DNA Methylation Microarray Data Reduction for Co-Methylation Analysis

by

Evan Gatev

Ph.D., Yale University, 2001

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF
THE REQUIREMENTS FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

in

THE FACULTY OF GRADUATE AND POSTDOCTORAL STUDIES
(Bioinformatics)

THE UNIVERSITY OF BRITISH COLUMBIA
(Vancouver)

May 2020

The following individuals certify that they have read, and recommend to the Faculty of Graduate and Postdoctoral Studies for acceptance, the dissertation entitled:

DNA Methylation Microarray Data Reduction for Co-Methylation Analysis

submitted by Evan Gatev in partial fulfillment of the requirements for the degree of Doctor of Philosophy in Bioinformatics

Examining Committee:

Michael S. Kobor, Medical Genetics
Co-supervisor

Sara Mostafavi, Statistics
Co-supervisor

Matthew Lorincz, Medical Genetics
University Examiner

Angela Devlin, Pediatrics
University Examiner

Additional Supervisory Committee Members:

Paul Pavlidis, Bioinformatics
Supervisory Committee Member

Martin Hirst, Microbiology and Immunology
Supervisory Committee Member

Alexandre Bouchard-Cote, Statistics
Supervisory Committee Member

Abstract

DNA Methylation (DNAm) is an epigenetic modification that is present across the human genome, primarily in the context of CpG di-nucleotides. In human population studies, high throughput bead chip microarray assays are the prevalent way to simultaneously measure the methylation state of many thousands of genomic CpG sites. Proximal genomic CpGs have correlated methylation state within a single cell and often function as a single biological unit. The prevailing common methylation state of such multiple CpGs within a common biological unit has been the subject of intense study, due to its immediate relevance for gene expression regulation and ultimately for health and disease. I designed and implemented a method for a biologically motivated DNAm array data reduction, which constructs co-methylated regions (CMRs), while incorporating information about the genomic CpG background from the reference human genome annotation. The method relies on the correlations of CpG methylation across individuals for proximal CpG probes. The method aims for enhanced statistical performance in terms of statistical power and specificity, including for downstream applications. For example, Epigenome Wide Association Studies (EWAS), an important such application, often places the focus on group "hits" with multiple adjacent CpGs that are significant, because their gnomic proximity makes it more likely that the detected correlations are not spurious. The CMRs capture such groups and I showed that the CMRs constructed in whole blood public data have high statistical specificity in the context of EWAS for chronological age and biological sex. When the composite CMR methylation measures were used to perform EWAS for age and sex, they had high sensitivity and specificity, including uncovering additional associated CpGs

not detected by conventional EWAS. The utility of the data reduction method was further discussed within the broader context of applying machine learning algorithms for high dimensional DNAm array data analysis.

Lay Summary

DNA Methylation is a chemical modification occurring at millions of sites across the human genome. It is important for the establishment of the epigenome, which specifies how different types of cells emerge from the same genome. Microarray assays measure the DNA methylation at hundreds of thousands of genomic sites simultaneously, producing high-dimensional data. Since multiple genomic regions containing proximal methylated sites function biologically as units, their statistical analysis requires intricate analytics. This work develops a method for DNA methylation array data reduction to enhance downstream analyses in terms of statistical power, specificity and biological interpretation. The method is motivated by biological findings and uniquely incorporates information from the whole human genome sequence to identify co-methylated regions. This is based on the inter-individual correlation of the methylation status of qualified proximal genomic sites. Several applications illustrate the utility of the method and an open source software implementation is available.

Preface

My contribution to the identification and design of the research program detailed in this thesis is as follows. I identified the need for a biologically motivated data reduction of DNA Methylation array data and designed the method and software to address this need. I performed all parts of the CoMeBack development, implementation and applications in Chapter 2, which has been published the journal "Bioinformatics" (Evan Gatev, et. al, CoMeBack: DNA methylation array data analysis for co-methylated regions, *Bioinformatics*, btaa049, https://doi.org/10.1093/bioinformatics/btaa049). I analyzed all the research data in Chapters 2 and 3. Chapter 4 of this thesis discusses my contributions to research projects that have appeared in the following publications:

1. Roberts, A. et al., Exposure to childhood abuse is associated with human sperm DNA methylation, Translational Psychiatry vol 8, Article number: 194 (2018)

   My contribution was the development of the DNA methylation based predictor for childhood abuse, including the feature selection and the application of the machine learning algorithm that produced the predictor for childhood abuse.

2. Morin, A. et al. Maternal blood contamination of collected cord blood can be identified using DNA methylation at three CpGs, Clinical Epigenetics, volume 9, Article number: 75 (2017)

My contribution was in the development of the DNA methylation based predictor for maternal blood contamination, including performing feature selection, generating the thresholds for screening, and the application of the machine learning algorithms that produced the final predictor.

3. McEwen, Lisa M., et al. "DNA methylation signatures in peripheral blood mononuclear cells from a lifestyle intervention for women at midlife: a pilot randomized controlled trial." Applied Physiology, Nutrition, and Metabolism, vol. 43, no. 3, 2018, p. 233+

   My contribution was the development of the DNA methylation based predictor for percent change in body weight, including the feature selection and the application of the machine learning algorithm that produced the predictor.

Table of Contents

List of Tables

Analysis	GEO Dataset	Tissue	N	Females	Males	Age Range	Reference
	GSE55763	Whole blood	2,669	860	1809	31-75	Lehne et al., 2015
CMR Construction,	GSE84727	Whole blood	377	92	285	25-66	Hannon et al, 2016
sCMR Discovery,	GSE80417	Whole blood	224	117	107	26-79	Hannon et al, 2016
Sex Predictor	GSE111629	Whole blood	175	68	107	26-55	Horvath et al, 2015
Construction	GSE72680	Whole blood	350	244	106	26-77	Grady Trauma Project
sCMR Validation	GSE125105	Whole blood	697	388	309	17-87	Arloth & Binder 2019
	GSE79100	Kidney	31	16	15		Winterpatch & Lukassen 2018
	GSE80261	Buccal	96	57	39		Portales-Casamar et al 2016
sCMR Tissue	GSE61258	Liver	79	34	45		Horvath et al 2014
Concordance	GSE64509	Brain	25-41	17-22	8-11		Horvath et al 2015
	GSE87640	Immune cells	20	8	12		Ventham et al. 2016
Sex Predictor Validation	GSE132203	Whole blood	794	571	223	18-76	Grady Trauma Project

List of Figures

List of Symbols

β Beta, a measure of percentage DNA Methylation at a CpG site.

M, a logit transformation of the DNA Methylation β

L1, the penalty using absolute value of the coefficient estimates in penalized regression

L2, the penalty using squared value of the coefficient estimates in penalized regression

List of Abbreviations

DNAm DNA Methylation

CpG Cytosine-guanine dinucleotide unit

ML Machine Learning

450K Illumina Infinium HumanMethylation450 BeadChip array

EPIC Infinium MethylationEPIC BeadChip Kit

EWAS Epigenome-wide association study

PC Principal Component

PCA Principal Component Analysis

FDR False Discovery Rate

DMR Differentially methylated region

mQTL Methylation quantitative trait loci

SNP Single nucleotide polymorphism

CA Childhood Abuse

RF Random Forest

EN Elatic Net

BMI Body Mass Index

PA Physical Activity

PBMC Peripheral Blood Mononuclear Cells

CMR Co-methylated region

sCMR Sex-associated co-methylation region

TF Transcription Factor

TFBS Transcription Factor Binding Site

WB Whole Blood

Glossary

CoMeBack: Co-methylation with genomic CpG background.

Acknowledgements

I thank Dr Kobor and Dr Mostafavi for their time and humor.

I thank my mother for inspiring me and for her support of this endeavor.

I thank my friends for making me wiser and the other players for making me stronger.

Dedication

To S. P. Capt. Harlock

“Some illusions you just can’t give up. Everybody has one – if you are human.”

Chapter 1: Introduction

1.1 DNA Methylation in Humans and its Measurement with Bead Chip Microarrays

DNA methylation (DNAm), where a methyl group is covalently attached to the 5' carbon of the Cytosine base, is present across the genome of humans and primates, primarily in the context of cytosine-phosphate-guanosine (CpG) di-nucleotides. Diverse DNAm patterns are associated with the establishment and maintenance of cellular identity within the tissues of higher organisms[1-4]. Moreover, at the molecular level DNAm has been implicated in establishing, altering and maintaining the chromatin state and in the regulation of gene transcription[5, 6]. In addition to its role in developmental biology, recent evidence suggests that DNAm patterns in a given tissue might also associate with persistent environmental exposures on the one hand, and health and disease on the other hand, thus making this a promising approach for study in various epidemiological contexts[7-9].

The human genome contains about 28M CpG sites, with most of these methylated. A global hypo-methylation trend with ageing has been observed for adults, with the exception of certain genomic contexts, like CpG islands in gene promoters; for these, the opposite trends have been observed[10, 11]. Moreover, biological sex differences in humans, known to range from gross anatomy and physiology to the molecular level, include divergent epigenetic profiles, encompassing DNAm[12-20].

Several recent studies have documented associations between thousands of CpG sites and sex, as well as differentially-methylated regions (DMRs) that include imprinted loci, as well as many autosomal regions[21-23]. In such studies, the most widely used approach for high throughput measurement of DNAm at a subset of the millions of genomic CpG sites consists of bead chip microarrays.

The most popular DNA methylation arrays have been the Illumina Infinium HumanMethylation450 BeadChip (450K)[24]and EPIC[25]array platforms. These arrays measure 485,512 (450K) and 865,859 (EPIC) sites, respectively, out of the 28 million possibly methylated CpG sites in the human genome. In this dissertation, the bulk of the data were obtained from publicly available datasets[26]that were generated with the 450K platform.

Illumina's DNAm microarray technology relies on "quantitative genotyping of C/T single nucleotide polymorphisms (SNPs) introduced following bisulfite (BS) conversion", which basically converts unmodified cytosines to uracil while cytosines with DNA modifications remain unaffected. Notably, this particular readout of the BS conversion does not distinguish between DNA methylation (5mC) and hydroxy-methylation (5hmC). The DNAm arrays target CpG sites, along with a small number of non-CpGs (CHH and CHG), with oligomer "probes" adhered to "beads", which are randomly arranged within wells on the surface of each array [25]. Two types of probes are used by the Infinium platforms, Type 1 and Type 2, designed to

efficiently target different genomic context. Type 1 probes have two different probe sequences per CpG site, for methylated versus unmethylated CpGs. Type 2 probes have only one probe sequence per CpG site. The hybridization to probes of a bisulphite converted DNA fragment enables single base extension with a labelled nucleotide matching the nucleotide immediately upstream of the target CpG site. Such incorporation of a nucleotide results in fluorescent signal detection in either the red or the green channel, depending on the probe type and methylation signal. The intensity of fluorescence is translated into a level of DNAm for each CpG site - either as a β value, a number between 0 = not methylated, and 1 = fully methylated, or logit-transformed β values, called M values, which are recommended for linear statistical analyses[27]. These DNAm measurements characterize a single individual with high dimensional data, where each measured array probe represents a separate dimension characterizing a single individual.

To see how intermediate β values between 0 and 1 arise from measurements in bulk tissue, such as whole blood, which consists of a mixed pool of cells of different cell types, recall that the DNAm state of a single CpG site can be either unmethylated, $\beta=0$, or methylated, $\beta=1$. Thus, in a single diploid human cell, the DNAm state of a single CpG site can be either unmethylated on both alleles ($\beta=0$), methylated on one of the alleles ($\beta=0.5$), or methylated on both alleles ($\beta=1$). In a bulk tissue sample, the measured DNAm is an average across the mixed pool of different cells, and the average may have any value between zero and one. In this way the β measures

DNAm percentage in a sample[28]. Similarly, for such bulk tissue samples consisting of pools of different types of cells, differences in measured DNAm across individuals reflect either cell pool-average differences in DNAm across the cell types in the sample, or different mixes of cell types across individuals, or both.

1.2 High Dimensional DNA Methylation Array Data.

Even though the DNAm array methylation state measurements have sparse coverage of the 28 milllion genomic CpGs, there is still a very large number of measurements performed, some of which may yield values that are correlated due to some common underlying biological mechanism affecting a set of CpG sites, usually within a close genomic proximity, in concert [29, 30]. Moreover, while there are hundreds of thousands of CpGs on the DNAm arrays, many of those tend to have little variability in the same tissue across individuals [31], so a "natural" data reduction occurs, that can be exploited in downstream analyses. Such findings point towards a general approach of biologically-motivated DNAm array data reduction, where the key requirement would be the identification of redundant DNAm measurements associated with the same biological process, so that their common signal can be preserved in the reduced data, while statistical noise is filtered out. One powerful approach to such identification of redundant measurements is to integrate existing domain knowledge about DNA methylation into the data reduction algorithm[32]. For example, this is the tack of the "Co-

Methylation with genomic CpG Background" (CoMeBack) method described in the next chapter.

When DNAm array data is analyzed along with other phenotypic data, for example disease state, there are two principal ways to proceed: First, DNAm is modeled as the independent variable that is explained by the phenotype variable within a linear regression specification. In this case, the phenotype and other intrinsic variables, like age and sex, are the explanatory variables on the righthand side of the equation in the linear regression. Such association studies, termed epigenome-wide association studies (EWAS) [33, 34], have to cope with multiple tests of statistical hypotheses[35] when a large set of measurements from the array data are used in separate statistical tests. Second, the phenotype variable itself can be explained by multiple DNAm explanatory variables. In this case, one constructs a "predictor," for the phenotype using DNAm measurements obtained from the array, but the large number of array measurements poses several challenges, for example overfitting a linear regression specification for typical sample sizes that are much smaller than the number of DNAm array measurements[36]. In both scenarios, data reduction can be useful, whether it is biologically motivated, such as removing tissue-specific low-variability probes[31], or statistical, for example by focusing on estimated clusters or modules[37].

To better understand how a careful DNAm data reduction may be useful in both scenarios discussed above, I proceed to briefly discuss two important statistical notions in the context of DNAm array data. The two concepts are the "curse of dimensionality" [38] and the signal to noise ratio.

1.2.1 The Curse of Dimensionality for DNA Methylation Array Data

High dimensional data present unique analytical challenges that are never easy to deal with. We discuss the "curse of dimensionality" [36] in the context of high dimensional DNAm array data within the predictor scenario, where hundreds of thousands of DNAm measurements are potential candidates for predicting the phenotype of interest. Intuitively, high dimensional DNAm variability implies that two individuals are virtually never "similar" for sample sizes much smaller than the high dimension of hundreds of thousands of DNAm measurements. This, in turn, makes it difficult to detect patterns and associations in the data.

The curse of DNAm array dimensionality stems from two related problems. First, one expects that the DNAm probes that are truly associated with most phenotypes of interest are only a very small subset of the hundreds of thousands of DNAm array measurements. Second, the associations for these true signals with the phenotype are relatively weak in terms of explained variability (also known as R-squared) [39]. The hurdle lies in picking out the relatively few informative probes among the many

thousands of "noise" measurements, some of which may be spuriously correlated with the phenotype of interest.

For DNAm predictor scenarios, the curse of dimensionality is manifested as "over-fitting", manifested as improving accuracy in the training data when additional DNAm probes are used to predict the phenotype, while in new data, testing such expanded predictors yields increasing error as additional DNAm probes are included to "improve" the predictor [36]. In this scenario, one has to somehow select informative DNAm probes, or equivalently, filter out the noisy ones, which represents a form of data reduction. One recent example of such data reduction through algorithmic feature selection can be found in the "epigenetic clock" [40] which first selects and then uses only 353 DNAm probes out of the 27K measurements on the recently retired Illumina 27K array, in order to predict chronological age. We note that this data reduction is rather extreme, as there are several thousand DNAm array probes that are weakly associated with age and hence they represent candidates for an age predictor [41]. More generally, this example illustrates how careful data reduction that preserves the most informative DNA methylation measurements can be useful for the construction of parsimonious predictors for phenotype, that are based on DNAm array measurements.

1.2.2 Signal-to-Noise ratio

To further understand the potential of careful DNAm array data reduction, we consider a stylized EWAS scenario, where the methylation beta is modeled as "explained by," or regressed on, the phenotype variable of interest and intrinsic measures such as age, sex and ethnicity [33]. With the vast number of DNAm betas measured on the array, multiple tests for associations are bound to yield some false positives due to spurious correlation that occurs by chance in a given, typically small, sample. In this scenario, data reduction may be useful in two related ways.

First, performing fewer tests with the reduced data lowers the multiple test correction burden, which may help discovery when the data reduction preserves the statistical signal, defined as the variance of the expected methylation beta conditional on the explanatory phenotypic variables [36]. Less obvious, good data reduction may help improve statistical power, when the data reduction procedure boosts the signal-to-noise ratio, where noise is defined as the variance of the measurement error that is independent of the explanatory variables [36].

Data-driven reduction approaches such as removing non-variable probes [31], or focusing on estimated clusters or modules [37], aim to achieve these goals using the statistical variability of the data. An alternative approach would be to identify multiple DNAm array probes that are redundant in the sense that they measure the same biological effect. Subsequently, aggregating these redundant probe betas in a

single “summary” composite beta, by way of a weighted sum, would reduce the statistical noise for that composite DNAm measure. Intuitively, if the statistical errors of the redundant measurements were independent of each other, they would cancel out to some extent as they are added together. As a result, the statistical error of the single composite measure would have lower variance, which in turn would improve the signal-to-noise ratio versus that of each individual redundant probe. In this way, a sensible data reduction method may improve statistical power, or even enable the downstream applications of more flexible, in terms of number of parameters, statistical models that may otherwise overfit the data when the signal-to-noise ratio is low[36]. I now proceed to briefly outline the emerging data reduction methods in the context of DNAm array data analyses.

1.3 Data Reduction in DNAm Analyses

The data reduction methods that have been applied for DNAm analysis and can be considered as falling into two groups, based on their placement within the analytic pipeline. First, we can distinguish data reduction for exploratory data analysis, where broad global patterns of variability are extracted and summarized. The widely used Principal Component Analysis (PCA) is the archetypal example, along with mainstream alternatives like Latent Factor Analysis, Clustering, and others[36]. These exploratory analyses broadly correspond, but are not identical, to the “unsupervised approach” category in Machine Learning (ML) parlance[42]. Next, there are more targeted analytical pipelines that focus on phenotypes beyond DNA

methylation, like the construction of sparse epigenetic predictors. These analyses broadly correspond to the class of "supervised" ML algorithms, that use the phenotype "label" as an additional input. A number of well-regarded general purpose ML algorithms that have been applied to analyze DNAm array data, including elastic net[40], LASSO[43], random forests[44], sparse PCA[45] and others[46]. Such methods typically include a data reduction step as a key element of the algorithm. However, since such general-purpose ML methods are designed to have wide applicability in diverse domains, their data reductions are not rooted in specific biological models.

In the context of exploratory DNAm data analysis, data reduction may be incorporated within more complex ML algorithms that aim to characterize specific DNAm patterns that may be present in the data. For example, a sparse version of the PCA[47] incorporates a data reduction, so that the output transformation removes, or collapses, many, or most of the input DNAm measurements. In this case, the objective is to go beyond characterizing the global variability patterns, to extract only the strongest pervasive patterns as a more limited subset of the entire large set of DNAm array probe measurements. A typical use-case is illustrated by the widely used lists of top candidates for further study reported in academic papers.

On the other hand, DNAm analysis in a supervised ML context may involve constructing predictors or risk scores for disease like cancer or neurological disorders [48-50], imputation of unmeasured phenotype[51], or forensic epigenetics[52]. Consider for example the "elastic net" ML algorithm[53] has become widely used in DNAm array data analyses to produce epigenetic predictors. The elastic net uses phenotypic information beyond DNA methylation to reduce the DNAm array data. The data reduction step in such an algorithm aims to decrease the overall number of initial DNAm measurement inputs to a smaller subset of robust signals that is best suitable to prediction or imputation of the phenotype of interest. This intermediate step typically works by removing the measurements that either resemble noise, or are mutually redundant.

In addition to statistical ML data reduction techniques, DNAm analytical pipelines often reduce the data based on biological considerations, for example by removing invariable probes[31], as stated in section 1.2.2. The justifications for such variability filtering, which can be traced back to the gene expression literature [54-57], typically include arguments that small effects are not biologically informative or reproducible, while they also hamper the statistical analyses through a high multiple test correction burden. The widespread application of such mainstream biologically motivated pre-processing procedures indicates that there is room for a DNAm array data reduction that integrates epigenetic knowledge to produce output that is both biologically informative and useful as input to downstream analytic pipelines. One

possibility for such biologically-motivated data reduction involves focusing of co-methylated regions, encompassing groups of (adjacent) array probes, versus individual CpGs.

1.4 Analysis of Co-Methylated Regions: Considerations and Methods

Existing dimension reduction methods for DNAm array data that output co-methylated regions (CMRs) or DMRs, can be considered to fall into two categories based on their key assumptions. First, "within an individual" methods compress methylation ratio (beta) levels within each individual, while the other group of "across individuals" methods exploit the correlations of the methylation levels to other variables between individuals. DMR methods like Bumphunter [58] smooth over beta values of adjacent CpGs within an individual, while methods like DMRcate [59] and A-clustering [60] consider correlations of CpG betas with external phenotypic variables, or correlations between CpG betas, respectively. Consider the above two categories in some more detail.

As stated in section 1.1, in bulk tissue, the DNAm betas are often at intermediate levels between zero and one when they reflect the aggregation of stochastic methylation state across individual cells. As a result, regions with multiple proximal CpG sites that function as biological units, for example enhancers, may have intermediate methylation values for the adjacent CpG sites that vary from site-to site within an individual. To the extent that such values are redundant

measurements of the common DNAm state of the single biological unit, combining these signals may reduce the stochastic noise and improve the statistical performance in downstream analyses. This is the basic intuition behind methods like Bumphunter [58] that smooth the beta values over multiple adjacent sites within each individual.

On the other hand, when the prevailing co-methylation state of a multi-CpG region, that functions as a biological unit, varies across individuals due to phenotype, this inter-individual variation of the co-methylation state would be reflected in the single CpG sites that would be correlated across individuals. This presents an opportunity to statistically identify co-methylated regions, by examining adjacent CpG correlations across-individuals. The identified regions can subsequently be evaluated in terms of their biological role by using existing knowledge about functional genomic features, like enhancers, that may be tissue-specific [61, 62].

Across-individual correlations of adjacent CpGs can be evaluated directly as an alternative to "within-individual" smoothing of the DNAm betas at adjacent CpG sites, with the advantage of avoiding "over-smoothing", or more generally, choosing parameters for smoothing strength and window size. In general, algorithmic parameters for DNAm beta smoothing may not be easy to estimate, as they typically vary across genomic and tissue context, as well as across populations and conditions. For example, co-methylated regions vary due to diseases such as

cancers, which may be heterogeneous, or polygenic, or more generally, difficult to characterize extensively at the molecular level. Such complications would tend to preclude the inference of universally applicable hyper-parameters for algorithmic beta smoothing. These considerations help explain the motivation for methods like A-clustering [60] and DMRcate [59] that are based entirely on across-individual correlations of adjacent CpG sites.

On the other hand, when one evaluates the correlation of adjacent array probe betas across individuals, the genomic context, in particular the genomic density of CpG sites not measured on an array, can be useful because it has been shown to affect the strength of the beta correlations within individuals [29, 30]. Consequently, if genomic CpG density is sufficiently high, then the CpGs inter-individual correlation would tend to be above the genomic context-dependent background levels[30]. Thus, adjacent array probes that measure a subset of such dense genomic CpG sites would tend to have measured betas that are correlated across individuals.

Based on these considerations, it is possible that a method for a biologically motivated data reduction, which constructs co-methylated regions while also incorporating a-priori CpG coordinates from the reference human genome sequence, may achieve statistical performance improvement, as well as facilitate enhanced interpretability in terms of existing knowledge about annotated genomic features. For example, EWAS studies often focus on "group hits" where there are

multiple adjacent CpG probes that are significantly associated with the variable of interest[59, 63], because their genomic proximity makes it more likely that they function as an individual biological unit and hence the detected correlations are not spurious.

Biological considerations as the ones discussed above can form the basis of a "pre-processing" DNAm array data reduction approach that incorporates the genomic proximity of array probes as a key element, along with the complete set of all genomic CpGs annotated in the reference human genome sequence. In contrast to statistical approaches like A-clustering[60] or DMRcate[59] that only use across-individual correlations, such data reduction would be based on a simple biological model. As a form of general pre-processing, it would have the advantage of offering flexibility regarding downstream analytical pipelines, so that unsupervised, supervised, or other methods can be used following such biologically-motivated DNAm array data reduction.

1.5 Thesis Objectives

The overarching objective of this dissertation was to develop a powerful, easy to use method and software for DNAm microarray data reduction that can enhance discovery and specificity, and to demonstrate its utility through several biologically important applications. To accomplish this, I undertook the following studies described in chapters 2, 3 and 4. Chapter 2 describes the biological data reduction

algorithm “Co-Methylation with genomic CpG Background” (CoMeBack) design and development in detail, and demonstrates an application for co-methylated region (CMR) construction in whole blood and the CMRs’ biological characterization in terms of chromatin state, known mQTLs and transcription factor binding sites. The chapter also provides examples of possible downstream analyses, including an EWAS application for CMRs associated with chronological age. While this application was meant to illustrate an example use of CoMeBack within a typical pipeline, it is also biologically important, because of the clinical relevance of age and ageing to health and disease[11].

The subsequent chapter 3 presents an extensive application of the CoMeBack method for identification of co-methylated regions associated with Sex, which is among the few phenotypes that have well established DNAm signatures[12, 21]. The comprehensive characterization of sex-associated CMRs showcases how the CoMeBack CMRs can be used as an upstream data reduction to enhance discovery and statistical specificity within a rich downstream analytical pipeline. The application of CoMeBack to identify CMRs associated with sex was motivated by the biological importance of the Sex phenotype in humans, as well as by its feasibility using publicly available data to form large aggregated samples.

In the next chapter 4 of this thesis I discuss three other mainstream machine learning (ML) algorithms that I have applied in different projects that involved

DNAm data analysis. I used these ML algorithms for sparse predictor construction and also to perform feature extraction. These applications addressed the specific objectives and context of the respective studies. In the chapter, I discuss the performance of the constructed predictors, as well as performance limitations, along with considerations relevant to the choice of a ML algorithm in these cases.

The final chapter 5 presents a discussion, including future research directions.

Chapter 2: CoMeBack: DNA Methylation Array Data Analysis for Co-Methylated Regions

2.1 Introduction

DNA methylation (DNAm) describes the covalent chemical attachment of a methyl group to the 5' carbon of the cytosine nucleotide, typically next to a guanine nucleotide, referred to as a CpG site. Diverse DNAm patterns are associated with the establishment and maintenance of cellular identity within the tissues of higher organisms. Moreover, at the molecular level, DNAm has been implicated in establishing, altering and maintaining the chromatin state, and in the regulation of gene transcription[5, 64]. In addition to its role in developmental biology, recent evidence suggests that DNAm patterns in a given tissue may also associate with persistent environmental exposures, as well as overall health and disease, making this an intriguing mechanism to study in various epidemiological contexts [7-9, 65]. Although substantial amounts of data have been generated, few DNAm associations have been replicated across different studies [34]. This may in large part be due to the complexity of various environments across populations studied, how a given environmental factor is measured and defined, and the power needed to detect true statistical associations.

Most DNAm association studies, termed epigenome-wide association studies (EWAS), utilize array technology, with the most popular being the Illumina Infinium 450K (450K) [24] and EPIC [25] array platforms. These arrays measure

485,512 (450K) and 865,859 (EPIC) sites, out of the 28 million possibly methylated sites in the human genome. The sample sizes required to overcome multiple test corrections are likely much larger than the vast majority of EWASs contain to date [39, 65]. There is a need to address the high dimensionality of DNAm array data in a biologically driven way, in order to detect any true statistical associations. Here we propose a method that clusters DNAm array data by exploiting the biological nature of DNAm where certain regions of proximally located DNAm sites exhibit correlated methylation state [29, 30]. Among such DNAm patterns, the simplest contain groups of proximal CpG sites, as seen in CpG islands, broadly defined to have high CpG content within a short region of DNA [66]. These CpGs have been shown to typically function together as a unit [67-71] and have been investigated in terms of their correlated methylation status and whether they are joined together by various biological mechanisms [58-60, 72-75]. While less than 10% of the approximately 28 million CpGs in the human genome are in CpG dense regions, they are enriched in the promoters and transcription start sites of developmental and housekeeping genes, implying potential biological relevance [67]. The selection and distribution of CpGs measured on a particular DNAm array is not arbitrary, as DNAm arrays are designed to be enriched for probes measuring such sites with predicted or established functional importance. For example, the 450K array interrogates the methylation status at many thousands of groups of proximal CpG sites whose methylation is correlated across individuals [76, 77].

Based on this knowledge, we combine such sites into regional units to reflect the biology of DNAm, and independent of any variables of interest, making this method unique amongst currently available DNAm regions-based analyses. While specialized methods for DNAm array data analysis have accounted for the spatial correlations of proximal CpG DNAm levels, by either adjusting their p-values, or smoothing over their methylation levels in EWAS [33, 58], or by identifying differentially methylated regions [59, 60, 72-75], or by correlation assessment visualization [77], few tools implement a universal, biologically driven, unsupervised pre-processing approach for clustering adjacent DNAm array probes. For example, while unsupervised dimension reduction methods like RPMM [78] aim to extract a very small number of latent variables based on sophisticated statistical models, our method instead uses a simple biological model of genomic CpG proximity to produce a large number of probe clusters. In this way, our method has some similarity only to the A-clustering (Aclust) method [60] which also produces many clusters based on correlated methylation. However, Aclust does not incorporate any information about unmeasured genomic CpGs and instead relies solely on a probe distance window, which is optional. As a result of this different design, Aclust would group correlated probes that would not be eligible for consideration by CoMeBack, and vice versa. CoMeBack also differs from methods that estimate differentially methylated regions (DMRs), for example DMRcate [59], because CoMeBack is an unsupervised method that does not use any phenotypic information to construct

the regions, and instead relies only on the correlated methylation state of certain eligible adjacent array probes.

Our method was designed to construct co-methylated regions (CMRs) with high specificity, biological interpretability and enhanced downstream discovery of statistical associations with multiple adjacent CpG sites. For these objectives, we addressed three related issues. First, two adjacent array probes may be correlated when they belong to the same contiguous CpG site region. We determine which probes potentially measure a single functional unit by considering all genomic CpG sites not measured on the array, referred to hereafter as background CpGs. Second, by jointly considering adjacent DNAm sites, we "borrow" statistical power across correlated sites to improve the specificity and downstream discovery of statistical associations with variables of interest [63]. Constructing CMRs results in data dimension reduction that implies a lower multiple-test correction burden than when using individual probes in downstream analyses, like EWAS. Finally, combining sites into regions results in expanded overlaps with annotated genomic features relative to the individual sites from the CMR, potentially facilitating biological interpretation in downstream feature enrichment analyses.

2.2 Methods

2.2.1 The CoMeBack Algorithm: Rationale and Design

Since neighboring CpGs are often in the same methylation state within an

individual [29, 30], we conjectured that adjacent array probes may have correlated methylation levels across individuals. This was the reflection of existing findings [29, 30, 63] that within an individual, genomic CpGs within 400bp of each other are highly likely to be in the same methylation state, with the probability declining to an overall background of about 74%, as genomic distance between CpGs increased beyond 2Kbp. Then, since array probe coverage of genomic CpGs is sparse, our algorithm would evaluate two contiguous array probes with adjacent genomic coordinates only if the reference human genome annotation contains a "chain" of unmeasured genomic CpGs of a specified density between them. We set the density of genomic CpGs needed to "chain" proximal array probes to be at least one CpG every 400bp. This density was chosen due to the likelihood of high correlation for such proximal CpGs, as previously reported [29, 30, 63]. If such a chain of unmeasured CpGs was present in the reference genome build, the two array probes that define the ends of the chain were considered to be likely correlated above background levels, so long as they were less than 2Kb apart. By using unmeasured intermittent CpGs from the human reference genome to link array probes, we could avoid using a fixed length window for array probes, adopting instead a sliding window over all the genomic CpGs, including those not measured on the array.

The underlying rationale for this design of CoMeBack was to reduce the false positives for correlated adjacent probes, while improving the detection of possibly biologically relevant co-methylation. For example, as depicted in Figure

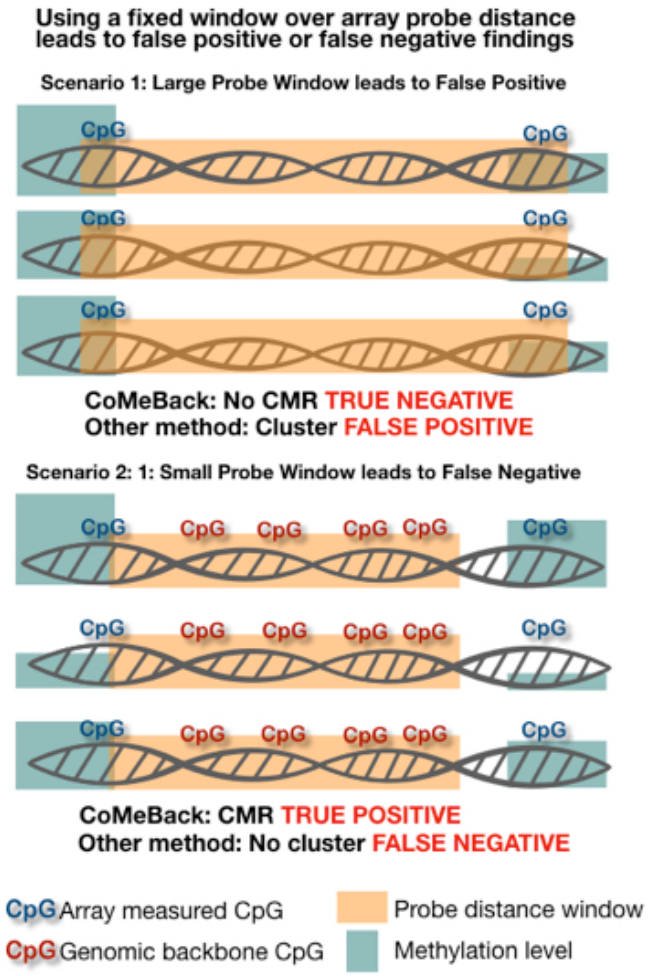


Figure 1. In contrast to existing methods, CoMeBack reduces false discoveries by using a sliding window over genomic CpG distance.

1, by using a wide array probe window of 2Kbp, a researcher may evaluate the correlation of two probes with no other intermittent unmeasured CpGs, and if the probes are spuriously correlated, which is likely considering their distance, they would still be grouped together producing a false positive. Such cases would be avoided by the CoMeBack algorithm, as the correlation of the two probes would

be evaluated only if the reference human genome annotation showed that there are at least four intermittent unmeasured CpGs linking the probes with a chain of at least one CpG per 400bp. Continuing with the example, if the researcher attempts to mitigate the false positive rate and selects an alternative, narrower array probe window of 1Kbp instead, then a false negative result would occur for any two adjacent array probes that measure the ends of a CpG island spanning even a single base pair over the 1Kbp probe window: the probes would not be examined for correlation, even though these sites are likely functioning as a unit. In contrast, CoMeBack would use the chain of contiguous unmeasured genomic CpGs to slide the genomic CpG window until it incorporates the next array probe. The main assumption underlying our algorithm is that genomic regions with a "dense" CpG background, defined here as at least one CpG per 400bp, are more likely to function as, or to mark a co-methylated biologically regulated unit, and hence their correlated methylation status is less likely to be spurious. This assumption is consistent with, and motivated by, the existing empirical findings, as discussed above.

Once genomic CpGs are chained together, CoMeBack estimates DNAm correlation across individuals for all adjacent array probes in the chain, which is declared a CMR if all pairs of adjacent probes are correlated above a given threshold. Thus, although genomic CpG coverage on the array is generally sparse, two adjacent probes that are more than 400bp away, but less than 2Kb away, may still be incorporated into a given CMR, so long as they are chained by the

presence of unmeasured CpGs with a density of at least one CpG per 400bp, and they are correlated across individuals above a sample size dependent threshold.

2.2.2 The CoMeBack Algorithm: Conceptualization and Implementation

Figure 2 illustrates these elements of CoMeBack by representing the algorithm as three conceptual stages. First, for a given DNAm array dataset as input, the reference human genome was scanned between every two adjacent array probes that are no more than 2Kbp apart, for the presence of an unmeasured genomic CpG chain of at least one CpG per 400bp. In this way, the algorithm aimed to minimize calls due to spurious correlations when two adjacent array probes are not linked by an unmeasured CpG chain. Note that the first stage of CoMeBack did not use the actual methylation data; instead it depends only on the CpG genomic location of the provided array probes, and the reference human genome build.

The second stage used the DNAm data to evaluate across-individual correlations for the adjacent array probes linked by an unmeasured CpG chain. If the estimated array probe correlation was above the threshold, the two adjacent probes were included together within a co-methylated region (CMR), and the next adjacent array probe was evaluated, in the same way, for inclusion in the current CMR. We also implemented an optional constraint requirement on the DNAm levels of adjacent CMR probes, so that the absolute difference between the median levels is below a user-specified value.

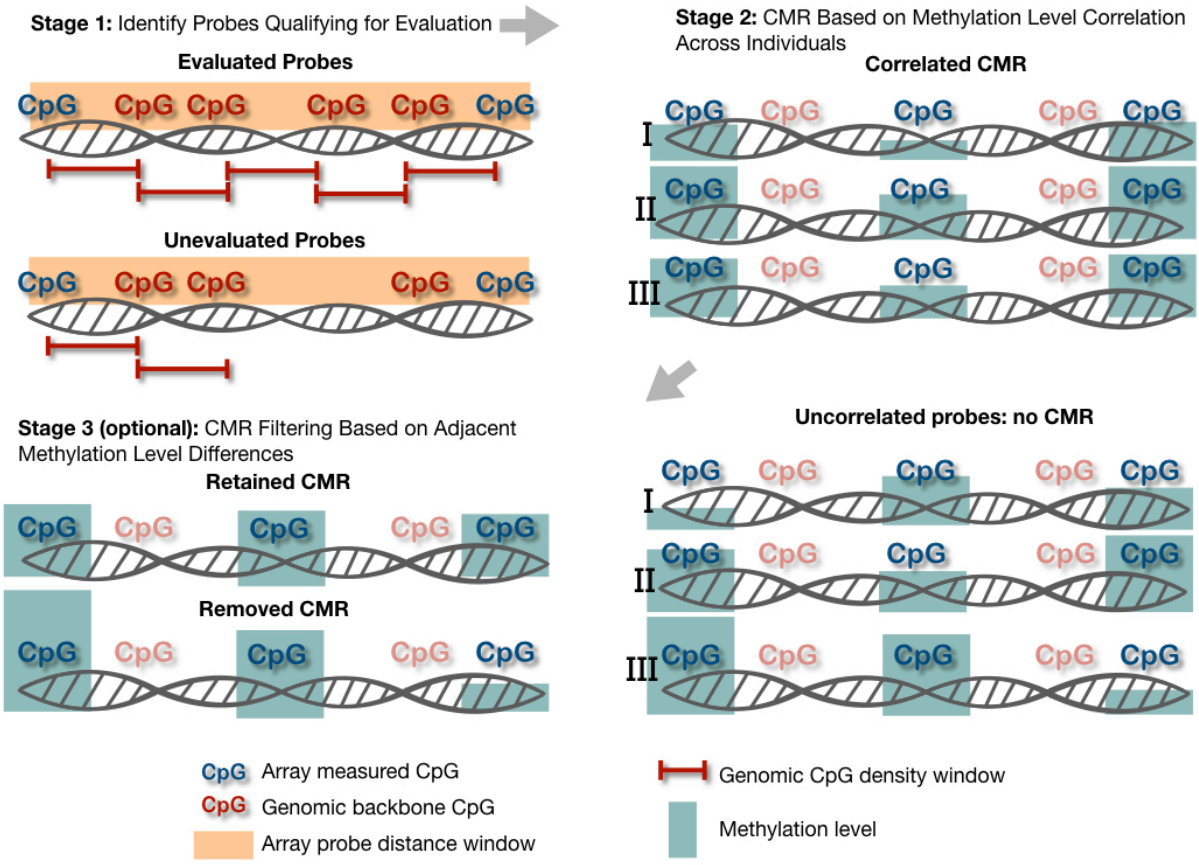


Figure 2. CoMeBack has 3 stages, where DNAm data are used at the second stage, along with unmeasured annotated genomic CpGs, to construct co-methylated regions (CMRs) from correlated array probe.

To evaluate the relationship between correlated genomic CpGs within individuals and the correlation of these CpG sites across individuals, we performed a simulation, to empirically determine a default guidance specification for a correlation threshold parameter dependent on sample size.

While this guidance can be useful in cases of small samples, the user may wish to fix the correlation threshold parameter to a higher value, if the objective is to

detect CMRs with certain minimum DNAm covariation across individuals. Since CMRs can be expected to differ across tissues, guidance for future estimation of reference CMRs in a certain tissue would include using a sample size of at least N=500, with a correlation threshold of at least 15%.

In summary, the algorithm initialized a new CMR with a single array probe, subsequently incorporating additional proximal adjacent array probes, if the following three conditions were met: 1) the genomic distance between two adjacent probes was less than 2Kbp, or as set by the user, and 2) the reference human genome build annotates unmeasured intermittent genomic CpGs between two adjacent array probes with a 400bp-density, and 3) the DNAm correlation (Pearson, Spearman or Kendal, as set by the user) between any two adjacent probes was above a sample size dependent, or user-defined threshold. Optionally, the user can further constrain the absolute difference between adjacent probes' median DNAm levels below a specified threshold. When these thresholds were met, the current CMR was declared finished and a new CMR was initialized, using the next adjacent probe as its starting first probe. Array probes that did not meet these thresholds were considered "singleton" non-CMR array probes.

Finally, the optional third stage of CoMeBack estimates, per individual, a composite methylation measure for each CMR. By default, the composite CMR methylation was defined as the scores of the first principal component of the CMR probes' DNAm levels, which is simply a weighted average, normalized by the

sum of the loadings of the probes in the CMR. This composite measure has the same scale as the individual probes and can be interpreted as a summary measure of the individual probes' DNAm level when the optional constraint for adjacent CMR probes to have similar measures has been included. The user also has the option to alternatively set equal weights for the CMR composite measure, corresponding to an average CMR probe methylation, or to use the probe median methylation as the composite measure.

We designed our method for conceptual simplicity and ease of use, via an R-software package, so that an investigator would input DNAm values and receive CMRs specific to their data as output. In addition, users could optionally calculate composite CMR methylation measures, defined to meaningfully aggregate the multiple methylation states of the individual CpG sites grouped within each CMR. The CMRs and their methylation measures can then be used for common downstream applications such as EWAS and PCA, amongst others. We illustrate below such applications after applying CoMeBack to construct reference CMRs for whole blood.

2.2.3 Reference CMRs for Whole Blood

2.2.3.1 Public Data Pre-Processing

We constructed whole blood CMRs using Illumina 450K array data from a large, ethnically heterogeneous aggregated cohort (N=5,191), comprised of several publicly available datasets (GSE55763, GSE84727, GSE80417, GSE111629,

GSE72680) [73, 79-81]. First, intra-dataset normalization was performed and batch effects were corrected within each cohort, using the ComBat function from the R-package sva [82]. The datasets were then merged and corrected for inter-dataset batch effects using the same function. The subset of probes that were measured in all datasets was filtered by removing XY chromosome binding probes, non-CpG probes, cross-hybridizing probes [83, 84], and probes containing, or immediately adjacent to a single nucleotide polymorphism (SNP) with a minor allele frequency ≥ 5%, as annotated by the Illumina manifest and as reported previously [84]. Finally, probes present on the 450K array but absent from the EPIC array were removed to allow CoMeBack performance to be directly compared between both platforms. This pre-processing resulted in 404,779 CpG sites that served as the input to CoMeBack. As whole blood is composed of multiple cell types, each with distinct DNAm patterns, where individuals vary in cell type proportions, we aimed to ensure that the estimated CMRs were not primarily capturing cell-type heterogeneity [85]. To accomplish this, we included cell type proportions predicted with the Houseman method [28], as implemented in the R-package minfi[85], as covariates in a linear regression model to adjust the data for downstream analyses. The residuals from this model were used to estimate the CMRs in whole blood.

For the newer Illumina EPIC platform, we also constructed whole blood CMRs using a single large cohort (N=795) from a publicly available dataset (GSE132203) [86] that was pre-processed similarly to the other public datasets above.

2.2.3.2 CMR Construction in Whole Blood Public Data

CMRs were constructed by estimating correlations across individuals in cell-type corrected whole blood data. These CMRs were used to estimate composite methylation measures for each individual in the DNAm data before cell-type correlation. Reference CMRs were constructed using a stringent minimum correlation cut-off of 30% in order increase the likelihood for replication and offset the imprecise correlation estimates typically obtained from smaller datasets, such as those prevalent in public data repositories. The maximum probe distance was set at 2Kbp, and an alternative CMR construction was also performed at 1Kbp maximum probe distance.

2.2.3.3 Robustness of CMRs

To assess how reproducible the constructed CMRs were in the presence of biological and technical variability, processed data was split into two equal-sized, random sub-samples with the algorithm run separately on each one. This process was repeated five times. We then evaluated how many of the constructed CMRs were identical across both datasets, and how many CMRs had at least a 2-probe overlap with CMRs constructed in the other sub-sample.

2.2.4 Characterizing the Reference CMRs

2.2.4.1 Sparse Coverage of Genomic CpGs with Array Probes

To assess the sparsity of array probe coverage of genomic CpGs included within the estimated CMRs, the number of array probes per CMR, the median CMR base

pair (bp) length, and the median density of background CpGs were characterized. For each n-probe CMR (n=2,3,..), we reported two metrics: the median bp length of all n-probe CMRs, divided by the number n, as well as the median number of background CpGs for all n-probe CMRs, divided by n.

2.2.4.2 Chromatin State Enrichment

We sought to determine the chromatin state of the genomic regions overlapping the probes of a CMR. To accomplish this, the Roadmap Epigenomics [61] ChromHMM [62] 18-state model for PBMCs was used to estimate the overlap enrichment of genomic regions in different chromatin states within the CMRs spanning the genomic coordinates from the first to the last probe, versus the non-CMR probe regions, two base-pairs in length, that span CpGs assayed in the non-CMR singleton probes. The overlaps with chromatin states were counted using the R-package GenomicRanges [87], with no-overlaps counted as zeroes and overlaps counted as ones.

2.2.4.3 Transcription Factor Binding Sites

CMR methylation state could potentially be affected by, or itself may affect, the binding of transcription factors (TFs) through their binding sites [88]. Hence, we examined CMR enrichment for known transcription factor binding site motifs, as compiled in the HOCOMOCO v11 database [89]. We scanned regions spanning 200bp on either side of a probe's assayed CpG, using the tool FIMO from the MEME software suite [90]. For any enriched binding site motifs, we checked for

any known effects of DNAm on TF binding affinity, using as reference the reported binding specificities of full-length human transcription factors and extended DNA binding domains to (un)methylated DNA for 542 transcription factors [91].

2.2.4.4 Genetic Control

Since the CMR CpGs are variable across individuals by design, we examined their enrichment for 34,391 known whole blood DNA methylation quantitative trait loci (mQTLs), as reported in the ARIES study [92]. CMRs were annotated as containing versus not containing an mQTL probe and enrichments were estimated using Fisher's exact test. We considered the possibility that for an mQTL, the observed methylation levels of adjacent array probes may be very different when the genetic variant affects only one probe, but not the other. Focusing on potential mQTLs with minor allele frequency (MAF) above 5%, we examined the corresponding to 95^{th} quantiles for the absolute difference between adjacent probes' betas. The mQTL enrichment in CMRs where the maximum of the absolute difference over all pairs of adjacent probes was greater than 10%, was determined against the background of all CMRs.

2.2.5 CMR-based EWAS and Comparison with DMRcate

To illustrate potential downstream applications of the CMRs estimated in whole blood, we conducted a CMR EWAS for chronological age using CMR composite betas, defined in section 2.2.2 above, estimated before cell-type correction. A linear model

for age was estimated with the CMR composite betas, as output by the third stage of CoMeBack, using the bioinformatically predicted cell type counts as covariates, and constraining their coefficient estimates to be between zero and one. Training data consisted of a random sub-sample of half the data, while the remaining observations were used as a testing hold-out sample for validation. To avoid any circular use of data in our analysis, the CMRs used for the CMR EWAS were constructed in the training data only. We restricted the age of the subjects to be between 20 years and 90 years to avoid potential non-linear relationships with epigenetic changes during early life and old age as previously observed [40]. Multiple-testing correction was performed using the Benjamini-Hochberg (BH) [35] false discovery rate (FDR) control at one percent. We applied an ad-hoc biological effect size filter, where the implied change in composite CMR beta over 60 years of age difference would be at least five percent. In addition to the hold-out sample validation, we also used the EPIC dataset for the validation of CMRs identified to be associated with age.

2.2.6 Comparison with Existing Methods

2.2.6.1 Comparison with A-Clustering

We compared the CoMeBack CMRs to the clusters output by the A-Clustering (Aclust) algorithm [60]. Aclust starts by placing each probe in its own cluster, subsequently merging clusters without considering genomic CpG density or the distance between correlated probes. An extra pre-processing step (termed "d-base-pair-merge" DBPM) can be included, that uses a probe distance window to

force all probes between two correlated probes to be included in a cluster, even if the intervening probes are not correlated. Aclust was run on our sub-samples, with a distance (defined as one minus the correlation) threshold of 0.7 that matches the 0.3 correlation cut-off used for CoMeBack), and including its DBPM pre-processing step with 2Kbp window (selected again to match the probe window size used in CoMeBack). We evaluated how many of the Aclust clusters were identical with the CMRs constructed in the same data. Enrichment of the Aclust clusters for chromatin states was also performed, similar to the CMR enrichment analysis above.

2.2.6.2 Comparison with DMRcate

Since the CMR-based EWAS is an important application of CoMeBack CMRs, we compared these results to the output of DMRcate [59], which was used to estimate differentially methylated regions (DMRs) associated with chronological age. We used DMRcate with FDR of 1% to match CoMeBack, and with the recommended default parameters. We considered all DMRs that overlapped with probes included in the CMRs, examined the validation of these DMRs in the hold-out testing data, and compared this to the validation of the CMRs that had significant age associations.

2.2.7 Principal Component Analysis of Composite CMR Methylation

We performed a principal component analysis (PCA) of the whole blood CMRs composite betas and compared them to the PCA of the individual probes that

composed the CMRs. In both cases, the PC scores were regressed on the estimated cell type counts, age and sex variables, to assess whether the associations of phenotypic variables of interest with global DNA methylation patterns captured by the top PCs of CMR probes would be retained, corresponding to associations found with the PCs of the CMR composite betas. We also considered variable probes alone, defined as probes where the methylation beta difference between the 99- and 1-percentile was at least 5%, and performed the comparison for the CMRs that contained at least one such variable probe.

2.2.8 Reproducibility and Software Availability

To enable uptake of our method and to facilitate reproducibility of our results, the CoMeBack open source R-package is publicly available at bitbucket.com/flopflip/comeback. Any details of interest can be further examined directly in the source code.

2.3 Results

2.3.1 Identification of CMRs in Whole Blood

The CoMeBack algorithm constructs dataset-specific CMRs taking as input all genomic CpGs and DNAm array measurements. It can be expected that a subset of the CMRs constructed using different datasets will be unique to a particular tissue, genetic background or environment, but there may also be some "reference" CMRs which will be common amongst diverse datasets, at least

within a given tissue. Reference whole blood CMRs that contain at least one variable probe for 2Kb and 1Kb probe windows respectively are available upon request, for future research that may incorporate the CoMeBack CMRs in downstream analyses. Additional reference CMRs for the EPIC platform are available upon request.

To assess the replicability across studies of the estimated CMRs, we used five-fold cross-validation across two equal-sized random splits of our data (see Methods). On average, there was an 78% identity of the estimated CMRs and 92% percent of the CMRs estimated in the training data had at least two probes in the testing CMRs.

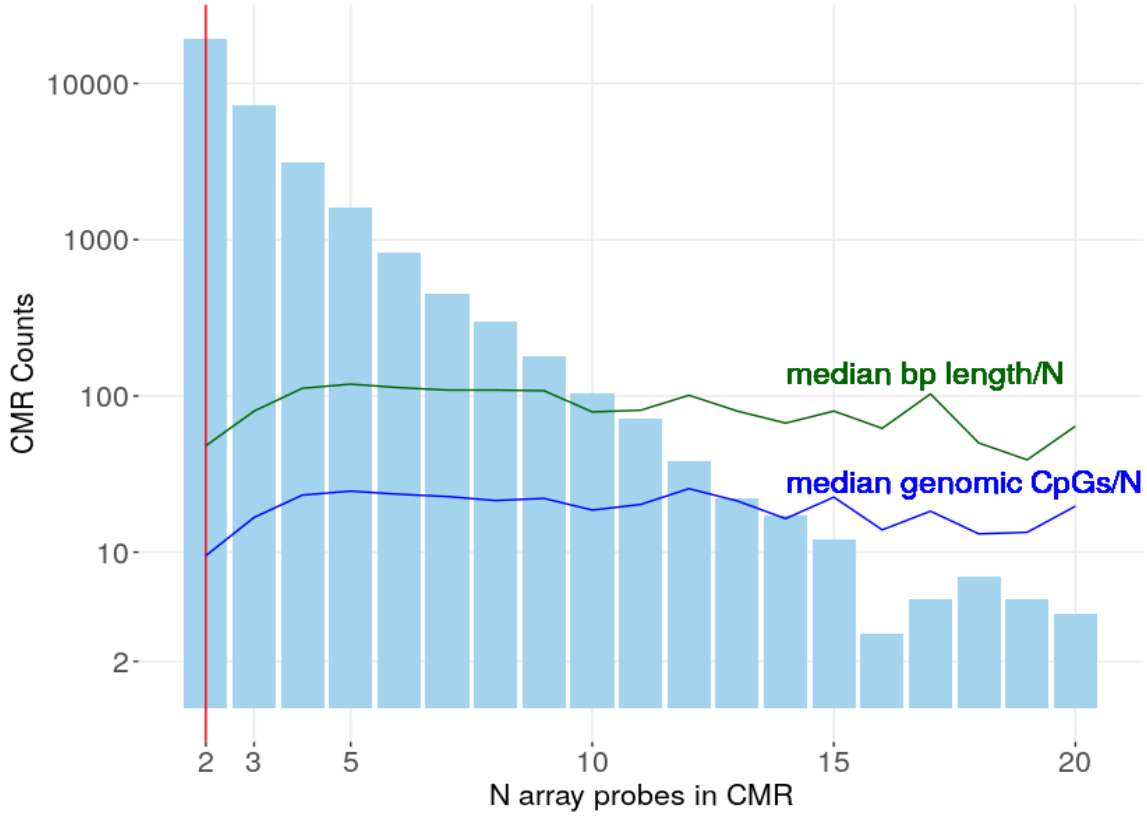
2.3.2 CMR Characteristics

2.3.2.1 Sparse Coverage of Genomic CpGs with Array Probes

Using the whole blood dataset we compiled from public data, CoMeBack constructed 33,572 CMRs. These CMRs included 97,424 probes, comprising approximately 24% of all 450K probes used as input. Figure 3 depicts the characterizations of CMRs in terms of: the number of CMRs with a given number of array probes per CMR; median size, measured in bp per CMR probe, and median number of genomic CpGs per CMR probe. The median CMR had two probes, and 80% of all CMRs had three or less probes. There were relatively few large CMRs that included more than 20 array probes, with the largest CMR containing 61 CpG probes. The 2-probe CMRs had a median of 19 background

CpGs, and a median total length of 96 bp. For the 3-probe CMRs, the median background CpG number was 50, with a median length of 239 bp. Overall, the density of array probes across CMRs of different size varied around 75 bp per CMR probe and 18 background CpGs per CMR probe. The relatively flat green and blue density lines in Figure 3 indicate that the density of array probes versus background CpGs was similar across CMRs with differing numbers of probes.

Figure 3. Most CMRs included three or less probes.



The red line is the median CMR. The green line is the median size, measured in bp per CMR probe. The blue line is the median number of genomic CpGs per CMR probe.

2.3.2.2 Chromatin State Enrichment

To investigate the functional relevance of CMR construction, we analyzed enrichment of chromatin states within the CMRs. Figure 4 shows how several chromatin states were enriched in CMR versus non-CMR probes. All enhancer states were enriched, especially gene and bivalent enhancers, as well as Polycomb repressed states.

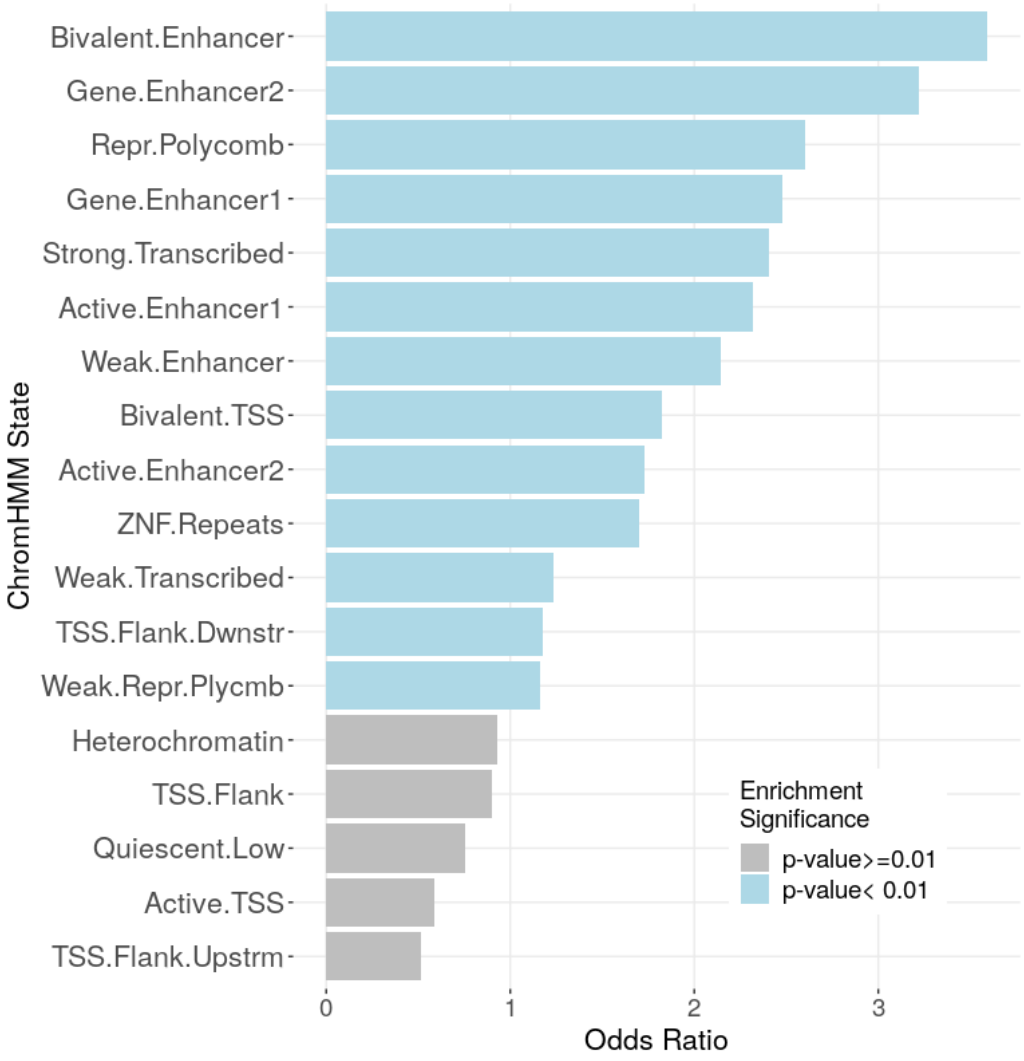


Figure 4. Whole Blood CMRs were enriched for enhancer chromatin states. Enrichment for overlaps of CMRs with chromatin-state regions (ChromHMM 18-state model) versus overlaps for non-CMR singleton probes.

2.3.2.3 Other Enrichments: Transcription Factor Binding Sites and mQTLs

To investigate any potential role of CMRs in the regulation of gene expression, we examined the enrichment of CMRs for known transcription factor binding site (TFBS) motifs. Out of the 404 motifs considered, 98 were present in CMR probe-spanning regions (see methods), with 38 motifs enriched in the CMRs as shown in Figure 5.

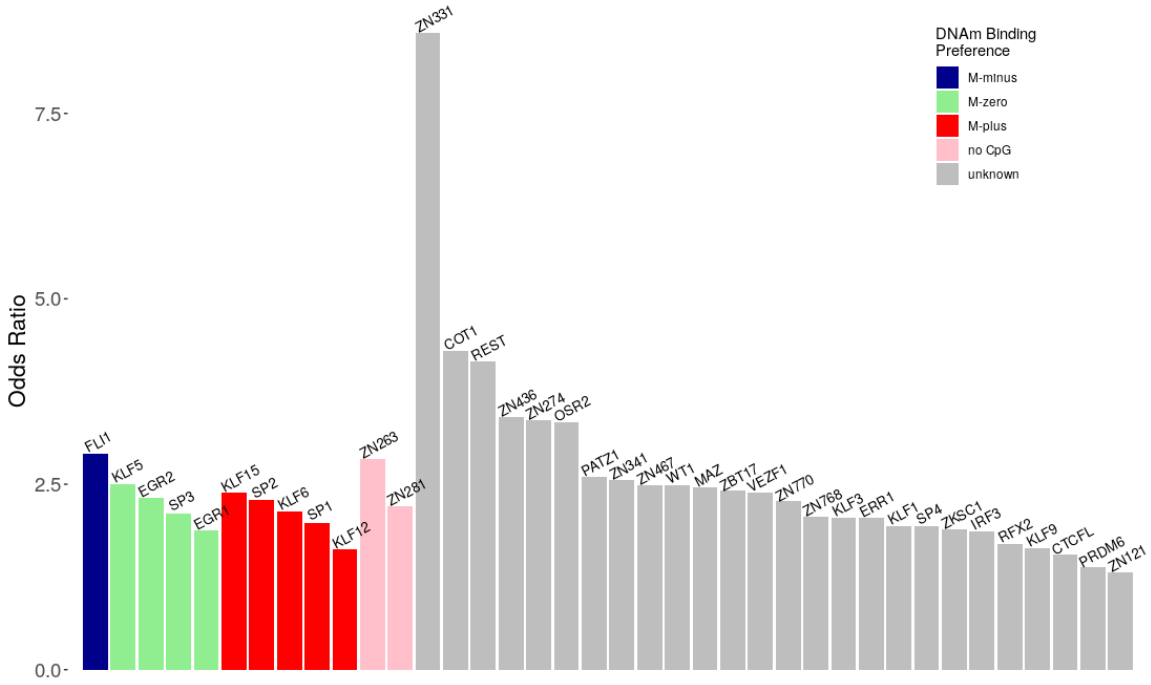


Figure 5. Significantly enriched (at 1%, Fisher exact test) TFBS motifs
The colors indicates TF preference for (non) methylated CpG, labeled M-plus(minus).

Several of the transcription factors (TFs) which bind these enriched TFBSs are known to function in blood cells, for example the KLF and SP family transcription factors. Moreover, while most of the 38 enriched motifs did not have a known affinity preference for positive or negative DNAm, we found six TFs whose binding affinity is affected by DNAm status of the CpGs contained within the motif.

Five of these TFs were “M-plus” (see [91]), the label for TFs that have increased binding affinity to methylated motif CpGs, while one had decreased, “M-minus”, binding affinity to methylated motif CpGs..

Considering the potential genetic influence on CMRs, we found that the majority of the CMRs did not include a known whole blood mQTL, as depicted in Figure 6.

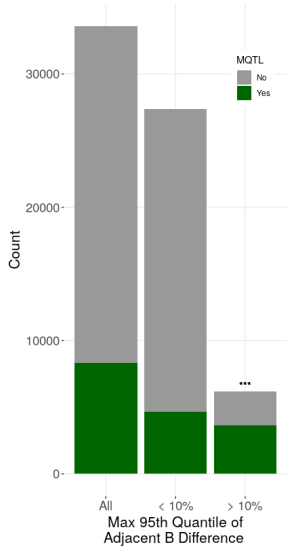


Figure 6. Enrichment for mQTL probes identified in the ARIES dataset.

Next, we found that CMRs where the maximum of the 95th quantiles for the difference between adjacent probe medians was greater than 10%, were enriched for whole blood mQTLs. Conversely, CMRs where the maximum of the 95th quantiles for the difference between adjacent probe medians was less than 10%, were not enriched for mQTLs.

2.3.3 CMR-based EWAS Application: DNAm Association with Chronological Age

A widely used study design, EWAS, relates genome-wide DNAm of individual probes to phenotypic variables or health outcomes [33] by using linear regressions, resulting in a large number of tests being performed. To assess the utility of CMRs for EWAS, we tested CMR associations with chronological age, which correlates with both DNA methylation and predicted cell type proportions [40, 93, 94]. For this CMR-based EWAS, we used the whole blood CMRs composite DNAm values, estimated in the training dataset, and regressed them on age and cell count covariates, followed by validation in the testing data (see Methods).

Focusing on statistically significant CMRs, where the calculated change of composite DNAm values over 60 years of age was at least five percent, resulted in 1,332 "large effect" CMRs (out of 18,388 CMRs with BH FDR below one percent) comprised of 4,660 probes. Of these CMRs, 911 had increasing and 421 had decreasing methylation with age. In general, there was a trend observed where increasing DNAm with age was more frequent amongst CMRs with large effect sizes.

Figure 7 shows that of 1,332 significant CMRs, 1,291 were validated (2.4% false positives), while in the hold-out test data, 1,325 were validated (0.5% false positives), with a high correlation between the effect sizes in the training and

testing data (R-squared of 99%), as depicted in Figure 8. The top age-associated CMRs included well-known age-related genes that contain multiple CpGs which were highly correlated. The best documented example of this being the 3-probe CMR within *ELOVL2*, for which methylation in this gene has been repeatedly found to associate with age [95]. The validated age-associated CMRs are available upon request.

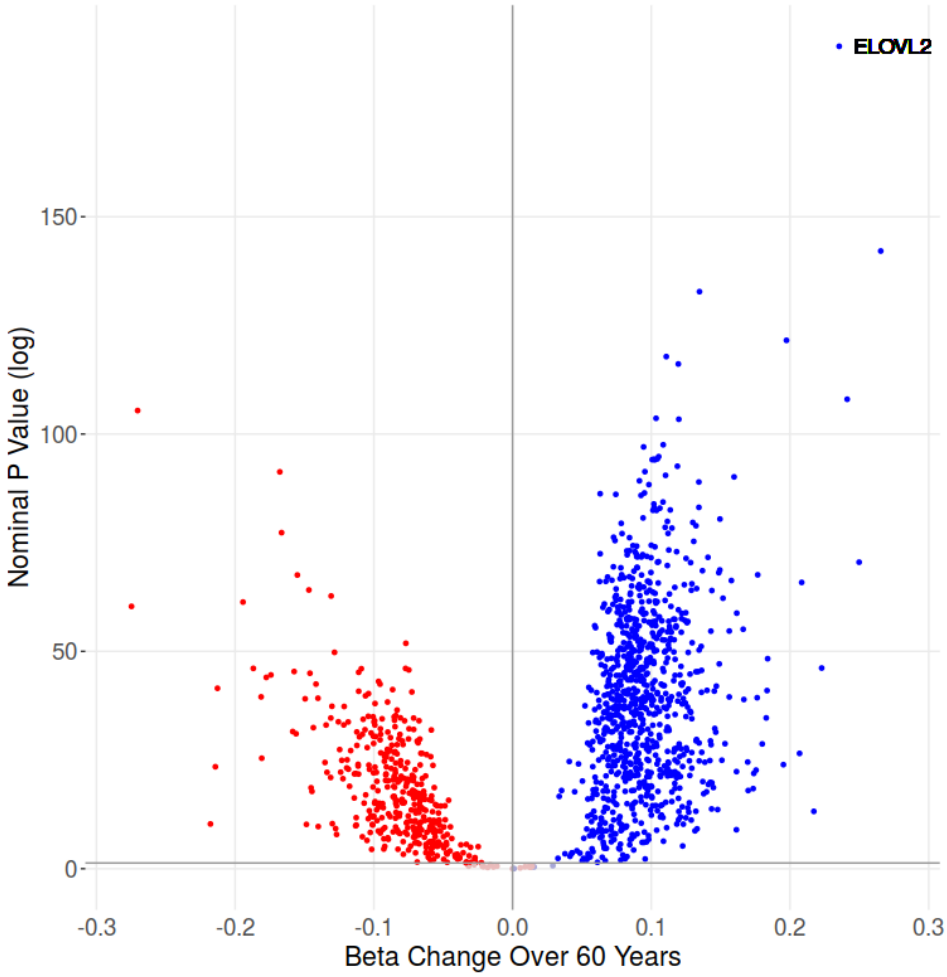


Figure 7. CMRs associated with age were well validated on the EPIC platform. Volcano plot of the DNA methylation (beta) changes with age, in the EPIC validation dataset, of the CMRs age effect size against p-values on negative log scale. Hits with increased (decreased) methylation are blue (red).

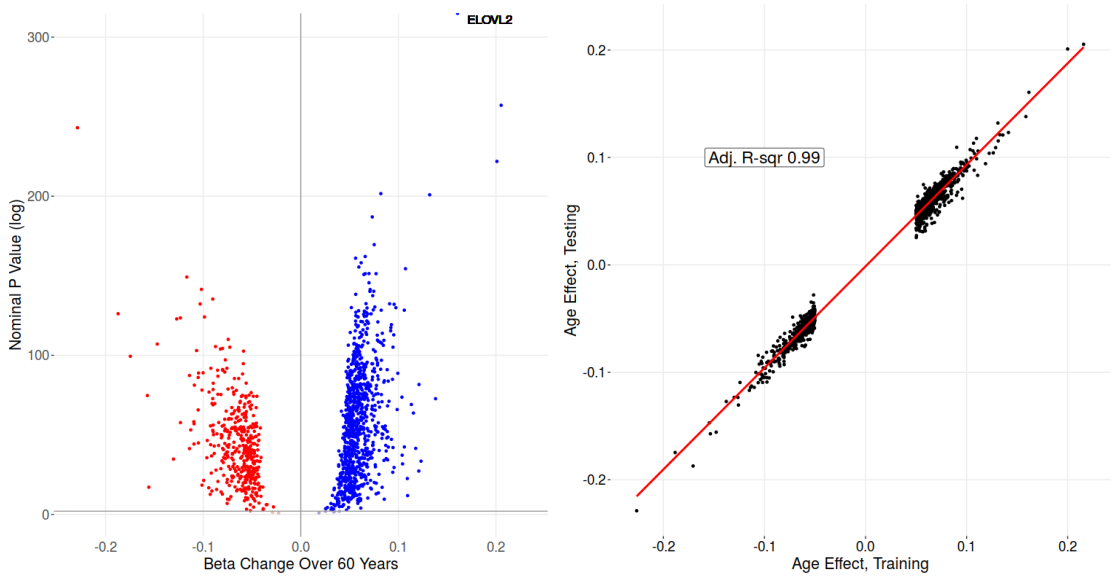


Figure 8. Validation of CMRs associated with chronological age.

Left: Volcano plot of the DNA methylation changes with age, in the validation dataset, of the CMRs age effect size against p-values on negative log scale. Higher (lower) methylated hits are blue (red). Right: Regression of age effect sizes in testing vs. training data for the significant CMRs.

To test for potential missing individual probe hits due to the CoMeBack data reduction, as well as the potential for additional discoveries from the CMR-EWAS, we considered standard EWAS of all the individual probes. Focusing on "large effect" probes for which the calculated change of composite DNAm measures over 60 years was at least five percent, as above, there were 3,769 probes significant in the training data, of which, 3,755 were validated in the testing data. Of these probes, 3,714 (99%) were within a significant CMR. In terms of CMR-only discoveries, 29 of the validated significant "large effect" CMRs had no individually significant "large effect" probes, with example shown in Figure 9.

Figure 9. Example of a validated significant CMR, where the individual probes were not significant in standard site by site EWAS with multiple-test correction.

We characterized the enrichment of whole blood CMRs associated with age versus the CMR background, using the ChromHMM genomic chromatin state regions in PBMCs as described in the Methods above, as depicted in Figure 10. Relative to all CMR background, the age-associated whole blood CMRs were significantly enriched for quiescent low-transcribed, weak Polycomb repressed, and weakly transcribed chromatin states.

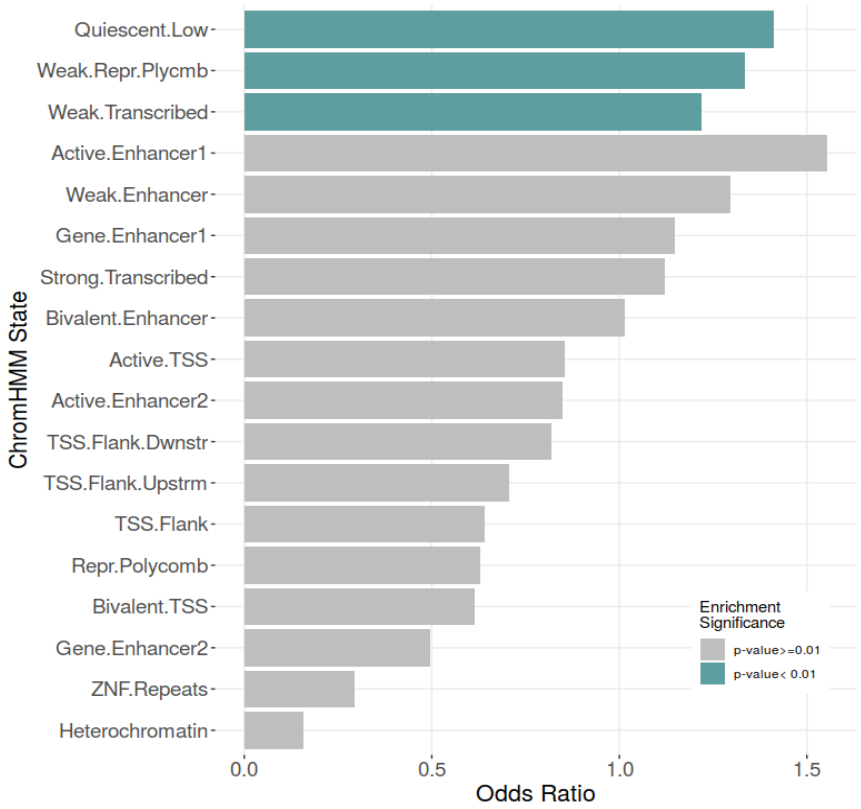


Figure 10. Enrichment of significant age-associated CMRs for genomic chromatin states.

2.3.4 PCA of Whole Blood CMRs

To determine whether the composite CMR methylation data reduction may be useful for unsupervised exploratory data analysis, we considered all phenotypic variables that were available in all the publicly available datasets, including the bioinformatically predicted cell type proportions. Figure 11 shows the principal components (PCs) of the composite CMR methylation in comparison to the PCs for the individual probes included in these CMRs. Trying to assess the potential ability of CMR data reduction to detect global methylation pattern associations, we examined whether the global patterns observed with the CMR PCs were

similar to the ones for the PCs of the individual probes included in the CMRs.

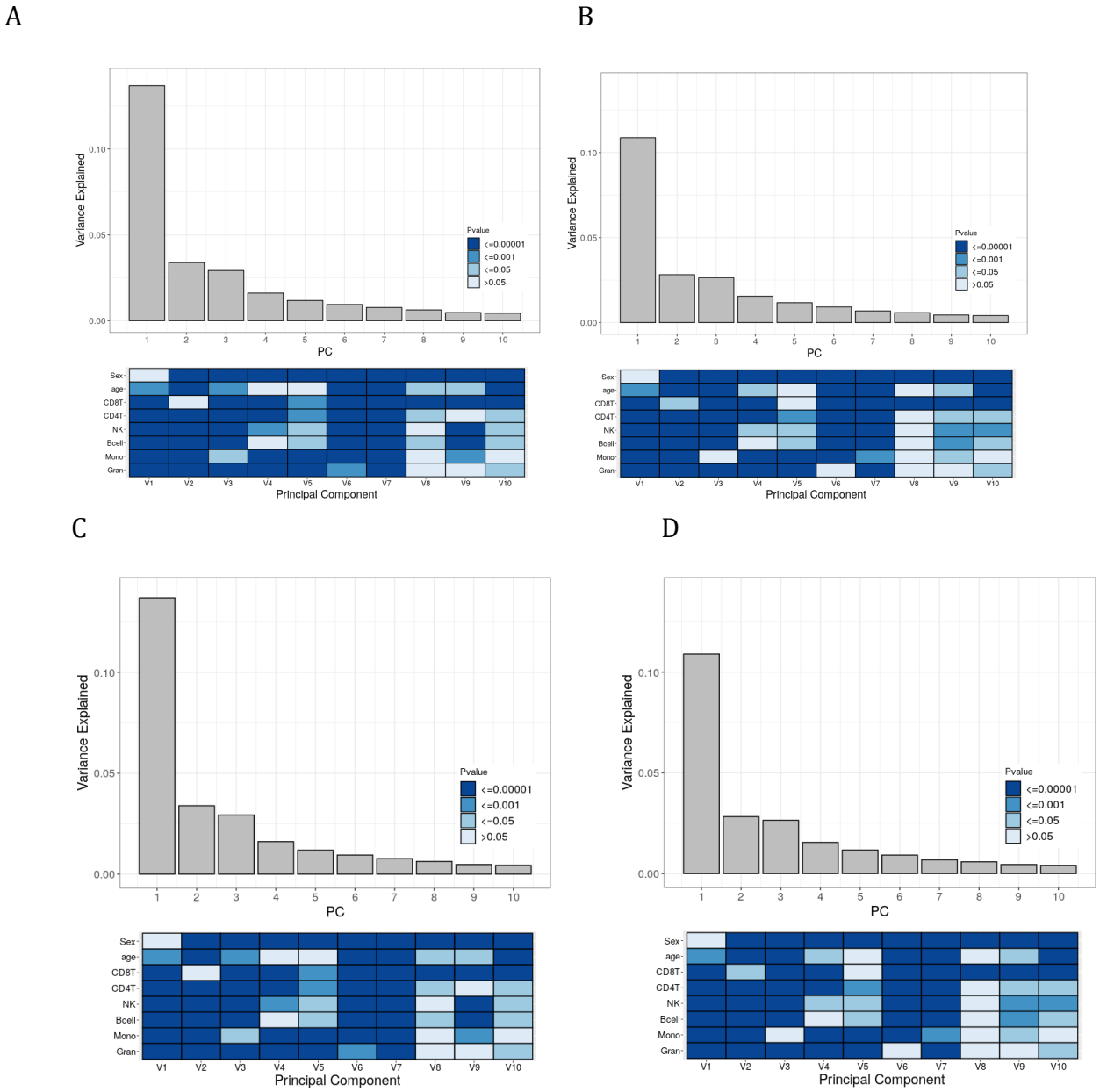


Figure 11. Associations for PCs of the CMRs were consistent with those of the individual probes' PCs.

PC analysis of whole blood CMRs data (A) versus PCs for the individual probes included in the CMRs (B). Regression of PC scores on estimated cell-type counts, age and sex showed that global CMR methylation patterns associated with these variables are consistent with the patterns present in the individual probes' PCs. (C): PCs for CMRs that have at least one variable probe, compared to PCs for these variable probes, shown in (D).

Considering all CMRs the first PC for the CMR data explained about 13% of the total variance, as seen in Figure 11, while the second PC added another 3%. In comparison, the first PCs for the individual probe data explained 11% of the variance, with the next PCs explaining comparably less variation than the corresponding CMR PCs, consistent with a lower total variability of the reduced CMR data.

For both the CMR and the individual probes' PCs, most of the top PCs' scores were associated with predicted cell type proportions, age and sex, and association patterns across the PCs were generally similar between CMRs and individual probes. For example, age and sex associations were weaker than the predicted cell type proportion associations in the first PC scores for both CMRs and probes, while the second PCs had weak association with specifically CD8T cell type proportions. Figure 11 highlights similar results when considering CMRs that contained variable probes.

2.3.5 Comparison with Existing Methods

2.3.5.1 Comparison with A-Clustering

Figure 12 depicts comparisons with the Aclust, which estimated 33,292 clusters, similar to the number of CMRs produced from CoMeBack. Figure 12(left) shows that the intersection of clusters identified by Aclust with the CMRs was only 16,958 common clusters (51%). The majority of these common clusters contain exactly two correlated probes. The difference in probe sets captured by the two

methods is even more pronounced for larger sets, with Aclust calling substantially more clusters with three or more probes than CoMeBack as shown in Figure 12(right).

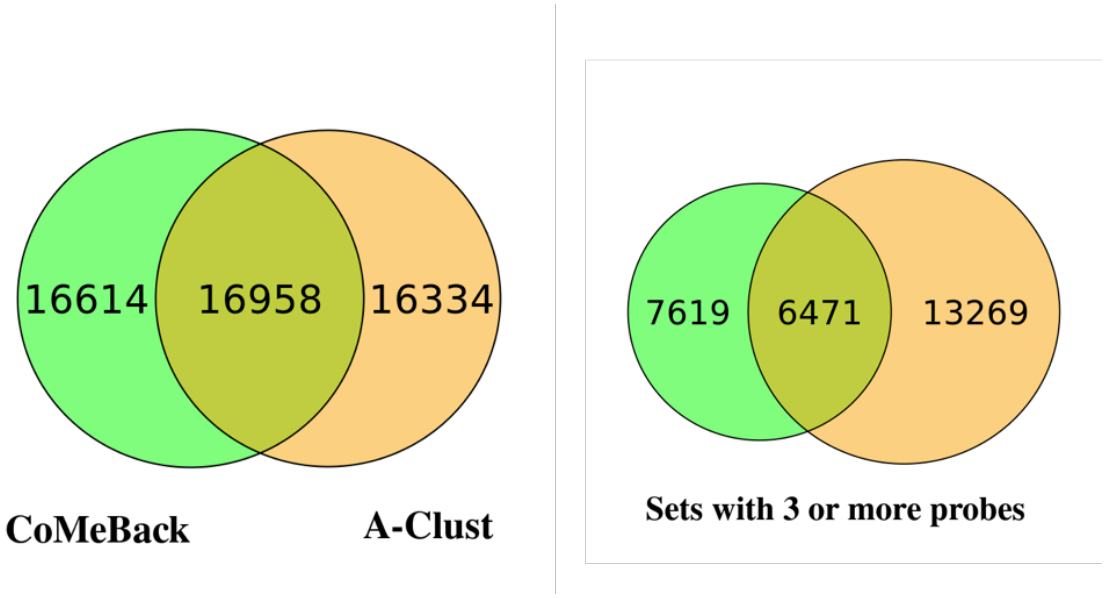


Figure 12. CoMeBack differs from A-Clustering.
Left: CoMeBack versus Aclust, all CMRs and clusters. Right: CMRs and A-clusters containing 3 or more probes.

We also considered the enrichment of the Aclust clusters for chromatin states, as we did for the CMRs above, depicted in Figure 13. Unlike the CMRs, the Aclust clusters were enriched for most chromatin states, including heterochromatin and quiescent low-transcribed states, suggesting that the probes composing the Aclust clusters may be more heterogeneous than those within CMRs, in terms of overlapping genomic regions in multiple chromatin states. To further probe this possibility, we examined how many Aclust clusters were entirely within a single

ChromHMM state, as opposed to overlapping multiple states. We found that only 64% of the Aclusters were in a single chromatin state, versus 81% of the CMRs.

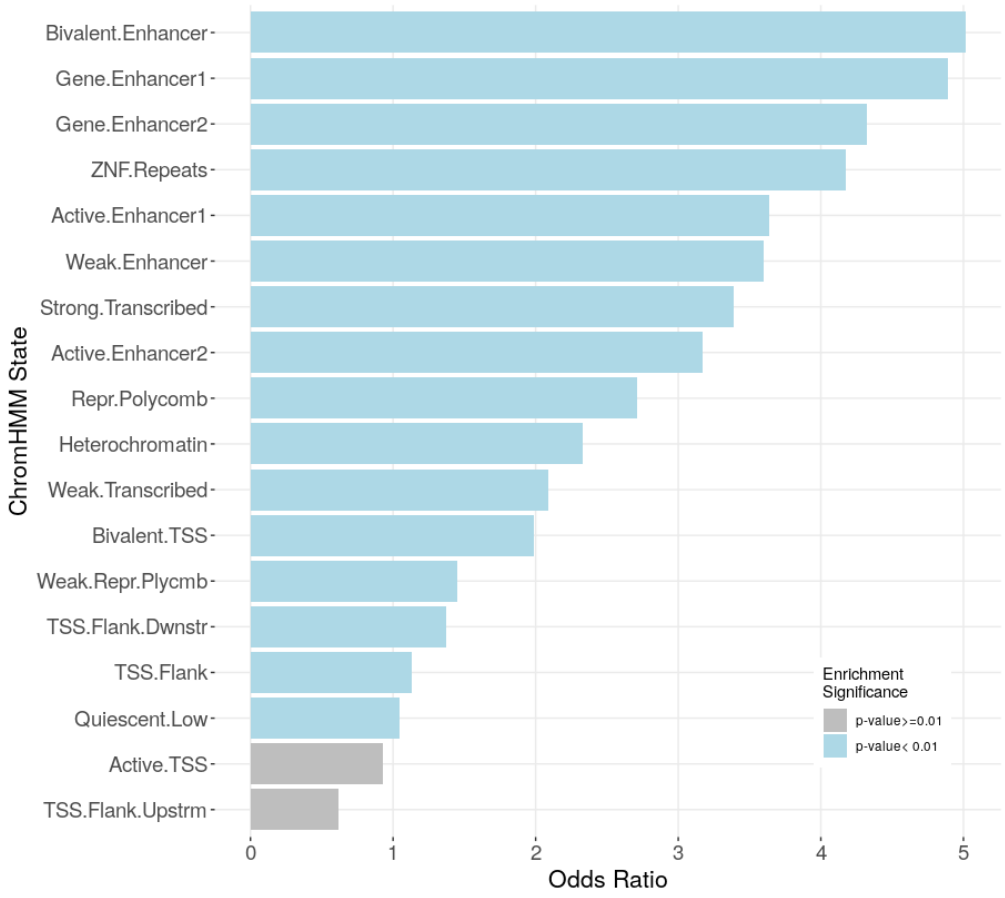


Figure 13. Aclust clusters were enriched for most chromatin states, including heterochromatin and quiescent low-transcribed states.

2.3.5.2 CMR EWAS Comparison with DMRcate

Since the CMR-based EWAS is a major application of CoMeBack, we also considered how it compared to existing tools designed specifically for differentially methylated region (DMR) detection. We compared the CMR-EWAS with a widely-adopted, recent, method developed for differential methylation

analysis, DMRcate [59]. Using chronological age as the variable of interest, DMRcate estimated 24,110 DMRs, in the training data, of which 14,291 DMRs (59%) contained at least one probe present in any CMR that was constructed in the training data. We considered how many of these DMRs had sets of probes corresponding exactly to an entire CMR, to find only 441 out of the 24,110 (2%) were among the 18,388 CMRs significant for age in the training data, as shown in Figure 14(left).

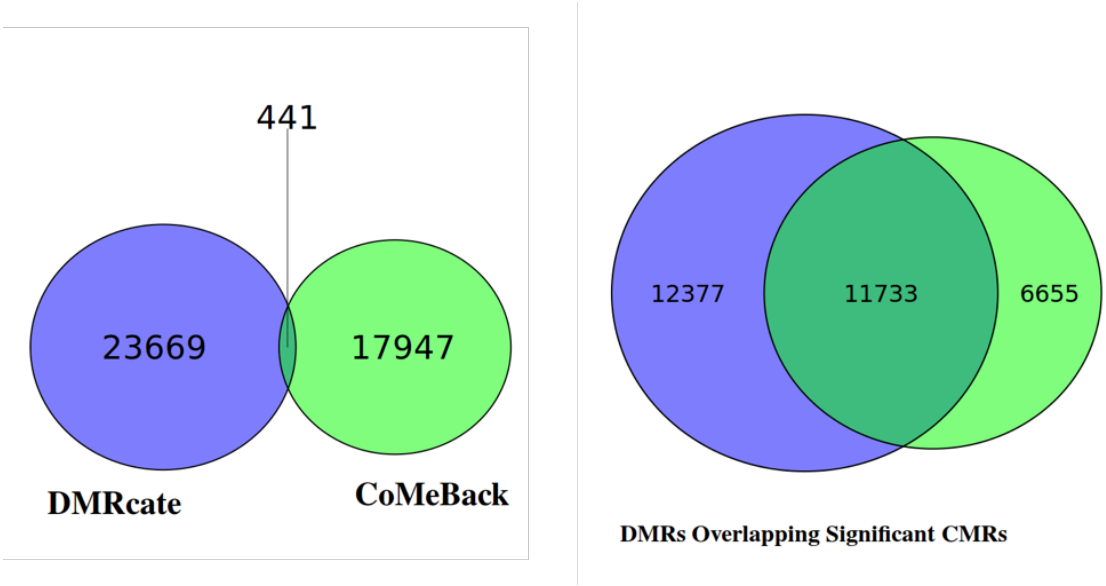


Figure 14. CoMeBack versus DMRcate.
Left: Age-associated DMRs that have the same probes as CMRs. Right: DMRs that overlap age-associated CMRs.

On the other hand, when considering partial overlaps of the DMRs probe sets with the CMRs, we found that 11,733 (49%) DMRs had overlaps, but were not identical, with the age-significant CMRs, as shown in Figure 14(right). Finally, of

the 1,332 significant CMRs with large effects, 1,289 (97%) were included as a probe subset within a DMR called by DMRcate.

2.4 Discussion

We envisioned our method as offering a biologically-driven analysis that reduces the dimensionality of the data, by constructing regions of correlated DNAm, while considering the locations of background genomic CpGs. This makes CoMeBack distinctive from existing, pure correlation-based methods, such as A-clustering [60], which conversely allows for the clustering of non-proximal correlated probes, or of proximal non-correlated probes, and does not account for background CpG density. CoMeBack may be viewed as a method for evaluating which adjacent CpG sites can be grouped together as a single unit, whose methylation state is related and potentially involved in transcriptional regulation. Our approach can enhance standard individual CpG EWAS by uncovering statistical associations with multiple probes with high specificity, while also offering as output genomic regions that may be easily interpretable in terms of biological function.

Our goal was to implement a useful, convenient tool for unsupervised data reduction that can be applied upstream of existing analytic pipelines, potentially lowering false positive rates, as multiple adjacent sites are associated to a phenotypic variable of interest. Our biologically motivated data pre-processing approach outputs CMRs for further downstream analyses and interpretation. The basis of the CMR construction was genomic CpG proximity, differing from existing methods that use array-probe proximity or solely correlations. The premise of our

approach was to anchor the correlated CMR probes into putative biological mechanisms acting to jointly methylate adjacent CpG sites, by using genomic background CpGs to define density and proximity. We hypothesized that this reduced the discovery of spurious correlations that may be present in the adjacent, yet sparse array probes.

To illustrate an analytic pipeline that builds on CoMeBack, we constructed reference CMRs in whole blood, followed by CMR characterization in terms of chromatin state and transcription factor binding sites. The reference CMRs estimated in whole blood, including those associated with chronological age, demonstrated that while there are fewer CMRs compared to non-CMR singleton probes, the CMRs were likely biologically significant. These CMRs were enriched for key regulatory elements including, several types of enhancers, mQTLs, and binding site motifs for transcription factors, some of which preferentially bind methylated CpGs [91]. The enrichment results suggest that CMRs with DNAm that varies across individuals are enriched in regulatory elements involved in transcription, which is consistent with prior findings from whole genome bisulfite sequencing experiments [61]. It is notable that several TFs whose binding affinity is affected by DNAm status of the motif's CpGs are "M-plus" [91], which is the label for TFs that have increased binding affinity to methylated motif CpGs. Such potential M-plus transcription factor binding suggests a mechanism for gene expression regulation that is different from the current working model of promoter CpG island methylation that suppresses the binding of M-minus transcription factors. The mQTL enrichment

suggests that genetic variability is an important driver of across-individual DNAm variability [96], and genetic drivers may affect multiple adjacent CpGs simultaneously.

Next, we note that in the case of chronological age CMR-EWAS, the results were broadly consistent with existing reports of life-long methylation gains in CpG islands, as well as with the observed higher proportion of sites with increasing methylation among the most highly replicated sites that are strongly associated with age [11]. The enrichment results for the age-associated CMRs for chromatin states were consistent with existing results showing loss of methylation in passive chromatin state regions with age in whole blood [11]. The comparison with DMRcate showed that DMRs called by DMRcate had somewhat lower reproducibility than the CMRs, while on the other hand, virtually all of the "large effect" CMRs were validated by DMRcate, in the sense that they were contained within one of the larger DMRs called by DMRcate. Finally, the PCA results were consistent with prior findings showing that cell-type composition and age are the major drivers of global DNAm patterns in whole blood [11].

CoMeBack is distinct amongst currently existing methods in its approach for DNAm data reduction, as CMR clustering is guided by the background genomic CpG density. To further differentiate CoMeBack from similar methods that involve correlation-based clustering, we briefly discuss below some key conceptual and methodological differences from the most similar alternative method, Aclust. While Aclust uses correlations of adjacent probes to output clusters, it has different

objectives and methodology, and perhaps most importantly, it produced substantially different output. Unlike CoMeBack, the clustering in Aclust is just one part of the Aclust DMR detection pipeline, where the Aclust clustering step aims to find correlated probes that are appropriate for the multivariate specification used in the Aclust downstream DMR estimation. The Aclust algorithm does not take into account the unmeasured genomic CpGs used by CoMeBack, nor the distance between the correlated probes, unless an extra pre-processing step (DBPM) is included. However, this step uses a fixed probe window to merge "all the sites (probes) wedged in between" two probes that are correlated, potentially resulting in non-correlated intermittent probes being force-merged within a cluster. As a result of these conceptual and implementation differences, the output of the two algorithms, when applied to the same data, was found to have substantial differences in terms of the actual clusters generated, including their characteristics, such as single chromatin state overlaps.

Considering downstream analyses, while the CMR-based EWAS is a major application of CoMeBack, our algorithm is different from tools like DMRcate [59], in that it is not guided by phenotypic variables, like environments, age or health outcomes, when constructing the CMRs. Rather, CMR EWAS proceeds after CMRs are constructed, with a concomitant reduction in the multiple-testing correction penalty. To further differentiate CoMeBack from DMRcate, we note that unlike DMRcate DMRs, CMR probes were required to have positive correlation and the

methylation levels across a CMR's probes can optionally be constrained to be similar to each other.

Overall, our CMR EWAS analysis illustrated how the use of CMRs for EWAS could enable the detection of additional sets of co-methylated CpGs that are correlated with a phenotype of interest. Given the ease of use of the CoMeBack software, we believe that CMR-based EWAS can usefully supplement standard single probe EWAS. While the CMR output may enhance EWAS discovery, validation and interpretation, it may also be used in other non-EWAS downstream analyses, like PCA, or networks of the constructed CMRs. We also note that singleton probes that are not included in CMRs can be easily identified and used in downstream analyses, in conjunction with the CMRs. For downstream analysis that focuses on CpGs with variable DNAm across individuals, a user might choose to utilize all probes measured on the array, even those that do not meet the thresholds required for inclusion into a CMR.

We also illustrated a CoMeBack application for unsupervised exploratory analysis, using the composite CMR DNAm values for PCA. With respect to predicted cell-type proportions, age and sex, the resulting global DNA methylation patterns typically observed when analyzed in a probe specific way, were retained when evaluated using CMRs. This concordance indicated that this type of exploratory analysis can also be informative in the case of other phenotypic variables of interest, with regressions of CMR PC scores used to assess the associations of phenotype with global DNAm patterns captured by the top PCs.

Overall, we hope that the applications of CoMeBack illustrated in this paper demonstrate its usefulness for enhancing different analytic pipelines for DNAm array data. The user-friendly CoMeBack software is freely available as an open-source R-package.

Chapter 3: CoMeBack Application for Characterization of Autosomal Co-methylated Regions associated with Sex

3.1 Background

Biological sex is defined by the genetic complement of sex chromosomes (X and Y chromosomes); human males have a 46,XY chromosome complement while females possess a 46,XX. Beyond genetic differences, males and females also differ at the anatomical, physiological, and molecular levels [12-20, 97-99]. Several studies using DNA methylation (DNAm) arrays have documented associations between sex and methylation at thousands of CpG sites, as well as larger differentially methylated regions (DMRs) on the autosomes, with some of these including imprinted loci, perhaps not surprisingly given the underpinning biology [21-23]. Moreover, in Epigenome-Wide Association Studies (EWAS) using DNAm array data [24, 25, 33, 34] , it is now a standard practice to adjust for sample sex and age covariates when investigating associations with a phenotype or a disease. Yet, our understanding of the role of DNAm associations with sex remains limited, both in terms of functional significance and how they are established.

The reasons for this incomplete characterization include limitations from the tissue-specificity of DNAm, as well as its plasticity in response to environmental factors, which are difficult to control for in typical DNAm study designs [100-102]. Thus, DNAm studies within a tissue and a specific condition typically have modest sample sizes, low statistical power and specificity, and their findings cannot be generalized to other tissues or diseases. On the other hand, DNAm biology involves prevalent

co-methylated genomic regions that function as biological units [68-71, 103], such as CpG islands in gene promoters [66] and imprinted regions [104]. Hence, studying such co-methylated regions would enhance our current understanding of the role of DNAm in sexual dimorphism.

While most previous studies on the relationship between sex and DNAm in humans [12-23] were limited in scope or statistical power, their focus has been mainly on associations of single CpG loci, as measured by individual probes on DNAm arrays. Few recent studies [21, 105] have estimated a small number of sex DMRs by using the sex-phenotype information to aggregate adjacent individually significant probes and impute DMRs. There is a need for further characterization of such sex-associated co-methylated regions, as that would facilitate the functional interpretation of DNAm associations with sex.

The developmental establishment of sex differences in DNAm is of general interest in the field of Developmental Origins of Health and Disease (DOHaD) [106-108], and several recent studies have focused on DNAm during the puberty transition [109-112]. These studies have focused on individual CpGs measured on high-throughput DNAm arrays, either across the genome, or in candidate regions related to sex hormones. Further characterization of changes over time in the DNAm status of co-methylated genomic regions, across all autosomes, may help to elucidate the potential functional role of DNAm during early life development in general and during the puberty transition in particular.

3.2 Results

3.2.1 Discovery and validation of sex-associated CMRs in whole blood

To identify genomic regions that show sexually dimorphic DNAm, we generated an aggregate discovery cohort of 3,795 normative adult whole blood (WB) samples by combining Infinium HumanMethylation450 BeadChip array data from publicly available datasets (GSE55763, GSE84727, GSE80417, GSE111629, GSE72680) [73, 79-81]. In total, 2,414 males and 1,381 females 25-80 years of age and of diverse genetic backgrounds were merged followed by cell-type correction using the Houseman method [28] (Table 1).

Using the CoMeBack algorithm to identify co-methylated regions (CMRs) [113], autosomal DNAm sites were grouped based on correlation and CpG background density yielded 34,568 WB CMRs in the aggregate discovery cohort. Of these, 337 genomic CMRs, a total of 1285 probes, were sexually dimorphic (denoted sCMRs). More specifically, these sCMRs showed significant differences in DNAm between males and females with false discovery rate [35] (FDR)<0.05 and absolute composite CMR beta [113] difference>4%, and were composed of probes that all showed the same direction of change between males and females.

Analysis	GEO Dataset	Tissue	N	Females	Males	Age Range	Reference
	GSE55763	Whole blood	2,669	860	1809	31-75	Lehne et al., 2015
CMR Construction,	GSE84727	Whole blood	377	92	285	25-66	Hannon et al, 2016
sCMR Discovery,	GSE80417	Whole blood	224	117	107	26-79	Hannon et al, 2016
Sex Predictor	GSE111629	Whole blood	175	68	107	26-55	Horvath et al, 2015
Construction	GSE72680	Whole blood	350	244	106	26-77	Grady Trauma Project
sCMR Validation	GSE125105	Whole blood	697	388	309	17-87	Arloth & Binder 2019
	GSE79100	Kidney	31	16	15		Winterpatch & Lukassen 2018
	GSE80261	Buccal	96	57	39		Portales-Casamar et al 2016
sCMR Tissue	GSE61258	Liver	79	34	45		Horvath et al 2014
Concordance	GSE64509	Brain	25-41	17-22	8-11		Horvath et al 2015
	GSE87640	Immune cells	20	8	12		Ventham et al. 2016
Sex Predictor Validation	GSE132203	Whole blood	794	571	223	18-76	Grady Trauma Project

Table 1. Public datasets used for discovery and validation of sCMRs.

The 337 sCMRs were validated in a separate cohort (GSE125105) of 312 males and 387 females aged 17-87 that was processed independently as described in the Methods section. After quality control (see Methods), 1191 out of 1285 sCMR probes (92%) remained in the validation dataset. To retain as many sCMRs as possible for validation, we chose to include, sCMRs represented by at least one probe in the validation data, resulting in 334 sCMR. Of these, 305 sCMRs had at least one significant (nominal $p<0.05$) probe with the remaining CMR probes showing the same direction of sex-biased DNAm between males and females (Figure 1). These 305 validated sCMRs ranged in size from 2-15 probes (7-2609 bps), encompassed a total 1,174 probes, and were detected across all autosomes (Suppl. File 1). In total, 235 sCMRs had higher DNAm levels in females compared to males, while 70 had higher levels in males compared to females.

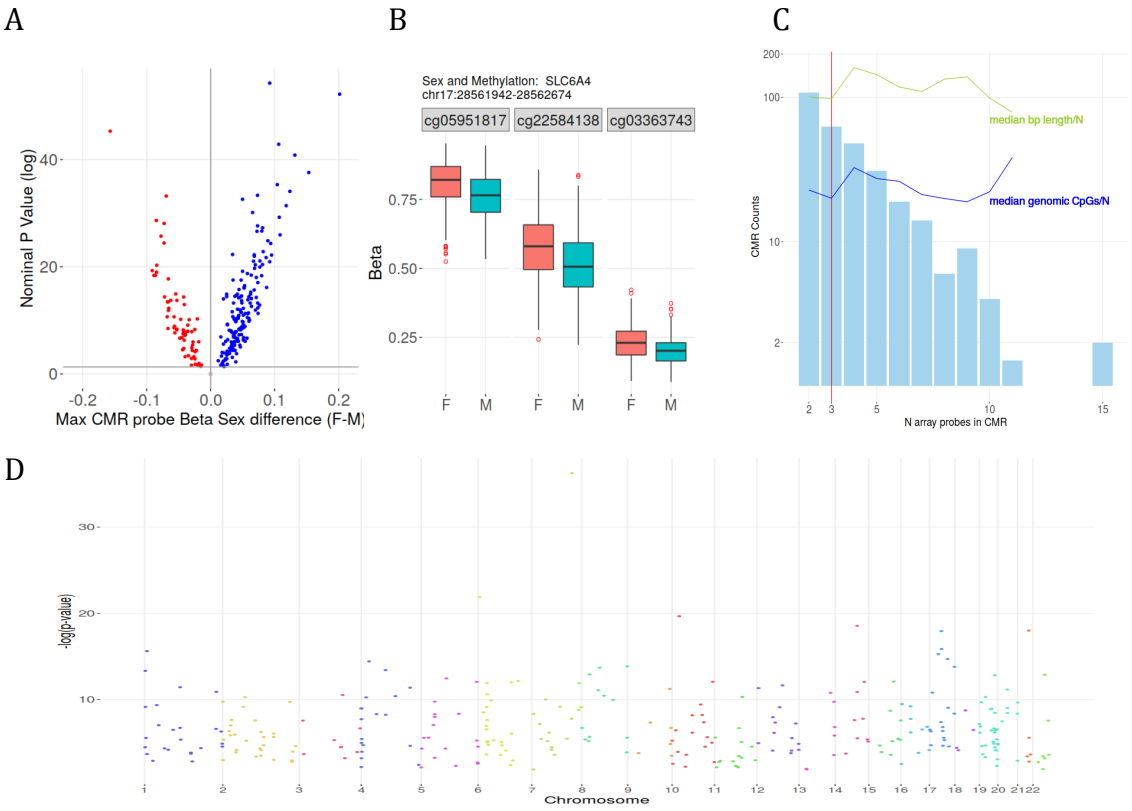


Figure 15. The validated sex-associated CMRs were present in all autosomes.

A. Volcano plot of the validated sCMRs, with FDR of 5%. B. An example of a validated top hit, SLC6A4. C: Number of N-probes sCMRs with median bp length and number of CpGs; red line is median of 3 probes. D: The autosomes are plotted proportional to their length, with sCMRs mapped to genomic coordinate position. Chromosomes 1,2, and 19 had the highest absolute number of sCMRs, while chromosomes 17, 19 and 22 had the highest density of sCMRs, defined as number of sCMRs divided by chromosome length.

3.2.1.1 sCMRs were not Enriched for GO and KEGG terms

Closer inspection of the 305 validated robust sCMRs revealed that they were associated with 167 genes (see Methods). Focusing on genes overlapping sCMRs revealed that they were not enriched for any particular GO term or KEGG pathway. Nevertheless, careful inspection of the list of sCMR associated-genes revealed genes involved in sex biology, sex-linked phenotypes, or steroid hormone biology. For example, sCMRs were observed overlapping the estrogen receptor gene (*ESR1*), as

well as the Cytochrome P450 1B1 (*CYP1B1*) gene, which is involved in the metabolism of hormones [114]. Furthermore, a sCMR was also observed overlapping the *SLC6A4* gene (Figure 1,B), which encodes a serotonin transporter that has been implicated in a range of mental health conditions [115-118] including depression and social anxiety. The sCMR overlapping the *SLC6A4* gene had two probes with more than 5% difference in median probe methylation between males and females, and the maximum absolute difference was 7.4%.

3.2.2 Characterization of sCMRs

3.2.2.1 Enrichments of genomic features with relevance to Sex

3.2.2.1.1 sCMRs were enriched for several chromatin states

Having identified and validated 305 sCMRs in WB, we then sought to characterize them further. Using the ChromHMM algorithm [62], which defines chromatin states, we observed that sCMRs were significantly enriched in ZNF repeats, polycomb repressive elements, heterochromatin, bivalent enhancers, and transcription start site (TSS) characteristics (TSS flanking, active TSS, TSS flanking downstream, and bivalent TSS) compared to the entire 34,568 CMR background (Figure 16 A).
Upon dividing the set of all 305 validated sCMRs based on whether they had higher DNAm levels in females compared to males, or vice versa, different chromatin states were found to be enriched in the two subsets (Figure 16 B). For sCMRs with higher levels in females, the highest enrichments were observed in heterochromatin, ZNF repeats and Polycomb-repressed states, while for sCMRs with higher levels in males,

the most enriched states were TSS upstream and downstream and Type 2 Active Enhancers [62].

3.2.2.1.2 sCMRs were enriched for a few TFBS motifs

To investigate the potential role of sCMRs in regulation of gene expression, we examined the enrichment of sCMRs for known transcription factor binding site (TFBS) motifs. Out of the 404 motifs determined by HOCOMOCO, a comprehensive database of transcription factors [89], 98 were found in sCMRs probe-spanning regions (see methods), and 9 motifs were enriched in the sCMRs (Figure 16,C). Of note, the motif with highest enrichment was for SPI1, which is not shown in Figure 2,C due to a much higher odds ratio compared to the other elements.

Several of the enriched TFs are known to function in blood cells, including the KLF and SP family of transcription factors [119]. Moreover, 3 TFs (KLF6, KLF12 and KLF15) show increased binding to methylated DNA (see [91]), while 1 TF (SP3) prefers to bind unmethylated DNA.

We considered sCMRs that had higher median DNAm in females compared to males and vice versa (see above). Performing a TFBS motif enrichment analysis separately for these two sets of sex-biased sCMRs (Figure 16 D), we found 7 motifs enriched for at least one of the sexes. One M-plus TFBS motif, KLF15 was enriched for sCMRs with higher DNAm in females, while two M-plus TFs, KLF6 and SP1, were enriched among the sCMRs with higher DNAm in males. Several TFBS motifs with unknown

methylation preference were enriched differently in sCMRs with higher median DMAm in males versus females.

3.2.2.1.3 sCMRs were enriched for sex hormone-related CpGs

To investigate the potential functional relevance of the sCMRs in sexual maturation, we considered whether sCMRs were enriched for CpGs with known DNAm associations to changes in reproductive hormones over the puberty transition in boys. We found that the sCMRs were significantly enriched, at 5% level, for CpGs associated with changes in DNAm over the puberty transition in boys for all five sex hormones considered (Figure 16 E). Specifically, for three out of the five reproductive hormones considered (Inhibin B, anti-Müllerian hormone and Testosterone), the enrichment was significant at the more stringent 1% level. The sCMRs were also enriched for probes in the entire set of CpGs reported in to be associated with any of the five hormones [109].

Considering the direction of sex-biased DNAm in sCMRs (Figure 16 F), we found that CpGs associated with sex hormone changes in boys were enriched in sCMRs with higher median methylation in males. On the other hand, CpGs associated with changes in follicle-stimulating hormone in boys were not significantly enriched in sCMRs that had higher median female methylation, as one might expect from a hormone integral to female ovulation.

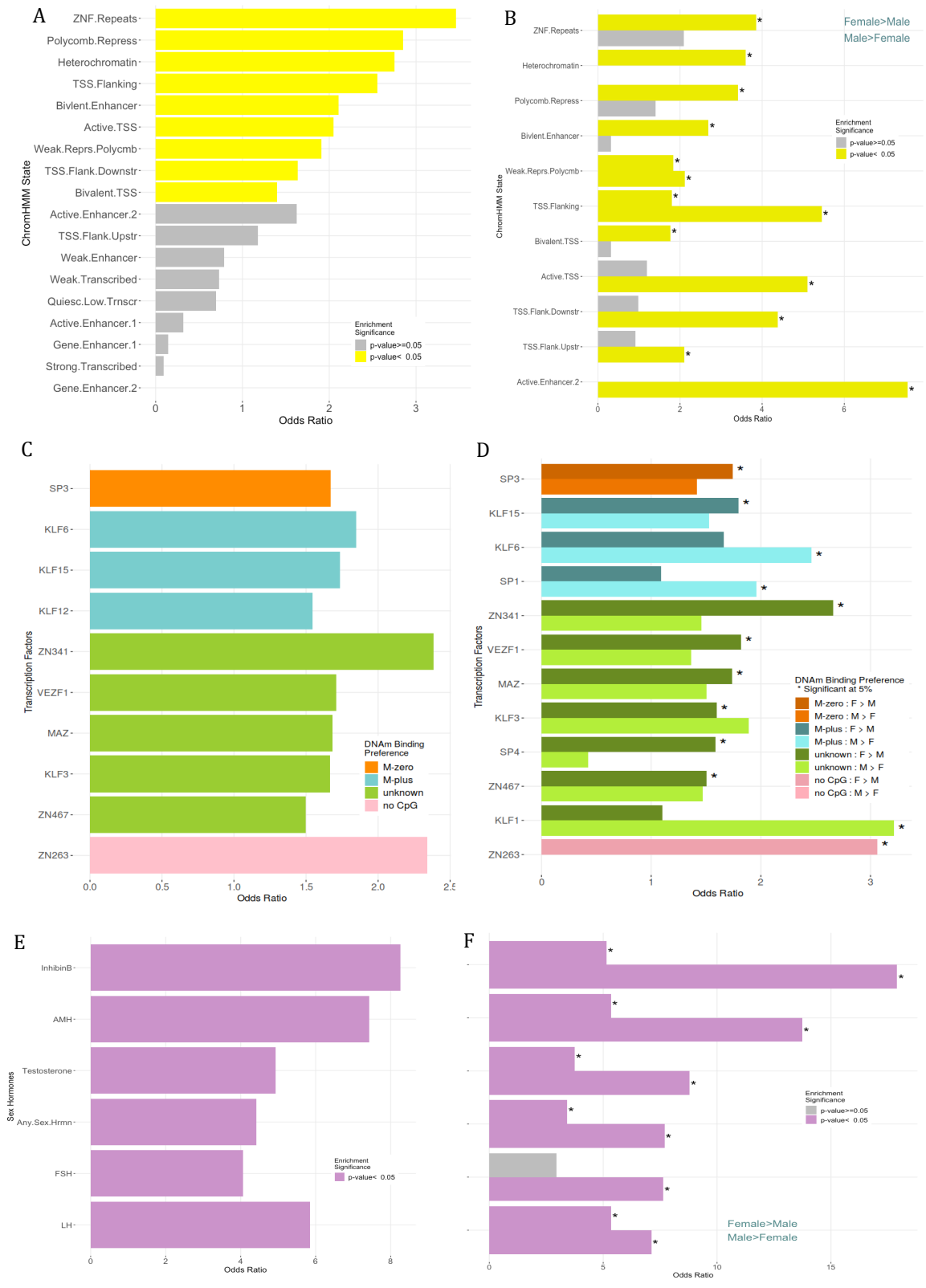


Figure 16. The sCMRs were enriched for chromatin states and genomic features.

A, B. Enrichment in chromatin states related to the transcription start site and weak Polycomb repressed. C, D. Enrichment for several transcription factor binding site (TFBS) motifs, including several involved in blood physiology, as well as TFs whose binding affinity depends on DNAm status. Transcription factors SPI1 (not shown) had odds ratio above 30. E,F. Enrichment for CpGs associated with sex hormone changes in boys over the puberty transition.

3.2.2.1.4 sCMRs were not enriched for imprinted genes or for imprinting control centers

Genomic imprinting is the DNAm-mediated process of monoallelic gene silencing depending on the parent-of-origin [104]. Although DMRs associated with imprinted genes have been widely reported, studies have also shown that DNAm levels at imprinted loci can often vary by offspring sex [120-124]. This led us to investigate whether the 305 validated sCMRs overlapped with CpGs from imprinted genes that exhibit sex-specific DNAm differences. We identified that only a single sCMR contained probes associated with a known maternally imprinted gene, *NLRP2.* We next sought to identify whether the validated sCMRs were also enriched for the 45 imprinting control centers (ICR) [125, 126]. Interestingly, only 4 sCMRs overlapped with an ICR, and there was no significant enrichment against the CMR background.

3.2.2.1.5 sCMRs were enriched for lncRNAs

We observed that among the most significant large effect sCMRs there were several long non-coding RNAs (lncRNAs), which seemed intriguing in light of recent evidence linking lncRNA to DNAm [127-129] and to sex maturation in model organisms [130, 131]. Therefore, we examined whether the 305 validated sCMRs overlapped the 9,066 lncRNAs that have been explored in relation to DNAm changes

in cancer [132]. Overall, we found a total of 69 sCMRs (23%) overlapped lncRNAs regions, revealing a significant enrichment for lncRNAs in sCMRs.

3.2.2.1.6 sCMRs were enriched for methylation quantitative trait loci

Considering the influence of genetic variation on DNAm patterns within a tissue, we anticipated that the majority of the sCMRs will include a known methylation quantitative trait locus (mQTL) (Figure 17). Using the most conservative set of mQTLs that included probes from both the ARIES [92] and the McRae [133] studies, we found that 51% of the sCMRs contained a probe that was among the 29,130 probes identified in both studies. Comparing these numbers against the WB CMRs background revealed a significant enrichment for mQTL within sCMRs.

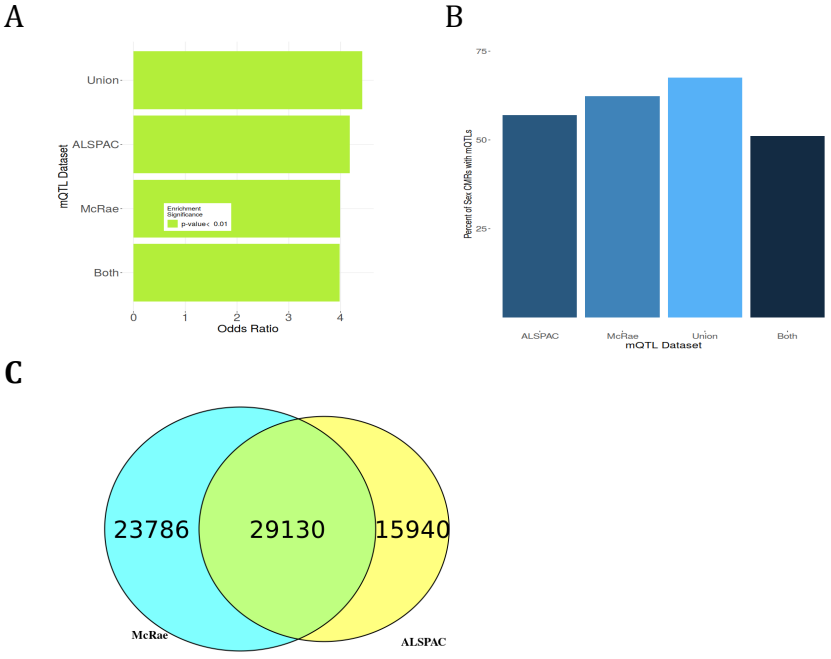


Figure 17. mQTLS were enriched in the sCMRs.

A: Enrichment. B. Number of WB mQTLs in the two different studies. C. Previously reported mQTLs from the two studies.

3.2.2.1.7 sCMRs were not enriched for genes differentially expressed between sexes

A recent study [134] identified sex-specific differences in gene expression in WB and PBMC samples, allowing us to determine if these overlapped sCMRs. Interestingly, three sCMRs overlapped with two genes differentially expressed between the sexes: *PER3* and *NLRP2,* the latter being the same gene which was previously identified as the only imprinted autosomal gene associated with an sCMR probe (see section 3.2.2.1.4 above).

3.2.3 Developmental establishment of sCMRs in whole blood

Given that sCMRs were identified in adults (25-80 years), we next investigated if sex-specific DNAm difference at these CpG sites were also observed in younger samples. Using the ARIES cohort [108] which includes 484 males and 487 females sampled at age 0, 7 and 15 years, (see Methods), we observed that most of the sCMRs showed sex-specific differences in DNAm from birth (Figure 18 A). Nevertheless, 10 sCMRs contained probes that were significant only at the later time points (7 and 15) suggesting that a handful of sCMRs may develop during this time period (Figure 18 B).

A

B

Figure 18. Several sCMRs were established over the period 7-15 years.

A. Common sCMRs containing at least one significant probe at different ages. B. Example, one of the10 sCMRs that were significant at 7 and 15 years, but not at 0 years.

3.2.4 WB sCMR concordance across different tissues and cancer

3.2.4.1 sCMRs had notable concordance across tissues

DNAm patterns vary substantially by tissue and cell types [1, 4, 135]. Using publicly available GEO datasets, we examined DNAm profiles of the 305 validated sex-specific WB sCMRs across multiple somatic tissues and immune cell types including buccal, kidney, liver, brain, monocytes, CD4 and CD8 T cells. (Figure 19). Overall, 32-75% of the 305 validated sCMRs were also determined to be sex-specific sCMRs in the investigated tissues based on a 5% nominal significance threshold. In

particular, over 75% of the WB sCMR were present in buccal tissue, which is among the tissues most commonly assayed for DNAm state.

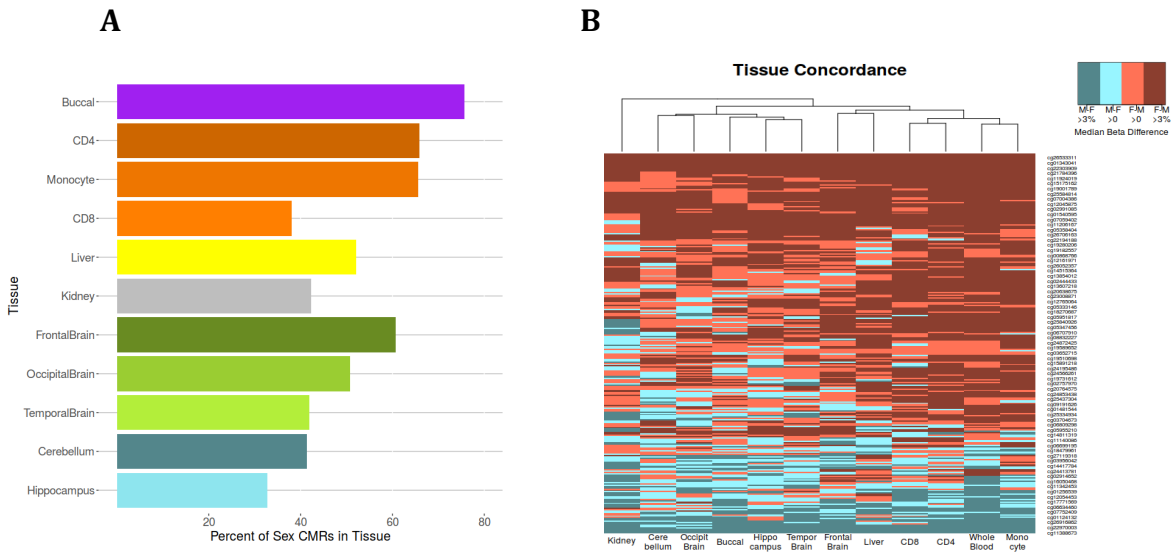


Figure 19. Whole blood sCMRs were concordant across multiple tissues.

A: Percentage of the 305 validated sCMRs that contained a CpG that was significantly associated with sex at the 5% significance level. B: concordance of sCMR median beta value difference across tissues. The median beta differences were calculated per sCMR, as the median female beta minus the median male beta and are shown here for the sCMR probe that was most significantly differentially methylated in whole blood.

The sCMRs beta differences between males and females were notably consistent in direction and magnitude of sex-biased DNAm across tissues. Moreover, 17 sex-specific WB sCMRs, containing 85 CpG probes were significant in all the investigated tissues and immune cell types, despite small sample sizes (Figure 20). Although not

significantly enriched for any gene ontology terms, some of the 17 sCMRs overlap with known genes, including *MYF5*, *SOD3*, *OOEP*, *DDX43*, *SOGA3*, and *PXDNL*.

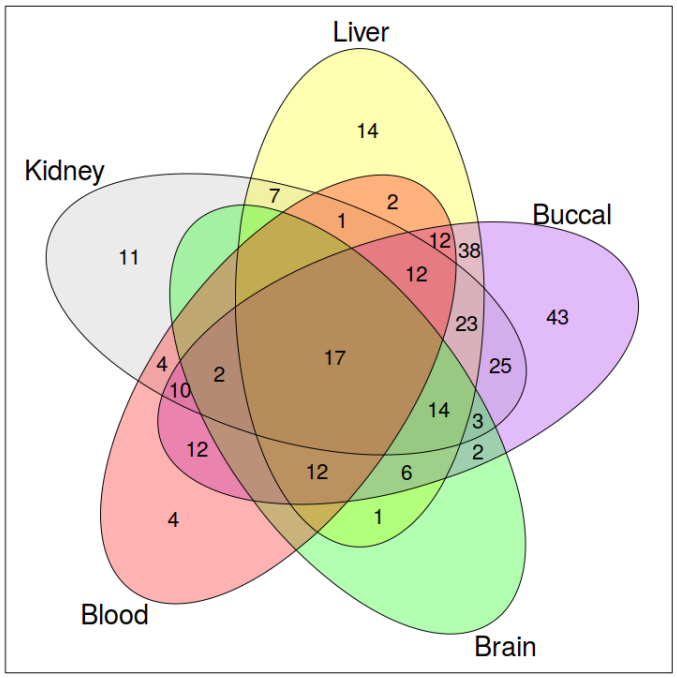


Figure 20. Several WB sCMRs were significantly associated with sex across different tissues despite small sample sizes for most tissues.
Brain sCMRS shown were common to all Brain regions. Blood CMRs shown were common across all blood cell types examined.

3.2.4.2 sCMRs could separate primary breast cancer samples based on estrogen and progesterone status

In an effort to determine the persistence of sCMRs to disease status, we turned our attention to one of the most extreme cases of DNAm dysregulation: cancer. Taking advantage of publicly available datasets from The Cancer Genome Atlas (TCGA) Research Network across a variety of cancer types (see methods), we found that most sCMR probes showed disrupted patterns of sex-specific DNAm pattern

compared to normative blood samples (Figure 21 A). Although these differences suggest that disease may dysregulate sex-specific effects on DNAm, one caveat of this analysis is that the cancer samples were derived from a variety of tissue types, many of which were not included in our across-tissue comparisons, due to lack of publicly available data for healthy subjects. Nevertheless, given the link between sCMR probes and genes related to steroid hormone biology, we found that they could clearly separate primary breast cancer samples based on estrogen and progesterone status (Figure 21 B).

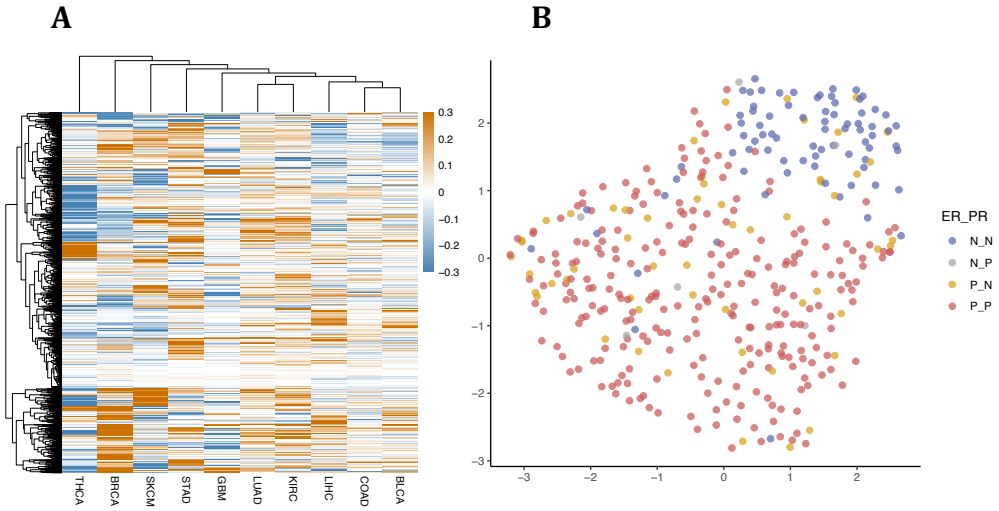


Figure 21. sCMR probes differentiated breast cancer samples based on estrogen and progesterone status.

A. Lack of sCMR methylation concordance across difference cancers. B. sCMR probes differentiated breast cancer samples based on estrogen and progesterone status. N/P are negative/positive for Estrogen Receptor (ER)/Progesterone Receptor (PR) respectively.

3.2.5 An Autosomal Predictor of Sex with Good Performance

We used the 1174 autosomal CpG probes from the 305 validated sCMRs to construct a sex predictor using a machine learning method, elastic net regression [53]. The predictor had 87 probes in 71 sCMRs. The predictor's performance was tested in an

independent Illumina Infinium Human Methylation EPIC beadchip microarray dataset (Figure 22). The predictor performed well, with an area under the receiver operating curve (AUC) of 99.8%, and overall accuracy of 99%.

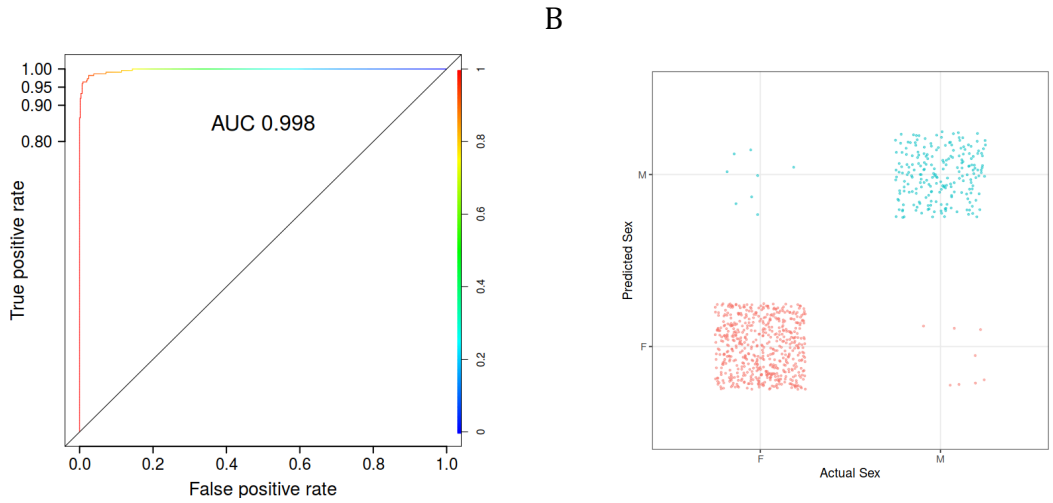


Figure 22. The sex predictor achieved high accuracy in an independent EPIC dataset.
A. Performance was assessed using area under the receiver operating curve (AUC).
B. Confusion table showing predictor accuracy for Males versus Females.

Next, to explore pan-tissue and pan-array technology applicability, with possible utility for other DNAm assays, such as pyrosequencing, we considered the feasibility of a sparse predictor that uses only a very small set of probes as inputs. We constructed and investigated the performance of two different sparse predictors. First, we constructed a "minimal" predictor by selecting among the same set of 1174 probes from all validated sCMRs but imposed a stronger regularization for the elastic net algorithm (see Methods). Next, a "pan-tissue-sCMR (PTS)" predictor was constructed from CpGs only from the 17 sCMRs that were found to contain a significant probe across all tissues investigated in this study (Figure 23).

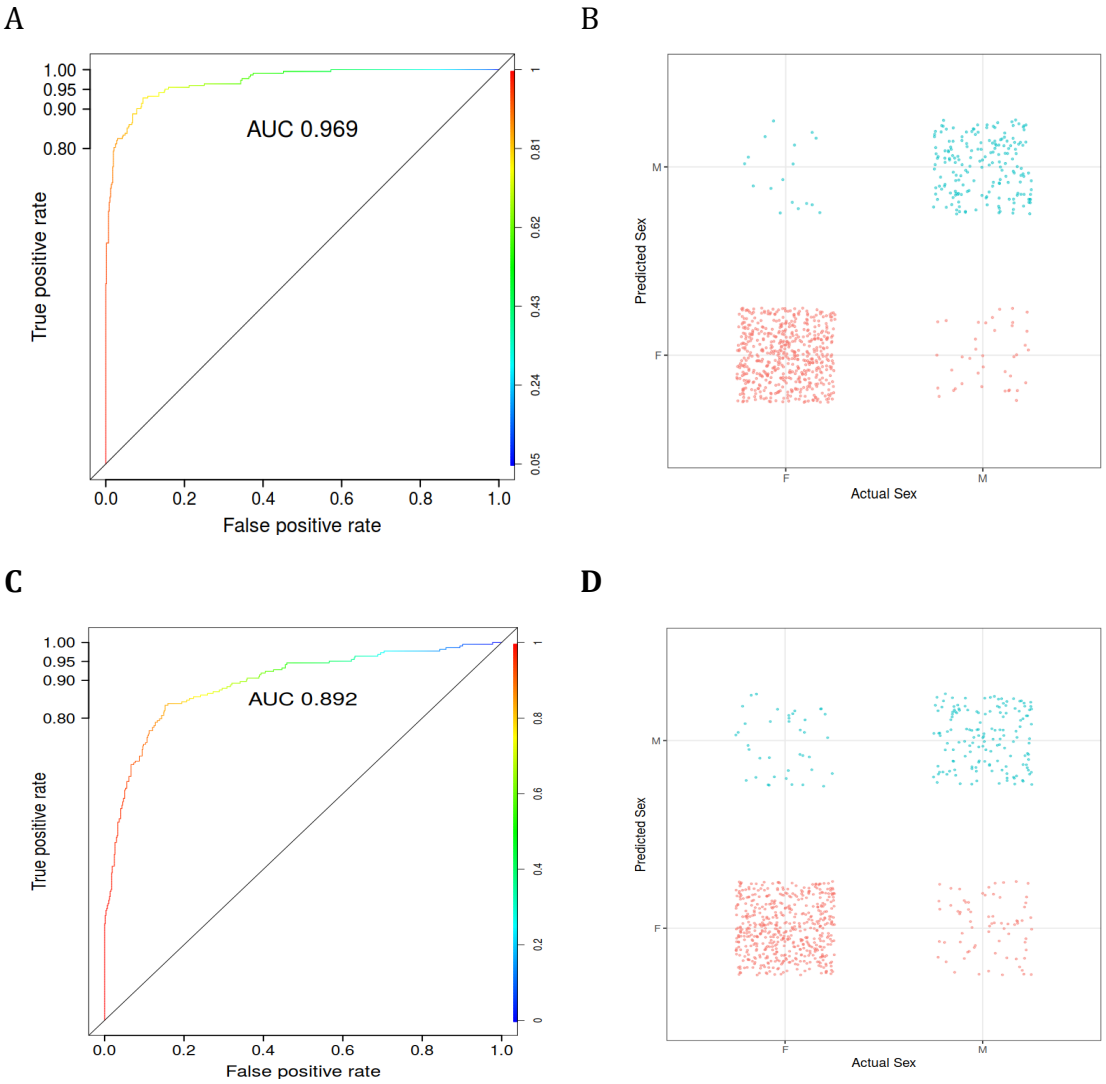


Figure 23. The sparse sex predictors achieved good accuracy in independent data. A,B Minimal Predictor. C,D, PTS predictor. Rigth: Performance was assessed using area under the receiver operating curve (AUC). Left: Confusion tables showing predictor accuracy for males versus females.

The minimal predictor (Figure 23 A, B) relied on 21 CpG probes and its performance was still adequate, with an area under the receiver operating curve (AUC) of 96% and overall accuracy of 90% in the testing EPIC data.

The PTS predictor (Figure 23 C,D) used 75 probes as input and still had reasonable performance, with an area under the receiver operating curve (AUC) of 92% and overall accuracy of 89% in the testing DNAm data.

3.3 Discussion

We aimed to characterize co-methylated autosomal genomic regions spanning multiple DNAm array probes with sex-specific beta methylation ratios. We used pre-defined reference WB CMRs [113] and interrogated their association with sex in adults, and at ages 0, 7 and 15. It is reasonable to focus on genomic regions spanning multiple CpGs as opposed to individual CpGs because groups of proximal CpGs, as captured by the CMRs, have been shown to frequently function as biological units [68-71, 103], for example CpG islands in gene promoters [66]. Additionally, previous studies have already characterized individual sex-associated CpGs, so we sought to enhance the characterization of sex-associated DNAm changes by focusing on genomic regions. Our choice of methodology was motivated by previous results showing that using CMRs for EWAS can improve specificity and substantially improve statistical power [113].

3.3.1 Characteristics of sCMRs

We discovered and validated 305 WB sCMRs with large sex differences in DNAm beta values. These sCMRs contained 1174 CpGs, including a number of CpG sites identified in previous studies to be differentially methylated by sex [21, 105] (Figure 24). In addition, multiple tissues and immune cell types including buccal, kidney,

liver, brain, monocytes, CD4 and CD8 T cells also exhibited similar sex-specific DNAm differences for at least 30% of the 305 validated WB sCMRs.

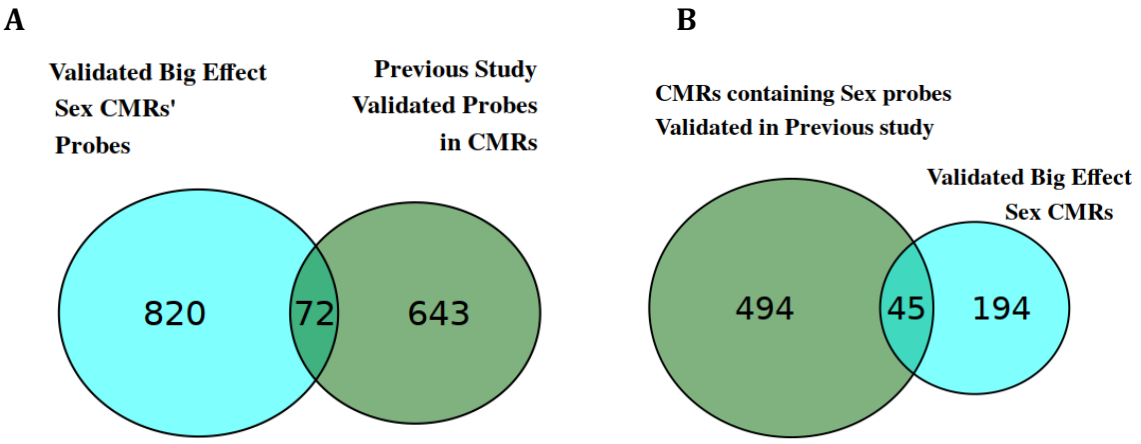


Figure 24. sCMR probes had limited overlap with previous studies of individual probes. A. The sCMRs included 72 previously reported sex-associated individual CpGs. B. Among the 305 validated sCMR, 45 included an individual CpG validated in a previous study.

Using a longitudinal youth cohort with data from three time points, we found that 282 of the 305 validated sCMRs (92%) contained Cps that were significantly associated with sex at age 0,7, and 15 years, while 10 sCMRs were only significant at ages 7 and 15, but were not significant at birth. These sCMRs were consistent with prior findings, as they included known CpGs that change their methylation status during puberty [109].

3.3.2 Sex-specific methylation patterns

We examined common genomic feature annotations for the 305 validated sCMRs in an attempt to glean some insight into their potential role in the establishment of the

sex phenotype and their significance in sex-specific disease etiology. Interestingly, TSS-related and Polycomb-Repressed chromatin state regions, as well as blood and sex-related TFBS and lncRNAs were enriched among the sCMRs. Several of these TFs have binding affinity that is dependent on DNAm methylation [91]. Next, CpGs associated with sex hormone changes in boys over puberty, as reported previously [109], were enriched in CMRs for all five hormones considered, suggesting that DNAm may play a role in the gene expression regulation related to sex hormones. These findings suggest that sCMRs may overlap a subset of genomic regions that have a functional role in the establishment and maintenance of sex phenotype.

We also found that the validated sCMRs were enriched for WB mQTLs identified in previous studies and these enrichments were regardless of which particular mQTL study was used as a reference. These results suggest that for a subset of the sCMRs containing mQTLs, genetic variability may drive the differential DNAm pattern across the sexes. Finally, it was intriguing that the sCMRs were enriched for lncRNAs, indicating that there may be another sex-specific layer in the known interplay between lncRNAs and DNAm in the establishment of chromatin state and downstream gene expression [127-129].

3.3.3 Autosomal Predictor of Sex

The autosomal sex predictor that was constructed from the validated sCMR probes exhibited good performance and would be useful for identifying samples with mis-labeled sex, where XY chromosome array probes are not available, such as from most processed publicly available datasets. Moreover, using only the sCMRs that

were found to be significantly differentially methylated in all tissues still produced a predictor with reasonable performance that would be usable across different tissues. Finally, the sparse 21-probe predictor would be suitable for identifying candidate regions for pyrosequencing, either for validation of sCMRs, or for calling of sample sex.

3.3.4 Strengths and limitations

We focused on pre-defined CMRs in whole blood, in order to determine genomic regions, as opposed to individualCpGs, that may be differentially methylated. We used CoMeBack CMRs [113] which have been show to improve power and specificity, and confirmed that our significant CMRs in the discovery cohort were validated in an independent dataset. On the other hand, our choice to focus on the 34,568 reference CMRs, which contain about 25% of all probes from the 450K array that are not cross-hybridizing (100,633 probes), also means that single probes not within a CMR were not considered for sex-specific DNAm in this study. However, we note that about one hundred thousand of these singleton probes are known to be invariable [31] across individuals. Moreover, previous studies [21, 105] have already considered single CpG EWAS for sex in a large cohort, which we considered for our comparison to the sCMRs.

3.4 Conclusion

We aimed to enhance and extend the characterization of sex-specific DNAm by studying the associations between sex and a set of pre-defined co-methylated

regions, that were estimated agnostic to phenotypic information, in a large sample of WB DNAm data collected with the Illumina Infinium Human Methylation 450K microarray platform. The WB dataset was compiled from five publicly available sources and included only healthy (control) individuals from multiple ethnicities spanning a wide age range between 25-80 years. We validated our sex-associated DNAm findings in three independent datasets.

We took advantage of the longitudinal ARIES cohort to examine the presence of sex-associated CMRs at birth (age 0) and at ages 7 and 15 years. We examined the mainstream publicly available genome annotations and database resources to investigate any enrichments of the validated sCMRs for genomic features, including chromatin states, as well as transcription factor binding site motifs and known genes related to sexual dimorphism.

Co-methylated regions associated with sex included many loci with substantial DNAm differences and enrichments for chromatin states, mQTLS and other genomic features, including ones related to sex biology. A substantial proportion of the WB sCMRs also showed sex-specific DNAm differences in other somatic tissues. Such sCMRs were prevalent across all autosomes, hence their characteristics should be considered in future candidate gene studies and epigenome wide association studies. Finally, we showed that it is possible to use a limited subset of these sCMRs to predict sample sex based on autosomal DNAm.

3.5 Methods

3.5.1 Study populations

3.5.1.1 Discovery

We constructed a large, ancestrally diverse aggregate cohort of 5,191 normative adult whole blood samples (GSE55763, GSE84727, GSE80417, GSE111629, GSE72680) [73, 79-81]. To minimize technical effects, all GEO datasets were pre-processed uniformly. Filtered and normalized DNAm data was used when available. Beta value distributions of Type 1 and Type 2 probes were plotted to confirm if the probe type differences on the 450k array were corrected. Beta-Mixture Quantile (BMIQ) normalization [136] was used to correct for variations resulting from probe biases if needed. Samples outside the 25-80 years age range or labeled as diseased were excluded from the analysis. Samples likely to be sex-mislabels (see sex check using XY probes below) were also removed from the discovery cohort. In detail, the estimation of the sex associations for adults used the CMRs composite betas, defined as the scaled score of the first principal component of the CMR probe betas [113], and it was restricted to the age range of 25-80 years, to avoid the known non-linearity [40] outside this range in the specification for the association between methylation and chronological age. The sample size for this range was 4,605 individuals, whose CMR betas were used to identify sCMRs. We then removed individuals that had a disease and retained only healthy controls from each dataset and we also removed sex-mislabeled samples, resulting in the final set of 3,795 individuals.

To evaluate whether samples from each dataset had been assigned the correct sex based on chromosome complement, samples were subjected to hierarchical clustering based on the beta values from XY probes. Genetic sex was evaluated for each sample by cluster membership, irrespective of the sample's annotated sex in available metadata. Samples were subsequently clustered on beta values at a subset of 5 probes mapping to the *XIST* promoter (cg03554089, cg12653510, cg05533223, cg11717280, cg20698282) [137]. In all datasets two primary clusters corresponding to male and female samples were observed in both clustering checks (XY and *XIST* probes), samples that clustered with those of the opposite sex in either sex check were removed from further analyses. For male samples that clustered with samples of their own sex when considering all XY probes, but clustered with female samples when considering only *XIST* probes, we assessed sex chromosome copy number to rule out 47,XXY chromosome complements with the conumee R package [138]. In total, 10 sex mislabeled samples were identified, corresponding to a prevalence rate of 0.2%, which is much lower than reported prevalence of sex mislabeling in gene expression studies [139].

3.5.1.2 Validation

We validated the significant sCMRs (FDR<0.05) with large sex effect sizes, defined as at least 4% beta difference, in two publicly available cohorts. First, we used a large cohort of adult individuals with publicly available DNAm data (GSE125105) [140], which was pre-processed (see below) as the other public datasets included in the discovery dataset. Next, we validated the sCMRs in the 0, 7 and 15-year old subjects

of the Accessible Resource for Integrated Epigenomic Studies (ARIES) cohort [141], a subsample of 1018 mother–child pairs from the Avon Longitudinal Study of Parents and Children (ALSPAC), a population-based birth cohort [142-144]. The ARIES cohort had 883 individuals (438 male and 445 female) at age 0, 971 individuals (484 male and 487 female) at age 7 and 938 individuals (463 male and 475 female) at age 15 years, where the minimum subject overlap was 854 individuals present at both ages 0 and 15.

3.5.2 Data preprocessing and quality assurance

First, intra-dataset batch effect correction was performed within each cohort, using the R-package sva [82]. The discovery datasets were then merged and corrected for inter-dataset batch effects. The subset of probes that were measured in all datasets was filtered by removing XY chromosome binding probes, non-CpG probes, cross-hybridizing probes [83, 84], and probes containing, or immediately adjacent to a single nucleotide polymorphism (SNP) with a minor allele frequency ≥ 5%, as annotated by the Illumina manifest and [84]. Finally, probes present on the 450K array but absent from the EPIC array were removed to allow comparisons, without modifications, to the newer EPIC arrays. This pre-processing resulted in 404,779 CpG sites. The EPIC samples were pre-processed similarly.

3.5.3 CMR Estimation

We used the whole blood CMRs for the Illumina 450K array constructed with the CoMeBack algorithm [113] with Spearman correlation of 30% and the other parameters at default values, in the discovery dataset described above. As whole

blood is composed of multiple cell types with distinct DNAm patterns, and cell type proportions vary across individuals, we aimed to ensure that the estimated CMRs are not primarily capturing cell-type heterogeneity [85]. Thus, we adjusted the methylation beta ratios for cell type proportions predicted with the Houseman method [28], as implemented in the R-package minfi [85], within a constrained linear regression model. We used the CoMeBack CMRs constructed in the cell-type corrected whole blood data and then used these CMRs to estimate the composite CMR betas in the uncorrected data, as described in [113].

3.5.4 Statistical analysis

We used the CMR composite beta, defined as the weighted sum of the probe betas as calculated with the CoMeBack algorithm [113], within a linear regression specification. The weights were defined previously [113], as the scaled loadings of the first principal component. The linear specification included estimated blood cell type counts, w_k ($\Sigma_k w_k=1$), as well as sex and age, and their interaction:

$$\beta_{CpG\_i} = \Sigma_k w_k + Sex + Age + Sex*Age + \varepsilon$$

We estimated the least-squares specification above while constraining the cell-type specific beta estimates to be between zero and one.

3.5.5 CMR Characteristics: Chromatin State Enrichment

We sought to determine the chromatin state of the genomic regions overlapping

the probes of a CMR. To examine the chromatin state of the CMRs, we used the Roadmap Epigenomics [61] ChromHMM [62]18-state model for PBMCs to estimate enrichment for overlaps with genomic regions in different chromatin states, of the CMRs spanning the genomic coordinates from the first to the last probe, versus the non-CMR probe regions, two base-pairs long, that span CpGs assayed in the non-CMR singleton probes. The overlaps with chromatin states were counted with the R-package GenomicRanges [87], with no-overlaps counted as zeroes and overlaps counted as ones.

3.5.6 Transcription Factor Binding Sites Enrichment

CMR methylation state could potentially be affected by, or itself may affect, the binding of transcription factors (TFs) [88, 91] through their binding sites. Hence, we examined CMR enrichment for known transcription factor binding site motifs, as compiled in the HOCOMOCO v11 database [89]. We scanned regions spanning 200bp on either side of a probe's assayed CpG, using the tool FIMO from the MEME software suite [90]. For any enriched binding site motifs, we checked for any known effects of DNAm on TF binding affinity, using as reference the reported binding specificities of full-length human transcription factors and extended DNA binding domains to (un)methylated DNA for 542 transcription factors [91].

3.5.7 Enrichment for sex hormone-related CpGs

We considered whether the sCMRs were enriched in CpGs with known statistical association to changes in reproductive hormones over the puberty transition. We used the top 999 CpGs identified in [109] as significantly associated (at FDR of 5%) with changes in reproductive hormones among boys during the puberty transition (the study found no significant probes for girls). These significant probes comprised 999, 492, 403, 282, and 218 CpGs associated with changes in testosterone (TS), follicle-stimulating hormone (FSH), luteinizing hormone (LH), anti-Müllerian hormone (AMH), and Inhibin B (InhB), respectively.

3.5.8 lncRNAs Enrichment

We considered whether the sCMRs were enriched in lncRNAs probes that were compiled in a previous study of DNAm changes in cancer [132]. The lncRNA list included 9,066 lncRNAs, and each lncRNA had an associated set of 450K probes. We considered the intersections of these probe-sets with the CMRs and performed an enrichment test using Fisher's exact test.

3.5.9 Enrichment of imprinted genes and imprinting control centers

We tested whether the sCMRs were enriched for probes from a known set of imprinted genes. For this, we downloaded a list of 107 genes from 'Geneimprint' by July 2019 [145] where the genes were filtered to only those with status 'imprinted' in Homo sapiens. Of these genes, 90 had available information in ENSEMBL [146]. We used the genomic ranges for these imprinted genes obtained from ENSEMBL to count intersection with the CMRs.

We also considered enrichment for the 45 imprinting control centers reported in [125, 126]. The genomic ranges were obtained from [126]. We performed the enrichment test using Fisher's exact test.

3.5.10 mQTLs Enrichment

We examined the sCRMs enrichment for known whole blood DNA methylation quantitative trait loci (mQTLs), as reported in the ARIES study[92]and the McRae study [133]. The CMRs were annotated as containing versus not containing an mQTL probe and enrichments were estimated using Fisher's exact test.

3.5.11 Enrichment of GO and KEGG terms

We considered whether the sCMRs were enriched in GO and KEGG terms using the missMethyl R-package [147]. The sCMR genomic regions, spanning the genomic coordinates from the first to the last probe, were annotated for overlaps with the genomic regions of all known genes as annotated by ENSEMBL [146].

3.5.12 CMRs Established over the 0 to 7 to 15 years period

We used the ARIES cohort [108] to investigate which CpGs among the 1174 probes included in the 305 validated sCMRs had significant sex differences. We used a linear model separately in the age 0, 7 and age 15 years sub-samples. We tested all probes from the validated sCMRs using a Welch test and an sCMR was considered validated if it contained at least one significant probe in the ARIES cohort.

3.5.13 Whole Blood sCMRs Across different tissues

We utilized several publicly available datasets from GEO [26] (Kidney: GSE79100, Buccal: GSE80261, Liver: GSE61258, Brain: GSE64509, and immune cell types:

GSE87640) to examine tissue specificity of the 305 validated sex-specific whole blood sCMRs. These samples were run on the 450K array that quantifies DNAm at 485,512 CpG across the genome. To minimize technical effects, all GEO datasets were pre-processed uniformly. Filtered and normalized DNAm data was used when available. Beta value distributions of Type 1 and Type 2 probes were plotted to confirm if the probe type differences on the 450k array were corrected. Beta-Mixture Quantile (BMIQ) normalization [136] was used to correct for variations resulting from probe biases if needed. Further, only" control" samples from each dataset were included in the analysis (Kidney: 31 samples, Buccal: 96 samples, Liver: 79 samples, Cerebellum: 32 samples, Frontal cortex: 41 samples, Hippocampus: 25 samples, Immune cell types: 20 samples). As part of sample quality check, we confirmed reported sex identity using the XY probes and reassigned sex labels if samples clustered incorrectly. We imputed at the median any probes that had a single missing value, and removed probes with more than one missing values. Probe Filtering resulted in variable subsets of CpGs associated with the 1174 CpGs linked with the validated 305 sCMRs (1171 Kidney CpGs , 1142 Buccal CpGs, full set of 1174 for all other tissues).

3.5.14 Cancer data analysis

Level 3 450K methylation data for 10 different TCGA cohorts (BRCA, COAD, LUAD, GBM, STAD, KIRC, LIHC, BLCA, THCA, SKCM - Thyroid carcinoma, Breast invasive carcinoma, Skin Cutaneous Melanoma, Stomach adenocarcinoma, Glioblastoma multiforme, Lung adenocarcinoma, Kidney renal clear cell carcinoma, Liver

hepatocellular carcinoma, Colon adenocarcinoma, and Bladder Urothelial Carcinoma respectively) were retrieve from Firebrowse reporsitory (firebrowse.org, version 2016_01_28). Only tumor samples were retained for the analysis.

R package umap [148] was used to obtain the Uniform Manifold Approximation and Projection (UMAP) plot with the following parameters: random state=123, n_neighbors=40, min_dist=0.2.

3.5.15 Autosomal Predictor of Sex

We used the 1174 probes from the 305 validated sCMRs to construct a sex predictor using the elastic net machine learning algorithm, whose parameters were tuned with 7-fold cross-validation. The predictor was tested in an independent dataset, GSE132203, where DNAm was measured on the EPIC platform, which was processed similarly to the other datasets, as described above. We removed the samples from the GSE132203 data that were also present in the GSE72680 dataset. For the sparse predictor, we increased the shrinkage penalty parameter by ten times the cross-validated value for the shrinkage parameter that was one standard error higher than the minimum. For the PTS predictor, we used the 85 probes that were in the 17 sCMR that had at least one significant probe in all tissues considered.

3.5.16 Software

Preprocessing, quality control, analysis, replication and enrichment analyses were done using R[149], version 3.6.2.

Chapter 4: Applications of Mainstream Machine Learning Algorithms for DNA Methylation Array Data Analysis

4.1 Background

While data reduction of the high dimensional DNAm array measurements may offer advantages, as discussed in chapter 2, some research questions may call for a different approach, where using the non-reduced data as input would be appropriate. For example, the construction of a DNAm based predictor for a condition or disease, that a-priori is expected to include many variably methylated sites, may be better addressed with mainstream Machine Learning (ML) methods that use the complete data as input, while such methods may also perform data reduction or feature extraction as part of the complete prediction algorithm.

ML techniques involve model-free prediction of outcomes (also known in the ML literature as "labeling of examples") [36], after tuning of the algorithmic parameters on a training dataset, without an explicit biological or sometimes even a statistical model. In contrast to biologically motivated data reduction like CoMeBack (chapter 2), where the output is intended for further downstream analyses, ML predictive algorithms produce output that is the end-goal of their application. In particular, such algorithms typically use the full high-dimensional data for algorithmic training and tuning. Subsequently, the performance of the particular ML algorithm is evaluated based on predictive accuracy in new data, as prediction is the end-goal of such ML applications in DNAm data.

Such a strategy using a general purpose ML algorithm was successfully pursued by Horvath, who constructed an "epigenetic clock" predictor for chronological age [40] that has been widely adopted. He extracted a relatively large set of 353 probes out of about 27 thousand array probes, by applying the elastic net (EN) ML algorithm [53]. These probes were used for an "age predictor" with a simple linear specification. This intuitive and easy-to-use linear predictor was quickly adopted in the community, with the main application centering on finding various statistical associations of phenotypic variables with the residual statistical noise of the clock, that was labeled "epigenetic age acceleration." Typically, the newly found statistical associations are interpreted in terms of their biological significance.

This example illustrates an "all-in-one" approach to DNAm predictor construction, where the ML algorithm incorporates a data reduction step as part of the construction. In this case, by shrinking some of the linear coefficient estimates - those below the pre-specified threshold parameter - to zero, the EN algorithm drops uninformative measurements, effectively reducing the high dimensional DNAm data as part of the predictor construction. On the other hand, I showed in chapter 3 that such general purpose ML algorithms as the EN also can be applied successfully downstream of the CoMeBack "pre-processing" data reduction.

This chapter discusses three applications of mainstream ML algorithms in 450K array data, focusing on how the choice of the particular algorithm used for phenotype prediction was motivated by the characteristics of the data and the objectives of the study. This discussion may help guide the choice of ML algorithms in future applications.

4.2 Overview of three mainstream ML Algorithms that I have been applied to DNAm array data.

This section contains a brief discussion of the basic structure of three mainstream ML algorithms that I have applied successfully for DNAm array data analysis: the LASSO, Elastic Net (EN) and Random Forest (RF) algorithms[36].

While the basic structure of the LASSO and the elastic net algorithms consists of a penalized linear regression specification [36], they differ in the type of penalty used for producing a subset of "shrunk" coefficient estimates, which has implications for the suitability of the algorithm for the intended DNAm analysis application. In particular, both algorithms include an "L1" penalty on the absolute values of the linear coefficient estimates, which results in the truncation of small absolute value coefficients to zero, and thus can be used for auxiliary array probe-feature selection. This property is desirable in scenarios where the sample size is small, and hence variance of the estimates in new data and overfitting are the major concerns[36].

In the case of DNAm data, typically there are no a-priori biological reasons to assume that the relationships of DNAm to phenotype are the truly linear [88]. Hence, adopting the simple linear specification implies that a strong bias of the phenotype estimator may be acceptable in lieu of its lower variance in new data. As a way to achieve such low variance, the penalized regression specification reduces the high dimensional DNAm array data, where the number of array probes, p, is much greater than the number of observations, N, that is, p>>N. As typical DNAm datasets with relatively small sample size, N, they require some form of penalty for the number of estimated parameters in order to avoid overfitting and in the case of the linear specification used in the LASSO and the EN algorithms, this involves shrinking the absolute value of the coefficients. The EN uses the L2 penalty of the ridge regression[36], which squares the coefficient estimates, and it adds this L2 penalty, with some weight, to the L1 lasso penalty. In this way, the feature selection of the LASSO, which drops probe with small estimated coefficients, is preserved, while strong signals (with large coefficient estimates) for array probes that are correlated may still be retained because of the use of the ridge-like L2 penalty.

The above discussion of the difference between the two linear specifications estimated with the LASSO versus the EN suggests that the choice of one over the other would be driven primarily by sample size considerations. The LASSO choice may be more appropriate for extremely small samples, like pilot studies, where a sparse strong signal may be sought or expected, while the elastic net, which retains

some redundant correlated signals may be more appropriate for somewhat larger samples, where performance may be meaningfully improved without significant overfitting concerns by including additional parameter estimates. In particular, the EN is also a more suitable choice than the LASSO when applied downstream of CoMeBack pre-filtering, because the probes in the CoMeBack CMRs are correlated by construction, and the inclusion of multiple probes from the same informative CMR may be expected to improve predictor performance. This reasoning underpinned the choice of elastic net for the sex predictors constructed in Chapter 3, as well as in one of the applications discussed below in this chapter.

The two linear specifications of the LASSO and the EN can be contrasted with the random forest (RF) ML algorithm, which draws on the recursive partitioning notion [36] and may be appropriate for some classification applications to DNAm array data. RF is an ensemble method [36], which combines recursive partitioning trees with randomization, so that over-fitting may be reduced in moderate sized datasets. When the number of probes (features in ML parlance) is large, as in DNAm arrays, and many weakly informative features are expected, an ensemble method like RF has been shown to often have good performance [36]. For the smaller statistical samples of DNAm array data observations, the choice of tree depth would be driven by over-fitting considerations and in cases where the hypothesis considers many weakly informative probes that are not expected to interact, “stumps” (single splits)

would be specified for the RF algorithm, that are conceptually similar to a "main effects only" linear model.

Considering probe-feature extraction as another form of data reduction, it is notable that RF may also be used to extract the DNAm probes that were most informative for the classification task at hand. This subset of probes could then be used within a simple recursive partitioning framework [36], so that only a limited number of probes, with their cut-off methylation levels, is used for phenotype classification. For example, if a small number of "diagnostic" probes is desired for classification, a simple committee vote decision rule can be used, that is based on the top most important probes as identified by the RF algorithm. I pursued such approach for detecting maternal blood contamination in cord blood [44].

4.3 Applying Random Forest for Detection of Maternal Blood Contamination in Cord Blood

The maternal blood contamination study [44] focused on selecting a small number of probes for detection of maternal blood contamination in cord blood. The goal of the study was to select a very small set of informative probes that may be suitable for pyrosequencing measurement of CpG methylation. The measured DNAm levels would be compared against a threshold level and would be used to decide on the presence of maternal blood contamination. For completeness, I briefly describe the

study design, before discussing my chosen approach to address the problem and the relevant results.

Neonatal cord blood is a well-interrogated tissue in epigenetic population studies, as it may be expected to be informative about early human development and also it is ready available. Such studies, when using DNAm array data, are often hindered by the introduction of maternal blood during labor, or by cross-contamination during sample collection. When considerable maternal contamination is present, its DNAm could interfere with the DNAm signal arising from the cord blood, or it may introduce spurious DNAm signals arising purely from the maternal blood.

In particular, the study discussed here was motivated by the discovery of maternal contamination of cord blood in a cohort of 150 neonates that were assayed with the Illumina 450K DNAm array. The contamination was initially identified by the uncharacteristic X chromosome DNAm patterns in 17 male neonates. The study then exploited the fact that DNAm exhibits substantial differences between neonates and adults [150, 151] and hence observed contamination, in terms of measured DNAm signal, would be different between male and female neonates. Specifically, since the X chromosome DNAm has highly distinct patterns for males versus females, the DNAm of XX blood from female mothers mixed with the blood of XY male neonates would be more distinct from the a mix with XX female neonates.

The main objective of the study was to identify a small panel of CpGs that can be used to identify contaminated cord blood samples. The study design was based on filtering of DNAm array probes, including an application of the RF algorithm, to select 10 informative CpGs that can discriminate between (un)contaminated samples. These 10 CpGs were used for a predictor of maternal contamination that was validated in an independent dataset of 189 additional samples, as well as by performing pyrosequncing assays in house.

In more detail, the informative CpG site selection involved a two-stage data reduction approach, where the first stage looked at individual probe associations, to identify the probes that had the highest discriminatory power for contaminated samples. This stage, which can be viewed as an initial filtering step in the spirit of dimension reduction, was performed in order to reduce the number of RF inputs to only the stronger statistical signals. While these filtered probes were all highly statistically significant, they were still potentially redundant and the goal for the RF filtering that I performed as the second stage, was to further reduce the number of CpG sites to a very small set, that would also be suitable for pyro-sequencing assessment. In this case, I received a list of 2,250 candidate CpGs that were obtained by collaborators who had performed an initial filtering based on a linear model specification. My task was to identify a very small set of 10 top candidate CpGs that would be used within a final predictor to call maternal contamination.

More specifically, my objective for the predictor was to produce a very small set of 10 independent DNAm signals that may be used for committee voting to decide on the contamination status outcome. Since the sample size in this study was relatively small (N=60), I chose the RF approach because it could incorporate multiple predictors in a non-linear way, while also being robust to over-fitting [36]. The relatively small size of the dataset also dictated my choice to use stumps (single node splits) for the RF. While the predictive performance of the algorithm was used for assessment of the feasibility of contamination detection based on DNAm, its ancillary output of probe-feature importance informed the choice of top candidates for the proposed pyro-sequencing application, including the in-house validation.

In more detail, when the RF algorithm was used to produce a prediction for maternal blood contamination from the filtered set of 2250 CpGs that were output from the first filtering stage, it ranked the CpG sites by the mean decrease in accuracy, which is a measure of their relative importance for the predictor. This probe importance output of the RF algorithm was the basis for the probe-feature selection that identified the top candidate probes. For each of the selected 10 probes with top importance, a threshold value was determined for separating contaminated from non-contaminated samples. These cut-off values were determined by performing binary recursive partitioning for each top probe [36].

In summary, the algorithm was trained with 60 samples and the top 10 probe features in terms of relative importance were the input for the final predictor. These sites were also examined and curated by epigenetics experts for their suitability for pyro-sequencing assays. The three CpG sites: cg25556035, cg15931839, and cg02812891 chosen had the best discrimination between contaminated and non-contaminated male samples, and were CpG sites for which robust pyrosequencing assays were feasible (Figure 25).

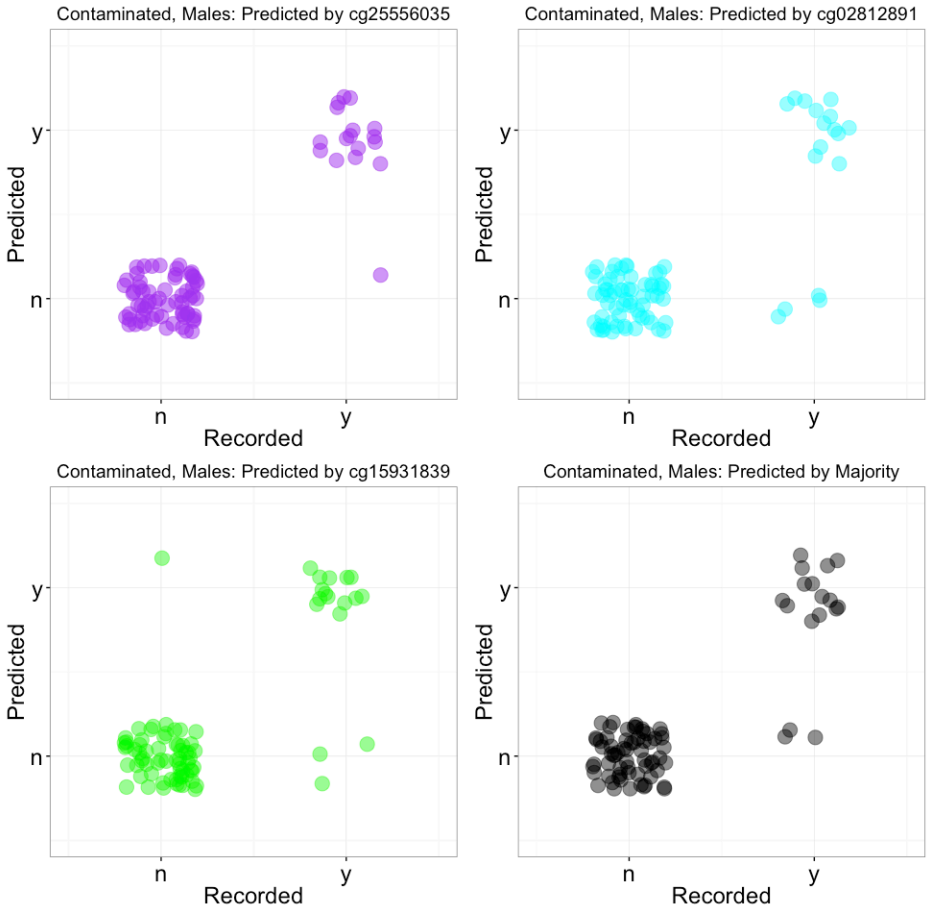


Figure 25. The maternal contamination predictor had good performance. Confusion tables for the 3 individual probes used for the predictor, and for the committee majority vote (bottom right).

The 3 probes were used as "committee members" to make contamination calls based on pyrosequencing assays, so that a majority vote of at least 2 probes would determine the call outcome. A more strict call would require all 3 committee members to agree on the call.

The subsequent predictor validation by pyrosequncing assay in an independent cohort of 189 individuals demonstrated its robust performance and confirmed that the RF algorithm application was appropriate for the construction of this predictor of maternal contamination of cord blood based on DNAm. We now turn to the application of another ML algorithm, the EN, for prediction of phenotype, based on DNAm array data.

4.4 Applying the Elastic Net algorithm to construct a childhood abuse exposure predictor in Sperm

My objective for the study of exposure to childhood abuse (CA) [43] was to construct a sparse predictor of CA based on DNAm in human sperm. This small size pilot study was motivated by previous findings [152] indicating that differential DNAm associated with CA may be expected at multiple loci in Sperm tissue. Given the very small sample size, I chose to construct a sparse predictor that would be expected to avoid severe overfitting based on a penalized generalized linear model, with an EN algorithm that combines L1 and L2 penalty (see description above). Below, I briefly describe the study design for completeness, followed by discussion

of the ML approach chosen to address the problem and a summary of the predictor results.

CA has been associated with differences in DNAm in multiple tissues[152]. The study discussed here was the first one to examine the association of CA with DNA methylation in human sperm. The study design included, first, a composite CA measure where physical, emotional, and sexual abuse in childhood that was measured on a discrete scale and was characterized as none, medium, or high. Next, DNAm was assayed using HumanMethylation450 BeadChips in 46 sperm samples from 34 men in a longitudinal non-clinical cohort. As a main focus, the study considered the associations of the DNAm principal components (PCs) with CA, as well as differentially methylated regions (DMRs) for CA. As a secondary objective, this study aimed to identify a small panel of CpGs that can be used to predict CA in sperm samples. For the purpose of constructing the predictor, the CA measure was dichotomized as High versus Low, where the Low category included the "none" and "medium" levels of the composite CA measure. Given the focus on DMRs associated with CA, I applied the EN algorithm to construct a CA predictor for sperm DNAm.

The EN algorithm, which incorporated a data reduction step, identified multiple CpG sites predictive of CA. The use of the EN instead of the LASSO algorithm was guided by the study design, which aimed to compare the CpG sites selected for the predictor with the sites identified in a separate DMR analysis. To this end, the EN was the

appropriate choice, because it was more likely than the LASSO to retain correlated CpGs within a DMR due to the L2 penalty, which is not used by the LASSO. In this application, an EN penalized generalized linear regression was estimated both for the input set of all array probes and for the restricted set of probes in the most significant DMRs, using the dichotomized CA variable mentioned above as the outcome variable. The output of the algorithm consisted of a predictor model that used an extremely small set of DNAm probes.

Specifically, when applied to the full data, the EN approach identified probes cg02622647, cg04703951, and cg17369694 as the most useful ones for classifying male participants with none or medium versus high abuse exposure. These probes correctly classified 79% of participants (12 true positives, five false positives, 15 true negatives, and two false negatives) in the training data.

Next, I refined the predictor construction by filtering of the input probes and using only the probes from the 12 most significant DMRs. The resulting predictor included 4 probes (Figure 26) and had improved performance of 88% accuracy in the training data. Note that the use of filtered probes from DMRs for the 4-probe CA predictor construction, which led to a performance improvement, is conceptually similar to the construction of the Sex predictor in Chapter 3, where only the probes of the sex-associated CMRs were used to construct a sparse predictor for biological sex.

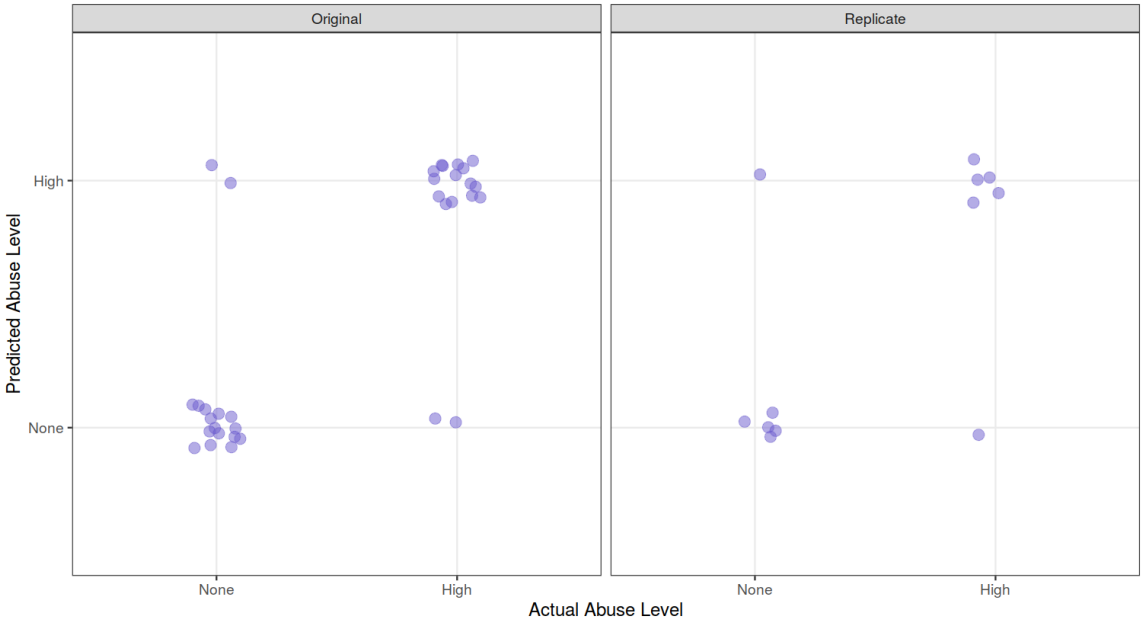


Figure 26. DNA methylation predictor for childhood abuse exposure.

The predictor based on four probes (cg02622647, cg09926099, cg00537837, cg20333904) selected from the four most-significant DMRs (NDUFA10, MIR5093, LRRK1 and ARL17A) achieved accuracy of 88% (p-value of 0.002) in predicting low versus high CA exposure (left panel). The predictive accuracy was similar in the replication set (right panel)

As no public datasets with measured sperm DNAm and CA were available, I could not test directly the predictor's ability to predict CA status in an independent cohort. However, the constructed 3-probe predictor was applied to three independent datasets (GSE108058[153], GSE102970[154], and GSE64096[155]) to ascertain whether the prevalence of abuse estimated with this predictor was approximately the same as the prevalence in the whole cohort from which the pilot study sample was drawn, where high abuse prevalence was 29%. In the three independent

datasets examined, the predicted CA prevalence was 30%, 35%, and 25%, respectively, which was similar to the main cohort prevalence.

While this study also demonstrates the limitations of performance evaluation when independent testing datasets are not available, such scenario is not atypical, given the limited availability of public DNAm data with detailed phenotypic measurements. In the next section, we now consider a recent application of the LASSO algorithm for predictor construction, where performance evaluation was also limited by testing data availability.

4.5 Applying the LASSO algorithm to construct a body weight percentage change predictor

The physical activity (PA) intervention study[41], investigated DNAm patterns before and after a lifestyle intervention from a 6-month pilot randomized control trial. In this study, my objective was to construct a sparse predictor of body weight change over a 6-month period based on the subject's DNAm patterns. I used the LASSO algorithm for the feature selection of a very small subset of the 450K array probes. While the LASSO model may incur high bias (see Chapter 1) due to its sparse linear specification, the main consideration driving its choice in this study was the very small training sample size, with its associated danger of over-fitting [36] a complicated model with large number of estimated parameters. Below, I briefly describe the study design, before proceeding to summarize the main

properties and performance results for the constructed weight percentage change predictor.

This PA intervention study explored potential epigenetic mechanisms underlying the many health benefits conferred by PA, by examining the DNAm patterns of 20 healthy postmenopausal community-dwelling women aged 55 to 70 years, who did not have a mobility disability. The study design was a pilot randomized controlled trial over 6 months, which tested the association between longitudinal measures of DNAm and changes in several objective measures, including PA and weight loss. The study subjects were allocated to an intervention group, or to a control group, with both groups receiving different monthly group-based health- related education sessions. Specifically, the randomized controlled trial consisted of nine 2-hour sessions focused on reducing sedentary behavior for the intervention group, versus six 1-hour sessions focused on other topics, for the control group. Samples of peripheral blood mononuclear cells (PBMCs) were collected both at baseline before the intervention and then at 6 months, after the trial was completed. The samples were interrogated with the Illumina 450k Methylation array to quantify genome-wide DNAm. The goals of the study were to examine potential associations between the 6-month lifestyle intervention and epigenetic changes, and to determine if it was possible to construct an epigenetic predictor of the intervention.

One of the main objectives was to construct a sparse predictor for weight loss, using only a very small number of CpG probes. These probes would be examined further for their biological function. This additional interrogation would include comparisons with the existing literature and annotations, to check whether the probes are known from prior studies focusing on body mass index (BMI) measures, or if they may be interpretable in terms of known genes or regulatory regions relevant to BMI. To construct a DNAm predictor for weight percentage change over the intervention, I applied the LASSO ML algorithm.

In this study, the absence of a "golden standard" for evaluation, combined with the limited knowledge of relevant biological mechanisms involving DNAm, suggested that in order to avoid over-fitting in an exploratory analysis, the probe selection with the LASSO had to be performed with heavy regularization due to the very small sample size (N=20). Hence, the penalty parameter for the LASSO regularization (determining how small the absolute value of the estimate has be for it to be set to zero) was set at a large value in order to retain only the most relevant probe measurements. In this way, the use of a very small set of probes within a simple linear specification aimed to reduce the variance of the weight loss predictor in new data.

The application of the LASSO in the PA intervention study resulted in identification of five CpG sites, (cg17920653, cg25134701, cg24088639, cg22664307, and

cg08104023), whose base line DNAm was able to predict the percent body weight change over the 6-month period, while controlling for baseline weight. The performance of the predictor was adequate, given the very small size of the sample used for the construction: the correlation between the predicted weight loss and actual weight loss was 74% (Figure 27). Thus, the accuracy of the DNAm array data based predictor was limited by adequate training data availability. Hence, the main purpose of this construction was to explore, as proof of concept, the potential feasibility of constructing a relatively accurate predictor when more training data become available.

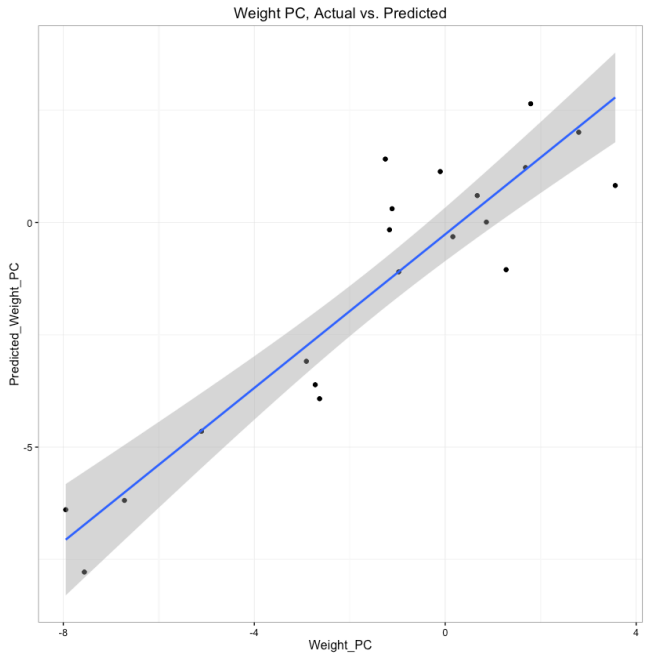


Figure 27. Performance of the weight percentage change predictor.
Fitted values of weight percentage change (PC) from a robust linear model plotted against actual measured values. Grey shading shows the 95% confident intervals.

4.6 Conclusions

The three applications discussed above illustrated how mainstream ML algorithms can be applied in the context of DNAm data to address a particular set of research questions. Such algorithms can be useful for prediction or imputation of unmeasured phenotype based on DNAm array data. While data reduction of the high-dimensional DNAm array measurements is a central component of two of the algorithms discussed above, it is embedded within the algorithm and the reduced data typically are not the main intended output of the algorithm. The choice of using a mainstream ML approach, as opposed to a biologically motivated data reduction or filtering, such as CoMeBack, can be guided by the particular study objectives. As the above examples showed, mainstream ML algorithms can be applied successfully in studies with a narrow focus on phenotype prediction from DNAm array data. The choice of a particular ML algorithm, for example LASSO vs EN, may be driven primarily by the quantity and quality of training DNAm data, as well as by biological considerations, such as focusing on genomic regions, as opposed to individual CpG sites.

Chapter 5: Conclusions and Future Directions

5.1 DNAm array Data Reduction with CoMeBack may improve specificity, power and biological interpretation

In this thesis I developed a biologically motivated method for DNAm microarray data reduction, examined its performance and demonstrated its applications using publicly available data. The need for DNAm array data reduction stems from the high dimensional nature of the data, which presents substantial analytical challenges. Moreover, biological hypotheses about the role of DNAm in gene expression regulation often entail a low expected number of significant CpG sites, possibly with small effects.

The CoMeBack method discussed in this thesis was based on existing biological findings about correlations among proximal genomic CpG sites and hence it approached the data reduction problem in a way that is different from data driven methods relying purely on statistical models. The main motivation for incorporating biological findings was to reduce the false positive findings that occur due to spurious correlations, and to increase statistical power by improving the signal to noise ratio.

The applications of CoMeBack to public data demonstrated that the method achieved good statistical power and specificity, including new findings in CMR-based EWAS about chronological age and biological sex, that were not discovered by

standard single CpG site analysis. Moreover, the reference whole blood CMRs that were constructed with public data were extensively characterized in terms of chromatin state, mQTLs and transcription factor bindings sites, highlighting the utility of the data reduction for downstream analysis with various analytical pipelines.

The CoMeBack algorithm was implemented in an easy to use, open source R package that has useful functionalities for analysis of both 450K and EPIC arrays and offers the user flexibility in terms of setting the key algorithmic parameters. Future developments of the CoMeBack method may lead to further improvements in terms of available functionality and performance. In particular, with additional public data becoming available, there are several areas where I plan to continue development of the CoMeBack methodology.

5.2 Reference CMRs for different tissues

The whole blood reference CMRs that were constructed with public 450K data can be used successfully for downstream CMR-EWAS application, as I demonstrated in chapter 3, which characterized the CMRs associated with Sex. Such applications showcase the utility of reference CMRS for additional tissues that can be used to address tissue-specific biological hypotheses. The growing number of publicly available DNAm datasets for different tissues offers the exciting prospect of constructing additional reference CMR sets for multiple tissues. Moreover, going

forward, the use of the newer Illumina EPIC platform, which doubles the number of assayed genomic CpG sites, will afford a more extensive characterization of the CMRs than the one with the recently retired 450K platform.

5.3 Using extended annotation data for CMR construction

The CoMeBack algorithm used genomic CpG coordinates to implement a biologically meaningful filtering of correlated CpG sites that aimed to reduce false positives arising from spurious correlations. It is conceivable that this simple biological model may be extended to incorporate additional functional annotations of genomic features, as they become available for different tissues. For example, the genomic coordinates of tissue-specific enhancers and other chromatin states may be useful for a more refined filtering of correlated CpG sites. As such data become more readily available, I plan to incorporate an optional, more flexible use of extended genomic annotations within the CoMeBack software, with the goal of further improving the statistical specificity of future tissue-specific reference CMRs constructed with publicly available EPIC data.

5.4 Adding convenience functionalities for Visualization and Integration with other softwares

The CoMeBack software outputs CMRs that are easy to analyze within downstream analytical pipelines, as was demonstrated in chapters 2 and 3. It is conceivable that

some of these downstream characterizations of CMRs may become standard practice, like enrichments of a subset of CMRs of interest for genomic features like imputed or measured chromatin states, or TF binding sites. In such ase, it may be desirable to incorporate such additional functionality within the software, in order to facilitate the downstream analysis. Moreover, built-in visualization of CMRs and their enrichments may improve the appeal of the software and increase its adoption, as researchers may prefer a more integrated solution that also addresses the need to present their results. As I proceed with further applications of CoMeBack in my own research, I plan to incorporate the most commonly used downstream analyses within the next versions of the software, while also taking into account any external user feedback that may appear on the software website.

5.5 Computational performance improvements

While CoMeBack has reasonable performance on modern computers, the future planned extensions of the functionality mentioned above are expected to place an additional computational burden, for example by running multiple CpG filters based on various genomic annotations. Such potential developments would call for additional optimizations of the software in order to cope with larger datasets and potentially more sophisticated filtering of correlated CpGs. If improved computational performance becomes desirable in the future, I plan to offer a parallelized version of the software. In addition to that, I plan to perform some

additional computational experiments to explore the potential of using optional graphics processing unit (GPU) acceleration to speed up the algorithm. It is my hope that such meaningful improvements would increase the appeal and adoption of the CoMeBack software.

Bibliography

1. Yuen,R.K., Neumann,S.M., et al. (2011) Extensive epigenetic reprogramming in human somatic tissues between fetus and adult. *Epigenetics & chromatin*, 4, 7.

2. Kim,M. and Costello,J. (2017) DNA methylation: an epigenetic mark of cellular memory. *Exp.Mol.Med.*, 49, e322.

3. Slieker,R.C., Bos,S.D., et al. (2013) Identification and systematic annotation of tissue-specific differentially methylated regions using the Illumina 450k array. *Epigenetics & Chromatin*, 6, 26.

4. Hannon,E., Lunnon,K., et al. (2015) Interindividual methylomic variation across blood, cortex, and cerebellum: implications for epigenetic studies of neurological and neuropsychiatric phenotypes. *Epigenetics*, 10, 1024-1032.

5. Jaenisch,R. and Bird,A. (2003) Epigenetic regulation of gene expression: how the genome integrates intrinsic and environmental signals. *Nat.Genet.*, 33, 245-254.

6. Smith,Z.D. and Meissner,A. (2013) DNA methylation: roles in mammalian development. *Nature Reviews Genetics*, 14, 204-220.

7. Ladd-Acosta,C. (2015) Epigenetic Signatures as Biomarkers of Exposure. *Current Environmental Health Reports*, 2, 117-125.

8. Marsit,C.J. (2015) Influence of environmental exposure on human epigenetic regulation. *J.Exp.Biol.*, 218, 71-79.

9. Joubert,B.R., Felix,J.F., et al. (2016) DNA Methylation in Newborns and Maternal Smoking in Pregnancy: Genome-wide Consortium Meta-analysis. *Am.J.Hum.Genet.*, 98, 680-696.

10. McClay,J.L., Aberg,K.A., et al. (2014) A methylome-wide study of aging using massively parallel sequencing of the methyl-CpG-enriched genomic fraction from blood in over 700 subjects. *Hum.Mol.Genet.*, 23, 1175-1185.

11. Jones,M.J., Goodman,S.J., et al. (2015) DNA methylation and healthy human aging. *Aging cell*, 14, 924-932.

12. Zhang,F.F., Cardarelli,R., et al. (2011) Significant differences in global genomic DNA methylation by gender and race/ethnicity in peripheral blood. *Epigenetics*, 6, 623-629.

13. Liu,J., Morgan,M., et al. (2010) A Study of the Influence of Sex on Genome Wide Methylation. *PloS one*, 5, e10028.

14. El-Maarri,O., Becker,T., et al. (2007) Gender specific differences in levels of DNA methylation at selected loci from human total blood: a tendency toward higher methylation levels in males. *Hum.Genet.*, 122, 505-514.

15. El-Maarri,O., Walier,M., et al. (2011) Methylation at Global LINE-1 Repeats in Human Blood Are Affected by Gender but Not by Age or Natural Hormone Cycles. *PLoS One*, 6.

16. Tapp,H.S., Commane,D.M., et al. (2013) Nutritional factors and gender influence age-related DNA methylation in the human rectal mucosa. *Aging Cell*, 12, 148-155.

17. Boks,M.P., Derks,E.M., et al. (2009) The relationship of DNA methylation with age, gender and genotype in twins and healthy controls. *PLoS ONE*, 4, e6767.

18. Sarter,B., Long,T.I., et al. (2005) Sex differential in methylation patterns of selected genes in Singapore Chinese. *Hum.Genet.*, 117, 402-403.

19. McCarthy,N.S., Melton,P.E., et al. (2014) Meta-analysis of human methylation data for evidence of sex-specific autosomal patterns. *BMC genomics*, 15, 981.

20. Bernardino,J., Lombard,M., et al. (2000) Common methylation characteristics of sex chromosomes in somatic and germ cells from mouse, lemur and human. *Chromosome Research*, 8, 513-525.

21. Singmann,P., Shem-Tov,D., et al. (2015) Characterization of whole-genome autosomal differences of DNA methylation between men and women. *Epigenetics & chromatin*, 8, 43.

22. Lin,S., Liu,Y., et al. (2019) Sex-related DNA methylation differences in B cell chronic lymphocytic leukemia. *Biology of Sex Differences*, 10, 2.

23. García-Calzón,S., Perfilyev,A., et al. (2018) Sex Differences in the Methylome and

Transcriptome of the Human Liver and Circulating HDL-Cholesterol Levels. *None*, 103, 4395-4408.

24. Bibikova,M., Barnes,B., et al. (2011) High density DNA methylation array with single CpG site resolution. *Genomics*, 98, 288-295.

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

26. Barrett,T., Wilhite,S.E., et al. (2013) NCBI GEO: archive for functional genomics data sets--update. *Nucleic Acids Res.*, 41, D991-D995.

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

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

29. Eckhardt,F., Lewin,J., et al. (2006) DNA methylation profiling of human chromosomes 6, 20 and 22. *Nat.Genet.*, 38, 1378-1385.

30. Hui,T., Cao,Q., et al. (2018) High-Resolution Single-Cell DNA Methylation Measurements Reveal Epigenetically Distinct Hematopoietic Stem Cell Subpopulations. *Stem Cell Reports*, 11, 578-592.

31. Edgar,R.D., Jones,M.J., et al. (2017) An empirically driven data reduction method on the human 450K methylation array to remove tissue specific non-variable CpGs. *Clinical Epigenetics*, 9, 11.

32. Cappelli,E., Felici,G., et al. (2018) Combining DNA methylation and RNA sequencing data of cancer for supervised knowledge extraction. *BioData Mining*, 11, 22.

33. Rakyan,V.K., Down,T.A., et al. (2011) Epigenome-wide association studies for common human diseases. *Nature Reviews Genetics*, 12, 529-541.

34. Flanagan,J.M. (2015) Epigenome-Wide Association Studies (EWAS): Past, Present, and

Future. In Verma,M. (ed.), *Cancer Epigenetics: Risk Assessment, Diagnosis, Treatment, and Prognosis.* Springer New York, New York, NY, pp. 51-63.

35. Benjamini,Y. and Hochberg,Y. (1995) Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, 57, 289-300.

36. Hastie,T., Tibshirani,R., et al. (2009) The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition. Anonymous Springer-Verlag, New York.

37. Langfelder,P. and Horvath,S. (2008) WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics*, 9, 559.

38. Bellman,R.E. (1961) Adaptive control processes: a guided tour. Anonymous Princeton, N.J., Princeton University Press.

39. Tsai,P. and Bell,J.T. (2015) Power and sample size estimation for epigenome-wide association scans to detect differential DNA methylation. *Int J Epidemiol*, 44, 1429-1441.

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

41. McEwen,L.M., O'Donnell,K.J., et al. (2019) The PedBE clock accurately estimates DNA methylation age in pediatric buccal cells. *Proc.Natl.Acad.Sci.USA*, 201820843.

42. Chapelle, Olivier, Bernhard Schölkopf and Alexander Zien. Semi-Supervised Learning. *The MIT Press*.

43. Roberts,A.L., Gladish,N., et al. (2018) Exposure to childhood abuse is associated with human sperm DNA methylation. *Translational psychiatry*, 8, 194-11.

44. Morin,A.M., Gatev,E., et al. (2017) Maternal blood contamination of collected cord blood can be identified using DNA methylation at three CpGs. *Clinical Epigenetics*, 9, 75.

45. Rahmani,E., Zaitlen,N., et al. (2016) Sparse PCA corrects for cell type heterogeneity in epigenome-wide association studies. *Nature Methods*, 13, 443-445.

46. Capper,D., Jones,D.T.W., et al. (2018) DNA methylation-based classification of central nervous system tumours. *Nature*, 555, 469-474.

47. Zou,H., Hastie,T., et al. (2006) Sparse Principal Component Analysis. *Journal of Computational and Graphical Statistics*, 15, 265-286.

48. Crowgey,E.L., Marsh,A.G., et al. (2018) Epigenetic machine learning: utilizing DNA methylation patterns to predict spastic cerebral palsy. *BMC Bioinformatics*, 19, 225.

49. Bartlett,C.L., Glatt,S.J., et al. (2019) Machine Learning and Feature Selection for the Classification of Mental Disorders from Methylation Data. , 311-321.

50. Jurmeister,P., Bockmayr,M., et al. (2019) Machine learning analysis of DNA methylation profiles distinguishes primary lung squamous cell carcinomas from head and neck metastases. *Science Translational Medicine*, 11, eaaw8513.

51. Angermueller,C., Lee,H.J., et al. (2017) DeepCpG: accurate prediction of single-cell DNA methylation states using deep learning. *Genome Biol.*, 18, 67.

52. Vidaki,A., Ballard,D., et al. (2017) DNA methylation-based forensic age prediction using artificial neural networks and next generation sequencing. *Forensic Science International: Genetics*, 28, 225-236.

53. Hui Zou and Trevor Hastie. (2005) Regularization and Variable Selection via the Elastic Net. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 67, 301-320.

54. Tsai,C., Wang,S., et al. (2005) Sample size for gene expression microarray experiments. *Bioinformatics*, 21, 1502-1508.

55. Stretch,C., Khan,S., et al. (2013) Effects of Sample Size on Differential Gene Expression, Rank Order and Prediction Accuracy of a Gene Signature. *PloS one*, 8, e65380.

56. Zheng-Bradley,X., Rung,J., et al. (2010) Large scale comparison of global gene expression patterns in human and mouse. *Genome Biol.*, 11, R124.

57. Maleki,F., Ovens,K., et al. (2019) Size matters: how sample size affects the reproducibility

and specificity of gene set analysis. *Hum Genomics*, 13.

58. Jaffe,A.E., Murakami,P., et al. (2012) Bump hunting to identify differentially methylated regions in epigenetic epidemiology studies. *Int J Epidemiol*, 41, 200-209.

59. Peters,T.J., Buckley,M.J., et al. (2015) De novo identification of differentially methylated regions in the human genome. *Epigenetics & chromatin*, 8, 6.

60. Sofer,T., Schifano,E.D., et al. (2013) A-clustering: a novel method for the detection of co-regulated methylation regions, and regions associated with exposure. *Bioinformatics*, 29, 2884-2891.

61. Kundaje,A., Meuleman,W., et al. (2015) Integrative analysis of 111 reference human epigenomes. *Nature*, 518, 317-330.

62. Ernst,J. and Kellis,M. (2012) ChromHMM: automating chromatin state discovery and characterization. *Nat Methods*, 9, 215-216.

63. Ong,M. and Holbrook,J.D. (2014) Novel region discovery method for Infinium 450K DNA methylation data reveals changes associated with aging in muscle and neuronal pathways. *Aging Cell*, 13, 142-155.

64. Bell,J.T., Pai,A.A., et al. (2011) DNA methylation patterns associate with genetic and gene expression variation in HapMap cell lines. *Genome Biol*, 12, R10.

65. Liu,D., Zhao,L., et al. (2019) EWASdb: epigenome-wide association study database. *Nucleic Acids Res*, 47, D989-D993.

66. Bird,A., Taggart,M., et al. (1985) A fraction of the mouse genome that is derived from islands of nonmethylated, CpG-rich DNA. *Cell*, 40, 91-99.

67. Deaton,A.M. and Bird,A. (2011) CpG islands and the regulation of transcription. *Genes & development*, 25, 1010-1022.

68. Esteller,M. (2002) CpG island hypermethylation and tumor suppressor genes: a booming present, a brighter future. *Oncogene*, 21, 5427-5440.

69. Gardiner-Garden,M. and Frommer,M. (1987) CpG Islands in vertebrate genomes. *Journal of Molecular Biology*, 196, 261-282.

70. Serge Saxonov, Paul Berg, et al. (2006) A Genome-Wide Analysis of CpG Dinucleotides in the Human Genome Distinguishes Two Distinct Classes of Promoters. *Proceedings of the National Academy of Sciences of the United States of America*, 103, 1412-1417.

71. Zhang,M.Q. and Ioshikhes,I.P. (2000) Large-scale human promoter mapping using CpG islands. *Nature Genetics*, 26, 61-63.

72. Assenov,Y., Müller,F., et al. (2014) Comprehensive Analysis of DNA Methylation Data with RnBeads. *Nat Methods*, 11, 1138-1140.

73. Horvath,S. and Ritz,B.R. (2015) Increased epigenetic age and granulocyte counts in the blood of Parkinson's disease patients. *Aging (Albany NY)*, 7, 1130-1142.

74. Morris,T.J., Butcher,L.M., et al. (2014) ChAMP: 450k Chip Analysis Methylation Pipeline. *Bioinformatics (Oxford, England)*, 30, 428-430.

75. Rakyan,V.K., Beyan,H., et al. (2011) Identification of Type 1 Diabetes–Associated DNA Methylation Variable Positions That Precede Disease Diagnosis. *PLoS genetics*, 7, e1002300.

76. Liu,Y., Li,X., et al. (2014) GeMes, Clusters of DNA Methylation under Genetic Control, Can Inform Genetic and Epigenetic Analysis of Disease. *Am J Hum Genet*, 94, 485-495.

77. Martin,T.C., Yet,I., et al. (2015) coMET: visualisation of regional epigenome-wide association scan results and DNA co-methylation patterns. *BMC bioinformatics*, 16, 131.

78. Houseman,E.A., Christensen,B.C., et al. (2008) Model-based clustering of DNA methylation array data: a recursive-partitioning algorithm for high-dimensional data arising as a mixture of beta distributions. *BMC Bioinformatics*, 9, 365.

79. Hannon,E., Dempster,E., et al. (2016) An integrated genetic-epigenetic analysis of schizophrenia: evidence for co-localization of genetic associations and differential DNA methylation. *Genome Biol*, 17.

80. Kilaru,V. (2015) GSE72680, DNA Methylation of African Americans from the Grady Trauma Project.

81. Lehne,B., Drong,A.W., et al. (2015) A coherent approach for analysis of the Illumina HumanMethylation450 BeadChip improves data quality and performance in epigenome-wide association studies. *Genome Biol*, 16.

82. Leek,J.T., Johnson,W.E., et al. (2012) The sva package for removing batch effects and other unwanted variation in high-throughput experiments. *Bioinformatics*, 28, 882-883.

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

84. Price,M.E., Cotton,A.M., et al. (2013) Additional annotation enhances potential for biologically-relevant analysis of the Illumina Infinium HumanMethylation450 BeadChip array. *Epigenetics & chromatin*, 6, 4.

85. Aryee,M.J., Jaffe,A.E., et al. (2014) Minfi: a flexible and comprehensive Bioconductor package for the analysis of Infinium DNA methylation microarrays. *Bioinformatics*, 30, 1363-1369.

86. Kilaru,V. (2019) GSE132203, DNA Methylation (EPIC) from the Grady Trauma Project.

87. Lawrence,M., Huber,W., et al. (2013) Software for Computing and Annotating Genomic Ranges. *PLOS Computational Biology*, 9, e1003118.

88. Jones,P.A. (2012) Functions of DNA methylation: islands, start sites, gene bodies and beyond. *Nature Reviews Genetics*, 13, 484-492.

89. Kulakovskiy,I.V., Vorontsov,I.E., et al. (2018) HOCOMOCO: towards a complete collection of transcription factor binding models for human and mouse via large-scale ChIP-Seq analysis. *Nucleic Acids Res*, 46, D252-D259.

90. Bailey,T.L., Boden,M., et al. (2009) MEME Suite: tools for motif discovery and searching. *Nucleic Acids Research*, 37, W202-W208.

91. Yin,Y., Morgunova,E., et al. (2017) Impact of cytosine methylation on DNA binding specificities of human transcription factors. *Science*, 356.

92. Shihab, H. A. et al. (2016) ARIES mQTL results.

93. Day,K., Waite,L.L., et al. (2013) Differential DNA methylation with age displays both common and dynamic features across human tissues that are influenced by CpG landscape. *Genome Biol*, 14, R102.

94. Johansson,Å, Enroth,S., et al. (2013) Continuous Aging of the Human DNA Methylome Throughout the Human Lifespan. *PLoS One*, 8.

95. Spólnicka,M., Pośpiech,E., et al. (2018) DNA methylation in ELOVL2 and C1orf132 correctly predicted chronological age of individuals from three disease groups. *Int J Legal Med*, 132, 1-11.

96. Smith,A.K., Kilaru,V., et al. (2014) Methylation quantitative trait loci (meQTLs) are consistently detected across ancestry, developmental stage, and tissue type. *BMC Genomics*, 15, 145.

97. Klein,S.L. and Flanagan,K.L. (2016) Sex differences in immune responses. *Nature Reviews Immunology*, 16, 626-638.

98. Mittelstrass,K., Ried,J.S., et al. (2011) Discovery of Sexual Dimorphisms in Metabolic and Genetic Biomarkers. *PLoS Genet*, 7.

99. Boraska,V., Jerončić,A., et al. (2012) Genome-wide meta-analysis of common variant differences between men and women. *Human Molecular Genetics*, 21, 4805-4815.

100. Michels,K.B., Binder,A.M., et al. (2013) Recommendations for the design and analysis of epigenome-wide association studies. *Nature Methods*, 10, 949-955.

101. Mansell,G., Gorrie-Stone,T.J., et al. (2019) Guidance for DNA methylation studies: statistical insights from the Illumina EPIC array. *BMC Genomics*, 20, 366-366.

102. Jin,Z. and Liu,Y. (2018) DNA methylation in human diseases. *Genes & Diseases*, 5, 1-8.

103. Bird,A. (2002) DNA methylation patterns and epigenetic memory. *Genes & development*, 16, 6-21.

104. Plasschaert,R.N. and Bartolomei,M.S. (2014) Genomic imprinting in development, growth, behavior and stem cells. *Development*, 141, 1805-1813.

105. Shah,S., McRae,A.F., et al. (2014) Genetic and environmental exposures constrain epigenetic drift over the human life course. *Genome Res.*, 24, 1725-1733.

106. Patrycja,A.J. and Deborah,M.S. (2019) Nutritional adversity, sex and reproduction: 30 years of DOHaD and what have we learned? *J.Endocrinol.*, 242, T51-T68.

107. Ngo,S. and Sheppard,A. (2015) The role of DNA methylation: a challenge for the DOHaD paradigm in going beyond the historical debate. *Journal of Developmental Origins of Health and Disease*, 6, 2-4.

108. Dunn,E.C., Soare,T.W., et al. (2019) Sensitive Periods for the Effect of Childhood Adversity on DNA Methylation: Results From a Prospective, Longitudinal Study. *Biological Psychiatry*, 85, 838-849.

109. Almstrup,K., Lindhardt Johansen,M., et al. (2016) Pubertal development in healthy children is mirrored by DNA methylation patterns in peripheral blood. *Sci Rep*, 6.

110. Bessa,D.S., Maschietto,M., et al. (2018) Methylome profiling of healthy and central precocious puberty girls. *Clinical Epigenetics*, 10, 146.

111. Thompson,E.E., Nicodemus-Johnson,J., et al. (2018) Global DNA methylation changes spanning puberty are near predicted estrogen-responsive genes and enriched for genes involved in endocrine and immune processes. *Clinical Epigenetics*, 10, 62.

112. Chen,S., Mukherjee,N., et al. (2017) Consistency and Variability of DNA Methylation in Women During Puberty, Young Adulthood, and Pregnancy. *Genet Epigenet*, 9, 1179237X17721540.

113. Gatev,E., Gladish,N., et al. (2020) CoMeBack: DNA Methylation Array Data Analysis for

Co-Methylated Regions. *Bioinformatics*.

114. Niwa,T., Murayama,N., et al. (2015) Regioselective hydroxylation of steroid hormones by human cytochromes P450. *Drug Metab. Rev.*, 47, 89-110.

115. Frick,A., Åhs,F., et al. (2015) Serotonin Synthesis and Reuptake in Social Anxiety Disorder: A Positron Emission Tomography Study. *JAMA Psychiatry*, 72, 794-802.

116. Forstner,A.J., Rambau,S., et al. (2017) Further evidence for genetic variation at the serotonin transporter gene SLC6A4 contributing toward anxiety. *Psychiatr. Genet.*, 27, 96-102.

117. Bleys,D., Luyten,P., et al. (2018) Gene-environment interactions between stress and 5-HTTLPR in depression: A meta-analytic update. *Journal of Affective Disorders*, 226, 339-345.

118. Calabrò,M., Mandelli,L., et al. (2020) Psychiatric disorders and SLC6A4 gene variants: possible effects on alcohol dependence and alzheimer's disease. *Mol.Biol.Rep.*, 47, 191-200.

119. Anonymous https://www.genecards.org/.

120. Vidal,A.C., Murphy,S.K., et al. (2013) Associations between antibiotic exposure during pregnancy, birth weight and aberrant methylation at imprinted genes among offspring. *Int J Obes (Lond)*, 37, 907-913.

121. Talens,R.P., Jukema,J.W., et al. (2012) Hypermethylation at loci sensitive to the prenatal environment is associated with increased incidence of myocardial infarction. *Int J Epidemiol*, 41, 106-115.

122. Murphy,S.K., Adigun,A., et al. (2012) Gender-specific methylation differences in relation to prenatal exposure to cigarette smoke. *Gene*, 494, 36-43.

123. Tobi,E.W., Lumey,L.H., et al. (2009) DNA methylation differences after exposure to prenatal famine are common and timing- and sex-specific. *Hum Mol Genet*, 18, 4046-4053.

124. Barker,D.J.P. and Thornburg,K.L. (2013) Placental programming of chronic diseases, cancer and lifespan: A review. *Placenta*, 34, 841-845.

125. Court,F., Tayama,C., et al. (2014) Genome-wide parent-of-origin DNA methylation analysis

reveals the intricacies of human imprinting and suggests a germline methylation-independent mechanism of establishment. *Genome Res*, 24, 554-569.

126. Pervjakova,N., Kasela,S., et al. (2016) Imprinted genes and imprinting control regions show predominant intermediate methylation in adult somatic tissues. *Epigenomics*, 8, 789-799.

127. Zhao,Y., Sun,H., et al. (2016) Long noncoding RNAs in DNA methylation: new players stepping into the old game. *Cell Biosci*, 6.

128. Colognori,D., Sunwoo,H., et al. (2019) Xist Deletional Analysis Reveals an Interdependency between Xist RNA and Polycomb Complexes for Spreading along the Inactive X. *Molecular Cell*, 74, 101-117.e10.

129. Schertzer,M.D., Braceros,K.C.A., et al. (2019) lncRNA-Induced Spread of Polycomb Controlled by Genome Architecture, RNA Abundance, and CpG Island DNA. *Molecular Cell*, 75, 523-537.e10.

130. Kiontke,K.C., Herrera,R.A., et al. (2019) The Long Non-Coding RNA lep-5 Promotes the Juvenile-to-Adult Transition by Destabilizing LIN-28. *Developmental Cell*, 49, 542-555.e9.

131. Lawson,H., Vuong,E., et al. (2019) The Makorin lep-2 and the lncRNA lep-5 regulate lin-28 to schedule sexual maturation of the C. elegans nervous system. *Elife*, 8.

132. Wang,Z., Yang,B., et al. (2018) lncRNA Epigenetic Landscape Analysis Identifies EPIC1 as an Oncogenic lncRNA that Interacts with MYC and Promotes Cell-Cycle Progression in Cancer. *Cancer Cell*, 33, 706-720.e9.

133. McRae,A.F., Marioni,R.E., et al. (2018) Identification of 55,000 Replicated DNA Methylation QTL. *Scientific Reports*, 8, 1-9.

134. Bongen,E., Lucian,H., et al. (2019) Sex Differences in the Blood Transcriptome Identify Robust Changes in Immune Cell Proportions with Aging and Influenza Infection. *Cell Reports*, 29, 1961-1973.e4.

135. Reinius,L.E., Acevedo,N., et al. (2012) Differential DNA Methylation in Purified Human

Blood Cells: Implications for Cell Lineage and Studies on Disease Susceptibility. *PLoS One*, 7.

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

137. Cotton,A.M., Price,E.M., et al. (2015) Landscape of DNA methylation on the X chromosome reflects CpG density, functional chromatin state and X-chromosome inactivation. *Hum Mol Genet*, 24, 1528-1539.

138. Hovestadt, Volker and Marc Zapatka. (2017) conumee: Enhanced copy-number variation analysis using Illumina DNA methylation arrays version 1.20.0 from Bioconductor. , 2020.

139. Toker,L., Feng,M., et al. (2016) Whose sample is it anyway? Widespread misannotation of samples in transcriptomics studies. *F1000Res*, 5.

140. Arloth J,B.E. (2019) Epigenome analysis of depressed and control subjects.

141. Relton,C.L., Gaunt,T., et al. (2015) Data Resource Profile: Accessible Resource for Integrated Epigenomic Studies (ARIES). *Int J Epidemiol*, 44, 1181-1190.

142. Boyd,A., Golding,J., et al. (2013) Cohort Profile: The 'Children of the 90s'—the index offspring of the Avon Longitudinal Study of Parents and Children. *Int J Epidemiol*, 42, 111-127.

143. Fraser,A., Macdonald-Wallis,C., et al. (2013) Cohort Profile: The Avon Longitudinal Study of Parents and Children: ALSPAC
mothers cohort. *Int J Epidemiol*, 42, 97-110.

144. Golding,J., Pembrey,M., et al. (2001) ALSPAC--the Avon Longitudinal Study of Parents and Children. I. Study methodology. *Paediatr Perinat Epidemiol*, 15, 74-87.

145. Anonymous Geneimprint.

146. Zerbino,D.R., Achuthan,P., et al. (2018) Ensembl 2018. *Nucleic Acids Res*, 46, D754-D761.

147. Phipson,B., Maksimovic,J., et al. (2016) missMethyl: an R package for analyzing data from Illumina's HumanMethylation450 platform. *Bioinformatics*, 32, 286-288.

148. Konopka,T. (2019) umap: Uniform Manifold Approximation and Projection. , 0.2.3.1.

149. R Core Team and R Foundation for Statistical Computing, Vienna, Austria. (2019) R: A language and environment for statistical computing. , 3.6.2.

150. Heyn,H., Li,N., et al. (2012) Distinct DNA methylomes of newborns and centenarians. *PNAS*, 109, 10522-10527.

151. Florath,I., Butterbach,K., et al. (2014) 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.

152. Lutz,P.-. and Turecki,G. (2014) DNA methylation and childhood maltreatment: From animal models to human studies. *Neuroscience*, 264, 142-156.

153. Jenkins,T.G., Liu,L., et al. (2018) Pre-screening method for somatic cell contamination in human sperm epigenetic studies. *Syst Biol Reprod Med*, 64, 146-155.

154. Wu,H., Estill,M.S., et al. (2017) Preconception urinary phthalate concentrations and sperm DNA methylation profiles among men undergoing IVF treatment: a cross-sectional study. *Hum. Reprod.*, 32, 2159-2169.

155. Jenkins,T.G., Aston,K.I., et al. (2015) Intra-sample heterogeneity of sperm DNA methylation. *Mol. Hum. Reprod.*, 21, 313-319.