clsa élcv
Canadian Longitudinal Study on Aging
Étude longitudinale canadienne sur le vieillissement

Genome-Wide DNA Methylation Profiling on 1,478 Participants using Illumina Infinium MethylationEPIC BeadChip Microarray Technology
Data Support Document

David Tse Shen Lin, Julia L. MacIsaac, Katia E. Ramadori, Chris P. Verschoor, Lisa M. McEwen, Michael S. Kobor, Andrew Paterson, Divya Joshi and Cynthia Balion

Abstract

DNA methylation represents one of the most well-studied epigenetic marks, which has previously been shown to influence gene expression without altering the genomic sequence within an individual. Investigating this mechanism within a population can provide insight into how the environment, including the process of aging, can influence cellular function and potentially an individuals' risk of adverse health outcomes. In this data release, we profiled genome-wide DNA methylation in peripheral blood mononuclear cells (PBMCs) isolated from 1,478 selected participants enrolled in the Canadian Longitudinal Study on Aging (CLSA) using the Illumina Infinium MethylationEPIC BeadChip microarrays (hereinafter referred to as *EPIC arrays*), which provides quantitative measurements of DNA methylation at 862,927 CpG sites and 2,932 CHH sites throughout the human genome. We performed data preprocessing that included sample- and array-based quality assessments, probe filtering, outlier analyses, data normalization, batch corrections, and cell-type estimations and adjustments. An exemplar epigenome-wide association study (EWAS) on chronological age using the preprocessed methylation data (1,445 participants, 783,136 loci) validated numerous previously reported age-associated DNA methylation changes. In addition to the methylation raw data, we also provide the preprocessed methylation dataset, as well as estimated epigenetic ages calculated using the established Horvath, Hannum, PhenoAge, and GrimAge epigenetic clock algorithms in this data release. Qualified researchers can access this epigenetics data release via the CLSA Data Access portal.

Contents

1.0 INTRODUCTION

1.1 Epigenetics and DNA methylation

Epigenetics is broadly defined as heritable changes in gene expression without alterations in the underlying DNA genomic sequence[1]. Epigenetics manifests through several mechanisms, including histone modifications, non-coding RNA elements, and DNA methylation (DNAm), which represents one of the best studied epigenetic modification in mammals to-date[2,3]. DNAm refers to the addition of a methyl (-CH_3) group, that generally occurs on the 5' position of cytosines at predominantly cytosine-guanine dinucleotide (CpG) sites[2,3]. DNAm represents highly dynamic molecular marks that are associated with environmental factors, as well as development, health, and aging[4]. In fact, DNAm data has been used to train and build indices, often referred to as "epigenetic clocks", that can accurately estimate an individual's "biological age" that is highly correlated to their chronological age[5–7]. Therefore, examination of DNAm in population studies can shed light into mechanisms behind how the environment and the process of aging can influence cellular functions, and potentially health outcomes.

Participants from the Canadian Longitudinal Study on Aging (CLSA), a national long-term study that follows 50,338 men and women (n = 30,097 Comprehensive cohort and n = 21,241 Tracking cohort) between 45 and 85 years of age for at least 20 years, were selected for DNAm profiling for researchers to critically examine the relationship between epigenetics, biological aging, and health outcomes[8]. Of the 30,097 participants in the Comprehensive cohort at baseline, 23,492 participants had provided blood and urine samples and had availability of EDTA whole blood and Buffy coat. Of these 23,492 participants, 6,268 participants had fasted for 5 or more hours, and of which, 3,000 were selected for genomics and metabolomics analyses. Additionally, from the remaining 20,492 participants, 7,000 participants were selected for genomics and metabolomics analyses. Of the 10,000 participants, a sub-sample of 1,500 participants were subsequently selected for epigenetic analysis, of which 1,478 participants were successfully assayed, and 1,445 participants passed stringent quality control assessments **(Figure 1)**. All sample selections were made to reflect the distribution of Comprehensive cohort by age, sex, and data collection site. This data support document outlines the procedures followed to generate, quality control, and preprocess the DNAm data available in the CLSA epigenetics data release.

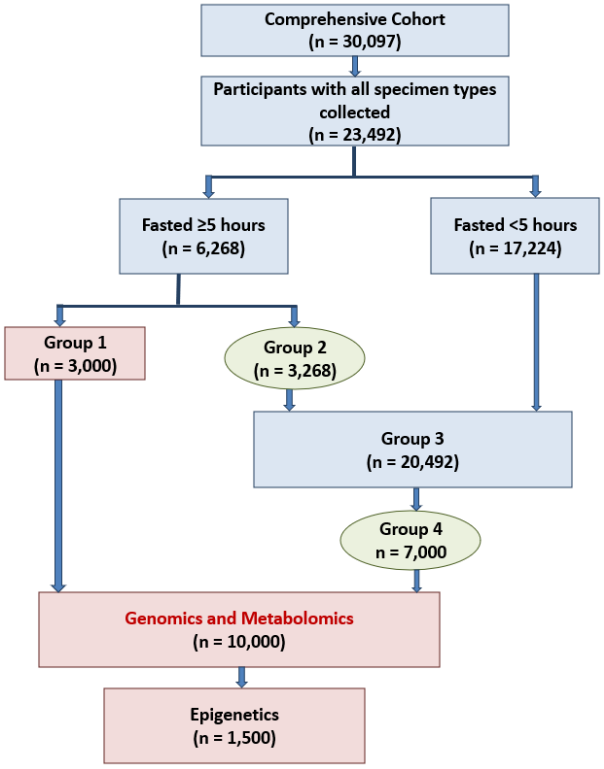

Figure 1. Process of selection of study participants for epigenetic analysis.

1.2 The Illumina Infinium MethylationEPIC BeadChips

The Illumina Infinium MethylationEPIC BeadChip arrays (hereon referred to interchangeably as *EPIC arrays*), the platform used in this study, is the latest human DNA methylation array developed and released by *Illumina* Inc[9]. The EPIC arrays are capable of quantitatively interrogating DNA methylation levels on over 850,000 CpG sites across the genome, including all designable RefSeq genes, with CpG Island shores, non-island CpGs, CpG islands outside of coding regions and miRNA promoter regions represented. In addition, non-CpG sites found to be differentially methylated in human stem cells, as well as regions identified in GWAS studies to be disease-associated, are included[9]. This high genomic coverage, combined with the capacity for high sample-throughput (96-384 samples/week) and standardized analyses pipelines, makes the EPIC chip the ideal platform to carry out genome-wide DNA methylation studies in large-scale population-based cohorts.

1.3 About this data release

This data release contains the raw DNA methylation data derived from peripheral blood mononuclear cell (PBMC) samples for 1,478 successfully assayed CLSA participants across 865,918 total genomic sites (consisting of 862,927 CpG probes, 2,932 CHH probes and 59 SNP probes), quantitatively measured using the EPIC arrays. To facilitate users who are less familiar with DNA methylation data preprocessing, we also provide a set of beta values for 1,445 samples at 783,136 CpGs/CHHs which passed stringent quality control assessments (described below) and have been subsequently batch corrected and corrected for blood cell type variations.

The EPIC chip annotation file can be directly accessed and downloaded at

ftp://webdata2:webdata2@ussd-ftp.illumina.com/downloads/productfiles/methylationEPIC/infinium-methylationepic-v-1-0-b4-manifest-file-csv.zip[10]. In addition to the site-specific DNA methylation data, this data release also contains epigenetic age-related measurements for all 1,445 participants, calculated from the raw DNA methylation data that were background-subtracted and colour-corrected using Illumina *GenomeStudio* software.

1.3.1 CLSA Epigenetic Data Files and File Formats

The following data are available for access by researchers that have followed the policies and guidelines set forth by the CLSA:

1. The raw DNA methylation data containing beta values, representing percent methylation (0% - 100%; scaled to a numeric value between 0 – 1) at each of the 865,918 probes for all 1,478 participants on the EPIC arrays (provided as .idat files or as a beta-value matrix derived from the MethylumiSet object in the CSV file format).
2. The color-corrected/background-subtracted, probe-filtered, sample outlier-removed, inter- and intra-sample normalized, batch and blood cell-type corrected beta-values at 783,136 probes for 1,445 participants as a beta-value matrix in the CSV format. Note the normalization procedures do not transform the data, beta-values remain on a scale of 0-1.
3. Epigenetic-age related measures for all 1,445 participants from the Horvath DNAmAge clock calculator, using the *GenomeStudio* colour-corrected and background-subtracted methylation data as input. All columns are of datatype numeric. This is a CSV file with the following column headings:

    1. DNAmAge (DNAmAge2) – This is the absolute DNA methylation/epigenetic age estimates calculated based on the Horvath 353-CpG Pan-Tissue clock sites in the units of biological years. This value usually shows a high correlation with a given individual's chronological age and is in theory unconfounded with cell type proportions[5].
    2. Age Acceleration Difference (AgeAccelerationDiff) – Absolute difference between chronological age and DNAmAge for an individual – i.e., is calculated as (DNAmAge - Chronological Age).
    3. Age Acceleration Residual (AgeAccelerationResidual) – Represents an epigenetic age acceleration measure defined as residual from regressing DNAmAge on chronological age – this is typically used as the universal measure of epigenetic age acceleration[5].
    4. Intrinsic Epigenetic Age Acceleration (IEAA) – Represents an epigenetic age acceleration estimate measure that is attributed to *intrinsic changes within the cells* regardless of cell type proportions in a given sample. IEAA is measured by accounting for both an individual's chronological age and blood cell type proportions[11].
    5. Extrinsic Epigenetic Age Acceleration (EEAA) – Represents an epigenetic age acceleration estimate measure that is attributed to *age-related changes in blood cell type composition*. EEAA is more related to immune system aging[11].
    6. Hannum Epigenetic Age (Hannum_Age2) – Epigenetic Age calculated based on 71 CpG sites as defined by Hannum et al. This clock was developed using whole blood samples[6].
    7. PhenoAge (DNAmPhenoAge) – The PhenoAge estimate is based on a phenotypic age score derived from chronological age and nine clinically relevant blood biomarkers.[7]
    8. GrimAge (DNAmGrimAge) – The GrimAge estimate was developed using DNA methylation-based surrogates for 7 age-related plasma proteins and a DNA methylation-based estimator of smoking pack-years.[49] Age and sex were included as covariates.

1.4 DNA extraction, data processing, and reporting

1.4.1 Sample storage and DNA extraction

Peripheral blood multinuclear cell (PBMC) fractions were isolated from whole blood draws by Ficoll separation, and approximately 400,000 – 1,000,000 PBMCs were aliquoted from each sample, frozen in 200µL phosphate-buffered saline (PBS) and stored in 2D barcode screw-top storage tubes in -80°C long-term. Samples were transferred to LN2 storage up to one week until shipment to the epigenomics facility, where the samples were placed into -80°C storage immediately upon receipt until further processing. Genomic DNA was extracted from 48 PBMC resuspension samples at a time using a QIAsymphony workstation. Briefly, samples were retrieved from the ultra-low freezer, thawed at room temperature for 5-10 minutes, then transferred to 2mL Sarstedt PP microtubes compatible for DNA extractions with the QIAsymphony workstation using the Qiagen DNA Midi kits and the custom program “BC400 CR22014 ID2282”. Genomic DNA were eluted in 50µL of Qiagen elution buffer (EB; 10mM Tris-Cl, pH 8.5), and quantified on a Nanodrop-8000 using 1µL from each sample. All extracted genomic DNA samples displayed good quality as measured by absorbance (A260/A280 = 1.6-2.0; A260/A230 > 1.5). DNA concentrations were subsequently normalized to 20ng/µL with Qiagen EB and stored in -20°C until bisulfite conversion, as described in the next section.

1.4.2 DNA methylation Arrays with EPIC Arrays

The process of DNA methylation profiling starts with bisulfite conversion of the extracted, high-purity genomic DNA samples. This initial step converts unmethylated cytosines to uracils, which further converts into thymines through polymerase chain reactions (PCRs), while leaves methylated cytosines intact[12]. The conversion of methylation status into sequence information leads to distinguishing reads on the EPIC arrays. Bisulfite conversions were carried out using the Zymo EZ DNA Methylation kits in 16 batches of ~96 samples per batch, following manufacturer’s instructions, using 750ng genomic DNA as input for each sample. Bisulfite-converted DNA samples were eluted in 12µL of M-Elution buffer (Zymo), quantified using a Nanodrop-8000, and stored in -80°C until EPIC arrays processing.

160ng of each bisulfite-converted DNA samples were used as the input for EPIC arrays, which were processed following manufacturer’s instructions, in 16 separate batches (~96 samples per batch, with the exception of batch 16, which contains only 80 samples). To control for data quality, each batch contained one control sample that was derived from a single bisulfite conversion reaction as a technical replicate, placed at a random position. Note, to account for potential positional bias throughout the experiments, all sample placements were randomized once at the bisulfite conversion procedure, and again during the array procedures.

1.4.3 Data acquisition, export, and reporting

Processed EPIC arrays were scanned using Illumina *iScan* and the *iScan Control Software* within 24 hours of completion of each array plate of 96 samples. The *iScan* reads and stores sample intensity information at each probe for every sample in the *.idat* format, which is provided as the raw data in this data release. Following the completion of scanning for all samples, all *.idat* files were imported into Illumina *GenomeStudio* software, with the colour-correction and background-subtraction options selected. Beta-values representing the percentage of methylation of each probe at a given sample were calculated in *GenomeStudio* using the following formula:

$$= \frac{Max(SignalB, 0)}{Max(SignalA, 0)+Max(SignalB, 0)+100}$$

For Infinium I assays, signal A and signal B are produced by two different bead types (representing the methylated & unmethylated probes, respectively), and reported in the same color. For Infinium II assays, signal A corresponds to the signal in the Red channel and signal B corresponds to the signal in the Green channel[13,14]. As each probe is represented by multiple beads on the arrays, an average beta-value is reported for each locus. All reported beta values fall between the range of 0 – 1, represent percentage methylation from 0% to 100%, with no additional transformation procedures performed on the data.

2.0 SAMPLE- AND MARKER-BASED QUALITY CONTROL

In the first stage of methylation data inspection, we perform basic quality controls after Illumina *GenomeStudio* data import to determine whether any given sample has underperformed, either due to inadequacies of the input genomic DNA quantity/quality, or the array procedures. These quality parameters include: the examination of overall sample array intensities, global methylation profiles, probe detection-p values, performance of the control probes (focusing on only bisulfite conversion probes in this report). We also ascertain sample identities by metadata cross-validation using participant sex and genetic information. The details of these procedures are described in these sections.

2.1 Sample-based quality control: Overall sample qualities

2.1.1 Sample methylation intensity inspections

We inspected the log-transformed median intensity values of both methylated (M) and unmethylated (U) channels in all samples, where good quality samples are expected to cluster together, and failed samples tend to cluster out and have lower median intensities. These intensity signals were derived from all data in the form of a collective *RGChannelSet* object (R Package: *minfi*; v1.30.0), which stores all raw green and red channel signal intensities of each sample[15]. Using the recommended threshold cutoff of *10.5*, all samples displayed good signal intensities above this value.

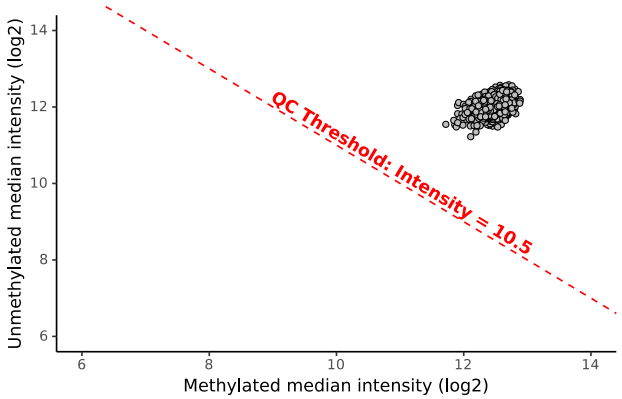

Figure 2. Median intensities (log2-transformed values) of the methylated and unmethylated signals for all 1,478 assayed CLSA PBMC samples on the EPIC arrays. All samples passed the QC threshold of 10.5.

2.1.2 Overall sample methylation profiles – Beta-value distributions

To perform further quality control inspections, we first created a *MethylumiSet* object (R package: *Methylumi*; v2.30.0)[16], which stores information including the average beta values at each defined probe, the PhenoData and the FeatureData (the probe annotations). We extracted the beta values using the *betas* functions and created a density plot for each sample. We observe bimodal distributions for all samples, with peaks approaching beta values of 0 and 1 (**Figure 3**). This is typically expected from the methylation array data and reflects the biology – most of the measured sites are either overall hypo-or hyper-methylated: Since the array is enriched for promoter-associated CpG islands and enhancers, these are typically refractory to DNA methylation across healthy individuals[17]. On the other hand, the EPIC array also targets a high amount of intergenic regions, which are often methylated[18]. Because none of the samples exhibit severe atypical distributions at this point, we retain all samples after this analysis.

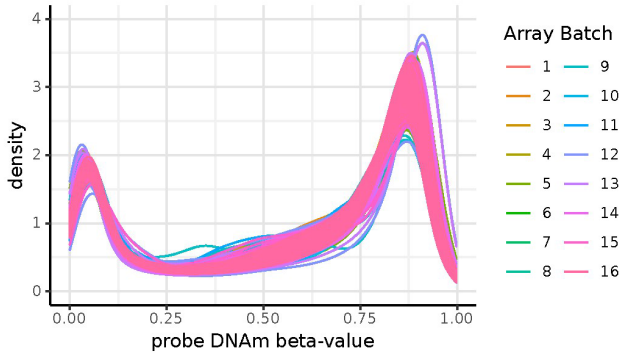

Figure 3. Beta-value distributions of all 1,478 samples represented by density plots. Each line depicts the beta-value profile of a unique sample, colored by the array batch in which the sample was processed.

2.1.3 Sample average detection p-values

The detection *P*-value is defined as 1 minus the *P*-value computed from the background model, characterizing the chance that the signal was distinguishable from negative controls[19]. We use the average detection *P*-value aggregated from all available probes as one of our first quality controls metrics and set an arbitrary threshold at 0.005, as recommended by *Illumina*. **Figure 4** shows that all of the samples assayed achieved an average detection *P*-value under 0.005, therefore, all samples are retained at this point.

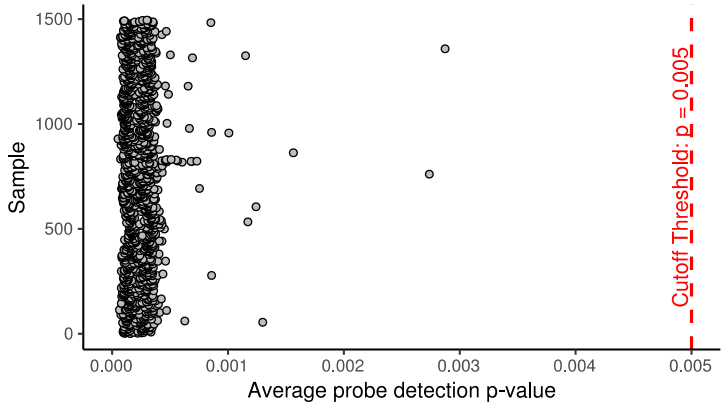

Figure 4. Average detection p-value calculated from all EPIC chip probes for all 1,478 assayed samples. Red dashed line indicates the quality control cutoff set at detection p = 0.005.

2.1.4 Control probe performances & Sample bisulfite conversion efficiencies

For comprehensive quality control purposes, we examined the set of built-in control probes available on the EPIC arrays. These different internal probes can be used to assess metrics that are array-dependent (array staining, single-base pair extension, hybridization) and sample-dependent (i.e., specificity)[20]. We note no quality outliers using these criteria (*data not shown*).

An additional sample-dependent quality control is bisulfite conversion efficiency, where low conversion efficiency can represent a source of error in the final array analyses results. Therefore, we assessed the bisulfite conversion efficiency of each sample, using the *bscon* function in the R *WateRmelon* (v1.28.0) package, which estimates this parameter based on a set of built-in control probes on the arrays[21]. We set a stringent threshold of 85% conversion efficiency based on the recommendation of Wong *et al*.[22], and identified four samples that fell below this standard (**Figure 5**): **54, 752, 946, and 1341**. These four samples were filtered out and excluded from the downstream preprocessing.

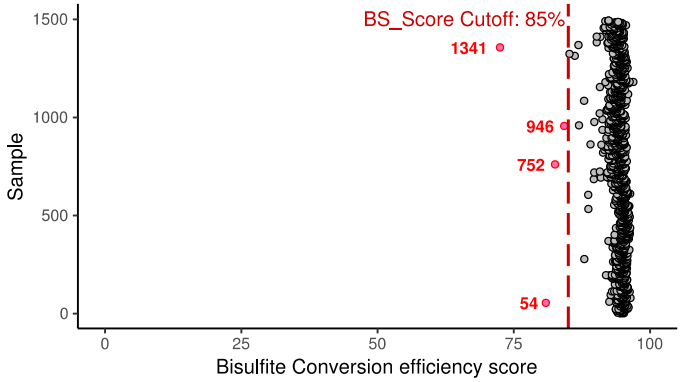

Figure 5. Bisulfite conversion efficiency of all 1,478 assayed samples inferred from the control probes as calculated by the *bscon* function in the *wateRmelon* package. Red dashed line indicates the quality control cutoff set at 85%. Four samples below this threshold are flagged in red.

2.1.5 Control sample performances: Correlations & multi-dimensional scaling

We examined the Pearson's pairwise correlation coefficients between the 16 control samples that were run throughout the experiment (one per batch of 96 samples). The resulting coefficient values when comparing each of the replicates to replicate 1 in the first array batch were all greater than r = 0.9967 (equivalent to R2 > 99.3%), well above the expected 98% as advised by Illumina[9]. All other samples show Pearson's correlations of r <0.9917 (**Figure 6A**).

Lastly, we carried out multi-dimensional scaling (MDS) on the raw methylation data to infer sample relations. This was done with the *plotSampleRelation* function in the *Lumi* package (v2.36.0)[23]. A plot of the top two principal components show that methylation data is mainly segregated by sample sex on principal component 2, which accounts for 12.4% of the methylation variance. As expected, the 16 control samples are clustered tightly together, confirming the robustness in the array results and the consistency between the different array runs (**Figure 6B**).

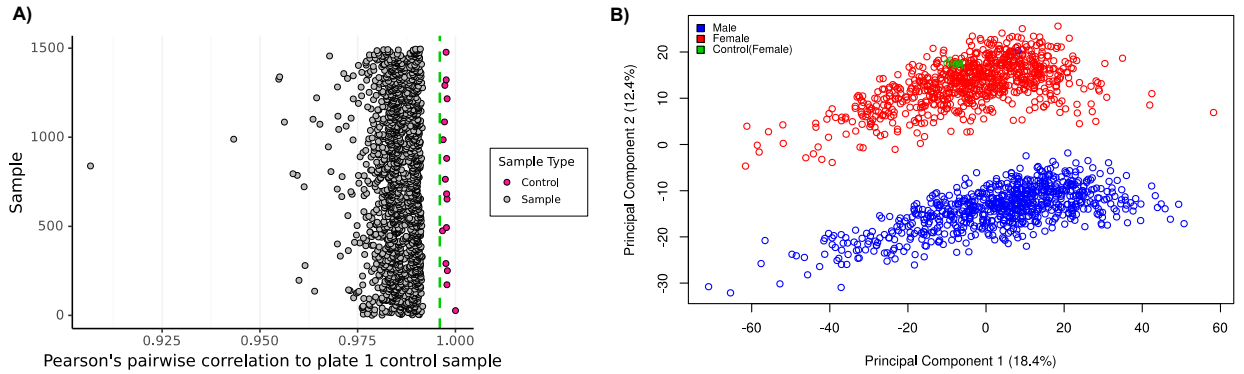

Figure 6. Performances of the control samples in the CLSA DNAm array dataset. **A)** A dot plot showing Pearson's pairwise correlation values of each sample, computed from all available beta-values, using the control sample on plate 1 as reference. Green dashed line: $r = 0.995$. **B)** A plot showing the top two principal components derived from the MDS algorithm on all CLSA samples using raw DNA methylation data. Each dot represents a unique sample colored by participant sex information, and control sample status.

2.2 Sample-based quality control: Sample identity matching

2.2.1 Sex matching

Global methylation profiles specific to the X chromosome exhibit dimorphic patterns in males versus females. It is known that in biological females, >60% probes exhibit intermediate methylation (beta value between 0.2 and 0.7) regardless of cell types, leading to a global trimodal beta-value distribution[24]. Biological males, on the other hand, display low methylation levels in a majority of these probes and an overall bimodal distribution[24]. In addition, we would expect biological females to display non-specific methylation on Y-chromosomal probes as they are absent, while exhibiting the expected bimodal distributions in males. We found almost all assayed individuals matched in biological sex based on the sex chromosome beta-value distribution patterns, with the exception of two samples: Samples **#918** and **#1057**, both of which exhibited a "female-like" X-chromosome but a "male-like" Y-chromosome methylation profile (**Figure 7**). Because these findings do not suggest a complete mismatch (**importantly, they matched by SNPs, as discussed in the next section**), but rather the possibility of trisomy of the sex chromosomes (i.e., XXY), and these samples showed the typical bimodal distributions in the autosomes, we flagged these two samples but have left them in for sample preprocessing. *Note that depending on the users' study questions and analysis models, they can decide whether these samples should be included in their studies at their own discretions.*

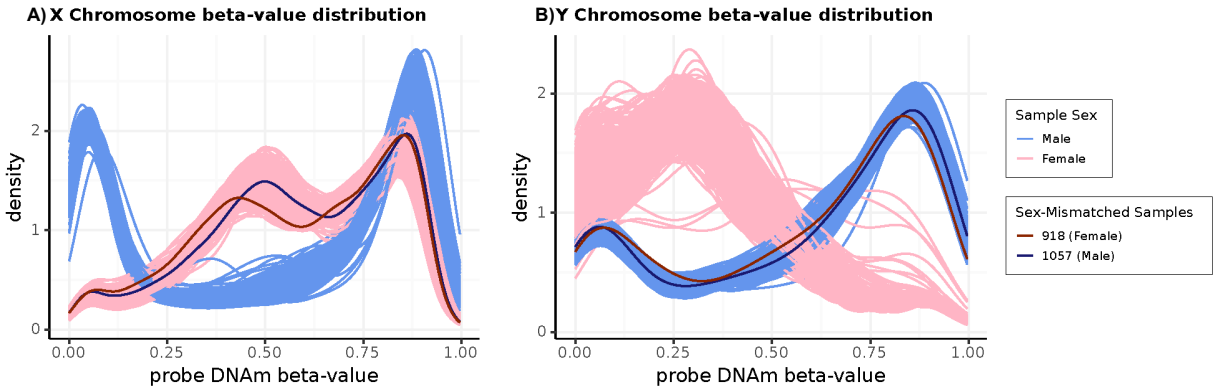

Figure 7. Beta-value distribution of DNA methylation on the X and Y chromosomes of all 1,478 participants in the CLSA study cohort. **A)** A density plot showing beta-value distributions of each sample on the X chromosome; **B)** A density plot showing beta-value distributions of each sample on the Y chromosome. Each line represents a unique sample, colored by reported sex. Two samples (#918, #1057) showing unexpected profiles are colored in maroon and dark blue, respectively in both plots.

2.2.2 SNPs marker matching

The EPIC chip contains 59 single nucleotide polymorphism (SNP) probes for genetic fingerprinting purposes[18]. To ensure the absence of any cross-contamination or identity mismatching in the assayed samples, we leveraged the CLSA genetic dataset[25] to perform an identity check by marker matching. We extracted 57 SNPs that are present in both datasets (In the genetic data, genotyped using the UK Biobank Axiom Arrays: 5 of the EPIC chip SNPs were directly measured, and 52 were imputed), and matched the sample identities using this information. Specifically, allele calls were extracted from the CLSA *.bgen* files for SNPs of interest, while the EPIC SNPs were converted to allele calls based on their beta values to facilitate cross-checking. No-calls in either dataset were dropped on a sample-to-sample basis. To account for possible call inaccuracies to genetic imputation and/or SNPs probes that failed to be called on the EPIC arrays, we elected for an arbitrary threshold of <85% match for designation of sample mismatches. Overall, over 99% of the samples showed >90% matches between these SNP calls, confirming identity match across the platforms based on genetic fingerprinting.

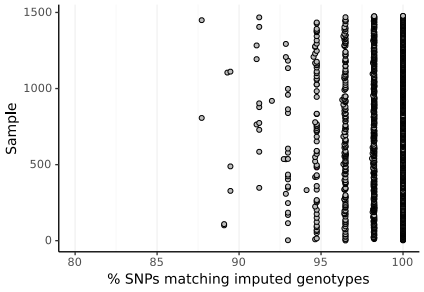

Figure 8. SNP measurements on the EPIC arrays confirm the identity of all 1,478 CLSA participants via cross-checking with the imputed genotypes in the CLSA genotyping dataset. Each dot represents a unique sample. % SNPs matching was calculated as (matching SNPs / all SNP data available), represented as percentages.

2.3 Marker-based quality control

2.3.1 Probe filtering

To get the dataset to the normalization stage, we first remove probes on the EPIC arrays whose profiles deviate from the expected bimodal distributions. These include the 59 SNP probes inspected for genetic fingerprinting as described in section 2.2.2, as well as 19,627 probes found on the sex chromosomes, which were illustrated in section 2.2.1.

In addition, we remove a small subset of probes that were annotated to provide unreliable measurements due to flaws in their design[18]. These include: 43,254 *cross-reactive* probes that have >47 base pair homology with an off-target sites; as well as a total of 12,791 probes on the EPIC arrays which are *polymorphic* either at the targeted CpG sites (12,378 probes), or at the single base extension sites for Infinium Type I probes (413). As some of these annotated probes are located on the X and Y chromosomes, there were 41,773 cross-reactive and 11,520 polymorphic autosomal probes that were taken out in this step of the data processing.

2.3.2 Probe quality control using *pfilter*

We assessed the performance quality and detection *P*-value at the probe level to determine underperforming probes across a large proportion of samples, prompting for their removal prior to data normalization. This was carried out using the *pfilter* function (R package: *wateRmelon*)[21], with the default quality thresholds. *pfilter* reported 8,561 probes that had beadcounts <3 in >5% of the samples, and 1,242 additional sites exhibiting detection p-values > 0.05 in >1% of the samples. These 9,803 probes with sub-par qualities were dropped out from the final dataset.

A summary of the number of total probes removed and remaining in the dataset is presented in **Figure 9.**

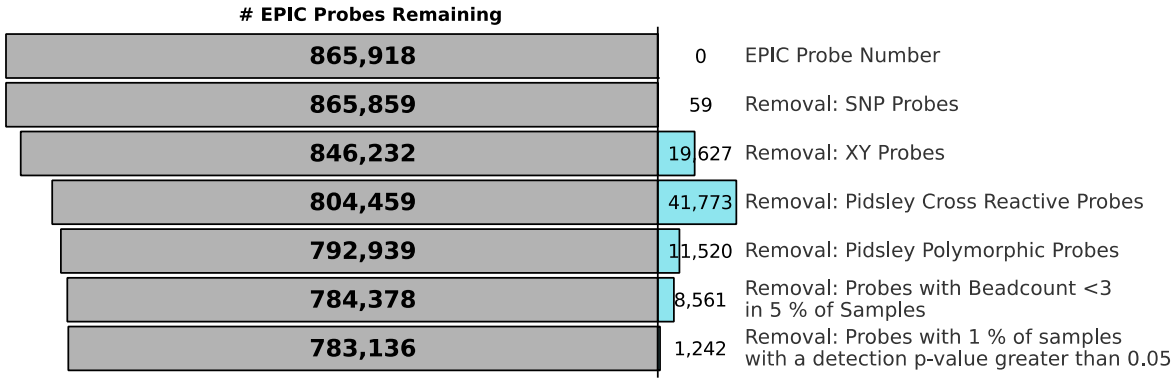

Figure 9. A probe attrition plot of the CLSA DNA methylation dataset summarizing the number of probes remaining after each stage of the probe filtering process.

2.4 Summary of sample & marker-based quality control

We removed 4 samples due to low bisulfite conversion efficiencies as measured by the array control probes, and 82,782 probes due to them being SNPs, located on the sex chromosomes, or having poor designs or performances. 1,474 samples (with 16 additional control samples) and 783,136 probes remain at this stage.

3.0 OUTLIER SAMPLE ANALYSES

3.1 Imputation of missing beta-values

Following filtering of poorly designed and underperforming probes, we use the remaining probes to perform some analyses that informs us of statistical outlying samples. The samples that are defined as outliers at this step will be removed from the subsequent sample normalizations so as not bias the global beta-value distributions.

Prior to carrying out these outlier analyses, we first imputed missing beta-values in the dataset at this stage. To infer these missing values, we first confirmed that the NA count is present in <5% for every sample and every remaining CpG features, then we elected to impute with the *impute.knn* function (R package: *impute*; v1.58.0) using the default settings, which uses the average of k = 10 nearest neighbours[26]. The resulting dataset was verified to contain no missing values and is used for principal component analysis (PCA), as described in the next section.

3.2 Principal component analysis (locfdr)

As a first approach to identify potential statistically outlying samples, we performed a principal component analysis (PCA) based on the remaining 1,490 (including controls) samples and 783,136 probes, and inspected sample distributions for outliers in the first PC as described in Hannum *et al.*[6] Briefly, each sample is converted into a z-score statistic based on the squared distance of its first PC from the population mean. The z-statistic was then converted to a local false-discovery rate (locfdr) by the Gaussian cumulative distribution, and the Benjamini-Hochberg procedure. Using this method, 72 Samples falling below a converted FDR threshold of 0.2 were designated as statistical outliers.

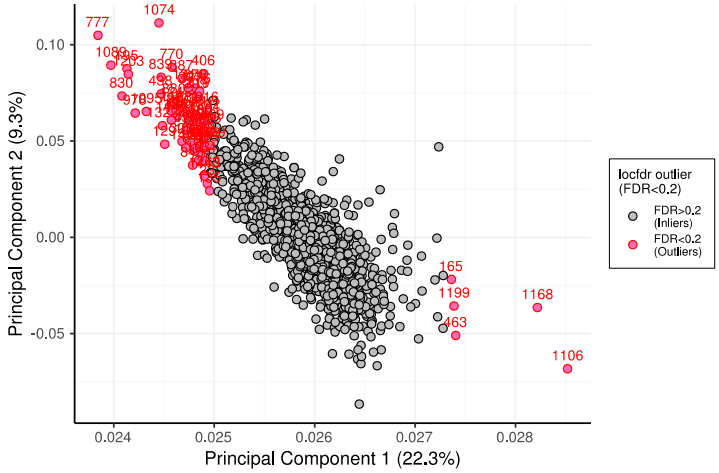

Figure 10. Identification of statistical outliers in the CLSA DNA methylation dataset by principal component analysis. A PCA was conducted using all probes remained after probe filtering, and the top two PCs are shown here. Locfdr outliers were computed from the first PCs and samples identified as statistical outliers (local false discovery rate <0.2) are colored and tagged in red.

3.3 Beta-value correlation inspections

Next, we examined the sample relationships by calculating pairwise Pearson's correlations between all samples based on all the probes. We expect significant statistical outliers to display low levels of correlations ($r < 0.60$) with a large proportion of the rest of samples. This could either be due to either severely subpar sample quality issues or extreme cell type differences from the normal distribution of all samples. Clustering showed that these outliers cluster out based on their sample relations – this is illustrated as a heatmap representing sample-sample correlation values in **Figure 11**, where we again show the controls clustering together (a red block showing a cluster of samples displaying high correlations, near the center of the figure). 20 samples clustering out to the right are flagged as statistical outliers.

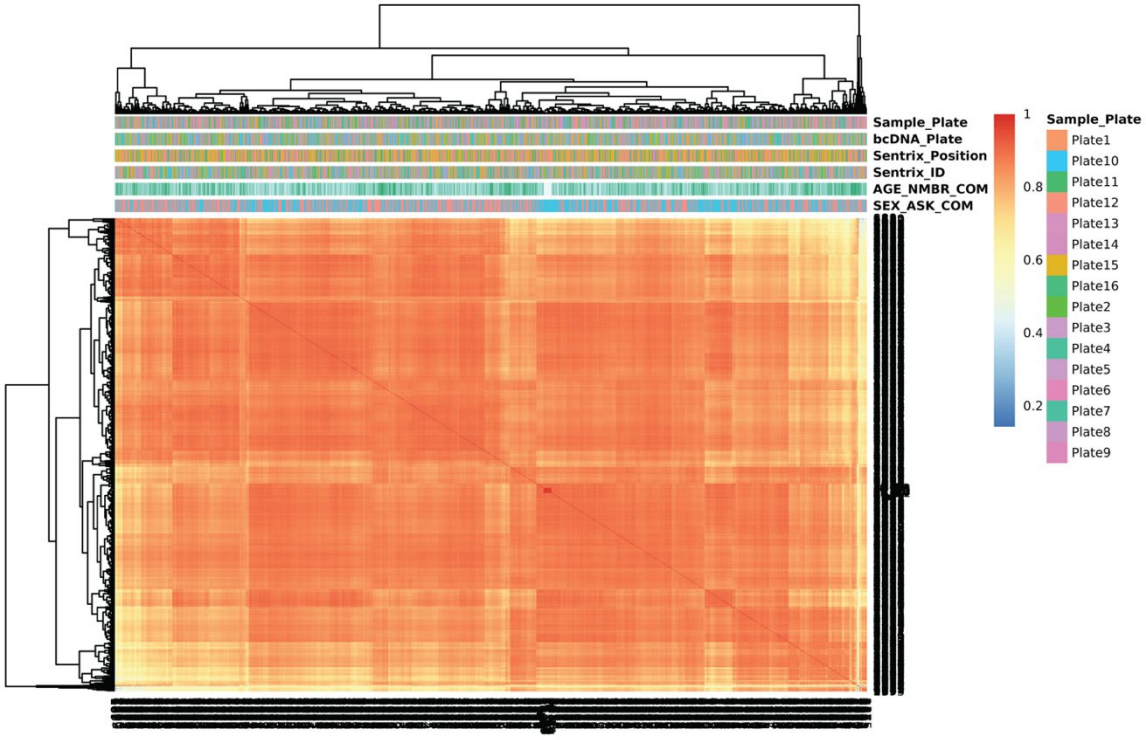

Figure 11. Identification of statistical outliers in the CLSA DNA methylation dataset by overall sample beta value correlations. Supervised clustering on filtered dataset was performed using sample beta value correlations, and each square on the heatmap represents the Pearson's pairwise correlation value between two samples represented on the row and column. Samples displaying consistently lower correlations to the majority were flagged as outliers.

3.4 Additional outlier analyses

3.4.1 *outlyx (wateRmelon)*

Outlyx is a built-in *wateRmelon*[21] function that takes a MethylumiSet object as an input and identifies outliers based on two outlier detection methods: 1) IQR, where the outliers are calculated based on interquartile ranges (default setting = 2); and 2) *PCOut*, where the outliers are determined using distance measures with a modified version of *PCOut* function from *mvoutlier*. We flagged a sample as an *outlyx* outlier when both IQR and *PCOut* are tagged as true and identified 14 outliers using this method.

3.4.2 *detectOutlier (lumi)*

The *detectOutlier* function built-in to the *lumi*[23] package identifies outlier sample based on distance to the cluster center, assuming all samples represent a single cluster and that the distance from the center sample is Gaussian distributed. Using the default threshold, we identified 30 outliers using this method. **Figure 12** shows the beta distributions of the inliers and outliers identified using this method.

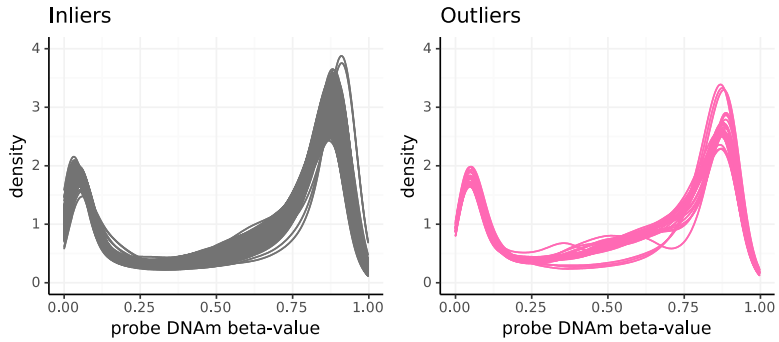

Figure 12. Identification of statistical outliers in the CLSA DNA methylation dataset by the *detectOutlier* function of the lumi package. Density plots of beta-value distributions of the inliers versus outliers as identified by this method are shown side-by-side, where each line represents a unique sample.

3.5 Summary of outlier analyses

To determine the final statistical outliers for preprocessing, we summarized the outlying samples determined by each method used. To balance between stringency of filtering and retaining samples with reasonable qualities, we decided to not filter out samples that were identified as outlying by only a single method.

Table 1 lists 29 samples that were detected as “outliers” in at least two of the different methods used. These samples were removed from the downstream normalization and subsequent sample preprocessing procedures.

	Outlier Detection Method Used			
FINAL_ID	locfdr	beta correlation	outlyx	lumi
124	Outlier			Outlier
129	Outlier		Outlier	
135	Outlier	Outlier		Outlier
195	Outlier	Outlier	Outlier	Outlier
277		Outlier		Outlier
387	Outlier			Outlier
397	Outlier			Outlier
406	Outlier			Outlier
660		Outlier		Outlier
714		Outlier		Outlier
770	Outlier			Outlier
777	Outlier	Outlier		Outlier
786		Outlier		Outlier
798	Outlier			Outlier
830	Outlier	Outlier	Outlier	Outlier
839	Outlier			Outlier
880	Outlier			Outlier
978	Outlier	Outlier	Outlier	Outlier
1062		Outlier		Outlier
1074	Outlier	Outlier	Outlier	Outlier
1089	Outlier	Outlier	Outlier	Outlier
1095	Outlier		Outlier	
1199	Outlier		Outlier	
1203	Outlier		Outlier	Outlier
1208		Outlier		Outlier
1269	Outlier	Outlier		Outlier
1310		Outlier		Outlier
1314	Outlier			Outlier
1323	Outlier	Outlier	Outlier	Outlier

TABLE 1. A summary of the 29 outlying samples that were detected as statistical outliers in at least two of the different methods used in our pipeline.

4.0 DATA NORMALIZATION

4.1 Inter-sample normalization: Quantile

Data normalization was performed in two stages. In the first stage, we normalized between samples using the inter-sample quantile normalization method, as this aims to remove unwanted technical variations without taking away the biological signal. Since all samples on the arrays were derived from the same tissue (PBMCs), we expect the general global methylation distributions to be similar between all samples, which is the underlying assumption of quantile normalization.

Quantile normalization is a nonlinear transformation which takes the average of each quantile across samples as the reference, and replaces each intensity score with this mean to force the observed distributions to be identical to the average[27]. We performed quantile normalization on the array beta-values using the *betaqn* function (*wateRmelon* package)[21,28]. As shown in **Figure 13**, quantile normalization forced identical beta-value distributions, as expected. However, this normalization does not affect Infinium probe-type distributions (**Figure 14**).

4.2 Intra-sample normalization: Beta Mixture-Interquantile (BMIQ)

Like its predecessor the Illumina Infinium HumanMethylation450K beadarrays, the EPIC arrays are composed of two different probe designs: Infinium Type I, which employs two probes (methylated and unmethylated) per CpG locus; and Infinium Type II, which measures methylation at a given CpG locus using a single probe[18]. The smaller subset of Type I probes on the EPIC arrays allows for methylation measurements at CpG dense regions due to its design characteristics[18]. However, due to differences in their inherent designs, the two probe types display distinct distributions, with Type II probes exhibiting a much lower dynamic range, and reported to have more biased results[29]. It is important to normalize the two probe types – by making the two distributions more similar, it will eliminate enrichment bias towards type I probes in supervised analyses, and minimize their technical variations which may compound regional analyses in genomic regions covered by both probe types[29].

To address the differences in the probe type distributions, we performed the intra-array normalization method, Beta Mixture Quantile Dilation (“BMIQ”), which decomposes the density profiles of type I and type II probes by fitting a beta-mixture model of the unmethylated (beta-value < 0.25), hemimethylated (0.25 < beta-value < 0.75), and methylated (beta-value > 0.75) states, then uses a quantile normalization to fit the beta-value distribution of the type II probes to that of type I probes[29]. As shown in **Figure 13**, the beta value distributions between samples remain relatively tight after BMIQ adjustments, with the peaks shifted and heightened as a result of adjusting the type II probe distributions. This shift of type II to type I probe distributions is evident when inspecting the density plots of the probes separately pre- and post-BMIQ, an effect that was not impacted by quantile normalization (**Figure 14**).

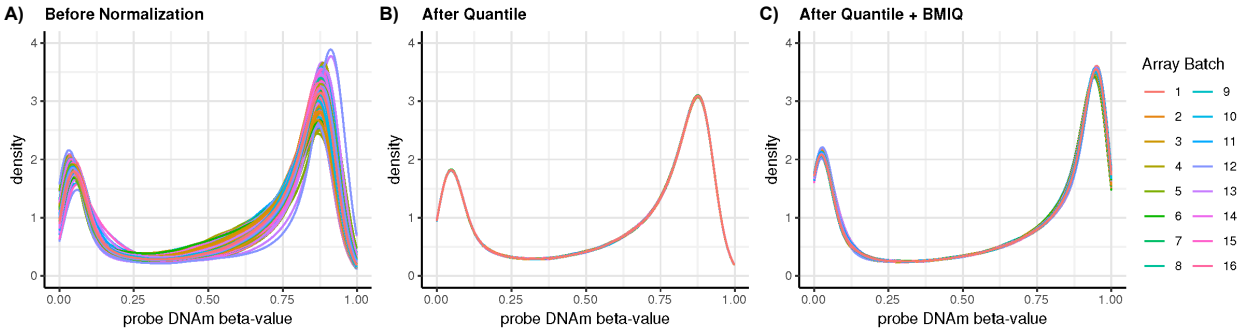

Figure 13. Density plots of beta-value distributions of the 1,445 CLSA DNAm samples that passed QC before and after data normalization. Each panel shows the beta-value distributions of all samples: **A)** Before data normalization; **B)** after inter-sample quantile normalization; and **C)** after inter-sample quantile and intra-sample BMIQ normalizations. Each sample is represented by a line, colored by their respective array batch.

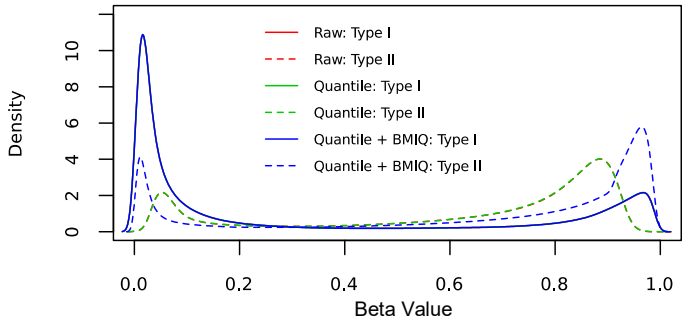

Figure 14. Probe Type beta-value density distributions of all 1,445 CLSA DNAm samples that passed QC before and after data normalization. Each line represents the average beta-distribution of all samples, colored by preprocessing stage (note that for "Type I" – all three solids lines overlay; while for "Type II", the red dashed line representing raw data is overlaid by the green dashed line, representing quantile-normalized data).

5.0 TECHNICAL BATCH CORRECTIONS

5.1 Inspection of batch effects

Normalization, in theory, should get rid of most of the technical batch variations present in the data. To ascertain this, we performed another PCA using the normalized sample beta-values at this stage of data preprocessing and examined the amount of variations associated with each of the top 10 PCs in a scree plot. We then performed multiple association tests to check for significant correlations between each PC and the variables (technical and biological) of interest to determine, if the variability in DNA methylation in any PC, is significantly attributable to batch effects. In this study, we considered three batch variables: bisulfite conversion batch, the array chip, and the array row. **Figure 15** shows the results of this analysis: While the first PC, representing 20.4% of the adjusted DNAm variance, is not associated with the technical variables, the second PC, which represents 8.6% adjusted DNAm variance, is moderately associated with array chip and bisulfite conversion batch. In addition, the technical variables are associated with PCs 4-10, despite these each representing only a small fraction of the DNAm variation. We will use the *ComBat* function to perform batch corrections in the next section.

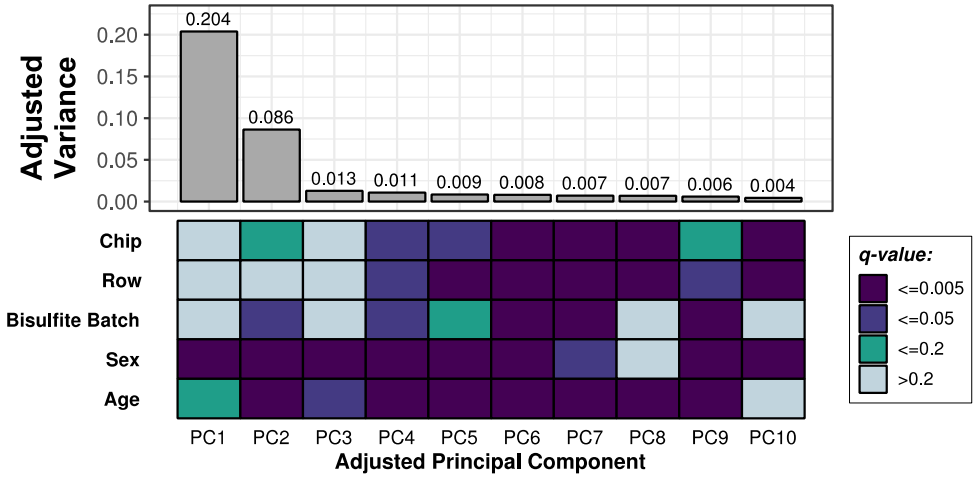

Figure 15. A heat-scree plot showing the relations between study technical ("Chip", "Row", and "Bisulfite Batch") and biological ("Sex" and "Age") variables and each principal component of the methylation data after data normalization. Top Panel is a bar/scree plot representing the percent of adjusted variance associated with each methylation data principal component, and the bottom panel is a heatmap showing the strength of association (multiple test corrected) between the study variables with each PC.

5.2 Technical batch corrections using *ComBat*

ComBat was developed by Johnson, Rabinovic, and Li[30]. Its purpose is to correct for effects of known microarray batches using parametric and non-parametric empirical Bayes frameworks, essentially by adjusting the data for each CpG to match a common cross-batch mean that is estimated using samples from all batches. The batches that are commonly under consideration for microarray studies are: the array chip, each of which holds 8 samples; and the array row, each row representing one sample on a chip. In addition to these two technical variables, the bisulfite conversion batch, which is unconfounded from the array batches, also showed some significant contributions to batch effects in this study (**Figure 15**), and has been suggested as a source of variation to take into consideration[31].

We used the *ComBat* function from the R package *sva* (v3.32.1) to adjust for beta value differences between samples, taking these three technical batches (row, chip, bisulfite batch) into consideration, and performing sequential *ComBat* procedures to correct for these batch effects. After each correction, the batch effects were examined using heat-scree plots (Section 5-1). **Figure 16** shows the final results after adjusting for all three technical batch variables and confirms the lack of association of array arrays and rows with the top 10 PCs following batch adjustments; while bisulfite conversion batch still appears to be strongly associated with PC8, this PC only represents 0.5% of the DNA methylation data variance, which is a sharp decline compared to before batch correction.

We note that sex appears to be consistently strongly associated with the top PCs. We therefore opted to perform a quick visual check on these associations by constructing violin plots of male versus female distributions for each of the top 3 PCs to confirm that while sex is associated with the top 3 PCs (accounting for a total of ~30% methylation variance), the distributions of the sexes are not completely segregated in each of these PCs (**Figure 17**). These data demonstrate that sex should be accounted as an important covariate in study models, because there is evidence for sex differences in DNA methylation of autosomal probes.

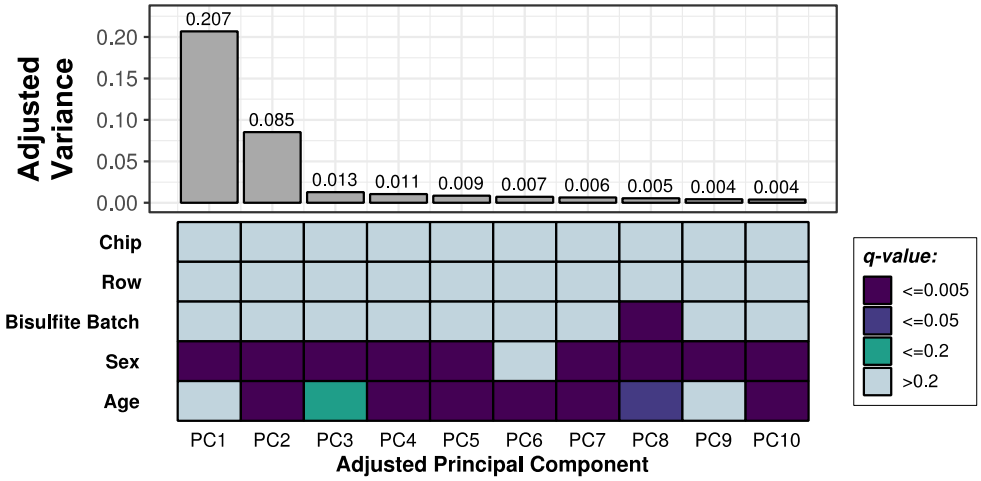

Figure 16. A heat-scree plot showing the relations between study technical ("Chip", "Row", and "Bisulfite Batch") and biological ("Sex" and "Age") variables and each principal component of the methylation data after batch correction with *ComBat*. Top Panel is a bar/scree plot representing the percent of adjusted variance associated with each methylation data principal component, and the bottom panel is a heatmap showing the strength of association (multiple test-corrected) between the study variables with each PC.

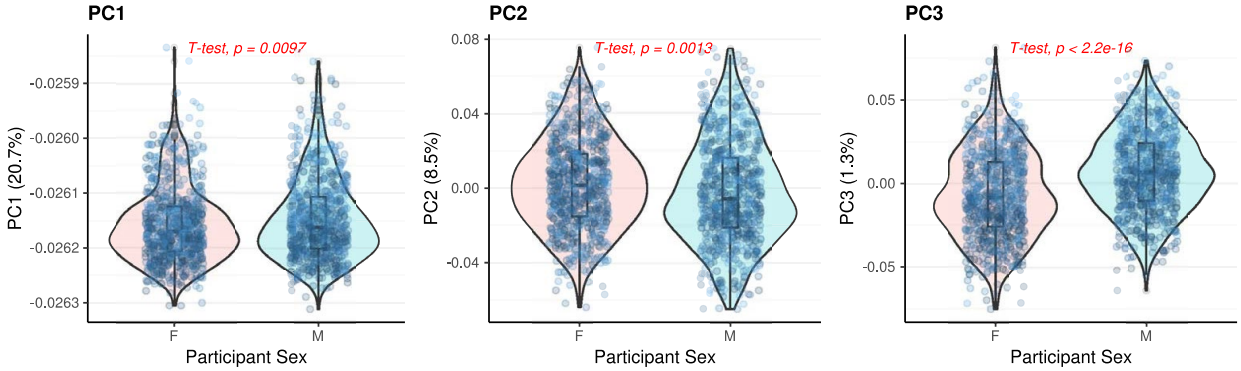

Figure 17. Violin plots showing the distribution of participant sex with respect to the top 3 principal components derived from a PCA analysis on the post-*ComBat* methylation data of the CLSA cohort. Two-tailed *t*-test statistics were performed between sexes in each PC.

6.0 BLOOD CELL TYPE CORRECTIONS

6.1 Bioinformatic estimations of blood cell types

As each cell type is known to be associated with a distinct DNA methylation signature, DNA methylation analyses in tissues are always compounded by the potential differences in cell type proportions between samples[32–34]. Since PBMCs are made up of different subtypes of immune cells, such as the B lymphocytes, T lymphocytes, and monocytes, it is important to address the cell proportions in each sample to minimize cell type bias. This can usually be achieved by either FACS sorting of the original specimen, or bioinformatically using a method known as cell-type deconvolution, as described below.

To estimate the immune cell type proportions within each PBMC sample, we performed bioinformatic cell type prediction based on an established reference-based algorithm[32]. In short,

this method involves constrained projection of the DNA methylation profile in question onto a "reference" DNA methylation set comprised of signatures derived for each sorted immune cell type[35]. This allows for inference of the estimated proportions of five major white blood cell subtypes found in PBMC specimens: B lymphocytes, CD4+ T lymphocytes, CD8+ T lymphocytes, Natural Killer (NK) lymphocytes, monocytes. The method also allows an estimation of granulocyte proportions which, although should be absent from the PBMC fractions, may be still present in a miniscule fraction owing to the cell separation procedure.

Figure 18 summarizes the cell count estimates for the remaining samples, where each colored dot represents the estimated proportion of the respective cell type for an individual, and the bar graphs illustrate the range of the cell proportions across all individuals. We observe zero granulocyte proportions for almost all individuals, confirming the purity of the PBMC isolations. Meanwhile, the rest of the estimated cell type proportions roughly conform to those expected in normal human samples, but with a large spread across the cohort[36].

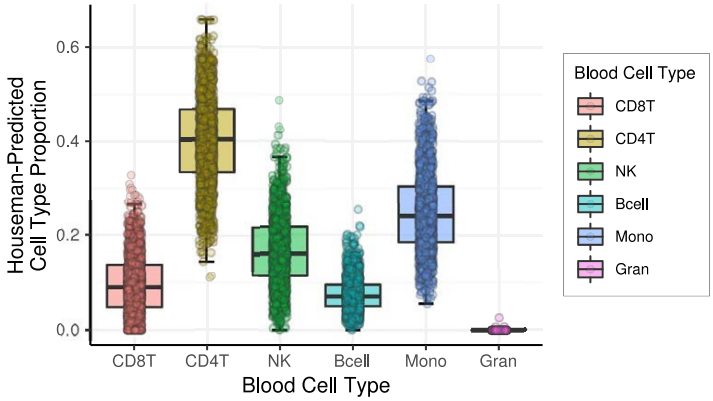

Figure 18. A boxplot showing the range of estimated cell type proportions (using the reference-based Houseman method) of the CLSA PBMC samples. Each overlaying dot represents an individual sample. Whiskers are plotted where? Different software plots whiskers are different locations.

To demonstrate that cell type signatures are indeed highly associated with sample methylation profiles, we included the estimated cell proportions into our post-*ComBat* heat-scree plot. **Figure 19** illustrates that the top 10 PCs from the batch-corrected methylation data are significantly and consistently associated with the various predicted blood cell types. Therefore, one should take caution and address the cell type bias in the downstream analyses, either by including the estimated proportions into the final study model as covariates, or by adjusting the beta values with respect to these cell type differences, as described in the next section.

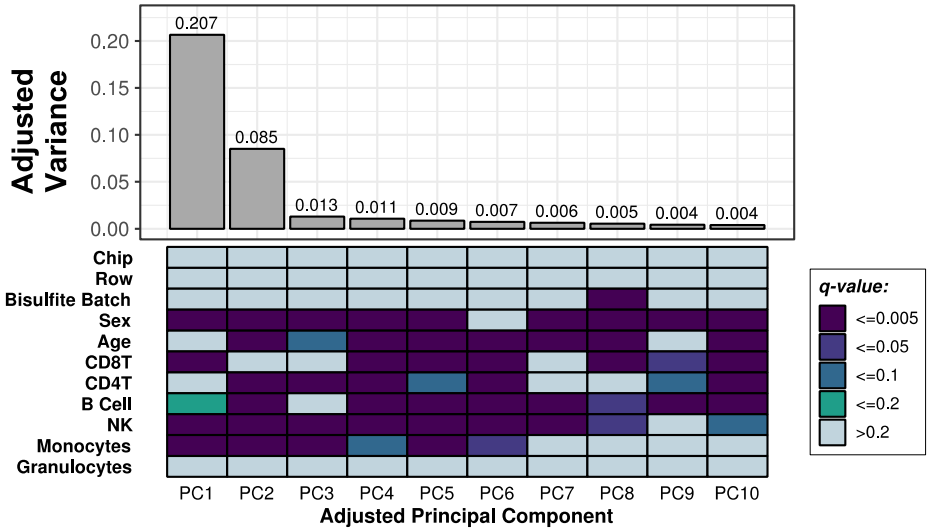

Figure 19. A heat-scree plot showing the relations between study variables, estimated cell types, and each principal component of the methylation data after batch correction with *ComBat*. Top Panel is a bar/scree plot representing the percent of adjusted variance associated with each methylation data principal component, and the bottom panel is a heatmap showing the strength of association (multiple test corrected) between the study variables with each PC.

6.2 Beta value adjustment to differential estimated cell type counts by linear regression

In this data release, we used a regression-based approach as described in Jones *et al.*[37] to provide users with beta-values that are adjusted for cell types. Briefly, we fitted a linear model on the DNA methylation outcome using the six estimated proportions of cell subtypes as additive variables for each probe, then we extracted residuals from the resulting linear models. These residuals represent DNA methylation variability that are unexplained by cell type composition, and therefore, may be attributed to other phenotypic variable of interest. Finally, we added the residuals of each regression model to the mean beta value of each probe across all samples to obtain the final "adjusted" methylation data. Note importantly, because the operational definition of a beta value being between 0-1, here we forced the adjusted values >1 to be 0.999 and those <0 to be 0.001, representing fully methylated and unmethylated state of the given probe, respectively.

Following this adjustment procedure, we see that the cell type proportions have been "regressed out" when we performed another round of reference-based cell type predictions (**Figure 20**), as the participants now exhibit comparable cell type proportions if these proportions were re-estimated using the cell type adjusted set of beta values. Furthermore, PCA shows that the cell types are no longer associated with the top PCs; importantly, the adjusted variances of the top 2 PCs have been reduced to 6.8% and 1.2% percent, respectively, demonstrating the importance of correcting for cell type variations associated with the data prior to further analyses (**Figure 21**). Nevertheless, the density plots of global beta values for all the samples remained as bimodally distributed (**Figure 22**). The provided cell-type corrected beta-values should now be appropriate for further data analyses without the need to include cell type proportions as covariates into the study models.

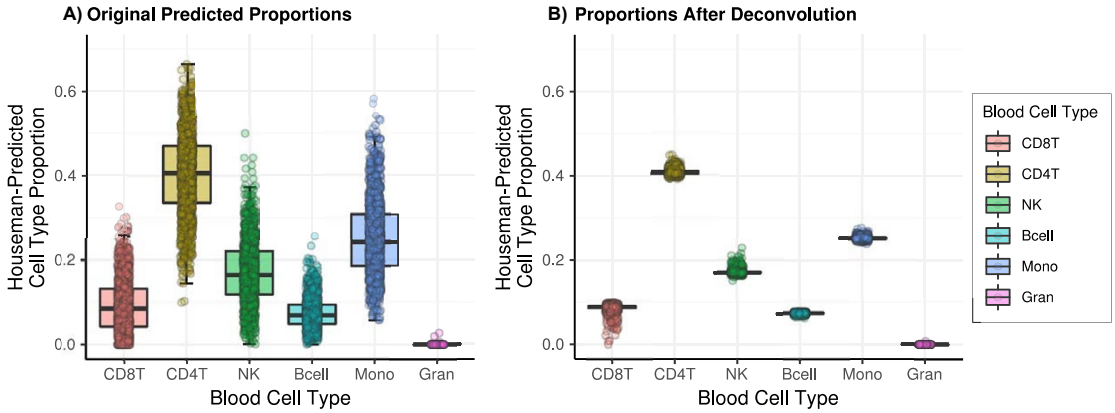

Figure 20. A boxplot showing the range of estimated cell type of the CLSA PBMC samples **A)** before; and **B)** after cell-type adjustment by linear regression. Each overlaying dot represents an individual sample.

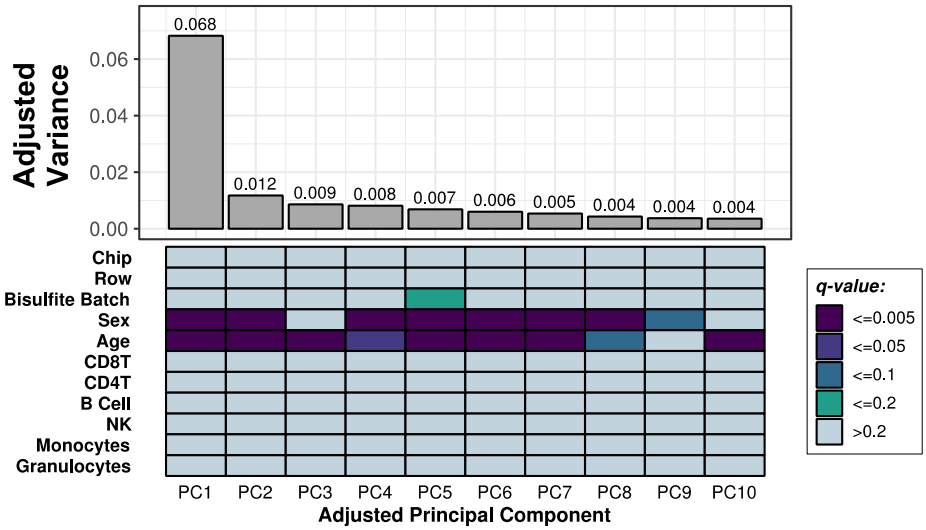

Figure 21. A heat-scree plot showing the relations between study variables, estimated cell types, and each principal component of the methylation data after cell-type adjustment. Top Panel is a bar/scree plot representing the percent of adjusted variance associated with each methylation data principal component, and the bottom panel is a heatmap showing the strength of association (multiple test-corrected) between the study variables with each PC.

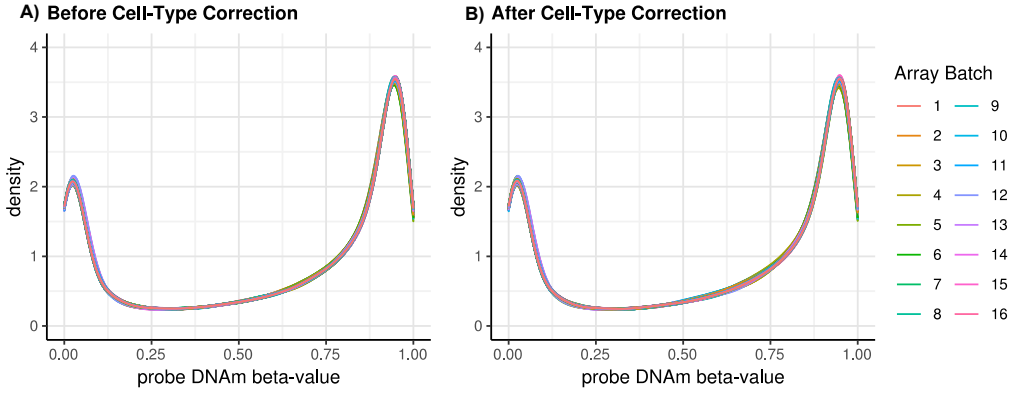

Figure 22. Density plots of beta-value distributions of the CLSA DNAm samples **A)** before; and **B)** after cell-type adjustment by linear regression. Each sample is represented by a line, colored by their respective array batch.

7.0 EXEMPLAR EPIGENOME-WIDE ASSOCIATION STUDY (EWAS) WITH CHRONOLOGICAL AGE

7.1 Introduction

In this section, we performed an example EWAS with chronological age using the preprocessed CLSA DNAm dataset to demonstrate the quality and utility of the data release. As several studies have demonstrated replicable associations between DNA methylation patterns and chronological aging in human blood samples, we will also compare our results with these past reports to illustrate data quality.

7.2 Chronological age EWAS

To ascertain the quality of the preprocessed CLSA data and examine whether there is an association between DNA CpG loci and chronological aging in our dataset, we used the preprocessed CLSA data and fitted each remaining CpG probe to the following linear model assigning age as the main effect, using participant reported sex as a covariate. Beta-values were converted to log-transformed methylation *M-values* in our study model as it is more appropriate for the homoscedasticity assumption in a multiple regression model[14]:

lm(*M value* ~ Chronological Age + Participant Sex)

Statistical significance of each CpG was assessed using multiple-test corrected (Benjamini-Hochberg) association *p-values* and represented as false discovery rates (FDRs)[38]. We also used the biological effect size (delta-beta)[39], calculated as the difference in methylation beta-values between the highest and lowest chronological age within the cohort, to further define significantly associated loci. Using the FDR statistics and the delta-beta value, we constructed a volcano plot as shown in **Figure 23** as a graphic representation of our model results. The list of CpGs that met the statistical thresholds of FDR < 1e-100 and biological effect size threshold of >5% delta-beta is presented in **Table 2**.

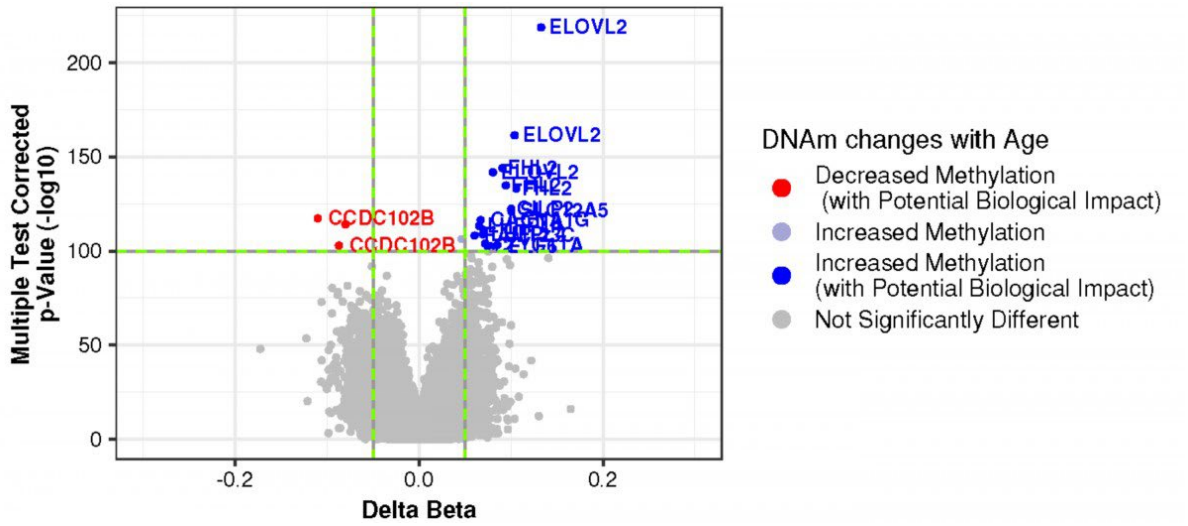

Figure 23. A volcano plot showing CpG loci on the EPIC arrays whose methylation status are significantly associated with chronological aging in the preprocessed CLSA DNAm dataset, as filtered by effect size (Delta Beta > 5%) and statistical significance (FDR < 1e-100). Loci passing the significant thresholds are colored and annotated by their associated genes.

TABLE 2. List of CpG loci on the EPIC arrays that show DNAm changes significantly associated with chronological age in the CLSA DNAm dataset. In addition to Beta, SE would be useful as well. For coordinate, please specify which build of the human genome (GRCh37/38)?

CpG	Chr	Coordinate	Nominal P-value	FDR	Trend with Age	Delta Beta	Associated Gene	Genomic Context
CpGs exhibiting hypermethylation with increasing age								
cg06784991	1	53308768	1.59E-108	6.90E-104	Hypermethylation	0.085	ZYG11A	Body
cg06639320	2	106015739	4.10E-139	5.36E-134	Hypermethylation	0.106	FHL2	TSS200
cg17268658	2	106015745	3.59E-150	9.36E-145	Hypermethylation	0.091	FHL2	TSS200
cg22454769	2	106015767	9.29E-141	1.45E-135	Hypermethylation	0.094	FHL2	TSS200
cg23606718	2	131513927	1.22E-113	6.35E-109	Hypermethylation	0.060	FAM123C	5'UTR
cg24866418	3	9594082	2.25E-114	1.26E-109	Hypermethylation	0.069	LHFPL4	Body
cg12841266	3	9594093	8.40E-119	5.49E-114	Hypermethylation	0.065	LHFPL4	Body
cg06570224	3	157812475	1.71E-116	1.03E-111	Hypermethylation	0.070	(Intergenic)	
cg16867657	6	11044877	1.89E-225	1.48E-219	Hypermethylation	0.133	ELOVL2	TSS1500
cg24724428	6	11044888	8.09E-168	3.17E-162	Hypermethylation	0.104	ELOVL2	TSS1500
cg21572722	6	11044894	7.23E-148	1.41E-142	Hypermethylation	0.080	ELOVL2	TSS1500
cg08637691	9	134989631	1.12E-107	4.16E-103	Hypermethylation	0.078	(Intergenic)	
cg27099280	15	72612204	7.25E-108	2.84E-103	Hypermethylation	0.074	CELF6	1stExon
cg11071401	17	48637194	5.29E-122	4.14E-117	Hypermethylation	0.067	CACNA1G	TSS1500
cg07544187	19	19651235	4.06E-128	4.54E-123	Hypermethylation	0.100	CILP2	Body
cg17110586	19	36454623	2.04E-109	9.38E-105	Hypermethylation	0.072	(Intergenic)	
cg07547549	20	44658225	4.45E-127	4.36E-122	Hypermethylation	0.100	SLC12A5	Body
CpGs exhibiting hypomethylation with increasing age								
cg26947034	7	33935438	9.21E-120	6.56E-115	Hypomethylation	-0.080	(Intergenic)	
cg19283806	18	66389420	3.23E-108	1.33E-103	Hypomethylation	-0.087	CCDC102B	5'UTR
cg13552692	18	66389447	5.40E-123	4.70E-118	Hypomethylation	-0.110	CCDC102B	5'UTR

All 17 hypermethylated and 3 hypomethylated CpGs passing our thresholds replicated top hits reported from previous studies. These top age-associated CpGs included loci associated with the promoters of two genes, FHL2 and ELOVL2, whose increased methylation levels have been well-documented as associated with chronological aging[40–42]. We also replicated hypermethylation trends for specific loci associated with ZYG11A, FAM123C, LHFPL4, CELF6, CACNA1G, CILP2, and SLC12A5[41,43–45]. In addition, we also observed the previously-reported age-related hypomethylation in the CCDC102B 5' UTR region[43].

Figure 24 highlights some of the CpG-specific methylation trends associated with chronological age found in the CLSA DNAm dataset.

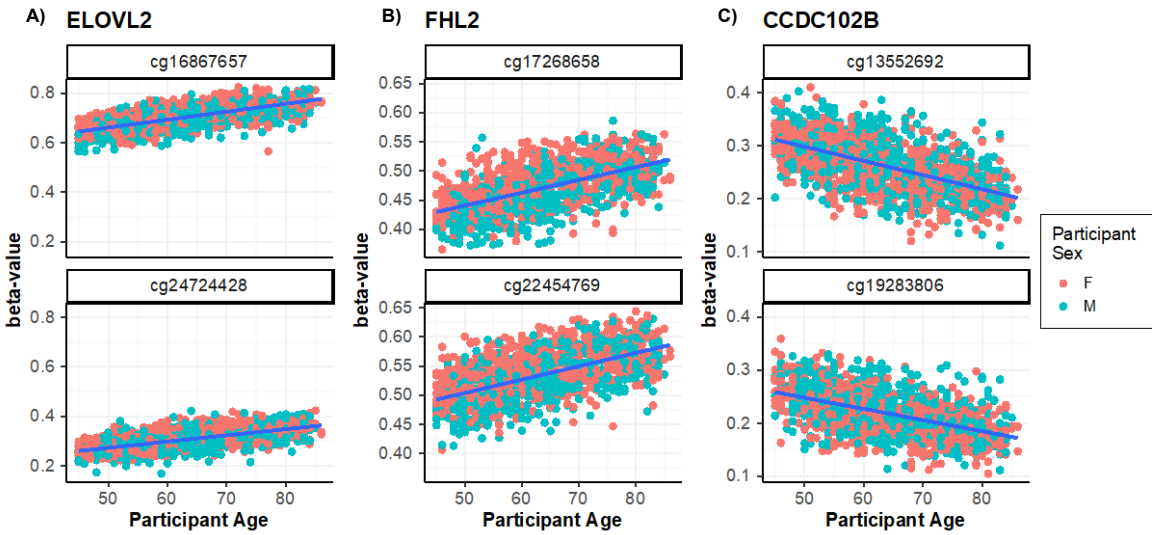

Figure 24. Dot plots showing changes in DNA methylation levels (beta-value) in the CLSA DNAm dataset with respect to participant ages at selected loci associated with **A)** ELOVL2 promoter; **B)** FHL2 promoter; and **C)** CCDC102B 5'UTR regions. Blue line in each panel represents predicted fit of each linear model (DNAm ~ Age).

8.0 EPIGENETIC AGE CALCULATIONS USING DNA METHYLATION DATA

8.1 Introduction

This CLSA epigenetics data release also provides epigenetic age estimations using the Horvath, the Hannum, PhenoAge, and GrimAge epigenetic clocks, which are the first and second generation established epigenetic age algorithms, for 1,445 participants assayed. These epigenetic clocks represent a biomarker of aging and health[5,6]. The pan-tissue Horvath epigenetic clock algorithm, originally developed based on the DNA methylation status of 353-CpG sites on the Illumina 450K arrays, is capable of predicting the "epigenetic age" that in theory correlates highly with an individual's chronological age[5]; Epigenetic age acceleration as determined by Horvath clock has been associated to conditions including: Down Syndrome, Huntington's disease, HIV infections, obesity, and lifetime stress[46]**.** The Hannum clock was developed with exclusively blood samples and uses a different set of 71 CpGs to estimate ages[6]. Although the EPIC chip lacks some of the original sites from the 450K arrays used to develop these clocks, it has been shown that the platform still provides an accurate age estimation[47]. The DNAm PhenoAge estimate is based on a phenotypic age score derived from chronological age and nine clinically relevant blood biomarkers, including albumin, creatinine, serum glucose, C-reactive protein, lymphocyte percentage, mean cell volume, red cell distribution width, alkaline phosphatase, and white blood cell count.[7] The DNAm PhenoAge was trained to predict all-cause mortality. The DNAm GrimAge estimate was developed using DNA methylation-based surrogates for seven age-related plasma proteins, including adrenomedullin, beta-2-microglobulin, cystatin-C, growth differentiation factor 15, leptin, plasminogen activator inhibitor 1, and tissue inhibitor metalloproteinases 1 (TIMP-1) along with a DNA methylation-based estimator of smoking pack-years.[49] Age and sex were included as covariates. The DNAm GrimAge was trained to predict time-to-death.

8.2 Calculation of epigenetic ages

As of September 2021, DNA methylation age estimates were recalculated using the updated online DNAm Age calculator. To calculate the epigenetic ages, we followed instructions as provided by the Horvath DNAmAge website[50]. Briefly, we used *GenomeStudio* colour-corrected and background-subtracted beta-value matrix as previously recommended[47], and subsetted to 30,084 CpGs provided in the Horvath annotation file to use as our data input. In addition to the 353 and 71 CpG sites for the Horvath and Hannum clocks, respectively, the additional CpGs are imperative for data normalization purposes for the prediction algorithm. Once the data was properly reshaped, they were submitted to the DNAmAge website together with a sample annotation file specifying the chronological age, biological sex, and tissue type (PBMC) of each individual, with the "Normalize Data" and "Advanced Analysis" options selected. The Horvath DNAm Age calculator estimates Hannum age and sample cell type proportions to allow calculations of the extrinsic and intrinsic epigenetic age accelerations[11] (EEAA and IEAA, respectively), and includes some additional calculations not discussed in this data release.

For Hannum age, data preprocessing and quality control were performed using the ENmix R package, and the resulting beta values were used to calculate the age estimates provided in the data release.

The DNAm PhenoAge estimate is based on a phenotypic age score derived from chronological age and nine clinically relevant blood biomarkers: albumin, creatinine, serum glucose, C-reactive protein, lymphocyte percentage, mean cell volume, red cell distribution width, alkaline phosphatase, and white blood cell count.[7] Each clock was derived using weighted combinations of CpG sites, with methylation beta values normalized using the Noob normalization approach.

The DNAm PhenoAge estimates were obtained from the online DNAmAge calculator hosted by the Clock Foundation. The DNAm GrimAge estimate was developed using DNA methylation-based surrogates for seven age-related plasma proteins: adrenomedullin, beta-2-microglobulin, cystatin-C, growth differentiation factor 15, leptin, plasminogen activator inhibitor 1, and tissue inhibitor metalloproteinases 1 (TIMP-1) along with a DNA methylation-based estimator of smoking pack-years.[49] Age and sex were included as covariates. CpG weights were applied to Noob-normalized beta values. DNAm GrimAge estimates were obtained via the DNAmAge online calculator provided by the Clock Foundation.

Note: In the future, if the DNA methylation age calculator is updated, it may result in different DNA methylation age values for the clocks.

8.3 Examination of epigenetic age estimates

We examined the epigenetic age calculation results by performing linear regressions of the epigenetic ages against the participant chronological ages and calculated the Pearson correlation coefficients. As expected, all clocks displayed highly significant positive correlations between the measures (**Figure 25**: Horvath: $r = 0.863$, $p < .0001$; Hannum: $r = 0.865$, $p < .0001$, PhenoAge: $r = 0.818$, $p < .0001$, GrimAge: $r = 0.896$, $p < .0001$).

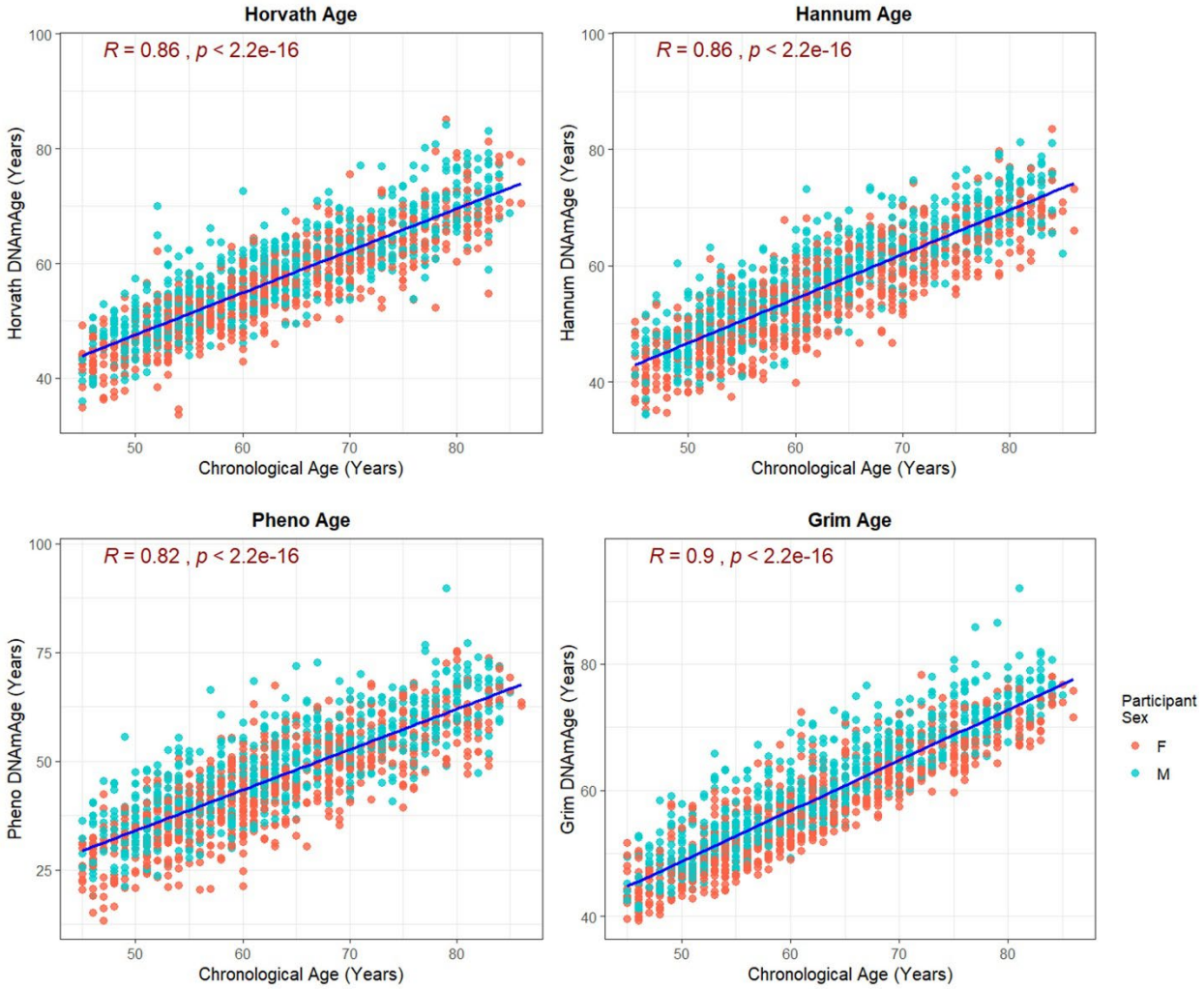

Figure 25. Dot plots showing relationships between **A)** Horvath DNAm Age; **B)** Hannum DNAm Age; **C)** Pheno DNAmAge; and **D)** Grim DNAmAge, and participant chronological age in years. Blue line in each panel represents predicted fit of each linear model (DNAm Age ~ Chronological Age).

REFERENCES

1. Bird, A. Perceptions of epigenetics. *Nature* **447**, 396–398 (2007).

2. Zhu, H., Wang, G. & Qian, J. Transcription factors as readers and effectors of DNA methylation. *Nat. Rev. Genet.* **17**, 551–565 (2016).

3. Islam, S. A., Lussier, A. A. & Kobor, M. S. Epigenetic analysis of human postmortem brain tissue. *Handb. Clin. Neurol.* **150**, 237–261 (2018).

4. Luo, C., Hajkova, P. & Ecker, J. R. Dynamic DNA methylation: In the right place at the right time. *Science* **361**, 1336–1340 (2018).

5. Horvath, S. DNA methylation age of human tissues and cell types. *Genome Biol.* **14**, R115 (2013).

6. Hannum, G. *et al.* Genome-wide methylation profiles reveal quantitative views of human aging rates. *Mol. Cell* **49**, 359–367 (2013).

7. Levine, M. E. *et al.* An epigenetic biomarker of aging for lifespan and healthspan. *Aging* **10**, 573–591 (2018).

8. Researchers | Canadian Longitudinal Study on Aging. https://www.clsa-elcv.ca/researchers.

9. Infinium MethylationEPIC Data Sheet | Illumina. https://science-docs.illumina.com/documents/Microarray/infinium-methylation-epic-data-sheet-1070-2015-008/Content/Source/Microarray/Infinium/MethylationEPIC/infinium-methylation-epic-data-sheet.html.

10. Infinium MethylationEPIC BeadChip Product Files. https://support.illumina.com/array/array_kits/infinium-methylationepic-beadchip-kit/downloads.html.

11. Chen, B. H. *et al.* DNA methylation-based measures of biological age: meta-analysis predicting time to death. *Aging* **8**, 1844–1865 (2016).

12. Clark, S. J., Statham, A., Stirzaker, C., Molloy, P. L. & Frommer, M. DNA methylation: bisulphite modification and analysis. *Nat. Protoc.* **1**, 2353–2364 (2006).

13. GenomeStudio Methylation Module v1.8 User Guide (11319130). 114.

14. Du, P. *et al.* Comparison of Beta-value and M-value methods for quantifying methylation levels by microarray analysis. *BMC Bioinformatics* **11**, 587 (2010).

15. Package ‘minfi’. http://bioconductor.org/packages/release/bioc/manuals/minfi/man/minfi.pdf.

16. Package ‘methylumi’. https://www.bioconductor.org/packages/release/bioc/manuals/methylumi/man/methylumi.pdf.

17. Long, H. K., King, H. W., Patient, R. K., Odom, D. T. & Klose, R. J. Protection of CpG islands from DNA methylation is DNA-encoded and evolutionarily conserved. *Nucleic Acids Res.* **44**, 6693–6706 (2016).

18. Pidsley, R. *et al.* Critical evaluation of the Illumina MethylationEPIC BeadChip microarray for whole-genome DNA methylation profiling. *Genome Biol.* **17**, 208 (2016).

19. Kuan, P. F., Wang, S., Zhou, X. & Chu, H. A statistical framework for Illumina DNA methylation arrays. *Bioinforma. Oxf. Engl.* **26**, 2849–2855 (2010).

20. Infinium HD Methylation Assay Reference Guide (15019519). https://support.illumina.com/content/dam/illumina-support/documents/documentation/chemistry_documentation/infinium_assays/infinium_hd_methylation/infinium-methylation-assay-reference-guide-15019519-07.pdf.

21. Package 'wateRmelon'. https://www.bioconductor.org/packages/release/bioc/manuals/wateRmelon/man/wateRmelon.pdf.

22. Wong, C. C., Pidsley, R. & Schalkwyk, L. C. The wateRmelon Package. https://bioconductor.org/packages/release/bioc/vignettes/wateRmelon/inst/doc/wateRmelon.pdf.

23. Package 'lumi'. https://www.bioconductor.org/packages/release/bioc/manuals/lumi/man/lumi.pdf.

24. Joo, J. E. *et al.* Human active X-specific DNA methylation events showing stability across time and tissues. *Eur. J. Hum. Genet. EJHG* **22**, 1376–1381 (2014).

25. Forgetta, V. *et al.* The Canadian Longitudinal Study on Aging. https://www.clsa-elcv.ca/doc/2748.

26. Package 'impute'. https://www.bioconductor.org/packages/release/bioc/manuals/impute/man/impute.pdf.

27. Bolstad, B. M., Irizarry, R. A., Astrand, M. & Speed, T. P. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinforma. Oxf. Engl.* **19**, 185–193 (2003).

28. Pidsley, R. *et al.* A data-driven approach to preprocessing Illumina 450K methylation array data. *BMC Genomics* **14**, 293 (2013).

29. Teschendorff, A. E. *et al.* A beta-mixture quantile normalization method for correcting probe design bias in Illumina Infinium 450 k DNA methylation data. *Bioinforma. Oxf. Engl.* **29**, 189–196 (2013).

30. Johnson, W. E., Li, C. & Rabinovic, A. Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostat. Oxf. Engl.* **8**, 118–127 (2007).

31. Price, E. M. & Robinson, W. P. Adjusting for Batch Effects in DNA Methylation Microarray Data, a Lesson Learned. *Front. Genet.* **9**, 83 (2018).

32. Houseman, E. A. *et al.* DNA methylation arrays as surrogate measures of cell mixture distribution. *BMC Bioinformatics* **13**, 86 (2012).

33. Reinius, L. E. *et al.* Differential DNA methylation in purified human blood cells: implications for cell lineage and studies on disease susceptibility. *PloS One* **7**, e41361 (2012).

34. Jaffe, A. E. & Irizarry, R. A. Accounting for cellular heterogeneity is critical in epigenome-wide association studies. *Genome Biol.* **15**, R31 (2014).

35. Koestler, D. C. *et al.* Blood-based profiles of DNA methylation predict the underlying distribution of cell types: a validation analysis. *Epigenetics* **8**, 816–826 (2013).

36. Kleiveland, C. R. Peripheral Blood Mononuclear Cells. in *The Impact of Food Bioactives on Health: in vitro and ex vivo models* (eds. Verhoeckx, K. et al.) 161–167 (Springer International Publishing, 2015). doi:10.1007/978-3-319-16104-4_15.

37. Jones, M. J., Islam, S. A., Edgar, R. D. & Kobor, M. S. Adjusting for Cell Type Composition in DNA Methylation Data Using a Regression-Based Approach. in *Population Epigenetics: Methods and Protocols* (eds. Haggarty, P. & Harrison, K.) 99–106 (Springer, 2017). doi:10.1007/7651_2015_262.

38. Benjamini, Y. & Hochberg, Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *J. R. Stat. Soc. Ser. B Methodol.* **57**, 289–300 (1995).

39. Bush, N. R. *et al.* The biological embedding of early-life socioeconomic status and family adversity in children's genome-wide DNA methylation. *Epigenomics* **10**, 1445–1461 (2018).

40. Garagnani, P. *et al.* Methylation of ELOVL2 gene as a new epigenetic marker of age. *Aging Cell* **11**, 1132–1134 (2012).

41. Florath, I., Butterbach, K., Müller, H., Bewerunge-Hudler, M. & Brenner, H. Cross-sectional and longitudinal changes in DNA methylation with age: an epigenome-wide analysis revealing over 60 novel age-associated CpG sites. *Hum. Mol. Genet.* **23**, 1186–1201 (2014).

42. Slieker, R. C., Relton, C. L., Gaunt, T. R., Slagboom, P. E. & Heijmans, B. T. Age-related DNA methylation changes are tissue-specific with ELOVL2 promoter methylation as exception. *Epigenetics Chromatin* **11**, 25 (2018).

43. Park, J.-L. *et al.* Identification and evaluation of age-correlated DNA methylation markers for forensic use. *Forensic Sci. Int. Genet.* **23**, 64–70 (2016).

44. Bacos, K. *et al.* Blood-based biomarkers of age-associated epigenetic changes in human islets associate with insulin secretion and diabetes. *Nat. Commun.* **7**, 11089 (2016).

45. Tajuddin, S. M. *et al.* Novel age-associated DNA methylation changes and epigenetic age acceleration in middle-aged African Americans and whites. *Clin. Epigenetics* **11**, 119 (2019).

46. Quach, A. *et al.* Epigenetic clock analysis of diet, exercise, education, and lifestyle factors.

Aging **9**, 419–446 (2017).

47. McEwen, L. M. *et al.* Systematic evaluation of DNA methylation age estimation with common preprocessing methods and the Infinium MethylationEPIC BeadChip array. *Clin. Epigenetics* **10**, 123 (2018).

48. Horvath, S. *et al.* Epigenetic clock for skin and blood cells applied to Hutchinson Gilford Progeria Syndrome and ex vivo studies. *Aging* **10**, 1758–1775 (2018).

49. Lu, A. T. *et al.* DNA methylation GrimAge strongly predicts lifespan and healthspan. *Aging* **11**, 303–327 (2019).

50. DNA Methylation Age Calculator. https://dnamage.genetics.ucla.edu/.