Evaluation of Whole-Genome DNA Methylation Sequencing Library Preparation Protocols

Jacob Morrison (jacob.morrison@vai.org)
Van Andel Research Institute https://orcid.org/0000-0001-8592-4744
Julie M. Koeman
Van Andel Research Institute
Benjamin K. Johnson
Van Andel Research Institute
Kelly K. Foy
Van Andel Research Institute
Wanding Zhou
The Children's Hospital of Philadelphia
David W. Chesla
Spectrum Health System
Larissa L. Rossell
Spectrum Health System
Emily J. Siegwald
Spectrum Health System
Marie Adams
Van Andel Research Institute
Hui Shen
Van Andel Research Institute https://orcid.org/0000-0001-9767-4084

Research Article

Keywords: DNA Methylation, Epigenetics, Whole genome bisulfite sequencing, Enzymatic methylation sequencing, Fallopian tube

Posted Date: March 1st, 2021

DOI: https://doi.org/10.21203/rs.3.rs-249202/v1

Abstract

Background: With rapidly dropping sequencing cost, the popularity of whole-genome DNA methylation sequencing has been on the rise. Multiple library preparation protocols exist, but a systematic evaluation and benchmarking of their performance against each other is currently lacking. We have performed 22 whole-genome DNA methylation sequencing experiments on fresh frozen human samples, and extensively benchmarked common library preparation protocols for whole-genome DNA methylation sequencing, including three traditional bisulfite-based protocols and a new enzyme-based protocol. Additionally, different input DNA quantities were compared for two kits compatible with a reduced starting quantity. In addition, we also present bioinformatic analysis pipelines for sequencing data from each of these library types. Results: An assortment of metrics were collected for each kit, including raw read statistics, library quality and uniformity metrics, cytosine retention, and CpG beta value consistency between technical replicates. Overall, the NEBNext Enzymatic Methyl-seq kit performed quantitatively better than the other three protocols at two different DNA input amounts. Additionally, the results for the different input amounts were generally consistent across all metrics. Conclusions: Based on these results, we recommend use of the NEBNext Enzymatic Methyl-seq kit for whole-genome DNA methylation sequencing. Further, a general bioinformatic pipeline is applicable across the four protocols, with the exception of extra trimming needed for the Swift Bioscience's Accel-NGS Methyl-Seq protocol to remove the Adaptase sequence.

Full Text

This preprint is available for download as a PDF.

Figures

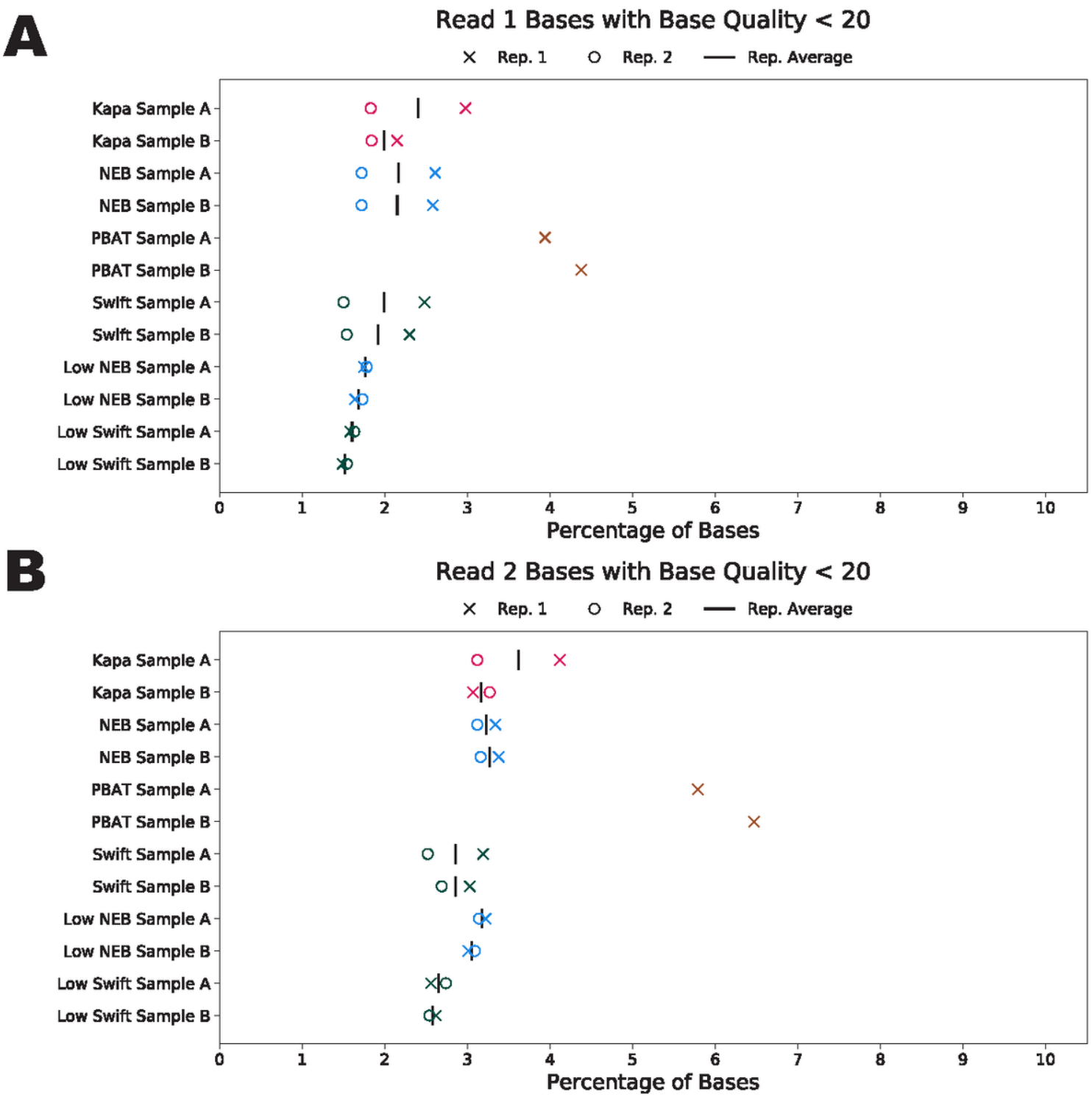

Figure 1

Raw read statistics for each protocol. (A) Percentage of bases with base quality < 20 for read 1. (B) Percentage of bases with base quality < 20 for read 2.

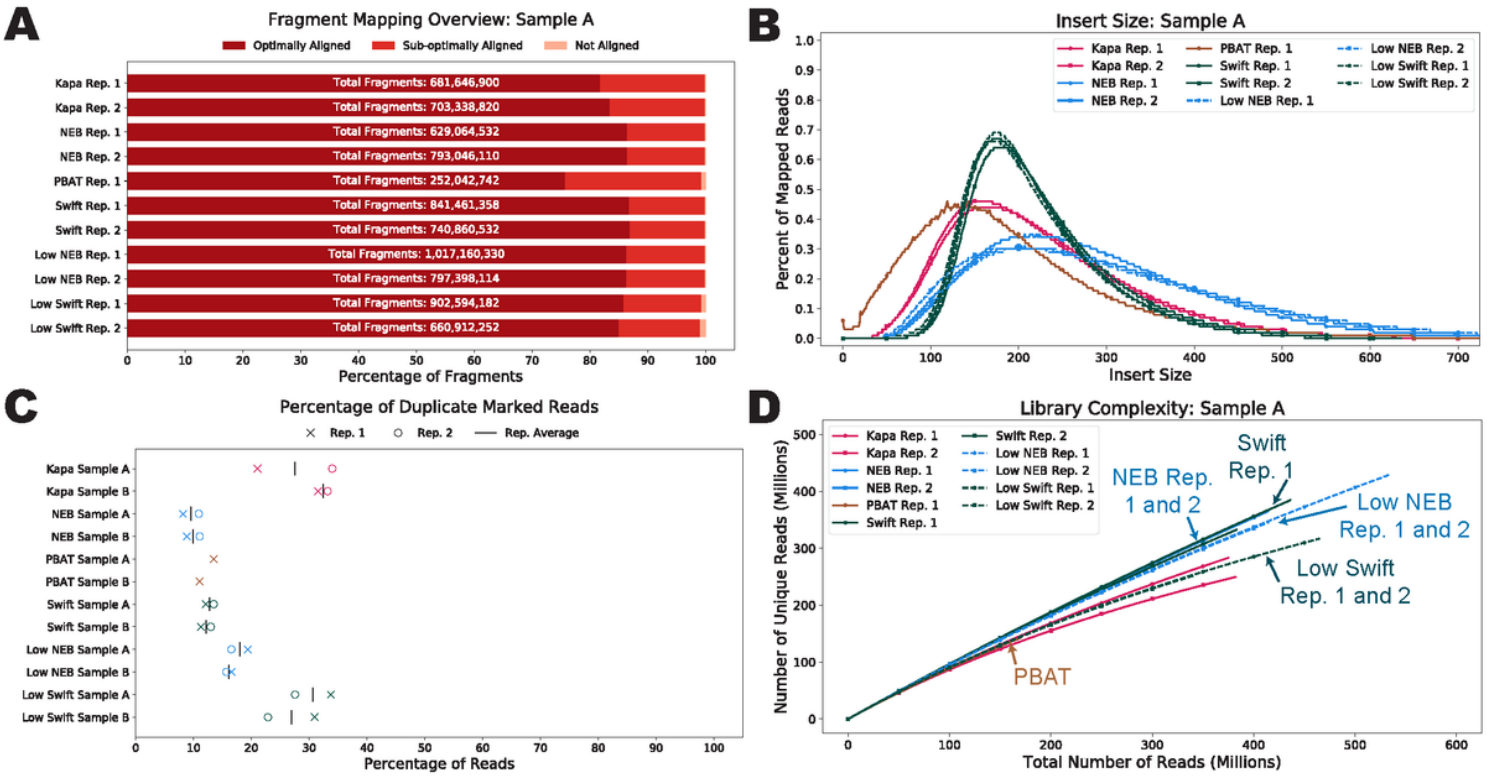

Figure 2

Library quality metrics for each protocol for Sample A. (A) The percentage of optimally, sub-optimally, and not aligned read fragments for each protocol. Note, read fragments treat reads 1 and 2 as separate entities, as it is possible that one read in the pair is mapped, while the other is not. (B) Insert size distribution. (C) Duplicate rate for reads with MAPQ ≥ 40. (D) The library complexity, which is a function of the duplicate rate. Metrics for Sample B are shown in Additional file 1: Figure S2.

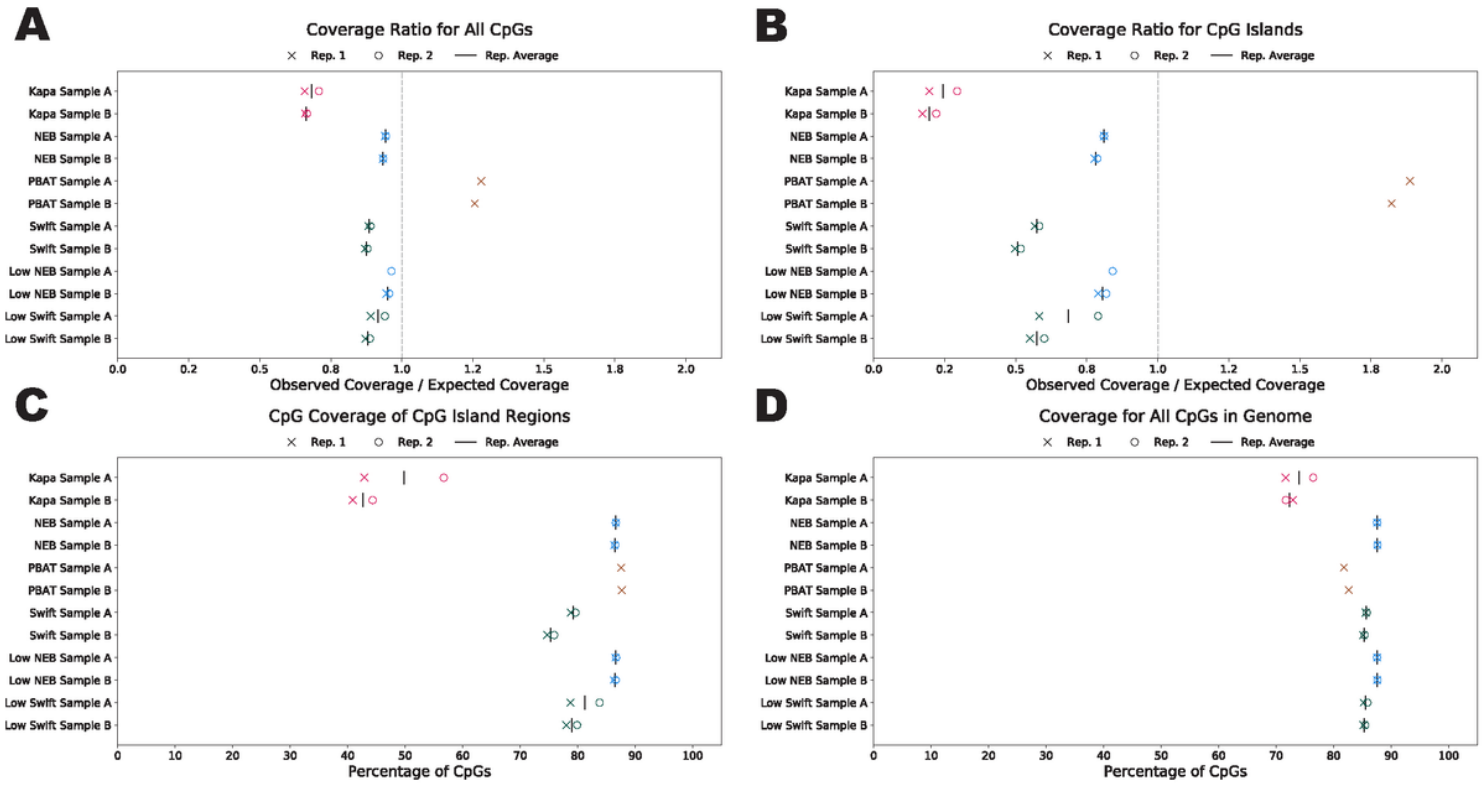

Figure 3

Library uniformity as measured by coverage of various ge- nomic element categories. Ratio of observed coverage to expected coverage for (A) all CpGs and (B) CpG islands. (C) Percentage of all CpGs covered by at least one unique read with MAPQ ≥ 40. (D) Percentage of CpGs in CpG islands covered by at least one unique read with MAPQ ≥ 40. Note, for (C) and (D), all libraries were downsampled to be comparable to PBAT (150 million reads, or ∼4.8X coverage, per sample, see Methods for details); therefore, any differences are not likely confounded by sequencing depth. As expected, this coverage will be substantially higher at increased depth.

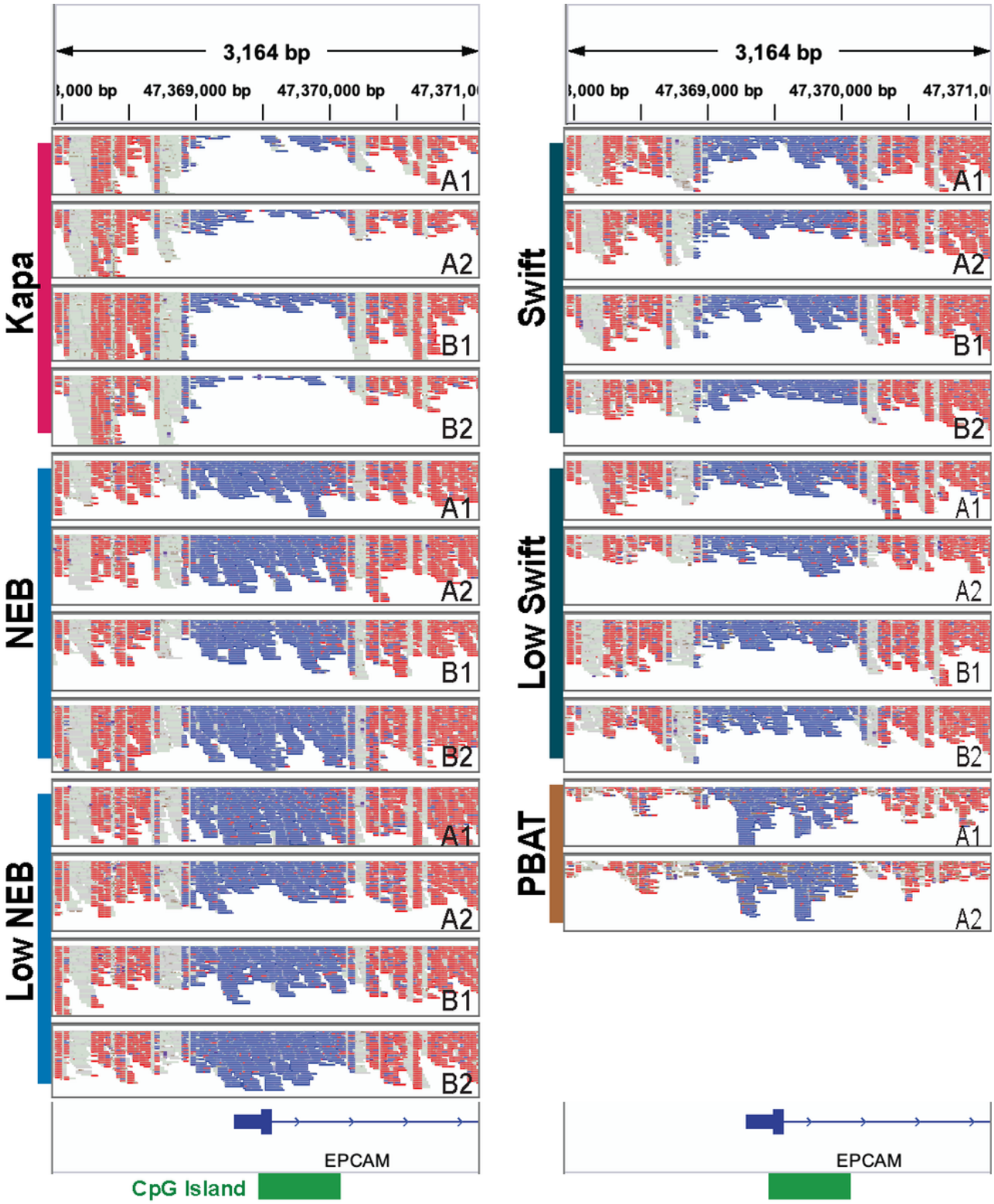

Figure 4

The EPCAM promoter region as a representative example for data generated with the protocols. The aligned reads tracks are taken from the Integrated Genomics Viewer (IGV) [45] in the bisulte mode mode where red represents an unconverted cytosine and blue represents a converted cytosine. Each panel represents one sample, with A and B denoting the biological replicates and 1 and 2 the technical replicates for each library construction protocol. The shown region is 1500 bp upstream and downstream

of exon 1. The location of a CpG island is indicated with a green box on the bottom. Note, the strands for the PBAT samples have been flipped in silico before being displayed to account for the strand definition in the Miura and Ito protocol. The strands in the PBAT protocol are opposite from what is expected by IGV, as well as the definition used by the other protocols.

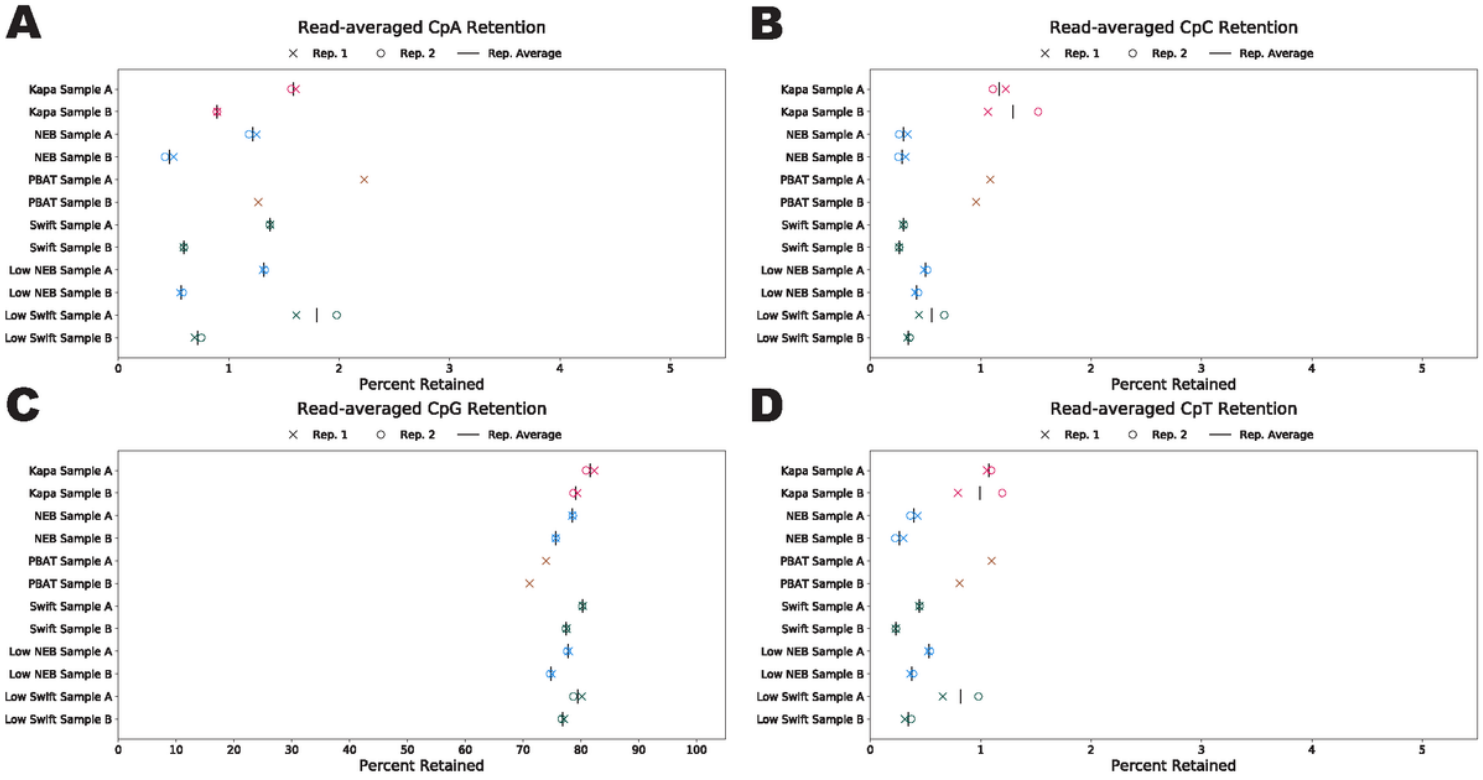

Figure 5

Read-averaged cytosine retention by dinucleotide context, namely (A) CpA, (B) CpC, (C) CpG, and (D) CpT. In each panel two technical replicates are shown for each biological replicate. The x-axis denotes percent retention, with a scale of 0-5% for CpH panels and 0-100% for the CpG panel.

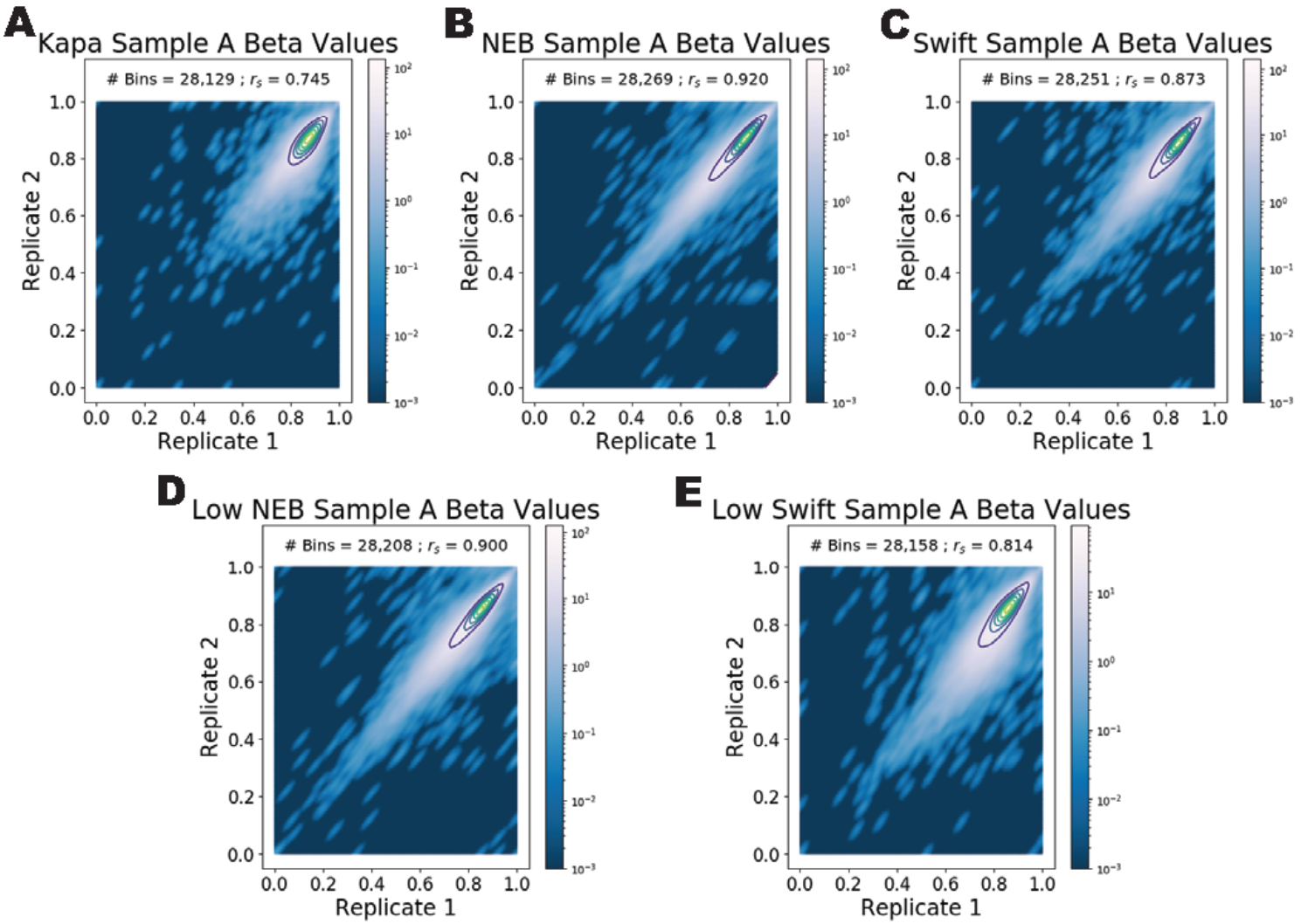

Figure 6

The NEB protocol has the highest correlation of beta values between Sample A technical replicates. The Spearman correlation coef- cient, rs, between the two replicates is listed in each gure, along with the number of 100kb bins used in calculating the coefficient. Note, all libraries were downsampled to be comparable to PBAT; therefore, any differences are not likely confounded by sequencing depth. Overall low correlation values are due to low coverage from downsampling.

Supplementary Files

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

- additionalfile1.pdf
- wgmskitcomparisonv1.3.bib
- bmcart.cls
- wgmskitcomparisonv1.3.tex
- vancouver.bst