

Methylation reference datasets from Quartet DNA materials for benchmarking epigenome sequencing

Corresponding Author: Dr Yuanting Zheng

This file contains all reviewer reports in order by version, followed by all author rebuttals in order by version.

Version 0:

Reviewer comments:

Reviewer #1

(Remarks to the Author)

The manuscript entitled "Quantitative methylation reference datasets of Quartet DNA reference materials for benchmarking genome-wide epigenome sequencing" by Guo et al. presents the results of thorough comparisons of whole genome DNA methylation assays across various methodologies and laboratories. The authors utilized quartet samples (father, mother, and their monozygotic twins), generating a total of 108 libraries examined using two different analytical pipelines. This is one of the largest DNA methylation studies analyzing the same samples using massively parallel sequencing.

The authors identified two main points from their analysis: 1) rather than excessively focusing on conversion controls, laboratories should prioritize sequencing depth, and 2) lab-specific variability poses a significant challenge to responsibility in WGBS. The authors proposed a minimum CpG depth of 20x, which optimally balances strand-specific qualitative consistency and quantitative precision.

Despite their tremendous efforts and extensive datasets, the interpretations lean heavily toward purely technical observations, with very little discussion regarding methodological suggestions or tips to improve the quality or reliability of the libraries. Moreover, the key points made by the authors are not particularly novel. Additionally, although it may be outside the scope of their manuscript, there is minimal discussion of biological implications, even though the authors employed a unique sample set (quartet samples) with deep sequencing. By failing to address biological implications or provide useful tips for generating or conducting genome-wide DNA methylation analysis, the manuscript loses significant value.

(Remarks on code availability)

Reviewer #2

(Remarks to the Author)

Guo, Chen, Zhang, Zhang and colleagues present a comprehensive reference dataset for genome-wide DNA methylation profiling. They leverage lymphoblastoid reference cell lines from the "Quartet" consortium as a well-defined source of reference materials and apply the three major sequencing approaches (WGBS, EM-seq, and TAPS) to these materials to generate genome-wide DNA methylation profiles. By including multiple laboratories and technical replicates, they generate a comprehensive, high-quality dataset, which they analyzed meticulously and use to define a consensus reference dataset. They go on to utilize this reference dataset to assess the performance of additional assays, including long-read sequencing technology. Overall, the dataset is impressive, and the analyses are thorough and described in sufficient detail. The integration of reference-dependent and reference-independent metrics and the emphasis on strand-specific methylation, add interesting insights for the field. I therefore believe the work constitutes a valuable contribution and a useful foundation for benchmarking and QC of DNA methylation profiling.

I only have one major concern (which splits into three sub-concerns), which pertains to the availability of the data and materials. For this reference to indeed be useful, it is inherently critical that both of these are readily available to researchers aiming to use the reference.

1. Materials: I have not found any mention in the manuscript about how materials would be available to researchers. I saw on

the website that materials could be ordered, which seems good, but this procedure should be briefly mentioned in the paper and links provided.

2. Data: Data availability statements like “data supporting findings of this study are available from the corresponding authors” are not acceptable for a publication of a dataset. The manuscript mentions several accession numbers for NODE (in the manuscript) and GSA (in the reporting form), but neither are accessible to me. The Quartet Data Portal does not yet contain any downloads for the reference dataset. Other references are nicely accessible via the website, so it’s probably the authors’ intention to add the methylation reference after publication, but in its current state, it is impossible for me as a reviewer to assess whether this will be done and whether the provided data would be appropriate. Third, even if data are added to the Quartet website, the permanence of this resource is unclear to me. As bespoke scientific databases often cease to exist or are no longer maintained after projects run out, it is unclear whether long-term accessibility of the data can be guaranteed. It would therefore be advisable to deposit a copy of the processed reference dataset also in accepted DOI-providing repository such as Figshare or Zenodo (note, I’m not suggesting to upload all raw data – assuming the GSA accession will become available after publication, this would be sufficient).

3. Workflow: Can the authors provide guidance (code?) detailing how exactly researchers would use the reference dataset to assess the performance of new data? The consensus building workflow is rigorous but also complex and the code the authors provide (although generally in a good shape) doesn’t make it clear enough how to proceed.

Minor Comments:

4. I’m still puzzled by the authors’ finding that increasing sequencing depth yields improved quantitative correlations but decreased qualitative concordance (page 7), and that quantitative correlation between batches is high while qualitative overlaps are low (page 8). Can the authors provide additional explanations to the results and discuss the interpretation in the discussion?

5. Some of the figures are very dense and will be difficult to interpret for readers, perhaps the authors can consider removing superfluous information (either completely or move it to the supplementary materials). For instance, Fig. 5d/e show rankings of methods in many different types of genomic regions and by different metrics. I believe most readers will be most interested in which methods do best and the distinction between region types makes fairly little difference, and the ranking by PCC and RMSE seems identical I think. One option to simplify the figure could be to show two waterfall bar plots of the average Recall and the average PCC across all region types, making it straightforward to see the best and worst methods and even to quantitatively compare the metrics (which is impossible with the rescaled values that are only encoded by color at the moment). The current plots could still be shown in the supplement, if the authors believe the complete information is critical.

6. Fig. 2a: Typo “dpeth” instead of “depth” (twice).

7. Fig. 2b: It’s unclear from the figure that the x-axis represents sequencing depth. Add a label and tick marks.

8. Fig. 5d: Misalignment between letter annotations “E”, “T”, “E”, etc. and the rows of the plots.

(Remarks on code availability)

I had a look at the repository shared via the private link to Code Ocean. The folder structure is clear and it’s straightforward to discover the code chunks relevant to different figures. A README file provides instructions for execution. Code for “upstream analysis”, which refers to time-consuming execution of raw data processing pipelines, is also provided. I had a quick look at a few code files and the code looks valid. I have not executed the code.

Reviewer #3

(Remarks to the Author)

While the manuscript presents a useful effort and dataset there are important issues with the main conclusions drawn, that limit the interpretation of results:

1. The authors state: “Notably, all replicates exhibited substantial inter-strand methylation differences (absolute delta methylation $\geq 10\%$ at $1\times$) across protocols (Fig. 2B), challenging conventional strand-merging practices”

My interpretation of Figure 2A is that the observed “strand bias” is likely just sampling noise. Fig2A shows that the absolute methylation deviation between strands decreases with increasing coverage (as would be expected) and is still decreasing at their cutoff coverage. This argues against true biological differences between the strands, and hence suggests that we should in fact pool +ve and -ve strands, as opposed to treating them separately as the authors suggest.

2. The authors state: “While increasing sequencing depth reduced qualitative concordance (Jaccard index), it improved quantitative agreement (Pearson Correlation Coefficient (PCC) $\geq 0.9\dots$)” and “Technical replicates within labs showed low reproducibility (Jaccard index, Fig. S4)”

It seems very counterintuitive that qualitative concordance would go down when quantitative concordance goes up. Is the Jaccard index being used to see if the same methylation values are reported across strand? If so, increasing coverage would clearly give more accurate values, but also make it less likely that they are numerically identical (e.g. 0.222 is not equal to 0.223). If instead, Jaccard is being used to assess coverage concordance (which makes more sense) this should be clearly explained.

(Remarks on code availability)

Version 1:

Reviewer comments:

Reviewer #1

(Remarks to the Author)

Thank you for clarifying the scope of the study and adding biological interpretations. The observation that monozygotic twins had a larger difference in PC1 compared to their parents is unexpected, but it seems very interesting. Additionally, the revised abstract helped me understand the primary objective of this project.

I have a few suggestions regarding the data visualizations. The authors used colors like cyan, pale teal, orange-yellow, yellow, and red-orange to represent different samples; however, similar colors are used for different purposes in the figures, such as pipelines in Fig. 3, CGI in Fig. 5, and Type in Fig. 6. I suggest that the authors avoid reusing colors to help reduce confusion. Fig. 1 uses different color codes for the samples. Additionally, the line width does not match the figure size in some figures (Fig. 4C and 4D); therefore, the lines overlap and are not visible.

(Remarks on code availability)

Reviewer #2

(Remarks to the Author)

The authors have addressed all my concerns.

Thank you for clarifying the apparent discrepancy between the qualitative / quantitative analyses. I think the sentences the authors added to the manuscript will avoid misunderstandings by future readers.

Very minor remark: the software checklist states that the software was available under the GNU General Public License v3, while the LICENSE file in the CodeOcean repository refers to the MIT License. Both should be synchronized.

(Remarks on code availability)

I have briefly had a look at the files provided on CodeOcean, which seems well-structured and understandable. The authors have made a very good effort to provide reproducible code.

Reviewer #3

(Remarks to the Author)

The authors have addressed my concerns from my original review, and I believe it significantly improves the manuscript.

(Remarks on code availability)

The code is well documented and readable.

Open Access This Peer Review File is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.

In cases where reviewers are anonymous, credit should be given to 'Anonymous Referee' and the source.

The images or other third party material in this Peer Review File are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.

To view a copy of this license, visit <https://creativecommons.org/licenses/by/4.0/>

Point-by-point response to the referees' comments

Reviewer #1 (Remarks to the Author):

The manuscript entitled “Quantitative methylation reference datasets of Quartet DNA reference materials for benchmarking genome-wide epigenome sequencing” by Guo et al. presents the results of thorough comparisons of whole genome DNA methylation assays across various methodologies and laboratories. The authors utilized quartet samples (father, mother, and their monozygotic twins), generating a total of 108 libraries examined using two different analytical pipelines. This is one of the largest DNA methylation studies analyzing the same samples using massively parallel sequencing.

The authors identified two main points from their analysis: 1) rather than excessively focusing on conversion controls, laboratories should prioritize sequencing depth, and 2) lab-specific variability poses a significant challenge to responsibility in WGBS. The authors proposed a minimum CpG depth of 20×, which optimally balances strand-specific qualitative consistency and quantitative precision.

Despite their tremendous efforts and extensive datasets, the interpretations lean heavily toward purely technical observations, with very little discussion regarding methodological suggestions or tips to improve the quality or reliability of the libraries. Moreover, the key points made by the authors are not particularly novel. Additionally, although it may be outside the scope of their manuscript, there is minimal discussion of biological implications, even though the authors employed a unique sample set (quartet samples) with deep sequencing. By failing to address biological implications or provide useful tips for generating or conducting genome-wide DNA methylation analysis, the manuscript loses significant value.

Reply:

We sincerely thank the reviewer for recognizing the scale and rigor of our study. We understand the reviewer's perspective that the interpretations lean heavily toward

technical observations and that there is minimal discussion of extensive methodological tips for experimental procedures or biological implications. We agree that these aspects are crucial for advancing the field and are often expected in methylation studies.

We agree with the reviewer's suggestion and have made initial efforts to provide methodological recommendations based on the current dataset. These results serve as a simplified example of how reference datasets can be used for benchmarking analysis pipelines. We have added a brief discussion section (page 12, lines 300–307) demonstrating an example of how this reference dataset can be used to compare different analysis pipelines, which partially addresses the need for illustrating the practical utility of our resource for methodological assessment.

"Using reference-dependent metrics, we compared two distinct analysis pipelines across 11 experimental batches. In over half of the batches, both pipelines showed no significant differences (Wilcoxon test) in performance metrics (**Fig. S13**). However, for WGBS and EM-seq data, Pipeline P2 (based on BWA-METH) outperformed P1 (based on Bismark) with higher Recall and PCC and lower RMSE (Batch A2 and A7). For TAPS, ONT, and PacBio data, pipelines exhibited minimal differences, potentially due to TAPS pipeline similarity and smaller batch sizes in ONT/PacBio (4 libraries per batch versus 12 in others). This pilot test demonstrates our reference datasets' utility as a ground-truth for evaluating analysis pipelines or library performance, enabling field-wide methodological calibration and improvement."

Our work consists in providing the necessary tools (the reference materials, reference dataset, and metrics) for the community to rigorously benchmark their own methods and pipelines, which in turn can inform future methodological improvements and best practices. In parallel, we are conducting a follow-up study involving a comprehensive inter-laboratory proficiency test using Quartet DNA reference materials. Our next study aims to identify key technical factors that influence DNA methylation measurements and to develop best-practice recommendations for generating or conducting genome-wide DNA methylation analysis.

We agree that some of the key points reported in our study—such as the influence of mean CpG depth and inter-lab variability—align with prior work, our study provides

independent validation across diverse platforms and reveals novel technical nuances. By leveraging well-characterized reference materials and a rigorously designed benchmarking framework, our study serves to independently validate earlier conclusions and reinforce their generalizability across platforms and laboratories. Importantly, our study also identifies previously underappreciated technical issues, such as strand-specific biases and inconsistencies in spike-in conversion efficiency, which have received limited attention in the literature. These novel observations not only deepen our understanding of library construction variability but also provide actionable insights for improving future experimental and analytical practices.

We also fully appreciate the reviewer's suggestion and agree that biological insights derived from such a unique sample set are of broad interest to the field, although they fall outside the defined scope and central aim of this resource-focused study. In response, we have added a brief biological observation in the revised manuscript (page 7, lines 155–163):

"We noticed that the two LCLs derived from the monozygotic twin daughters (D5 and D6) exhibited consistently large differences in the PC1 space across eight out of nine batches, although one might expect the methylation profiles of identical twins to show the highest similarity among all six pairs of the Quartet sample groups (e.g. F7 vs. D6, M8 vs D6, etc.). This twin-specific separation was already observed in our earlier transcriptomic analyses³⁷, and is now recapitulated in the methylome, indicating that the divergence reflects a genuine biological signal rather than technical noise. The consistency of these observations across both transcriptome and methylome studies further supports the stability and utility of the Quartet reference materials for investigating biological variation across multi-omics layers."

Additionally, we have included an example differential methylation CpG (DMC) analysis in the revised manuscript (page 10, lines 219–230):

" Using the strand-specific, absolute quantitative methylation data we generated, we demonstrated how researchers can derive differential methylation profiles tailored to their study objectives through straightforward data processing. Specifically, we applied a filtering strategy that excludes discordant CpG sites (with strand differences $\geq 20\%$),

merged concordant sites from both strands, and treated strand-only sites as concordant for differential methylation CpG analysis. Setting a threshold of Δ methylation% $\geq 20\%$, we observed that the number of DMCs was the highest between D5 and F7 and the lowest between D6 and M8, consistent with the PCA results (**Fig. S10**). Due to differential B cell subtype selection and cell culture effects, the largest differences were not necessarily between parents, nor were the smallest differences always between the monozygotic twins — a pattern also consistent with our previous transcriptomic findings. This example highlights not only the biological consistency of the Quartet reference materials across omics layers but also their value for cross-omics benchmarking and biological studies."

While a comprehensive exploration of these biological implications falls beyond the scope of this resource-focused study, we believe that our findings offer a valuable foundation for future biological investigations using this set of reference materials.

In summary, we would like to respectfully clarify that the primary objective and contributions of our current manuscript are distinct. Our aim is not to provide a comprehensive guide on executing methylation experiments or specific analysis workflows or to report novel biological findings. Instead, our central goal is to establish and provide a robust, well-characterized set of reference materials and a corresponding reference dataset (ground-truth) derived from a unique and deeply sequenced sample set. This resource is specifically designed to serve as a standard benchmark for evaluating the performance, accuracy, and quality of various DNA methylation assays and bioinformatics analysis pipelines using a set of objective and well-defined metrics.

As stated in the introduction of ISO 33405:2024, "The production of reference materials (RMs) is a key activity for the improvement and maintenance of a worldwide coherent measurement system. As detailed in ISO 33403, RMs with different characteristics are used in measurements, such as calibration, quality control, proficiency testing, and method validation, as well as for the assignment of values to other materials." As a part of the Multi-omics Quartet Project, our team has long recognized the importance of developing and applying reference materials, reference datasets, and quality control metrics. To this end, the Quartet Project has been dedicated

to enhancing the reliability, reproducibility, and integration of multi-omics data (across genomics, transcriptomics, proteomics, and metabolomics). These efforts have been published as part of a Nature Collection (<https://www.nature.com/collections/ibadeahigd>), highlighting their foundational role in omics benchmarking.

We believe that the novelty and significant value of our manuscript lie precisely in the creation and rigorous validation of this critical reference resource and benchmarking framework for methylation. In a field with diverse technologies and evolving analytical methods, a standardized ground-truth is indispensable for objective comparisons, method development, and quality control, which ultimately contributes to the reliability and reproducibility of methylation research.

We hope this study, together with ongoing efforts in the Quartet Project, will provide valuable infrastructure for the epigenomics community, supporting method development, quality assurance, and cross-platform integration. We greatly appreciate the reviewer's constructive feedback, which has helped us clarify the scope and contributions of our work.

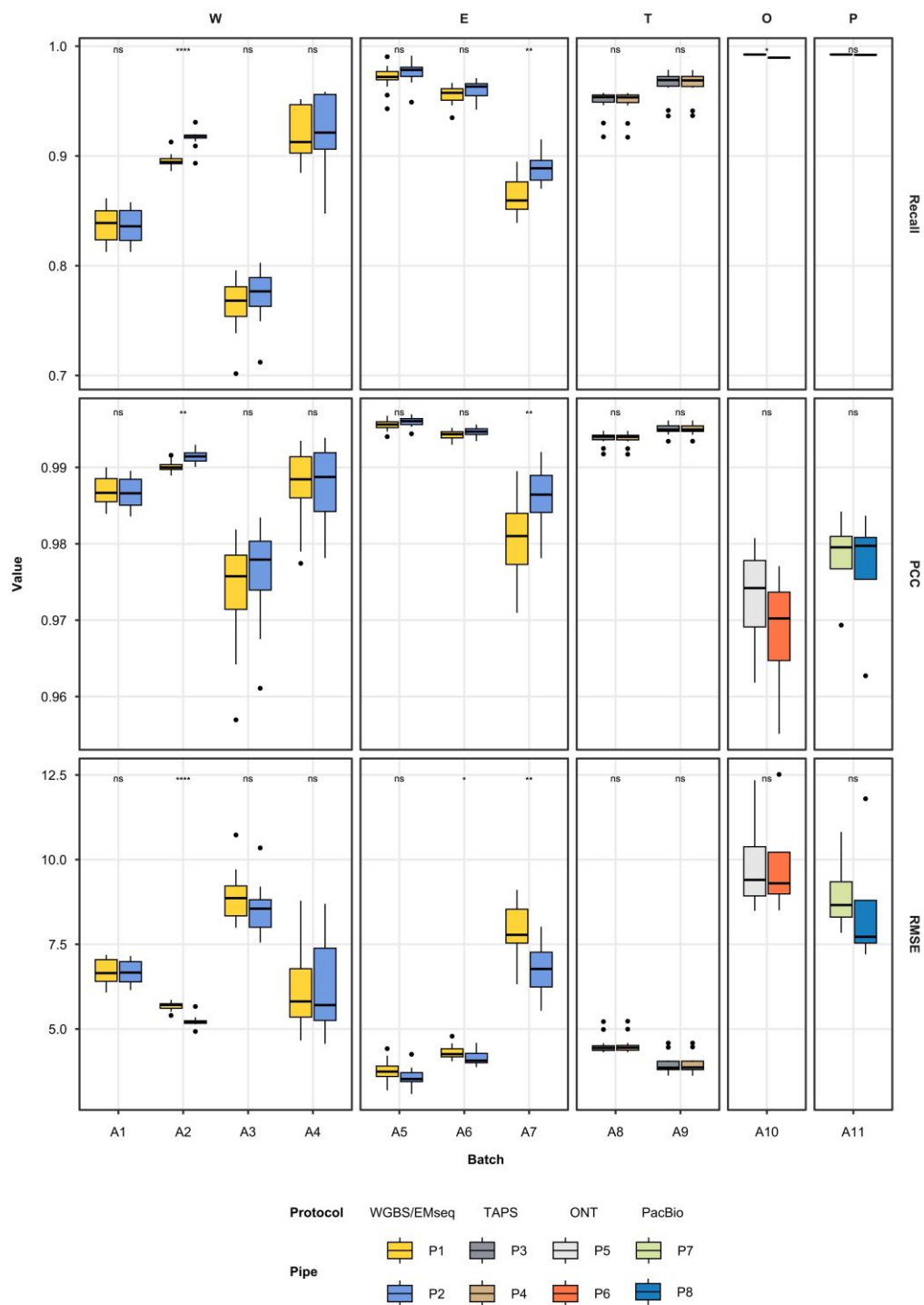


Figure S13. Performance assessment based on reference datasets across different regions.

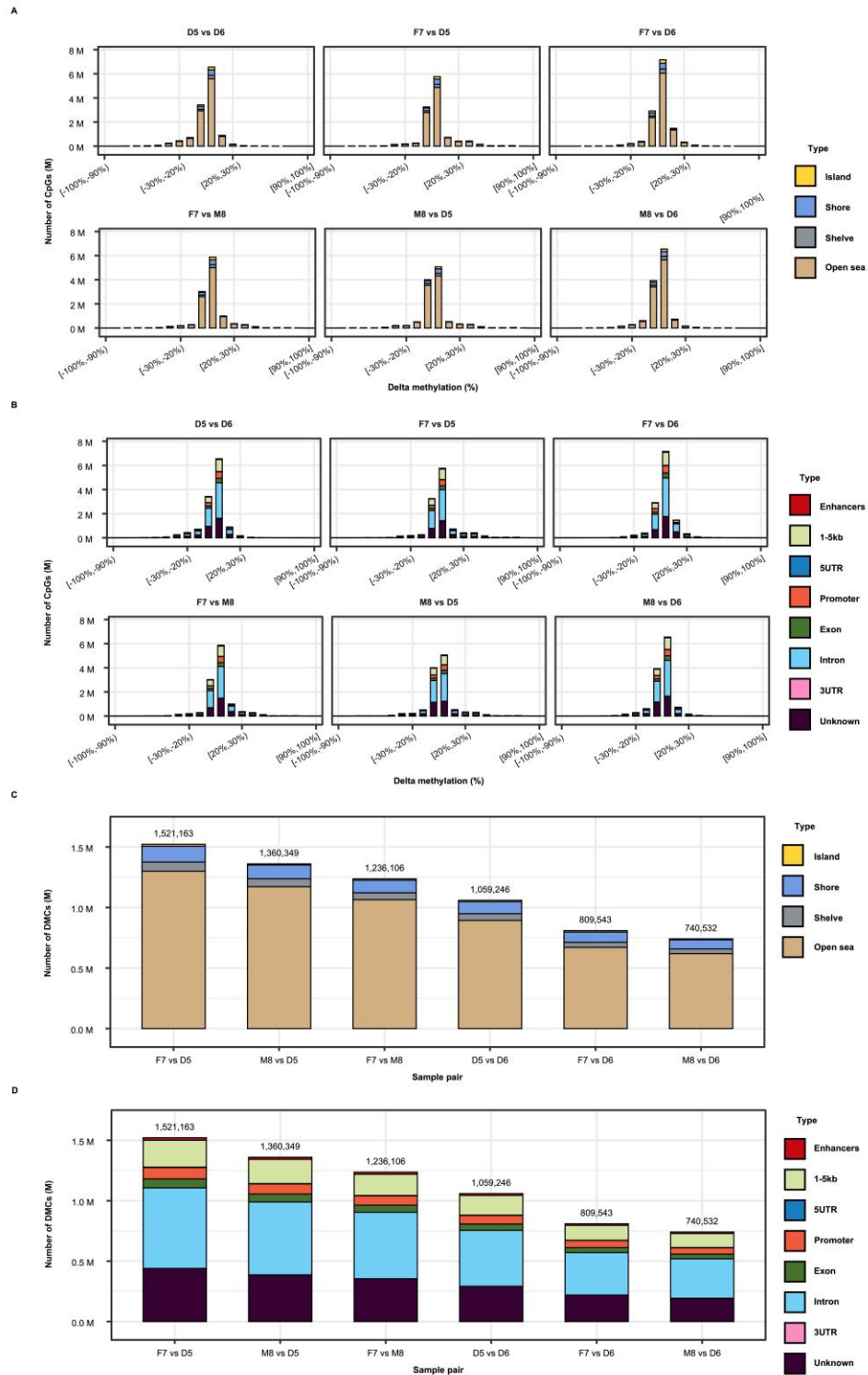


Figure S10. Annotation distribution across different sample pairs based on reference datasets.

Reviewer #2 (Remarks to the Author):

Guo, Chen, Zhang, Zhang and colleagues present a comprehensive reference dataset for genome-wide DNA methylation profiling. They leverage lymphoblastoid reference cell lines from the “Quartet” consortium as a well-defined source of reference materials and apply the three major sequencing approaches (WGBS, EM-seq, and TAPS) to these materials to generate genome-wide DNA methylation profiles. By including multiple laboratories and technical replicates, they generate a comprehensive, high-quality dataset, which they analyzed meticulously and use to define a consensus reference dataset. They go on to utilize this reference dataset to assess the performance of additional assays, including long-read sequencing technology. Overall, the dataset is impressive, and the analyses are thorough and described in sufficient detail. The integration of reference-dependent and reference-independent metrics and the emphasis on strand-specific methylation, add interesting insights for the field. I therefore believe the work constitutes a valuable contribution and a useful foundation for benchmarking and QC of DNA methylation profiling.

Reply:

We sincerely appreciate the reviewer’s positive evaluation and insightful feedback.

I only have one major concern (which splits into three sub-concerns), which pertains to the availability of the data and materials. For this reference to indeed be useful, it is inherently critical that both of these are readily available to researchers aiming to use the reference.

1. Materials: I have not found any mention in the manuscript about how materials would be available to researchers. I saw on the website that materials could be ordered, which seems good, but this procedure should be briefly mentioned in the paper and links provided.

Reply:

We are truly thankful for the constructive comments. We have updated the Materials section to explicitly state that all reference materials are available for order through our

institutional repository, with access details and ordering links provided as follows (page 27; lines 506-509):

“Quartet DNA reference materials can be requested from the Quartet Data Portal (<https://chinese-quartet.org/>) or through the National Sharing Platform for Reference Materials (<https://www.ncrm.org.cn/Web/Home/EnglishIndex>) under the Administrative Regulations of the People’s Republic of China on Human Genetic Resources.”

2. Data: Data availability statements like “data supporting findings of this study are available from the corresponding authors” are not acceptable for a publication of a dataset. The manuscript mentions several accession numbers for NODE (in the manuscript) and GSA (in the reporting form), but neither are accessible to me. The Quartet Data Portal does not yet contain any downloads for the reference dataset. Other references are nicely accessible via the website, so it’s probably the authors’ intention to add the methylation reference after publication, but in its current state, it is impossible for me as a reviewer to assess whether this will be done and whether the provided data would be appropriate. Third, even if data are added to the Quartet website, the permanence of this resource is unclear to me. As bespoke scientific databases often cease to exist or are no longer maintained after projects run out, it is unclear whether long-term accessibility of the data can be guaranteed. It would therefore be advisable to deposit a copy of the processed reference dataset also in accepted DOI-providing repository such as Figshare or Zenodo (note, I’m not suggesting to upload all raw data – assuming the GSA accession will become available after publication, this would be sufficient).

Reply:

We sincerely appreciate the valuable comments offered by the reviewer. We have removed the non-specific data availability statement and now provide explicit access details for all datasets (page 33; lines 680-685):

“The raw sequence data used in this paper have been deposited in the Genome Sequence Archive (GSA) (<https://ngdc.cncb.ac.cn/gsa-human/s/zlvSm7V1>). All reference datasets can be accessed in Figshare (DOI:

<https://doi.org/10.6084/m9.figshare.29481713.v1>) or the Quartet Data Portal (<https://reference-datasets.chinese-quartet.org/index.html?prefix=develop/Methylation/v20250409/>) under the Administrative Regulations of the People's Republic of China on Human Genetic Resources.”

We have confirmed that all the aforementioned links are currently accessible, and the official GSA links and NODE links will be made public after the article is accepted in principle.

3. Workflow: Can the authors provide guidance (code?) detailing how exactly researchers would use the reference dataset to assess the performance of new data? The consensus building workflow is rigorous but also complex and the code the authors provide (although generally in a good shape) doesn't make it clear enough how to proceed.

Reply:

We thank the reviewer for this constructive suggestion. In response, we have deposited a fully documented, executable workflow on Code Ocean, including a detailed README file and step-by-step implementation scripts, enabling researchers to directly apply our reference dataset to evaluate the performance of new data (Compute Capsule's provisional DOI: [10.24433/CO.1027879.v1](https://doi.org/10.24433/CO.1027879.v1)).

Minor Comments:

4. I'm still puzzled by the authors' finding that increasing sequencing depth yields improved quantitative correlations but decreased qualitative concordance (page 7), and that quantitative correlation between batches is high while qualitative overlaps are low (page 8). Can the authors provide additional explanations to the results and discuss the interpretation in the discussion?

Reply:

We thank the reviewer for this critical observation and clarify that our intended reference was not to increase sequencing depth yields but rather to elevate the detection depth **cutoff**.

To avoid ambiguity, we have revised the Results part in the manuscript accordingly (page 7; lines 134-140):

"We use the Jaccard index metric to quantify site detection consistency, reflecting the fraction of sites jointly covered by the compared datasets. In contrast, the Pearson Correlation Coefficient (PCC) measures methylation level consistency at shared sites. While increasing the cutoff of detection depth reduced qualitative concordance (Jaccard index), it improved quantitative agreement (Pearson Correlation Coefficient (PCC) ≥ 0.9 , excluding outlier batch A1 at 10 $\times$), demonstrating the critical role of cytosine depth cutoffs on methylation measurement reliability (**Fig. 2B; Fig. S3**)."

We provide explanations to the results (page 8, lines 171-173):

"Notably, while the Jaccard index quantifies the proportion of CpG sites jointly detected above the 20 $\times$ cut-off, PCC evaluates methylation concordance specifically at these shared high-coverage (above 20 $\times$) CpG sites. "

We discuss the interpretation in the Discussion section (page 14, lines 359-361):

"As the detection cutoff for cytosine in CpG context increases, the number of cytosines detected on both strands decreases, and the absolute strand bias at these intersecting sites becomes smaller. "

5. Some of the figures are very dense and will be difficult to interpret for readers, perhaps the authors can consider removing superfluous information (either completely or move it to the supplementary materials). For instance, Fig. 5d/e show rankings of methods in many different types of genomic regions and by different metrics. I believe most readers will be most interested in which methods do best and the distinction between region types makes fairly little difference, and the ranking by PCC and RMSE seems identical I think. One option to simplify the figure could be to show two waterfall bar plots of the average Recall and the average PCC across all region types, making it straightforward to see the best and worst methods and even to quantitatively compare the metrics (which is impossible with

the rescaled values that are only encoded by color at the moment). The current plots could still be shown in the supplement, if the authors believe the complete information is critical.

Reply:

We sincerely appreciate the reviewer's constructive suggestions on improving the readability of the figures. In direct response, we have streamlined Fig. 5 by: (1) relocating the original panels 5d/e (detailed regional/metric rankings) to Supplementary Figure S11 to preserve comprehensive data; (2) creating two new box plots in the main Fig. 5d/e where Fig. 5d analyzes methylation accuracy across 11 batches using three reference-dependent metrics (Recall, PCC, RMSE) with Wilcoxon tests comparing CGI vs. non-CGI regions, while Fig. 5e maintains identical axis structure but contrasts promoter vs. enhancer regions - both pairs selected for their established biological significance in epigenetic regulation.

We provide modifications to the Results (page 11, lines 272-288):

“When stratifying by CpG island (CGI) versus non-CGI contexts, reference-dependent metrics revealed distinct technical patterns across all 11 batches. Recall and RMSE consistently demonstrated statistically significant differences between regions in every batch (Wilcoxon test, $p < 0.05$), with CGI regions exhibiting superior detection rates (higher Recall) and enhanced quantitative precision (lower RMSE) (**Fig. 5D**). In contrast, PCC showed significant batch-dependent variability—displaying non-significant differences in only two batches—attributable to its sensitivity to distributional characteristics of methylation data, wherein CGI and non-CGI regions manifest fundamentally distinct quantitative profiles. This metric-specific inconsistency underscores the recommendation to prioritize Recall and RMSE for robust quality assessment.

Comparative analysis of enhancer versus promoter regions—biologically pivotal regulatory elements—revealed more complex technical behavior (**Fig. 5E**). Significant Recall differences emerged in 7 of 11 batches, while RMSE variations occurred in 7 different batches. Notably, PCC and RMSE showed concordant directional trends (high PCC with low RMSE) in the majority of batches, demonstrating consistent quantitative

reliability across these functionally distinct elements despite underlying biological heterogeneity. Collectively, these context-stratified analyses validate our reference materials' capacity to detect nuanced, region-specific technical biases—a critical prerequisite for developing precision epigenomic workflows.”

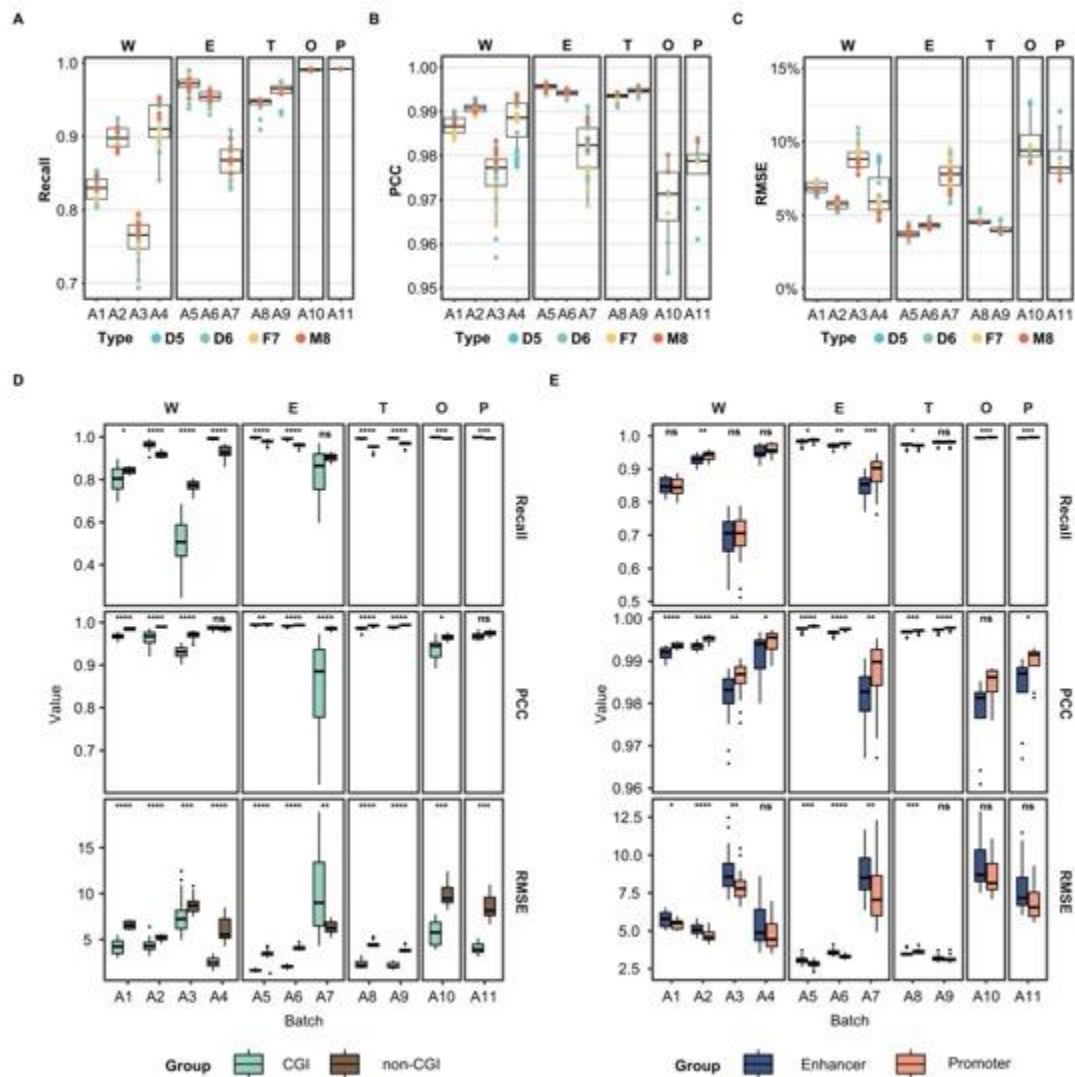


Fig. 5 Performance assessment based on reference datasets across protocols.

6. Fig. 2a: Typo “dpeth” instead of “depth” (twice).

Reply:

Thank you and we have made the corrections.

7. Fig. 2b: It's unclear from the figure that the x-axis represents sequencing depth. Add a label and tick marks.

Reply:

We appreciate the suggestions and have made the corrections (Fig. 2 Strand bias existed in each library across all protocols).

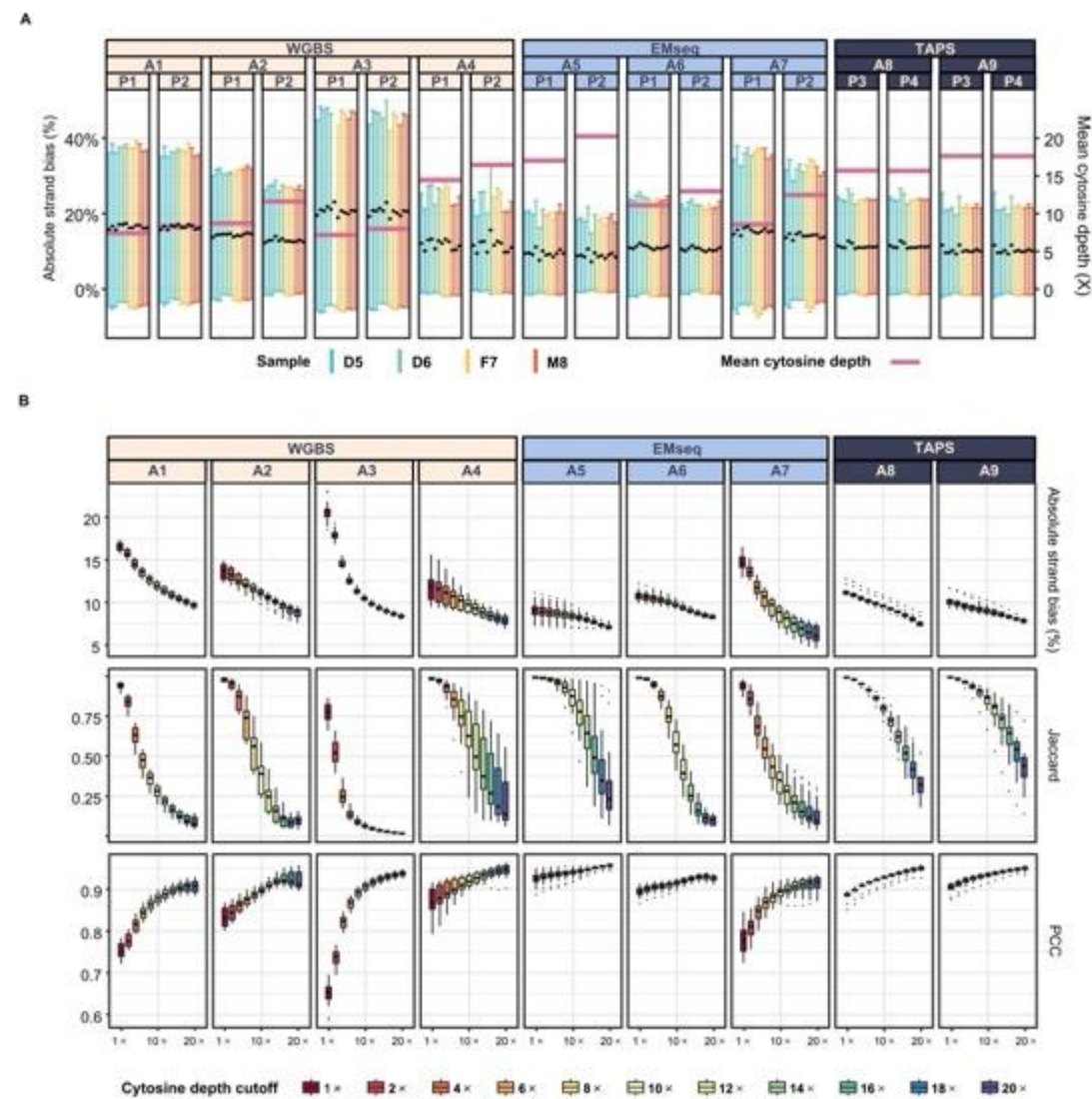


Fig. 2 Strand bias existed in each library across all protocols

8. Fig. 5d: Misalignment between letter annotations “E”, “T”, “E”, etc. and the rows of the plots.

Reply:

We appreciate the comments and have made the corrections. We have moved Fig. 5d to Fig S12 (Performance assessment based on reference datasets across different regions).

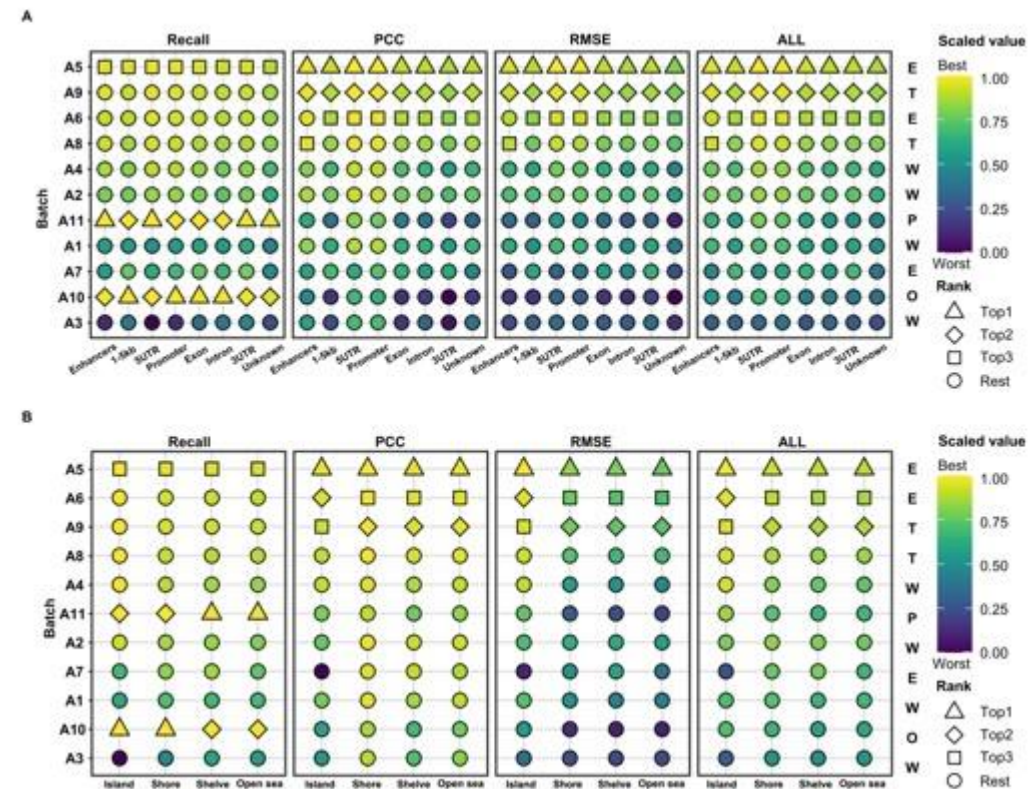


Fig. S12 Performance assessment based on reference datasets across different regions

Reviewer #2 (Remarks on code availability):

I had a look at the repository shared via the private link to Code Ocean. The folder structure is clear and it's straightforward to discover the code chunks relevant to different figures. A README file provides instructions for execution. Code for "upstream analysis", which refers to time-consuming execution of raw data processing pipelines, is also provided. I had a quick look at a few code files and the code looks valid. I have not executed the code.

Reply:

We are deeply grateful for the reviewer's efforts and valuable feedback.

Reviewer #3 (Remarks to the Author):

While the manuscript presents a useful effort and dataset there are important issues with the main conclusions drawn, that limit the interpretation of results:

1. The authors state: "Notably, all replicates exhibited substantial inter-strand methylation differences (absolute delta methylation $\geq 10\%$ at $1\times$) across protocols (Fig. 2B), challenging conventional strand-merging practices"

My interpretation of Figure 2A is that the observed "strand bias" is likely just sampling noise. Fig2A shows that the absolute methylation deviation between strands decreases with increasing coverage (as would be expected) and is still decreasing at their cutoff coverage. This argues against true biological differences between the strands, and hence suggests that we should in fact pool +ve and -ve strands, as opposed to treating them separately as the authors suggest.

Reply:

We sincerely appreciate the reviewer's insightful interpretation and agree that we should merge strands in appropriate CpG detection coverage.

We have made this revision in the manuscript (page 7, lines 141-143):

"All replicates exhibited substantial inter-strand methylation differences (absolute delta methylation $\geq 10\%$ at $1\times$) across protocols (**Fig. 2B**), indicating that strand bias is a consistent technical variation when quantification is not filtered by detection coverage."

2. The authors state: "While increasing sequencing depth reduced qualitative concordance (Jaccard index), it improved quantitative agreement (Pearson Correlation Coefficient (PCC) $\geq 0.9...$)" and "Technical replicates within labs showed low reproducibility (Jaccard index, Fig. S4)"

It seems very counterintuitive that qualitative concordance would go down when

quantitative concordance goes up. Is the Jaccard index being used to see if the same methylation values are reported across strand? If so, increasing coverage would clearly give more accurate values, but also make it less likely that they are numerically identical (e.g. 0.222 is not equal to 0.223). If instead, Jaccard is being used to assess coverage concordance (which makes more sense) this should be clearly explained.

Reply:

We thank the reviewer for this critical insight, and this apparent discrepancy was also noted by Reviewer #2. We confirm that the Jaccard index was indeed applied to assess coverage concordance (specifically, the proportion of overlapping CpG sites detected between two nreplicates, calculated as $|A \cap B| / |A \cup B|$), not methylation value identity.

To avoid ambiguity, we have revised the manuscript to explicitly state (page 7, lines 134-140):

“We use the Jaccard index metric to quantify site detection consistency, reflecting the fraction of sites jointly covered by the compared datasets. In contrast, the Pearson Correlation Coefficient (PCC) measures methylation level consistency at shared sites. While increasing the cutoff of detection depth reduced qualitative concordance (Jaccard index), it improved quantitative agreement ($PCC \geq 0.9$, excluding outlier batch A1 at 10×), demonstrating the critical role of cytosine depth cut-offs on methylation measurement reliability (**Fig. 2B; Fig. S3**).”

This clarification resolves the apparent paradox: raising the depth cutoff inevitably detects more unique sites in each replicate, leaving fewer sites in common; this lowers the Jaccard index. Meanwhile, the higher coverage at the retained sites improves methylation measurement precision, driving the PCC upward.

Point-by-point response to the referees' comments

Reviewer #1 (Remarks to the Author):

Thank you for clarifying the scope of the study and adding biological interpretations. The observation that monozygotic twins had a larger difference in PC1 compared to their parents is unexpected, but it seems very interesting. Additionally, the revised abstract helped me understand the primary objective of this project.

I have a few suggestions regarding the data visualizations. The authors used colors like cyan, pale teal, orange-yellow, yellow, and red-orange to represent different samples; however, similar colors are used for different purposes in the figures, such as pipelines in Fig. 3, CGI in Fig. 5, and Type in Fig. 6. I suggest that the authors avoid reusing colors to help reduce confusion. Fig. 1 uses different color codes for the samples. Additionally, the line width does not match the figure size in some figures (Fig. 4C and 4D); therefore, the lines overlap and are not visible.

Reply:

We are grateful for the reviewer's guidance, which was instrumental in clarifying the manuscript. We sincerely thank the reviewer for the careful feedback on our visualizations. We have harmonized the colour palette to avoid reuse across different encodings (samples, pipelines, CGI, types), added consistent in-figure keys, and adjusted line widths in Figs. 4C–4D to prevent overlap and improve clarity. We appreciate the reviewer's attention to detail, which has strengthened the presentation.

Reviewer #2 (Remarks to the Author):

The authors have addressed all my concerns.

Thank you for clarifying the apparent discrepancy between the qualitative / quantitative analyses. I think the sentences the authors added to the manuscript will avoid misunderstandings by future readers.

Very minor remark: the software checklist states that the software was available under the GNU General Public License v3, while the LICENSE file in the CodeOcean repository refers to the MIT License. Both should be synchronized.

Reply:

We sincerely appreciate the reviewer's positive evaluation and insightful feedback. We have synchronized the software licensing information: both the manuscript (Code Availability) and the Reporting Summary now specify the MIT License for our code

Reviewer #2 (Remarks on code availability):

I have briefly had a look at the files provided on CodeOcean, which seems well-structured and understandable. The authors have made a very good effort to provide reproducible code.

Reply:

We are deeply grateful for the reviewer's efforts and valuable feedback.

Reviewer #3 (Remarks to the Author):

The authors have addressed my concerns from my original review, and I believe it significantly improves the manuscript.

Reply:

We sincerely thank the reviewer for confirming that our revisions addressed the original concerns and significantly improved the manuscript; we appreciate the constructive feedback that guided these changes.

Reviewer #3 (Remarks on code availability):

The code is well documented and readable.

Reply:

We are deeply grateful for the reviewer's efforts and valuable feedback.