Comparison of Methylation Capture Sequencing and Infinium EPIC Methylation Array in Peripheral Blood Mononuclear Cells

Chang Shu
Yale University School of Medicine https://orcid.org/0000-0002-3730-5102
Xinyu Zhang
Yale University School of Medicine
Bradley E. Aouizerat
New York University College of Dentistry
Ke Xu (ke.xu@yale.edu)
https://orcid.org/0000-0002-6472-7052

Research

Keywords: methylation capture sequencing, EPIC, DNA methylation, Peripheral Blood Mononuclear Cells

Posted Date: October 29th, 2020

DOI: https://doi.org/10.21203/rs.3.rs-33940/v2

Version of Record: A version of this preprint was published on November 23rd, 2020. See the published version at https://doi.org/10.1186/s13072-020-00372-6.

Abstract

Background: Epigenome-wide association studies (EWAS) have been widely applied to identify methylation CpG sites associated with human disease. To date, the Infinium Methylation EPIC array (EPIC) is commonly used for high-throughput DNA methylation profiling. However, the EPIC array covers only 30% of the human methylome. Methylation Capture bisulfite sequencing (MC-seq) captures target regions of methylome and has advantages of extensive coverage in the methylome at an affordable price.

Methods: Epigenome-wide DNA methylation in four peripheral blood mononuclear cell samples was profiled by using SureSelectXT Methyl-Seq for MC-seq and EPIC platforms separately. CpG site-based reproducibility of MC-seq was assessed with DNA sample inputs ranging in quantity of high (> 1000ng), medium (300-1000ng), and low (150ng-300ng). To compare the performance of MC-seq and the EPIC arrays, we conducted a Pearson correlation and methylation value difference at each CpG site that was detected by both MC-seq and EPIC. We compared the percentage and counts in each CpG island and gene annotation between MC-seq and the EPIC array.

Results: After quality control, an average of 3,708,550 CpG sites per sample was detected by MC-seq with DNA quantity >1000ng. Reproducibility of MC-seq detected CpG sites was high with strong correlation estimates for CpG methylation among samples with high, medium, and low DNA inputs ($r > 0.96$). The EPIC array captured an average of 846,464 CpG sites per sample. Compared with the EPIC array, MC-seq detected more CpGs in coding regions and CpG islands. Among the 472,540 CpG sites captured by both platforms, methylation of a majority of CpG sites was highly correlated in the same sample (r: 0.98~0.99). However, methylation for a small proportion of CpGs (N=235) differed significantly between the two platforms, with differences in beta values of greater than 0.5.

Conclusions: Our results show that MC-seq is an efficient and reliable platform for methylome profiling with a broader coverage of the methylome than the array-based platform. Although methylation measurements in majority of CpGs are highly correlated, a number of CpG sites show large discrepancy between the two platforms, which warrants further investigation and needs cautious interpretation.

Introduction

The rapid increase in the number of epigenome-wide association studies (EWAS) have successfully identified differentially methylated CpG sites that are associated with environmental exposures and diseases (1-6). Such DNA methylation marks have been used as biomarkers for diagnosing, subtyping, and monitoring disease progression(7-11). One of the most popular and affordable methods to profile epigenome-wide DNA methylation are array-based platforms, primarily the Illumina Human Methylation 450K (450K) and Infinium MethylationEPIC (EPIC) BeadChips (Illumina Inc, San Diego, CA). These arrays both utilize Illumina's beadchip technology that does not require polymerase chain reaction (PCR), but is subject to dye intensity biases between the two platforms(12). These arrays have limited coverage of the

methylome and can only detect up to 870,000 CpGs across the epigenome, leaving a large proportion of CpG sites unmeasured. Moreover, the EPIC array offers improved but still suboptimal coverage of regulatory elements(13). Whole genome bisulfite sequencing (WGBS) is able to capture more than 28 million CpGs, but the feasibility remains low for the population-based EWAS due to high cost and large genomic DNA input requirements to compensate for degradation during DNA bisulfite treatment. Alternatively, Methylation Capture Sequencing (MC-seq) is able to detect DNA methylation at single-nucleotide resolution utilizing a targeted next-generation sequencing approach(14). It permits profiling of significantly more CpG sites than the EPIC array, requires less genomic DNA input than WGBS, and less expensive than WGBS, but can be susceptible to bias due to the presence of PCR duplicates. Feature-to-cost comparisons among different platforms can help understand the utilities of each platform and provide guidance for investigators in choosing a methylation profiling platform.

A few studies have compared the CpG coverage, reproducibility, and performance of array-based and MC-seq platforms (15-17). Teh et al compared MC-seq and the 450K array in seven DNA samples extracted from saliva (15). A recent study compared the EPIC array and TruSeq targeted bisulfite sequencing in four cord blood DNA samples (17). However, no comparisons of MC-seq and array-based methylome profiling of peripheral blood mononuclear cell (PBMC) has been reported. Here, we profiled the DNA methylome in PBMC using the Agilent SureSelect Methyl-Seq platform and compared the results to the EPIC array in DNA samples extracted from PBMCs.

Methods

Methylation Capture Sequencing (MC-seq)

DNA Samples Description. DNA was extracted from de-identified PBMC collected from four individuals. Genomic DNA quality was determined by estimating the A260/A280 and A260/A230 ratios by spectrophotometry and concentration by fluorometry. DNA integrity and fragment size were confirmed using a microfluidic chip run on an Agilent Bioanalyzer. To assess the reproducibility of MC-seq by DNA quantity, DNA samples from each participant were profiled in triplicate times with high (>1000ng), medium (300-1000ng) and low (150-300ng) DNA input. In total, 12 DNA samples were measured by MC-seq. Bisulfate conversion was conducted for each DNA sample as described below.

Methyl-Seq Target Enrichment Library Prep. Indexed paired-end whole genome sequencing libraries were prepared using the SureSelect XT Methyl-Seq kit (Agilent, part#G9651B). Genomic DNA was sheared to a fragment length of 150-200 bp using focused acoustic energy delivered by the Covaris E220 system (Covaris, part#500003). Fragmented sample size distribution was determined using the Caliper LabChip GX system (PerkinElmer, Part#122000). Fragmented DNA ends were repaired with T4 DNA Polymerase and Polynucleotide Kinase and “A” base was added using Klenow fragment in a single reaction followed by AMPure XP bead-based purification (Beckman Coulter, part#A63882). The methylated adapters were ligated using T4 DNA ligase followed by AMPure XP bead purification. Quality and quantity of adapter-ligated DNA were assessed using the Caliper LabChip GX system. Samples yielding >350 ng were

enriched for targeted methylation sites by using the custom SureSelect Methyl-Seq Capture Library. Hybridization was performed at 65°C for 16 hours using a C1000 Thermal Cycler (BIO-RAD, part# 1851197). Once the enrichment was completed the samples were mixed with streptavidin-coated beads (Thermo Fisher Scientific, part#65602) and washed with a series of buffers to remove non-specific bound DNA fragments. DNA fragments were eluted from beads with 0.1 M NaOH. Unmethylated C residues of enriched DNA were modified by bisulfite conversion using the EZ DNA Methylation-Gold Kit (Zymo Research, part#D5005). The SureSelect enriched, bisulfite-converted libraries were PCR amplified using custom-made indexed primers (IDT, Coralville, Iowa). Dual-indexed libraries were quantified by quantitative polymerase chain reaction (qPCR) using the Library Quantification Kit (KAPA Biosystems, Part#KK4854) and inserts size distribution was assessed using the Caliper LabChip GX system. Samples with a yield of ≥2 ng/ul were proceeded to sequencing.

Flow Cell Preparation and Sequencing. Sample concentrations were normalized to 10 nM and loaded onto an Illumina NovaSeq flow cell at a concentration that yields 40 million passing filter clusters per sample. Samples were sequenced using 100bp paired-end sequencing on an Illumina HiSeq NovaSeq according to Illumina standard protocol. The 10bp dual index was read during additional sequencing reads that automatically follows the completion of the first read. Data generated during sequencing runs were simultaneously transferred to the Yale Center for Genome Analysis high-performance computing cluster. A positive control (prepared bacteriophage Phi X library) provided by Illumina was spiked into every lane at a concentration of 0.3% to monitor sequencing quality in real time.

Preprocessing and Quality Control. Signal intensities were converted to individual base calls during a run using the system's Real Time Analysis (RTA) software. Sample de-multiplexing and alignment to the human genome was performed using Illumina's CASAVA 1.8.2 software suite. The sample error rate was required to be less than 1% and the distribution of reads per sample in a lane was required to be within reasonable tolerance.

Quality control (QC) on MC-seq was conducted following standard procedure as previously described(18). Quality of sequence data was examined by using *FastQC* (ver. 0.11.8). Adapter sequences and fragments at 5’ and 3’ (phred score <30) with poor quality were removed by *Trim_galore* (ver. 0.6.3_dev). We used Bismark pipelines (ver. v0.22.1_dev) to align the reads to the bisulfite human genome (hg19) with default parameters(19). Quality-trimmed paired-end reads were transformed into a bisulfite converted forward strand version (C->T conversion) or into a bisulfite treated reverse strand (G->A conversion of the forward strand). Duplicated reads were removed from the Bismark mapping output and CpG, CHG and CHH (where H = A, T or C) were extracted.

All CpG sites were grouped by sequencing coverage, also known as read depth. The groups with coverage of 1x to 100x were used to test the relationship between coverage and number of CpG sites. Only the CpG sites with coverage > 10x depth were used for final comparisons to ensure MC-seq data quality. Genes were annotated using Homer *annotatePeaks.pl*, including intergenic, 5’UTR, promoter, exon, intron, 3’UTR, transcription start site (TTS), and non-coding categories. CpG island, shore, shelf, and open sea

annotation was defined by locally developed bash and R scripts based on genomic coordinates (hg19) of CpG islands from the UCSC genome browser. Definition of CpG shores was defined as up to 2 kb from CpG islands and CpG shelf was defined as up to 2 kb from a CpG shore.

Assessment of Reproducibility. We assessed CpG- and participant-based reproducibility for MC-seq among 12 samples with DNA quantity of high, medium, and low input in two ways. First, CpG-based reproducibility was assessed by calculating Pearson correlations using the CpG sites in common from the samples from the same participant with different input DNA quantities. Scatterplots were rendered showing 10,000 randomly selected common CpG sites comparing samples with high and medium, high and low and medium and low DNA input. Second, participant-based reproducibility was assessed by comparing methylation profiles among pairs of participants using the samples with high DNA input, by calculating Pearson correlations of common CpG sites.

EPIC Array Data Preprocessing

The Infinium Methylation EPIC array (Illumina, San Diego, CA, USA) was used to measure PBMC DNA methylation profiles from the same four participants. These four samples with DNA input of 1000 ng were preprocessed using standard procedures as previously described (20). Briefly, the predicted sex based on methylome was consistent with self-reported sex for all samples. All samples had a call rate greater than 0.15. A total of 19,090 CpG sites on X chromosomes and 537 CpG sites on Y chromosomes were filtered. Probes within 10 base pairs of a single nucleotide polymorphism were removed. A total of 846,464 CpG sites passed quality control.

Comparison of Methylation at Each CpG site Between MC-seq and EPIC Array

The overall distribution of gene annotation in terms in relation to CpG island and genetic region between MC-seq and EPIC array data among samples from the four participants were compared. CpG sites common between MC-seq and EPIC array assays were defined according to genomic coordinates. Pearson correlation and the absolute beta-difference value (Db) were calculated among common CpG sites between MC-seq methylation percentage values and EPIC methylation beta values by using R (ver. 3.5.1). If median Db of the common CpG site was > 0.1, it was defined as a discordant CpG pair between MC-seq and EPIC; otherwise, the CpG site was defined as a concordant CpG pair. The density plot of Db and a Manhattan plot showing the distribution of Db across epigenome were illustrated. Scatterplots were rendered showing the correlation of b values from 10,000 randomly selected CpG sites measured by both MC-seq and EPIC array.

Results

MC-seq Overview and Reproducibility

In MC-seq, all sequences were efficiently mapped to the reference genome with greater than 89% mapping efficiency. Interestingly, the number of non-CpG sites was significantly greater than the number of CpG

sites. Among all detected methylation sites by MC-seq, 11% were CpG sites, 65% were CHH sites, and 24% were CHG sites (**Figure 1a)**.

Figure 1b shows the relationship of the number of detected CpG sites and depth of sequence coverage by MC-seq in one sample. The depth of read at which the majority of sites were sequenced was estimated to be approximately 10X coverage, observed as the inflection point of the distribution of **Figure 1b**. An increase of depth only slightly increased the capture of CpG sites and the inflection point is on 10X coverage, consistent with previous literature(15, 17). Thus, the number of CpG sites with coverage ≥ 10x from MC-seq was used in subsequent analyses.

After quality control filtering, MC-seq captured an average of 2,878,207 methylation CpG sites with coverage ≥10x among the 12 DNA samples, with an average of 3,708,550 CpG sites among samples with high DNA input (>1000ng), an average of 3,046,172 CpG sites among samples with medium DNA input (300-1000ng), and an average of 1,879,898 CpG sites among samples with low DNA input (150-300ng) (**Figure 1c** and **Table 1**). Despite the fact that the detected number of CpG sites varied depending on DNA input quantity, CpG-based correlation among the common CpG sites between samples with high and medium, high and low DNA input quantity exceeded r>0.95. Correlations of common CpG sites between medium and low DNA inputs was also high with r in 0.92-0.94 (**Table 2**). **Figure 1d** shows the scatterplot of 10,000 randomly selected common CpGs between samples with high and medium, high and low, and medium and low DNA input quantity. Pair-wise participant-based correlations were high as r>0.98 among common CpG sites (**Table 3**). Overall, MC-seq exhibited good reproducibility. The methylation profile generating in high DNA input from each participant was used for subsequent analyses.

Table 1: Detected CpG number by DNA amount in MC-seq with coverage ≥ 10X

DNA amount	Participant ID	CpG number	Average CpG Number
Low	S1	1,774,940	1,879,898
	S2	1,831,086	
	S3	2,154,732	
	S4	1,758,834	
Medium	S1	2,768,456	3,046,172
	S2	3,338,200	
	S3	3,119,259	
	S4	2,958,772	
High	S1	3,406,879	3,708,550
	S2	3,642,776	
	S3	3,722,552	
	S4	4,061,994	
Total average			2,878,207

Table 2: Comparison of MC-seq between samples with high, medium and low DNA input amount

Participant ID	DNA Amount			
	High	Medium	Common CpG	Pearson Correlation
S1	3,406,879	2,768,456	2,747,844	0.984
S2	3,642,776	3,338,200	3,283,296	0.984
S3	3,722,552	3,119,259	3,101,938	0.977
S4	4,061,994	2,958,772	2,957,239	0.979
	High	**Low**	**Common CpG**	
S1	3,406,879	1,774,940	1,771,936	0.960
S2	3,642,776	1,831,086	1,829,919	0.966
S3	3,722,552	2,154,732	2,153,175	0.974
S4	4,061,994	1,758,834	1,758,622	0.963
	Medium	**Low**	**Common CpG**	
S1	2,768,456	1,774,940	1,745,241	0.942
S2	3,338,200	1,831,086	1,827,536	0.943
S3	3,119,259	2,154,732	2,135,980	0.939
S4	2,958,772	1,758,834	1,744,416	0.928

Table 3: Overlap of detected CpG across samples with high DNA input amount by MC-seq

Participant ID 1	Participant ID 2	Common CpG	Pearson R
S1	S2	3,336,037	0.980
S1	S3	3,350,314	0.976
S1	S4	3,394,970	0.982
S2	S3	3,519,772	0.978
S2	S4	3,613,753	0.982
S3	S4	3,676,406	0.978

Distribution of Methylome Regions by MC-seq and EPIC

We compared genome-wide DNA methylation captured by MC-seq and by EPIC array in the four high DNA input samples. An average of 3,708,550 CpG sites were detected by MC-seq and 846,464 CpG sites by EPIC array. Overall, MC-seq detected 11.5 times more CpG sites in exons and 10.2 times more CpG sites in 5' UTR region compared to the EPIC array, and 4.8 to 8.9 times more CpG site in other categories of genomic regions by MC-seq compared to EPIC array. However, the proportion of CpGs out of all CpGs successfully measured that map to gene regions in MC-seq as compared to the EPIC array did not significantly differ between these two platforms. For example, the proportion of CpG sites in transcription termination site (TTS) regions were similar between two platforms. MC-seq showed slightly greater proportions of CpG sites in 5'UTR and exon regions while the EPIC array detected a greater proportion of CpG sites in promoter regions (**Figure 2a**). In terms of CpG sites in relation to CpG islands including open seas, shelves, and shores, MC-seq detected 10.9 times more CpG sites located on CpG islands and 5.4-6.2

times more on other regions compared with the EPIC array. The proportion of CpG islands detected by MC-seq was greater than by the EPIC array (42% versus 29%) while the EPIC array detected a modestly higher percentage of CpG sites located in open seas than the MC-seq (39% versus 31%) (**Figure 2b**).

Comparison of Common CpG sites Measured by MC-seq and EPIC

A total of 472,540 CpG sites were measured by both platforms. Overall, the correlations of these shared CpG sites was high, ranging from r=0.983 to 0.985 across the four samples (**Figure 3a**). **Figure 3b** presents the distribution of the absolute difference of methylation b values between MC-seq and EPIC. A small proportion of CpG sites (1.4%) were discordant (i.e., Db > 0.1), while 98.6% of CpG sites were concordant (i.e., Db < 0.1). **Figure 3a** presents the concordant (blue) and discordant CpG sites (green) between MC-seq and EPIC for participant S1 (**Figure 3a**). The 60,753 discordant CpG sites appeared to be randomly distributed across the epigenome (**Figure S1**). Among the discordant CpG sites, we identified 239 CpG sites with highly discrepant methylation (i.e., Db>0.5) (**Table S1)**. Participants S2, S3, S4 has similar distribution of concordant and discordant plots as participant S1 and in shown in **Figure S2**.

Density plots of methylation b showed bimodal distribution using both the MC-seq and the EPIC array platforms (**Figure 3c**). Density of methylated CpG sites was slightly higher than the density of unmethylated CpG sites on both platforms. However, the two peaks in the EPIC array density plot were closer than the two peaks in the MC-seq density plot (**Figure 3c**), indicating that MC-seq captures a higher dynamic range (i.e., more methylated and unmethylated) of CpG sites than the EPIC array. Participants S2, S3, S4 has similar density plots and in shown in **Figure S3**.

Discussion

We profiled the same PBMC samples using the MC-seq and EPIC array platforms and compared their performance. Our results show that the Agilent SureSelect Methyl-Seq targeted enrichment platform produced high quality DNA methylation sequencing data at single base-pair resolution. MC-seq can reliably detect CpG sites with DNA input quantities as low as 300ng. Overall, MC-seq detected 3-4 times more CpG sites than the EPIC array; however, the proportion of CpG sites mapped on functional genomic regions was similar between the two platforms. Methylation at a majority of CpG sites between the two platforms was highly correlated while methylation at a low percentage of CpG sites differed significantly between the two platforms. Specifically, we found that methylation at 239 CpG sites differed significantly between the two platforms with absolute Db values greater than 0.5, which suggests that these CpG sites should be interpreted with caution in EWAS studies.

Our results show that MC-seq produces highly reliable CpG site methylation estimates across the genome. The observed CpG-based reproducibility is high, suggesting that technical variation on CpG calls is low. Inter-personal methylation variation is important for EWAS analysis. We found that our participant-based methylation on common CpG sites across four participants are also highly correlated, which further demonstrates the high reproducibility of this platform.

One disadvantage of sequencing-based approaches is the requirement for a larger quantity of input DNA than array-based approaches for methylation profiling. The recommended input DNA for Agilent SureSelect platform is 1ug while input DNA quantity for EPIC array can be as low as 250ng. Input DNA quantity is one important consideration influencing study design and methylation assay platform selection for population-based EWAS. Agilent has reported that DNA quantity can be as low as 250ng for SureSelect sequencing(14). To examine whether DNA quantity impacts the performance of MC-seq and to test whether low input DNA quantity also produces reliable CpG detection, we compared the capacity of CpG site detection across three different DNA input quantities. We found that medium DNA input quantity (i.e., 300ng to 1000ng) reliably detected CpG sites is comparable to the number of CpG sites captured by high DNA input quantity (i.e., greater than 1000ng). Low DNA input quantity (i.e., less than 300ng) detected the lowest number of CpG sites compared with high and medium DNA input quantity. For samples with low DNA input quantity, additional PCR cycles are needed to ensure post-capture library yield that results in extensive duplicate reads. In the four low DNA input samples, the duplicate rate exceeds 80%. Thus, removing duplicate reads is an important step in the QC process for MC-seq. We found that the number of CpG sites in low DNA input samples without duplicated reads still is significantly higher than the number of CpG sites detected by the EPIC array.

Consistent with previous reports, we found that methylation at the majority of CpG sites measured by both approaches (>98%) is highly consistent between MC-seq and array-based methods. However, we identified 1.4% of CpG sites with discrepancies in CpG methylation that exceeds 10%. More importantly, 239 out of 60,753 discordant CpG sites had methylation differences exceeding 50%. These CpG sites are located on 159 gene regions (**Table 4**). Some of these genes have been previously reported to be associated with diseases. For example, *SLC45A4* was reported to harbor an epigenetic marker for adiposity(21). The methylation b differs on the CpG site of this gene by as much as 0.63 between the two platforms. We have also identified those CpG sites that showed less but still apparent discrepancy between the two assay platforms (i.e., absolute difference of beta values between 0.1-0.5). The top 100 CpG sites discrepant in a range of 0.1-0.4 between two platforms are presented in **Table S2** to allow investigators to consider this potential source of bias in EWAS findings. The discrepancy might be due to bias in the performance of the beadchip assay at these positions, sequence context-dependent impacts on the performance of sequencing, batch effects, or a combination of these possibilities. This large discrepancy warrants further investigation and interpretation of findings at these CpG sites must be interpreted with caution.

One of the limitations of this study is the small number of participants used to estimate inter-sample variability. A previous study used a benchmark approach to evaluate performance of different platforms (17) and concluded that the EPIC array performed better than the MC-seq platform. However, the study did not remove duplicate reads as part of their data processing, which may have comprised the QC for MC-seq data processing as discussed above. Future studies including benchmarking using a larger sample size could further improve the analysis of platform performance. Of note, MC-seq detected high percentages of CHG and CHH sites across four methylome, which is consistent with previous reports(15). The significances of those methylation sites warrant further investigation.

New approaches to measurement of DNA methylation continue to emerge that may warrant similar investigation in an ongoing effort to provide users with empiric comparisons to inform decisions about platform selection. One recent approach is enzymatic methyl-sequencing (EM-seq) (e.g., NEBNext EM-seq by New England Biolabs, Ipswich, MA) (22). The input genomic DNA, requirement is low 10-200 ng and EM-seq has comparable performance to WGBS (22), but its performance in relation to array- or capture sequencing-based approaches has not been reported. Should EM-seq gain popularity, it would be important to directly compare the performance of MC-seq and EM-seq to provide empiric evidence to users to inform platform selection.

Nevertheless, we have demonstrated that MC-seq is an efficient, reliable, and affordable platform that allows medium input quantity of DNA input (i.e., >300ng), which is equivalent to DNA input required for EPIC array. MC-seq has the advantage of capturing significantly more CpG sites than the EPIC array. Although methylation measurements between the two platforms are highly consistent, we have identified a small number of CpG sites that must be interpreted with caution if they are associated with a trait of interest because they showed significant discrepancies between the two platforms.

Conclusions

Our results show that MC-seq is an efficient and reliable platform for methylome profiling with a broader coverage of the methylome than the array-based platform. Although methylation measurements in majority of CpGs are highly correlated, a number of CpG sites show large discrepancy between the two platforms, which warrants further investigation and needs cautious interpretation.

List Of Abbreviations

EM-seq: Enzymatic Methyl-seq

EPIC: Illumina Infinium MethylationEPIC Beadchip

EWAS: Epigenome-Wide Association Studies

MC-seq: Methylation Capture Sequencing

PBMC: Peripheral Blood Mononuclear Cell

PCR: polymerase chain reaction

QC: Quality Control

RTA: Real Time Analysis

TTS: Transcription Termination Site

WGBS: Whole Genome Bisulfite Sequencing

Declarations

Ethics approval and consent to participate

The study was approved by the committee of the Human Research Subject Protection at Yale University and the Institutional Research Board Committee of the Connecticut Veteran Healthcare System. De-identifiable samples were from Women's Interagency HIV Study cohort. All participants provided written consents.

Acknowledgement

The project was supported by the National Institute on Drug Abuse (R03DA039745, R01DA038632, R01DA047063, R01DA047820). The authors appreciate the support of the Yale Center of Genomic Analysis and Women's Interagency HIV Study.

Authors' contributions

CS contributed to data analysis and the first draft of manuscript. XZ contributed to data processing, quality control, analysis, and manuscript preparation. BA was involved in manuscript preparation and provided peripheral blood monocyte cells. KX contributed to study design, analytical strategies, and manuscript preparation. All authors read and approved the final manuscript.

Competing Interests

The authors declare that they have no competing interests.

Data Availability

All methylation data from MC-seq and EPIC platforms is deposited in GEO (GSE152922).

References

1. Bakusic J, Schaufeli W, Claes S, Godderis L. Stress, burnout and depression: A systematic review on DNA methylation mechanisms. Journal of Psychosomatic Research. 2017;92:34-44.
2. Kraiczy J, Nayak KM, Howell KJ, Ross A, Forbester J, Salvestrini C, et al. DNA methylation defines regional identity of human intestinal epithelial organoids and undergoes dynamic changes during development. Gut. 2019;68(1):49-61.
3. Lam K, Pan K, Linnekamp JF, Medema JP, Kandimalla R. DNA methylation based biomarkers in colorectal cancer: a systematic review. Biochimica et Biophysica Acta (BBA)-Reviews on Cancer. 2016;1866(1):106-20.
4. Li M, D'Arcy C, Li X, Zhang T, Joober R, Meng X. What do DNA methylation studies tell us about depression? A systematic review. Translational psychiatry. 2019;9(1):68.

5. Nano J, Ghanbari M, Wang W, de Vries PS, Dhana K, Muka T, et al. Epigenome-Wide Association Study Identifies Methylation Sites Associated With Liver Enzymes and Hepatic Steatosis. Gastroenterology. 2017;153(4):1096-106.e2.
6. Teroganova N, Girshkin L, Suter CM, Green MJ. DNA methylation in peripheral tissue of schizophrenia and bipolar disorder: a systematic review. BMC genetics. 2016;17(1):27.
7. Delpu Y, Cordelier P, Cho W, Torrisani J. DNA methylation and cancer diagnosis. International journal of molecular sciences. 2013;14(7):15029-58.
8. Figueroa ME, Lugthart S, Li Y, Erpelinck-Verschueren C, Deng X, Christos PJ, et al. DNA methylation signatures identify biologically distinct subtypes in acute myeloid leukemia. Cancer cell. 2010;17(1):13-27.
9. Holm K, Hegardt C, Staaf J, Vallon-Christersson J, Jönsson G, Olsson H, et al. Molecular subtypes of breast cancer are associated with characteristic DNA methylation patterns. Breast cancer research. 2010;12(3):R36.
10. Berdasco M, Esteller M. Clinical epigenetics: seizing opportunities for translation. Nature Reviews Genetics. 2019;20(2):109-27.
11. Mohammad HP, Barbash O, Creasy CL. Targeting epigenetic modifications in cancer therapy: erasing the roadmap to cancer. Nat Med. 2019;25(3):403-18.
12. Dedeurwaerder S, Defrance M, Bizet M, Calonne E, Bontempi G, Fuks F. A comprehensive overview of Infinium HumanMethylation450 data processing. Briefings in Bioinformatics. 2013;15(6):929-41.
13. Pidsley R, Zotenko E, Peters TJ, Lawrence MG, Risbridger GP, Molloy P, et al. Critical evaluation of the Illumina MethylationEPIC BeadChip microarray for whole-genome DNA methylation profiling. Genome Biol. 2016;17(1):208-.
14. Wang JZ, Pabon C, Napier M. Agilent SureSelectXT Methyl-Seq Applications with Low-Input DNA and Smaller Capture Libraries 2017 [Available from: https://www.agilent.com/cs/library/applications/5991-7838EN.pdf.
15. Teh AL, Pan H, Lin X, Lim YI, Patro CP, Cheong CY, et al. Comparison of Methyl-capture Sequencing vs. Infinium 450K methylation array for methylome analysis in clinical samples. Epigenetics. 2016;11(1):36-48.
16. Sun Z, Cunningham J, Slager S, Kocher JP. Base resolution methylome profiling: considerations in platform selection, data preprocessing and analysis. Epigenomics. 2015;7(5):813-28.
17. Heiss JA, Brennan KJ, Baccarelli AA, Tellez-Rojo MM, Estrada-Gutierrez G, Wright RO, et al. Battle of epigenetic proportions: comparing Illumina's EPIC methylation microarrays and TruSeq targeted bisulfite sequencing. Epigenetics. 2020;15(1-2):174-82.
18. Wreczycka K, Gosdschan A, Yusuf D, Gruning B, Assenov Y, Akalin A. Strategies for analyzing bisulfite sequencing data. J Biotechnol. 2017;261:105-15.
19. Krueger F, Andrews SR. Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications. bioinformatics. 2011;27(11):1571-2.

20. Zhang X, Hu Y, Justice AC, Li B, Wang Z, Zhao H, et al. DNA methylation signatures of illicit drug injection and hepatitis C are associated with HIV frailty. Nature Communications. 2017;8(1):2243.
21. Lillycrop KA, Garratt ES, Titcombe P, Melton PE, Murray RJ, Barton SJ, et al. Differential SLC6A4 methylation: a predictive epigenetic marker of adiposity from birth to adulthood. International Journal of Obesity. 2019;43(5):974-88.
22. Williams L, Bei Y, Church HE, Dai N, Dimalanta ET, Ettwiller LM, et al. Enzymatic Methyl-seq: the next generation of methylome analysis. NEB expressions. 2019.

Figures

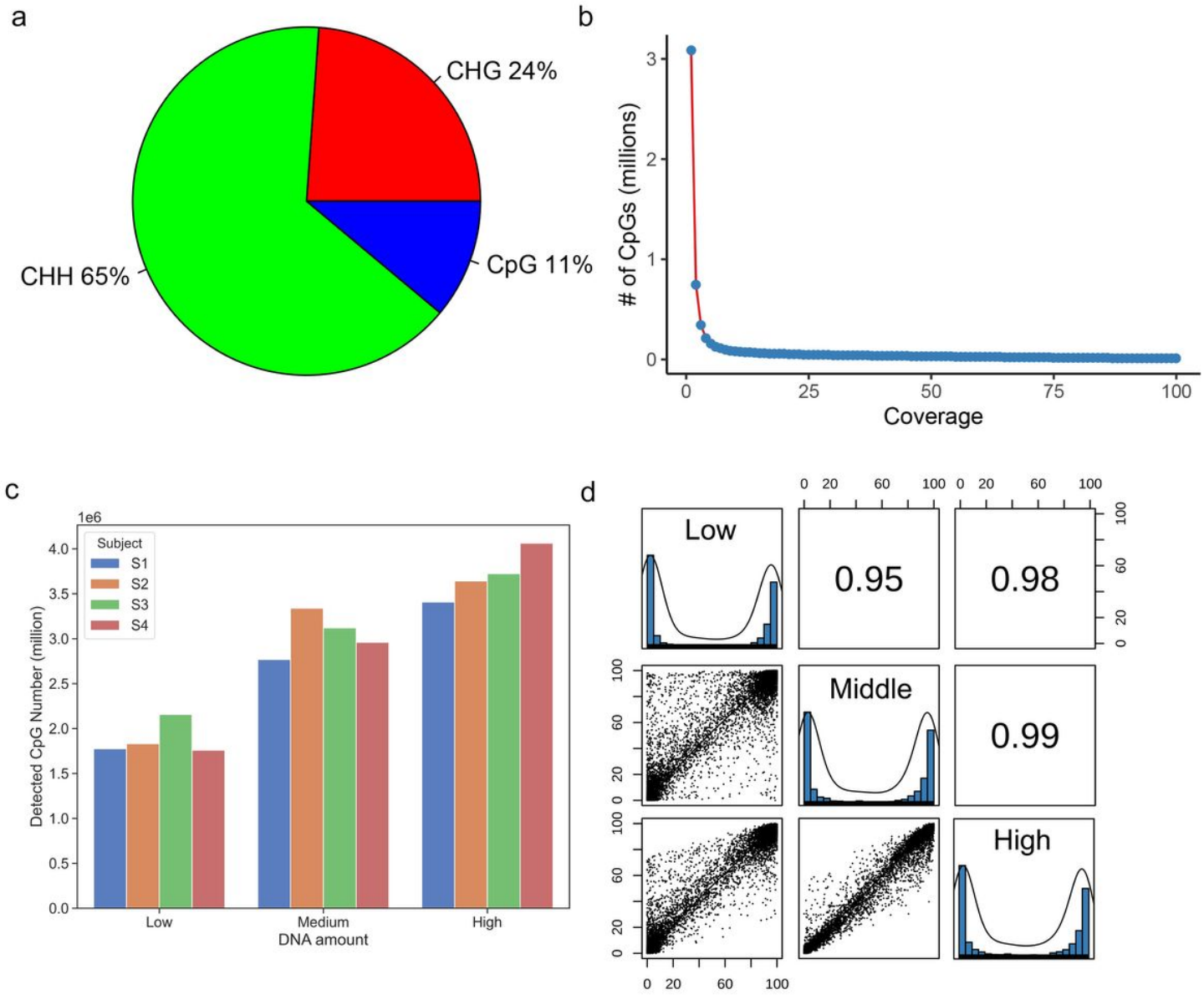

Figure 1

Methylation Capture Sequencing (MC-seq). a. Distribution of methylation sequence context (CpG, CHH, CHG); b. Coverage depth versus a number of detected CpG sites; c. Detected CpG sites in low, medium, and high DNA inputs for four participants using MC-seq with minimum coverage ≥10X; d. Scatterplots comparing 10,000 randomly selected common CpG sites among samples with high, medium and low DNA input quantity and their Pearson correlations.

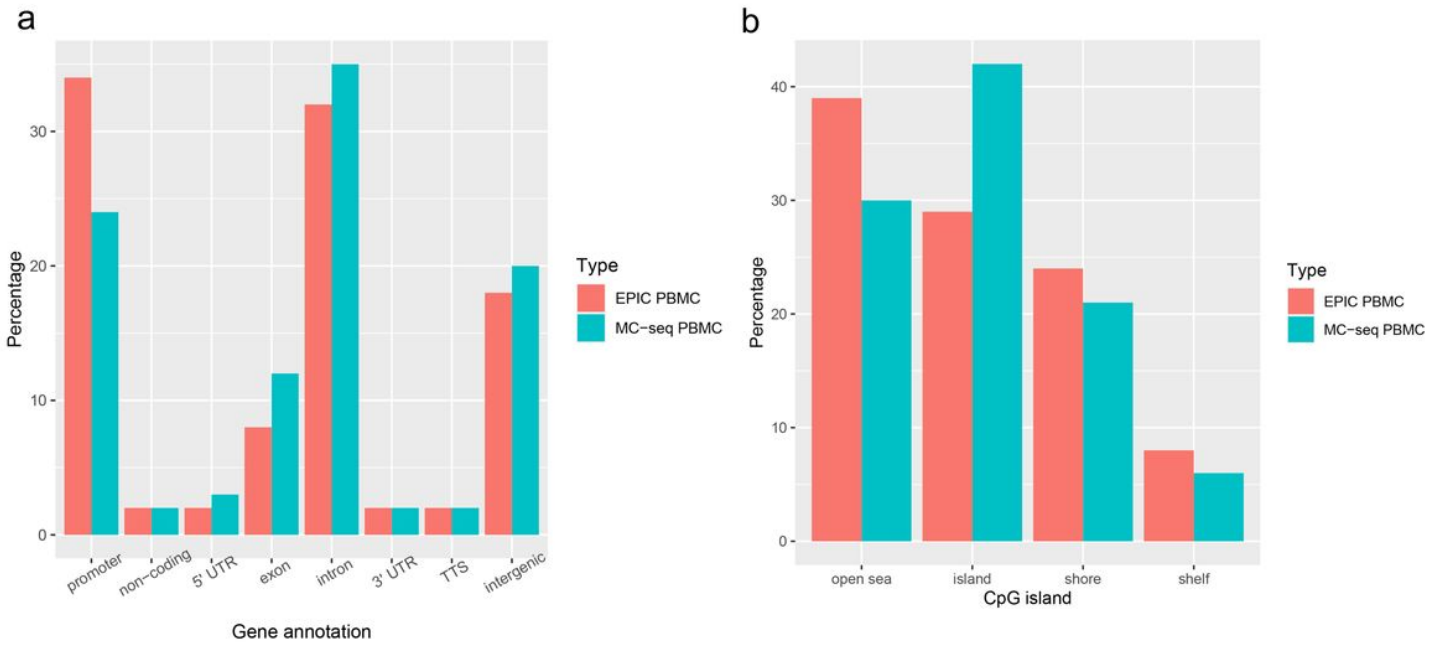

Figure 2

Comparison of CpG proportion in epigenomic regions between MC-seq and EPIC. a. Distribution of genomic regions (intergenic, promoter, 5'UTR, exon, intron, non-coding, 3'UTR, transcription termination site (TTS), non-coding) b. Distribution of CpG position relative to CpG islands (CpG island, shore, shelf, open sea,).

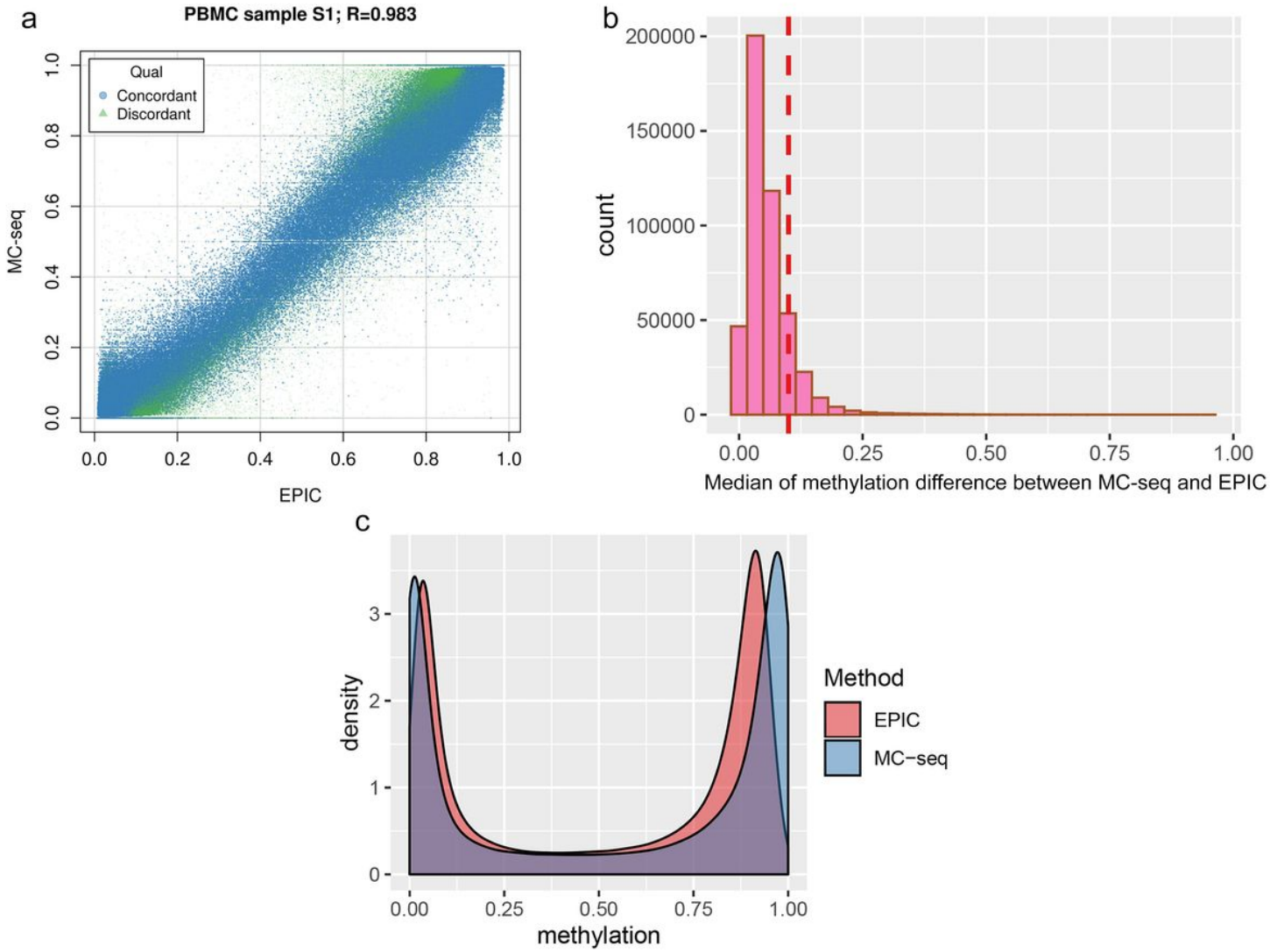

Figure 3

Comparing methylation values among common CpG sites between MC-seq and EPIC. a. Correlation of methylation values measured by MC-seq and EPIC array among common CpG sites in participant S1. Blue dots represent concordant CpGs with □□< 0.1 between the two platforms and green dots represent discordant quality with □□ ≥0.1; b. The distribution of median □□ in common CpG sites between MC-seq and EPIC array. The red dotted line represents □□=0.1 as a cutoff for concordant CpG site between two platforms; c. The density plot of methylation values among common CpG sites profiled by MC-seq and EPIC array in participant S1.

Supplementary Files

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

- Table05282020supp.xlsx
- FigureS1.pdf

- Figure2final.pdf
- FigureS3.pdf