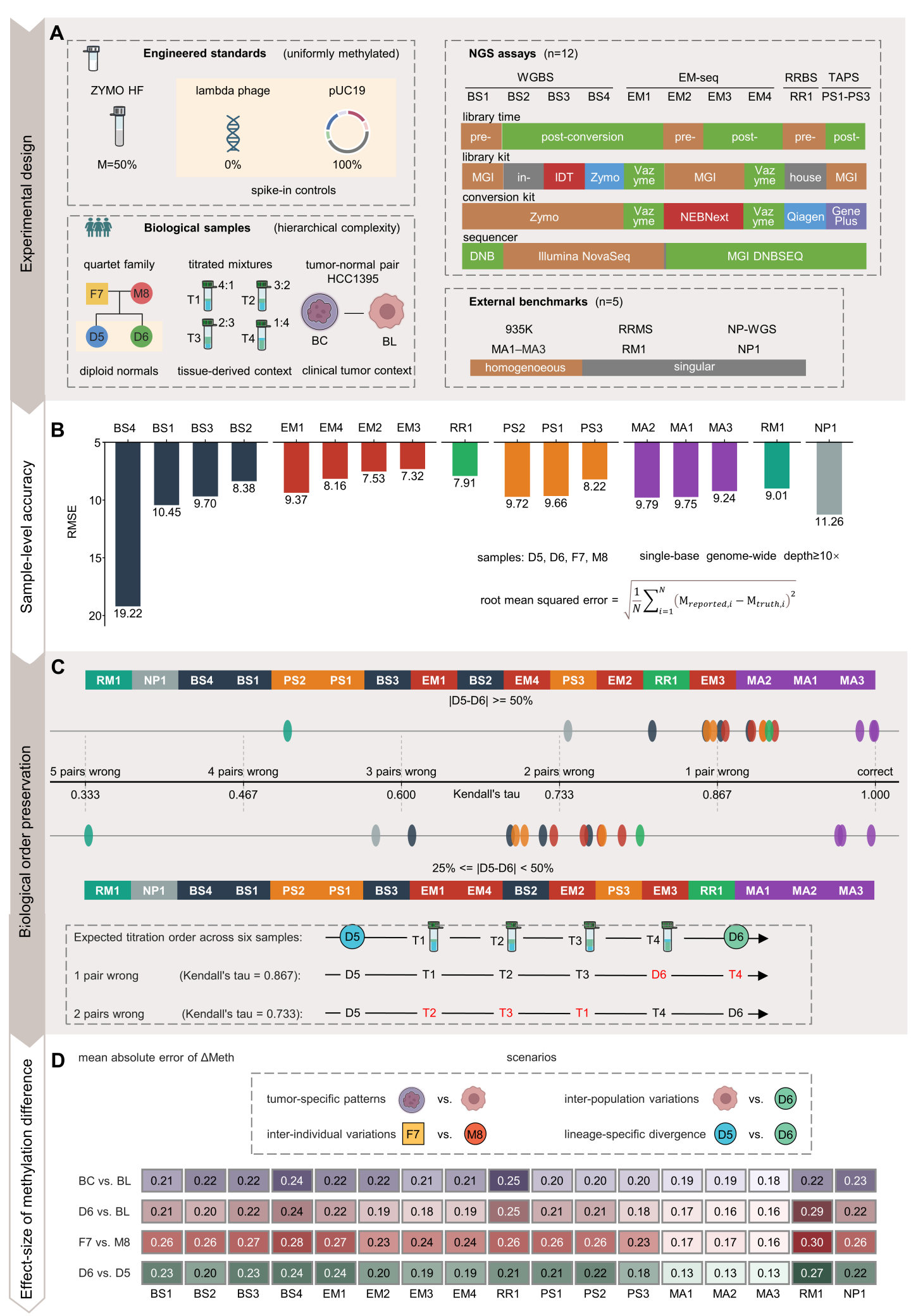
Fig. 4. Localized difficulties in the HCC1395 tumor genome. **A**: The distribution of copy number and ploidy bins of the BC reference sample, inferred from the Oxford Nanopore whole genome sequencing data. **B**: Critical Difference (CD) diagram of the Nemenyi test showing statistical comparison of all ploidy bins for non-targeted NGS data. The same letter indicates the same Nemenyi group. Shared letters mean no significant separation. A lower mean rank and assignment to group “a” both denote greater quantitative difficulty. **C**: Heatmap of BC-BL delta RMSE across ploidy bins, in which the high values indicate increased errors. **D**: Distribution and decision tree for classifying GIAB strata using WGBS, EM-seq, and TAPS assays. For each assay and sample, stratum-localized RMSE and sample-level global RMSE were averaged across technical replicates. Common difficulty was defined as the median regional penalty across the five background samples (BL, D5, D6, F7, and M8). The BC-extra penalty was calculated by subtracting the median common difficulty from the localized penalty

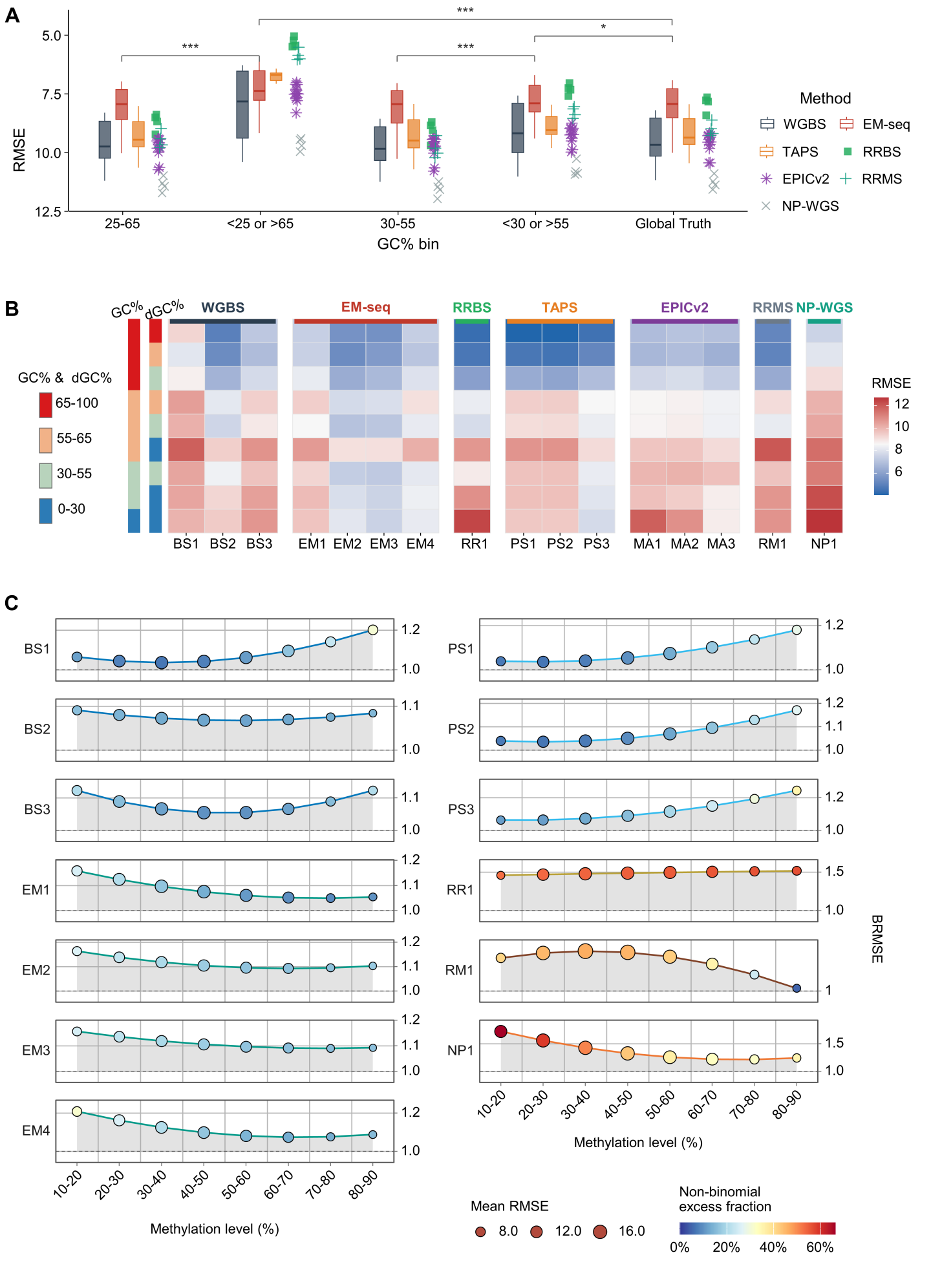
in BC (HCC1395) sample. The BC-extra penalty quantified the additional localized error relative to the non-tumor background after counting for sample-level RMSE baselines. For every stratum, a method "passed" when ≥ 2 assays exceeded the practical cutoff of 2.33 RMSE units. Penalty was calculated by subtracting sample-level global RMSE from the RMSE within the evaluated stratum, after averaging the technical replicates. The common difficulty was defined as the median regional penalty across the five background samples, including BL, D5, D6, F7, and M8. A method passed the criterion when at least two assays exceeded the 2.33 RMSE cutoff.

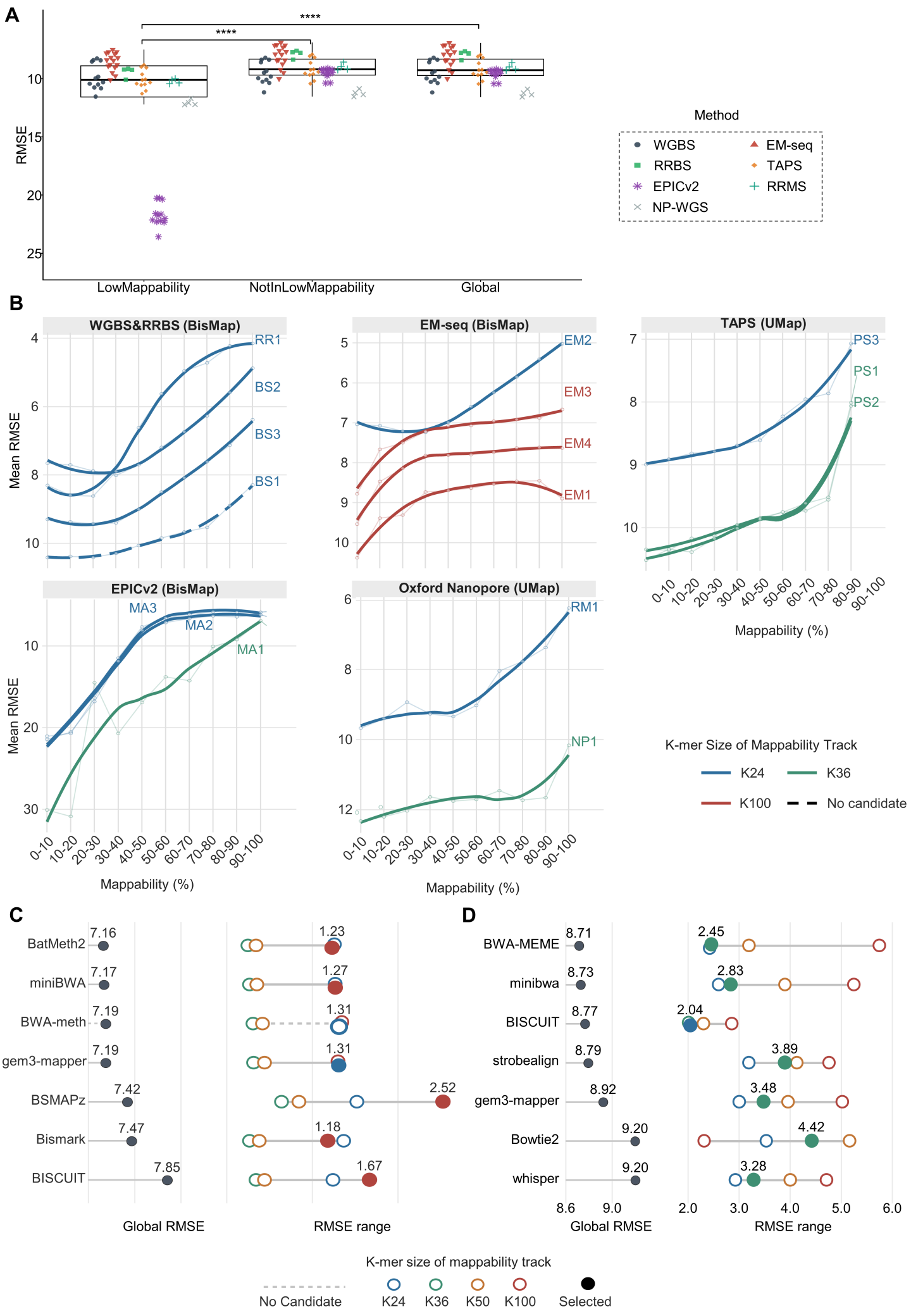
Fig. 5. Quantitative accuracy of different benchmarking scopes and conventional data quality metrics. **A**: RMSE across BLUEPRINT-covered, EPICv2-targeted, and genome-wide truth-set regions. **B**: Correlation between RMSE and quantitative reproducibility (Spearman's rho) across scopes. **C**: Conversion efficiency measured by mitochondrial DNA, spiked-in lambda phage DNA and pUC19 plasmid DNA, and quantitative bias at cytosines with intermediate methylation levels from ZYMO standards.

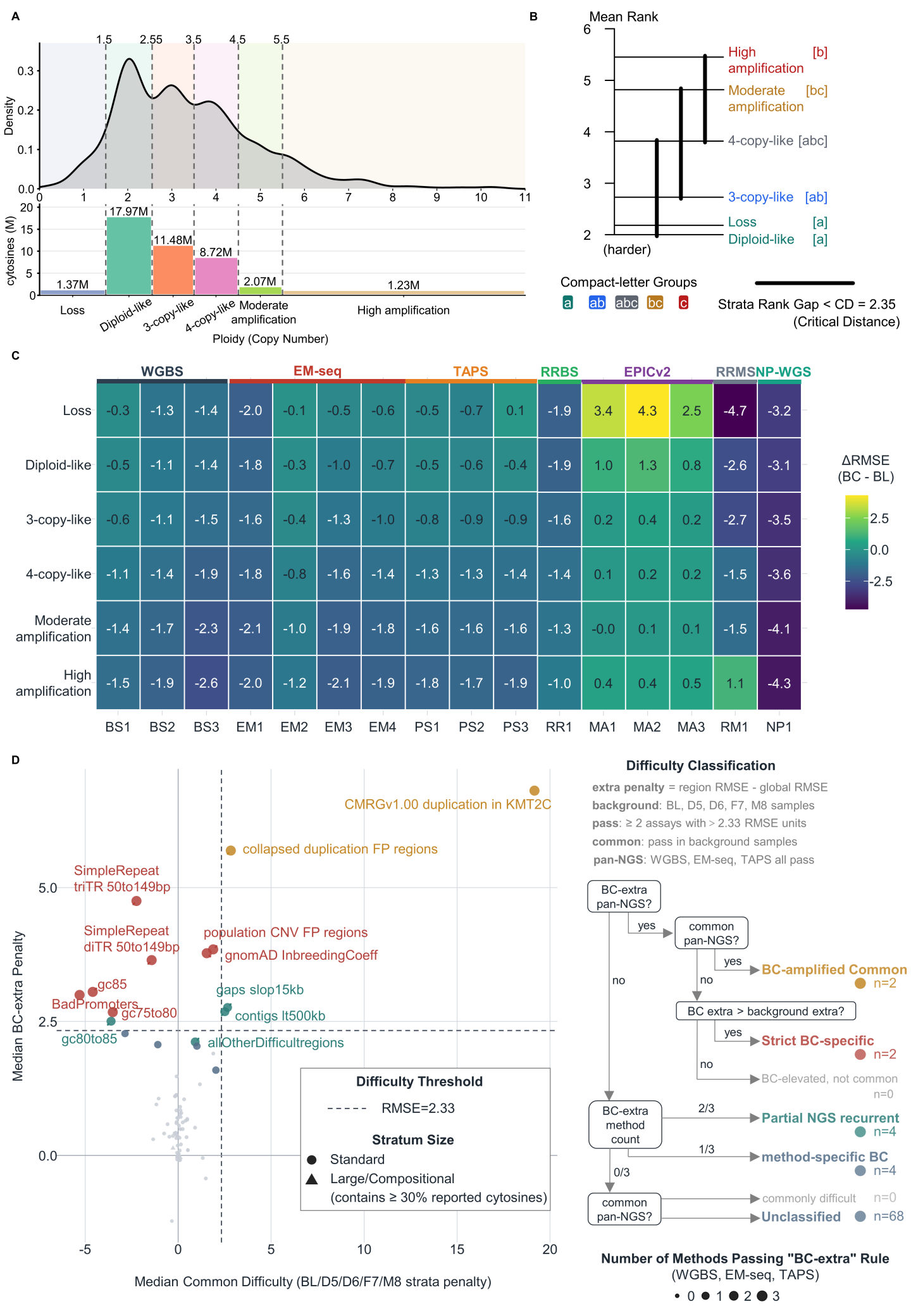
Fig. 6. Stratification of CpGs by MethRidgeQC and its comparsion with depth filtering. **A**: MethRidgeQC assigns CpG cytosine measurements to four reliability classes on D6_1 libraries (the first technical replicate of D6 sample), including high (H), moderately high (MH), moderately low (ML), and low (L). Left panel shows the proportions of four classes. Their quantitative accuracy (RMSE) values were annotated in the blocks. Right panel shows the distribution of the RMSE values of the four classes and two combined groups. Dashed vertical lines mark the corresponding sample-level

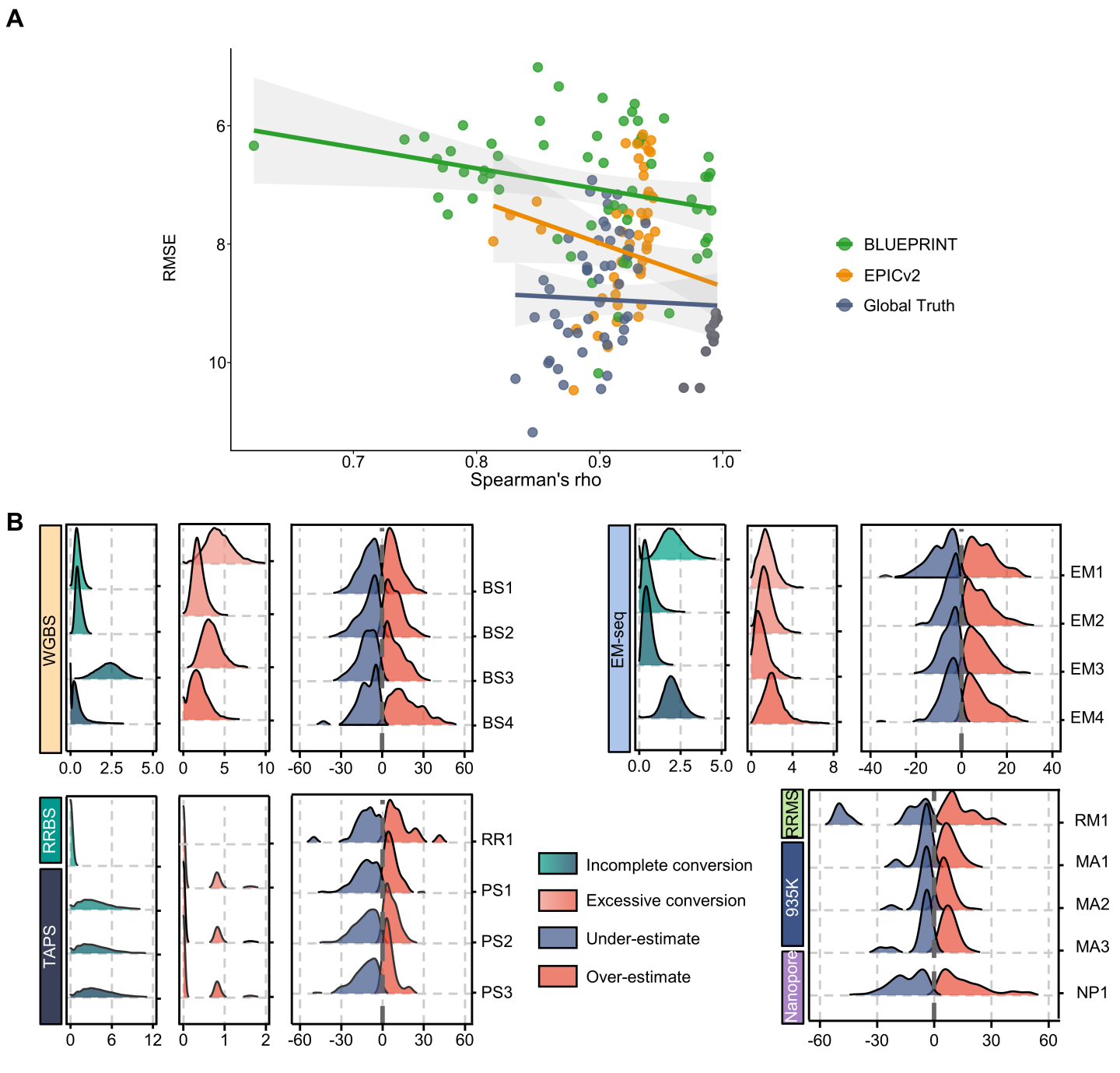
RMSE for each assay. **B**: Relative RMSE reduction obtained by retaining H+MH measurements compared with filtering solely by depth of ≥5× or ≥10×. Results are shown separately for each assay across D5, D6, F7, M8, BC, and BL samples. C: CpG cytosine counts and class-specific RMSE values within the 5–10× and ≥10× depth strata on D6_1 sample. Patterned bars show cytosine counts on a log10 scale, with patterns identifying reliability classes. Circles below the shared baseline show the corresponding RMSE, with larger values plotted farther downward and fill color encoding RMSE.

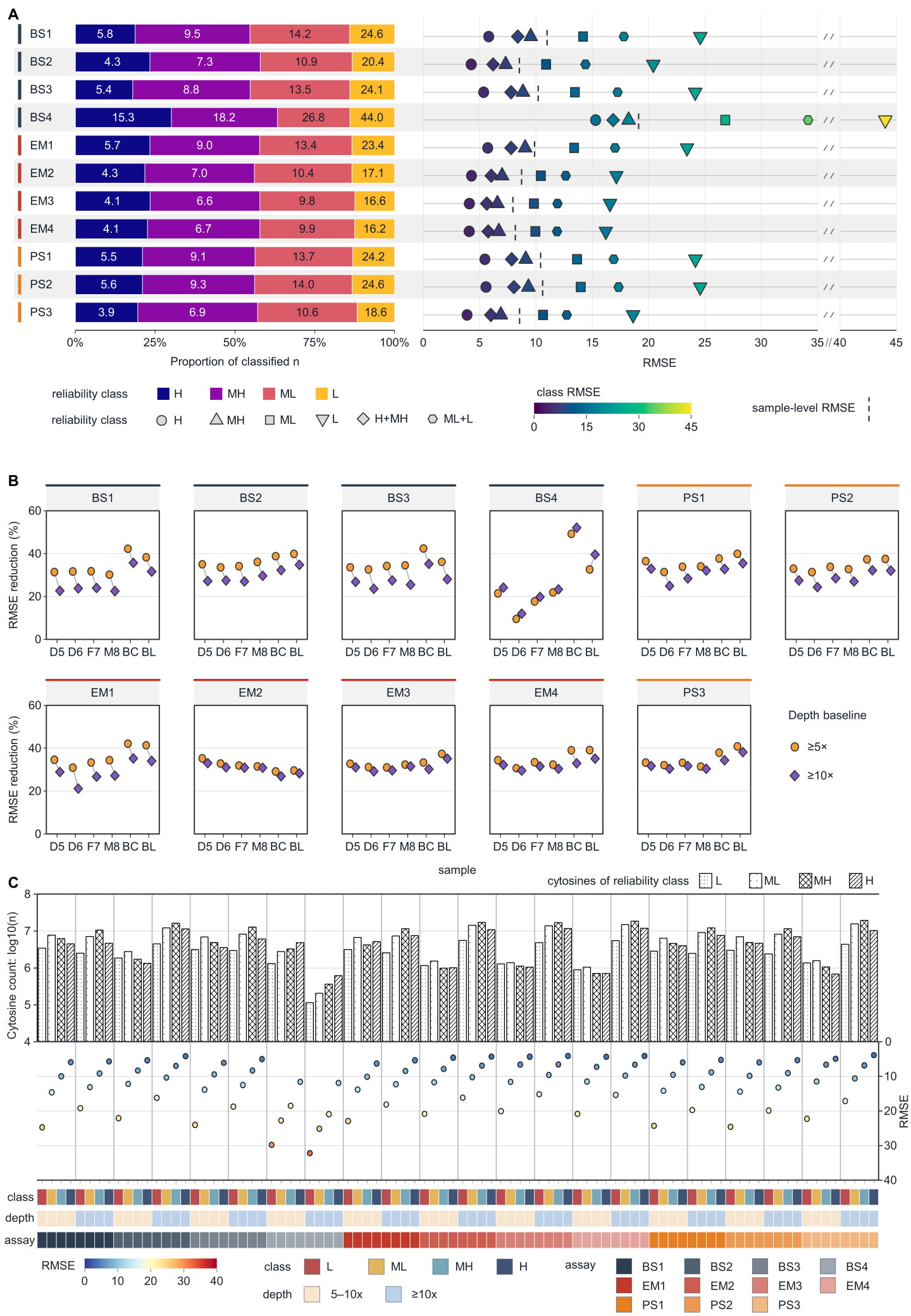

Experimental design
A
Engineered standards (uniformly methylated)
ZYMO HF
M=50%
lambda phage
0%
pUC19
100%
spike-in controls
Biological samples (hierarchical complexity)
quartet family
F7
M8
D5
D6
diploid normals
titrated mixtures
T1 4:1
T2 3:2
T3 2:3
T4 1:4
tissue-derived context
tumor-normal pair HCC1395
BC
BL
clinical tumor context
NGS assays (n=12)
WGBS
EM-seq
RRBS
TAPS
BS1 BS2 BS3 BS4 EM1 EM2 EM3 EM4 RR1 PS1-PS3
library time
pre- post-conversion pre- post- pre- post-
library kit
MGI in- IDT Zymo Vazyme MGI Vazyme house MGI
conversion kit
Zymo Vazyme NEBNext Vazyme Qiagen Gene Plus
sequencer
DNB Illumina NovaSeq MGI DNBSEQ
External benchmarks (n=5)
935K MA1–MA3
RRMS RM1
NP-WGS NP1
homogeneoeus
singular
Sample-level accuracy
B
BS4 BS1 BS3 BS2 EM1 EM4 EM2 EM3 RR1 PS2 PS1 PS3 MA2 MA1 MA3 RM1 NP1
RMSE
5
10
15
20
19.22
10.45
9.70
8.38
9.37
8.16
7.53
7.32
7.91
9.72
9.66
8.22
9.79
9.75
9.24
9.01
11.26
samples: D5, D6, F7, M8
single-base genome-wide depth≥10×
root mean squared error = $\sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(M_{reported,i} - M_{truth,i}\right)^2}$
Biological order preservation
C
RM1 NP1 BS4 BS1 PS2 PS1 BS3 EM1 BS2 EM4 PS3 EM2 RR1 EM3 MA2 MA1 MA3
|D5-D6| >= 50%
5 pairs wrong
4 pairs wrong
3 pairs wrong
2 pairs wrong
1 pair wrong
correct
0.333
0.467
0.600
Kendall's tau
0.733
0.867
1.000
25% <= |D5-D6| < 50%
RM1 NP1 BS4 BS1 PS2 PS1 BS3 EM1 EM4 BS2 EM2 PS3 EM3 RR1 MA1 MA2 MA3
Expected titration order across six samples: D5 T1 T2 T3 T4 D6
1 pair wrong (Kendall's tau = 0.867): D5 T1 T2 T3 D6 T4
2 pairs wrong (Kendall's tau = 0.733): D5 T2 T3 T1 T4 D6
Effect-size of methylation difference
D
mean absolute error of ΔMeth
scenarios
tumor-specific patterns vs.
inter-individual variations F7 vs. M8
inter-population variations vs. D6
lineage-specific divergence D5 vs. D6
BC vs. BL 0.21 0.22 0.22 0.24 0.22 0.22 0.21 0.21 0.25 0.20 0.20 0.20 0.19 0.19 0.18 0.22 0.23
D6 vs. BL 0.21 0.20 0.22 0.24 0.22 0.19 0.18 0.19 0.25 0.21 0.21 0.18 0.17 0.16 0.16 0.29 0.22
F7 vs. M8 0.26 0.26 0.27 0.28 0.27 0.23 0.24 0.24 0.26 0.26 0.26 0.23 0.17 0.17 0.16 0.30 0.26
D6 vs. D5 0.23 0.20 0.23 0.24 0.24 0.20 0.19 0.19 0.21 0.21 0.22 0.18 0.13 0.13 0.13 0.27 0.22
BS1 BS2 BS3 BS4 EM1 EM2 EM3 EM4 RR1 PS1 PS2 PS3 MA1 MA2 MA3 RM1 NP1


A
RMSE
5.0
7.5
10.0
12.5
***
***
***
*
25-65
<25 or >65
30-55
<30 or >55
Global Truth
GC% bin
Method
WGBS
EM-seq
TAPS
RRBS
EPICv2
RRMS
NP-WGS
B
GC%
dGC%
WGBS
EM-seq
RRBS
TAPS
EPICv2
RRMS
NP-WGS
GC% & dGC%
65-100
55-65
30-55
0-30
BS1
BS2
BS3
EM1
EM2
EM3
EM4
RR1
PS1
PS2
PS3
MA1
MA2
MA3
RM1
NP1
RMSE
12
10
8
6
C
BS1
BS2
BS3
EM1
EM2
EM3
EM4
PS1
PS2
PS3
RR1
RM1
NP1
BRMSE
10-20
20-30
30-40
40-50
50-60
60-70
70-80
80-90
Methylation level (%)
Mean RMSE
8.0
12.0
16.0
Non-binomial excess fraction
0%
20%
40%
60%


A
****
****
RMSE
10
15
20
25
LowMappability
NotInLowMappability
Global
Method
WGBS
RRBS
EPICv2
NP-WGS
EM-seq
TAPS
RRMS
B
WGBS&RRBS (BisMap)
EM-seq (BisMap)
TAPS (UMap)
EPICv2 (BisMap)
Oxford Nanopore (UMap)
Mean RMSE
RR1
BS2
BS3
BS1
EM2
EM3
EM4
EM1
PS3
PS1
PS2
MA3
MA2
MA1
RM1
NP1
0-10
10-20
20-30
30-40
40-50
50-60
60-70
70-80
80-90
90-100
Mappability (%)
K-mer Size of Mappability Track
K24
K36
K100
No candidate
C
BatMeth2 7.16
miniBWA 7.17
BWA-meth 7.19
gem3-mapper 7.19
BSMAPz 7.42
Bismark 7.47
BISCUIT 7.85
1.23
1.27
1.31
1.31
2.52
1.18
1.67
Global RMSE
RMSE range
D
BWA-MEME 8.71
minibwa 8.73
BISCUIT 8.77
strobealign 8.79
gem3-mapper 8.92
Bowtie2 9.20
whisper 9.20
2.45
2.83
2.04
3.89
3.48
4.42
3.28
8.6
9.0
2.0
3.0
4.0
5.0
6.0
Global RMSE
RMSE range
K-mer size of mappability track
No Candidate
K24
K36
K50
K100
Selected


A
1.5
2.55
3.5
4.5
5.5
Density
cytosines (M)
17.97M
11.48M
8.72M
2.07M
1.37M
1.23M
Loss
Diploid-like
3-copy-like
4-copy-like
Moderate amplification
High amplification
Ploidy (Copy Number)
B
Mean Rank
High amplification [b]
Moderate amplification [bc]
4-copy-like [abc]
3-copy-like [ab]
Loss [a]
Diploid-like [a]
(harder)
Compact-letter Groups
a ab abc bc c
Strata Rank Gap < CD = 2.35
(Critical Distance)
C
WGBS
EM-seq
TAPS
RRBS
EPICv2
RRMS
NP-WGS
Loss -0.3 -1.3 -1.4 -2.0 -0.1 -0.5 -0.6 -0.5 -0.7 0.1 -1.9 3.4 4.3 2.5 -4.7 -3.2
Diploid-like -0.5 -1.1 -1.4 -1.8 -0.3 -1.0 -0.7 -0.5 -0.6 -0.4 -1.9 1.0 1.3 0.8 -2.6 -3.1
3-copy-like -0.6 -1.1 -1.5 -1.6 -0.4 -1.3 -1.0 -0.8 -0.9 -0.9 -1.6 0.2 0.4 0.2 -2.7 -3.5
4-copy-like -1.1 -1.4 -1.9 -1.8 -0.8 -1.6 -1.4 -1.3 -1.3 -1.4 -1.4 0.1 0.2 0.2 -1.5 -3.6
Moderate amplification -1.4 -1.7 -2.3 -2.1 -1.0 -1.9 -1.8 -1.6 -1.6 -1.6 -1.3 -0.0 0.1 0.1 -1.5 -4.1
High amplification -1.5 -1.9 -2.6 -2.0 -1.2 -2.1 -1.9 -1.8 -1.7 -1.9 -1.0 0.4 0.4 0.5 1.1 -4.3
BS1 BS2 BS3 EM1 EM2 EM3 EM4 PS1 PS2 PS3 RR1 MA1 MA2 MA3 RM1 NP1
ΔRMSE (BC - BL)
2.5
0.0
-2.5
D
CMRGv1.00 duplication in KMT2C
collapsed duplication FP regions
SimpleRepeat triTR 50to149bp
SimpleRepeat diTR 50to149bp
population CNV FP regions
gnomAD InbreedingCoeff
gc85
BadPromoters
gc75to80
gc80to85
gaps slop15kb
contigs lt500kb
allOtherDifficultregions
Median BC-extra Penalty
Median Common Difficulty (BL/D5/D6/F7/M8 strata penalty)
Difficulty Threshold
RMSE=2.33
Stratum Size
Standard
Large/Compositional
(contains ≥ 30% reported cytosines)
Difficulty Classification
extra penalty = region RMSE - global RMSE
background: BL, D5, D6, F7, M8 samples
pass: ≥ 2 assays with > 2.33 RMSE units
common: pass in background samples
pan-NGS: WGBS, EM-seq, TAPS all pass
BC-extra pan-NGS?
yes
common pan-NGS?
BC-amplified Common n=2
BC extra > background extra?
Strict BC-specific n=2
BC-elevated, not common n=0
BC-extra method count
2/3
Partial NGS recurrent n=4
1/3
method-specific BC n=4
0/3
commonly difficult n=0
Unclassified n=68
Number of Methods Passing "BC-extra" Rule
(WGBS, EM-seq, TAPS)
0 1 2 3


A
RMSE
6
8
10
0.7
0.8
0.9
1.0
Spearman's rho
BLUEPRINT
EPICv2
Global Truth
B
WGBS
BS1
BS2
BS3
BS4
EM-seq
EM1
EM2
EM3
EM4
RRBS
TAPS
RR1
PS1
PS2
PS3
Incomplete conversion
Excessive conversion
Under-estimate
Over-estimate
RRMS
935K
Nanopore
RM1
MA1
MA2
MA3
NP1


A
BS1 5.8 9.5 14.2 24.6
BS2 4.3 7.3 10.9 20.4
BS3 5.4 8.8 13.5 24.1
BS4 15.3 18.2 26.8 44.0
EM1 5.7 9.0 13.4 23.4
EM2 4.3 7.0 10.4 17.1
EM3 4.1 6.6 9.8 16.6
EM4 4.1 6.7 9.9 16.2
PS1 5.5 9.1 13.7 24.2
PS2 5.6 9.3 14.0 24.6
PS3 3.9 6.9 10.6 18.6
0% 25% 50% 75% 100%
Proportion of classified n
0 5 10 15 20 25 30 35 // 40 45
RMSE
reliability class H MH ML L
reliability class H MH ML L H+MH ML+L
class RMSE
0 15 30 45
sample-level RMSE
B
BS1 BS2 BS3 BS4 PS1 PS2
EM1 EM2 EM3 EM4 PS3
RMSE reduction (%)
0 20 40 60
D5 D6 F7 M8 BC BL
Depth baseline
≥5×
≥10×
sample
C
cytosines of reliability class L ML MH H
Cytosine count: log10(n)
4 5 6 7 8
RMSE
0 10 20 30 40
class
depth
assay
RMSE 0 10 20 30 40
class L ML MH H
depth 5–10x ≥10x
assay BS1 BS2 BS3 BS4 EM1 EM2 EM3 EM4 PS1 PS2 PS3

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementaryData.zip