

RESEARCH PAPER

 OPEN ACCESS

Comparison of Methyl-capture Sequencing vs. Infinium 450K methylation array for methylome analysis in clinical samples

Ai Ling Teh^{a,*}, Hong Pan^{a,b,*}, Xinyi Lin^{a,*}, Yubin Ives Lim^{a,c}, Chinari Pawan Kumar Patro^a, Clara Yujing Cheong^a, Min Gong^a, Julia L. Maclsaac^e, Chee-Keong Kwoh^b, Michael J. Meaney^{a,d}, Michael S. Kobor^e, Yap-Seng Chong^{a,c}, Peter D. Gluckman^{a,f}, Joanna D. Holbrook^a, and Neerja Karnani^{a,c}

^aSingapore Institute for Clinical Sciences, A*STAR, Singapore; ^bSchool of Computer Engineering, Nanyang Technological University, Singapore; ^cYong Loo Lin School of Medicine, National University of Singapore, Singapore; ^dLudmer Center for Neuroinformatics & Mental Health, Douglas University Mental Health Institute, McGill University, Montreal, Quebec Canada; ^eCentre for Molecular Medicine and Therapeutics, Child and Family Research Institute, Department of Medical Genetics, University of British Columbia, Vancouver, BC, Canada; ^fCentre for Human Evolution, Adaptation and Disease, Liggins Institute, University of Auckland, Auckland, New Zealand

ABSTRACT

Interindividual variability in the epigenome has gained tremendous attention for its potential in pathophysiological investigation, disease diagnosis, and evaluation of clinical intervention. DNA methylation is the most studied epigenetic mark in epigenome-wide association studies (EWAS) as it can be detected from limited starting material. Infinium 450K methylation array is the most popular platform for high-throughput profiling of this mark in clinical samples, as it is cost-effective and requires small amounts of DNA. However, this method suffers from low genome coverage and errors introduced by probe cross-hybridization. Whole-genome bisulfite sequencing can overcome these limitations but elevates the costs tremendously. Methyl-Capture Sequencing (MC Seq) is an attractive intermediate solution to increase the methylome coverage in large sample sets. Here we first demonstrate that MC Seq can be employed using DNA amounts comparable to the amounts used for Infinium 450K. Second, to provide guidance when choosing between the 2 platforms for EWAS, we evaluate and compare MC Seq and Infinium 450K in terms of coverage, technical variation, and concordance of methylation calls in clinical samples. Last, since the focus in EWAS is to study interindividual variation, we demonstrate the utility of MC Seq in studying interindividual variation in subjects from different ethnicities.

ARTICLE HISTORY

Received 16 September 2015
Revised 4 December 2015
Accepted 10 December 2015

KEYWORDS

DNA methylation;
Epigenome-wide association study; Infinium HumanMethylation450 array; Methyl-Capture Sequencing; next generation sequencing

Introduction

Epigenetic modifications involve heritable and *de novo* changes in chromatin structures, and can sometimes be influenced by the underlying DNA sequence. DNA methylation is one of the most extensively studied epigenetic modifications and involves the addition of a methyl group to the fifth carbon of the cytosine nucleotide. DNA methylation patterns are relatively stable and conserved during cellular division, but can be altered significantly through development, metabolic disorders and pathologies, or by external effectors such as nutrition, pollution, and stress.^{1–3} As DNA methylation marks reflect both environmental and genetic influences,^{4–7} they provide insight into how extrinsic and intrinsic stimuli can interact to induce physiological or pathological changes. Epigenome-wide association studies (EWAS), in which DNA methylation profiles of large numbers of clinical samples are assayed, are being increasingly employed to uncover biological mechanisms that underlie health outcomes.^{8–11}

The success of EWAS interrogating DNA methylation marks faces two main challenges. First, unlike genome-wide association studies (GWAS), these marks are generally tissue


specific and thus assaying clinical samples relevant to the disease is critical. Second, the choice of assay platform to employ, which is influenced by the trade-off between cost-effectiveness and genomic coverage, can influence the findings from the study. Practical and ethical limitations for collection of relevant tissue biopsies have led to the use of peripheral tissue surrogates such as buccal swabs and blood samples. Buccal swabs are generally preferred as they are relatively homogenous in cellular composition, and their collection procedure is non-invasive and hence extendible to pediatric cohorts. Blood profiles are dominated by leucocyte DNA, which are more labile to external influences such as acute infection. In addition, Lowe et al.¹² showed that buccal is a more informative surrogate tissue in non-blood related diseases/phenotypes. This observation has also been reiterated in the work of Smith et al.¹³ that highlighted the advantages of methylome profiling of saliva over blood in EWAS of psychiatric traits.

Many methods are available to interrogate DNA methylation in clinical samples at single-base resolution. These methods can be broadly classified into 2 categories: microarray- and

CONTACT Neerja Karnani  Neerja_Karnani@sics.a-star.edu.sg

Present Address: C.P.K. Patro, Social and Cognitive Computing Department, Institute of High Performance Computing, A*STAR, Singapore

*These authors contributed equally to this work.

 Supplemental data for this article can be accessed on the publisher's website.

Published with License by Taylor & Francis Group, LLC © Ai Ling Teh, Hong Pan, Xinyi Lin, Yubin Ives Lim, Chinari Pawan Kumar Patro, Clara Yujing Cheong, Min Gong, Julia L. Maclsaac, Chee-Keong Kwoh, Michael J. Meaney, Michael S. Kobor, Yap-Seng Chong, Peter D. Gluckman, Joanna D. Holbrook, and Neerja Karnani

This is an Open Access article distributed under the terms of the Creative Commons Attribution-Non-Commercial License (<http://creativecommons.org/licenses/by-nc/3.0/>), which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited. The moral rights of the named author(s) have been asserted.

next-generation sequencing (NGS)-based. Microarray-based technologies, such as Infinium HumanMethylation27¹⁴ and Infinium HumanMethylation450 (Infinium 450K),¹⁵ use a fixed number of probes to survey specific genomic loci across the genome. Infinium 450K has thus far been the most widely used method in EWAS due to its low cost, modest DNA requirement, and significantly reduced sample processing time, making high-throughput processing of large numbers of clinical samples possible. However, microarray-based platforms are unable to expand beyond genomic regions dictated by both the number and specificity of probes, thereby limiting the exhaustive screening of the genome for epigenetically altered loci. Of the different next generation sequencing (NGS)-based technologies, whole-genome bisulfite conversion (WGBS)¹⁶⁻¹⁸ is generally regarded as the gold standard because it provides the highest genomic coverage; however, its substantial cost and processing time renders it unfeasible in EWAS where large numbers of samples have to be assayed. To reduce the cost and processing time, NGS alternatives to WGBS use various approaches to target specific loci in the epigenome. Reduced-representation bisulfite sequencing (RRBS) platforms^{19,20} use cytosine methylation-specific restriction enzyme digestion; affinity-enrichment platforms, such as Methylated DNA Immunoprecipitation Sequencing (MeDIP-Seq),²¹⁻²³ utilize methylation-specific antibodies to extract methylated-cytosine-containing DNA fragments after DNA fragmentation. Compared to RRBS platforms, MeDIP-Seq allows broader coverage of the genome but fails to provide methylation measures at base-pair resolution. Both platforms, however, introduce bias toward CpG-rich repeats.^{24,25} Moreover, both technologies are limited in regions surveyed due to the available restriction digestion enzymes or antibodies. Another affinity-enrichment platform, Methyl-Capture Sequencing (MC Seq) utilizes target-specific bait sequences, circumventing this restriction through bait design and allows for epigenome-wide surveying of specific genomic loci of physiological and clinical interest. MC Seq thus presents an attractive cost-effective alternative to uncover novel disease-associated genomic loci in EWAS, and overcomes the limitations of lower genome coverage (Infinium 450K), high cost and processing time (WGBS), while avoiding overrepresentation of repeated (RRBS) and methylated regions (MeDIP-Seq).

The performance of MC Seq has been examined in a few studies. Hing et al.²⁶ demonstrated the use of MC Seq in surveying different tissues in the mouse genome. Li et al.²⁷ provided a comprehensive evaluation of the technical performance of MC Seq, illustrating the method's accuracy through comparison with WGBS and reproducibility of methylation values across technical replicates in maize samples and human cell lines. They also used buccal epithelial samples and showed that MC Seq can detect allelic DNA methylation at selected imprinted DMRs. However, no evaluation was performed on other MC Seq methylation sites, nor did they examine interindividual variation, which is the main variable of interest in EWAS, and likely to require higher resolution, as interindividual differences are usually smaller than those of allelic imbalance. Allum et al.²⁸ recently reported the utility of MC Seq in investigating interindividual variation using an adipose tissue customized panel. An EWAS study of plasma triglyceride levels

was performed using adipose tissue derived from obese subjects with comparison of methylation values from MC Seq with those generated by WGBS or Infinium 450K.

There are 3 major issues that remain to be addressed in considering the utility of the MC Seq approach for studies that relate variation in DNA methylation to phenotype. First, no study has yet examined the performance of MC Seq in investigating interindividual variation from routinely used clinical samples such as buccal swabs. Second is the issue of the required amount of DNA material. Hing et al.²⁶ utilized a capture-then-bisulfite-convert approach (Agilent SureSelect Human Methyl-Seq), using 3 μg of starting DNA, while Li et al.²⁷ and Allum et al.²⁸ employed a bisulfite-convert-then-capture assay (Roche NimbleGen), using as little as 500 ng of DNA.²⁷ Both Li et al.²⁷ and Allum et al.²⁸ also compared the performance of a capture-then-bisulfite-convert approach using 3 μg of starting DNA with that of a bisulfite-convert-then-capture assay. Thus, the feasibility of employing a capture-then-bisulfite-convert approach with MC Seq using smaller quantities of genomic DNA has not been investigated either. Finally, there is, as yet, no direct comparison between MC Seq employing a capture-then-bisulfite-convert approach with Infinium 450K. This study provides a comparison of methods to help choose an appropriate platform for future EWAS. We first evaluate the technical performance of MC Seq in human buccal DNA samples. Second, we provide a comprehensive comparison of Infinium 450K and MC Seq, across key functional genomic regions in context of their coverage and concordance of methylation calls in clinical samples. Last, as proof-of-principle, in a small sample, we examine the utility of MC Seq in studying interindividual variation and demonstrate that methylation values from MC Seq can distinguish subjects from different ethnic groups. We find that the performance of MC Seq with a capture-then-bisulfite-convert approach, using either 1 μg or 3 μg of DNA, was similar; hence, MC Seq can be used with starting DNA quantity comparable to the requirement for Infinium 450K. We also show that, compared to Infinium 450K, MC Seq provides increased coverage of the epigenome and, hence, the detection of more genomic sites showing interindividual variation. However, the application of MC Seq to EWAS with small effect sizes will only be feasible if the inter-group differences exceed the technical variation.

Results

Overview of MC Seq analysis of human buccal epithelial samples

The DNA methylome of 7 buccal epithelium samples were generated using MC Seq with a capture-then-bisulfite-convert approach with 3 μg of genomic DNA. Mapping efficiency, sequence duplication rates and sequence bait specificities are summarized in Table 1. On an average, 38 million pair-end reads were generated per sample, of which 34 million aligned uniquely to the bisulfite-converted human reference genome (hg19/GRCH37). Twenty-eight million reads remained after removal of duplicated reads; 93% and 97% of the reads were found within the target region or within 200 bases of the target region, respectively. Bait capture efficiency was high as reads

Table 1. Summary of sequence alignment and duplicate rates for 7 buccal epithelium samples from MC Seq.

	F1_Ind	F2_Ind	F3_Cau	F4_Chi	F5_Chi	F6_Chi	M7_Chi	Average
Raw sequence reads	32,974,840	89,531,096	34,696,225	21,464,654	41,325,121	52,461,681	36,269,247	44,103,266
Sequence pairs analysed in total	27,882,564	77,296,641	29,464,393	18,066,637	34,755,956	44,321,376	30,962,012	37,535,654
Number of paired-end alignments with a unique best hit	25,782,034	65,302,698	27,722,823	16,721,943	32,445,266	40,018,396	29,287,695	33,897,265
Mapping efficiency (%)	92.5	84.5	94.1	92.6	93.4	90.3	94.6	91.7
Duplicate (%)	7.04	45.76	5.74	5.12	6.42	7.72	5.31	11.87
Sequence pairs after removing duplicate	23,966,529	35,419,351	26,131,257	15,865,465	30,362,840	36,928,166	27,733,649	28,058,180
Reads in targeted region (%)	93.44	89.64	93.70	93.65	92.95	94.36	93.90	93.09
Reads in targeted regions 200 bp (%)	97.13	92.91	97.01	97.12	96.98	97.35	97.17	96.52

within 200 bases of targeted regions were observed to be >25-fold more abundant than the genomic background.

SureSelect Human Methyl-Seq panel (Agilent Technologies) was designed to capture 3.2 million CpGs within its baits target region, of which 51%, 19%, 5%, and 25% belonged to CpG islands, shores, shelves, and open seas, respectively (Fig. 1a); 25%, 3%, 15%, 34%, 2%, 2%, and 19% of the CpGs belonged to promoter, 5'-UTR, exon, intron, 3'-UTR, TTS, and intergenic regions (Fig. 1b). With 200 bases flanking both sides of the target regions, MC Seq would capture 4.8 million CpGs, with similar CpG content distribution (Fig. 1a) and similar genomic features distribution (Fig. 1b). In our buccal epithelium samples, we detected an average of 2.6 million CpGs with at least 10X coverage, the majority of which (89%) were within the target region. The number of detected CpGs located on shores, shelves, and open seas was comparable to the expected (maximum possible within target region), while the number of detected CpGs belonging to CpG islands was lower than expected (Fig. 1a, Supplementary Fig. 1). The observed distribution of CpGs within genomic features was similar to the expected (Fig. 1b, Supplementary Fig. 2).

Fig. 1c shows the chromosome-wise distribution of CpGs detected by MC Seq, normalized by the number of CpGs in the human epigenome on the chromosome, for all 7 samples. As expected, only the male sample (M7_Chi) showed representation of CpGs on the Y chromosome and this sample had a relatively poorer coverage of the X chromosome than the female samples (Fig. 1c). All samples also showed comparatively lower coverage on chromosome 18 (Fig. 1c), which is a consequence of the bait design (data not shown). In addition to methylation sites within the CpG context, MC Seq also detected 5.7 million and 15.1 million methylation sites within CHG and CHH context with at least 10X reads coverage (Fig. 1d, Supplementary Fig. 3).

Reproducibility of MC Seq

To investigate the technical variation of MC Seq in buccal samples, 4 of the 7 samples were investigated in duplicate. Mapping efficiency, sequence duplication rates, and sequence bait specificities for the 4 replicate samples are summarized in Supplementary Table 1. Sample-based Pearson correlation between replicates was high ($R=0.9812$ with minimum of 10X reads coverage, Fig. 1e, Supplementary Fig. 4). Due to the large number of CpG sites (>2 million) used to calculate the sample-based correlation, it is not uncommon for the sample-based correlation to be high (>0.90). However, a high sample-based correlation might not translate into high probe-based

calculation because the probe-based calculation is calculated using fewer data points (the number of samples performed in replicates). For example, Allum et al.²⁸ observed sample-based correlations exceeding 0.90 in all their comparisons, but the average probe-based correlation was only 0.2. In studying inter-individual variation, the probe-based correlation might be more important than the sample-based correlation. Reports of environmentally driven interindividual variation in methylation have typically detailed changes between 0–5%. The Infinium 450K can detect differences of 20% in methylation with 99% confidence.¹⁵ As a more sensitive measure of technical variation, we report the absolute difference in methylation and the cumulative percentage of probes showing absolute differences in methylation within 5%, 10%, 20%, and 50%, respectively (Fig. 1f, Supplementary Fig. 5). With a minimum of 10X reads coverage, 63%, 21%, 12%, and 3% of probes showed absolute differences in methylation within 0–5%, 5–10%, 10–20%, and 20–50%, respectively (Fig. 1f, Supplementary Fig. 5). The performance was improved at 30X, where 71%, 20%, 9%, and 1% of probes showed absolute differences in methylation within 0–5%, 5–10%, 10–20%, and 20–50%, respectively (Fig. 1f, Supplementary Fig. 5). Further increases in reads coverage improved the performance minimally. Replicates clustered together in hierarchical clustering (Fig. 1g).

Performance of MC Seq at 3 μ g and 1 μ g was similar

To investigate the performance of MC Seq in buccal samples using lower quantity of genomic DNA, 2 (F5_Chi and M7_Chi) of the 7 samples were repeated with MC Seq using 1 μ g of genomic DNA. Mapping efficiency, sequence duplication rates and sequence bait specificities for the 2 samples performed using 1 μ g of genomic DNA are summarized in Supplementary Table 2 and are similar to those performed using 3 μ g of genomic DNA (Table 1). MC Seq performed similarly using lower quantities of genomic DNA (Table 2). The comparison of MC Seq at 1 μ g and 3 μ g (Fig. 2) mirrors the comparison between technical replicates at 3 μ g (Fig. 1e–f), with similar error rates shown in Figs. 1 and 2. Correlation of methylation values derived using 1 μ g and 3 μ g of genomic DNA was high ($R=0.9842$ with minimum of 10X reads coverage, Fig. 2a, Supplementary Fig. 6). With a minimum of 10X coverage, 67%, 20%, 11%, and 3% of probes showed absolute differences in methylation within 0–5%, 5–10%, 10–20%, and 20–50%, respectively (Fig. 2b, Supplementary Fig. 7). Performance was improved at 30X reads coverage, where 74%, 19%, 7%, and 1% of probes showed absolute differences in methylation within 0–5%, 5–10%, 10–20%, and 20–50%, respectively

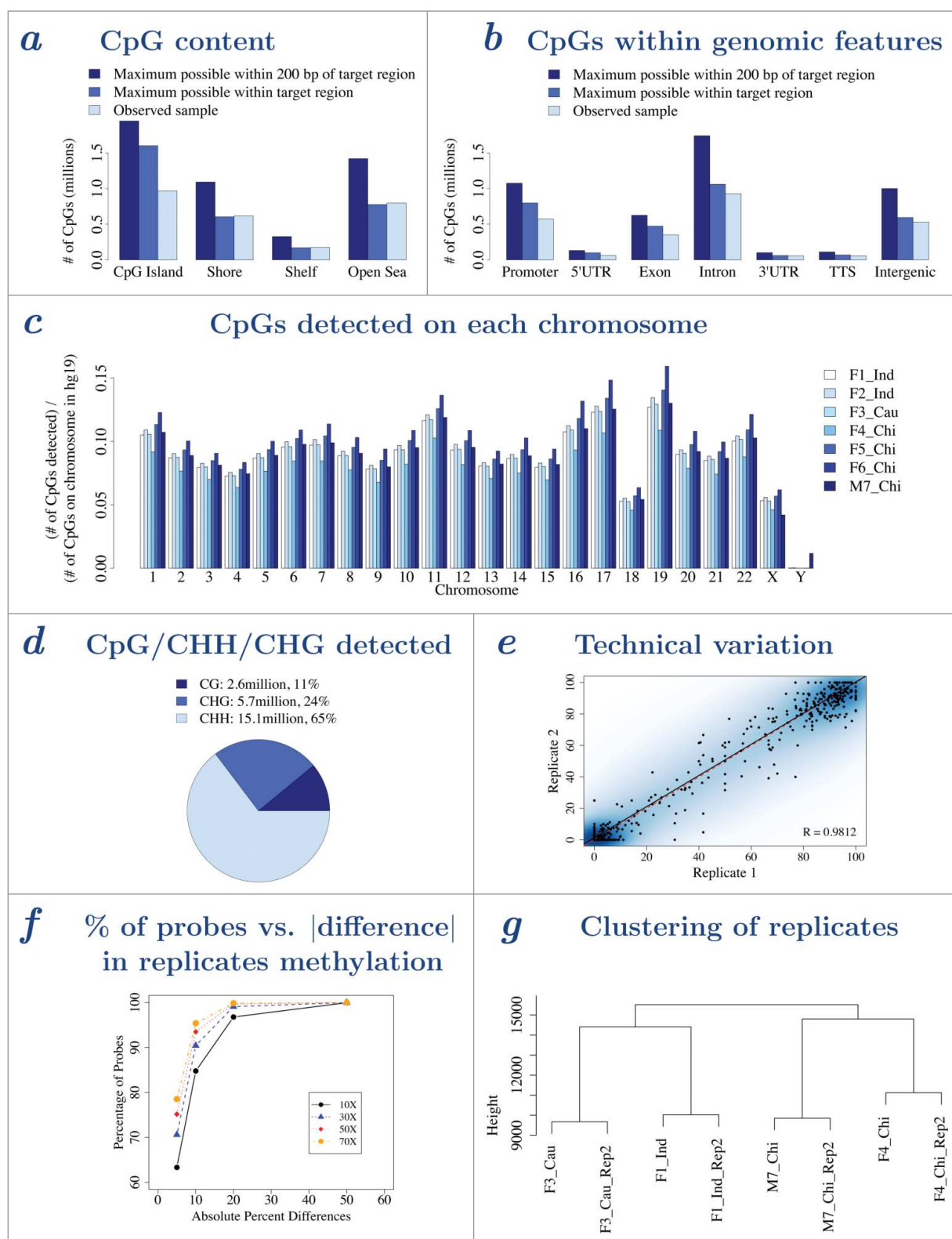


Figure 1. Analysis of MC Seq data. a – CpG content distribution of CpGs (island, shore, shelf, and open sea). Height of bar graphs represent number of CpGs covered by MC Seq, where dark, medium and light blue bars represent maximum possible within 200 bases of target region list, maximum possible within target region list and observed CpGs (for one sample), respectively. b – Functional genomic distribution of CpGs (promoter, 5'-UTR, exon, intron, 3'-UTR, TTS, and intergenic). Height of bar graphs represent number of CpGs covered by MC Seq, where dark, medium and light blue bars represent maximum possible within 200 bases of target region list, maximum possible (within target region list) and observed CpGs (for one sample), respectively. c – Chromosome distribution of CpGs for all 7 buccal samples. Height of bar graphs show number of CpGs detected at each chromosome normalized by number of CpGs on the chromosome in human epigenome (hg19). Male sample M7_Chi shows a peak in normalized number of CpGs detected on Y chromosome. d – Pie chart showing number and percentage of methylation sites detected within CpG, CHG, and CHH context for one sample. e – Pearson correlation and scatterplot of methylation values from MC Seq for replicate 1 (horizontal axis) and replicate 2 (vertical axis) for one sample. Color represents density of CpG sites, with darker blue indicating higher density of CpG sites and lighter blue indicating lower density of CpG sites. Five hundred randomly selected CpG sites are shown as black points. Dotted line gives $y=x$ line, solid line gives best-fit line; overlapping lines indicate high concordance between replicates. f – Cumulative percentage of probes (vertical axis) vs. absolute difference in methylation between replicates (horizontal axis), at $\geq 10X$ (solid line), $\geq 30X$ (dashed line), $\geq 50X$ (dotted line) and $\geq 70X$ (dotted-dashed line) reads coverage, for one sample. g – Hierarchical clustering analysis of replicates show that replicates cluster together. Corresponding plots for a-b and d-g for other samples are provided in Supplementary Figs. 1–5.

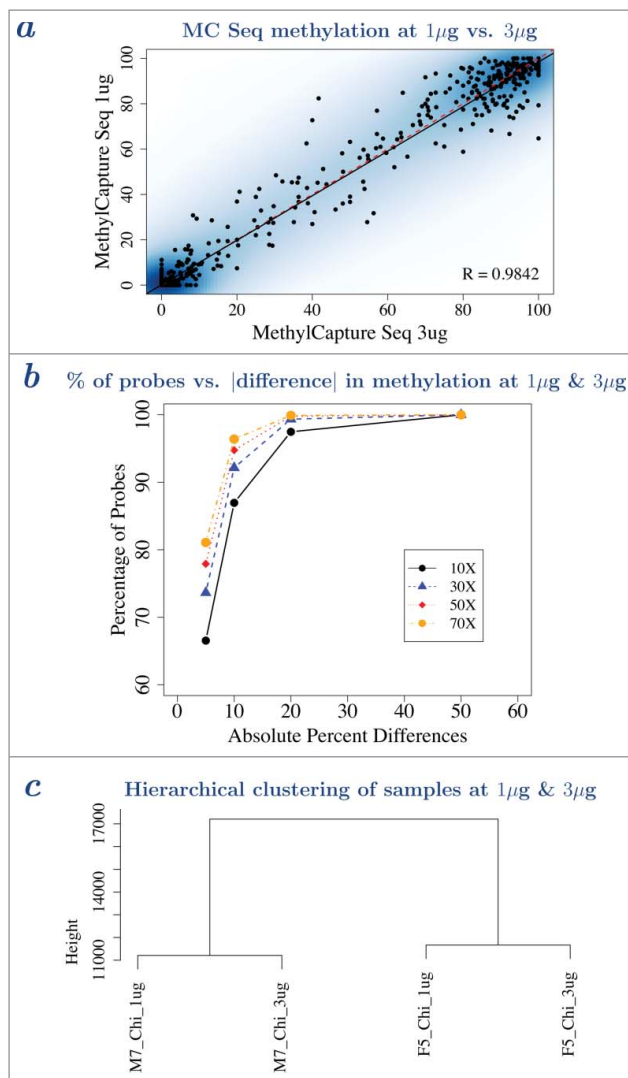


Figure 2. Performance of MC Seq at 3 μg and 1 μg were similar. a – Pearson correlation and scatterplot of methylation values from MC Seq at 3 μg (horizontal axis) and 1 μg (vertical axis) for one sample. Color represents density of CpG sites, with darker blue indicating higher density of CpG sites and lighter blue indicating lower density of CpG sites. Five hundred randomly selected CpG sites are shown as black points. Dotted line gives $y=x$ line, solid line gives best-fit line; overlapping lines indicate high concordance at 3 μg and 1 μg . b – Cumulative percentage of probes (vertical axis) vs. absolute difference in methylation between 3 μg and 1 μg (horizontal axis), at $\geq 10\text{X}$ (solid line), $\geq 30\text{X}$ (dashed line), $\geq 50\text{X}$ (dotted line) and $\geq 70\text{X}$ (dotted-dashed line) reads coverage, for one sample. c – Hierarchical clustering analysis shows that corresponding samples at 3 μg and 1 μg cluster together. Corresponding plots for a-b for other samples are provided in Supplementary Figs. 6–7.

Table 2. Comparison of performance of MC Seq using 1 μg and 3 μg of genomic DNA. First two columns give number of CpGs observed at 1 μg and 3 μg , third column gives the number of common CpGs observed at both. Last two columns give the correlation between methylation values at 1 μg and 3 μg .

Using only CpG sites with methylation in range 1–99% (reads coverage $\geq 10\text{X}$)					
Sample ID	MC Seq 1 μg	MC Seq 3 μg	Common	Pearson R	Spearman R
F5_Chi	2,394,229	1,955,175	1,754,997	0.9658	0.9286
M7_Chi	2,069,434	1,765,358	1,523,200	0.9713	0.9250
Average	2,231,832	1,860,267	1,639,099	0.9686	0.9268
Using CpG sites with methylation in range 0–100% (reads coverage $\geq 10\text{X}$)					
Sample ID	MC Seq 1 μg	MC Seq 3 μg	Common	Pearson R	Spearman R
F5_Chi	3,283,332	2,761,030	2,738,079	0.9808	0.9328
M7_Chi	2,964,101	2,601,001	2,549,734	0.9842	0.9223
Average	3,123,717	2,681,016	2,643,907	0.9825	0.9275

(Fig. 2b, Supplementary Fig. 7). Further increase in reads coverage to 50X showed only slight improvement, where 78%, 17%, 5% and 0.2% of probes showed absolute differences in methylation within 0–5%, 5–10%, 10–20%, 20–50%, respectively (Fig. 2b, Supplementary Fig. 7). Furthermore, hierarchical clustering analysis performed on the 4 samples (F5_Chi at 1 μg and 3 μg and M7_Chi at 1 μg and 3 μg) showed that the corresponding 1 μg and 3 μg samples clustered together (Fig. 2c).

Methylation values from MC Seq and Infinium 450K were highly correlated and both gave a bimodal distribution

To compare MC Seq and Infinium 450K, the same 7 buccal samples were also interrogated using Infinium 450K, with replicates for 4 of the 7 samples (Table 3). Technical replicates showed high concordance in methylation values, and more than 99% of probes passed quality control, indicating high performance of Infinium 450K (Supplementary Fig. 8, Supplementary Table 3). The sample-based correlation between methylation values from MC Seq and Infinium 450K was high and increased slightly with increasing coverage ($R=0.9776$, 0.9819, and 0.9840 at 10X, 30X, and 50X, respectively), while the number of CpGs detected decreased significantly (Number of CpGs = 2.6, 1.3, and 0.6 million at 10X, 30X, and 50X, respectively) with increasing coverage (Fig. 3a–c, Supplementary Figs. 9–11). With a minimum of 10X coverage, 57%, 24%, 15%, and 4% of sites showed 0–5%, 5–10%, 10–20%, and 20–50% absolute differences in methylation between the 2 platforms, respectively (Fig. 3d, Supplementary Fig. 12). Concordance between MC Seq and Infinium 450K was slightly improved at 30X, where 60%, 23%, 14%, and 3% of probes showed absolute differences in methylation within 0–5%, 5–10%, 10–20%, and 20–50%, respectively (Fig. 3d, Supplementary Fig. 12). The distribution of methylation values from both platforms followed a bimodal distribution, with the range for methylation values derived from Infinium 450K slightly condensed compared to those derived from MC Seq (Fig. 3e–f, Supplementary Figs. 13–14).

Methylation values from both MC Seq and Infinium 450K distinguish interindividual variation

A key interest in EWAS is to identify methylation sites that show appreciable interindividual variation with the outcome/phenotype

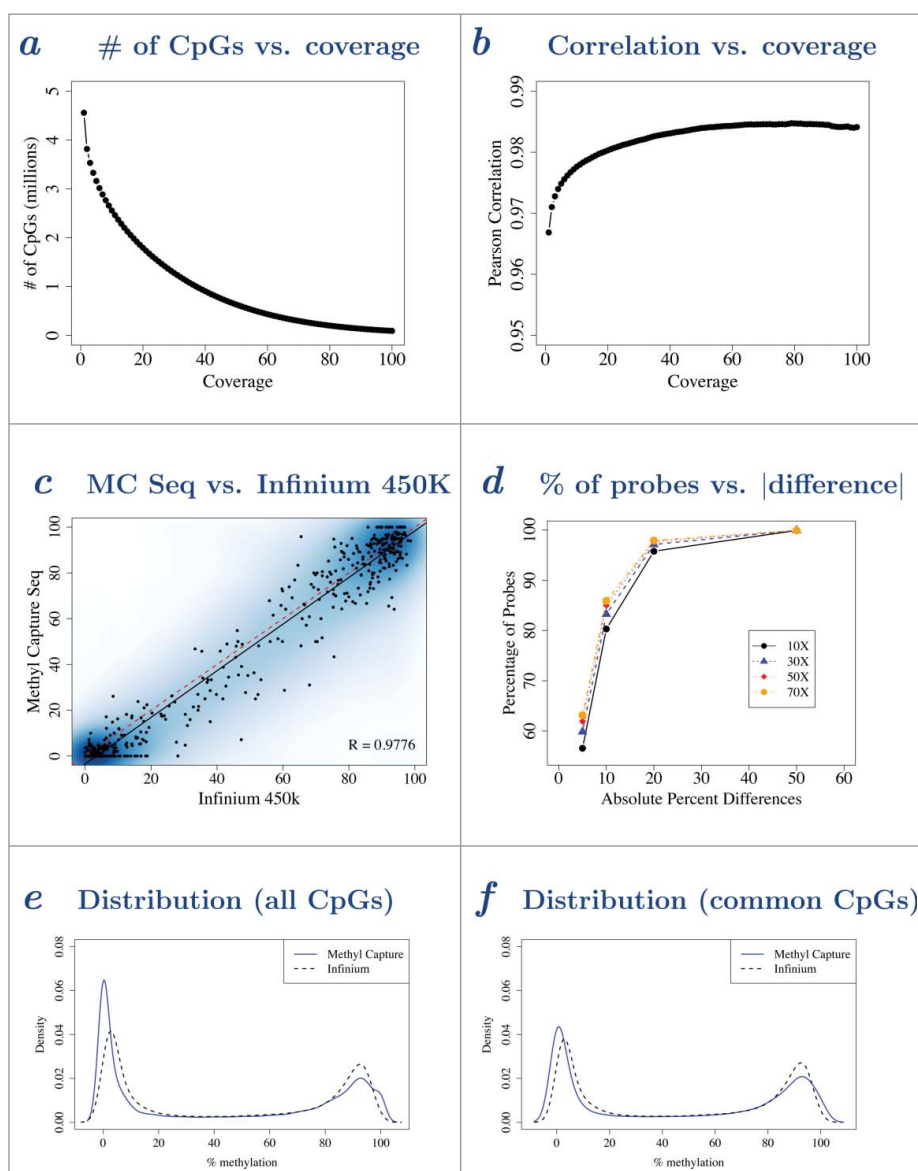


Figure 3. Methylation values from MC Seq and Infinium 450K were highly correlated and both gave a bimodal distribution. a – Observed number of CpGs (vertical axis) from MC Seq for one sample at different MC Seq reads coverage (horizontal axis). As reads coverage increases (left to right), number of CpGs decreases (top to bottom). b – Pearson correlation (vertical axis) between methylation values from MC Seq and Infinium 450K at the same CpG sites, at different MC Seq reads coverage (horizontal axis) for one sample. As reads coverage increases (left to right), Pearson correlation increases (bottom to top). c – Scatterplot of methylation values from MC Seq ($\geq 10X$, vertical axis) and Infinium 450K (horizontal axis) at the same CpG sites for one sample. Color represents density of CpG sites, with darker blue indicating higher density of CpG sites and lighter blue indicating lower density of CpG sites. Five hundred randomly selected CpG sites are shown as black points. Dotted line gives $y=x$ line, solid line gives best-fit line; parallel lines indicate high correlation between methylation values from the 2 platforms; slight vertical shift indicates a small systematic bias. d – Cumulative percentage of probes (vertical axis) vs. absolute difference in methylation between MC Seq and Infinium 450K (horizontal axis), at $\geq 10X$ (solid line), $\geq 30X$ (dashed line), $\geq 50X$ (dotted line) and $\geq 70X$ (dotted-dashed line) reads coverage, for one sample. e – Distribution of methylation values for *all* CpGs from MC Seq ($\geq 10X$, solid line) and Infinium 450K (dotted line) for one sample. f – Distribution of methylation values for *common* CpGs from MC Seq ($\geq 10X$, solid line) and Infinium 450K (dotted line) for one sample. Corresponding plots for other samples are provided in Supplementary Figs. 9–14.

of interest. As a proof-of-principle to demonstrate that MC Seq can detect significant interindividual variation that is related to phenotype, we clustered our 7 multi-ethnic samples using unsupervised hierarchical clustering. As it has been previously reported that the most variable CpGs can be influenced by genotype,²⁹ we performed clustering analysis using only the most variable CpGs (CpG sites with interquartile range $> 20\%$). MC seq detected 3.7 times more variable CpG sites than Infinium 450K (7,880 CpG sites for MC Seq at $\geq 30X$ reads coverage vs. 2130 CpG sites for Infinium 450K), indicating the gain in genomic coverage. Both platforms clustered samples by ethnicity (Fig. 4) with generally high

confidence,³⁰ and the results were robust to the clustering method and distance metric used (Supplementary Figs. 15–16). For MC Seq, 4393 (56%) and 4712 (60%) out of 7,880 CpGs were located within 1 bp and 10 bp of SNPs, while 1178 (55%) and 1290 (61%) out of 2130 CpGs were located within 1 bp and 10 bp of SNPs for Infinium 450K. We further annotated these 7,880 CpGs (for MC Seq) and 2130 CpGs (for Infinium 450K), and compared them to the overall distribution of CpGs assayed by the 2 platforms (Supplementary Figs. 17–18). For both platforms, these highly variable CpGs were more likely to be located in intronic and intergenic regions, and were less likely to be located in promoter and exon

Table 3. Comparison of MC Seq and Infinium 450K.

Sample ID	%meth 0–100 (reads coverage $\geq 10X$)			%meth 1–99 (reads coverage $\geq 10X$)		
	Number of CpGs shared with Infinium	Pearson R	Spearman R	Number of CpGs shared with Infinium	Pearson R	Spearman R
F1_Ind	334,056	0.9765	0.9187	242,915	0.9653	0.9085
F2_Ind	335,875	0.9747	0.9244	250,070	0.9638	0.9194
F3_Cau	336,496	0.9743	0.9268	250,050	0.9620	0.9178
F4_Chi	306,293	0.9679	0.9195	211,389	0.9487	0.9027
F5_Chi	348,414	0.9719	0.9253	262,547	0.9589	0.9126
F6_Chi	365,053	0.9767	0.9276	277,924	0.9670	0.9221
M7_Chi	338,700	0.9802	0.9177	245,214	0.9709	0.9122
Average	337,841	0.9746	0.9229	248,587	0.9624	0.9136
Common	291,087					

regions (Supplementary Fig. 17). They were also more likely to be located in CpG shelves and open seas, and less likely to be in CpG islands (Supplementary Fig. 18).

MC Seq provides denser coverage of the epigenome

We further compared MC Seq and Infinium 450K in context of their coverage of the epigenome (Table 4, Fig. 5, Supplementary Figs. 19–20). There are 28 million CpGs in the human

epigenome (hg19). As mentioned above, MC Seq was designed to assay up to 3.2 million CpGs within the target regions (11% of all CpGs in epigenome). In our buccal epithelia samples, we detected an average of 2.6 million CpGs (9% of all CpGs in epigenome) and 1.4 million CpGs (5% of all CpGs in epigenome)

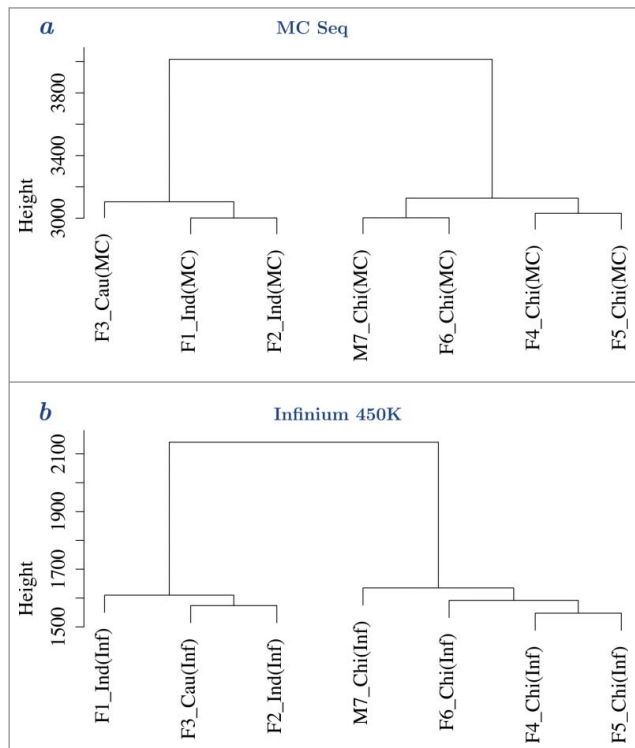


Figure 4. Hierarchical clustering analysis of methylation values showed clinical samples clustered by ethnicity. a – Hierarchical clustering analysis of all 7 samples profiled using MC Seq, using most variably methylated probes, e.g., probes with interquartile range $>20\%$ (autosomal sites). Clustering was performed using Euclidean distance and “ward.D” method in R. b – Hierarchical clustering analysis of all 7 samples profiled using Infinium 450K, using most variably methylated probes, e.g., probes with interquartile range $>20\%$ (autosomal and non cross-reactive sites). Clustering was performed using Euclidean distance and “ward.D” method in R. Hierarchical clustering analysis using other distance metrics and agglomeration methods, are reported with their approximately unbiased p-values, in Supplementary Figs. 15–16.

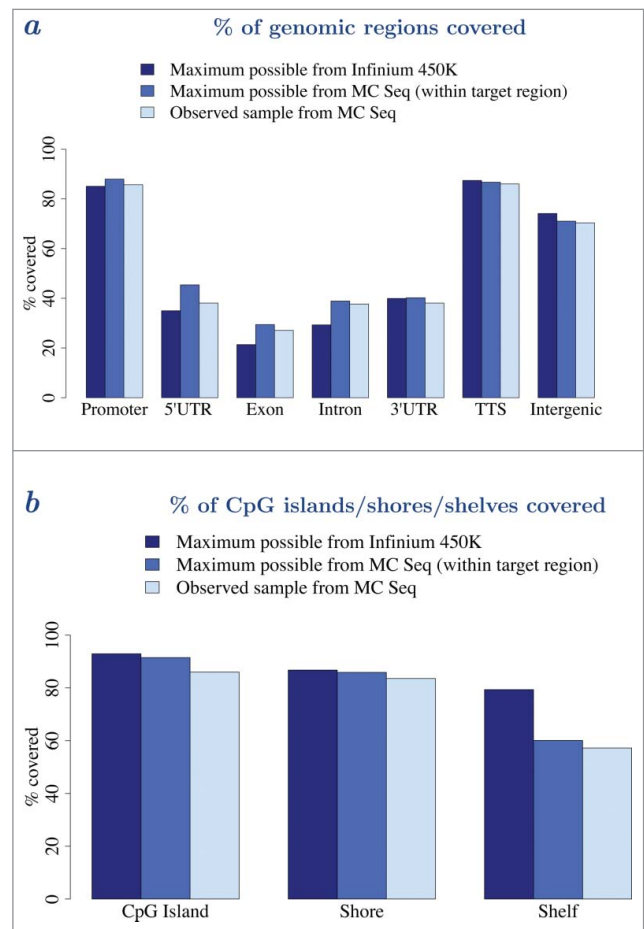


Figure 5. MC Seq provides denser coverage of the epigenome. a – Genomic coverage (percentage covered) of unique genes (promoter, 5'-UTR, exon, intron, 3'-UTR, TTS, and intergenic regions) by Infinium 450K (dark blue), MC Seq, maximum possible (medium blue), and MC Seq, observed for one sample at $\geq 10X$ (light blue), respectively. b – CpG coverage (percentage covered) of CpG islands, shores, and shelves, by Infinium 450K (dark blue), MC Seq, maximum possible (medium blue), and MC Seq, observed for one sample at $\geq 10X$ (light blue), respectively. Corresponding plots for other samples are provided in Supplementary Figs. 19–20.

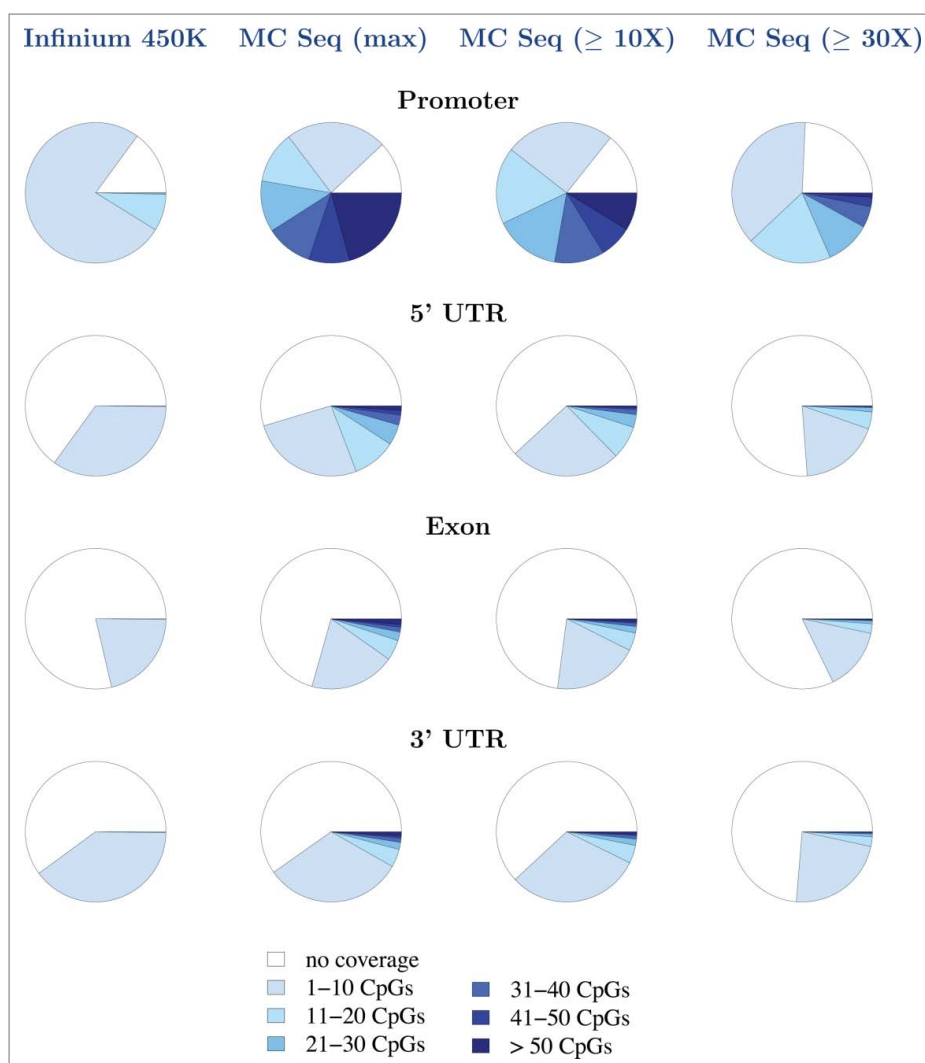


Figure 6. MC Seq provides denser coverage of the epigenome. Genomic coverage (*density of coverage*) of unique genes (promoter, 5'-UTR, exon, and 3'-UTR) by Infinium 450K (first column), MC Seq, maximum possible (second column), MC Seq, observed for one sample at $\geq 10X$ (third column), and MC Seq, observed for one sample at $\geq 30X$ (fourth column), respectively. Density of coverage for remaining regions (intron, TTS, and intergenic regions; CpG islands, shores, and shelves) are provided in Supplementary Figs. 21–22.

with at least 10X and 30X coverage, respectively. By contrast Infinium 450K arrays assay only 1.7% of the CpGs present in the human epigenome. MC Seq platform includes 395,817 (82%) of the 482,421 CpGs on Infinium 450K. Both platforms

provided coverage for 91–93% of all 29K CpG islands in the human epigenome. Infinium 450K covered 87% and 79% of CpG shores and shelves respectively, while MC Seq covered 86% and 60% of CpG shores and shelves, respectively (Fig. 5b,

Table 4. Summary of CpGs detected by MC Seq and Infinium 450K.

Sample ID	Number of CpGs Covered at any depth	Number of CpGs in Target Regions	%meth 0–100 (reads coverage $\geq 10X$)		Number of CpGs detected on Infinium	%meth 1–99 (reads coverage $\geq 10X$)
			Number of CpGs covered	Number of CpGs covered within target regions		Number of CpGs covered
F1_Ind	4,559,599	3,013,570	2,556,468	2,286,416	485,071	1,750,064
F2_Ind	5,693,640	3,028,601	2,663,413	2,332,705	485,050	1,874,231
F3_Cau	4,533,681	3,011,822	2,570,213	2,294,401	484,977	1,794,456
F4_Chi	4,305,522	2,980,378	2,234,838	2,058,696	485,024	1,460,009
F5_Chi	4,800,089	3,030,079	2,761,030	2,413,251	485,096	1,955,175
F6_Chi	4,802,595	3,060,089	3,004,290	2,610,047	485,086	2,139,697
M7_Chi	4,527,777	3,018,753	2,601,001	2,316,914	485,313	1,765,358
Average	4,746,129	3,020,470	2,627,322	2,330,347	485,088	1,819,856
Common	3,583,577		2,050,654		440,594	983,944

Supplementary Figs. 20). Both platforms are similar in the percentage of promoter regions covered; MC Seq and Infinium 450K covered 85% and 88% of all promoter regions, respectively (Fig. 5a, Supplementary Figs. 19). The proportions of TSS and 3'UTR regions covered by both platforms were also similar (87% for TSS and 40% for 3'UTR). MC Seq covered a significantly higher proportion of exon, intron, and 5'UTR regions (MC Seq vs. Infinium 450K: 29% vs. 21% for exon, 45% vs. 35% for 5'UTR, and 39% vs. 29% for intron). Even though the percentage of CpG islands and promoter regions covered by the 2 platforms are similar, MC Seq covered these regions at a much higher density, i.e., MC Seq assays more CpGs for the same promoter region when compared to Infinium 450K (Fig. 6, Supplementary Figs. 21–22). Likewise, MC Seq generally assays other genomic regions at much higher density (Fig. 6, Supplementary Figs. 21–22). Compared to these theoretical values of coverage for MC Seq calculated based on the bait design, we observed similar coverage of the epigenome (percentage covered and density of coverage) at 10X, but reduced coverage at 30X (Supplementary Figs. 19–22). However, to increase the coverage of the epigenome at 30X, more sequencing could be performed. We also found that the percentage of CpGs located near SNPs for both platforms were similar, e.g., 4% within 1 bp and 10% within 10 bp (data not shown).

Discussion

High-throughput methylome profiling of human samples is an important and evolving approach to investigate the role of epigenetic alterations in health. We established that MC Seq with a capture-then-bisulfite-convert approach is an attractive alternative to the widely used Infinium 450K array. We demonstrated that the performance of MC Seq at 1 μ g and 3 μ g were similar, and thus concluded that MC Seq can be effectively carried out using a capture-then-bisulfite-convert approach with 1 μ g of genomic DNA, making these 2 technologies comparable in terms of the amount of DNA required (Infinium 450K also requires 1 μ g of starting DNA). We also compared the 2 platforms in terms of their coverage of the epigenome and showed that MC Seq surveys a higher density of CpGs than Infinium 450K in key functional genomic regions and provides improved coverage of the human epigenome. Additionally, MC Seq also allows detection of 20.8 million methylation sites within the non-CpG context (CHG + CHH), which is a magnitude higher than the 450K, which offers only \sim 3091 such sites. Finally, we demonstrated as proof-of-principle, in a small sample, that methylation values from MC Seq could distinguish subjects from different ethnic groups.

The Infinium 450K can detect differences of 20% in methylation with 99% confidence.¹⁵ In our study of MC Seq, we observed that with a high reads coverage (\geq 30X), 91% of the detected CpGs had technical variation within 10%, while 9% of the detected CpGs had technical variation exceeding 10%, suggesting that MC Seq would be less sensitive than Infinium 450K in detecting small interindividual/inter-group differences in methylation. This would be an important consideration for studies with small effect sizes (5–10% inter-group differences).

We compared methylome profiles generated using a capture-then-bisulfite-convert approach with those generated using Infinium 450K. A comparison of the capture-then-bisulfite-convert approach with the bisulfite-convert-then-capture approach (Roche NimbleGen) has previously been conducted.^{27,28} Importantly, the capture-then-bisulfite-convert approach captures only one strand of the DNA, while the bisulfite-convert-then-capture approach captures both DNA strands, and thus the convert-first approach allows for profiling of genetic variation and detection of allele-specific DNA methylation in imprinted regions or hemi-methylation. For example, Li et al.²⁷ has illustrated the recapturing of allele specific information using the convert-then-capture approach; Allum et al.²⁸ compared genotype calls from the convert-then-capture approach to those from Illumina BeadChip array and found a 99% genotype concordance between the 2 methods.

There are 5 practical considerations in choosing a platform (MC Seq vs. Infinium 450K) to employ in an EWAS. These factors include (i) desired genomic coverage (MC Seq has higher density of CpGs, and hence higher coverage of variable sites); (ii) expected inter-group differences in methylation or sensitivity desired (Infinium 450K has higher sensitivity for subtle variations in methylome); (iii) cost (Infinium 450K has lower cost) and; (iv) coverage of methylation sites in non-CpG context (MC Seq has a higher capacity). The fifth factor for consideration is whether an investigator wishes to customize and specify methylation sites of interest, which can be only done through bait design in MC Seq. For example, an investigator may wish to include regions previously identified in other studies. Additionally, MC Seq baits can be customized for the organism or tissue of interest, similar to what was done by Hing et al.²⁶ for the mouse genome and Allum et al.²⁸ for adipose tissue. If the customized baits are designed to cover a smaller fraction of the genome, then the read coverage could be increased with no extra cost, potentially leading to higher accuracy, although as shown in this manuscript the gain in accuracy is limited.

Another practical consideration in employing MC Seq is the amount of sequencing required and minimum sequencing depth to filter methylation sites for analysis. This decision should be guided by statistical power calculations and would depend on (i) technical variation of MC Seq; (ii) expected inter-group differences and; (iii) study sample size. For detecting large inter-group differences, the required amount of sequencing and sequencing depth would be less than that for detecting small inter-group differences. Nevertheless, we find that at 10X reads coverage, 84% of the detected CpGs had technical variation within 10%. The technical variation was reduced at 30X reads coverage, where 91% of the detected CpGs had technical variation within 10%. The modest improvement in performance with a 3-fold increase in sequencing depth suggests that this is a feature of the platform. This is further supported by the fact that further increases in reads coverage did not decrease technical variation significantly (at 50X, 94% of CpGs had technical variation within 10%). The technical variation of MC Seq that we report here could be used to guide sample size determination in future EWAS. However, as technical variation of the assay would ultimately depend on experimental conditions, inclusion of replicate samples in each study would help the investigator determine the technical variation and the

final minimum sequencing depth to filter methylation sites for analysis (e.g., the inter-group differences has to exceed the technical variation).

This study has several strengths. First, we conducted an extensive analysis of MC Seq in buccal samples, a non-invasive clinical sample ubiquitously collected across cohorts and popular among those studying a pediatric population. Previous investigations of MC Seq were in mouse,^{26,27} maize,²⁷ cell lines,²⁷ and adipose tissue.^{27,28} Second, we demonstrated that MC Seq could be used effectively with a capture-then-bisulfite-convert approach using only 1 μg of DNA. Other studies have used a capture-then-bisulfite-convert approach with 3 μg of DNA,²⁶⁻²⁸ which has a number of limitations. Third, it is the first study investigating the utility of MC Seq in studying inter-individual variation from buccal samples from different ethnic groups. The only other study investigating interindividual variation with MC Seq was done on adipose tissue.²⁸ Last, we provided a comprehensive evaluation between MC Seq and Infinium 450K comparing the strengths and weaknesses of each platform for future EWAS. We demonstrated that MC Seq with a capture-then-bisulfite-convert approach could be employed with 1 μg of DNA. Li et al.²⁷ showed that a bisulfite-convert-then-capture assay could be used with as little as 500 ng of DNA. It will be of interest to investigate if MC Seq approach can be used at lower quantities of DNA (≤ 500 ng) in the future.

In summary, we find that MC Seq is an attractive alternative platform to Infinium 450K, for interrogating DNA methylation at single-base resolution in large number of clinical samples. Both platforms can be deployed with 1 μg of DNA. MC Seq provides denser coverage of the epigenome but the use of MC Seq in EWAS with small effect sizes will only be feasible if the inter-group differences exceed the technical variation.

Material and Methods

Collection and processing of buccal samples

Buccal epithelium was collected with informed consent from 7 volunteers comprising of 3 different ethnicities (Indian – F1_Ind and F2_Ind; Caucasian – F3_Cau; Chinese – F4_Chi, F5_Chi, F6_Chi and M7_Chi); one of the 7 samples (M7_Chi) was male. Ethnicity was self-reported. The buccal epithelium collection was done using SK-2 Isohelix swabs following manufacturer's instructions (Isohelix, UK) and stored at -80°C until further use. DNA was extracted using the Isohelix Xtreme DNA Isolation kit (XME-50, Isohelix, UK), with minor modifications to manufacturer's protocol. Following isolation, DNA was quantified using a NanoDrop spectrophotometer (ND-2000, NanoDrop, USA) and Quant-It Picogreen dsDNA Assay (P11496, Life Technologies, USA). DNA integrity was also confirmed by gel electrophoresis. We investigated the performance of MC Seq using 3 μg of genomic DNA for all 7 samples with replicates for 4 samples (F1_Ind, F3_Cau, F4_Chi and M7_Chi). DNA methylation profiling using MC Seq was also conducted using 1 μg of genomic DNA for 2 samples (F5_Chi and M7_Chi). For comparison, all 7 samples were

also profiled using 1 μg of DNA on Infinium 450K, with replicates for 4 samples (F2_Ind, F5_Chi, F6_Chi and M7_Chi).

MC Seq sample preparation and sequencing

Genomic libraries were prepared using the SureSelect^{XT} Methyl-Seq Target Enrichment System for Illumina Multiplexed Sequencing (Agilent Technologies). Briefly, 1 μg or 3 μg of genomic DNA per sample were randomly sheared via ultra sonication and DNA fragments between 150–200 bp were extracted. Sample DNA then underwent end repair, adapter ligation, hybridization to SureSelect^{XT} Methyl-Seq Capture Library, streptavidin bead enrichment, bisulfite conversion, PCR amplification and were uniquely indexed using a 6-letter sequencing tag following the manufacturer's protocol. Sample genomic libraries were then pooled and multiplexed in 4 separate lanes using 100 bp paired-end sequencing (Illumina HiSeq2000).

Processing of MC Seq data

Quality control of read sequences was performed using *FastQC*. *Trim Galore!* was then used to trim/remove reads with phred score < 30 and/or read length < 70 bases and *FastQC* was run on the trimmed sequences to verify quality control. The number of reads in the target region was determined using *bedtools intersect* (v2.24.0) command and we report number of reads within the target region as well as within 200 bases of the target region. Quality-trimmed paired-end reads were aligned to the reference human genome (hg19) using *Bismark* (v0.13.0)³¹ and *bowtie2* (v2.2.4),³² using default parameters. Duplicated reads were removed using *Bismark deduplication* tool. Methylation values were made using *Bismark methylation extractor*. Only methylation sites that were on the negative strand were retained. Sites that were on-target were determined using the *bedtools intersect* command; a site was considered on-target if it was within the target region.

Annotation of CpG sites from MC Seq

To annotate CpG sites assayed by MC Seq (SureSelect Human Methyl-Seq panel), the full list of 28 million CpGs within the human genome (hg19) was downloaded from Saffery et al.³³ From this list of 28 million CpGs, we determined CpGs that were (i) within the bait design target regions and (ii) within 200 bases of bait design target regions. These CpGs, which represent the maximum possible CpGs that can be captured by MC Seq, were then annotated in terms of their genomic features (promoter, 5'-UTR, exon, intron, 3'-UTR, TTS and intergenic) and CpG content (island, shores, shelves, open seas) using *Homer annotatePeaks* function (hg19).³⁴ These CpGs represent the theoretical maximal possible. For comparison, we also used *Homer annotatePeaks* to annotate CpGs that we observed in our samples with a minimum of 10X reads coverage.

Evaluation of technical variation and performance using lower quantities of DNA for MC Seq

To examine technical performance of MC Seq (profiled using 3 μg of genomic DNA), we first assessed reproducibility across replicates with Pearson correlation and scatterplots, using methylation sites covered with at least 10X reads coverage. Second, we calculated absolute differences in methylation values between replicates as a function of reads coverage. Third, hierarchical clustering analysis was used to ensure that replicates clustered together, using methylation sites covered with at least 10X reads coverage. Similarly, to compare the performance of MC Seq using 3 μg or 1 μg of genomic DNA, we utilized (i) Pearson correlation and scatterplots; (ii) absolute differences in methylation values at different reads coverage and; (iii) hierarchical clustering analysis, using methylation sites covered with at least 10X reads coverage.

Infinium 450K sample preparation and hybridization

DNA methylation profiling using Infinium 450K BeadChip arrays was performed following the manufacturer's protocol. After DNA extraction from buccal epithelium samples, 1 μg of DNA was treated with sodium bisulfite using the Zymo EZ-DNA kit (Zymo Research, Orange, CA, USA) according to manufacturer's instructions. Bisulfite-conversion was confirmed using methylation-specific PCR. Bisulfite-treated genomic DNA was then isothermally amplified at 37°C for 22 hours, enzymatically fragmented, purified and hybridized onto the Illumina Infinium HumanMethylation450 BeadChips (Illumina Inc., CA, USA) arrays at 48°C for 18 hours. The arrays were then washed and scanned using the Illumina iScan system (Illumina Inc., CA, USA) according to the manufacturer's instructions.

Processing of Infinium 450K data

Processing was carried out using an in-house protocol.³⁵ Signal intensities and raw methylation values were extracted from GenomeStudio™ without any data processing. Probes with data from 2 beads or fewer or with signal detection p-values exceeding 0.01 (calculated using signal versus background of the individual bead intensities) were removed. Signal intensities from the green and red channel signals were normalized and background (negative probe control values) subtraction performed. Methylation β values were then derived as the ratio of methylation probe intensity to overall intensity. Methylation β values were processed to scale the range of Type II probes to that of Type I probes.³⁶

Comparative analysis of MC Seq and Infinium 450K

MC Seq and Infinium 450K were compared using 2 metrics: (i) methylation values and (ii) genome coverage. Firstly, we compared methylation values from Infinium 450K and MC Seq at the same CpG sites over different MC Seq reads coverage, using Pearson correlation, scatterplots, and distribution plots of the methylation values. Secondly, we compared the coverage of the human epigenome by both platforms in terms of the proportions/numbers of regions

covered and the density of coverage. For MC Seq, we determined coverage of CpG/genomic regions using (i) all hg19 CpGs within the bait design target regions (representing maximal possible coverage that could be obtained from MC Seq) and (ii) observed CpGs from buccal samples. For computing coverage of CpG islands/shores/shelves, genomic coordinates (hg 19) of CpG islands were downloaded from UCSC genome browser. CpG shores were defined as up to 2 kb from CpG islands and CpG shelves were defined as up to 2 kb from CpG shores. The number of probes from each platform that overlapped with each distinct CpG island/shore/shelf were determined using *bedtools intersect* command. Likewise, to determine gene-centric coverage, we downloaded and processed lists of genomic feature regions (promoter, 5'-UTR, exon, intron, 3'-UTR, TTS) for unique genes from UCSC (hg19) and determined their overlap with CpGs covered by both platforms using *bedtools intersect* command. We also determined the number of CpG sites located near SNPs for both platforms. The *dbsnp142Common* table was downloaded from UCSC genome browser and we further restricted the analysis to SNPs with *r_s* numbers, of high quality (weight=1) and that were reported by the 1000 genomes project to have a minor allele frequency of at least 1%. We then determined the number of CpGs that had SNPs within 1 bp or 10 bp of the CpG site using *bedtools intersect* command. Lastly, to illustrate the utility of both methods for distinguishing interindividual variation, unsupervised hierarchical clustering analysis was performed, using most variably methylated CpGs (interquartile range >20%). For Infinium 450K, methylation sites from sex chromosomes and probes known to be cross-hybridizing^{37,38} were excluded in the hierarchical clustering analysis. For MC Seq, only probes from autosomal chromosomes and those with a minimum read coverage of 30X were used. Clustering was performed using euclidean distance and "ward.D" method in R. We also present clustering results using other distance metrics (Manhattan distance) and agglomeration methods (single, average, and complete), together with their approximately unbiased *P*-values³⁰ in the Supplementary Material.

Disclosure of potential conflicts of interest

YSC has received reimbursement for speaking at conferences sponsored by companies selling nutritional products. He is part of an academic consortium that has received research funding from Abbott Nutrition, Nestec, and Danone. The other authors declare no competing interests.

Acknowledgments

This work was supported by the Strategic Positioning Fund (SPF002_G00056) provided by the Agency for Science, Technology and Research (A*STAR), Singapore. Additional funds were provided by Singapore Institute for Clinical Sciences (SICS) –A*STAR.

References

1. Zhang G, Pradhan S. Mammalian Epigenetic Mechanisms. *IUBMB Life* 2014; 66:240-56; PMID:24706538; <http://dx.doi.org/10.1002/iub.1264>
2. Messerschmidt DM, Knowles BB, Solter D. DNA methylation dynamics during epigenetic reprogramming in the germline and preimplantation embryos. *Genes Dev* 2014; 28:812-28; PMID:24736841; <http://dx.doi.org/10.1101/gad.234294.113>

3. Bird A. DNA methylation patterns and epigenetic memory. *Genes Dev* 2002; 16:6-21; PMID:11782440; <http://dx.doi.org/10.1101/gad.947102>
4. Pembrey M, Saffery R, Bygren LO, Epidemiology NiE. Human trans-generational responses to early-life experience: potential impact on development, health and biomedical research. *J Med Genet* 2014; 51:563-72; PMID:25062846; <http://dx.doi.org/10.1136/jmedgenet-2014-102577>
5. Liu Y, Li X, Aryee MJ, Ekström TJ, Padyukov L, Klareskog L, Vandiver A, Moore AZ, Tanaka T, Ferrucci L. GeMes, clusters of DNA methylation under genetic control, can inform genetic and epigenetic analysis of disease. *Am J Hum Genet* 2014; 94:485-95; PMID:24656863; <http://dx.doi.org/10.1016/j.ajhg.2014.02.011>
6. Zeisel SH. Epigenetic mechanisms for nutrition determinants of later health outcomes. *Am J Clin Nutr* 2009; 89:1488S-93S; PMID:19261726; <http://dx.doi.org/10.3945/ajcn.2009.27113B>
7. Heijmans BT, Tobi EW, Stein AD, Putter H, Blauw GJ, Susser ES, Slagboom PE, Lumey LH. Persistent epigenetic differences associated with prenatal exposure to famine in humans. *Proc Natl Acad Sci U S A* 2008; 105:17046-9; PMID:18955703; <http://dx.doi.org/10.1073/pnas.0806560105>
8. Mill J, Heijmans BT. From promises to practical strategies in epigenetic epidemiology. *Nat Rev Genet* 2013; 14:585-94; PMID:23817309; <http://dx.doi.org/10.1038/nrg3405>
9. Rakyan VK, Down TA, Balding DJ, Beck S. Epigenome-wide association studies for common human diseases. *Nat Rev Genet* 2011; 12:529-41; PMID:21747404; <http://dx.doi.org/10.1038/nrg3000>
10. Portela A, Esteller M. Epigenetic modifications and human disease. *Nat Biotechnol* 2010; 28:1057-68; PMID:20944598; <http://dx.doi.org/10.1038/nbt.1685>
11. Feinberg AP. Epigenomics reveals a functional genome anatomy and a new approach to common disease. *Nat Biotechnol* 2010; 28:1049-52; PMID:20944596; <http://dx.doi.org/10.1038/nbt1010-1049>
12. Lowe R, Gemma C, Beyan H, Hawa MI, Bazeos A, Leslie RD, Montpetit A, Rakyan VK, Ramagopalan SV. Buccals are likely to be a more informative surrogate tissue than blood for epigenome-wide association studies. *Epigenetics* 2013; 8:445-54; PMID:23538714; <http://dx.doi.org/10.4161/epi.24362>
13. Smith AK, Kilaru V, Klengel T, Mercer KB, Bradley B, Conneely KN, Ressler KJ, Binder EB. DNA extracted from saliva for methylation studies of psychiatric traits: evidence tissue specificity and relatedness to brain. *Am J Med Genet B Neuropsychiatr Genet* 2015; 168B:36-44; PMID:25355443; <http://dx.doi.org/10.1002/ajmg.b.32278>
14. Bibikova M, Le J, Barnes B, Saedinia-Melnyk S, Zhou L, Shen R, Gunderson KL. Genome-wide DNA methylation profiling using Infinium (R) assay. *Epigenomics* 2009; 1:177-200; PMID:22122642; <http://dx.doi.org/10.2217/epi.09.14>
15. Bibikova M, Barnes B, Tsan C, Ho V, Klotzle B, Le JM, Delano D, Zhang L, Schroth GP, Gunderson KL, et al. High density DNA methylation array with single CpG site resolution. *Genomics* 2011; 98:288-95; PMID:21839163; <http://dx.doi.org/10.1016/j.ygeno.2011.07.007>
16. Cokus SJ, Feng S, Zhang X, Chen Z, Merriman B, Haudenschild CD, Pradhan S, Nelson SF, Pellegrini M, Jacobsen SE. Shotgun bisulphite sequencing of the Arabidopsis genome reveals DNA methylation patterning. *Nature* 2008; 452:215-9; PMID:18278030; <http://dx.doi.org/10.1038/nature06745>
17. Lister R, Pelizzola M, Downen RH, Hawkins RD, Hon G, Tonti-Filippini J, Nery JR, Lee L, Ye Z, Ngo QM, et al. Human DNA methylomes at base resolution show widespread epigenomic differences. *Nature* 2009; 462:315-22; PMID:19829295; <http://dx.doi.org/10.1038/nature08514>
18. Lister R, O'Malley RC, Tonti-Filippini J, Gregory BD, Berry CC, Millar AH, Ecker JR. Highly integrated single-base resolution maps of the epigenome in Arabidopsis. *Cell* 2008; 133:523-36; PMID:18423832; <http://dx.doi.org/10.1016/j.cell.2008.03.029>
19. Lee YK, Jin S, Duan S, Lim YC, Ng DP, Lin XM, Yeo GS, Ding C. Improved reduced representation bisulfite sequencing for epigenomic profiling of clinical samples. *Biol Proced Online* 2014; 16:1; PMID:24406024; <http://dx.doi.org/10.1186/1480-9222-16-1>
20. Meissner A, Gnirke A, Bell GW, Ramsahoye B, Lander ES, Jaenisch R. Reduced representation bisulfite sequencing for comparative high-resolution DNA methylation analysis. *Nucleic Acids Res* 2005; 33:5868-77; PMID:16224102; <http://dx.doi.org/10.1093/nar/gki901>
21. Magdalena J, Goval JJ. Methyl DNA immunoprecipitation. *Methods Mol Biol* 2009; 567:237-47; PMID:19588096; http://dx.doi.org/10.1007/978-1-60327-414-2_15
22. Jacinto FV, Ballestar E, Esteller M. Methyl-DNA immunoprecipitation (MeDIP): hunting down the DNA methylome. *Biotechniques* 2008; 44:35, 7, 9 passim; PMID:18254377; <http://dx.doi.org/10.2144/000112708>
23. Down TA, Rakyan VK, Turner DJ, Flicek P, Li H, Kulesha E, Graf S, Johnson N, Herrero J, Tomazou EM, et al. A Bayesian deconvolution strategy for immunoprecipitation-based DNA methylome analysis. *Nat Biotechnol* 2008; 26:779-85; PMID:18612301; <http://dx.doi.org/10.1038/nbt1414>
24. Harris RA, Wang T, Coarfa C, Nagarajan RP, Hong C, Downey SL, Johnson BE, Fouse SD, Delaney A, Zhao Y, et al. Comparison of sequencing-based methods to profile DNA methylation and identification of monoallelic epigenetic modifications. *Nat Biotechnol* 2010; 28:1097-105; PMID:20852635; <http://dx.doi.org/10.1038/nbt.1682>
25. Bock C, Tomazou EM, Brinkman AB, Muller F, Simmer F, Gu H, Jager N, Gnirke A, Stunnenberg HG, Meissner A. Quantitative comparison of genome-wide DNA methylation mapping technologies. *Nat Biotechnol* 2010; 28:1106-14; PMID:20852634; <http://dx.doi.org/10.1038/nbt.1681>
26. Hing B, Ramos E, Braun P, McKane M, Jancic D, Tamashiro KL, Lee RS, Michaelson JJ, Druley TE, Potash JB. Adaptation of the targeted capture Methyl-Seq platform for the mouse genome identifies novel tissue-specific DNA methylation patterns of genes involved in neurodevelopment. *Epigenetics* 2015; 10:581-96; PMID:25985232; <http://dx.doi.org/10.1080/15592294.2015.1045179>
27. Li Q, Suzuki M, Wendt J, Patterson N, Eichten SR, Hermanson PJ, Green D, Jeddeloh J, Richmond T, Rosenbaum H, et al. Post-conversion targeted capture of modified cytosines in mammalian and plant genomes. *Nucleic Acids Res* 2015; 43:e81; PMID:25813045; <http://dx.doi.org/10.1093/nar/gkv244>
28. Allum F, Shao X, Guenard F, Simon MM, Busche S, Caron M, Lambourne J, Lessard J, Tandre K, Hedman AK, et al. Characterization of functional methylomes by next-generation capture sequencing identifies novel disease-associated variants. *Nat Commun* 2015; 6:7211; PMID:26021296; <http://dx.doi.org/10.1038/ncomms8211>
29. Teh AL, Pan H, Chen L, Ong ML, Dogra S, Wong J, MacIsaac JL, Mah SM, McEwen LM, Saw SM, et al. The effect of genotype and in utero environment on interindividual variation in neonate DNA methylomes. *Genome Res* 2014; 24:1064-74; PMID:24709820; <http://dx.doi.org/10.1101/gr.171439.113>
30. Suzuki R, Shimodaira H. Pvcust: an R package for assessing the uncertainty in hierarchical clustering. *Bioinformatics* 2006; 22:1540-2; PMID:16595560; <http://dx.doi.org/10.1093/bioinformatics/btl117>
31. Krueger F, Andrews SR. Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications. *Bioinformatics* 2011; 27:1571-2; PMID:21493656; <http://dx.doi.org/10.1093/bioinformatics/btr167>
32. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. *Nat Methods* 2012; 9:357-9; PMID:22388286; <http://dx.doi.org/10.1038/nmeth.1923>
33. Saffery R, Gordon L. Time for a standardized system of reporting sites of genomic methylation. *Genome Biol* 2015; 16:85; PMID:25924664; <http://dx.doi.org/10.1186/s13059-015-0654-9>
34. Heinz S, Benner C, Spann N, Bertolino E, Lin YC, Laslo P, Cheng JX, Murre C, Singh H, Glass CK. Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Mol Cell* 2010; 38:576-89; PMID:20513432; <http://dx.doi.org/10.1016/j.molcel.2010.05.004>
35. Pan H, Chen L, Dogra S, Teh AL, Tan JH, Lim YI, Lim YC, Jin S, Lee YK, Ng PY, et al. Measuring the methylome in clinical samples: improved processing of the Infinium Human Methylation450

- BeadChip Array. *Epigenetics* 2012; 7:1173-87; PMID:22964528; <http://dx.doi.org/10.4161/epi.22102>
36. Dedeurwaerder S, Defrance M, Calonne E, Denis H, Sotiriou C, Fuks F. Evaluation of the Infinium Methylation 450K technology. *Epigenomics* 2011; 3:771-84; PMID:22126295; <http://dx.doi.org/10.2217/epi.11.105>
37. Price ME, Cotton AM, Lam LL, Farre P, Emberly E, Brown CJ, Robinson WP, Kobor MS. Additional annotation enhances potential for biologically-relevant analysis of the Illumina Infinium HumanMethylation450 BeadChip array. *Epigenetics Chromatin* 2013; 6:4; PMID:23452981; <http://dx.doi.org/10.1186/1756-8935-6-4>
38. Chen YA, Lemire M, Choufani S, Butcher DT, Grafodatskaya D, Zanke BW, Gallinger S, Hudson TJ, Weksberg R. Discovery of cross-reactive probes and polymorphic CpGs in the Illumina Infinium HumanMethylation450 microarray. *Epigenetics* 2013; 8:203-9; PMID:23314698; <http://dx.doi.org/10.4161/epi.23470>