

UC Berkeley

UC Berkeley Previously Published Works

Title

Comparison of DNA methylation measurements from EPIC BeadChip and SeqCap targeted bisulphite sequencing in PON1 and nine additional candidate genes

Permalink

<https://escholarship.org/uc/item/0mr420s9>

Journal

Epigenetics, 17(13)

ISSN

1559-2294

Authors

Khodasevich, Dennis

Smith, Anna R

Huen, Karen

et al.

Publication Date



2022-12-09

DOI

10.1080/15592294.2022.2091818

Peer reviewed

Comparison of DNA methylation measurements from EPIC BeadChip and SeqCap targeted bisulphite sequencing in *PON1* and nine additional candidate genes

Dennis Khodasevich ^{a,*}, Anna R. Smith ^{a,b,*}, Karen Huen^a, Brenda Eskenazi^c, Andres Cardenas^{a,b,*}, and Nina Holland^{a,*}

^aDivision of Environmental Health Sciences, Children's Environmental Health Laboratory, School of Public Health, University of California, Berkeley, CA, USA; ^bCenter for Computational Biology, University of California, Berkeley, CA, USA; ^cCenter for Children's Environmental Health, School of Public Health, University of California, Berkeley, CA, USA

ABSTRACT

Epigenome-wide association studies (EWAS) are widely implemented in epidemiology, and the Illumina HumanMethylationEPIC BeadChip (EPIC) DNA microarray is the most-used technology. Recently, next-generation sequencing (NGS)-based methods, which assess DNA methylation at single-base resolution, have become more affordable and technically feasible. While the content of microarray technology is fixed, NGS-based approaches, such as the Roche Nimblegen, SeqCap Epi Enrichment System (SeqCap), offer the flexibility of targeting most CpGs in a gene. With the current usage of microarrays and emerging NGS-based technologies, it is important to establish whether data generated from the two platforms are comparable. We harnessed 112 samples from the Center for the Health Assessment of Mothers and Children of Salinas (CHAMACOS) birth cohort study and compared DNA methylation between the EPIC microarray and SeqCap for *PON1* and nine additional candidate genes, by evaluating epigenomic coverage and correlations. We conducted multivariable linear regression and principal component analyses to assess the ability of the EPIC array and SeqCap to detect biological differences in gene methylation by the *PON1*₋₁₀₈ single nucleotide polymorphism. We found an overall high concordance ($r = 0.84$) between SeqCap and EPIC DNA methylation, among highly methylated and minimally methylated regions. However, substantial disagreement was present between the two methods in moderately methylated regions, with SeqCap measurements exhibiting greater within-site variation. Additionally, SeqCap did not capture *PON1* SNP associated differences in DNA methylation that were evident with the EPIC array. Our findings indicate that microarrays perform well for analysing DNA methylation in large cohort studies but with limited coverage.

ARTICLE HISTORY

Received 2 March 2022
Accepted 15 June 2022



KEYWORDS

DNA methylation; SeqCap; NGS-based sequencing; EPIC microarray; *PON1*


Introduction

Epigenetic mechanisms, such as DNA methylation, histone modifications, and non-coding RNAs, regulate gene expression by altering DNA accessibility and chromatin structure[1]. DNA methylation is the most widely investigated epigenetic mechanism. It occurs when a methyl group is added at a cytosine nucleotide that precedes a guanine (CpG dinucleotides), influencing DNA function by altering transcriptional activity of a gene and chromatin accessibility and remodelling [1,2]. The human genome contains approximately 30 million CpG sites distributed

throughout several gene regions, including CpG islands, shores, shelves, and gene bodies. CpG islands are stretches of DNA with a high frequency of CpG dinucleotides that occur in proximity to gene promoter regions[3]. It was previously believed that the majority of functional changes occurred in CpG islands, but more additional research has shown that DNA methylation changes along CpG shores (regions within 2kb of islands) and within the gene body may also have functional effects on gene expression[4]. DNA methylation patterns are established during the prenatal period and vary by tissue and cell type

CONTACT Anna R. Smith  annsmi11@berkeley.edu  Division of Environmental Health Sciences, School of Public Health, University of California, Berkeley, 2121, Berkeley Way, #5302, Berkeley, CA 94704, USA

*These authors contributed equally to this work.

 Supplemental data for this article can be accessed online at <https://doi.org/10.1080/15592294.2022.2091818>

© 2022 The Author(s). Published by Informa UK Limited, trading as Taylor & Francis Group.
This is an Open Access article distributed under the terms of the Creative Commons Attribution-NonCommercial-NoDerivatives License (<http://creativecommons.org/licenses/by-nc-nd/4.0/>), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited, and is not altered, transformed, or built upon in any way.

with the potential to be influenced by some environmental exposures[5]. Differential DNA methylation is associated with ageing, cancer, and the development of cardiovascular and neurodegenerative diseases [6–10]. This has generated increased interest in studying DNA methylation changes that may act as a mechanism through which environmental exposures could influence gene expression and health[11].

Epigenome-wide association studies (EWAS) are widely used to examine associations between DNA methylation and either exposures or health outcomes. At present, the Illumina HumanMethylationEPIC BeadChip (EPIC) DNA microarray is the most widely used technology for conducting EWAS in epidemiological studies, as it efficiently measures methylation at >850,000 CpG sites throughout the genome[12]. Recently, targeted next-generation sequencing (NGS)-based technologies have been developed, which assess DNA methylation at single-base resolution[13], and are becoming more affordable and technically feasible. While the content of microarrays is pre-selected and fixed, NGS-based technologies, such as Roche Nimblegen, SeqCap Epi Enrichment System (SeqCap)[14], offer the flexibility of assessing methylation of most CpGs [15] and enable the reproducible targeting of selected genomic regions, up to 210 Mb, from bisulphite-treated genomic DNA[14].

Despite technological availability, literature comparing the use of the two platforms in large human cohort studies is sparse. One study utilizing DNA from whole blood samples from 96 participants found that overall bisulphite-based amplicon sequencing (BSAS), a method that combines bisulphite conversion with targeted amplification of regions of interest, transposome-mediated library construction, and benchtop NGS[16], correlated highly with EPIC arrays, however exhibited substantial variation in sites where the magnitude of change via the EPIC array was greater than 5%[17]. Another study utilizing cord blood DNA from 4 participants compared the EPIC DNA microarray to a NGS-based technology similar to SeqCap, Illumina TruSeq Methyl Capture EPIC Kit (TruSeq), and found that although TruSeq offered greater read depth and

excellent coverage, the methylation data generated from the EPIC array were more precise than that of TruSeq[18].

We aim to expand on prior studies [18] and compare the performance of the widely used EPIC BeadChip array with SeqCap sequencing methodology in terms of coverage, correlation between platforms, and identification of differential methylation by the *PON1* gene, which was previously characterized[19]. We will harness blood samples from children participating in the Center for the Health Assessment of Mothers and Children of Salinas (CHAMACOS) longitudinal birth cohort study and compare DNA methylation in 10 genes, including the well-studied *PON1* gene, five *PON1*-related genes (*PON2*, *PON3*, *ACHE*, *AHR*, *SP1*), two genes that regulate DNA methylation (*DNMT1*, *DNMT3B*), and two candidate genes (*TAPBP*, *VTRNA2-1*) from our prior analyses that may be of interest for future studies in environmental epidemiology.

Methods

Study participants

Participants were Mexican-American children from the CHAMACOS birth cohort study, which examined the impact of pesticides and other environmental and social exposures on the health and development of children living in the Salinas Valley, California. A detailed description of the CHAMACOS cohort and results from prior epigenetic analyses were previously published [12,20,21]. For this study, we included a convenience subset of data from 112 seven-year-old children, who had sufficient DNA samples and DNA methylation measured on both the EPIC and SeqCap platforms. Study protocols were approved by the University of California, Berkeley. Written informed consent and verbal assent were obtained from adults and children, respectively.

Gene selection

We examined DNA methylation in 10 genes relevant to environmental epidemiology, including the

Paraoxonase 1 (*PON1*) gene on chromosome 7, which has a broad spectrum of functions ranging from pesticide sensitivity to inflammation and can serve as a useful model for integrating genetic and epigenetic data[19]. Paraoxonase 2 (*PON2*) and Paraoxonase 3 (*PON3*) were chosen due to their proximity to *PON1* on chromosome 7 and their similar biological functions[22]. Acetylcholinesterase (*ACHE*), also located on chromosome 7, codes for the enzyme AChE, whose activity is associated with *PON1* due to the *PON1*-catalysed hydrolyzation of organophosphates[23]. Aryl Hydrocarbon Receptor (*AHR*), on chromosome 7, codes for the AHR transcription factor, which plays a role in the induction of *PON1* expression[24]. DNA methyltransferase 1 (*DNMT1*) on chromosome 19 and DNA methyltransferase 3 beta (*DNMT3B*) on chromosome 20 play an integral role in maintaining DNA methylation[25]. *SPI* on chromosome 12 encodes a transcription factor that positively regulates *PON1* transcription[26]. Finally, we examined two additional genes, *VTRNA2-1* (also known as *miR886*) on chromosome 5 and *TAPBP* on chromosome 6, which were differentially methylated region (DMR) hits of interest from our prior studies[27].

Blood collection and processing

Blood specimens were collected by venipuncture at the CHAMACOS field office by a paediatric phlebotomist. Whole blood was collected in BD vacutainers® (Becton, Dickinson and Company, Franklin Lakes, NJ) containing no anticoagulant. These samples were centrifuged, divided into serum and clot, and stored at -80°C at the Berkeley Public Health Biorepository, University of California, Berkeley. DNA was isolated from the banked blood clot samples using QIAamp DNA Blood Maxi Kits (Qiagen, Valencia, CA) according to the manufacturer's protocol with minor modifications, as previously described[28].

EPIC methylation analysis and data processing

DNA was normalized to 55 $\mu\text{g}/\text{ml}$, and bisulphite conversion was performed on 1 μg aliquots of DNA using Zymo Bisulphite conversion Kits

(Zymo Research, Orange, CA). DNA was whole genome amplified, enzymatically fragmented, purified, and applied to the Illumina Infinium MethylationEPIC BeadChips (Illumina, San Diego, CA) according to the Illumina methylation protocol [29–31]. BeadChip processing was performed using robotics, and the Illumina Hi-Scan system was used for analysis. DNA methylation was measured at 866,836 CpG sites on the EPIC BeadChip. Quality control (QC) measures were previously described [21] and included use of repeats, internal standards, and randomization of samples across chips and plates.

DNA methylation data was initially processed using the R package *minfi* (v.1.36.0)[32]. Briefly, sample quality was estimated using the function *getQC*, and any samples with weak median methylated/unmethylated signal strength were dropped. Cell type proportions (CD8 T-Cells, CD4 T-Cells, B Cell, Monocytes, Granulocytes, and Natural Killer Cells) were estimated using the function *estimateCellCounts* in *minfi*, which uses the Reinius adult reference data set. Probes were mapped to genome, and sex was estimated using the functions *mapToGenome* and *getSex*, respectively. Functional normalization was applied using the function *preprocessFunnorm*. Failed probes, as well as any samples with $\geq 1\%$ of probes below the limit of detection, were identified and removed from normalized data using the function *detectionP*. Beta values were calculated from normalized data using *getBeta*. Any zero values in the betas were set to half the minimum observed non-zero value. Signal levels of type I and II probes were normalized using *rcp* from the package *Enmix* (v1.26.10)[33]. Batch effects associated with plates used during the bisulphite conversion step were estimated and removed using the ComBat method[34], as implemented in the R package *sva* (v3.38.0)[35]. Within the 7-year samples utilized for this study, no samples were dropped due to poor signal quality.

SeqCap methylation analysis and data processing

Browser extensible data (BED) files for SeqCap were created for all genes of interest using the

table browser in the University of California, Santa Cruz (UCSC) Genome Browser Gateway[36]. For promoter regions, we chose 2000 bp upstream of the gene. Genome locations were converted from hg19 to hg38 using the UCSC Genome LiftOver tool[37].

A total of 90 μ L of DNA (normalized to 50ng/ μ L, 260/280 > 1.8) from each participant was sent to the Functional Genomics Laboratory at the University of California, Berkeley for DNA methylation analysis by the Roche Nimblegen, SeqCap Epi Enrichment System, following company protocol[14]. Raw sequencing data was then processed at the UC Davis Bioinformatics Core. Briefly, pair-end 151 bp raw sequencing reads were subjected to adapter removal by scythe-0.993. Bases that have qualities lower than 30 were trimmed using *sickle-1.33*, and trimmed reads that are less than 30 bp long were not considered for downstream analysis. The reads that have passed the above quality control were aligned to the human genome (version GRCh38), using *BSseeker2*[38]. *Bowtie2* [39] was used as the underlying aligner. Methylation status was identified using a built-in function in *BSseeker2*. The methylation results for the regions that include the selected genes, as well as 20 K bp up- and down-stream were aggregated using *bedtools-2.25.0*[40].

Determination of *PON1*₋₁₀₈ single nucleotide polymorphism

The promoter single nucleotide polymorphism (SNP), *PON1*₋₁₀₈ was genotyped using a fluorogenic allele-specific assay (Amplifluor, Chemicon, Temecula, CA), as described previously[19].

Statistical analyses

To assess the overall coverage and correlations, we calculated the Pearson correlation coefficient between EPIC array and SeqCap methylation values, for each CpG site common to both the EPIC array and SeqCap, across all 10 genes, to obtain an overall correlation per gene.

To assess the ability of the EPIC array and SeqCap to detect biological differences in DNA methylation, we used multivariable linear

regression to determine whether *PON1*₋₁₀₈ SNP is associated with *PON1* DNA methylation, using each of the technologies. We chose the functional *PON1*₋₁₀₈ SNP as a model for the detection of biologically relevant differential DNA methylation because the SNP is a known predictor of *PON1* gene methylation and affects *PON1* enzyme quantity[19]. DNA methylation data were expressed as *M*-values, which are calculated as the \log_2 ratio of the intensities of methylated to unmethylated probes[41]. SeqCap data featured a relatively large number of methylation measurements reading 0 or 1. To avoid the formation of infinite values, betas of 0 and 1 were converted to 0.01 and 0.99, respectively, prior to *M*-value conversion.

Linear regression models were fit for each CpG site using the following DNA methylation outcomes: (1) EPIC array DNA methylation values for *PON1* CpG sites also included on SeqCap (14 comparisons), (2) SeqCap DNA methylation values for *PON1* CpG sites also included on the EPIC array (14 comparisons), and (3) all SeqCap DNA methylation values for the *PON1* gene (255 comparisons). Linear regression models were run with site-specific *M*-value as the dependent variable and *PON1*₋₁₀₈ genotype and relevant covariates as the independent variables, which included batch and cell-type proportions in the EPIC models and cell-type proportions in the SeqCap models, since batch effects were not a concern for SeqCap. The purpose of the first two models is to compare the ability of EPIC and SeqCap to detect differences in associations between genotype and methylation using only sites available in both platforms, while the final model serves to elucidate the number of associations detected from SeqCap, which provides higher gene coverage. We assessed statistical significance after adjusting for multiple comparisons using the Bonferroni correction.

In addition, we conducted principal component analyses (PCA) using spectral decomposition on the covariance matrix of the centred DNA methylation data matrix, consisting of methylation measurements in the 14 selected CpG sites in *PON1*. In both EPIC and SeqCap analyses, the first two principal components were extracted and used for visualization. Given the strong associations between the *PON1*₋₁₀₈ SNP and DNA methylation

within *PON1*, we then utilized linear regression to determine whether the first two principal components could predict the identity of the *PON1*₋₁₀₈ SNP, using the SNP as the dependent variable, and the principal components and relevant covariates as the independent variables, which included batch and cell-type proportions in the EPIC models and cell-type proportions in the SeqCap models. All analyses were conducted in R (version 3.6.2 R Development Core Team).

Results

Study participants and design

Seven-year-old children from the CHAMACOS cohort with DNA methylation measured using both EPIC and SeqCap ($n=112$) were included in the coverage and correlation analyses. Of these, 48 were assigned male and 64 were assigned female at birth. After excluding one participant with missing *PON1*₋₁₀₈ promoter polymorphism data, the final sample size for analyses examining detection of biological differences was 111 participants, including 29 individuals with the CC genotype, 60 individuals with the CT genotype, and 22 individuals with the TT genotype. The CC genotype has been associated with higher levels of gene expression, and thus lower levels of methylation, the CT genotype has been associated with intermediate levels of methylation, and those with TT genotypes have been associated with the highest level of methylation[19].

Coverage and correlations

Table 1 describes the 10 genes included in the reproducibility analyses, number of CpG sites assessed, beta distribution, and DNA methylation correlation between EPIC and SeqCap. SeqCap greatly expanded the number of CpG sites available for analysis across the included genes. This increase in CpG site coverage ranged from a five-fold increase in *TAPBP* to an approximately forty-fold increase in *SP1*. Overall correlation of DNA methylation results between the two methods was relatively high ($r=0.84$), with the individual genes exhibiting a wide range of correlation coefficients, ranging from -0.02 (*TAPBP*) to 0.88 (*PON2*)

Table 1. Summary statistics of genes included in the analysis, beta distributions, and correlations between DNA methylation measured using EPIC and SeqCap.

Gene	Chr	EPIC		SeqCap β		Pearson correlation coefficient (r)
		sites (n)*	EPIC β mean (range)	SeqCap sites (n)	mean (range)	
Overall	–	249	–	5006	–	0.84
<i>ACHE</i>	7	34	0.35 (0.00–0.98)	500	0.28 (0–1)	0.82
<i>AHR</i>	7	15	0.92 (0.62–0.98)	436	0.87 (0–1)	0.02
<i>PON1</i>	7	14	0.82 (0.20–0.98)	255	0.74 (0–1)	0.49
<i>PON2</i>	7	21	0.38 (0.00–0.97)	248	0.32 (0–1)	0.88
<i>PON3</i>	7	31	0.44 (0.01–0.97)	328	0.39 (0–1)	0.84
<i>DNMT1</i>	19	49	0.72 (0.01–0.98)	1397	0.62 (0–1)	0.74
<i>DNMT3B</i>	20	31	0.54 (0.02–0.97)	884	0.44 (0–1)	0.74
<i>SP1</i>	12	21	0.26 (0.01–0.96)	795	0.20 (0–1)	0.88
<i>TAPBP</i>	6	26	0.04 (0.00–0.24)	126	0.01 (0–1)	-0.02
<i>VTRNA2-</i>	5	7	0.38 (0.02–0.69)	37	0.30 (0–1)	0.48

Notes: *EPIC sites correspond to those also included in SeqCap and not total sites available on the EPIC array.

Abbreviations: Chr, chromosome; EPIC, Illumina HumanMethylationEPIC BeadChip; SeqCap, Roche Nimblegen, SeqCap Epi Enrichment System.

(Table 1). Although there was a wide range of correlation coefficients within each gene for the two methods, the overall patterns of correlation between the two methods were consistent throughout all genes (Table 1). EPIC DNA methylation values within each site tended to be concentrated within a small range, while corresponding SeqCap values ranged more broadly and regularly featured outliers close to the extreme methylation values of 0 and 1 (Figure S1).

In *PON1*, SeqCap measured methylation at 255 CpG sites, compared to 19 sites with the EPIC array, and particularly increased coverage within the gene body (Figure 1). Corresponding boxplots displaying methylation from SeqCap of the other included genes are shown in Figure S2. Both EPIC and SeqCap reveal similar overall methylation patterns in *PON1*, with highly methylated outer sites and moderately methylated central sites of the gene (Figure 2). However, EPIC methylation values within each site exhibit lower variability than SeqCap methylation values within each site. This pattern of SeqCap data exhibiting higher within-site methylation measurement variability was consistent throughout all genes included in this study (Figure S3). The high overall correlation ($r=0.84$) between the two methods is largely driven by a high level of agreement in highly methylated sites and sites with the lowest methylation (Figure 3). Despite the high correlation, noticeable disagreement exists between the two methods in moderately

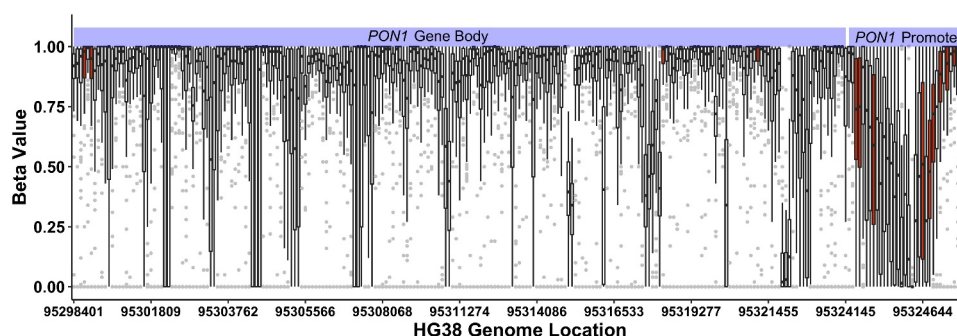


Figure 1. Complete CpG methylation distribution in the *PON1* gene, measured by the Roche Nimblegen, SeqCap Epi enrichment system (SeqCap). All CpG sites in *PON1* covered by SeqCap, organized by hg38 genome location. All CpG sites also included in the Illumina HumanMethylationEPIC BeadChip are shaded in red. Approximate locations of the promoter and gene body are labelled on the x-axis.

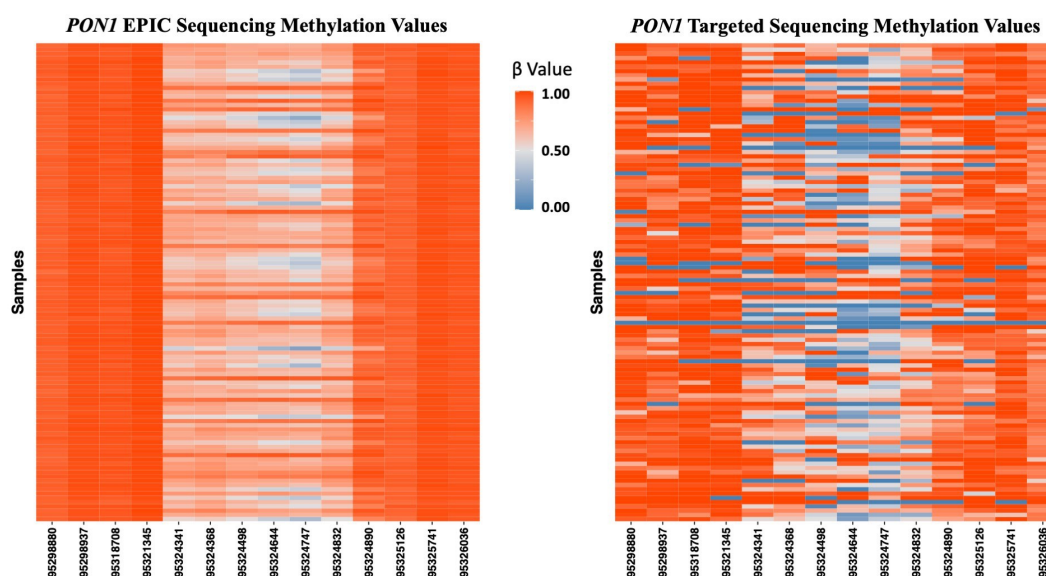


Figure 2. Distribution of methylation values in the *PON1* gene. Each row contains methylation measurements from a single participant and each column contains all measurements from a single CpG site, organized by position in the hg38 reference genome. CpG-site specific methylation values across all participants from EPIC data (left) and SeqCap targeted data (right). Colour scale corresponds to the methylation beta value, ranging from 0 (completely unmethylated) to 1 (completely methylated).

methylated sites. Additionally, the SeqCap measurements exhibit a skewed distribution towards 0 and 1 throughout the entire range of EPIC methylation measurements.

Detection of biological differences in DNA methylation

We aimed to characterize the utility of SeqCap to detect biologically relevant differences in DNA methylation by comparing its performance to that of the EPIC array. For this aim, we characterized associations between the functional *PON1*₋₁₀₈ SNP,

which exerts a strong effect on *PON1* enzyme expression, and DNA methylation within the *PON1* gene. Overall, SeqCap methylation measurements exhibit higher variability compared to the EPIC methylation measurements. SeqCap provides methylation measurements on CpG sites closer to the location of the *PON1*₋₁₀₈ SNP. EPIC contains measurements on CpG sites located 61 base pairs upstream and 85 base pairs downstream of the *PON1*₋₁₀₈ SNP, while SeqCap contains measurements on CpG sites located 9 base pairs upstream and 1 base pair downstream of the *PON1*₋₁₀₈ SNP. *PON1*₋₁₀₈ polymorphism-associated trends in *PON1*

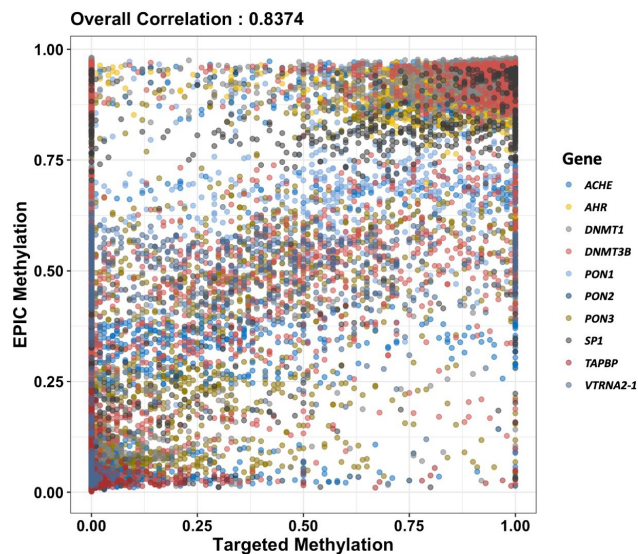


Figure 3. Direct comparisons between DNA methylation from the Roche Nimblegen, SeqCap Epi enrichment system (targeted Methylation; x-axis) and the Illumina HumanMethylationEPIC Bead (EPIC Methylation; y-axis) at 242 CpG sites from 112 participants. Pearson correlation coefficient of 0.84.

methylation are clearly defined in EPIC data, with participants harbouring the CC genotype exhibiting lower methylation values in the central *PON1* CpG

sites and samples of participants with the TT genotype exhibiting higher methylation values in the central *PON1* CpG sites (Figure 4). Polymorphism-associated trends were less pronounced in the SeqCap data, largely due to higher within-site variability.

Linear regression models, with methylation at each site as the dependent variable and the *PON1*₋₁₀₈ polymorphism genotype as the independent variable, support the visual trends from the differential methylation boxplots. The linear regression models reveal stronger associations between the *PON1*₋₁₀₈ polymorphism and methylation when using EPIC data (Table 2). The *PON1*₋₁₀₈ polymorphism was negatively associated with methylation at eight CpG sites when using EPIC data, all of which were located within the *PON1* promoter. The *PON1*₋₁₀₈ polymorphism was positively associated with methylation at two CpG sites when using the full SeqCap data, both of which were in the *PON1* gene body. No sites were statistically significant when using the subset SeqCap data. Extended model summaries with

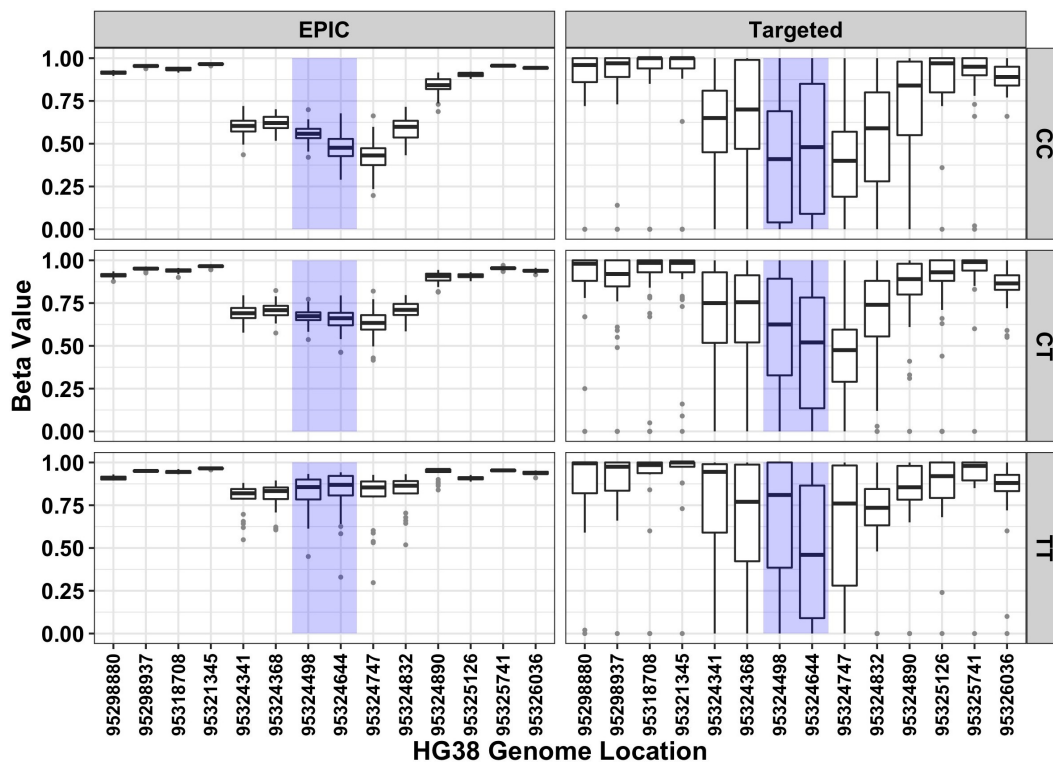


Figure 4. Differential methylation by *PON1*₋₁₀₈ single nucleotide polymorphism. Distribution of methylation values in *PON1* from Illumina HumanMethylationEPIC BeadChip (EPIC; left) and Roche Nimblegen, SeqCap Epi enrichment system (targeted; right), grouped by *PON1*₋₁₀₈ polymorphism. Blue region highlights the two closest EPIC CpG sites to the *PON1*₋₁₀₈ polymorphism.

Table 2. Linear regression model summaries examining associations between *PON1*₋₁₀₈ single nucleotide polymorphism and DNA methylation using two analytical platforms.

Dataset ¹	CpG sites (n)	Significant sites	Directionality of Association ⁴
EPIC ²	14	8	All negative
SeqCap (subset) ³	14	0	–
SeqCap (full) ³	255	2	All positive

¹Three different datasets: (1) EPIC: DNA methylation measured by Illumina HumanMethylationEPIC BeadChip and also measured by Roche Nimblegen, SeqCap Epi Enrichment System; (2) SeqCap (subset): DNA methylation measure by SeqCap and also measured by EPIC; (3) SeqCap (full): DNA methylation measured by complete Roche Nimblegen, SeqCap Epi Enrichment System dataset.

²Controlling for cell type proportions and batch.

³Controlling for cell type proportions.

⁴A negative association between genotype and methylation indicates that the presence of more C alleles is associated with lower methylation within the CpG site.

detailed coefficients, standard errors, and p-values for the significant sites are reported in Table S1.

Finally, we utilized principal component analysis (PCA) to compare the ability of DNA methylation measurements from SeqCap and EPIC, in the 14 CpG sites of interest, to predict *PON1*₋₁₀₈ polymorphism identity. While the first principal component in the EPIC PCA roughly categorizes participants by *PON1*₋₁₀₈ genotype, neither of the first two principal components in the SeqCap PCA corresponds to *PON1*₋₁₀₈ genotype (Figure 5). This visual trend is supported by regression models, in which the first principal component for EPIC predicted *PON1*₋₁₀₈ genotype, while the first principal component for SeqCap did not predict *PON1*₋₁₀₈ genotype (Table S2).

Discussion

We performed DNA methylation analyses of the *PON1* gene and nine additional genes in child leukocytes using the NGS-based technology, SeqCap, to generate DNA methylation data with base pair resolution. We compared our results with the same DNA analysed by the EPIC microarray, which queries ~850,000 CpG sites in the human genome [18] and is currently the most widely used platform for conducting EWAS in epidemiological studies. Since the samples were isolated from whole blood clots, the results of this study are generalizable to other studies analysing DNA methylation in leukocytes, which is a common sample type used in human studies [12,42].

We found an overall high concordance between SeqCap and EPIC DNA methylation, particularly among highly methylated and minimally methylated regions. However, noticeable differences were observed between the two methods in moderately methylated regions, with SeqCap measurements generally exhibiting greater within-site variation. The higher variability of methylation measurements in SeqCap became apparent when looking at a small number of candidate genes. The higher within-site methylation measurement variability in SeqCap inhibited the detection of biologically relevant differences in DNA methylation. While 8 of the 14 *PON1* CpG sites were significantly more methylated in association with the

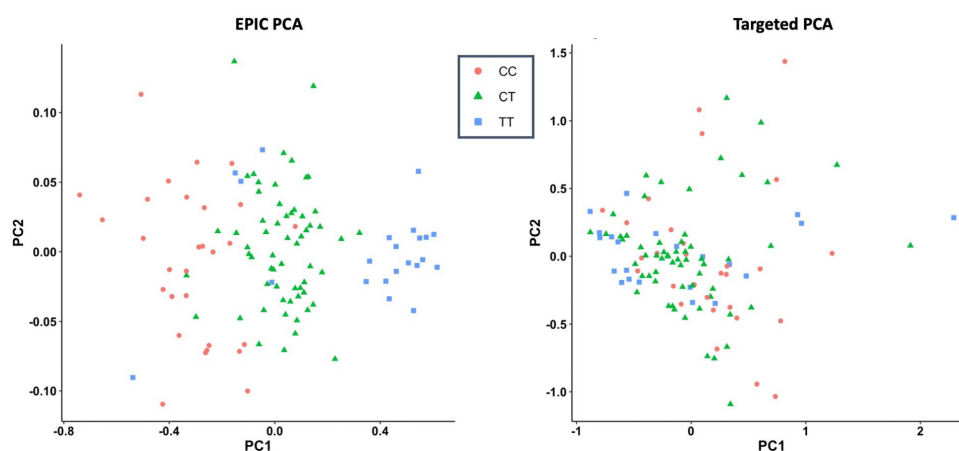


Figure 5. Principal component analysis of *PON1* methylation data. PCA was performed on both datasets, Illumina HumanMethylationEPIC BeadChip (EPIC PCA; left) and Roche Nimblegen, SeqCap Epi enrichment system (targeted PCA; right), using only the methylation data from the selected 14 CpG sites. First and second principal components are plotted from both analyses, with added colour based on *PON1*₋₁₀₈ polymorphism.

PON1₋₁₀₈ polymorphism (C alleles) when using EPIC data, none of those same sites at the *PON1*₋₁₀₈ were significant when using SeqCap data. However, the increased coverage from SeqCap revealed two novel CpG sites within the gene body that were significantly less methylated in association with the *PON1*₋₁₀₈ polymorphism. These findings are consistent with the current knowledge of the relationship between methylation and expression of the genes, because while more DNA methylation in the promoter region is usually associated with downregulated expression, higher DNA methylation in a gene body can be associated with higher gene expression [43,44]. This novel observation in *PON1* suggests that the substantial expansion of CpG site coverage in SeqCap provides the possibility of further novel findings in other genes that may be otherwise missed due to the limited coverage in EPIC arrays. However, this may be at the expense of greater accuracy as suggested by our data.

Several prior publications compared the use of arrays and targeted sequencing for the assessment of DNA methylation [13,15]. However, most of the studies have not conducted comparisons for scenarios encountered in environmental epidemiology, in which there are often large sample sizes and small effect sizes [18,42]. To address this gap in the literature, a recent study harnessed cord blood samples from the PROGRESS cohort to compare DNA methylation results from TruSeq with those obtained from the EPIC array, and evaluate both methods in terms of coverage, reproducibility, and identification of differential methylation by infant sex at birth[18]. Among many metrics, EPIC outperformed TruSeq in the identification of differential methylation by sex. The study suggested that although TruSeq offers greater read depth and coverage, it does not have the precision of microarrays, which are therefore still preferred for conducting EWAS in environmental epidemiology[18]. This finding was consistent with our study in which EPIC outperformed SeqCap in the identification of differential methylation by genotype.

Another study[17], using whole blood from adults participating in the Christchurch Health and Development longitudinal birth cohort study, aimed to determine whether the NGS-technology BSAS could validate prior EPIC array findings[45], have utilization as a replication, and/or expansion tool, as well as assess CpG sites residing in genomic regions not included on the EPIC microarray [17]. The study found that while BSAS validated EPIC array data at some loci and correlated across all loci, some individual loci did not validate, especially those in which the magnitude of change via the EPIC array was greater than 5%. The authors suggested that BSAS may serve as an investigative tool for specific genomic regions[17]. These findings were consistent with our study using whole blood, in which methylation levels correlated across loci as a whole but did not validate in those with moderate methylation in the EPIC array.

Studies using other biological matrices found more concordance between array and NGS-based methods, than those that used human blood from large cohort studies. For example, one recent experimental study utilized data from seven well-characterized human cell lines to compare DNA methylation results from a broad range of methylome sequencing assays and targeted approaches (NEBNext Enzymatic Methyl-Seq, Swift Biosciences Accel-NGS Methyl-Seq, SPLinted Ligation Adapter Tagging, NuGEN TrueMethyl oxBS-Seq, TruSeq, Nanopore) and the EPIC array. Overall high concordance between the various methods was observed with some notable differences in performance, as well as higher concordance between replicates when using microarray-based methods. There was also indication that a minimum amount of population variance may have driven poor concordance between assays in a subset of CpG sites[46]. In our study, genes containing CpG sites with a wide range of methylation (e.g., *SPI* and *PON2*) exhibited higher overall correlation coefficients than genes containing sites with similar methylation ranges (e.g., *AHR* and *TAPBP*).

Finally, another study examined the reproducibility and ability of the EPIC array, TruSeq, and HumanMethylation450 BeadChip (450 K) array to identify loci differentially methylated between

sample groups in a transformed prostate cancer cell line (LNCaP) and primary cell cultures of prostate epithelial cells (PrEC). The study found that methylation readings from the EPIC array and TruSeq were in high agreement, with Spearman correlation coefficients of 0.88 and 0.82 for the LNCaP and PrEC cell lines, respectively[30]. This was similar to the overall Pearson correlation observed in our study ($r = 0.84$).

The strength of our study includes the use of blood from a paediatric cohort study for the comparison of DNA methylation between two state-of-the-art technologies. In addition, we utilized the *PONI* gene, which was extensively studied in this cohort and serves as an ideal candidate to evaluate technological performance due to the strong associations between the *PONI*₁₀₈ SNP, gene methylation, and enzyme activity[19]. However, the major limitation was that while the EPIC array measured >850,000 CpG sites throughout the genome, our SeqCap assay was designed to measure CpG sites in 10 selected genes at base-pair resolution. This focus on a small number of candidate genes limited the comparison between the two methodologies to only 249 CpG sites located in 10 genes. The performance in other genomic regions or at larger scales may differ, so our results should be interpreted with caution. Although this reduced the number of CpG sites compared to an epigenome-wide approach, our approach more closely resembles a hypothesis-driven analysis of a specific selection of candidate genes. We also did not assess reproducibility, because while we had sample duplicates for DNA methylation measured on the EPIC array, we did not have duplicates for the SeqCap platform. In addition, we compared results from the EPIC array to just one candidate NGS-based technology, and it is possible that other technologies could have different performances[47].

In summary, our study agreed with prior validation studies, in that while SeqCap offered high coverage, the EPIC array outperformed in its ability to detect biologically meaningful differential DNA methylation across a range of methylation levels, which is often the goal in environmental epidemiology. While microarrays appear preferential for assessing differential methylation in cohort studies, NGS-based technologies could

supplementally identify differentially methylated loci not included on the microarray or obtain high coverage information on genes of interest. Since our study can only be generalized to the one NGS-based technology we assessed, future studies could compare the performance of the EPIC array to additional NGS-based technologies in large human cohort studies.

Acknowledgments

We would like to acknowledge and thank Philip Collender, MPH for processing and cleaning the EPIC array methylation data. We are also grateful to all the CHAMACOS participants, researchers, and field staff.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

This work was supported by the National Institutes of Health (R03 AG067064; R01 ES031259; R01 ES012503; R24 ES028529 and R01 ES021369).

Data availability statement

Due to the nature of this research, participants of this study did not agree for their data to be shared publicly, so supporting data is not available. Datasets generated and analysed during the current study are available from the corresponding author with appropriate permission from the CHAMACOS study team and investigators upon reasonable request and institutional review board approval.

ORCID

Dennis Khodasevich  <http://orcid.org/0000-0003-1412-8251>

Anna R. Smith  <http://orcid.org/0000-0003-1047-3859>

References

- [1] Handy DE, Castro R, Loscalzo J. Epigenetic modifications: basic mechanisms and role in cardiovascular disease. *Circulation*. 2011;123:2145–2156.
- [2] Smith A, Kaufman F, Sandy MS, et al. Cannabis exposure during critical windows of development: epigenetic and molecular pathways implicated in neuropsychiatric disease. *Curr Envir Health Rpt*. 2020;7:325–342.

- [3] Gardiner-Garden M, Frommer M. CpG islands in vertebrate genomes. *J Mol Biol.* **1987**;196:261–282.
- [4] Jones PA. Functions of DNA methylation: islands, start sites, gene bodies and beyond. *Nat Rev Genet.* **2012**;13:484–492.
- [5] Armstrong DA, Lesueur C, Conradt E, et al. Global and gene-specific DNA methylation across multiple tissues in early infancy: implications for children's health research. *FASEB J.* **2014**;28:2088–2097.
- [6] Robertson KD. DNA methylation and human disease. *Nat Rev Genet.* **2005**;6:597–610.
- [7] Kwok JB. Role of epigenetics in Alzheimer's and Parkinson's disease. *Epigenomics.* **2010**;2:671–682.
- [8] Low FM, Gluckman PD, Hanson MA. Developmental plasticity and epigenetic mechanisms underpinning metabolic and cardiovascular diseases. *Epigenomics.* **2011**;3:279–294.
- [9] Salameh Y, Bejaoui Y, El Hajj N. DNA methylation biomarkers in aging and age-related diseases. *Front Genet.* **2020**;11:171.
- [10] Wilkinson GS, Adams DM, Haghani A, et al. DNA methylation predicts age and provides insight into exceptional longevity of bats. *Nat Commun.* **2021**;12:1615.
- [11] Tammen SA, Friso S, Choi S-W. Epigenetics: the link between nature and nurture. *Mol Aspects Med.* **2013**;34:753–764.
- [12] Solomon O, MacIsaac J, Quach H, et al. Comparison of DNA methylation measured by Illumina 450K and EPIC BeadChips in blood of newborns and 14-year-old children. *Epigenetics.* **2018**;13:655–664.
- [13] Teh AL, Pan H, Lin X, et al. Comparison of methyl-capture sequencing vs. Infinium 450K methylation array for methylome analysis in clinical samples. *Epigenetics.* **2016**;11:36–48.
- [14] Sequencing R. SeqCapEpi CpGiant Probes. *Next Generation Sequencing (NGS) solutions* 2021. Accessed 2 March 2022; <https://sequencing.roche.com/content/rochesequence/en/support-resources/discontinued-products/seqcap-epi-cpgiant-enrichment-kit/resources.html>.
- [15] Sun Z, Cunningham J, Slager S, et al. Base resolution methylome profiling: considerations in platform selection, data preprocessing and analysis. *Epigenomics.* **2015**;7:813–828.
- [16] Masser DR, Stanford DR, Freeman WM. Targeted DNA methylation analysis by next-generation sequencing. *J Vis Exp.* **2015**;52488. DOI:10.3791/52488.
- [17] Noble AJ, Pearson JF, Boden JM, et al. A validation of Illumina EPIC array system with bisulfite-based amplicon sequencing. *PeerJ.* **2021**;9:e10762.
- [18] Heiss JA, Brennan KJ, Baccarelli AA, et al. Battle of epigenetic proportions: comparing Illumina's EPIC methylation microarrays and TruSeq targeted bisulfite sequencing. *Epigenetics.* **2019**;15:174–182.
- [19] Huen K, Yousefi P, Street K, et al. PON1 as a model for integration of genetic, epigenetic, and expression data on candidate susceptibility genes. *Environ Epigenet.* **2015**;1. DOI:10.1093/eep/dvv003.
- [20] Eskenazi B, Bradman A, Gladstone EA, et al. CHAMACOS, A longitudinal birth cohort study: lessons from the fields. *J Children's Health.* **2003**; 1: 3–27.
- [21] Yousefi P, Huen K, Schall RA, et al. Considerations for normalization of DNA methylation data by Illumina 450K BeadChip assay in population studies. *Epigenetics.* **2013**;8:1141–1152.
- [22] Draganov DI, Teiber JF, Speelman A, et al. Human Paraoxonases (PON1, PON2, and PON3) are lactonases with overlapping and distinct substrate specificities. *J Lipid Res.* **2005**;46:1239–1247.
- [23] Bryk B, BenMoyal-Segal L, Podoly E, et al. Inherited and acquired interactions between ACHE and PON1 polymorphisms modulate plasma acetylcholinesterase and Paraoxonase activities. *J Neurochem.* **2005**;92:1216–1227.
- [24] Gouédard C, Barouki R, Morel Y. Induction of the paraoxonase-1 gene expression by resveratrol. *Arterioscler Thromb Vasc Biol.* **2004**;24:2378–2383.
- [25] Rhee I, Bachman KE, Park BH, et al. DNMT1 and DNMT3b cooperate to silence genes in human cancer cells. *Nature.* **2002**;416:552–556.
- [26] Osaki F, Ikeda Y, Suehiro T, et al. Roles of Sp1 and protein kinase C in regulation of human serum paraoxonase 1 (PON1) gene transcription in HepG2 cells. *Atherosclerosis.* **2004**;176:279–287.
- [27] Solomon O, Yousefi P, Huen K, et al. Prenatal phthalate exposure and altered patterns of DNA methylation in cord blood. *Environ Mol Mutagen.* **2017**;58:398–410.
- [28] Holland N, Furlong C, Bastaki M, et al. Paraoxonase polymorphisms, haplotypes, and enzyme activity in latino mothers and newborns. *Environ Health Perspect.* **2006**;114:985–991.
- [29] Bibikova M, Barnes B, Tsan C, et al. High density DNA methylation array with single CpG site resolution. *Genomics.* **2011**;98:288–295.
- [30] Pidsley R, Zotenko E, Peters TJ, et al. Critical evaluation of the Illumina MethylationEPIC BeadChip microarray for whole-genome DNA methylation profiling. *Genome Biol.* **2016**;17:208.
- [31] Sandoval J, Heyn H, Moran S, et al. Validation of a DNA methylation microarray for 450,000 CpG sites in the human genome. *Epigenetics.* **2011**;6:692–702.
- [32] Aryee MJ, Jaffe AE, Corrada-Bravo H, et al. Minfi: a flexible and comprehensive bioconductor package for the analysis of Infinium DNA methylation microarrays. *Bioinformatics.* **2014**;30:1363–1369.
- [33] Niu L, Xu Z, Taylor JA. RCP: a novel probe design bias correction method for Illumina methylation BeadChip. *Bioinformatics.* **2016**;32:2659–2663.
- [34] Johnson WE, Li C, Rabinovic A. Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics.* **2007**;8:118–127.

- [35] Leek JT, Johnson WE, Parker HS, et al. The sva package for removing batch effects and other unwanted variation in high-throughput experiments. *Bioinformatics*. 2012;28:882–883.
- [36] UCSC Genome Browser. UCSC genome browser gateway. Accessed 2 March 2022. University of California Santa Cruz Genomics Institute. <https://genome.ucsc.edu/cgi-bin/hgGateway>.
- [37] UCSC Genome Browser. Lift genome annotations. Accessed 2 March 2022. University of California Santa Cruz Genomics Institute. <https://genome.ucsc.edu/cgi-bin/hgLiftOver>.
- [38] Guo W, Fiziev P, Yan W, et al. BS-Seeker2: a versatile aligning pipeline for bisulfite sequencing data. *BMC Genomics*. 2013;14:774.
- [39] Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. *Nat Methods*. 2012;9:357–359.
- [40] Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*. 2010;26:841–842.
- [41] Du P, Zhang X, Huang -C-C, et al. Comparison of beta-value and M-value methods for quantifying methylation levels by microarray analysis. *BMC Bioinformatics*. 2010;11:587.
- [42] Breton CV, Marsit CJ, Faustman E, et al. Small-magnitude effect sizes in epigenetic end points are important in children’s environmental health studies: The Children’s Environmental Health and Disease Prevention Research Center’s Epigenetics Working Group. *Environ Health Perspect*. 2017;125:511–526.
- [43] Jjingo D, Conley AB, Yi SV, et al. On the presence and role of human gene-body DNA methylation. *Oncotarget*. 2012;3:462–474.
- [44] Wen K. X, Milic, J., El-Khodor, B, et al. The role of DNA methylation and histone modifications in neurodegenerative diseases: a systematic review. *PLoS One*. 2016;11:e0167201.
- [45] Osborne AJ, Pearson JF, Noble AJ, et al. Genome-wide DNA methylation analysis of heavy cannabis exposure in a New Zealand longitudinal cohort. *Transl Psychiatry*. 2020;10:1–10.
- [46] Foox J, Nordlund J, Lalancette C, et al. The SEQC2 epigenomics quality control (EpiQC) study. *Genome Biol*. 2021;22:332.
- [47] Tanić M, Moghul, I., Rodney, S., et al. Comparison and imputation-aided integration of five commercial platforms for targeted DNA methylome analysis. *Nat Biotechnol*. 2022;1–10. DOI:10.1038/s41587-022-01336-9.