UC Berkeley

UC Berkeley Electronic Theses and Dissertations

Title

Determinants of Cord Blood DNA Methylation Variability in a Mexican-American Birth Cohort

Permalink

https://escholarship.org/uc/item/0p1563vg

Author

Yousefi, Paul Darius

Publication Date

2016

Peer reviewed|Thesis/dissertation

eScholarship.org Powered by the California Digital Library
University of California

Determinants of Cord Blood DNA Methylation Variability in a Mexican-American Birth Cohort

by

Paul Darius Yousefi

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Environmental Health Sciences

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Nina T. Holland, Chair
Professor Lisa F. Barcellos
Professor Brenda Eskenazi
Professor Alan E. Hubbard

Spring 2016

Determinants of Cord Blood DNA Methylation Variability in a Mexican-American Birth Cohort

© 2016

by Paul Darius Yousefi

Abstract

Determinants of Cord Blood DNA Methylation Variability in a Mexican-American Birth Cohort

by

Paul Darius Yousefi

Doctor of Philosophy in Environmental Health Sciences

University of California, Berkeley.

Professor Nina Holland, Chair

Epigenetic mechanisms, particularly DNA methylation, are a possible link between environmental and biological determinants of health. As the DNA methylome undergoes rearrangement *in utero* and is susceptible to environmental insults, it may be a mechanism explaining the developmental origins of human disease with public health importance. However, the epidemiologic studies needed to identify the role DNA methylation plays mediating environmental exposure disease risk still face several obstacles.

This dissertation addresses knowledge gaps impeding the rigorous adoption of genome-scale measures of site-specific DNA methylation, like the Illumina Infinium HumanMethylation450 (450K) BeadChip®, into epidemiologic study designs. We then investigate the impact of prenatal exposure to polybrominated diphenyl ethers (PBDEs) on DNA methylation of children at birth. PBDEs are a class of flame retardant chemicals widely used in U.S. consumer products over the last 40 years that have previously been associated with adverse neurobehavioral outcomes, obesity, and other effects.

Specifically, we aimed to:

1) Identify and minimize sources of technical variation for site-specific DNA methylation measured by 450K BeadChip assay in epidemiologic studies
2) Characterize sources of biological variation due to host factors (e.g. blood cell composition and sex) in measures of whole blood DNA methylation at birth
3) Determine whether prenatal exposure to PBDEs is associated with differential methylation patterns of CpG sites in umbilical cord blood

Our results identified that the newly proposed All Sample Mean Normalization (ASMN) procedure performed consistently well, both at reducing batch effects and improving replicate comparability compared to several other leading normalization methods. It can be successfully implemented in epidemiologic studies to enhance 450K DNA methylation data preprocessing.

In our examination biological variation, we found that a standard approach in epigenome-wide analysis – minfi white blood cell composition estimation – did not correlate well with white cell counts from newborns (ρ = -0.05 for granulocytes; ρ = -0.03 for lymphocytes), but improved

substantially ($\rho = 0.77$ for granulocytes; $\rho = 0.75$ for lymphocytes) in older children likely due to increasing similarity with minfi's adult reference data as children aged. This suggests that minfi may not currently be appropriate for analysis involving newborns or young children.

Additionally, results on DNA methylation differences by sex identified 3,031 differentially methylated positions (DMPs) and 3,604 sex-associated differentially methylated regions (DMRs) on autosomes that were mostly hypermethylated in girls compared to boys. Our hits were significantly enriched for gene ontology terms related to nervous system development and behavior.

Finally, we investigated the impact of exposure to PBDEs during the highly susceptible prenatal period on DNA methylation of Mexican-American children enrolled in the Center for Health Assessment of Mothers and Children of Salinas (CHAMACOS) at birth. We identified between 6 and 48 DMRs in umbilical cord blood associated with different measures of prenatal PBDE exposure. BDE's-47, -99 and Σ4BDE had fewer (from 6 to 9), mostly hypomethylated DMRs. Prenatal BDE-100 and -153 levels were associated with more DMRs (11 and 48 respectively) and the majority hypermethylated. The PBDE-DMRs we found were located in genes (e.g. *NRBP1, CDH9, NTN1, S100A13*) involved in biologically relevant functions (including axon guidance and tumor suppression) given the health effects observed in association with BDE exposure to date.

TABLE OF CONTENTS

ABSTRACT 1

TABLE OF CONTENTS i
ACKNOWLEDGEMENTS ii

CHAPTER 1 1
AIMS AND ORGANIZATION OF DISSERTATION

CHAPTER 2 3
BACKGROUND AND SIGNIFICANCE

CHAPTER 3 7
DESCRIPTION OF STUDY DESIGN AND METHODS

CHAPTER 4 11
Considerations for Normalization of DNA Methylation Data by Illumina 450K BeadChip Assay in Population Studies

CHAPTER 5 28
Estimation of blood cellular heterogeneity in newborns and children for epigenome-wide association studies

CHAPTER 6 38
Sex differences in DNA methylation assessed by 450K BeadChip in newborns

CHAPTER 7 53
Prenatal PBDE exposure and altered patterns of DNA methylation regions in cord blood

CHAPTER 8 67
CONCLUSION AND SYNTHESIS OF FINDINGS

CHAPTER 9 70
FUTURE DIRECTIONS

DISSERTATION PUBLICATIONS 72

REFERENCES 73

ACKNOWLEDGEMENTS

I want to thank my family: Molly Kauffman, Bill Bonnet, Manny, Laura, Alison, and Ashley Yousefi for their love and support.

I want to particularly thank my dissertation chair and adviser: Dr. Nina Holland. Her mentorship, enthusiasm and support have been an invaluable resource.

I want to thank the other members of my dissertation committee: Dr. Lisa F. Barcellos, Dr. Brenda Eskenazi, and Alan Hubbard for their mentorship and scientific guidance.

Thank you to all members of Nina Holland's lab group for making this research possible, including Karen Huen, Veronica Davé, Gwen Tindula, Jennifer Ames, Maria Escudero-Fung among others.

Thank you to current and former members of Lisa Barcellos' Genetic Epidemiology and Genomics Lab for collaboration and technical assistance including Hong Quach, Gary Artim, Xiaorong Shao, Emon Elboudwarej, Farren Briggs, Milena Gianfrancesco, Brooke Rhead, and Michaela George

Thank you to CHAMACOS participants and all those who helped make the CHAMACOS study possible in addition to Brenda Eskenazi, including Asa Bradman, Kim Harley, Robert Gunier, Jonathan Chevrier, and Katherine Kogut and to name a few.

Thank you to Environmental Mutagenesis & Genomic Society for their support and encouragement in my career as a young scientific investigator.

Finally, I would like thank the following individuals and organizations for financial support of this doctoral research including: Lois A. Rifkin, the UC-Berkeley Public Health Alumni Association, UC-Berkeley Graduate Division, Genetic and Environmental Toxicology Association, National Institute of Environmental Health Sciences and the Environmental Protection Agency.

CHAPTER 1

AIMS AND ORGANIZATION OF DISSERTATION

This dissertation investigated the hypothesis that environmental exposures during prenatal development, such as to flame retardant polybrominated diphenyl ether (PBDE) chemicals, may disrupt fetal patterns of a key epigenetic regulatory mechanism, DNA methylation. To rigorously evaluate this hypothesis, several knowledge gaps first had to be addressed. This included comparison and development of methods to minimize sources of technical variation in genomic-scale measurements of DNA methylation. Additionally, biological sources of DNA methylation variability from host factors like sex and blood cell type composition remained to be adequately characterized, particularly in young children.

The specific aims addressed by this dissertation are:

Aim 1: To identify and minimize sources of technical variation for site-specific DNA methylation measured by 450K BeadChip assay in epidemiologic studies.

Aim 2: To characterize sources of biological variation due to host factors (e.g. blood cell composition and sex) in measures of whole blood DNA methylation at birth.

Aim 3: To determine whether prenatal exposure to PBDEs is associated with differential methylation patterns of CpG sites in umbilical cord blood.

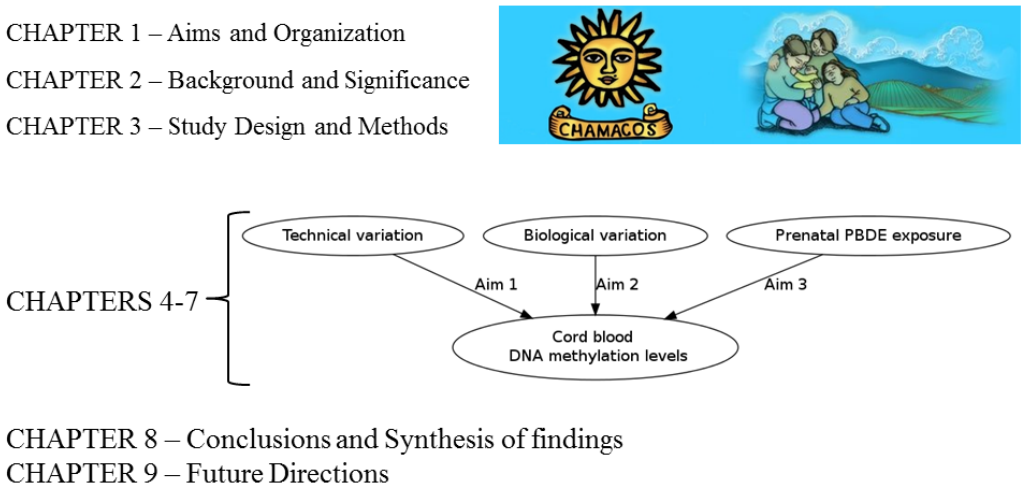


Figure 1. Flow chart of the dissertation's organization, with chapters 4-7 addressing each of the 3 aims.

Chapter 2 discusses the potential for differential health impact from exposures during the prenatal period in line with the "developmental origins of health and disease" hypothesis (DOHaD). Further, the evidence supporting DNA methylation as a possible mediator of such early life exposures is presented, as our the current methodological shortcomings for research in this area.

Chapter 3 presents the CHAMACOS study design and methods used in this dissertation. In brief, the CHAMACOS study is a longitudinal birth cohort designed to assess the health effects of pesticides and other exposures on child growth and development. Measurement of DNA methylation was performed in umbilical cord blood using the Illumina Infinium HumanMethylation450 (450K) BeadChip. PBDE measurements were performed in maternal serums samples (collected at 26 weeks gestation and delivery) using mass spectrometry.

Chapter 4 reports on our comparative assessment of 10 normalization procedures – including the newly proposed All Sample Mean Normalization (ASMN) method - to reduce technical variability and batch effects in 450K analysis in the context of a large population study. We also examined the performance of normalizations methods in combination with correction for the two types of Infinium chemistry typically utilized in 450K array pre-processing.

Chapter 5 compares a standard approach in epigenome-wide analysis – minfi reverse-estimation of white blood cell composition from 450K BeadChip data– to microscopic differential cell count (DCC) using histologically stained blood smear slides. We analyze samples taken both at birth and in 12 year-old children to evaluate the impact of age on these estimates.

Chapter 6 presents the results of our characterization of the contribution of a key biological host factor, child sex, to DNA methylation in umbilical cord blood across the genome. In addition to performing analysis at individual CpG sites considered independently, we implement newly development methods for identifying more functionally relevant differentially methylated regions. To interpret and contextualize our findings, we also perform pathway enrichment analysis.

Chapter 7 describes our DNA methylome wide assessment of the relationship between prenatal PBDE exposure and DNA methylation in umbilical cord blood. We perform de novo assessment to identify all potential confounding factors, resulting in adjustment for maternal years living in the United States and 450K BeadChip analysis batch. Again, we employ region-based estimation procedures in addition to considering CpGs individually.

Chapter 8 summarizes our findings, highlighting the knowledge gaps in the current literature of epidemiologic analysis of DNA methylation and how our results have sought to address them.

Chapter 9 suggests future research directions for building on the findings of our analysis estimating white blood cell composition, and expounding on the results of our observed associations between DNA methylation and both sex and PBDE exposure.

CHAPTER 2

BACKGROUND AND SIGNIFICANCE

Epigenetic markers, such as DNA methylation, represent important regulatory mechanisms for biological systems and are a potential additional layer in the central dogma of molecular biology.[1] DNA methylation has been the most extensively investigated epigenetic mechanism and refers to the methylation of a cytosine (C) base at its 5th carbon followed by a guanine (G) base in the DNA code, called a CpG site.[2] Such markers likely play significant roles in the etiology of common diseases important to public health.[3] Since DNA methylation can be altered by external factors and is readily assayable, it has potential to yield insights into the environmental contributions to disease that have been difficult to address systematically.[4] However, the epidemiologic studies needed to identify the role DNA methylation plays mediating environmental exposure on later-life disease risk still face several obstacles. The study proposed is designed to address some of the knowledge gaps impeding the rigorous adoption of genome-scale measures of site-specific DNA methylation into epidemiologic study designs.

PBDE exposure

One class of persistent organic pollutants (POPs), polybrominated diphenyl ethers (PBDEs), is of particular concern as Americans - and Californians especially - have been widely exposed to these compounds. They have been shown to have health effects in children, and have been associated with altered DNA methylation in early studies of global methylation of repetitive elements.[5] Specifically, PBDEs are brominated organic compounds that have been widely used as flame retardants in consumer products such as furniture, infant products, and electronics.[6] Originally issued in the 1970's, California's Technical Bulletin (TB 117) required that furniture, baby products, and other items resist open flames. The penta-BDE technical mixture (comprised of congeners BDEs -47 -99 -100 and -153), the compounds most commonly used to comply with TB-117 in products containing foam, were phased out in 2004 but continue to leach into the environment from older household items.[7] Children living in California have significantly higher serum PBDE levels compared to children living in other areas of the United States,[8,9] as well as children from Mexico and Europe.[10,11] Exposure to penta-PBDE, has been associated with poorer neurodevelopment,[12-15] effects on thyroid hormone,[16,17] decreased fertility,[18] and measures of obesity.[19]

DNA methylation and disease etiology

A growing number of human diseases have been associated with aberrant patterns of DNA methylation, including cancers, neurodegenerative diseases, diabetes, and other metabolic disorders.[20-25] Cancer cells are characterized by global hypomethylation and local hypermethylation of regulatory regions in tumor suppressor genes, a pattern that is also emerging in the etiology of other chronic disease.[20,22,26,27] Additionally, the early embryonic epigenome undergoes a wave of demethylation followed by *de novo* methylation, making pregnancy a particularly sensitive time during which methylation patterns can be dysregulated.[28,29] Disruption of these methylation patterns has been linked to aging and chronic disease development.[30-32] In line with the "developmental origins of health and disease" hypothesis (DOHaD) which posits that adverse intrauterine environments may affect health later in life,[33,34] maternal exposures, nutrition and other factors as hyperlipidemia, inflammation, and other metabolic imbalances

likely influence the fetal environment and impact programming of later life health status.[35] As such, epidemiologic studies investigating epigenetic perturbation as near as possible to the prenatal period are a high priority for determining the role of DNA methylation in the etiology of common diseases.

DNA methylation

The human genome contains about 30 million CpG sites distributed throughout several gene regions referred to as CpG islands, shores, shelves, and gene bodies. CpG islands are stretches of DNA with a high frequency of CpG dinucleotides that often occur in proximity to gene promoter regions.[36] DNA methylation changes across these functional regions may impact gene expression.[37,38] The amount and pattern of DNA methylation is modulated during the prenatal period and may vary by tissue and cell type.[39,40] Gain of DNA methylation (hypermethylation) can lead to gene silencing while loss (hypomethylation) leads to overexpression.[1]

Measurement of DNA methylation in epidemiologic studies

Several technologies, including next generation sequencing and genome-wide arrays, are currently available to study the DNA methylome.[2] However, sequencing technologies can be prohibitively expensive for use with population-based epigenome-wide association studies (EWAS), which often require analysis of hundreds of samples in large datasets over multiple time points. The Illumina Infinium HumanMethylation450 (450K) BeadChip has emerged as one of the preferred methodologies in epidemiologic analysis of DNA methylation because of its optimal combination of genome-wide coverage (99% of RefSeq genes), comprehensive representation of functional gene sub-regions, good reproducibility across other platforms (r=0.88 with Pyrosequencing),[41-43] and relative affordability. Before genome wide DNA methylation data can be successfully incorporated into epidemiologic studies, it is critical to minimize sources of technical variation and bias that increase with sample size. Experiments involving hundreds of samples must be run in several batches across a long time span, potentially exacerbating variation in instrumentation and assay chemistry. To date, several sources of technical variation with 450K BeadChip data have been identified, including background signal and performance differences between two types of assay chemistries used. Differences between the measurement of the two colored probes (red and green), including hybridization efficiency and chip scanning properties, can also introduce noise to methylation results. While several computational methods have been developed to remove such color channel variation,[44-46] these methodologies have yet to have their performance compared systematically.

Biological variation in DNA methylation by white blood cell type

Whole blood is a desirable matrix to use for population based EWAS as it is readily available and has been obtained for many human studies with a wide variety of initial aims.[40,47-49] Since DNA methylation may vary by cell type, analyses involving health outcomes or exposures that also co-vary with cell type may be confounded.[48-50] Several approaches have been proposed to address confounding bias in EWAS due to varying white blood cell type composition. In practice, the most common has been to estimate the relative percentages of different cell types and adjust for these as covariates in statistical analysis. The most reliable white blood cell count is performed either by automated hematology analyzer, as part of a complete blood count (CBC),

or retrospectively by microscopic differential cell count (DCC) using histologically stained blood smear slides.

However, since many epidemiologic studies do not have direct white blood counts, there is growing interest in computational approaches that estimate cell type proportions based on DNA methylation data. A method by Jaffe and Irizarry[50] (referred to here as the minfi method) that was adapted from the work of Houseman et al.[51,52] has seen widespread use. The minfi approach is tailored for use with 450K BeadChip data and uses a cell-sorted 450K reference dataset as the core of the estimation method. That reference data is taken from a study of six middle-aged (mean age=38±13.6) Swedish men[53] that may not be generalizable to other populations such as newborns which are known to have vastly different immune cell profiles than adults.[54-56] The cell type estimates produced by minfi also remain to be systematically validated against a gold standard cell count, such as CBC or DCC, which is necessary before they can used reliably in studies of DNA methylation in infants and children.

Biological variation in DNA methylation by sex

Beyond tissue type, host factors such as sex and age also contribute to inter-individual differences in DNA methylation. Previous studies of DNA methylation using the 27K BeadChip have reported autosomal differentially methylated positions (DMPs) or CpG sites with varying methylation between males and females, providing evidence that it will be important to adjust for sex in analysis of methylation data.[57-59] However, these studies did not account for the existence of non-specific probes for autosomal CpGs that cross-react with CpGs on sex chromosomes, thereby yielding false positives.[60] Recently, McCarthy et al. published a meta-analysis of 76 studies all using an earlier version of the Infinium methylation array (27K BeadChip) to identify sex-associated autosomal DMPs across specimens from multiple tissue types from adults and children.[61] After excluding the sex-biased cross-reactive probes, they identified 184 DMPs that were associated with sex. However, the work by McCarthy and colleagues had key limitations: they used the 27K BeadChip instead of the much larger 450K BeadChip, failed to account for differences in cell composition, and only attempted to identify individual DMPs rather than differentially methylated regions (DMRs). Thus, a rigorous genome-wide characterization sex differences in whole blood DNA methylation still remains to be performed.

Exposure and DNA methylation

The ultimate goal in evaluating and minimizing all sources of technical and biological variation, is to begin to tease out the impact that exogenous exposures have on the DNA methylome as a potential mediating mechanism of later life disease. Emerging literature suggests that exposure to POPs can contribute to altered DNA methylation. For example, polychlorinated biphenyls (PCBs) and several types of organochlorines (OCs) have been associated with DNA methylation in LINE-1 and ALU repetitive elements, used as biomarkers of DNA methylation globally across the genome, in four independent cross-sectional studies of adults.[62-65] Exposure to dichlorodiphenyl trichloroethane (DDT) has been seen to disrupt DNA methylation at several CpG islands in the hypothalamus of Wistar rats.[66] Additionally, alterations of DNA methylation patterns in the sperm of male rodent offspring following maternal exposure to DDT persisted for three generations, demonstrating trans-generational epigenetic reprogramming.[67] Fewer studies have looked specifically at PBDE exposure. One study in rats found that pre- and perinatal exposure to BDE-47 was associated with lower methylation in the frontal brain lobe at the L1Rn repetitive element.[68] In the only study done in humans, we found that maternal PBDE serum

concentrations were significantly associated with DNA methylation in the LINE-1 and ALU repetitive elements when accounting for co-exposure to DDT.[5] However, no studies to date have explored association between PBDE exposure and site specific DNA methylation across the genome.

Summary

DNA methylation is a fundamental mechanism in the regulation of biological development and function. Understanding how DNA methylation varies in response to exposure, especially in pregnancy when it is particularly susceptible to disruption, has the potential to provide substantial insight into the environmental contributions to common diseases. However, before epidemiologic studies can be conducted rigorously, key knowledge gaps concerning sources of technical and biological variation need to be addressed. The goal of this dissertation is to address several of these gaps and proceed to examine the effects of prenatal exposure to PBDEs on DNA methylation in newborns, reflecting environmental impact during one of the most vulnerable periods in child development.

CHAPTER 3

DESCRIPTION OF STUDY DESIGN AND METHODS

CHAMACOS study design

The Center for the Health Assessment of Mothers and Children of Salinas (CHAMACOS) study is a longitudinal birth cohort study of the effects of exposure to pesticides and environmental chemicals on the health and development of Mexican-American children living in the agricultural region of Salinas Valley, CA. Detailed description of the CHAMACOS cohort has previously been published.[69,70] Briefly, 601 pregnant women were enrolled in 1999-2000 at community clinics and 527 liveborn singletons were born. Follow up visits occurred at regular intervals throughout childhood. Eligible women were ≥18 years of age, <20 weeks gestation at enrollment, English or Spanish speaking, Medi-Cal eligible and planning to deliver at the county hospital. CHAMACOS mothers were primarily young (M=25.6±5.3 years), low-income, Mexican-born, Spanish-speaking women who were farm workers themselves or lived with farm workers in Salinas Valley, CA at the time of enrollment. Women were interviewed at ~13 and 26 weeks gestation, and after delivery. All human subject activities for this population were approved by the Committee for the Protection of Human Subjects at the University of California, Berkeley (Protocol ID's: 2010-03-949 and 2010-01-620). However, the current study did not conduct any additional work involving the use of new human subjects.

Blood collection

Maternal blood samples were obtained by a trained phlebotomist and cord blood specimens were collected by hospital staff at time of delivery. Blood samples were collected in vacutainers with and without heparin (green- and red-top), separated into clots, buffy coats and other blood components, aliquotted into several sub-samples, and stored at the appropriate temperature (-80°C) under stringent Standard Operating Procedures (SOPs).

DNA isolation

DNA isolation was performed from banked umbilical cord blood clot samples using QIAamp DNA Blood Maxi Kits (Qiagen, Valencia, CA) according to modified manufacturer's protocol.[71] All samples were checked for DNA quality and quantity by Nanodrop 2000 Spectrophotometer (Thermo Scientific, Waltham, MA). Those with good quality (260/280 ratio exceeding 1.8) were normalized to a concentration of 50ng/ul.

450K BeadChip DNA methylation analysis

DNA aliquots of 1ug were bisulfite converted using Zymo Bisulfite conversion Kits (Zymo Research, Orange, CA). DNA samples were whole genome amplified, enzymatically fragmented, purified, and applied to the 450K BeadChips according to the Illumina methylation protocol.[41,42] 450K BeadChips analyzed using the Illumina Hi-Scan system. DNA methylation was measured at 485,512 CpG sites. Each 450K BeadChip can fit n=12 samples in total, and these chips are typically run combined onto plates in sets of 8, for a batch of n=96 samples. The n=378 cord blood samples were run on 36 distinct BeadChips across 5 plates. BeadChips included on the same plate (up to n=8 BeadChips per plate) were assayed simultaneously, and time between plate runs was approximately one week using the same batch of all reagents and chips.

450K BeadChip data extraction and preparation

Control probe data extraction was performed using using Illumina GenomeStudio software (version XXV2011.1, Methylation Module 1.9). This dataset contains raw signal observations for all probes included as controls in the design of the 450K assay. Such control values include negative controls (for background subtraction), extension controls, staining controls, bisulfite controls, and n=93 normalization control probes among others. The normalization control probe pairs are targeted to non-variable regions of stable housekeeping genes and are the observations used to perform a class of color channel normalizations we refer to as reference normalization (RN)-factor based methods.

Sample data were also extracted using Illumina GenomeStudio. This provides raw intensities for both red and green color channels, detection p-values as a measure of assay performance, and β's calculated from raw signals at 485,512 assayed probes. Data cleaning performed prior to evaluation of different color channel normalization procedures included background correction of raw signal intensities according to Illumina recommendations.

For chapters 5-7, raw signal intensities were normalized for color channel bias using ASMN, the best performing method identified in Chapter 4. Differences in the two types of Infinium assay chemistry were then addressed by Beta Mixture Quantile (BMIQ) normalization.[72] Samples were retained only if 95% of sites assayed had detection P> 0.01. Probes where 95% of samples have detection P>0.01 are also dropped. Sites with annotated probe SNPs and with common SNPs (minor allele frequency >5%) within 50bp of the target identified in the MXL (Mexican ancestry in Los Angeles, California) HapMap population and X-chromosome cross-reactive probes (n=29,233) identified by Chen and colleagues[73] were excluded from analysis. Methylation values at all sites were logit transformed to the M-value scale to better comply with modeling assumption.[74] Additionally, for chapter 7 we removed methylation observations that were greater or less than three times the interquartile range for a given probe to reduce the potential to extreme outlying methylation values to drive associations in statistical models.

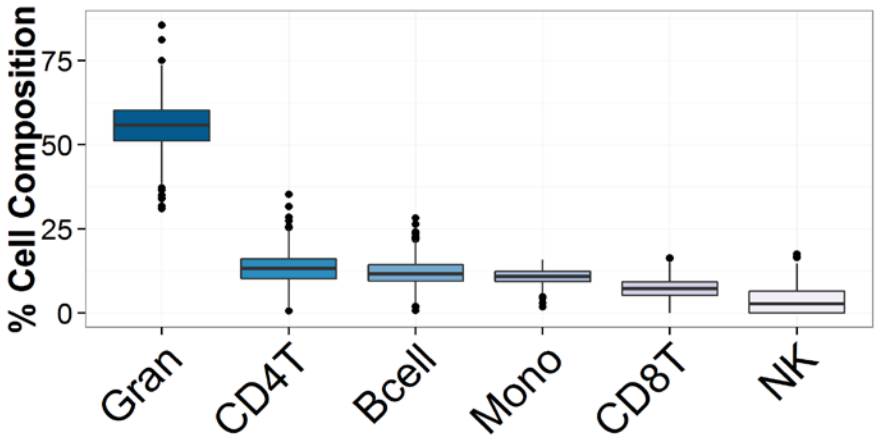


Figure 1. White blood cell type estimates by minfi for CHAMACOS newborns.

Differential white blood cell counts (DCC)

Whole blood smear slides were prepared from heparinized blood using the push-wedge blood smearing technique [75] and stored at -20°C until staining with DiffQuik® kit. This staining highlights cytoplasmic details and neurosecretory granules to characterize the differential white blood count. The slides were fixed for 15 minutes at room temperature, stained in the basophilic and eosinophilic dyes for five seconds each and washed after each staining period.Slides are

scored to count the frequency of five types of easily visually identifiable blood cells (lymphocytes, monocytes, neutrophils, eosinophils, and basophils) by Zeiss Axioplan light microscope with 100× oil immersion lens. Scoring is conducted at the perceived highest density of white blood cells using the standard battlement track scan method. At least 100 cells are scored for each slide following validation of reproducibility by the repeated scoring of 5 sets of 100 cells from the same slide.

Minfi cell count estimation

Raw 450K BeadChip signal intensities (stored as IDAT files) were read into the *minfi* (v1.10.2) Bioconductor R package[76] using the *read.450k.exp* function. Estimation of six different white blood cell types (CD8+ T and CD4+ T lymphocytes, CD56+ natural killer cells, CD19+ B cells, CD14+ monocytes, and granulocytes) was performed using the default implementation of the *estimateCellCounts* function. Briefly, this function takes a user-supplied target 450K BeadChip dataset, combines that with a cell-sorted reference dataset available in the *FlowSorted.Blood.450k* Bioconductor package (v1.2.0)[53,77] and normalizes the combined data. The reference dataset, n observations from $i = 6$ subjects at each of $j = 6$ different separated cell types, is subset to the 600 most cell type informative CpGs, called S_0. Individual level predictions of cell type proportion, Γ^*, are fit in the corresponding user submitted dataset, S_1, using the coefficients estimated in S_0 ($\tilde{B}_0$) by the following equation:

$$\Gamma^* = \left(\tilde{B}_0^T \tilde{B}_0\right)^{-1} \tilde{B}_0^T S_1$$

This yields an estimated percentage of all six white blood cell types (Figure 1) for every sample submitted in S_1. Detailed description has previously been published.[50,51]

PBDE measurement

Maternal serums samples (collected at 26 weeks gestation and delivery) were analyzed for PBDEs by Dr. Andreas Sjödin at CDC. PBDE congeners were measured using gas chromatography isotope dilution high resolution mass spectrometry (GC-IDHRMS).[78] We focussed on the four congeners comprising the penta-BDE mixture: BDE-47, -99, -100, and -153. These were detected in almost all mothers (>97% detection frequency). As previously reported, BDE - 47 had the highest mean concentration of the congeners measured (Figure 2).[9,14,79] PBDE concentrations are expressed on a serum lipid basis (nanograms per gram lipids). Total serum lipid concentrations were determined based on the measurement of triglycerides and total cholesterol using standard enzymatic methods (Roche Chemicals, Indianapolis, IN).[80] Data below the limit of detection (LOD) with no detectable signal were imputed from a log-normal

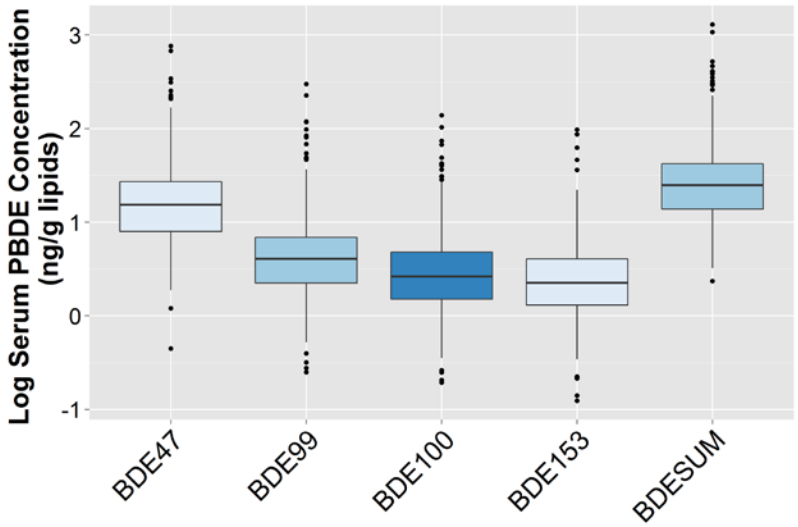


Figure 2. Prenatal PBDE serum levels in CHAMACOS mothers.

probability distribution.[81] QA/QC procedures included the use of blanks and spiked quality control samples in each set of unknowns.

CHAPTER 4

Considerations for Normalization of DNA Methylation Data by Illumina 450K BeadChip Assay in Population Studies.

ABSTRACT

Analysis of epigenetic mechanisms, particularly DNA methylation, is of increasing interest for epidemiologic studies examining disease etiology and impacts of environmental exposures. The Infinium HumanMethylation450 BeadChip® (450K), which interrogates over 480,000 CpG sites and is relatively cost effective, has become a popular tool to characterize the DNA methylome. For large scale studies, minimizing technical variability and potential bias is paramount. The goal of this paper was to evaluate the performance of several existing and novel color channel normalizations designed to reduce technical variability and batch effects in 450K analysis from a large population study.

Comparative assessment of 10 normalization procedures included the GenomeStudio® Illumina procedure, the lumi smooth quantile approach, and the newly proposed All Sample Mean Normalization (ASMN). We also examined the performance of normalizations in combination with correction for the two types of Infinium chemistry utilized on the 450K array. We observed that the performance of the GenomeStudio® normalization procedure was highly variable and dependent on the quality of the first sample analyzed in an experiment, which is used as a reference in this procedure. While the lumi normalization was able to decrease batch variability, it increased variation among technical replicates, potentially reducing biologically meaningful findings. The proposed ASMN procedure performed consistently well, both at reducing batch effects and improving replicate comparability.

In summary, the ASMN procedure can improve existing color channel normalization, especially for large epidemiologic studies, and can be successfully implemented to enhance a 450K DNA methylation data pipeline.

Introduction

Epigenetic mechanisms regulate gene expression without changes in DNA sequence and include DNA methylation, histone modifications, and non-coding RNAs .[82-84] Growing evidence shows that epigenetics may be an interface through which environmental exposures affect gene expression and adverse health.[85] DNA methylation, an addition of a methyl group at the carbon-5 position of cytosine (5mC) in CpG dinucleotides, is the best studied epigenetic mechanism. Several technologies, including next generation sequencing and genome-wide arrays, are currently available to study the DNA methylome.[2] However, sequencing technologies can be prohibitively expensive for use with population-based studies, which may require analysis of hundreds of samples in large datasets over multiple time points. Illumina's 450K Methylation array has emerged as one of the preferred methodologies to study DNA methylation because of its optimal combination of genome-wide coverage (99% of RefSeq genes), comprehensive representation of functional gene sub-regions, good reproducibility across other platforms (r=0.88 with Pyrosequencing)[41,43] and relative affordability.

Before sources of biological variability in DNA methylation can be accurately assessed, it is critical to minimize technical variance and bias. Experiments involving hundreds of samples need to be run in several batches across a long time span, potentially exacerbating variation in instrumentation and assay chemistry. Differences between the measurement of the two colored probes (red and green), including labeling hybridization efficiency and chip scanning properties, can also introduce noise to methylation results. The Illumina proprietary software package (GenomeStudio) adjusts for this variability of color signals across an experiment, which we refer to as the Illumina First Sample Normalization (IFSN). In addition, other normalization methodologies have recently been proposed, including smooth quantile normalization from the lumi R package[44] and other pipelines drawing on its infrastructure.[45,46]

Another class of adjustment that has received attention in the literature addresses the two different 450K Infinium array chemistries: Infinium I, which was previously implemented on an older Illumina methylation 27K array, and Infinium II, which was added as coverage expanded for the 450K array. Recent studies have demonstrated that the signals from the Infinium I and II assays are likely not completely comparable: Infinium I has a broader dynamic range of methylation values, called β's, that tend to be more stable and reproducible in comparison to Infinium II,[45,86] potentially introducing a source of bias based on the type of probe used. Several correction and normalization methodologies have been proposed to adjust for differences between the two 450K Infinium chemistries including peak-based correction (PBC),[86] subset quantile normalization (SQN and SWAN),[45,87] and beta-mixture quantile normalization (BMIQ).[88] The PBC approach has been criticized in two recent publications for poor performance when its strong assumptions of bi-modality in β distributions are not met.[45,87] However, a recent evaluation of the other available normalizations between Infinium I and II chemistries (SQN, SWAN, BMIQ) showed them to be comparable.[88] While initial assessment of each of these normalizations has been conducted,[88,89] including evaluation of reductions in batch effects, the sample sizes used in these publications (ranging from 6-85 analyzed on 1-8 chips) have not been sufficient to detect the type of batch variability likely to occur in large population studies.

A complete 450K data preparation pipeline for epidemiologic analysis ideally includes several distinct components, including: subtraction of background signal, color channel normalization, checks for bisulfite conversion and extension efficiency, removal of poor

performing CpG's and SNP associated probes, and adjustment for Infinium chemistry. In the current study we conduct a focused assessment of the performance of color channel normalization procedures, one key aspect of 450K data preparation. Our study evaluates 10 procedures: two existing normalizations (IFSN and lumi), several variations on the method used by IFSN normalization, called reference factor (RN) based normalizations, and proposed here a new optimized All Sample Mean Normalization (ASMN) procedure. Our analysis uses a large dataset of 432 samples (36 chips/5 plates) to identify which procedures most effectively minimize technical variation in population-based studies.

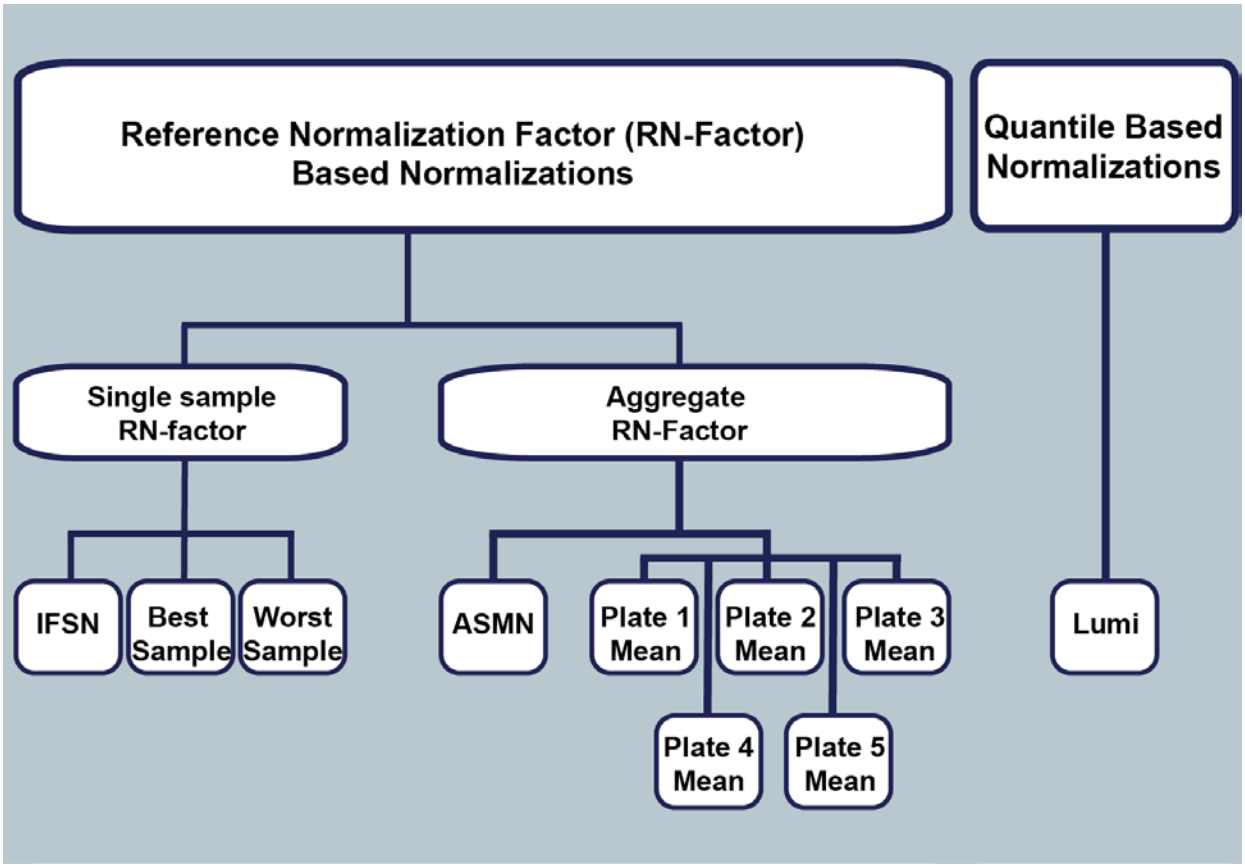


Figure 1. Flow chart of ten color channel normalization procedures implemented. Nine of those procedures were reference normalization factor (RN-factor) based methods that use the n=93 normalization control probes assayed in every sample on the 450K chip for adjustment. Of the RN-factor based methods, three methods used the RN-factors from a single sample: the Illumina first sample normalization (IFSN), the best performing sample normalization, and the worst performing sample normalization. The remaining six RN-factor based procedures use aggregated RN-factors across different groups of samples, including the mean RN-factors for each plate of the experiment (Plates1-5 Means) and the all sample mean normalization (ASMN) that uses the mean RN-factors for all experimental samples. The remaining normalization, the lumi procedure, uses a quantile-based methodology instead of RN-factors.

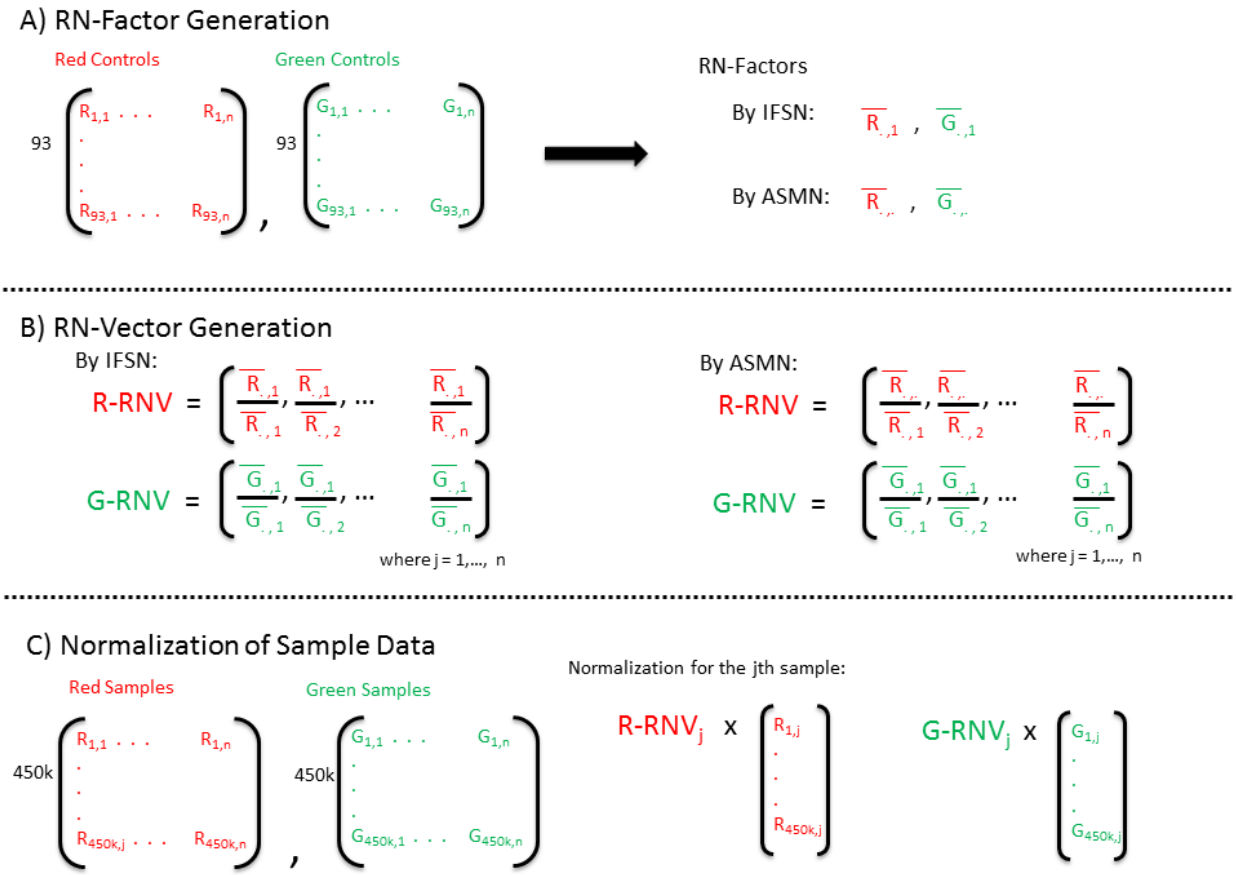


Figure 2. Reference normalization factor (RN-factor) based color channel normalization for the 450K methylation array. (A) The 450K chip includes n=93 normalization control probes in both assay colors (red and green). The mean values of these sites are used to create RN-factors for normalizing both color channels over all samples (i.e. an experiment). The Illumina first sample normalization (IFSN) method uses the first sample's mean red and green control probes as RN-factors ($\bar{R}_{.,1}$ and $\bar{G}_{.,1}$). The all sample mean normalization (ASMN) method instead uses the mean read and green control probes taken across all control sites and all samples in a given experiment ($\bar{R}_{.,.}$ and $\bar{G}_{.,.}$) as RN-factors. (B) A set of sample-wise normalization values, taken as the ratio of the RN-factor to each sample's mean control probe values, is then computed. This results in a vector of length n normalization values for each color channel (R-RNV and G-RNV). (C) Color channel normalization of sample data occurs by multiplying the each of the jth sample's red and green signals by the jth normalization value from the corresponding RN-vector (where j=1,2,…, n).

Results

To assess color channel normalization performance, 10 procedures were implemented on 450K data from a large epidemiologic cohort. These procedures fell into two distinct methodological categories: 1) reference normalization (RN)-factor based and 2) quantile based methods (Figure 1). The first category included nine variations of RN-factor based procedures and the second category was represented by the lumi smooth quantile normalization.

RN-factor based normalizations utilize the mean values from the red and green normalization control probes included on the 450K chip as RN-factors in their adjustment (Figure2A). These RN-factors are used to compute two vectors of length n (RN-vectors), containing the ratio of each sample's mean red and green control probe values to that of the RN-factor of the same color (Figure 2B). Sample normalization occurs by multiplying the *j*th sample's red and green signals by the corresponding elements from the red and green RN-vectors (Figure 2C and Methods).

Among the RN-factor based normalizations, procedures differed by which control probe observations were used to calculate the RN-factors (Figure1). There were two groups of RN-factor procedures: a) those using only the control probe values from a single sample (IFSN, best sample, worst sample) and b) those using an RN-factor aggregated across groups of samples (mean by each of 5 assay plates, ASMN). Figure 2 shows each step of RN-factor based normalization for both a single sample (IFSN) and an aggregate procedure (ASMN).

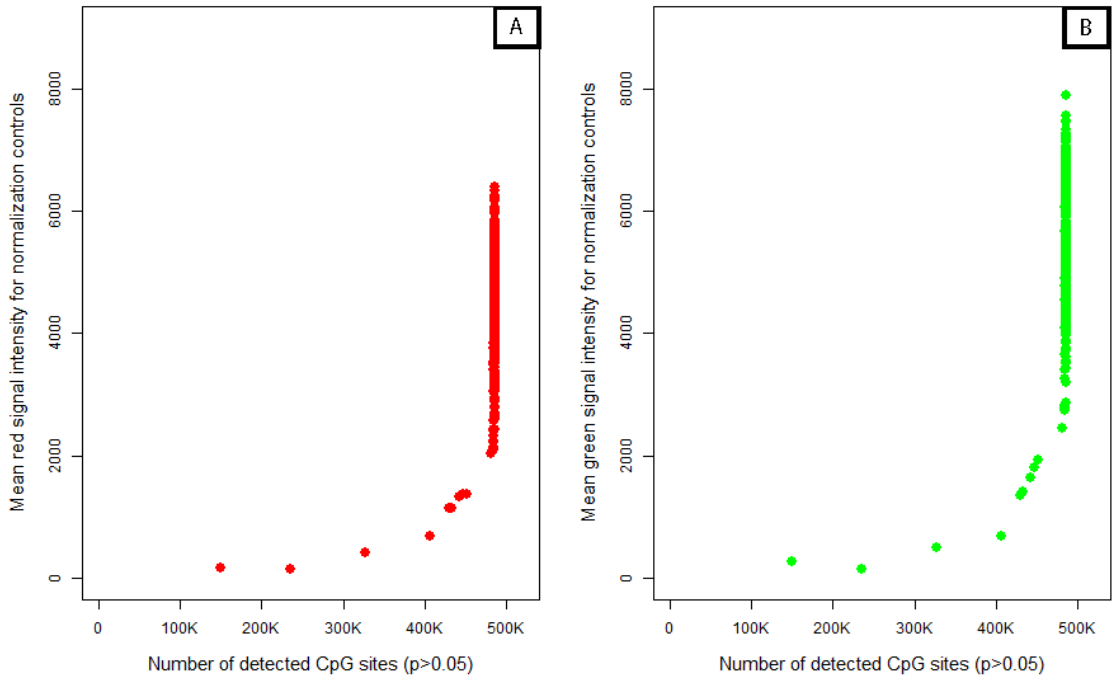


Figure 3. Plot of mean red (A) and green (B) signal intensity of normalization control probes (n=93) by number of detected CpG sites in the 450K array sample data (n=432). For both color channels, samples with lower intensity readings in their normalization control probes tended to have more poor performing CpG sites in their samples.

Performance of the normalization procedures was evaluated by three criteria. First, we assessed *the stability* of RN-factor based normalizations when using RN-factors from samples of varying quality, or when using RN-factors aggregated across batches (i.e. assay plates) or an entire experiment (i.e. ASMN). The other two criteria included evaluation of *repeatability of technical replicates* and *reduction in batch variation*.

RN-factor based normalization stability

The majority of samples from our cohort proved to be of good quality with less than one percent of CpG sites with detection p-values equal or greater than 0.05. However, nine of the 432 samples were considered of lower quality (>1% of CpG sites with detection p-values≥0.05). When we plotted the signal intensity of normalization control probe against quality of methylation calls (measured by number of detectable CpG sites), we found that samples with low red and green control signals also had lower quality methylation calls (Figure 3). The correlation between control probe signal intensity and number of detectable CpG sites was 0.76 ($p<0.0005$) for both red and green signals. If according to the Illumina IFSN algorithm (Figure 2) the first sample on which the entire experiment is normalized happens to be one of low quality, the overall results and interpretation of the data may be negatively affected. Thus, ASMN was developed to increase normalization stability and robustness by non-arbitrarily drawing on

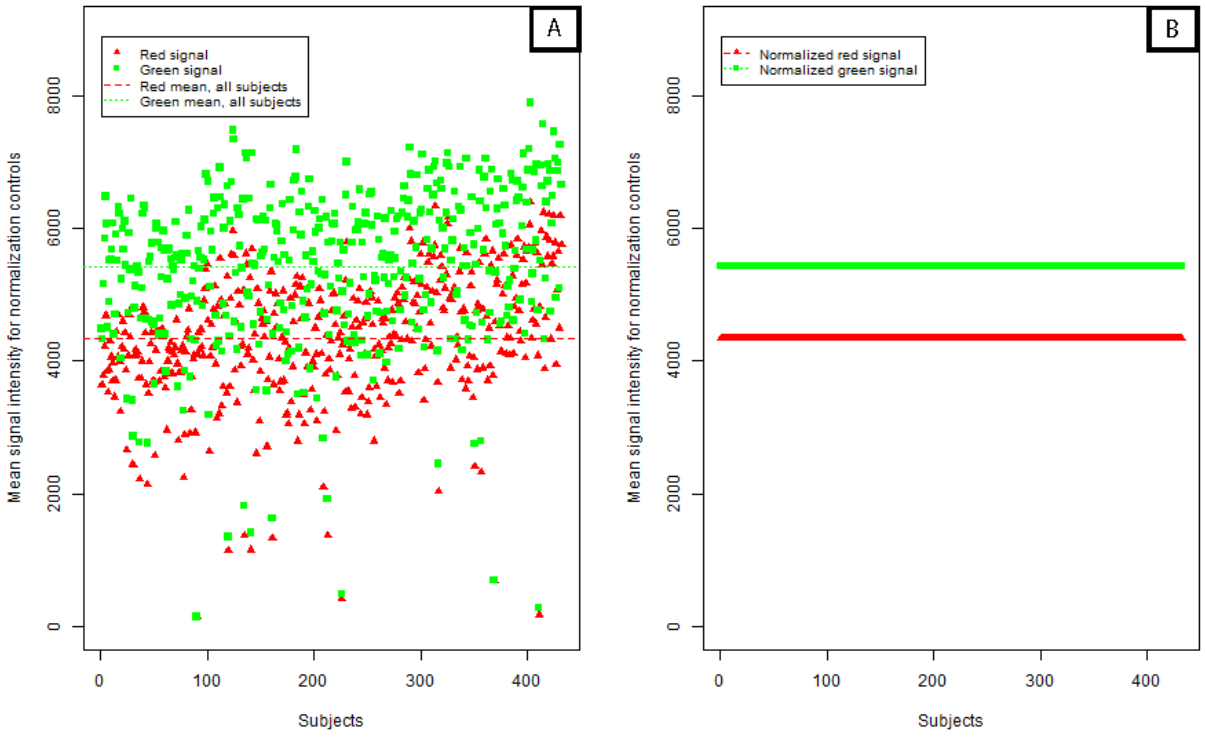


Figure 4. Mean control probe color signal intensity before and after normalization. (A) Distribution of mean green and red normalization controls (93 controls per signal color per sample) as included in the 450K chip over 432 DNA samples. Each point, red triangle or green square, represents the average of the normalization controls for that signal color per sample prior to implementation of color channel normalization. (B) Following adjustment using a reference normalization (RN) –factor based normalization, the average normalization controls for all samples are 'forced' to be the same level, making observations across samples comparable. Here, ASMN normalization was performed which uses the mean red and green signal for all samples for adjustment.

observations from all Illumina internal controls and study samples (described in detail in Methods).

Although we observed a positive association between control probe signal intensity and number of detectable CpG sites, this relationship appeared to exhibit a threshold effect (Figure 3). Samples with fewer numbers of detectable CpG sites also had lower mean red and green control probe values, but increases in probe signal intensities above 2,000 for red and 3,000 for green did not appear to contribute to additional gains in CpG detection. While both the number of detectable CpG sites and the mean control probe signal intensity provide information regarding assay quality, only the former was a measure designed for this purpose. This makes it difficult to distinguish what constitutes a 'better' sample or a 'better' mean control probe value from those above the control signal threshold. We implemented the ASMN procedure to draw from the central tendency of this distribution rather than the tail, since we did not have convincing evidence to prefer higher mean control probe values above the threshold. Further, drawing from the center of this distribution made the ASMN more stable and less susceptible to variation in sample quality. To confirm that the mean was an appropriate measure of central tendency, we also performed the normalization using median RN-factors, but obtained similar results (data not shown).

We found little available rationale, including information in the Illumina reference manual, to support the preferential use of control probe data from the first sample over other samples beyond convenience.[90] The IFSN approach carries an unstated assumption: the control probe values from any sample should perform equally well at reducing color bias and batch variation. This is not always the case, for instance Figure 4A shows a broad range of variability in the normalization control values among 432 samples analyzed in this study. Furthermore, this variability also suggests that the particular sample used for normalization may affect the normalization quality of all samples. After performing normalization, RN-factor based methods bring all data observations to the same scale. Figure 4B illustrates this effect by showing that after normalization, the previously dispersed normalization control values (Figure 4A) become standardized to the values of the red and green RN-factors.

In Figure 5, we compared the normalized β's given an unadjusted β

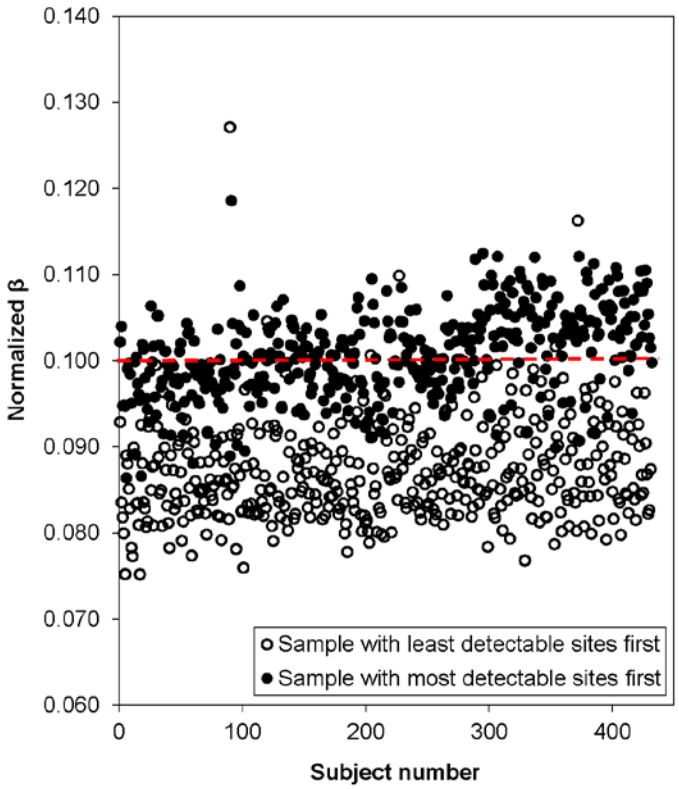


Figure 5. Plot of normalized DNA methylation (β's) given an unadjusted β of 0.1 (Signal A=5000 and Signal B=570) for all 432 samples. Open circles represent data normalized using the sample with the least detectable sites (sample 411, the lowest quality sample). Filled circles were normalized using the sample with the most detectable sites (sample 355, the highest quality sample).

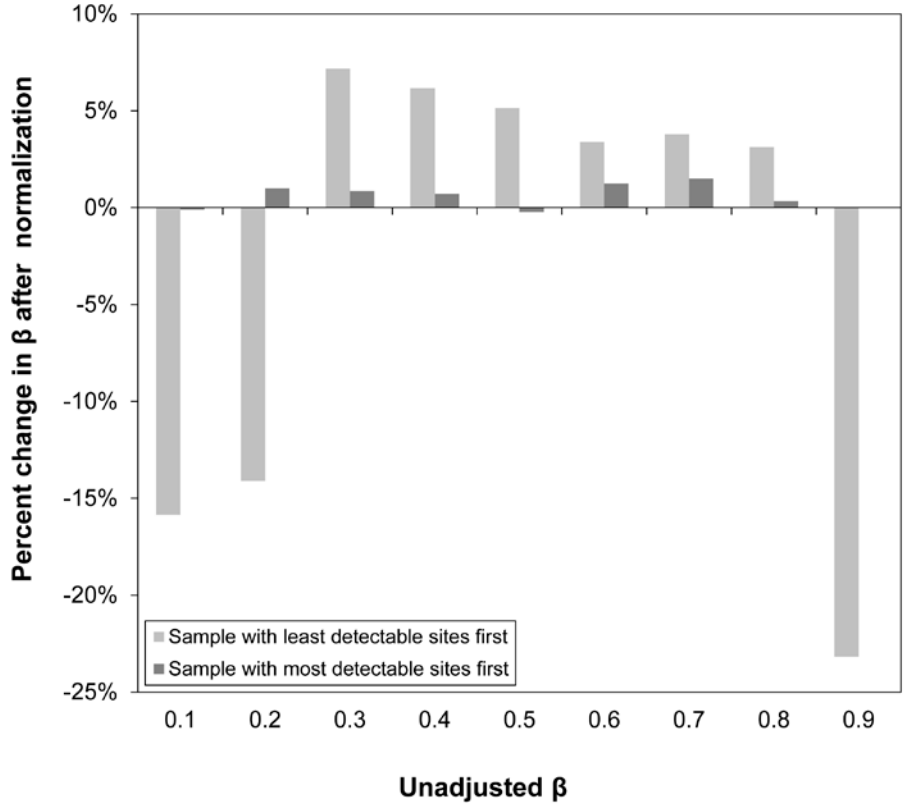


Figure 6. Average percent change of methylation values, β's, after normalization by best and worth performing samples. Mean percent change in β's, values ranging from 0.1 to 0.9, based on normalization by the lowest quality sample (largest amount of CpG sites with $p<0.05$) and the highest quality sample (least amount of CpG sites with $p<0.05$) over all samples (n=432). While normalization by the highest quality sample changed the β's only slightly (<10% on average), normalization by the lowest quality sample tended to change the low and high methylation β'substantially (>10% on average).

of 0.10 for all samples normalizing either on the lowest or highest quality sample. In general, normalization using the poorest quality sample (sample 411 in this dataset) yielded much lower methylation β's. Further, normalization using the poor quality samples led to much larger variability in β's (Figure 6 and Table 1), particularly at extreme methylation values. Table 1 shows that when we normalize using a high quality sample (e.g. sample 355), or if we normalize over a summary measure (mean over one plate or all samples), the β's do not change drastically after normalization and remain in the high, medium, and low range. However, when we normalize over the low quality sample, all three β's (low, medium, and high) decreased and the normalized value for high βs became much lower (~0.6 vs. ~0.9). These results could bias downstream analyses as the power to detect differences in methylation would be lessened, highlighting the importance of choosing a reliable normalization procedure based on high quality samples.

Repeatability and batch variability

When examining repeatability of replicates, we assessed the reduction in root mean squared error (root-MSE) between replicates by each normalization procedure (Table 2). All RN-factor based color channel normalization procedures resulted in lower mean root-MSE between replicates. The greatest reduction was a decrease of 10.83%, occurring with normalization using the RN-factors of the best performing sample in the experiment (sample 355 here). Those normalizations that used an aggregate RN-factor, such as the ASMN and the single plate-mean normalizations, each elicited similar reductions in root-MSE between replicates, all producing approximately a 10% reduction compared to un-normalized data. Of the color channel

Table 1. Reference normalization (RN) factors and methylation (β's) for a single sample by normalization procedure.

	Reference Normalization Factors		Calculated β's (Infinium I)			Calculated β's (Infinium II - red)		
	Red	**Green**	**High**	**Medium**	**Low**	**High**	**Medium**	**Low**
Plate 1	3878.3	5116.5	0.907	0.455	0.073	0.911	0.468	0.076
Plate 2	4254.6	5408	0.909	0.456	0.073	0.91	0.46	0.074
Plate 3	4145.7	5271.2	0.908	0.456	0.073	0.91	0.46	0.074
Plate 4	4720	5680.8	0.91	0.457	0.073	0.907	0.447	0.07
Plate 5	5041.4	5913	0.911	0.457	0.073	0.906	0.441	0.069
ASMN	4337.6	5429.6	0.909	0.456	0.073	0.909	0.456	0.073
IFSN	3633	4486	0.906	0.455	0.072	0.905	0.451	0.072
Sample 355	5480.5	6875.5	0.913	0.458	0.073	0.913	0.458	0.073
Sample 411	168.5	271.6	**0.627**	**0.318**	**0.05**	**0.684**	**0.375**	**0.064**

Red and green RN- factors were calculated using the mean signals of the 93 normalization controls (A and T signals for red factor and C and G signals for the green factor) over plate 1, 2, 3, 4 or 5, over all samples on all plates, using the first sample (sample 1), a high quality sample (most CpG sites detected, sample 355), or a low quality sample (least CpG sites detected, sample 411). Calculated methylation values (β's) were for one sample whose mean normalization signals were equal to the all plates values (4337.6 and 5429.6 for red and green, respectively and with fixed signal As and signal B corresponding to high, medium, and low β's. Signal A's were 400, 3000, and 5000 and signal B's were 5000, 2600, and 400 for high, medium, and low β's, respectively.
ASMN – All sample mean normalization; IFSN - Illumina first sample normalization; Sample 355- sample with the most detectable sites (high quality); Sample 411- sample with the least detectable sites (low quality).

normalizations evaluated, the lumi normalization reduced mean replicate root-MSE the least (0.58%). However, both datasets that utilized an additional normalization technique for Infinium chemistry adjustment (BMIQ) saw increases in root-MSE compared to un-normalized. While this effect was relatively small for the ASMN combined with BMIQ normalization (a 1.21% increase), the lumi normalization followed by BMIQ produced a sizeable increase in mean replicate root-MSE (11.96%), indicating a decrease in repeatability. The changes in correlation observed for each of the normalization procedures relative to un-normalized results largely followed similar trends as those observed for root-MSE (Table 2). However, due to the bounded nature of the correlation coefficient, the magnitude of the effect was not as large.

Visual assessment of batch variability also identified important differences between normalization procedures (Figure 7). Color channel normalization is expected to increase comparability of mean chip values and decrease batch variability over non-normalized β's, as seen in Figure 7A. Lumi smooth quantile normalization (Figure 7B) appears to retain many of the extreme points and batch trends observed in the raw β's, as does using the worst performing sample's RN-factor values, which also decreases the real scale of the β distribution (Figure 7C). The box plots of mean sample β using normalization by ASMN, shown in Figure 7D, demonstrate a reduction in the number of outlier samples and batch-related variability.

In the site-level analysis of batch-associated variability, the 'raw' un-normalized β's showed a relatively high percentage of CpG sites that were associated with the chip batch (12.8%) (Figure 8) compared to other normalizations. Other mean RN-factor based color channel normalization procedures, including ASMN and each of the plate mean RN-factor normalization procedures showed fewer batch associated sites than raw β's and the percentage of sites were

largely consistent across these procedures. When using only one sample's control probe values, sample quality appeared to influence the amount of batch variability across the experiment. For instance, the best performing sample (by fewest number of non-detectable CpG sites) and a well-performing first-experiment sample used in the IFSN, both had percentages of batch-association comparable with aggregate RN-factor based procedures. However, the worst performing sample had the highest level of batch association. The lumi procedure also showed a reduction in the percentage of batch-associated sites compared to un-normalized results, having even a slightly lower percentage than the aggregate RN-factor based procedures. Additionally, the number of batch-associated sites was further reduced for both the lumi and ASMN when they had been followed by the BMIQ adjustment for Infinium assay chemistry.

Finally, the ASMN normalization procedure has been compiled into an R package that will be freely available in an open-source distribution in the bioconductor repository for bioinformatics software (www.bioconductor.org).

Table 2. Repeatability of technical replicates by improvement of root mean squared error (root-MSE) and mean Spearman correlation (R2) compared for un-normalized results.

Normalization Method	% Change in Root-MSE	R^2 for Replicates	% Change in R^2
ASMN	-10.43	0.97	0.339
Plate 1	-10.72	0.97	0.339
Plate 2	-10.55	0.97	0.338
Plate 3	-10.5	0.97	0.338
Plate 4	-10.17	0.97	0.343
Plate 5	-10	0.97	0.348
IFSN	-9.91	0.97	0.342
Sample 411	-5.59	0.965	-0.115
Sample 355	-10.83	0.97	0.344
lumi	0.58	0.968	0.151
ASMN + BMIQ	1.21	0.965	-0.176
lumi + BMIQ	11.96	0.962	-0.428

Percentage change in root-MSE and R2 between 15 sets of replicate pairs and un-normalized results by the ten normalization procedures. The un-normalized root-MSE had a baseline value of 0.0499 methylation units (β's) and the un-normalized R2 was 0.9664. For root-MSE calculations, a pair of replicates was randomly chosen from two replicates sets that had more than two total samples and consistently evaluated across each normalization method. Normalization procedures included: All sample mean normalization (ASMN), normalization by reference normalization (RN) factor taken as the mean control probe values for each of the plates (Plate 1 – Plate 5) run, normalization by the worst performing sample's RN-factor (sample 411) and the best performing sample's RN-factor (sample 355), lumi smooth quantile normalization, and both the ASMN and lumi normalization followed by beta-mixture quantile normalization (BMIQ).

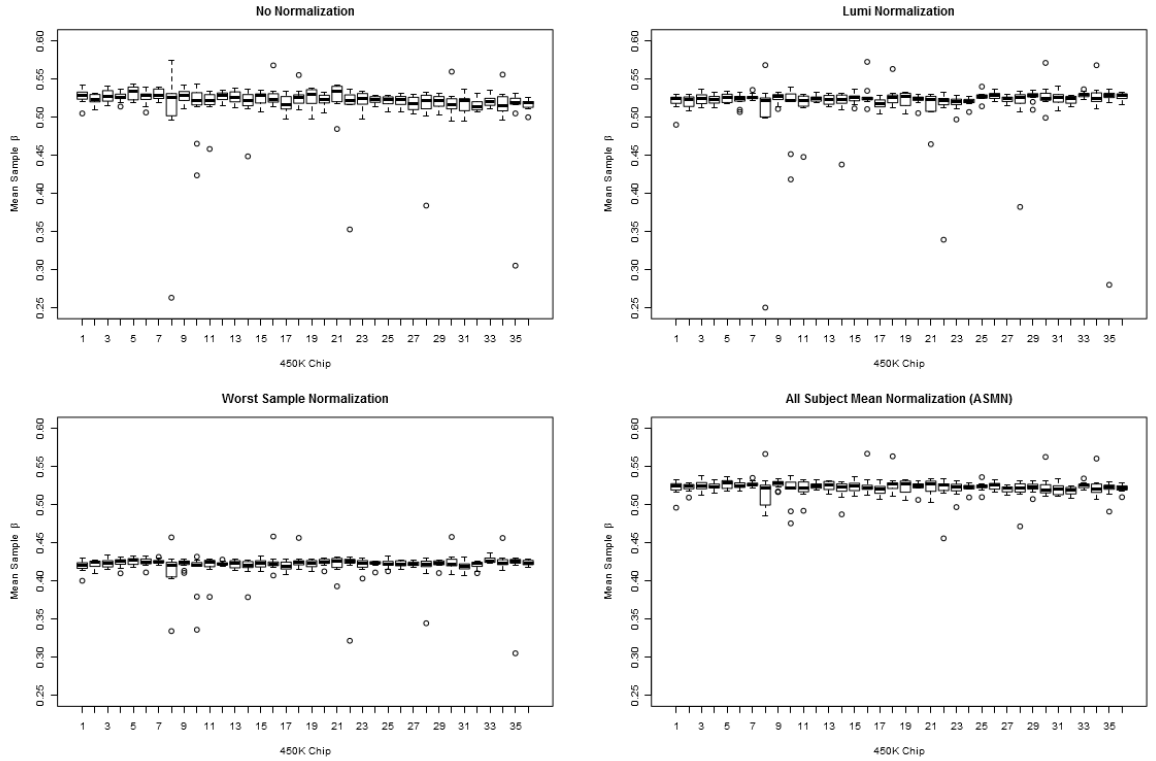


Figure 7. Box plots of sample mean methylation by normalization methods. Box plots of mean per-sample methylation (β) for all sites interrogated on the 450K array (n=485,512) by color channel normalization methods. Plots are shown for (A) un-normalized results and three different normalization methods, (B) lumi smooth quantile normalization, (C) normalization using the worst performing sample's reference normalization factor values (sample 411), and (D) using the all sample mean normalization (ASMN) method. Each chip assays twelve samples, so every box plot contains twelve observations in total.

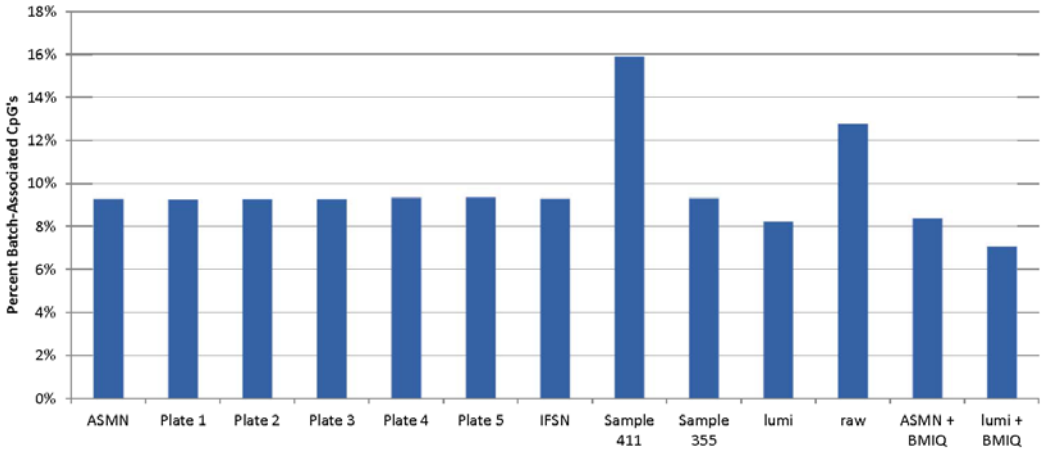


Figure 8. Percent of 450K array CpG sites associated with chip batch ($p<0.01$) shown by normalization method. Normalization methods include: All sample mean normalization (ASMN), normalization by RN-factors taken as the mean control probe values for each of the plates (1-5) run, normalization by the worst performing sample's reference normalization (RN) factor (sample 411) and the best performing sample's RN-factor (sample 355), lumi smooth quantile normalization, and both the ASMN and lumi normalization followed by beta-mixture quantile normalization (BMIQ). Batch association was evaluated by ANOVA for each of the n=485,512 CpG sites interrogated.

Table 3. Mean standard deviation (SD) and coefficient of variation (CV) between 15 sets of replicates by type of Illumina infinium chemistry (Inf I and InfII) and different normalization procedures.

	Mean SD, InfI (95%CI)	Mean SD, InfII (95%CI)	Mean CV, InfI (95%CI)	Mean CV, InfII (95%CI)
ASMN	0.0135 (0.0019, 0.0478)	0.0226 (0.0058, 0.0694)	19.0475 (0.4751, 71.5218)	7.1307 (1.0838, 25.2767)
Plate 1	0.0135 (0.0019, 0.0476)	0.0225 (0.006, 0.0694)	19.0033 (0.4618, 71.5156)	7.078 (1.0461, 25.1793)
Plate 2	0.0135 (0.0019, 0.0478)	0.0226 (0.0059, 0.0693)	19.0341 (0.4713, 71.5202)	7.1121 (1.0721, 25.254)
Plate 3	0.0135 (0.0019, 0.0478)	0.0226 (0.0059, 0.0695)	19.0347 (0.4721, 71.5195)	7.1174 (1.0725, 25.2498)
Plate 4	0.0134 (0.0018, 0.0479)	0.0227 (0.0057, 0.0695)	19.0808 (0.4852, 71.534)	7.1717 (1.1124, 25.3501)
Plate 5	0.0134 (0.0018, 0.048)	0.0227 (0.0056, 0.0695)	19.1025 (0.4922, 71.5366)	7.1975 (1.1308, 25.3863)
IFSN	0.0135 (0.0018, 0.0479)	0.0228 (0.0058, 0.0705)	19.0663 (0.4873, 71.5252)	7.1929 (1.0992, 25.2778)
Sample 411	0.0147 (0.0019, 0.0493)	0.026 (0.0061, 0.0772)	19.4851 (1.5131, 71.5281)	9.2926 (1.5752, 27.0743)
Sample 355	0.0135 (0.0019, 0.0478)	0.0224 (0.0058, 0.0684)	19.0388 (0.4665, 71.5322)	7.0819 (1.0766, 25.3046)
lumi	0.0134 (0.0017, 0.0529)	0.0251 (0.0056, 0.0839)	19.6198 (0.4354, 75.3505)	7.6002 (1.031, 25.1459)
Raw	0.0136 (0.0018, 0.0518)	0.0259 (0.0063, 0.0837)	18.6096 (0.4855, 71.2258)	7.7368 (1.1908, 25.0344)
ASMN + BMIQ	0.0135 (0.0019, 0.0478)	0.0241 (0.002, 0.083)	20.5427 (0.4805, 78.8597)	13.2042 (0.738, 56.3085)
lumi + BMIQ	0.0134 (0.0017, 0.0529)	0.0266 (0.0018, 0.0973)	19.6198 (0.4354, 75.3505)	13.2997 (0.7296, 54.9395)

Normalization procedures included: All sample mean normalization (ASMN), reference normalization (RN) factors taken as the mean control probe values for each of the plates (Plate 1 – Plate 5) run, normalization by the worst performing sample's RN-factors (sample 411) and the best performing sample's RN-factors (sample 355), lumi smooth quantile normalization, no normalization (raw), and both the ASMN and lumi normalization followed by beta-mixture quantile normalization (BMIQ).

Discussion

In this study, we implemented and evaluated the performance of 10 variations of color channel normalization for Illumina 450K methylation data from a large epidemiologic study. In addition to using two common color channel normalization procedures (IFSN and lumi), we also implemented our preferred new ASMN normalization procedure, and several additional strategies to evaluate the range of performance that could be achieved with RN-factor based procedures. We specifically examined the ability of these normalization procedures to reduce major sources of technical variability by assessment of a) batch effects and b) performance of technical replicates included in the experiment. We found that the ASMN procedure outperformed the Illumina recommended IFSN algorithm, and further, that ASMN consistently performed well while the performance of IFSN varied depending on sample quality. We observed comparable performance between normalizations using the RN-factors from the best performing sample and ASMN, while the latter had the added benefit of not relying on data mining. We also found that the ASMN procedure was better at increasing repeatability between technical replicates than the commonly used lumi approach and had similar benefits for reducing batch effects. Lastly, we confirmed that the advantages of ASMN normalization compared to lumi were retained even after adjustment for differences in Infinium chemistry using the popular BMIQ algorithm. These findings suggest that the ASMN procedure is an improvement over existing strategies for color channel normalization, especially for large epidemiologic studies. Thus, its implementation in conjunction with other data cleaning steps in any 450K methylation data pipeline is warranted.

Improved performance in repeatability and reduction of batch effects were observed for ASMN when compared to the IFSN procedure recommended by Illumina. While some of these gains in performance were relatively small in scale, as when comparing the number of batch associated CpG's found for each procedure, they were consistent across all performance

measures. Further, our parallel assessment of normalization by using the RN-factor values for both the best and worst performing samples showed the range of possible performance that could have been garnered with the IFSN strategy. While the best sample's performance was largely comparable with the ASMN, the worst sample's performance was drastically worse, even seeming to introduce batch variability compared to non-normalized results (Figure 8). This range of performance demonstrated that while some samples may perform satisfactorily when used in normalization, others may introduce bias to results. The likelihood of a poor performing normalization by the IFSN strategy is essentially a random draw from the range of sample qualities included in a given experiment. The ASMN procedure provides a convenient and more reliable alternative, since its performance is stable over a given experiment. In addition, the use of ASMN instead of normalization by RN-factors from the best performing sample provides a robust methodology that does not rely on prior access to data or data mining.

Comparison of the ASMN procedure to the lumi normalization showed that ASMN had increased repeatability across all metrics evaluated. In fact, lumi often performed only marginally better than using raw non-normalized results (Table 2). While lumi did not effectively improve repeatability, it did provide substantial reductions in batch effects, outperforming both ASMN and the best sample RN-factor normalizations in this regard. One possible explanation for this inconsistent performance may be over-fitting of the lumi algorithm, which aggressively coerces the distribution of normalization targets to have identical quantiles as the reference distribution. In turn, this may reduce the number of possible methylation values and minimize batch effects, even while not addressing the repeatability issues. Further, since the loss in batch variability does not co-occur with gains in repeatability, the apparent benefits of this approach may actually come at the cost of artificially reduced biological variability.

When we examined the performance of the lumi and ASMN procedures followed by adjustment for differences in Infinium chemistry using the BMIQ algorithm, we continued to observe benefits of using ASMN rather than lumi. While in general both lumi + BMIQ and ASMN + BMIQ performed well at reducing batch variability, neither of these combined strategies saw improved performance of technical replicates compared to datasets receiving only color channel normalization. Again, the lumi + BMIQ dataset exhibited the same trend seen in the lumi color normalization alone: much lower batch variability with increased variability between technical replicates. As such, it seems likely that the issue of the lumi algorithm over-fitting is retained even when followed by BMIQ normalization. The ASMN + BMIQ dataset, like the dataset receiving the ASMN normalization alone, had consistent performance in reducing technical variability. While some of the gain in repeatability between replicates afforded by the ASMN was lessened with addition of BMIQ, it was previously demonstrated that adjustment for Infinium chemistry is needed,[45,86,88] and thus BMIQ has to remain in the 450K data processing pipeline.

It is important to clarify that the assessment we present here is focused on performance of color channel normalization in particular, and isn't a comprehensive evaluation of all the processing steps needed prior to analysis of biological effects from 450K array data. Several additional data processing steps have been suggested in the literature and are freely available as R packages, including filtering out SNP-associated probes included in the 450K assay and adjusting for the Infinium I and II chemistries.[45,46,88] To confirm that improved performance would be retained in the context of a full pipeline, we also performed SNP-filtration prior to ASMN normalization and observed similar gains (data not shown). Our results indicate that color channel normalization should indeed be performed in addition to SNP-filtering and Infinium

chemistry adjustment (BMIQ), and should be included in any robust Infinium data processing pipeline.

While other studies have examined normalization strategies for 450K data, to date they have focused on reducing differences between the Infinium I and II chemistries, and have been under-powered to evaluate the batch effects that are likely to occur in large association studies.[45,86,87] A main advantage of our approach was a large sample size and inclusion of many technical replicates for rigorous evaluation of normalization performance. Only one other evaluation has been published to date examining normalization of data from the 450K assay for anything approaching a population study.[89] This study based on 85 samples, found that a pipeline that included lumi color adjustment, followed by BMIQ performed the best at reducing batch variability and increasing repeatability. However, Marabita et al. mostly directed their comparison to performance between Infinium I and II chemistry adjustment. Further, they did not consider an option we propose here as ASMN, which we tested alongside the lumi + BMIQ procedure (that they preferred). Our study's capacity to detect true batch effects was much larger than the Marabita study, which examined only 85 samples analyzed on eight BeadChips. Numerous BeadChips and plates analyzed in our study ($n_{samples}$=432, n_{chips}=36, n_{plates}=5) are more representative of the scale of batch effects that would be encountered in large population or case-control studies. Also, Marabita et al. only examined repeatability with n=16 total replicates (n=8 pairs) while our assessment included more than double that number of replicates (n=38 from n=15 samples).

In summary, we implemented the most comprehensive comparative evaluation of color channel normalization procedures for the 450K assay to date. The large sample size and the many technical replicates included in the analysis allowed for careful assessment of sources of technical variability, including those that are likely to be unique to large epidemiologic studies. Our results show that the ASMN normalization procedure that we introduced is an excellent alternative to the two leading color channel normalization strategies, Illumina's IFSN and lumi. ASMN reduced technical variability compared to the IFSN procedure and did not encounter the performance trade-offs of the lumi approach. As ASMN relies on a predefined measure of central tendency among control values, it is a stable and robust approach to normalization. Further, the ASMN procedure yielded reductions in technical variability beyond normalization for Infinium chemistry type alone by BMIQ. These findings suggest that, especially for large epidemiologic studies, the ASMN color channel normalization is a valuable component to be included in a 450K methylation data pipeline.

Materials and methods

Samples

DNA was isolated from a convenience set of blood clots from 408 healthy children participating in a longitudinal birth cohort study, using QIAamp DNA blood kits from Qiagen (Germantown, MD) according to the manufacturer's protocol. Following isolation, all samples were checked for DNA quantity and quality by Nanodrop 2000 spectrophotometer. Samples were retained if they produced high yield and good DNA quality (as assessed by 260/280 ratio exceeding 1.6) and concentrations were adjusted to 50ng/ul. DNA aliquots of 1ug were bisulfite converted using Zymo Bisulfite conversion Kits (Zymo Research, Orange, CA). Study protocols were approved by the University of California, Berkeley Committee for Protection of Human Subjects

Illumina Infinium HumanMethylation450 DNA Methylation Assay

DNA samples were whole genome amplified, enzymatically fragmented, purified, and applied to the 450K BeadChips according to the Illumina methylation protocol.[41,42] BeadChips were processed with robotics and analyzed using the Illumina Hi-Scan system. Each 450K BeadChip can fit n=12 samples in total, and these chips are usually run combined onto plates in sets of 8, for a batch of n=96 samples. To accommodate all of the samples analyzed in this experiment, 36 BeadChips were run across 5 plates. BeadChips included on the same plate (up to n=8 BeadChips per plate) were analyzed simultaneously, and time between plate runs was approximately one week using the same batch of all reagents and chips.

Data Extraction

Sample data were extracted using Illumina GenomeStudio software (version XXV2011.1, Methylation Module 1.9) methylation module. This provides raw intensities for both red and green color channels, detection p-values as a measure of assay performance, and β's calculated from raw signals for all samples at all 485,577 assayed probes. Data cleaning performed prior to evaluation of different normalization procedures included background correction of raw signal intensities according to Illumina recommendations using GenomeStudio software. The background is determined as the lowest 5^{th} percentile of the 600 negative controls included in the assay and was subtracted from the probe intensities. Also, the n=65 SNP probes noted in the Illumina manual were filtered out, leaving 485,512 CpG sites for analysis.

Simultaneous to sample data extraction, control probe data extraction was also performed using the GenomeStudio software. This additional matrix contains raw signal observations for all of the probes included as controls in the design of the 450K assay. Such control values include negative controls (for background subtraction), extension controls, staining controls, bisulfite controls, and n=93 normalization control probes among others. The normalization control probe pairs are targeted to non-variable regions of stable housekeeping genes and are the observations used to calculate the RN-factors used RN-factor based normalization procedures (Figure 2).

Quality Assurance/Quality Control (QA/QC)

Of the samples selected for analysis, 14 samples were randomly chosen to be included as technical replicates. Replicates were designed to maximize the capacity to detect multiple forms of bias across the experiment. As such, 3 pairs of samples were included as intra-chip replicates, 6 pairs were included as intra-plate replicates, 4 pairs were DNA isolation replicates, and 1 sample was an inter-plate replicate run 7 times across all plates of the experiment. Furthermore, one internal control sample (DNA from a Jurkat cell-line) was run on each sample plate, replicated 5 times in total. Including the Jurkat DNA, fifteen sets of replicates were included throughout the experiment, comprising n=38 QA/QC samples. The location of samples on assay wells for each of the Illumina BeadChips was randomized.

In addition to replicates, the Illumina GenomeStudio software provides an internal measure of assay quality for each CpG site interrogated: a detection p-value. This value represents the chance that the signals produced from a given site were not distinguishable from background. Thus, a small detection p-value would indicate that the fluorescent signals at a particular CpG site were likely above background levels. Illumina suggests using a detection p-value cutoff of 0.05 above which a CpG site should be excluded from analysis.

Color Channel Batch Normalization Procedures

Including all QA/QC samples, a total of n=432 samples were assayed. We refer to this as the total samples included in our "experiment." Further, we define "batch" effects as occurring at two different levels: 1) the BeadChip level (which includes observations from n=12 samples) and, 2) the plate level (which includes n=8 BeadChips and n=96 samples). Our experiment includes 36 BeadChip batches and 5 plate batches. Unless otherwise specified, all batch analysis was conducted at the BeadChip level.

To evaluate the performance of different procedures adjusting for color channel bias across batches, we implemented 10 different normalization procedures to background subtracted signals, creating a total of ten different datasets. These 10 procedures fell into two methodological categories: 1) reference normalization (RN)-factor based and 2) quantile based methods (Figure 1). The nine RN-factor based procedures utilize the values of the n=93 normalization control probe to construct RN-factors (Figure 2) and differ by which observations are used to calculate RN-factors. There were two groups of RN-factor based methods: 1) those using only the RN-factors from a single sample and 2) those using an aggregated RN-factor (Figure1).

Each of the 10 color channel normalization procedures are described below:

1. The Illumina first sample normalization (IFSN) is the standard color channel normalization recommended by Illumina. This procedure uses the mean of the first sample's normalization control probe values (both red and green) to calculate the RN-factors. Another single-sample RN-factor normalization was performed: one using the RN-factors for the best performing sample in the experiment (sample number 355). As described in QA/QC above, the best performing samples was determined by having the highest number of CpG sites meeting a detection p-value threshold less than 0.05.

3. A single-sample RN-factor normalization was also performed using the RN-factors for the worst performing sample in the experiment (sample number 411). The worst performing sample in the experiment was determined by having the least number of CpG sites meeting a detection p-value threshold less than 0.05.

4. The all sample mean normalization (ASMN) strategy that we developed uses the means of the RN-factors of all samples in the experiment (in this case n=432) as the RN-factors.

5-9. Beyond calculating the RN-factors as the mean over all samples in the experiment, we also performed normalization by averaging over different sub-groups within the experiment, namely each of the 5 plate-batches in which the experiment was run. RN-factors calculated as the mean RN-factors by each plate created 5 different mean-plate RN-factors and 5 output datasets. These procedures essentially set 1 plate batch as the baseline to which all other batches are normalized.

10. Lastly, one non RN-factor based color channel normalization, the lumi smooth quantile normalization procedure, was also implemented. This approach involves local polynomial smoothing followed by an interpolation step. The procedure assumes that the distributions of data within each color channels are identical and coerces the distribution of each target color channel to have identical quantiles to the reference distribution.

To further confirm the stability of the ASMN procedure, an additional dataset was generated which removed n=16,667 CpG sites that potentially include common (minor allele frequency >5%) SNPs prior to ASMN normalization. SNP list was obtained using the HapMap project population most comparable to our cohort.[91] All measures of normalization performance were retained following removal of possible SNP-associated CpGs (data not shown).

Adjustment for Infinium Chemistry

In addition to the ten datasets created by implementing different color channel batch normalization strategies, we also implemented an adjustment procedure (BMIQ) to account for the systematically different performances of the Infinium I and II chemistries to two of our color channel normalized datasets. We applied BMIQ to the ASMN and lumi normalized datasets (numbers 2 and 6 above) to evaluate how reduction of batch variability would be impacted by adding this needed correction for assay chemistry. The BMIQ normalization procedure is a model-based strategy that applies a three-state beta mixture model to assign methylation states, followed by quantile normalization using the parameters of these beta distributions.[88]

Statistical Analysis

After extraction of raw values was conducted using the Illumina Genome Studio software, all subsequent statistical analysis was performed using the R statistical computing software. The lumi smooth quantile normalization was implemented using the lumi package.[44] The BMIQ algorithm was implemented using the freely available code cited in Teschendorff et al.[88]

Repeatability was assessed by comparison of the performance of the 15 sets of technical replicates distributed broadly across all of the chips run for the experiment. We take our use of the term 'repeatability' from the Wild, Vineis, and Garte (2008) text, meaning the "ability to yield the same results… each time the test is conducted in the same laboratory."[92] Standard deviations and coefficients of variation were calculated for all CpG sites run on the Infinium assay (n= 485,512 CpG sites total). The means of these measures, taken for both Infinium I and II assays separately, were taken across all replicate sets for each of the color channel normalizations conducted as a measure of procedure stability (Table 3).

Further, the root mean squared error (root-MSE) was computed between all sets of technical replicates for each of the normalization procedures evaluated. This provided an estimate of technical error in the same scale as the measurement taken, in this case on the zero-to-one scale of methylation β's. For raw, un-normalized β values, the mean root-MSE among all 15 sets of technical replicates was 0.0499 β units. Using this value as a reference, we compared the mean replicate root-MSE across each of the different normalization procedures to this standard expressed as a percentage change from the mean root-MSE for the un-normalized dataset. Spearman correlation coefficients were also calculated for all replicates sets and averaged by normalization procedure as an additional measure of replicate comparability.

Batch variability was also evaluated for each of the normalization procedures implemented. Box plots of mean per-sample β for all sites interrogated on the 450K array were constructed to visualize trends in means by batch across the entire experiment. Plots are shown for three different color channel normalization procedures (lumi, worst sample RN-factor, and ASMN) by the Illumina chip batch on which they were analyzed (Figure 7). Beyond visual assessment of batch trends, a site-level analysis of batch-associated variability was conducted for each of the normalizations utilized. Batch variability across chips was evaluated by ANOVA for each of the CpG sites. A site was considered 'batch-associated' if the p-value associated with effect of analysis chip was less than or equal to 0.01. Levels of batch association were compared between each normalization procedure by taking the number of CpG sites meting the $p \leq 0.01$ criteria for batch association as a percentage of total sites on the 450K assay.

CHAPTER 5

Estimation of blood cellular heterogeneity in newborns and children for epigenome-wide association studies.

ABSTRACT

Confounding by cellular heterogeneity has become a major concern for epigenome-wide association studies (EWAS) in peripheral blood samples from population and clinical studies. Adjusting for white blood cell percentage estimates produced by the minfi implementation of the Houseman algorithm (minfi) during statistical analysis is now an established method to account for this bias in adults. However, minfi has not been benchmarked against white blood cell counts in children that may differ substantially from the reference dataset used in its estimation.

We compared estimates of white blood cell type percentages produced by two methods, minfi and differential cell count (DCC), in a birth cohort at two time points (birth and 12 years of age). We found that both minfi and DCC had similar trends as children aged, and neither count method had differed by sex among newborns ($p>0.10$). However, minfi estimates did not correlate well with DCC in samples from newborns ($\rho = -0.05$ for granulocytes; $\rho = -0.03$ for lymphocytes). In older children, correlation improved substantially ($\rho = 0.77$ for granulocytes; $\rho = 0.75$ for lymphocytes), likely due to increasing similarity with minfi's adult reference data as children aged.

Our findings suggest that the minfi method may provide suitable estimates of white blood cell composition for samples from adults and older children, but may not currently be appropriate for EWAS involving newborns or young children.

Introduction

Epigenome-wide association studies (EWAS) have increasingly been used to identify novel biological mechanisms that contribute to disease status or respond to environmental exposures. Several large-scale DNA methylation assays have been developed in recent years, including MeDIP, RRBS, and wgBS,[2,93-95] but due to its reliability, relatively low cost, and broad coverage the Illumina Infinium HumanMethylation450 BeadChip® (450K) has been widely adopted in population-based EWAS.[41,42,96,97]

Unlike genetics, epigenetic markers may change over time or in response to exposures. DNA methylation in particular undergoes widespread remodeling *in utero*,[4,82,98] For this reason and because early life exposures have been hypothesized to contribute differential risk towards later life ill health, performing EWAS at birth or in young children has been of great interest to investigators. Several EWAS, including those for prenatal exposure to smoking and arsenic,[99,100] have quantified DNA methylation in cord blood. This strategy to assess epigenetic perturbation as near as possible to the prenatal period remains a high priority in light of the fetal origins of human disease hypothesis.[39,101-103]

Whole blood is a desirable matrix to use for EWAS as it is readily available and has been obtained for many human studies with a wide variety of initial aims (including past genome-wide association studies (GWAS)).[40,47-49] However, as EWAS are more commonly performed in blood, there is growing awareness that heterogeneous white blood cell type populations may bias results due to confounding.[48,49] Since DNA methylation may vary by cell type, analyses involving health outcomes or exposures that also co-vary with cell type may be confounded. The consequence of such bias has been clearly demonstrated by Jaffe and Irizarry,[50] who found that many published associations between blood-based CpG methylation and age were no longer statistically significant after adjustment for cell composition.

Several approaches have been proposed to address confounding bias in EWAS due to varying white blood cell type composition. One method is to restrict DNA methylation measurement to isolated populations of white blood cells. In practice, this requires performing fluorescence-activated cell sorting (FACS) prior to DNA isolation and subsequently quantifying DNA methylation signal in isolated cell populations. While appealing theoretically, this approach is not feasible for large population-based studies that rely on banked samples.

One alternative involves estimation of the relative proportions of different cell types, allowing for statistical adjustment for cellular mixture during data analysis. The performance of this approach depends largely on the quality of the estimate of cell type proportions. The most reliable white blood cell count is performed either by automated hematology analyzer, as part of a complete blood count (CBC), or retrospectively by microscopic differential cell count (DCC) using histologically stained blood smear slides.

However, since many epidemiologic studies do not have direct white blood counts, there is growing interest in computational approaches that estimate cell type proportions based on DNA methylation data. In 2012, Houseman et al. were the first to develop such a computational method, using 27k BeadChip results from n=46 isolated white cell samples from an unknown number of blood donors as a reference dataset.[51,52] The updated version, produced by Jaffe and Irizarry[50] (referred to here as the minfi method), has seen the most widespread use because it was incorporated in a popular bioinformatic software pipeline for 450K data, and made several adjustments to specifically improve performance for 450K BeadChip data including the addition of a 450K BeadChip reference dataset (see methods for details). The minfi method is appealing

in the context of EWAS studies because it can be readily implemented with no additional cost or data collection. However, the cell type estimates produced by minfi have not yet been systematically validated against a gold standard cell count, such as CBC or DCC. Additionally, minfi uses a small (n=6) cell-sorted 450K dataset from middle-aged Swedish men in its estimation procedure that may not be an appropriate reference when estimating cell composition in infants and children.[53]

Here, we conduct a comparison of the estimates of the relative abundance of white blood cell types produced by two methods, minfi and DCC, with randomly selected samples from a large epidemiologic cohort followed by the Center for the Health Assessment of Mothers and Children of Salinas (CHAMACOS) study at birth and at 12 years of age with 450K BeadChip data. We report findings showing that reference data and other assumptions should be carefully considered prior to utilizing computationally derived white blood cell estimates in EWAs studies in cord samples.

Materials and methods

Study population

The CHAMACOS study is a longitudinal birth cohort study of the effects of exposure to pesticides and environmental chemicals on the health and development of Mexican-American children living in the agricultural region of Salinas Valley, CA. Detailed description of the CHAMACOS cohort has previously been published.[104,105] Briefly, 601 pregnant women were enrolled in 1999-2000 at community clinics and 527 liveborn singletons were born. Follow up visits occurred at regular intervals throughout childhood, including a visit at 12 years of age that included only male child participants. For this analysis, we include the subset of subjects that had 450K BeadChip data available at birth (n=151) and matched data for the 12-year follow up (n=60). DCC analysis included the subset of subjects with a whole blood smear slides available at the birth (n=111) and 12-year visits (n=45). Both newborns and 12 year olds included in the sample were healthy at the time of blood collection according to the study protocol, and confirmed by abstracted medical records and questionnaires. All subjects included in the subset were Latino in ancestry and 94.0% had at least one Mexican-born parent. Study protocols were approved by the University of California, Berkeley Committee for Protection of Human Subjects. Written informed consent was obtained from all mothers and assent was provided at the 12-year visit.

Blood collection and processing

Whole blood was collected in BD vacutainers (Becton, Dickinson and Company, Franklin Lakes, NJ) containing either heparin anticoagulant or no anticoagulant. Whole blood smear slides were prepared from heparinized blood using the push-wedge blood smearing technique[75] and stored at -20°C until staining. Aliquots of blood clot were stored at -80°C until DNA isolation.

DNA preparation

DNA isolation was performed using QIAamp DNA Blood Maxi Kits (Qiagen, Valencia, CA) according to manufacturer's protocol with small, previously described modifications .[71] Following isolation, all samples were checked for DNA quality and quantity by Nanodrop 2000

Spectrophotometer (Thermo Scientific, Waltham, MA). Those with good quality (260/280 ratio exceeding 1.6) were normalized to a concentration of 55ng/ul.

450K BeadChip DNA methylation analysis

DNA samples were bisulfite converted using Zymo Bisulfite Conversion Kits (Zymo Research, Irvine, CA), whole genome amplified, enzymatically fragmented, purified, and applied to Illumina Infinium HumanMethylation450 BeadChips (Illumina, San Diego, CA) according to manufacturer protocol. 450K BeadChips were handled by robotics and analyzed using the Illumia Hi-Scan system. DNA methylation was measured at 485,512 CpG sites.

White blood cell composition estimation

White cell composition was characterized by two different methods in whole blood.

Differential cell counts(DCC):

Whole blood smears were stained utilizing a DiffQuik® staining kit, a modern commercial variant of the Romanovsky stain, a histological stain used to differentiate cells on a variety of smears and aspirates. This staining highlights cytoplasmic details and neurosecretory granules, which are utilized to characterize the differential white blood count. The staining kit is composed of a fixative (3:1 methanol: acetic acid solution), eosinophilic dye (xanthene dye), basophilic dye (dimethylene blue dye) and wash (deionized water). For consistency and to ensure the best results the slides were all fixed for 15 minutes at 23°C (room temperature), stained in both the basophilic dye and eosinophilic dye for five seconds each and washed after each staining period to prevent the corruption of the dye.

Slides were scored for white blood cell type composition by Zeiss Axioplan light microscope with 100× oil immersion lens. Scoring was conducted at the perceived highest density of white blood cells using the standard battlement track scan method, which covers the entire width of a slide examination area. The counts for each of the five cell types (lymphocytes, monocytes, neutrophils, eosinophils, and basophils) were recorded by a dedicated mechanical counter. At least 100 cells were scored for each slide. Scoring reliability was initially validated by repeated scoring of 5 sets of 100 cells from the same slide with excellent reproducibility (CV≤5%).

Minfi cell count estimation:

Results from the 450K BeadChip analysis were stored as raw IDAT files, and read into the *minfi* (v1.10.2) Bioconductor R package[76] using the *read.450k.exp* function. Estimation of the six (CD8+ T and CD4+ T lymphocytes, CD56+ natural killer cells, CD19+ B cells, CD14+ monocytes, and granulocytes) different white blood cell types was performed using the default implementation of the *estimateCellCounts* function. Briefly, this function takes a user-supplied target 450K BeadChip dataset, combines that with the cell-sorted Reinius reference dataset available in the *FlowSorted.Blood.450k* Bioconductor package (v1.2.0) [53,77] and quantile normalizes the combined data. The reference dataset has n observations from $i = 6$ subjects at each of $j = 6$ different separated cell types. Six hundred cell type informative CpG sites are chosen in the reference dataset, by comparing the mean methylation in a given cell type to the mean methylation of all five remaining cell types for CpG sites assayed. One hundred CpGs are chosen to distinguish each of the j cell types. These represent the 50 CpG sites with the greatest T statistic that were hypermethylated and the 50 CpG sites that were most hypomethylated compared to other cells. This results in a $600 \times n$ matrix of CpG site methylation subset from

the full reference dataset, called S_0. Within S_0, the relationship between indicators for each cell type and DNA methylation is then estimated, producing a vector of coefficients, called $\tilde{B}_0$, of length j. Individual level predictions of cell type proportion, Γ^*, are then fit in the corresponding user submitted dataset, S_1, using the coefficients estimated in S_0 by the following equation:

$$\Gamma^* = \left(\tilde{B}_0^T \tilde{B}_0\right)^{-1} \tilde{B}_0^T S_1$$

Detailed description of the algorithm has previously been published.[50,51]

Statistical analysis

All statistical analyses were performed using R statistical computing software (v3.1.0).[106] Differences in means by age and sex were assessed by Mann-Whitney U-Test. The linear relationships between cell type estimates by the two methods were determined using the Spearman correlation coefficient.

Results

Estimates of white blood cell composition by the two different methods implemented, minfi and DCC, are summarized in Table I. The minfi method estimated the relative percentages of six white blood cell types (CD8+ T and CD4+ T lymphocytes, CD56+ natural killer cells, CD19+ B cells, CD14+ monocytes, and granulocytes) in samples from n=151 newborns and again for 60 of the same children at age 12. Microscopic differential cell count (DCC) of these newborns (n=111) and 12 year olds (n=45) used banked available whole blood smear slides to count the frequency of five types of easily visually identifiable blood cells (lymphocytes, monocytes, neutrophils, eosinophils, and basophils) (see Methods for details).

Table 1. Summary of white blood cell type percentage estimates by two methods, differential cell count (DCC) and minfi in newborn and 12 year old child blood samples.

	Newborns				12 year olds					
	N	Mean, %	Min	Max	N	Mean, %	Min	Max	P Value	
Minfi										
CD8+ T	151	7.1	1.6	13.5	60	14.3	4.6	24.7	<0.01	*
CD4+ T	151	13.7	3.5	25.6	60	13.9	2.4	27.1	0.90	
NK cells	151	4.2	0	17.4	60	8.5	0	21.4	<0.01	*
B Cells	151	12.2	1.9	24.1	60	9.4	5.1	16.7	<0.01	*
Monocytes	151	11.1	4.4	15.8	60	7.1	3.1	11.9	<0.01	*
Granulocytes	151	55.0	34.1	67.7	60	49.1	35	80.7	<0.01	*
DCC										
Lymphocytes	111	29.2	0.3	36	45	46.8	21.4	84.1	<0.01	*
Monocytes	111	6.9	3	10.8	45	6.2	2.8	12.2	<0.01	*
Neutrophils	111	60.4	54	67.2	45	42.7	9.7	68.5	<0.01	*
Eosinophils	111	2.9	0.5	5.3	45	4.1	0	15	0.20	
Basophils	111	0.2	0	2	45	0.1	0	3.9	0.04	*

*Cell types with mean percentages significantly different (p<0.05) between ages by Mann-Whitney U test.

Cell composition estimates by age and sex

By the minfi method, the mean percentage estimates of all cell types except CD4+T lymphocytes were significantly different between newborn and 12 year old samples (p<0.01)

(Table I). Estimates of granulocytes represented the largest percentage of cell types in newborn samples (mean=55.0%), while lymphocytes (CD8+ T and CD4+ T lymphocytes, and natural killer cells) were noticeably less frequent (mean=37.2%, p<0.01) (Figure 1A). In minfi estimates from 12 year olds, granulocyte and lymphocyte populations became much more comparable, with means 49.1% and 46.1% respectively (p=0.21).

By the DCC method, the mean percentage of all but one cell type (eosinophil granulocytes) also differed significantly between newborns and 12 year olds (Table I). For newborns, mean DCC counts were 63.6% for granulocytes and 29.2% for lymphocytes (p<0.01; Figure 1B). By 12 years of age, the gap in the frequency of cell types narrowed (46.9% and 46.8% respectively; p=0.57).

In newborns, there was no difference in cell type distributions by sex (p>0.10) by either minfi estimates (n=58 girls, n=93 boys; Figure 2A) or DCC direct analysis (n=58 girls, n=53 boys; Figure 2B). At age 12 only boys were sampled (N_{minfi}=60; N_{DCC} = 45) so the comparison by sex was not possible.

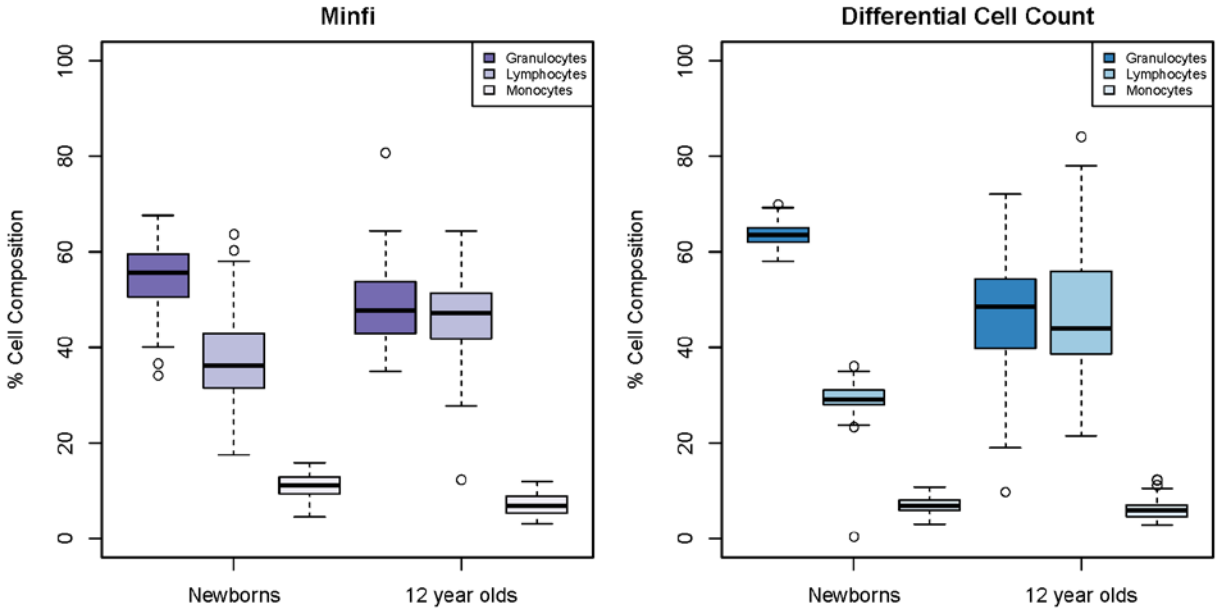


Figure 1. Box plots of percent cell composition estimated by minfi and differential cell count (DCC) in samples from newborns and 12 year olds. The minfi estimates are taken from n=151 newborns and n=60 12 year old boys, and have summed estimates of CD8+ T, CD4+ T, natural killer cells, and B cells into a single category of lymphocytes for comparison. The DCC estimates are taken from n=111 newborns and n=45 12 year olds, and have summed proportions of neutrophils, eosinophils, and basophils into category of granulocytes.

Comparison of cell composition estimates by minfi and DCC

Three cell type populations (Figure 3) were used for more direct comparison of the two methods of assessment of white blood cell composition. For the minfi method, the frequencies for CD8+ T, CD4+ T, natural killer cells, and B cells, were summed to give an estimate of lymphocytes. For DCC, proportions of neutrophils, eosinophils, and basophils were summed to give an estimate of granulocytes.

In samples from newborns, those estimates by minfi and DCC had poor linear correspondence with one another (Figure 3). In fact, the Spearman rank correlation coefficients

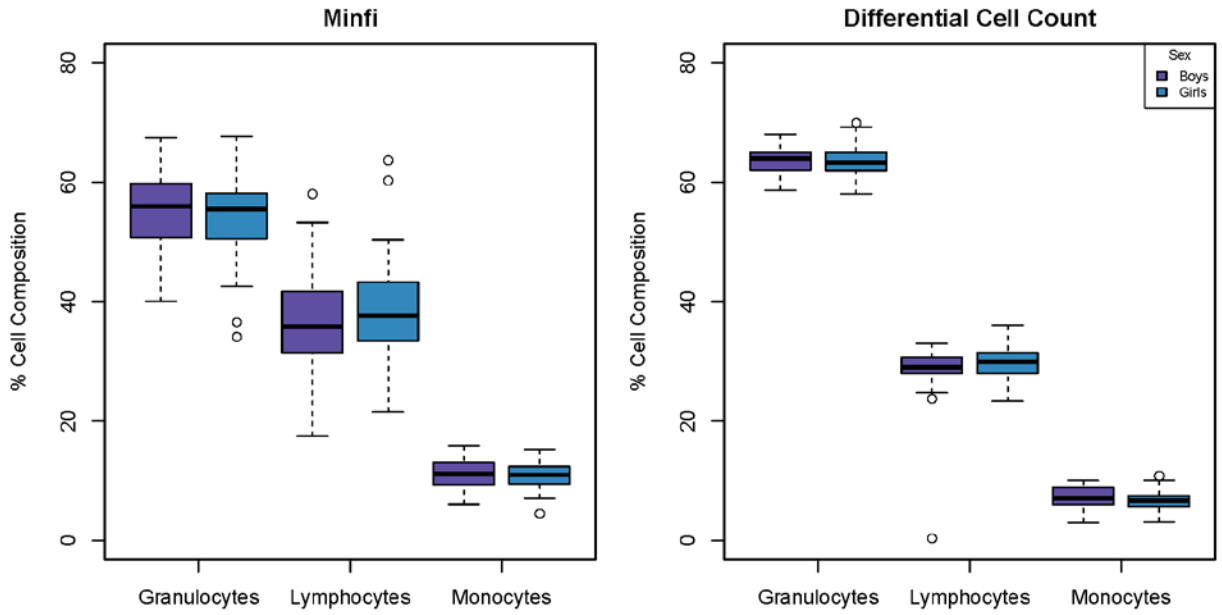


Figure 2. Box plots of percent cell composition estimated by minfi and differential cell count (DCC) for girls and boys in samples from newborns. The minfi estimates are taken from n=151 newborns (n=58 girls and n=93 boys). They have summed estimates of CD8+ T, CD4+ T, natural killer cells, and B cells into a single category of lymphocytes for comparison. The DCC estimates are taken from n=111 newborns (n=58 girls and n=53 boys). They have summed proportions of neutrophils, eosinophils, and basophils into category of granulocytes.

calculated between minfi estimates and direct DCC analysis of monocytes, granulocytes and lymphocytes ranged from -0.01 to -0.05 and were not statistically significant (Figure 3). In scatterplots showing comparison between the two methods, the minfi method appeared to overestimate the proportion of lymphocytes (mean = 37.2%) relative to that by DCC (mean = 29.2%) and reference levels from newborns that range from 19-29% (Figure 3A-C).[55,56] However, minfi also appears to overestimate the percent of monocytes (mean = 11.1%, reference = 5-7%) and gives a smaller percentage of granulocytes (minfi mean = 55.0%, DCC mean = 63.6%, reference = 32-83%). The minfi estimates also had greater variability in newborn samples, standard deviations (SDs) ranging from 2.3 to 8.8, compared to DCC estimates with SDs from 1.8 – 3.6.

However, the two estimates of cell counts were much more consistent in older children than in newborns. At 12 years of age, the means and standard deviations of all comparable cell populations, including granulocytes, lymphocytes and monocytes, were similar by both approaches (Figure 1). The Spearman correlation values for 12-year-old subjects ranged from 0.26 to 0.77 and were significant for granulocytes and lymphocytes (both $p<0.001$), and approached significance for monocytes ($p=0.08$) (Figure 3). Scatterplots comparing the estimates by minfi and DCC at 12 years of age also showed the trend between methods to be linear with comparable amounts of variance.

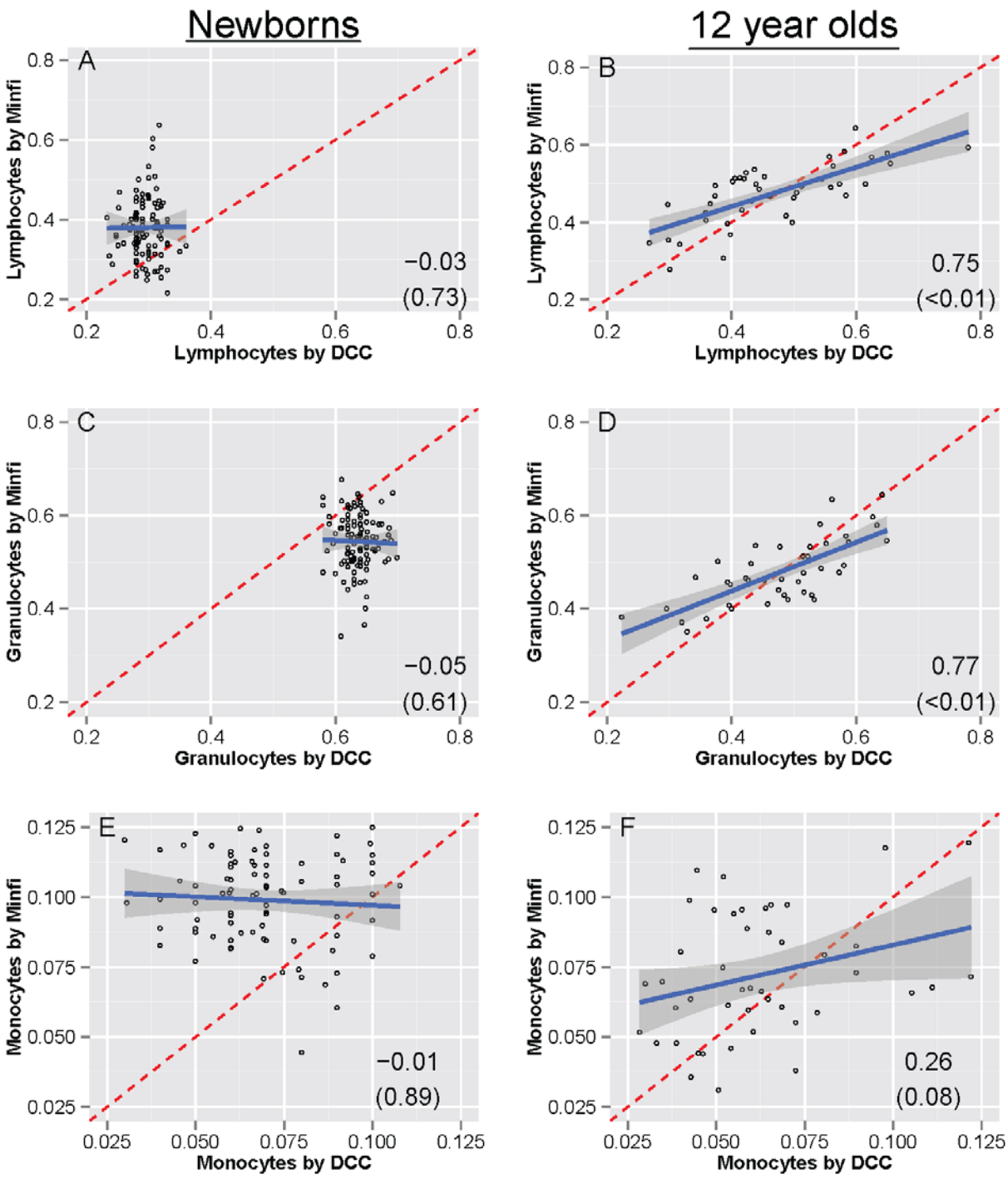


Figure 3. Scatter plot of cell type percentages by minfi and differential cell count (DCC) methods in cord samples for monocytes (A), granulocytes (B), and lymphocytes (C). Also, plots of 12 year samples for monocytes (D), granulocytes (E), and lymphocytes (F). Estimate of linear trend by regression shown in blue with 95% confidence interval in gray. Exact linear correlation, slope=1 and intercept=0, shown in dotted red for reference. Spearman rank correlation, ρ(P-value), shown in bottom right corner for each comparison.

Discussion

Here, we present a detailed comparison of the leading method for estimating white blood cell composition, minfi, against results from a well-established clinical cell counting procedure, DCC, in samples from Mexican-American children at two time points to produce adjustment covariates in a whole blood EWAS analysis. While both methods yielded similar results in 12 year olds, minfi estimates and DCC were quite different in newborns. Our findings suggest that the algorithms applied by minfi may not be appropriate for cell type estimation in newborns and young children.

Longitudinally, as children aged from birth to 12 years old, we observed similar trends by both minfi and DCC. Each found that granulocyte levels were higher than lymphocytes at birth, but became comparable with one another by 12 years of age (Figure 1). This change corresponds with age-specific reference values, which demonstrate high levels of granulocytes (32-83%) relative to lymphocytes (19-29%) in newborns.[55,56] The levels of these two cell types vary noticeably postnatally and through early childhood, reaching a peak difference at 24 months, but stabilizing to adult levels around 9-12 years of age (28-48% for lymphocytes and 33-76% for granulocytes).[55,56]

Among older children, we found that minfi and DCC estimates were consistent with one another and mean estimates by both methods fell within published age-specific reference values [55,56]. However, minfi and DCC estimates differed greatly in newborns. In fact, we saw a negative correlation and linear trend across methods for each comparable cell type (Figure 3 and Figure 2) suggesting that the algorithm implemented by minfi may have difficulty estimating cell composition in samples from newborns.

This deviation is likely explained by the reference dataset used in its prediction model, which is derived from six middle-aged Swedish men. The composition of this reference dataset is crucial to minfi's performance: it is used to both identify CpGs differentially methylated by cell type and fit a regression model to those informative sites. The coefficients estimated in the reference data, $\tilde{B}_0$, establish the linear relationship between methylation at the informative sites and cell composition, which is used for prediction in the target dataset. A key assumption of this method is that the magnitude of the $\tilde{B}_0$ is consistent between the reference and target data. In many situations, this may not be an unreasonable assumption to make. Since the sites used to fit $\tilde{B}_0$ are chosen by their association with cell type, one may expect them to perform cell type specific functions that would be consistent over time. However, given the poor performance of the minfi estimator in newborns, it seems likely that consistent effect of $\tilde{B}_0$ in an adult reference does not hold for young children and impacts the estimator's accuracy. White blood cell populations are still maturing in the early post-natal period and are known to change greatly in relative abundance.[54-56] Further, DNA methylation is known to vary greatly over embryogenesis and may still be changing during early life.[28,107] Should the relationship between these two factors be inconsistent between early childhood and later life, this would result in biased minfi estimates.

Similar bias could occur if the consistent effect assumption does not hold across other biological host factors, such as gender or racial/ethnic ancestry. Both leukocyte populations and DNA methylation are known to vary by such factors.[108-111] The current minfi reference data may be particularly susceptible to these forms of bias because all subjects are men of northern European descent. However, these biases are likely not as pronounced as those introduced by age

in young children since the minfi estimates in CHAMACOS boys of Mexican ancestry are relatively accurate at age 12. Filtering out sites that vary by ethnicity or sex when fitting $\tilde{B}_0$ could potentially reduce bias further, resulting in more accurate estimates of cell composition. Similarly, sites that vary by age could be excluded, an approach that has been used previously to identify candidate metastable epialleles.[112] However, given the lack of variation of the current Swedish male reference dataset over any of these potentially biasing factors, it is preferable to expand or generate a new reference that would have observations from early childhood, and that vary by race and ethnicity.

In conclusion, our comparison of the minfi method for estimating white blood cell composition against a cytological differential cell count demonstrates that minfi can robustly estimate cell populations in children as young as 12 years of age. However, minfi did not perform well in samples from newborns that are important targets of future EWAS because of interest in prenatal epigenetic changes due to exposure or physiological effect on future health. We hypothesize that this is due to low generalizability of the reference dataset currently used in the minfi estimation and suggest that improvement of this dataset would likely enhance its predictions in young children. We encourage using caution when applying the minfi method in populations that deviate substantially in white cell composition and/or methylation patterns from the current minfi reference data, such as by sex, racial/ethnic ancestry, and age in particular. Future work should explore further other factors, such as environmental exposures, that may also impact the validity of the minfi estimates.

CHAPTER 6

Sex differences in DNA methylation assessed by 450K BeadChip in newborns.

ABSTRACT

DNA methylation is an important epigenetic mark that can potentially link early life exposures to adverse health outcomes later in life. Host factors like sex and age strongly influence biological variation of DNA methylation, but characterization of these relationships is still limited, particularly in young children.

In a sample of 111 Mexican-American subjects (58 girls , 53 boys), we observed that ~3% of CpG sites interrogated by the 450K BeadChip in umbilical cord blood specimens were differentially methylated between girls and boys at birth (FDR $P<0.05$) after adjusting for cell composition. Of those CpGs, 3,031 were located on autosomes, and 82.8% of those were hypermethylated in girls compared to boys. Beyond individual CpGs, we found 3,604 sex-associated differentially methylated regions (DMRs) where the majority (75.8%) had higher methylation in girls. Using pathway analysis, we found that sex-associated autosomal CpGs were significantly enriched for gene ontology terms related to nervous system development and behavior. Among hits in our study, 35.9% had been previously reported as sex-associated CpG sites in other published human studies. Further, for replicated hits, the direction of the association with methylation was highly concordant (98.5-100%) with previous studies.

To our knowledge, this is the first reported epigenome-wide analysis by sex at birth that examined DMRs and adjusted for confounding by cell composition. We confirmed previously reported trends that methylation profiles are sex-specific even in autosomal genes, and also identified novel sex-associated CpGs in our methylome-wide analysis immediately after birth, a critical yet relatively unstudied developmental window.

Background

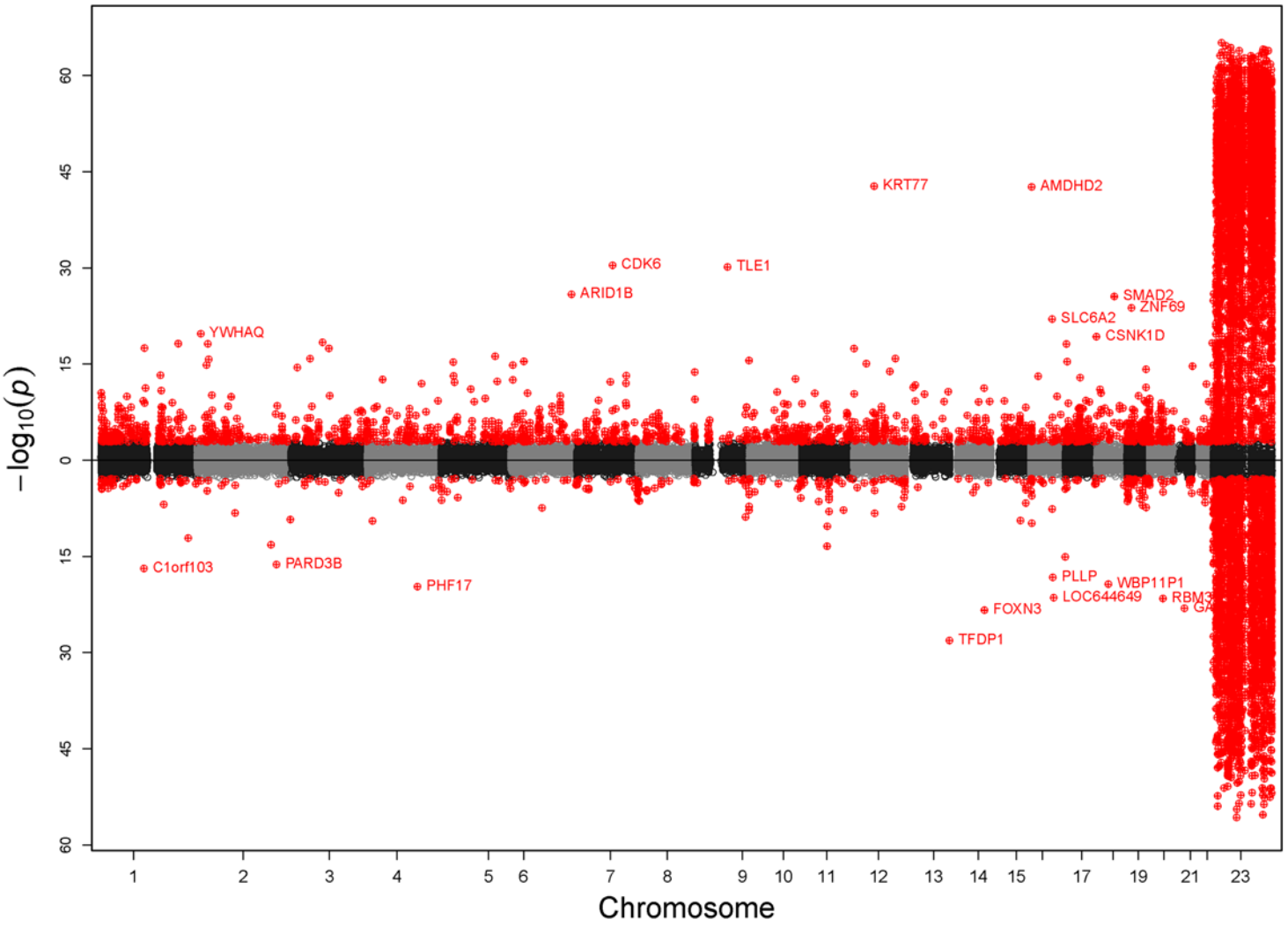
There is a growing interest in examining the role epigenetic marks like histone modifications, non- coding RNAs, and DNA methylation may play as biological mechanisms through which environmental exposures and other physiological and lifestyle factors can lead to disease. Unlike genetics, epigenetic markers modifications are dynamic and can change over time or in response to exposures. Furthermore, host factors such as sex and age also contribute to inter-individual differences in epigenetic markers.

Previous studies of DNA methylation using the Illumina 27K BeadChip methylation array have reported autosomal differentially methylated positions (DMPs) or CpG sites with varying methylation between males and females, providing evidence that it will be important to adjust for sex in analysis of methylation data.[113-118] However, these studies did not account for the existence of non-specific probes for autosomal CpGs that cross react with CpGs on sex chromosomes, thereby yielding false positives.[73] Recently, McCarthy et al. published a meta-analysis of 76 studies all using the 27K BeadChip array to identify sex-associated autosomal DMPs across specimens from multiple tissue types from adults and children.[110] After excluding the sex-biased cross-reactive probes, they identified 184 DMPs that were associated with sex.

While McCarthy et al. identified several interesting autosomal DMPs, their study focused on methylation assessed by the 27K BeadChip. In 2011, Illumina released a new version of their methylation array, the 450K BeadChip, which greatly expanded the number of CpGs interrogated to over 480,000 sites. Further, their approach was restricted to identification of individual DMPs rather than differentially methylated regions (DMRs). DMR-finding approaches have several advantages over considering CpG sites individually, including decreased likelihood of hits from technical artifacts and possibly improved functional impact of results.

As methylation is cell-type specific and immune cell profiles have been shown to vary between sexes, consideration of cell composition is of utmost importance in methylation studies.[119,120] Yet previous studies of sex-associated differences in methylation[113-118] haven't taken this into account in their analyses. White blood cell composition can be estimated from 450K BeadChip data computationally in adults,[50,51] but these estimates are not appropriate for use for young children in their current implementation.[121] As an alternative, differential cell count (DCC) can be employed to effectively determine such cell type proportions (% lymphocytes, monocytes, neutrophils, eosinophils, and basophils) in cord blood samples.

Here, we use the 450K BeadChip to assess sex differences in DNA methylation from umbilical cord blood from boys and girls participating in a large epidemiologic cohort followed by the Center for the Health Assessment of Mothers and Children of Salinas (CHAMACOS) study. We use DCCs to account for white blood cell composition. In addition to interrogating DMPs, we apply the newly released 'DMRcate' methodology[122] to identify sex-associated DMRs in newborns.



.

Figure 1. Manhattan plot for association between child sex (females) and DNA methylation at all 450K CpGs, adjusting for batch and cell composition by differential cell count (DCC). Associations where methylation was higher for girls relative to boys are plotted above the x-axis, while those with decreased methylation are plotted below. CpGs meeting FDR multiple testing threshold of (P<0.05) shown in red.

Table 1. Demographic characteristics of newborn CHAMACOS subjects, N=111.

	Boys, N=53		Girls, N=58		
	N	%	N	%	P Value*
Country of birth					0.18
Mexico	44	46.8	50	53.2	
United States	6	42.9	8	57.1	
Other	3	100	0	0	
Maternal age at delivery (years)					
18-24	26	55.3	21	44.7	0.54
25-29	16	43.2	21	56.8	
30-34	8	38.1	13	61.9	
35-45	3	50	3	50	
Family income					0.58
Below poverty threshold	32	50	32	50	
Above poverty threshold	21	44.7	26	55.3	
Maternal BMI (kg/m^2)					0.31
Normal	24	51.1	23	48.9	
Overweight	14	37.8	23	62.2	
Obese	14	56	11	44	
Smoking during pregnancy					0.07
No	51	50.5	50	49.5	
Yes	2	20	8	80	

	Boys, N=53	Girls, N=58	
	Mean ± SD	Mean ± SD	P Value**
Gestational age (weeks)	39.1 ± 1.4	38.7 ± 2.0	0.35
Birthweight (Kg)	3.5 ± 0.5	3.4 ± 0.6	0.35
Blood Cell Composition (%)			
Lymphocytes	28.7 ± 4.5	29.6 ± 2.5	0.15
Monocytes	7.1 ± 1.9	6.8 ± 1.8	0.48
Neutrophils	60.5 ± 1.8	60.4 ± 2.9	0.53
Eosinophils	3.1 ± 1.2	2.8 ± 1.0	0.44
Basophils	0.2 ± 0.3	0.3 ± 0.5	0.27

* P value from $\chi 2$ test for independence
** P value from Mann-Whitney U test

Results

Sex-associated differentially methylated positions in newborns

Analysis of DNA methylation differences between newborn boys and girls was performed by linear regression for 450K BeadChip CpGs among subjects with DCC measurements (n=111; 58 girls and 53 boys), adjusting for cell composition and batch (Table 1). After data cleaning, n=410,072 CpGs were analyzed, which excluded sites previously reported to exhibit sex-chromosome specific cross-reactivity.[73] Resulting p-values were plotted by chromosome, with sites having higher methylation levels in girls compared to boys plotted above the x-axis and those with lower levels plotted below (Figure 1). After adjustment for multiple

Table 2. Summary of sex-associated DMPs.

	All CpG sites	Autosomes		X chromosome	
		Count	%	Count	%
Hypermethylated in girls compared to boys	8,131	2,509	30.9	5,622	69.1
Hypomethylated in girls compared to boys	3,645	522	14.3	3,123	85.7
Total	11,776	3,031	25.7	8,745	74.3

Number of CpGs significantly hyper- and hypo-methylated in newborn girls compared to boys at FDR multiple testing threshold (q<0.05), for all CpGs, and then stratified by autosomes and the X chromosome.

testing (FDR p<0.05), we identified 11,776 CpGs that differed significantly by sex in newborns (Table 2). Of those hits, the majority of sites had higher methylation in girls compared to boys (69.0%). This trend was consistent on both the X chromosome (64.3 % of sites higher in girls) and in autosomes (82.8%). While the majority of hits were found on the X chromosome (74.3%), a substantial number were also identified on autosomes (3,031 or 25.7%; Table 2).

As differential hypermethylation is to be expected for girls due to X-inactivation,[123-125] we focused characterization of results on autosomal sites showing sex differences (Table 3 and Additional File 1). Most of these were located in CpG shores, islands and open sea (40.4%, 40.1%, and 15.4%, respectively) (Figure 2 and Table 4). In comparison, shelf regions had the

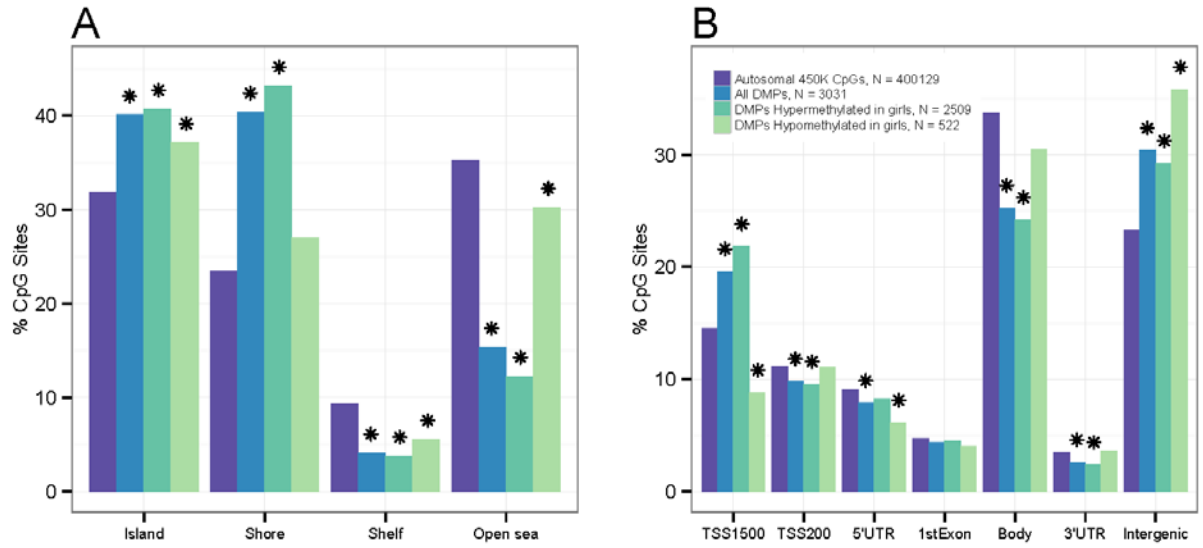


Figure 2. Percent of 450K CpGs (purple), and percent of all (blue), hypermethylated (dark green), and hypomethylated (light green) autosomal differentially methylated positions (DMPs) associated with sex (A&B). These percentages are given by island functional categories (island, shore, shelf, and open sea) in A, and gene functional categories (within 1500bp of a transcription start site (TSS), 200bp of a TSS, a 5' untranslated region (UTR), first exon, gene body, 3'UTR, and intergenic) in B. * indicates that the proportion of sites significantly altered compared to the coverage on the 450K BeadChip (P<0.05).

Table 3. Results for the top 30 gene-annotated autosomal DMPs associated with sex in CHAMACOS newborns.

ProbeID	Gene Symbol(s)	Chrom-osome	Position*	β_{Girl} (M-value)	P-value	Girl Mean (% methylation)	Boy Mean (% methylation)
cg03691818	KRT77	12	53085038	2.38	1.84E-43	11.9	3.0
cg26921482	AMDHD2	16	2570283	1.72	2.43E-43	43.8	20.1
cg17743279	CDK6	7	92463268	1.84	4.04E-31	3.1	1.0
cg07852945	TLE1	9	84303915	0.93	7.30E-31	14.0	7.8
cg26355737	TFDP1	13	114292172	-0.95	7.57E-29	80.6	88.5
cg25568337	ARID1B	6	157098338	0.68	1.36E-26	21.2	14.6
cg05100634	SMAD2	18	45457604	1.77	2.91E-26	2.4	0.9
cg03608000	ZNF69	19	11998623	0.86	1.76E-24	5.5	3.1
cg02325951	FOXN3	14	89878619	-0.72	4.04E-24	62.4	73.0
cg17612569	GABPA;ATP5J	21	27107221	-3.07	7.86E-24	1.0	6.8
cg04874129	SLC6A2	16	55690873	0.75	9.82E-23	14.1	9.0
cg08906898	RBM39	20	34319899	-0.92	2.66E-22	87.2	92.6
cg04946709	LOC644649	16	59789030	-0.73	3.58E-22	70.1	78.8
cg02989351	YWHAQ	2	9770584	0.37	1.90E-20	16.5	13.2
cg12204423	PHF17	4	129732568	-0.82	1.94E-20	5.0	8.3
cg25304146	WBP11P1	18	30092971	-0.49	4.64E-20	57.6	64.9
cg22345911	CSNK1D	17	80231263	1.18	5.43E-20	5.2	2.7
cg01906879	GBE1	3	81811016	0.54	4.17E-19	8.5	6.0
cg06152526	PLLP	16	57290525	-0.62	5.25E-19	69.8	77.5
cg04190002	SHANK3	22	51113604	0.40	5.35E-19	33.8	27.9
cg06644124	ZNF281	1	200379083	0.50	6.63E-19	31.4	25.1
cg07628841	GPN1;CCDC121	2	27851430	0.30	7.09E-19	42.3	37.3
cg23001456	KIAA0664	17	2615074	0.86	7.47E-19	4.5	2.8
cg26213873	CTTNBP2NL	1	112939056	0.41	3.34E-18	8.6	6.6
cg25438440	CLDND1	3	98241168	0.55	3.72E-18	6.1	4.2
cg07816873	ERC1	12	1100472	0.54	3.88E-18	23.4	17.5
cg24016844	C1orf103	1	111506641	-0.43	1.28E-17	7.4	9.6
cg11841231	PARD3B	2	205543309	-0.68	5.39E-17	79.8	85.9
cg13323902	VTRNA1-1	5	140090859	0.53	6.23E-17	26.0	19.8
cg12900929	PRDM4	12	108154862	0.83	1.41E-16	11.2	6.9

Regression coefficients, β_{girl}, are reported in M-value scale for the change in methylation of girls relative to boys. Girl and boy mean methylation levels are shown on the β value or % methylation.

* Positions shown for hg19 (Genome Reference Consortium GRCh37) genome assembly.

lowest percentage of hits (4.1%). To assess whether the overrepresentation of hits in CpG islands and shores was due to the design of the 450K BeadChip, we compared the number of hits in each functional category with the number of CpG sites included in the assay. Both shores and CpG islands were significantly overrepresented among all autosomal hits compared to the 450K background (χ^2=486.1, P<0.01 and χ^2=95.5, P<0.01), while shelves and the open sea hits were underrepresented (each with P<0.01). For CpG sites that were hypermethylated in girls compared to boys, we also observed overrepresentation in CpG islands and shores, and underrepresentation in shelf and open-sea locations (all P<0.01). Sites that were hypomethylated in girls compared to boys were underrepresented in the open sea (30.3%, P<0.01) and shelves (5.6%, P<0.01). Hypomethylated sites were enriched at islands (χ^2= 6.53, P=0.01), but did not deviate significantly from the 450K representation of shores (χ^2= 3.42, P=0.06).

The 11,776 CpG hits differentially methylated between newborn boys and girls were found in 2,250 unique genes, and 1,430 (63.6%) of these genes were located on autosomes. Many genes contained multiple significant sites, with an average of 4.7 CpGs per gene and a maximum of 114 CpGs. However, the largest portion of sex-associated autosomal hits (30.4%) was located in intergenic regions and seen at lower than expected frequency in gene bodies (P<0.01) (Figure 2). Near gene transcription starting points (TSS200, 5'UTR, and first exons), all categories were either lower than 450K CpG design frequencies or did not deviate from them significantly. Further upstream (TS1500), hits that were hypermethylated in girls were significantly enriched (χ^2= 108.5, P<0.01) while those showing decreased methylation were underepresented (χ^2= 13.3, P<0.01). At the end of genes (3'UTR), hits that had higher methylation for girls were underrepresented (2.4%, P<0.01), while hits having higher methylation for boys did not deviate from expected 450K frequencies (3.6%, p = 0.97).

Examining the autosomal genes containing sex-associated DMPs for enrichment of particular gene ontology (GO) terms identified 278 pathways that were significantly enriched (FDR P<0.05 and at least 5 genes per GO term) (Table 5). These enriched GO terms fell into several broad categories including: 1) nervous system development, 2) behavior, 3) cellular development processes, and 4) cellular signaling and motility (Additional file 2).

Table 4. DMPs by gene and CpG island annotation.

		Sex DMPs		
	Autosomal 450K	All DMPs	Hypermethylated	Hypomethylated
Island	127,488	1,216	1,022	194
Shore	93,981	1,225	1,084	141
Shelf	37,490	124	95	29
Open Sea	141,170	466	308	158
Total	400,129	3,031	2,509	522
TSS1500	58,088	594	548	46
TSS200	44,647	298	240	58
5'UTR	36,279	239	207	32
1stExon	18,842	134	113	21
Body	135,011	766	607	159
3'UTR	14,073	79	60	19
Intergenic	93,189	921	734	187
Total	400,129	3,031	2,509	522

Counts of all, hypermethylated, and hypomethylated autosomal CpGs associated with sex in newborns by proximity to CpG island and gene features. Count of autosomal and all 450K CpGs shown for reference.

Table 5. The top 30 differentially enriched gene ontology pathways among hits for sex in autosomal CpGs.

			Genes		
GO category	Category name	Total	Changed	P-value	FDR
GO:0043005	Neuron Projection	651	67	2.22E-05	0.00E+00
GO:0007275	Multicellular Organismal Development	4621	353	9.06E-12	4.91E-09
GO:0097458	Neuron Part	980	104	8.83E-11	8.47E-09
GO:0044767	Single-Organism Developmental Process	5402	394	2.05E-10	1.24E-08
GO:0048856	Anatomical Structure Development	4828	359	2.43E-10	1.24E-08
GO:0007268	Synaptic Transmission	692	78	1.50E-09	5.09E-08
GO:0048731	System Development	4093	311	5.69E-10	1.54E-07
GO:0007270	Neuron-Neuron Synaptic Transmission	129	26	3.30E-09	4.76E-07
GO:0007267	Cell-Cell Signaling	1192	115	3.51E-09	4.76E-07
GO:0043167	Ion Binding	6038	422	2.02E-08	1.25E-06
GO:0007626	Locomotory Behavior	186	30	1.25E-07	3.19E-06
GO:0032879	Regulation Of Localization	2010	168	4.15E-08	4.50E-06
GO:0044707	Single-Multicellular Organism Process	6462	438	2.47E-07	5.04E-06
GO:0009653	Anatomical Structure Morphogenesis	2484	196	2.97E-07	5.05E-06
GO:0044765	Single-Organism Transport	3605	262	1.79E-06	2.36E-05
GO:0044763	Single-Organism Cellular Process	11949	739	1.85E-06	2.36E-05
GO:0009790	Embryo Development	998	94	5.15E-07	4.66E-05
GO:1902578	Single-Organism Localization	3791	271	4.22E-06	4.79E-05
GO:0007399	Nervous System Development	2053	171	4.97E-08	4.81E-05
GO:0065008	Regulation Of Biological Quality	3239	236	4.77E-06	4.86E-05
GO:0051703	Intraspecies Interaction Between Organisms	40	11	6.02E-06	5.58E-05
GO:0009887	Organ Morphogenesis	918	87	9.65E-07	6.55E-05
GO:0048513	Organ Development	2958	222	9.66E-07	6.55E-05
GO:0036477	Somatodendritic Compartment	562	60	1.10E-06	8.50E-05
GO:0044459	Plasma Membrane Part	2279	174	1.90E-06	9.14E-05
GO:0016043	Cellular Component Organization	5410	366	1.68E-05	1.43E-04
GO:0030425	Dendrite	378	45	1.15E-06	1.70E-04
GO:0051705	Multi-Organism Behavior	79	15	2.27E-05	1.78E-04
GO:0005883	Neurofilament	9	6	2.54E-06	1.88E-04
GO:0035637	Multicellular Organismal Signaling	751	86	7.94E-08	2.00E-04

Table 6. Results for the top 30 gene-annotated autosomal DMRs associated with sex in CHAMACOS newborns.

Gene Symbol(s)	Chrom-osome	Start Position*	End Position*	# of Probes	MaxFC for girls compared to boys	Minimum P-value
KRT77	12	53084709	53085323	4	0.09	1.34E-132
AMDHD2,ATP6V0C	16	2569911	2571449	9	0.24	7.17E-129
PPP1R3G	6	5085986	5087749	7	0.16	3.35E-78
CDK6	7	92461971	92464481	14	0.02	1.89E-65
TFDP1	13	114291977	114292740	10	-0.08	4.54E-57
CYP1A1	15	75018150	75019376	26	0.12	6.24E-57
C6orf174	6	127796287	127797286	7	0.12	2.64E-54
SMAD2	18	45456441	45458698	11	0.02	1.47E-51
ARID1B	6	157097800	157099375	9	0.07	6.59E-48
PEX10	1	2344089	2347015	26	0.07	6.53E-46
ATP5J,GABPA	21	27106793	27108257	11	-0.06	9.90E-45
SLC6A2	16	55689865	55691102	9	0.05	1.00E-43
A1BG,NCRNA00181	19	58861502	58862398	6	0.11	1.14E-43
PHF17	4	129731835	129733574	8	-0.03	2.69E-43
NUPL1	13	25874859	25876335	14	0.11	8.61E-42
PPFIA3	19	49636270	49636594	3	0.15	2.68E-40
ZNF69	19	11998457	11999148	11	0.03	4.58E-39
REM1,NCRNA00028	20	30071726	30073576	9	0.11	7.17E-39
YWHAQ	2	9770130	9771347	7	0.03	1.00E-37
LOC644649	16	59788728	59790180	7	-0.09	1.74E-35
FOXN3	14	89878584	89878733	5	-0.11	2.22E-35
PXDNL	8	52320944	52322341	9	0.07	7.31E-35
SHANK3	22	51112536	51114364	4	0.06	1.86E-34
SHANK2	11	70672365	70673256	11	-0.06	1.33E-33
RBM39	20	34319899	34319989	2	-0.06	1.74E-33
GIPC2	1	78511140	78512129	12	0.06	1.75E-33
CSNK1D	17	80230660	80232440	12	0.03	1.82E-33
NAPSA	19	50860534	50862121	8	-0.10	2.28E-33
FBXO47	17	37123638	37124558	9	0.08	2.94E-31
CCDC121,GPN1	2	27850964	27852231	14	0.05	5.71E-31

Max fold changes (FC) reported in M-value scale for the change in methylation of girls relative to boys.
* Positions shown for hg19 (Genome Reference Consortium GRCh37) genome assembly.

Sex-associated differentially methylated regions in newborns

Additionally, identification of groups of CpGs with 450K BeadChip methylation differences between newborn boys and girls was performed using the DMR-finding algorithm *DMRcate*.[122,126] This approach identifies and ranks DMRs by Gaussian kernel smoothing of results from linear models for individual CpGs that were adjusted for cell composition and array batch (see methods for details). A total of 3,604 DMRs were significantly associated with sex in newborns after correcting for multiple testing (FDR p<0.05; Table 6 and Additional files 3-4). These spanned 2,608 genes and contained a total of 22,402 unique CpGs. The number of sites within the DMRs ranged from 2 to 99 CpGs, with 50% of DMRs containing 5 or more CpGs and 25% having 8 or more. Further, DMR length averaged 863.8bp, and ranged from 3bp to 16.5kb. Figure 3 shows the DNA methylation levels for boys and girls at two example top DMRs. Figure 3A shows 7 CpG sites in a DMR that had higher methylation for girls in a region spanning the *PPP1R3G* transcription factor on chromosome 6. While Figure 3B shows a 8 CpGs from a DMR with lower methylation among girls in the promoter of *PIWIL1*, which is an important gene for stem cell proliferation and inhibition of transposon migration.[127,128]

As with DMPs, the majority of sex-associated DMRs had higher methylation in girls compared to boys (75.8%; Table S1). This was true for both autosomes and sex chromosomes when considered individually, with 83.8 % and 58.5% of DMRs having higher methylation in girls, respectively. However, a greater total number of DMRs identified were located on autosomes (2,471 or 68.6%) compared to the X chromosome. Similarly, the 70.3% of the genes covered by sex-associated DMRs were located on autosomes. Further, while the *DMRcate* method does not constrain all CpGs within a DMR to have the same direction of association with the predictor of interest, we found that the majority of DMRs had 100% concordance across

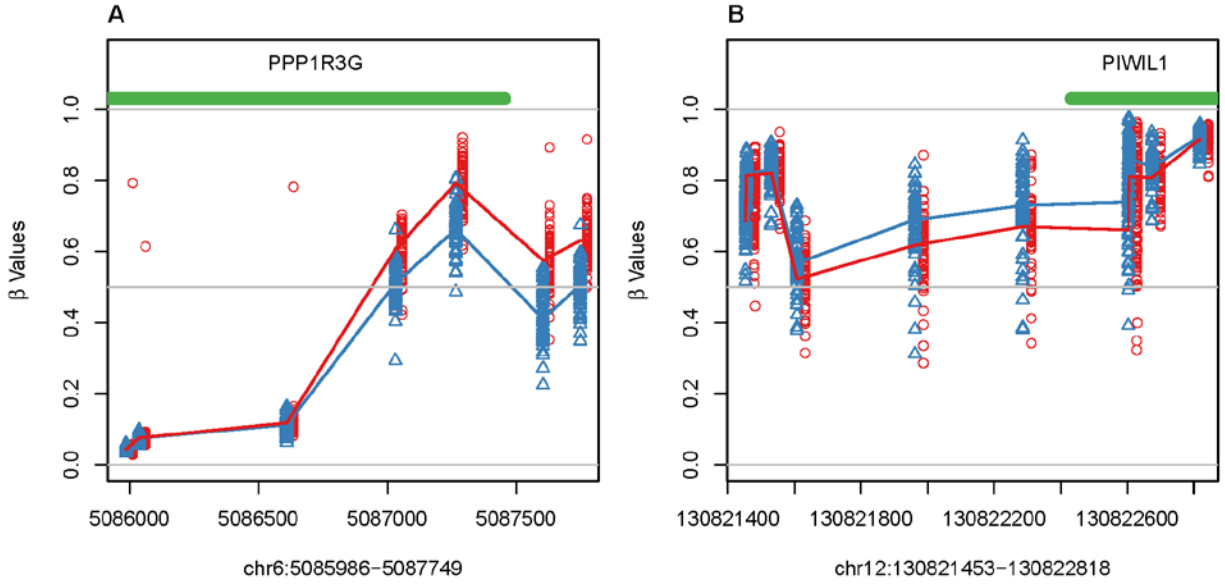


Figure 3. DNA methylation (β values) for CpG sites included in two top DMRs associated with child sex in newborns. One DMR (A) contains 7 CpG sites, is located on chromosome 6 and spans a 1,763bp region in the exon of PPP1R3G (chr6:5085986-5087749). The other (B) on chromosome 12 includes 8 CpGs over a 1,365bp region across the promoter and 1st exon of PIWIL1 (chr12:130821453−130822818). Girls are shown with red circles, boys with blue triangles, and median methylation per CpG by sex is shown by red and blue lines. Green lines show the genomic coordinates of exon regions for each gene shown.

CpGs in the direction of effect with sex (Additional file 5).

Comparison of the individual site results (DMPs) with the DMR findings revealed that of the 11,776 CpG sites associated with sex in the DMP analysis, 9, 941 (84.4%) were also included in a DMR. On autosomes, DMRs included 83.2% of sites found as sex-associated DMPs. Conversely, the DMRs added 12,461 total sites (11,719 on autosomes) that had not been found by DMP analysis alone.

Discussion

Here, we assessed methylation sex differences in newborns as determined by 450K BeadChip. Using reliable DCC estimates, our results are the first reported EWAS analysis by sex at birth that adjusted for confounding by cell composition. To our knowledge, we are also the first study to assess regions of differential methylation associated with sex in addition to considering all CpG sites individually. We identified a large numbers of X-chromosome CpG sites with higher methylation in girls, which is most likely attributable to X-inactivation.[124,129] Interestingly, we also demonstrated that a substantial number of autosomal sites and regions also appear hypermethylated in females (Figure 1 and Table 2).

To assess the consistency of our findings with those of prior analyses, autosomal CpG sites identified as differentially methylated by sex in the current analysis were compared to hits from the three most similar published studies to date (Table 7).[110,130,131] These studies differed from ours either in DNA methylation analysis platform (27K in McCarthy et al. 2014) or in tissue type used (Xu et al. 2014 in human prefrontal cortex and Hall et al. 2014 in pancreatic isolates). Although the meta-analysis performed by McCarthy et al. included some studies in umbilical cord blood, most of the studies were performed in adult tissues. Each study found between 184 and 614 autosomal CpG sites that were differentially methylated in association with sex (total of n=1,192 unique sites across all three studies). Our results replicated 428 (35.9%) of all hits, and 29.4% - 42.4% by different studies. Further, among replicated sites we observed 98.5-100% concordance in the direction of methylation differences. While there was substantial overlap between our autosomal sex-associated hits and these previously published results, 2,603 or 85.9% of our results are novel findings, some of which may be specific to the time point and tissue assessed (umbilical cord blood). Our larger number of hits is likely due to the increased coverage of the 450K BeadChip. In fact, when considered as a percentage of the number of sites analyzed, we observed a comparable portion of autosomal hits to that found by McCarthy and colleagues using the 27K platform (0.74% and 0.68% respectively; $P = 0.25$).

Importantly, the autosomal methylation increases we observed were most concentrated in CpG islands and shores (Figure 2A). As this trend was not evaluated in past studies, it should be explored and confirmed in additional datasets. Further, our findings that neurodevelopmental ontology terms were strongly enriched among our autosomal findings suggests that DNA methylation may contribute to differences in cognitive processes early in life. This is consistent with sex differences in brain development and rates of maturation that have previously been observed by magnetic resonance imaging in slightly older children (6-17 years of age)[132] and represent a possible regulatory mechanism requiring additional investigation.

Our autosomal hits included several genes already known to exhibit sex-specific functions. These included the male fertility and spermatogenesis related genes identified by McCarthy and colleagues (*DDX43*, *NUPL1*, *CRISP2*, *FIGNL1*, *SPESP1* and *SLC9A2*). One of our top hits showing increased methylation for girls (Table 3) included *SLC6A4*, Solute Carrier

Table 7. Comparison of CHAMACOS autosomal sex-associated CpG sites (n=3,031) with other published studies.

Study	Study Population	Tissue	Platform	Number of Autosomal Hits	Number of CHAMACOS Hits Replicated (%)	Percent of Concordance in Direction of Methylation Differences Among Replicated Sites
McCarthy et al. 2014	Meta analysis of 76 studies (n=6,795)	Multiple types	Illumina 27K	184	54 (29.4%)	100%
Xu et al. 2014	46 Caucasian adults	postmortem prefrontal cortex	Illumina 450K	614	260 (42.4%)	98.50%
Hall et al. 2014	87 Caucasian adults	pancreas	Illumina 450K	470	176 (37.5%)	100%

Family 6, that is involved in presynaptic reuptake of norepinephrine and has been implicated in several neurological disorders with sex-differences in prevalence.[133-135] Similarly, we observed novel sex differences in the *SHANK2* and *SHANK3* scaffolding protein genes that have been associated with autism spectrum disorders (Tables 3 and 6, Additional file 1).[136,137] Further, our hits included the homeobox containing transcription factor *EMX2*, Empty Spiracles Homeobox2, that is required for sexual differentiation and gonadal development[138] and we found to be hypermethylated among girls (Additional file 1).

The DMR analysis confirmed several trends observed by analyzing CpGs individually. In particular, DMR results again showed that girls tend to exhibit hypermethylation compared to boys. Also, many CpGs found to be autosomal DMPs were separately identified as being located within sex-associated DMRs. Besides confirming many of the findings in the DMP analysis, the application of DMR-finding substantially expanded the number of CpG sites considered significant. These results demonstrate that considering methylation over regions rather than single CpG sites may be a more effective way to identify differentially methylated sites and genes of interest.

Conclusions

We confirmed and expanded previously identified trends in autosomal and X-chromosome methylation sex differences during a previously unstudied window child development, immediately after birth, likely critical in establishing long term health. This strategy to assess epigenetic perturbation as near as possible to the prenatal period remains a high priority in light of the fetal origins of human disease hypothesis.[39,101,102,139]

Methods

Study population

The CHAMACOS study is a longitudinal birth cohort study of the effects of exposure to pesticides and environmental chemicals on the health and development of Mexican-American children living in the agricultural region of Salinas Valley, CA. Detailed description of the CHAMACOS cohort has previously been published.[104,105] Briefly, 601 pregnant women were enrolled in 1999-2000 at community clinics and 527 liveborn singletons were born. Follow up

visits occurred at regular intervals throughout childhood. For this analysis, we include the subset of subjects that had both 450K BeadChip data and differential cell count analysis available at birth (n=111). Mothers retained in the study subset had a mean age of 25.8 years (± 5.1 SD) at time of delivery. Study protocols were approved by the University of California, Berkeley Committee for Protection of Human Subjects. Written informed consent was obtained from all mothers.

Blood collection and processing

Cord blood was collected and stored in both heparin coated BD vacutainers (Becton, Dickinson and Company, Franklin Lakes, NJ) and vacutainers without anticoagulant at the same time. Blood clots from anticoagulant-free vacutainers were stored at -80°C and used for isolation of DNA for DNA methylation analysis. Heparinized cord blood was used to prepare whole blood cell slides blood using the push-wedge blood smearing technique[75] and stored at -20°C until staining for differential white blood cell count.

DNA preparation

DNA isolation was performed using QIAamp DNA Blood Maxi Kits (Qiagen, Valencia, CA) according to manufacturer's protocol with small, previously described modifications.[71] Following isolation, all samples were checked for DNA quality and quantity by Nanodrop 2000 Spectrophotometer (Thermo Scientific, Waltham, MA). Those with good quality (260/280 ratio exceeding 1.6) were normalized to a concentration of 50ng/ul.

450K BeadChip DNA methylation analysis

DNA samples were bisulfite converted using Zymo Bisulfite Conversion Kits (Zymo Research, Irvine, CA), whole genome amplified, enzymatically fragmented, purified, and applied to Illumina Infinium HumanMethylation450 BeadChips (Illumina, San Diego, CA) according to manufacturer protocol. Locations of samples from boys and girls were randomly assigned across assay wells, chips and plates to prevent any batch bias. 450K BeadChips were handled by robotics and analyzed using the Illumina Hi-Scan system. DNA methylation was measured at 485,512 CpG sites.

Probe signal intensities were extracted by Illumina GenomeStudio software (version XXV2011.1, Methylation Module 1.9) methylation module and back subtracted. Systematic QA/QC was performed, including assessment of assay repeatability, batch effects using 38 technical replicates, and data quality established as previously described.[140] Samples were retained only if 95% of sites assayed had detection P> 0.01. Color channel bias, batch effects and difference in Infinium chemistry were minimized by application of All Sample Mean Normalization (ASMN) algorithm,[140] followed by Beta Mixture Quantile (BMIQ) normalization.[88] Sites with annotated probe SNPs and with common SNPs (minor allele frequency >5%) within 50bp of the target identified in the MXL (Mexican ancestry in Los Angeles, California) HapMap population were excluded from analysis (n=49,748). Probes where 95% of samples had detection P>0.01 were also dropped (n= 460). Since our analysis was focused on CpG sites associated with sex, we excluded sites on the Y chromosome (n=95) and X-chromosome cross-reactive probes (n=29,233) identified by Chen and colleagues.[73] Remaining CpGs included 410,072 sites for analysis of sex. Methylation values at all sites were logit transformed to the M-value scale to better comply with modeling assumption,[74] but results were reported as β values or percent methylation when possible for ease of interpretation.

Differential cell counts

Whole blood smear slides were stained utilizing a DiffQuik® staining kit, a modern commercial variant of the Romanovsky stain, a histological stain used to differentiate cells on a variety of smears and aspirates. This staining highlights cytoplasmic details and neurosecretory granules, which are utilized to characterize the differential white blood count. The staining kit is composed of a fixative (3:1 methanol: acetic acid solution), eosinophilic dye (xanthene dye), basophilic dye (dimethylene blue dye) and wash (deionized water). For consistency and to ensure the best results the slides were all fixed for 15 minutes at 23°C (room temperature), stained in both the basophilic dye and eosinophilic dye for five seconds each and washed after each staining period to prevent the corruption of the dye.

Slides were scored for white blood cell type composition by Zeiss Axioplan light microscope with 100× oil immersion lens. Scoring was conducted at the perceived highest density of white blood cells using the standard battlement track scan method, which covers the entire width of a slide examination area. Counts for each of the five identifiable cell types (lymphocytes, monocytes, neutrophils, eosinophils, and basophils) were recorded by a dedicated mechanical counter. At least 100 cells were scored for each slide following validation of reproducibility by the repeated scoring of 5 sets of 100 cells from the same slide (CV≤5%).

DMP analysis

Association between sex at birth and differential 450K DNA methylation at individual CpGs was performed by linear regression, adjusting for DCC variables and analysis batch. This analysis was performed using R statistical computing software (v3.1.0).[141] Although DCC estimates were not significantly associated with sex, we chose to include them in the model because likelihood ratio tests showed that including them improved model fit for more than 2,000 of the CpG sites assessed by 450K BeadChip. We also examined gestational age and subject birthweight as possible covariates since both have been shown to be associated with DNA methylation,[142] and performed sensitivity analysis to assess their potential impact. However, neither was associated with child sex or contributed to improved model fit. P-values were corrected for multiple testing using a Benjamini-Hochberg (BH) FDR threshold of 0.05.[143]

Enrichment of annotated genomic features

Comparison of sex-DMP results to annotated function categories, including relation to genes (TSS1500, TSS200, 5'UTR, 1stExon, Body, 3'UTR, Intergenic) and CpG islands (Island, Shore, Shelf, Open Sea), was performed using UCSC Genome Browser annotations supplied by Illumina. A χ^2 test of independence with 1 degree of freedom was used to determine whether there was evidence of enrichment among DMP results (P value < 0.05).

DMR analysis

Identification of sex-associated DMRs was performed using the method described by Peters et al.[122] and implemented in the *DMRcate* Bioconductor R-package.[126] The approach begins by fitting a standard *limma* linear model to all CpG sites in parallel.[144] This model was parameterized identically to the DMP analysis with sex as the binary predictor of interest, adjusting for DCC variables and analysis batch. The CpG site test statistics were then smoothed by chromosome according to the *DMRcate* defaults, which employs a Gaussian kernel smoother with bandwidth λ = 1,000 base pairs (bp) and scaling factor C=2. The resulting kernel-weighted

local model fit statistics were compared to modeled values using the method of Satterthwaite[145] to produce p-values that are adjusted for multiple testing using a BH FDR threshold of 0.05.[143] Regions or DMRs were assigned by grouping FDR significant sites that are a maximum of λ bp from one another and contain at least two or more CpGs. Under this method, CpGs are collapsed into DMRs without considering the direction of the association with the predictor (i.e. sex). The minimum BH-adjusted p-value within a given DMR is taken as representative of the statistical inference for that region and the maximum fold change in methylation values (here on the M-value scale) summarizes the effect size.

Gene ontology analysis

Gene ontology term enrichment analysis was performed by DAVID,[146,147] WebGestalt (WEB-based Gene SeT AnaLysis),[148] and ConsensusPathDB,[149] using hypergeometric distribution to assess enrichment significance. Visualization of results and GO term categorization by semantic similarity dimension reduction was performed by REVIGO.[150]

Additional Files

All additional files can be accessed only and are permanently hosted at the url:

http://bmcgenomics.biomedcentral.com/articles/10.1186/s12864-015-2034-y

They include:

Additional file 1: ***Sex-associated autosomal DMPs.*** Results for all significant autosomal DMPs associated with sex in CHAMACOS newborns ranked by P value.

Additional file 2: ***Visualization of enriched gene ontology categories***. Gene ontology categories significantly enriched (PBH <0.05) in genes with sex-modified autosomal CpG sites.

Additional file 3: ***Summary of sex-associated DMRs.*** Number of DMRs significantly hyper- and hypo-methylated in newborn girls compared to boys at FDR multiple testing threshold (q<0.05), for all DMRs, and then stratified by autosomes and X chromosome.

Additional file 4: ***Sex-associated DMRs.*** Results for all significant DMRs associated with sex in CHAMACOS newborns ranked by P value.

Additional file 5: ***Distribution of effect direction concordance within DMRs.*** Histogram of percent concordance of direction of sex-association for CpGs within identified DMRs.

CHAPTER 7

Prenatal PBDE exposure and altered patterns of DNA methylation regions in cord blood.

ABSTRACT

Polybrominated diphenyl ethers (PBDEs) are flame retardants widely used in U.S. consumer products between the 1970's and 2004. Prenatal PBDE exposure has been associated with adverse neurobehavioral outcomes in children and thyroid disruption. DNA methylation, which undergoes rearrangement *in utero* and is susceptible to environmental insult, may be a biological mechanism mediating these effects.

To assess whether prenatal PBDE exposure was associated with altered patterns of cord blood DNA methylation in newborns from CHAMACOS, a Mexican-American birth cohort.

We examined the relationships between serum PBDE levels *in utero* and differential DNA methylation at individual CpG positions (DMPs) and genomic regions with multiple CpGs (DMRs), measured by 450K BeadChip in cord blood (N=313).

Assessment of genomic regions identified between 6 and 48 DMRs associated with different PBDE measures. BDE's-47, -99 and Σ4BDE had fewer (from 6 to 9), mostly hypomethylated DMRs. Prenatal BDE-100 and -153 levels were associated with more DMRs (11 and 48 respectively) and the majority hypermethylated. Many CpGs and DMRs were consistently identified across multiple PBDE exposure measures. However, naïve analysis of DNA methylation as independent CpGs did not identify DMPs in relation to PBDE exposure (FDR<0.05).

This is the first study of site specific and regional DNA methylation differences at birth associated with prenatal PBDE exposure. DMR analysis significantly expanded differentially methylated genes and pathways, including those related to neurological activity and thyroid function, suggesting novel epigenetic modifications potentially mediating the effects of PBDE exposure on health outcomes.

Introduction

Polybrominated diphenyl ethers (PBDEs) are a class of persistent organic pollutants (POPs) that have been commonly used as flame retardants in consumer products. From the 1970's through 2004, the penta- and octaBDE technical mixtures (comprised of congeners BDEs -47 -99 -100 and -153) was commonly applied to polyurethane foam and electronic devices to meet California's furniture flammability standards.[151,152] This has contributed to significantly higher serum PBDE levels in children from this state compared to children living in other areas of the United States (US),[153,154] as well as children from Mexico and Europe.[155,156]

PBDEs are highly lipophilic, bioaccumulative and have long half-lives.[157] While young children have some of the highest serum levels due to unique sources of postnatal exposure (e.g. breastfeeding, increased exposure to house dust, etc.), PBDEs are known to cross the placenta and can lead to exposure during crucial windows of cell differentiation during prenatal development.[153,158] To date, prenatal PBDE exposure has been associated with child neurodevelopment,[15,159,160] child BMI[161] and altered thyroid hormone levels during pregnancy.[162]

Epigenetic markers, particularly DNA methylation, regulate gene transcription and may mediate environmental impacts on health. The fetal DNA methylome undergoes rearrangement *in utero* and may be disturbed by xenobiotic exposures.[4,28,107] Thus, environmental alteration of the methylome is a possible mechanism contributing to the developmental origins of human disease.[163,164]

An emerging literature suggests that exposure to POPs may alter DNA methylation. For example, polychlorinated biphenyls (PCBs) and several types of organochlorines (OCs) have been associated with DNA methylation in LINE-1 and ALU repetitive elements in four independent cross-sectional studies of adults.[62,165-167] Exposure to dichlorodiphenyl trichloroethane (DDT) has been seen to disrupt DNA methylation at several CpG islands in the hypothalamus of Wistar rats.[168] Additionally, alterations of DNA methylation patterns in the sperm of male rodent offspring following maternal exposure to DDT persisted for three generations, demonstrating the potential for trans-generational epigenetic reprogramming.[169]

Few studies have looked specifically at PBDE exposure and DNA methylation. One in rats found that pre- and perinatal exposure to BDE-47 was associated with lower methylation in the frontal brain lobe at the L1Rn repetitive element.[170] In the only study in humans, we found that maternal PBDE serum concentrations were associated with differential DNA hypomethylation in the LINE-1 and ALU repetitive elements in cord blood and blood of 9 year old children adjusting for blood cell composition.[171]

Here, we report on the results of the first study to date to explore association between PBDE exposure and DNA methylation across the genome. Specifically, we examine the effect of prenatal PBDE serum levels from a large birth cohort, the Center for the Health Assessment of Mothers and Children of Salinas (CHAMACOS) study, on umbilical cord DNA methylation assessed by 450K BeadChip. In addition to interrogating individual DMPs, we seek to identify PBDE-DMRs using the 'DMRcate' methodology.[122]

Methods

Study participants

Subjects were participants in the CHAMACOS longitudinal birth cohort study examining the impact of pesticide and other environmental exposures on the health and development of

Table 1. Demographic characteristics of CHAMACOS newborns and their mothers, N=313.

	N	%
Child sex		
Boy	156	49.8
Girl	157	50.2
Maternal years in U.S.		
≤1	77	24.6
2–5	94	30
6–10	62	19.8
11+	43	13.7
Entire life	37	11.8
Maternal age at delivery (years)		
18-24	155	49.7
25-29	99	31.7
30-34	42	13.5
35-45	16	5.1
Family income		
Below poverty threshold	195	62.3
Above poverty threshold	118	37.7
Maternal pre-pregnancy BMI (kg/m2)		
Normal	117	38.6
Overweight	122	40.3
Obese	61	20.1
Smoking during pregnancy		
No	294	93.9
Yes	19	6.1

Mexican-American children living in the Salinas Valley, an agricultural region. Detailed description of the CHAMACOS cohort has previously been published.[104,105] Briefly, 601 pregnant women were enrolled in 1999-2000 at community clinics and 527 liveborn singletons were born. Women were interviewed at ~13 and ~26 weeks gestation, and after delivery. Characteristics of participating mothers and children are shown in Table 1. Mothers were mostly young (25.4 ± 4.8 years old), Mexican-born, and overweight or obese prior to pregnancy (with an average body mass index 26.8 ± 5.1). Very few reported having smoked at any time point during their pregnancy (6.1%) and most quit by the time their children were born. Their offspring were equally distributed by sex. The majority of children were born at term (93.9%) with an average birthweight of 3.5 ± 0.5 kg. For this study, 313 mother-child pairs with both prenatal PBDE exposure and 450K BeadChip data available in cord blood were included for analysis. The sub-cohort used for this study did not differ from the entire CHAMACOS cohort in demographic or exposure characteristics. Study protocols were approved by the University of California, Berkeley Committee for Protection of Human Subjects. Written informed consent was obtained from all mothers.

Blood collection and processing

Maternal blood samples were obtained by a trained phlebotomist at around 27 weeks from last menstrual period (mean±SD=26.9 ± 3.2 weeks) and at delivery. Cord blood specimens were collected by hospital staff at time of delivery. Blood samples were collected in vacutainers both with and without heparin (green- and red-top), and separated into clots, serum, buffy coats, and stored at -80°C until analysis.

PBDE exposure measurement

Maternal serum samples were analyzed at the Centers for Disease Control and Prevention (Atlanta, GA) for PBDEs and results were previously published.[160,162] Our analysis utilized the four PBDE congeners with the highest detection frequency (>97%): BDE-47, -99, -100, -153 and their sum. PBDE concentrations are expressed relative to levels of total serum lipids. Limits of detection (LODs) ranged from 0.2 to 0.7 ng/g lipid for all congeners except BDE-47, which

ranged from 0.3 to 2.6 ng/g lipid. Values below the LOD were assigned machine read values when available and otherwise were imputed from a log-normal distribution by the "fill-in" method described by Lubin and colleagues.[172] For subjects without prenatal samples (n=76), PBDE levels were measured in delivery blood and back-estimated based on paired samples at both 26 weeks and delivery ($r \geq 0.98$, $p < 0.001$) as previously described.[173] Measurements of umbilical cord PBDE levels were not available for this study, however prenatal maternal are considered representative surrogates of fetal levels.[174] Table 2 shows summaries of the distributions for each of the four prenatal PBDE serum congeners measured in our sample and their sum. The study distribution did not differ significantly from the full CHAMACOS sample. As previously reported[153,160,173] BDE-47 had the highest level (median= 15.3 ng/g lipid) in our sample, which was more than triple that of the next highest detected congener (BDE-99, median = 4.1 ng/g lipid)

DNA preparation

DNA isolation was performed from banked non-heparinized umbilical cord blood clot samples using QIAamp DNA Blood Maxi Kits (Qiagen, Valencia, CA) according to the manufacturer's protocol with minor modifications as previously described.[71]

450K BeadChip DNA methylation analysis

DNA aliquots of 1ug were bisulfite converted using Zymo Bisulfite conversion Kits (Zymo Research, Orange, CA). DNA was whole genome amplified, enzymatically fragmented, purified, and applied to the Illumina Infinium HumanMethylation450 BeadChips (Illumina, San Diego, CA) according to the Illumina methylation protocol.[41,42] 450K BeadChips were handled by robotics and analyzed using the Illumina Hi-Scan system. DNA methylation was measured at 485,512 CpG sites. Samples were run on 36 distinct BeadChips across 5 plate batches and were assigned randomized locations across chips and plates.

Probe signal intensities were extracted by Illumina GenomeStudio software (version XXV2011.1, Methylation Module 1.9) methylation module and background subtracted. Systematic QA/QC was performed, including assessment of assay repeatability and batch effects using 38 technical replicates as previously described.[140] Color channel bias, batch effects and difference in Infinium chemistry were minimized by application of the All Sample Mean Normalization (ASMN) algorithm,[140] followed by Beta Mixture Quantile (BMIQ) normalization.[88] Samples were retained only if 95% of sites assayed had detection P-values > 0.01. Sites with annotated probe SNPs and with common SNPs (minor allele frequency >5%) within 50bp of the target identified in the MXL (Mexican ancestry in Los Angeles, California) HapMap population and X-chromosome cross-reactive probes (n=29,233) identified by Chen and colleagues[73] were excluded from analysis. Probes where 95% of samples had detection P-values >0.01 were also dropped (n= 460). This left 410,121 CpGs for analysis. Methylation values at all sites were logit transformed to the

Table 2. Distributions of concentrations of PBDE congeners and their sum by lipid weight (ng/g lipid) in serum of mothers from the CHAMACOS study during pregnancy (n=313).

	Median	IQR	Min	Max
BDE-47	15.34	(7.99 , 27.11)	0.45	760.95
BDE-99	4.06	(2.23 , 6.86)	0.25	297.60
BDE-100	2.62	(1.51 , 4.80)	0.19	138.30
BDE-153	2.25	(1.30 , 4.07)	0.12	96.88
Σ4BDEs	24.78	(13.77 , 42.13)	2.35	1293.74

M-value scale to better comply with modeling assumption.[74]

To reduce the potential to extreme outlying methylation values to drive associations in statistical models, we removed methylation observations that were greater or less than three times the interquartile range for a given probe. The majority of probes (78.6%) had no observations removed and the remaining mean N per probe was 312 (range from 234 to 313); see Figure S1.

Statistical analysis

Analysis of differentially methylated positions (DMPs) associated with each of the five prenatal PBDE exposure measures was performed by generalized linear regression, adjusting for maternal years living in the US and 450K batch. PBDE exposure measurements were $\log_{10}$ transformed to approximate a normal distribution. Adjustment covariates were selected from among factors previously reported to show association with child DNA methylation level, including maternal age, parity, smoking during pregnancy, child sex, maternal years in the US, pre-pregnancy body mass index, gestational age, birthweight and white blood cell composition. White blood cell proportions have been previously described.[121] Bivariate associations between these factors and both PBDE exposure measurements and previously described DNA methylation of LINE-1 and ALU repetitive elements in cord blood[171] were used to assess potential for confounding. For results of these bivariate analyses, by linear regression, Student's T-test, and ANOVA where appropriate, see Figure S2. Maternal years living in the US were adjusted for as it was the only factor to show consistent association with PBDE exposures. P-values were corrected for multiple testing using a Benjamini-Hochberg (BH) FDR threshold of 0.05.[143]

Identification of genomic regions with more than one contiguous CpG site showing associations with PBDE exposure, or differentially methylated regions (DMRs), was performed by the method described by Peters et al[122] and implemented in the *DMRcate* Bioconductor R-package.[126] This model had identical parameterization to the linear model implemented for DMP analysis, with PBDE exposure measures as the predictors of interest and adjusting for maternal years living in the US and 450K batch. Sensitivity analysis adjusting for white blood cell composition[50] was also performed. This method first fits a standard *limma* linear model to all CpG sites in parallel[144] and then smooths the resulting CpG site test statistic by a Gaussian kernel. We implemented the *DMRcate* default smoothing parameters with bandwidth $\lambda = 1{,}000$ base pairs (bp) and scaling factor C=2. The resulting kernel-weighted local model fit statistics are then compared to modeled values using the Satterthwaite method[145] to produce p-values that are adjusted for multiple testing using a BH FDR threshold of 0.05.[143] DMRs are assigned by grouping FDR significant sites that are a maximum of λ bp from one another and contain at least two or more CpGs. Under this method, CpGs are collapsed into DMRs without considering the direction of the association with the exposure. The minimum BH-adjusted p-value within a given DMR is taken as representative of the statistical inference for that region and the maximum fold change in methylation values (here on the M-value scale) summarizes the effect size.

Results

Linear models for association between prenatal PBDE serum levels and individual CpG sites revealed no significant DMPs after accounting for multiple testing (FDR p<0.05). P-values at analyzed CpGs generally exhibited genomic deflation, with λ's ranging from 0.63 to 0.74, except in models with BDE-153 as the predictor, which showed slight inflation (λ = 1.03; see Figure S2). For all four congeners and their sum, coeffients for a small majority of CpG sites showed hypermethylation in relation to PBDE exposure, ranging from 53.3% of CpGs for BDE-47 to 66.3% for BDE-153. However, most CpGs with extreme p-value ($p< 1x10^{-4}$) were hypomethylated (from 82.4% for BDE-47 to 95.2% for BDE-99). An example volcano plot of the regression coefficients and p-values for models between BDE-153 levels and DNA methylation highlights these trends (Figure 1).

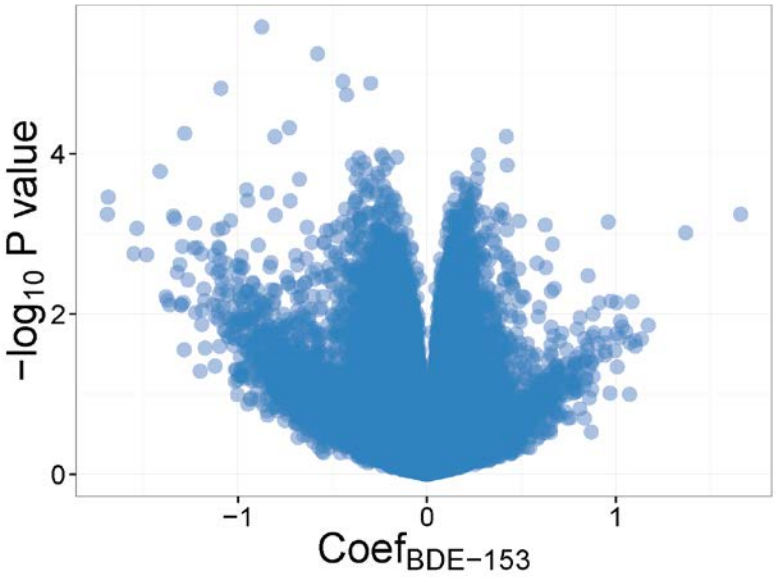


Figure 1. Volcano plot of the regression coefficients for association between prenatal log_{10}BDE-153 and newborn DNA methylation in M-values at 410,121 CpG sites by -log_{10} P-values. A majority of CpGs (272,095 or 66.3%) have positive coefficients in relation to BDE-153. However, 9 out of 10 extreme p-values ($p< 1x10^{-4}$) are from sites with decreasing methylation relative to BDE-153. As an example, the CpG with the most extreme p-value (cg14279439, $p= 2.62x10^{-6}$) had coefficient = -0.87, which for an IQR increase in log_{10}BDE-153 represents a -0.63decrease in M-value.

We also analyzed groups of CpGs using the DMR-finding algorithm *DMRcate*.[122] This method identifies and ranks DMRs by kernel smoothing results from linear models for individual CpGs that were adjusted for maternal years in the US and array batch (see Methods for details). We identified multiple DMRs for each PBDE congener and their sum (Table 3). BDE-153 produced the greatest number of significant relationships with DNA methylation in cord blood: 48 significant DMRs after correcting for multiple testing (FDR p<0.05). Figure 2 shows the p-values for individual CpG sites produced by the *DMRcate* method by chromosome. For other prenatal PBDE measures, the number of significant DMRs ranged from 27 for BDE-47 to 55 for BDE-100. The top 25 most significant PBDE-DMRs are shown in Table 4 (for complete results see Table S1). For all PBDE-DMRs identified, the mean number of CpGs included within a given DMR was 6.4 and varied from 2 to 20. Further, PBDE-DMRs had mean length of 48.4bp and ranged anywhere from 2 to 1,312bp.

For BDE's -47, -99 and Σ4BDE, the majority of DMRs identified in cord blood showed hypomethylation in association with prenatal PBDE concentrations, with percentages of DMRs hypomethylated at 66.7%, 77.8%, and 62.5% respectively. However, for the congeners with the most DMRs identified, BDE's-100 and -153, DMRs were mostly hypermethylated (respectively 54.5% and 79.2%). Although the *DMRcate* method does not constrain all CpGs within a DMR to

Table 3. Number and direction of PBDE associated DMRs and unique CpGs and genes by BDE congener.

	Number of DMRs	DMRs with decreased methylation (%)	DMRs with increased methylation (%)	Unique CpGs in all DMRs	Unique Genes in all DMRs
BDE-47	6	4 (66.7)	2 (33.3)	44	8
BDE-99	9	7 (77.8)	2 (22.2)	48	9
BDE-100	11	5 (45.5)	6 (54.5)	75	10
BDE-153	48	10 (20.8)	38 (79.2)	303	49
Σ4BDEs	8	5 (62.5)	3 (37.5)	51	10

have the same direction of association with the predictor of interest, we found that all but three DMRs had 100% concordance across CpGs in the direction of effect PBDE exposure (see Table S1). Many of the CpGs within PBDE-DMRs were identified in response to multiple PBDE congeners. Table 5 shows the count of DMR-CpGs found in common between in results from all combinations of prenatal PBDE exposure measures considered. Additionally, in sensitivity analysis adjusting for white blood cell composition, we found that 46% of probes were retained in DMRs called for BDE-153, 89% for BD-100, and 100% for BDEs-47, 99, and Σ4BDE.

The plots in Figure 3 show methylation at CpGs contained in two DMRs strongly associated with prenatal PBDE concentrations. Eleven CpGs included in a DMR hypermethylated in relation to all PBDE exposure measures tested spanned the tail of the coding region of *NRBP1* and the promoter and beginning of the coding region of *KRTCAP3* (Figure 3A). Figure 3B presents an example of a hypomethylated DMR with 7 CpGs located in the gene body of *S100A13*. In both DMRs, the absolute methylation levels vary across CpGs, however, the difference by exposure status is observed consistently.

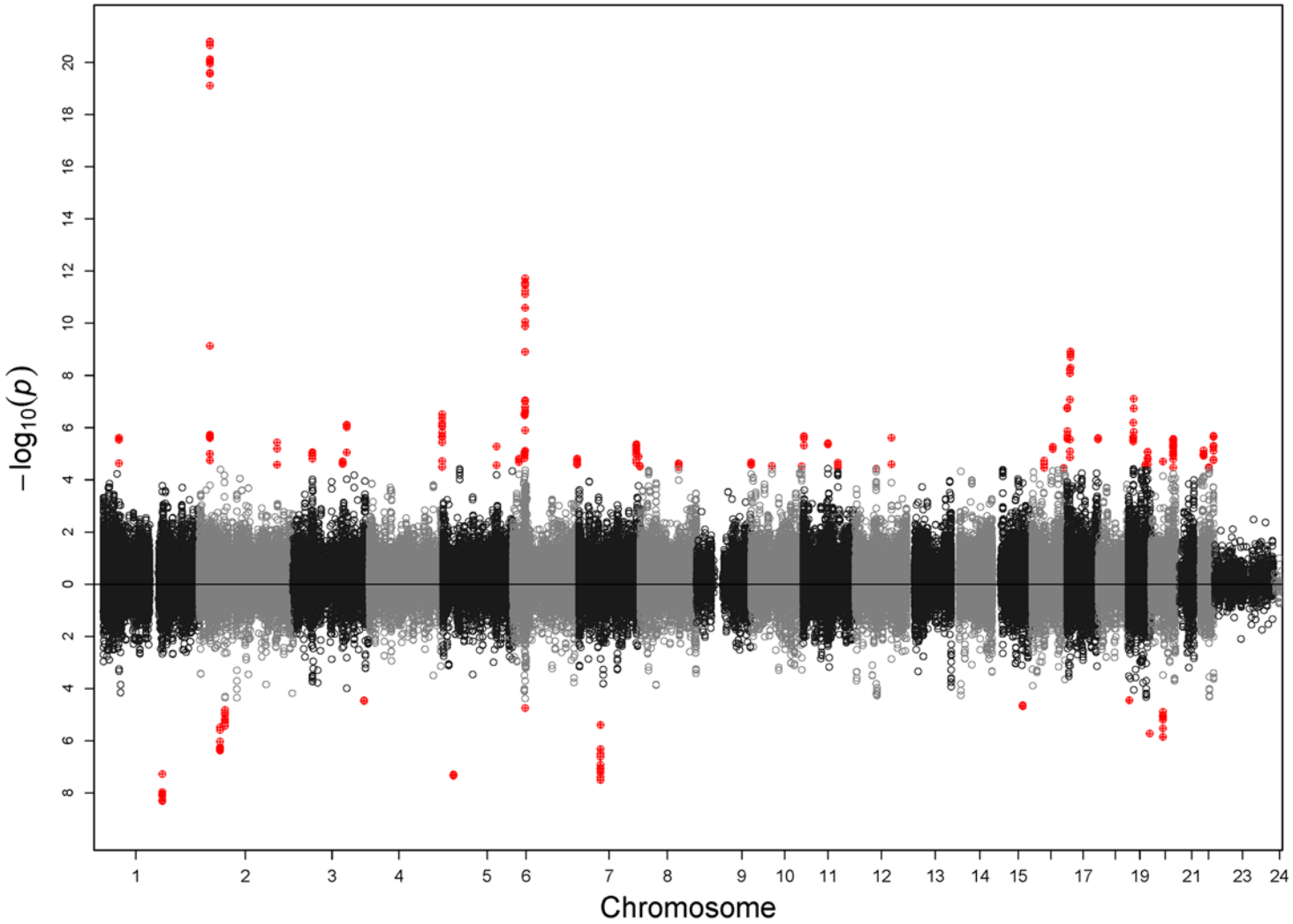


Figure 2. Manhattan plot for association between prenatal serum BDE-13 and DNA methylation as determined by the *DMRcate* DMR model, adjusting for batch. P-values are shown for all 450K CpGs prior to DMR-clustering. Sites having higher methylation levels in relation to prenatal serum BDE-153 levels are plotted above the x-axis and those with lower levels plotted below. CpGs meeting FDR multiple testing threshold of (P<0.05) shown in red.

Discussion

We assessed differences in cord blood DNA methylation at over 410,000 CpGs in association with exposure to PBDEs during prenatal development. To our knowledge, this is the first study to examine DNA methylation broadly across the genome in relation to PBDE exposure. When we considered groups of CpG sites jointly as DMRs related to PBDE concentrations *in utero*, we observed a substantial number of genomic regions exhibiting significant associations after adjusting for multiple testing. Differential methylation observed in genes, including those related to neurological activity and thyroid function, suggest novel epigenetic effects of PBDE exposure that may relate to or mediate health outcomes.

There is a growing body of research identifying DMRs associated with exposure or phenotypes.[175-177] In comparison to DMPs, analysis of DMRs provides important advantages. For instance, it utilizes the patterns of co-correlation between nearby CpG sites essentially taking advantage of existing epigenomic structure.[178,179] Additionally, considering regions rather than individual CpG sites also improves statistical power.[180,181] While some epidemiologic studies have begun to implement other DMR-finding algorithms,[177,182,183] few have applied *DMRcate*,[184,185] which has showed improved performance at DMR predictions compared to other methods (e.g. BumpHunter, comb-p, etc.) in simulated and whole-genome bisulfite sequencing data.[122] Our findings underscore the importance of considering the combined impact of CpGs in functional groupings when modeling the effect of exposure on DNA methylation.

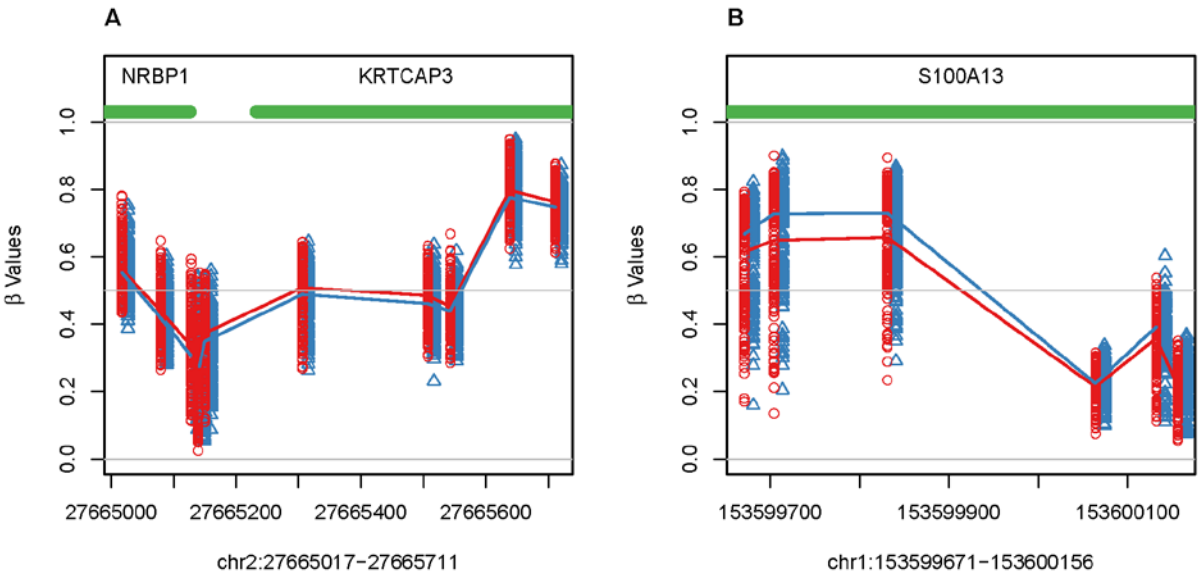


Figure 3. DNA methylation (β values) for CpG sites included in two top DMRs associated with prenatal PBDEs in cord blood. One DMR hypermethylated in relation to BDE-99 levels (A) contains 11 CpG sites, is located on chromosome 2 and spans a 694bp region spanning both *NRBP1* and *KRTCAP3* (chr2: 27665017-27665711). The other DMR hypermethylated relative to BDE-47 (B) on chromosome 1 includes 6 CpGs over a 485bp region in the gene body of *S100A13* (chr1: 153599671-153600156). For both plots, the methylation levels at each of the CpGs shown dichotomized by exposure groupings: those with above median BDE levels are plotted as red triangles, while those with below median exposure are presented as blue circles.

The DMRs we identified were located in biologically relevant genes given the health effects observed in association with BDE exposure to date. Neurotoxicity, neurodevelopment deficits, and cancer are leading health effects of concern related to BDE-exposure and several studies including our own have reported associations with these outcomes.[15,159,160,186] In our analysis, we observed the strongest and most consistent association across all congeners and their sum with hypermethylation of a DMR containing the *NRBP1* (Nuclear Receptor Binding Protein) gene across 10-12 CpG sites. This is a pseudokinase with a role in cell homeostasis and protein regulation.[187] *NRBP1* was recently shown to act as a tumor suppressor in cancer and has been associated with prostate and breast cancer.[188,189] We also found a consistent association of hypomethylation of a DMR containing *CDH9* (cadherin 9) with prenatal exposure to all 4 congeners and their sum. *CDH9* is a calcium-dependent cell adhesion protein in the central nervous system. A SNP residing between CDH9 and CDH10 has been associated with Autism spectrum disorder, social communication spectrum phenotypes,[190] and risk of suicide.[191] We also observed differential hypermethylation of *NTN1* (netrin 1). Along with other netrins, *NTN1* is involved in axon guidance and cell migration during development.[192] Recent studies report that it also participates in maintenance of the blood brain barrier in response to neuroinflammation.[193] Genetic variants of netrin and other genes involved in axon guidance have been associated with Parkinson's disease and amyotrophic lateral sclerosis.[194,195] DMRs containing *NTN1* were positively associated with prenatal exposure to both BDE-100 and BDE-153 and spanned across 5 CpGs.

Table 4. Results for the top 25 DMRs associated with 10-fold increase in prenatal maternal PBDE serum levels (ng/g lipid) in CHAMACOS cord blood.

Gene Symbol(s)	Chromosome	Start Position	End Position	# of Probes	BDE Congener	MaxFC	Minimum P-value
KRTCAP3,NRBP1	2	27664399	27665711	12	BDE-153	0.05	3.05E-16
KRTCAP3,NRBP1	2	27664399	27665711	12	BDE-100	0.05	8.87E-16
KRTCAP3,NRBP1	2	27664399	27665711	12	Σ4BDEs	0.04	1.84E-11
KRTCAP3,NRBP1	2	27664399	27665711	12	BDE-47	0.04	1.30E-09
LOC650226	7	56515510	56516504	10	BDE-47	-0.04	1.71E-08
GPSM3,NOTCH4	6	32164413	32165321	18	BDE-153	0.03	6.47E-08
LOC650226	7	56515510	56516504	10	Σ4BDEs	-0.04	9.90E-08
KRTCAP3,NRBP1	2	27665017	27665711	11	BDE-99	0.03	4.98E-07
LOC650226	7	56515510	56516504	10	BDE-99	-0.04	4.98E-07
LOC650226	7	56515510	56516504	10	BDE-100	-0.04	3.98E-06
NTN1	17	9018806	9019336	5	BDE-153	0.04	2.23E-05
SDHAP3	5	1594282	1595048	6	BDE-47	-0.06	4.82E-05
S100A13,S100A1	1	153599479	153600156	7	BDE-153	-0.08	7.23E-05
TMEM88	17	7758581	7759140	6	BDE-153	0.02	8.15E-05
CDH9	5	27038605	27038836	5	BDE-100	-0.04	1.48E-04
NOTCH4	6	32164801	32165321	9	BDE-100	0.02	1.74E-04
LOC650226	7	56515510	56516504	10	BDE-153	-0.04	3.50E-04
CDH9	5	27038605	27038836	5	BDE-153	-0.04	4.84E-04
SDHAP3	5	1594282	1595048	6	Σ4BDEs	-0.06	6.33E-04
MIR181C,MIR181D	19	13984091	13984736	3	BDE-153	0.03	7.14E-04
CDH9	5	27038605	27038836	5	Σ4BDEs	-0.03	9.71E-04
NXN	17	800431	800749	5	BDE-153	0.03	1.31E-03
PSORS1C3	6	31148332	31148748	13	BDE-100	0.03	1.37E-03
ZFP42	4	188916496	188917251	13	BDE-153	0.02	1.76E-03
PSORS1C3	6	31148332	31148748	13	BDE-153	0.03	1.76E-03

FC – fold change in DNA methylation M-value per $\log_{10}$ unit increase in PBDE conger

DMR results were biologically relevant given the existing epidemiologic research on the health effects of PBDE exposure and may suggest targets for future study. For example, we identified a 7 CpG DMR in the gene body of *S100A13*, or S100-family calcium binding protein A13, that was hypomethylated in response to BDE-153 exposure. *S100A13* is differentially expressed in thyroid tissue of both adults and developing fetuses,[196,197] and is thought to be involved in intercellular signaling.[198,199] Endocrine disrupting effects of PBDE exposure on thyroid hormone have been well established in animal studies[200-203] and been reported in epidemiologic literature.[204,205] Among pregnant mothers in our cohort, we've observed significant associations with serum PBDE levels and decreased thyroid-stimulating hormone[162] and suggestive decreases among their newborns.[206] Given these epidemiologic findings, further exploration of the role of *S100A13* and other DMR related genes in mediation of PBDE thyroid disruption is warranted.

Table 5. CpG overlap in DMRs by BDE congeners and their sum.

	BDE-47 (44 CpGs)	BDE-99 (48 CpGs)	BDE-100 (75 CpGs)	BDE-153 (303 CpGs)	Σ4BDEs (51 CpGs)
BDE-47 (44 CpGs)	44	32	44	38	44
BDE-99 (48 CpGs)		48	34	26	36
BDE-100 (75 CpGs)			75	64	49
BDE-153 (303 CpGs)				303	41
Σ4BDEs (51 CpGs)					51

Interestingly, while methylation at many genomic regions showed association across multiple PBDE congeners, others were only associated with specific congeners. In particular, BDE-153 was associated with the largest number of differentially methylated regions (N=48). This may be due to differences in BDE-153 chemistry: among the four congeners studied here it is the largest BDE by molecular weight, carries additional bromine groups, and has the longest half-life of all BDEs considered.[157,207,208] BDE-153 is also the only studied congener to be included in both the penta- and octaBDE technical mixtures.[209] For these reasons, BDE-153 may be more indicative of PBDE exposure over a longer time frame and/or from multiple sources, resulting in additional associations for this congener in particular. Alternately, the unique chemical properties of BDE-153 may produce biological effects distinct from other BDE congeners, as we have observed in the case of BMI and measures of obesity.[161]

Our study has several strengths and some limitations. CHAMACOS is a well-characterized prospective birth cohort study that is relatively homogeneous with regard to socioeconomic status, ancestry, and geographic distribution helped limit the potential for confounding bias. The detailed covariate data already available for this cohort further allowed

systematic assessment of potential confounders. The broad range of PBDE exposures experienced in this California cohort makes it particularly relevant for extrapolating effects to other populations. Assessment of DNA methylation by 450K BeadChip is state of the art and provides detailed characterization of methylation levels. One limitation of our DMP analysis is sample size. While our DMR analysis was powered to identify regional associations by capitalizing on knowledge of genomic structure, naïve DMP analysis didn't have power sufficient to identify results after accounting for multiple testing. Pooled meta-analysis for increased power wasn't possible at this time because to our knowledge no other birth cohorts currently have both 450K methylation data and PBDE data available. In the future, we hope to replicate our findings if/when such data become available. A general limitation of our study is that DNA methylation was assessed in blood as a non-invasive surrogate tissue for children's studies. There is a growing consensus that blood DNA is an appropriate and informative matrix for research on epigenetic programming of environmental exposures.[40,210,211] Given that we did not observe association between PBDE exposure and white blood cell composition and our results did not change appreciably after adjustment for white blood cell proportions, this is unlikely to present a source of bias. Another limitation is that our assessment only evaluates the relationship of prenatal PBDE exposure with DNA methylation in newborns. More studies will be needed to determine whether DMRs persist over time and mediate health risks in older children and adults. Further, the impact of methylation results (DMRs) on expression of biologically relevant genes (e.g. *NRBP1, CDH9, NTN1, S100A13*) remains to be assessed in future work.

Conclusions

In summary, we analyzed site specific and regional DNA methylation differences in human cord blood associated with prenatal maternal PBDE serum concentrations. We identified differential methylation at multiple genomic regions, some of which were consistently identified across several PBDE congeners. The differential methylation we observed in key genes and pathways reveals novel epigenetic impacts of PBDE exposure, and suggest potential biological mediators of the effects of PBDE exposure on health outcomes.

Supplemental material

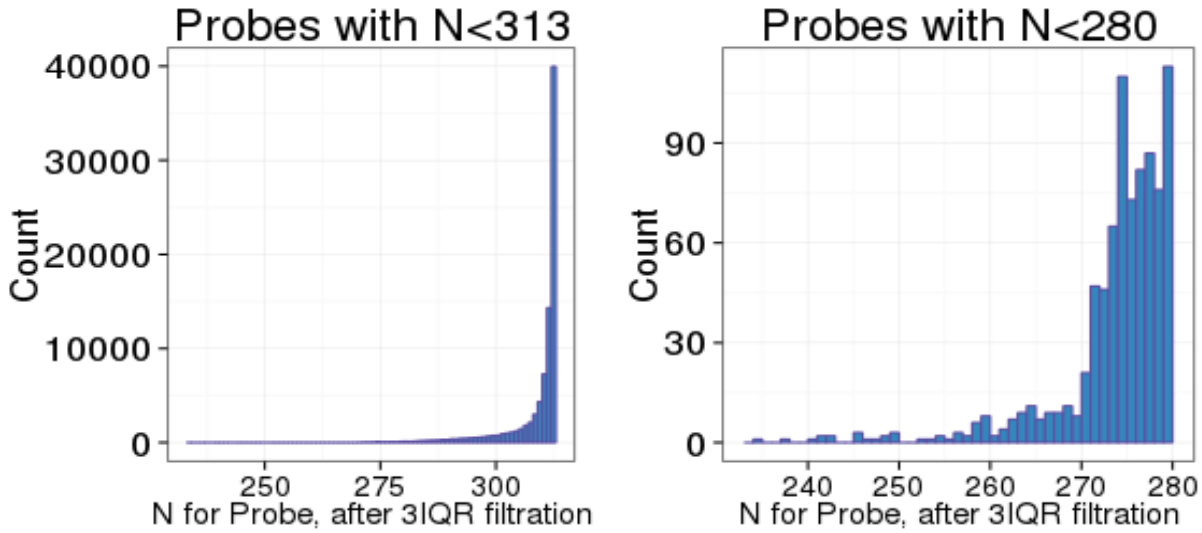


Figure S1. Histogram of sample N's of analyzed CpG sites following exclusion of outlying methylation values greater Figure S1. Histogram of sample N's of analyzed CpG sites following exclusion of outlying methylation values greater or less than 3times the CpG interquartile range (3IQR). Count is shown for all 87,784 CpGs (left panel) and restricted to the 838 CpGs with sample N's less than 280 (right panel). or less than 3times the CpG interquartile range (3IQR). Count is shown for all 87,784 CpGs (left panel) and restricted to the 838 CpGs with sample N's less than 280 (right panel).

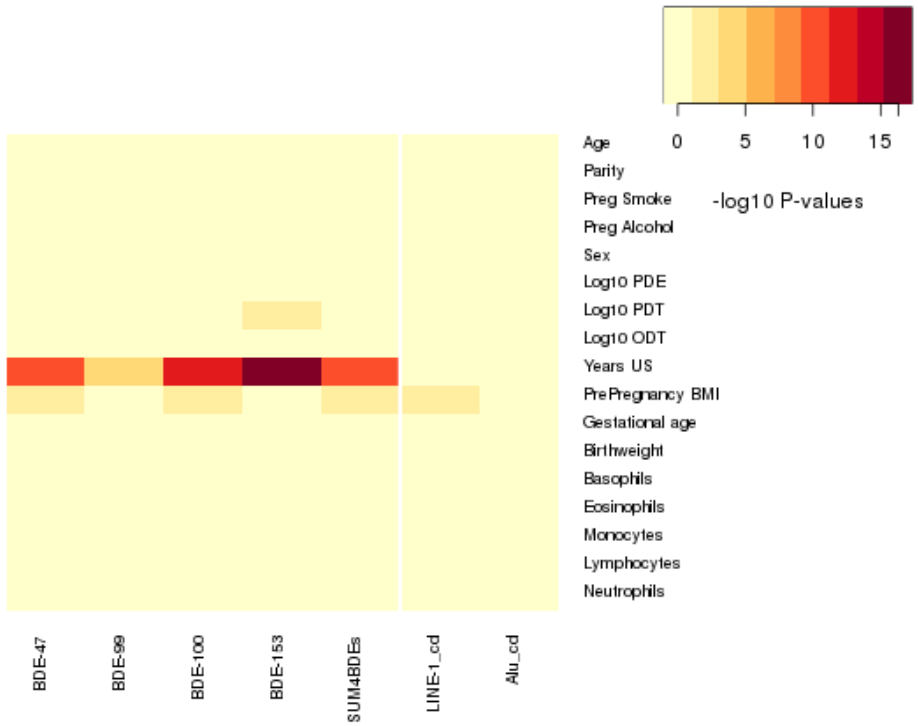


Figure S2. Heatmap of P-values from analysis of potential covariates for PBDE-EWAS modeling.

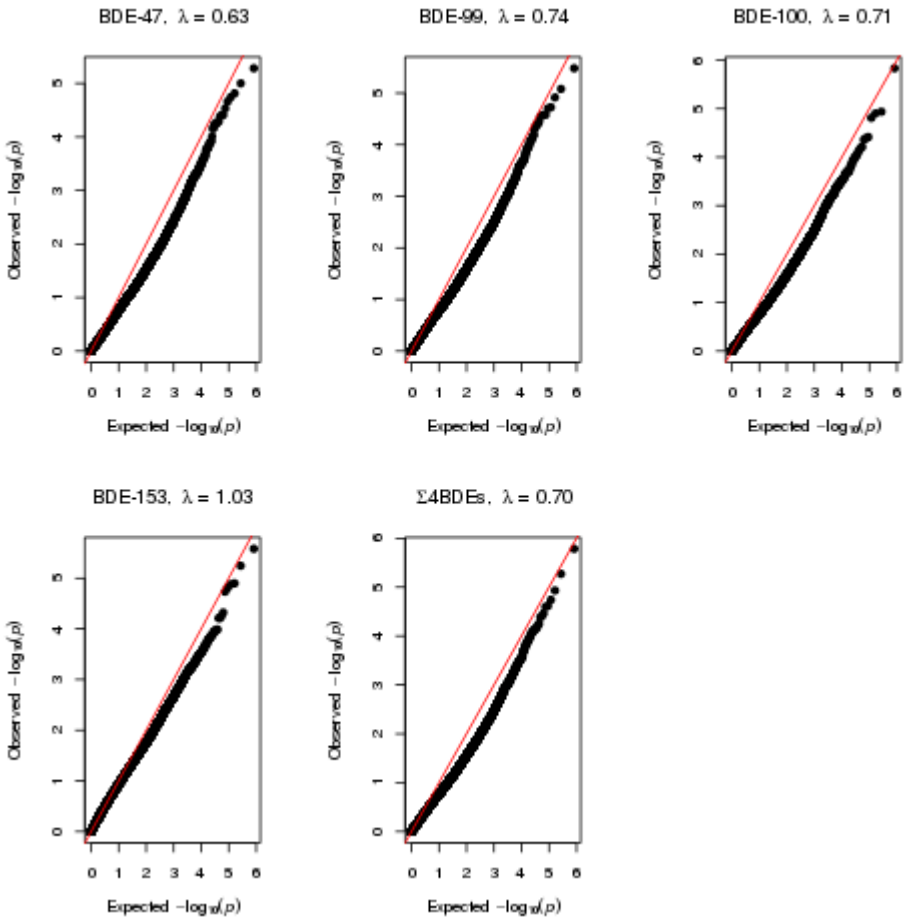


Figure S3. Quantile-quantile plots of observed and expected –log10 P-values and corresponding genomic inflation factors (λ's) from DMP models from 4 PBDE congeners and their sum.

CHAPTER 8

CONCLUSION AND SYNTHESIS OF FINDINGS

Epigenetic mechanisms, particularly DNA methylation, are a possible link between genetic and environmental determinants of health. As the DNA methylome undergoes rearrangement *in utero* and is susceptible to environmental disruption, it may be a mechanism mediating the developmental origins of human disease. Epidemiologic studies seeking to determine the impact of environmental exposures on genome-scale site-specific DNA methylation during the particularly vulnerable period of prenatal development face several methodological obstacles, many of which have yet to be rigorously addressed in the scientific literature.

In particular, this dissertation's specific aims were designed to address the following data and knowledge gaps:

- While several data-processing procedures have been proposed for 450K BeadChip results, little comparative assessment of performance measures, such as batch effect reductions and improved repeatability of technical replicates, has been conducted especially for the large sample sizes required in epidemiologic studies.

- The Illumina first sample normalization is the standard data-processing method in the field, however, the stability of its results when normalizing on samples of varying quality has not been evaluated.

- Statistical adjustment for white blood cell counts estimated by the Houseman algorithm (minfi) has become an established method to account for cell composition bias, but these estimates have been developed for a population of middle-aged white males and have not been validated for child samples.

- Previous studies have indicated some differential DNA methylation on autosomal chromosomes by sex in adults, but the extent of these differences are not well understood due to pervasive methodological shortcomings.

- Epigenome-wide evaluation of sex differences in DNA methylation have only considered CpG sites individually, instead of as regions which are likely more functionally relevant.

- Prenatal exposure to polybrominated diphenyl ether (PBDE) flame retardants has been associated with impaired child neurodevelopment and increased child BMI, but the biological mechanisms mediating these effects are not well understood.

- In the only study in humans to date, we previously found that maternal PBDE serum concentrations were associated with differential DNA hypomethylation in the LINE-1 and ALU repetitive elements in cord blood, but no site-specific assessment has been performed.

The birth cohort design and rich collection of data and biological specimens of the CHAMACOS study provided us with a unique opportunity to address these gaps in the use of 450K BeadChip data examine the effects of sex and prenatal PBDE exposure on DNA methylation.

Our key findings are:

- The ASMN procedure we proposed, and that has already been submitted to the public software database Bioconductor, performs consistently well, both at reducing batch effects and improving replicate comparability. It can be successfully implemented to enhance 450K DNA methylation data preprocessing. **Chapter 4.**

- Minfi white blood cell estimates did not correlate well with differential cell count (DCC) proportions in samples from newborns (ρ = -0.05 for granulocytes; ρ = -0.03 for lymphocytes): a deviation most likely explained by differences between DNA methylation profiles of our CHAMACOS newborn sample and the reference dataset used in the minfi prediction model, which is derived from six middle-aged Swedish men. **Chapter 5.**

- However, in older children correlation between minfi and DCC improved substantially (ρ = 0.77 for granulocytes; ρ = 0.75 for lymphocytes), probably due to increasing similarity with minfi's adult reference data as CHAMACOS children aged. This suggests that the minfi method may provide suitable estimates of white blood cell composition for samples from adults and older children, but may not currently be appropriate for analysis involving newborns or young children. **Chapter 5.**

- We observed a large number of CpG sites exhibiting differential DNA methylation (~3% of CpG sites interrogated by the 450K BeadChip) between girls and boys at birth (FDR $P<0.05$) after adjusting for cell composition. While some of these differentially methylated positions (DMPs) that were mostly hypermethylated in girls were located on the X-chromosome and are likely attributable to X-inactivation, we identified 3,031 CpGs that were located on autosomes. Further, of those 82.8% were hypermethylated in girls compared to boys. **Chapter 6.**

- Using pathway analysis, we found that sex-associated autosomal DMPs were significantly enriched for gene ontology terms related to nervous system development and behavior. This suggests that DNA methylation may contribute to differences in cognitive processes early in life and is consistent with sex differences in brain development and rates of maturation that have previously been observed by magnetic resonance imaging in slightly older children. **Chapter 6.**

- In addition to individual CpGs, we found 3,604 sex-associated differentially methylated regions (DMRs). The trend of hypermethylation in girls observed for individual site results was also seen in DMRs where the majority of DMRs (75.8%) also had higher methylation in girls. Besides confirming many of the findings in the DMP analysis, the application of DMR-finding substantially expanded the number of CpG sites considered significant. **Chapter 6.**

- Among the sex-associated DMPs we identified, we replicated 35.9% of all sex-associated CpG sites reported in previously published human studies. For these replicated hits, the direction of the association with methylation in our results was highly concordant (98.5-100%) with previous studies. While there was substantial overlap between our autosomal sex-associated hits and these previously published results, 2,603 or 85.9% of our results are novel findings. **Chapter 6.**

- We performed the first study to examine DNA methylation broadly across the genome (>410,000 CpGs) in relation to PBDE exposure. When we considered groups of CpG sites jointly as genomic regions we identified between 6 and 48 DMRs in umbilical cord blood associated with different measures of prenatal PBDE exposure. Many CpGs and DMRs were consistently identified across multiple PBDE exposure measures. **Chapter 7.**

- In our analysis of prenatal-PBDE related DMRS, certain BDE congeners, BDE's-47, -99 and Σ4BDE, had fewer (from 6 to 9), mostly hypomethylated DMRs. While prenatal BDE-100 and -153 levels were associated with more DMRs (11 and 48 respectively) and the majority hypermethylated. **Chapter 7.**

- The PBDE-DMRs we identified were located in genes (e.g. *NRBP1, CDH9, NTN1*, *S100A13*) involved in biologically relevant functions (including axon guidance and tumor suppression) given the health effects observed in association with BDE exposure to date, including neurotoxicity, neurodevelopmental deficits and cancer. **Chapter 7.**

- BDE-153 differs from the other congeners considered both in chemistry (it has the longest half-life of all BDEs considered) and in route of exposure (it was the only studied congener included in two different commercial products: penta- and octaBDE). Interestingly, we observed the greatest number of DMRs (N=48) associated with BDE-153, suggesting its biological effects may be distinct from other BDE congener. **Chapter 7.**

Overall, we have developed methods and performed analyses that have addressed key knowledge gaps in the nascent use of DNA methylation measurements in an epidemiologic context. We found evidence that existing approaches to normalization of 450K BeadChip data were not optimal and could vary greatly in practice, especially with the samples sized used population-based studies. In repeated measures of DNA methylation at birth and in older children, along with differential cell counts, we found that a standard approach in epigenome-wide analysis – minfi white blood cell composition estimation – may not be appropriate for use in analysis of cord blood. Additionally, we performed the most extensive characterization of the contribution of a key biological host factor, child sex, to DNA methylation to date.

Further, we utilized these developments to investigate the effects on genome wide site-specific DNA methylation associated with exposure to key chemicals of concern, PBDEs. This revealed key genes and biological pathways that may be altered by PBDE exposure in utero and which may be biological mediators of adverse health outcomes.

CHAPTER 9

FUTURE DIRECTIONS

Key next steps that should be explored to expand upon the work presented in this dissertation include:

- While we evaluated the impact of age on minfi cell composition estimates, more work is needed to explore further other factors, such as environmental exposures, that may also impact the validity of the minfi estimates.

- Given the lack of variation of the current Swedish male reference dataset used by minfi over several potentially biasing factors, it would be advisable to expand or generate a new reference that would have observations from early childhood, and that vary by race and ethnicity.

- The time frame of the analyses conducted in Chapters 6 and 7 should be expanded to incorporate older children:

    - Assessment of the temporal stability of sex-associated DNA methylation markers should also be performed by incorporating 450K BeadChip data from the same CHAMACOS children at older ages.

    - PBDE exposure in this population has also continued through childhood, so it will be important to incorporate PBDE exposures at later time periods to assess the effect of longitudinal exposure on DNA methylation.

- Additionally, the results of analysis of PBDE exposure on DNA methylation (**Aim 3**) should ideally be validated in a different study population. Cross-validating results and combining data in consortia is common practice for the acceptance of genome wide association studies (GWAS), but is only beginning to be performed in the field of epigenetics. While to our knowledge no other birth cohorts currently have both 450K methylation data and PBDE data available, eventually this added replication would help establish the generalizability of our results.

- The functional impact on gene expression of our findings on PBDE-DMRs, especially in the identified biologically relevant genes (e.g. *NRBP1, CDH9, NTN1, S100A13*), remains to be assessed in future work.

- Beyond examining the effect of exposure on DNA methylation, further analysis should begin to evaluate association between DNA methylation and health outcomes of concern for CHAMACOS children. Such health outcomes might include measures of obesity, neurodevelopmental deficits, hormonal irregularities, and/or onset of puberty.

- Following initial assessment of the relationships a) between DNA methylation and exposure, and b) between DNA methylation and health outcomes, epidemiologic inquiry should ultimately test whether DNA methylation operates as a biological mediator between exposure and outcome. The theoretical framework and estimation techniques for mediation analysis have recently received much attention in the causal inference literature and have improved substantially.[212,213] These new approaches, such as the estimation of the Natural Indirect Effect (NIE) and the application of Targeted Minimum Loss-Based Estimation (TMLE),[214] should be applied to determine whether DNA mediates the effect of PBDE exposure on health risks.

DISSERTATION PUBLICATIONS

1. **Considerations for normalization of DNA methylation data by Illumina 450K BeadChip assay in population studies. (Chapter 4)**

   Yousefi P, Huen K, Schall RA, Decker A, Elboudwarej E, Quach H, Barcellos L, Holland N.

   Epigenetics 8(11):1141-52. 2013.

2. **Estimation of blood cellular heterogeneity in newborns and children for epigenome-wide association studies. (Chapter 5)**

   Yousefi P, Huen K, Quach H, Motwani G, Hubbard A, Eskenazi B, Holland N.

   Environ Mol Mutagen 56(9):751-8. 2015.

3. **Sex differences in DNA methylation assessed by 450 K BeadChip in newborns. (Chapter 6)**

   Yousefi P, Huen K, Dave V, Barcellos L, Eskenazi B, Holland N.

   BMC Genomics 16(1):911. 2015.

4. **Prenatal PBDE exposure and altered patterns of DNA methylation regions in cord blood. (Chapter 7)**

   Yousefi P, Huen K, Bradman A, Eskenazi B, Holland N.

   Submitted to Environmental Health Perspectives, March 25, 2016

REFERENCES

1. Bird, A., *DNA methylation patterns and epigenetic memory.* Genes Dev, 2002. **16**(1): p. 6-21.
2. Laird, P.W., *Principles and challenges of genomewide DNA methylation analysis.* Nat Rev Genet, 2010. **11**(3): p. 191-203.
3. Foley, D.L., et al., *Prospects for epigenetic epidemiology.* Am J Epidemiol, 2009. **169**(4): p. 389-400.
4. Perera, F. and J. Herbstman, *Prenatal environmental exposures, epigenetics, and disease.* Reprod Toxicol, 2011. **31**(3): p. 363-73.
5. Huen, K., et al., *Effects of age, sex, and persistent organic pollutants on DNA methylation in children.* Environ Mol Mutagen, 2014. **55**(3): p. 209-22.
6. U.S. DHHS, *Toxicological profile for polybrominated biphenyls and polybrominated diphenyl ethers (PBBs and PBDEs),* A.f.T.S.a.D. Registry, Editor. 2004, US Department of Health and Human Services.
7. Dodson, R.E., et al., *After the PBDE phase-out: a broad suite of flame retardants in repeat house dust samples from California.* Environ Sci Technol, 2012. **46**(24): p. 13056-66.
8. Windham, G.C., et al., *Body burdens of brominated flame retardants and other persistent organo-halogenated compounds and their descriptors in US girls.* Environ Res, 2010. **110**(3): p. 251-7.
9. Bradman, A., et al., *Factors associated with serum polybrominated diphenyl ether (PBDE) levels among school-age children in the CHAMACOS cohort.* Environ Sci Technol, 2012. **46**(13): p. 7373-81.
10. Rose, M., et al., *PBDEs in 2-5 year-old children from California and associations with diet and indoor environment.* Environ Sci Technol, 2010. **44**(7): p. 2648-53.
11. Eskenazi, B., et al., *A comparison of PBDE serum concentrations in Mexican and Mexican-American children living in California.* Environ Health Perspect, 2011. **119**(10): p. 1442-8.
12. Gascon, M., et al., *Polybrominated diphenyl ethers (PBDEs) in breast milk and neuropsychological development in infants.* Environ Health Perspect, 2012. **120**(12): p. 1760-5.
13. Chao, H.R., et al., *Levels of breast milk PBDEs from southern Taiwan and their potential impact on neurodevelopment.* Pediatr Res, 2011. **70**(6): p. 596-600.
14. Eskenazi, B., et al., *In utero and childhood polybrominated diphenyl ether (PBDE) exposures and neurodevelopment in the CHAMACOS study.* Environ Health Perspect, 2013. **121**(2): p. 257-62.
15. Herbstman, J.B., et al., *Prenatal exposure to PBDEs and neurodevelopment.* Environ Health Perspect, 2010. **118**(5): p. 712-9.
16. Gascon, M., et al., *Effects of pre and postnatal exposure to low levels of polybromodiphenyl ethers on neurodevelopment and thyroid hormone levels at 4 years of age.* Environ Int, 2011. **37**(3): p. 605-11.
17. Chevrier, J., et al., *Polybrominated diphenyl ether (PBDE) flame retardants and thyroid hormone during pregnancy.* Environ Health Perspect, 2010. **118**(10): p. 1444-9.
18. Harley, K.G., et al., *PBDE concentrations in women's serum and fecundability.* Environ Health Perspect, 2010. **118**(5): p. 699-704.

19. Erkin-Cakmak, A., et al., *In Utero and Childhood Polybrominated Diphenyl Ether Exposures and Body Mass at Age 7 Years: The CHAMACOS Study.* Environ Health Perspect, 2015. **123**(6): p. 636-42.
20. Robertson, K.D., *DNA methylation and human disease.* Nat Rev Genet, 2005. **6**(8): p. 597-610.
21. Feinberg, A.P., *Phenotypic plasticity and the epigenetics of human disease.* Nature, 2007. **447**(7143): p. 433-40.
22. Esteller, M., *Epigenetics in cancer.* N Engl J Med, 2008. **358**(11): p. 1148-59.
23. Urdinguio, R.G., J.V. Sanchez-Mut, and M. Esteller, *Epigenetic mechanisms in neurological diseases: genes, syndromes, and therapies.* Lancet Neurol, 2009. **8**(11): p. 1056-72.
24. Ling, C. and L. Groop, *Epigenetics: a molecular link between environmental factors and type 2 diabetes.* Diabetes, 2009. **58**(12): p. 2718-25.
25. Rakyan, V.K., et al., *Identification of type 1 diabetes-associated DNA methylation variable positions that precede disease diagnosis.* PLoS Genet, 2011. **7**(9): p. e1002300.
26. Taby, R. and J.P. Issa, *Cancer epigenetics.* CA Cancer J Clin, 2010. **60**(6): p. 376-92.
27. Esteller, M., et al., *Detection of aberrant promoter hypermethylation of tumor suppressor genes in serum DNA from non-small cell lung cancer patients.* Cancer Res, 1999. **59**(1): p. 67-70.
28. Guo, H., et al., *The DNA methylation landscape of human early embryos.* Nature, 2014. **511**(7511): p. 606-10.
29. Smith, Z.D., et al., *DNA methylation dynamics of the human preimplantation embryo.* Nature, 2014. **511**(7511): p. 611-+.
30. Jung, M. and G.P. Pfeifer, *Aging and DNA methylation.* BMC Biol, 2015. **13**: p. 7.
31. Almen, M.S., et al., *Genome-wide analysis reveals DNA methylation markers that vary with both age and obesity.* Gene, 2014. **548**(1): p. 61-7.
32. Marioni, R.E., et al., *DNA methylation age of blood predicts all-cause mortality in later life.* Genome Biol, 2015. **16**: p. 25.
33. Tamashiro, K.L. and T.H. Moran, *Perinatal environment and its influences on metabolic programming of offspring.* Physiol Behav, 2010. **100**(5): p. 560-6.
34. Heindel, J.J., *The Obesogen Hypothesis of Obesity: Overview and Human Evidence*, in *Obesity Before Birth*, R.H. Lustig, Editor. 2011, Springer.
35. Faulk, C. and D.C. Dolinoy, *Timing is everything: the when and how of environmentally induced changes in the epigenome of animals.* Epigenetics, 2011. **6**(7): p. 791-7.
36. Gardiner-Garden, M. and M. Frommer, *CpG islands in vertebrate genomes.* J Mol Biol, 1987. **196**(2): p. 261-82.
37. Irizarry, R.A., et al., *The human colon cancer methylome shows similar hypo- and hypermethylation at conserved tissue-specific CpG island shores.* Nat Genet, 2009. **41**(2): p. 178-86.
38. Ball, M.P., et al., *Targeted and genome-scale strategies reveal gene-body methylation signatures in human cells.* Nat Biotechnol, 2009. **27**(4): p. 361-8.
39. Armstrong, D.A., et al., *Global and gene-specific DNA methylation across multiple tissues in early infancy: implications for children's health research.* FASEB J, 2014. **28**(5): p. 2088-97.
40. Chadwick, L.H., et al., *New insights and updated guidelines for epigenome-wide association studies.* Neuroepigenetics, 2014.

41. Bibikova, M., et al., *High density DNA methylation array with single CpG site resolution.* Genomics, 2011. **98**(4): p. 288-95.
42. Sandoval, J., et al., *Validation of a DNA methylation microarray for 450,000 CpG sites in the human genome.* Epigenetics, 2011. **6**(6): p. 692-702.
43. Roessler, J., et al., *Quantitative cross-validation and content analysis of the 450k DNA methylation array from Illumina, Inc.* BMC Res Notes, 2012. **5**(1): p. 210.
44. Du, P., W.A. Kibbe, and S.M. Lin, *lumi: a pipeline for processing Illumina microarray.* Bioinformatics, 2008. **24**(13): p. 1547-8.
45. Touleimat, N. and J. Tost, *Complete pipeline for Infinium((R)) Human Methylation 450K BeadChip data processing using subset quantile normalization for accurate DNA methylation estimation.* Epigenomics, 2012. **4**(3): p. 325-41.
46. Schalkwyk, L., et al., *wateRmelon: Illumina 450 methylation array normalization and metrics. R package version 0.99.16*. 2013.
47. Michels, K.B., et al., *Recommendations for the design and analysis of epigenome-wide association studies.* Nat Methods, 2013. **10**(10): p. 949-955.
48. Liang, L. and W.O.C. Cookson, *Grasping nettles: cellular heterogeneity and other confounders in epigenome-wide association studies.* Human Molecular Genetics, 2014. **23**(R1): p. ddu284-R88.
49. Lowe, R. and V.K. Rakyan, *Correcting for cell-type composition bias in epigenome-wide association studies.* Genome Medicine, 2014. **6**(3): p. 1-2.
50. Jaffe, A.E. and R.A. Irizarry, *Accounting for cellular heterogeneity is critical in epigenome-wide association studies.* Genome Biol, 2014. **15**(2): p. R31.
51. Houseman, E.A., et al., *DNA methylation arrays as surrogate measures of cell mixture distribution.* BMC Bioinformatics, 2012. **13**: p. 86.
52. Bibikova, M., et al., *Genome-wide DNA methylation profiling using Infinium® assay.* Epigenomics, 2009. **1**(1): p. 177-200.
53. Reinius, L.E., et al., *Differential DNA methylation in purified human blood cells: implications for cell lineage and studies on disease susceptibility.* PLoS One, 2012. **7**(7): p. e41361.
54. Cheng, C.K.-W., et al., *Complete blood count reference interval diagrams derived from NHANES III: stratification by age, sex, and race.* Laboratory hematology : official publication of the International Society for Laboratory Hematology, 2004. **10**(1): p. 42-53.
55. Dallman, P.R., *Blood and blood-forming tissues,* in *Paediatrics* A. Rudolph, Editor. 1977, Appleton-Century-Crofts: New York. p. 1109-1112.
56. Nathan, D.G. and F.A. Oski, *Hematology of Infancy and Childhood, 2nd Edition*. 2 ed. 1981. 1552-1574.
57. Liu, J., et al., *A study of the influence of sex on genome wide methylation.* PLoS One, 2010. **5**(4): p. e10028.
58. Boks, M.P., et al., *The relationship of DNA methylation with age, gender and genotype in twins and healthy controls.* PLoS One, 2009. **4**(8): p. e6767.
59. Tapp, H.S., et al., *Nutritional factors and gender influence age-related DNA methylation in the human rectal mucosa.* Aging Cell, 2013. **12**(1): p. 148-55.
60. Chen, Y.A., et al., *Discovery of cross-reactive probes and polymorphic CpGs in the Illumina Infinium HumanMethylation450 microarray.* Epigenetics, 2013. **8**(2): p. 203-9.

61. McCarthy, N.S., et al., *Meta-analysis of human methylation data for evidence of sex-specific autosomal patterns.* Bmc Genomics, 2014. **15**.
62. Rusiecki, J.A., et al., *Global DNA hypomethylation is associated with high serum-persistent organic pollutants in Greenlandic Inuit.*, in *Environ Health Persp*. 2008. p. 1547-1552.
63. Kim, K.Y., et al., *Association of low-dose exposure to persistent organic pollutants with global DNA hypomethylation in healthy Koreans.* Environ Health Perspect, 2010. **118**(3): p. 370-4.
64. Itoh, H., et al., *Association between serum organochlorines and global methylation level of leukocyte DNA among Japanese women: a cross-sectional study.* Sci Total Environ, 2014. **490**: p. 603-9.
65. Lind, L., et al., *Global DNA hypermethylation is associated with high serum levels of persistent organic pollutants in an elderly population.* Environ Int, 2013. **59**: p. 456-61.
66. Shutoh, Y., et al., *Low dose effects of dichlorodiphenyltrichloroethane (DDT) on gene transcription and DNA methylation in the hypothalamus of young male rats: implication of hormesis-like effects.* J Toxicol Sci, 2009. **34**(5): p. 469-82.
67. Skinner, M.K., et al., *Ancestral dichlorodiphenyltrichloroethane (DDT) exposure promotes epigenetic transgenerational inheritance of obesity.* BMC Med, 2013. **11**: p. 228.
68. Byun, H.M., et al., *Epigenetic effects of low perinatal doses of flame retardant BDE-47 on mitochondrial and nuclear genes in rat offspring.* Toxicology, 2015. **328**: p. 152-9.
69. Eskenazi, B., et al., *CHAMACOS, a longitudinal birth cohort study: lessons from the fields.* J Childrens Health, 2003. **1**: p. 3–27.
70. Eskenazi, B., et al., *Association of in utero organophosphate pesticide exposure and fetal growth and length of gestation in an agricultural population.* Environmental Health Perspectives, 2004. **112**(10): p. 1116-1124.
71. Holland, N., et al., *Paraoxonase polymorphisms, haplotypes, and enzyme activity in Latino mothers and newborns.* Environ Health Perspect, 2006. **114**(7): p. 985-91.
72. Teschendorff, A.E., et al., *A beta-mixture quantile normalization method for correcting probe design bias in Illumina Infinium 450 k DNA methylation data.* Bioinformatics, 2013. **29**(2): p. 189-196.
73. Chen, Y.-a., et al., *Discovery of cross-reactive probes and polymorphic CpGs in the Illumina Infinium HumanMethylation450 microarray.*, in *Epigenetics*. 2013. p. 203-209.
74. Du, P., et al., *Comparison of Beta-value and M-value methods for quantifying methylation levels by microarray analysis*, in *Bmc Bioinformatics*. 2010. p. 587.
75. Turgeon, M.L., *Clinical Hematology Theory and Procedures, FIFTH EDITION*. 2011. 40-45.
76. Aryee, M.J., et al., *Minfi: a flexible and comprehensive Bioconductor package for the analysis of Infinium DNA methylation microarrays.* Bioinformatics, 2014. **30**(10): p. 1363-1369.
77. Jaffe, A.E., *FlowSorted.Blood.450k: Illumina HumanMethylation data on sorted blood cell populations. R package version 1.2.0.*
78. Sjodin, A., et al., *Semiautomated high-throughput extraction and cleanup method for the measurement of polybrominated diphenyl ethers, polybrominated biphenyls, and polychlorinated biphenyls in human serum.* Anal Chem, 2004. **76**(7): p. 1921-7.

79. Castorina, R., et al., *Determinants of serum polybrominated diphenyl ether (PBDE) levels among pregnant women in the CHAMACOS cohort.* Environ Sci Technol, 2011. **45**(15): p. 6553-60.
80. Phillips, D.L., et al., *Chlorinated hydrocarbon levels in human serum: effects of fasting and feeding.* Arch Environ Contam Toxicol, 1989. **18**(4): p. 495-500.
81. Lubin, J.H., et al., *Epidemiologic evaluation of measurement data in the presence of detection limits.* Environ Health Perspect, 2004. **112**(17): p. 1691-6.
82. Foley, D.L., et al., *Prospects for epigenetic epidemiology.* Am J Epidemiol, 2009. **169**(4): p. 389-400.
83. Pennisi, E., *Behind the scenes of gene expression.* Science, 2001. **293**(5532): p. 1064-7.
84. Ho, S.M. and W.Y. Tang, *Techniques used in studies of epigenome dysregulation due to aberrant DNA methylation: an emphasis on fetal-based adult diseases.* Reprod Toxicol, 2007. **23**(3): p. 267-82.
85. Tammen, S.A., S. Friso, and S.W. Choi, *Epigenetics: The link between nature and nurture.* Mol Aspects Med, 2012.
86. Dedeurwaerder, S., et al., *Evaluation of the Infinium Methylation 450K technology.* Epigenomics, 2011. **3**(6): p. 771-84.
87. Maksimovic, J., L. Gordon, and A. Oshlack, *SWAN: Subset-quantile within array normalization for illumina infinium HumanMethylation450 BeadChips.* Genome Biol, 2012. **13**(6): p. R44.
88. Teschendorff, A.E., et al., *A Beta-Mixture Quantile Normalisation method for correcting probe design bias in Illumina Infinium 450k DNA methylation data.* Bioinformatics, 2012.
89. Marabita, F., et al., *An evaluation of analysis pipelines for DNA methylation profiling using the Illumina HumanMethylation450 BeadChip platform.* Epigenetics, 2013. **8**(3): p. 333-46.
90. Illumina (November, 2010, Accessed 5/1/2013) *GenomeStudio Methylation Module v1.8 User Guide*
91. *International HapMap Project.* *www.hapmap.ncbi.nlm.nih.gov*.
92. Vineis, P. and S. Garte, *Biomarker Validation*, in *Molecular Epidemiology of Chronic Diseases*, C. Wild, P. Vineis, and S. Garte, Editors. 2008, John Wiley & Sons, Ltd: Chichester, UK. p. pg72.
93. Weber, M., et al., *Chromosome-wide and promoter-specific analyses identify sites of differential DNA methylation in normal and transformed human cells.* Nat Genet, 2005. **37**(8): p. 853-862.
94. Meissner, A., et al., *Genome-scale DNA methylation maps of pluripotent and differentiated cells.* Nature, 2008. **454**(7205): p. 766-770.
95. Lister, R., et al., *Human DNA methylomes at base resolution show widespread epigenomic differences.* Nature, 2009. **462**(7271): p. 315-22.
96. Liu, Y., et al., *Epigenome-wide association data implicate DNA methylation as an intermediary of genetic risk in rheumatoid arthritis.* Nat Biotechnol, 2013. **31**(2): p. 142-147.
97. Teschendorff, A.E., et al., *An Epigenetic Signature in Peripheral Blood Predicts Active Ovarian Cancer.* PLoS One, 2009. **4**(12).
98. Hughes, V., *Epigenetics: The sins of the father.* Nature, 2014. **507**(7490): p. 22-4.

99. Joubert, B.R., et al., *450K Epigenome-Wide Scan Identifies Differential DNA Methylation in Newborns Related to Maternal Smoking during Pregnancy.* Environmental Health Perspectives, 2012. **120**(10): p. 1425-1431.
100. Koestler, D.C., et al., *Differential DNA methylation in umbilical cord blood of infants exposed to low levels of arsenic in utero.* Environ Health Perspect, 2013. **121**(8): p. 971-977.
101. Barker, D.J., *In utero programming of chronic disease.* Clin Sci (Lond), 1998. **95**(2): p. 115-28.
102. Essex, M.J., et al., *Epigenetic vestiges of early developmental adversity: childhood stress exposure and DNA methylation in adolescence.* Child Dev, 2013. **84**(1): p. 58-75.
103. Babenko, O., I. Kovalchuk, and G.A. Metz, *Stress-induced Perinatal and Transgenerational Epigenetic Programming of Brain Development and Mental Health.* Neurosci Biobehav Rev, 2014. **48C**: p. 70-91.
104. Eskenazi, B., et al., *CHAMACOS, a longitudinal birth cohort study: lessons from the fields.* J Childrens Healt, 2003. **1**: p. 3-27.
105. Eskenazi, B., et al., *Association of in utero organophosphate pesticide exposure and fetal growth and length of gestation in an agricultural population.* Environ Health Perspect, 2004. **112**(10): p. 1116-24.
106. Team, R.C., *R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL: http://www.R-project.org/.* 2013: Vienna, Austria.
107. Smith, Z.D., et al., *DNA methylation dynamics of the human preimplantation embryo.* Nature, 2014. **511**(7511): p. 611-5.
108. Hsieh, M.M., et al., *Prevalence of neutropenia in the U.S. population: age, sex, smoking status, and ethnic differences.* Ann Intern Med, 2007. **146**(7): p. 486-92.
109. Lim, E.M., et al., *Race-specific WBC and neutrophil count reference intervals.* Int J Lab Hematol, 2010. **32**(6 Pt 2): p. 590-7.
110. McCarthy, N.S., et al., *Meta-analysis of human methylation data for evidence of sex-specific autosomal patterns.*, in *BMC Genomics*. 2014. p. 981.
111. Adkins, R.M., et al., *Racial differences in gene-specific DNA methylation levels are present at birth.* Birth Defects Res A Clin Mol Teratol, 2011. **91**(8): p. 728-36.
112. Harris, R.A., D. Nagy-Szakal, and R. Kellermayer, *Human metastable epiallele candidates link to common disorders.* Epigenetics, 2013. **8**(2): p. 157-163.
113. Tapp, H.S., et al., *Nutritional factors and gender influence age-related DNA methylation in the human rectal mucosa.*, in *Aging Cell*. 2013. p. 148-155.
114. Fuke, C., et al., *Age related changes in 5-methylcytosine content in human peripheral leukocytes and placentas: an HPLC-based study.*, in *Ann. Hum. Genet.* 2004. p. 196-204.
115. Liu, J., et al., *A study of the influence of sex on genome wide methylation.*, in *PLoS ONE*. 2010. p. e10028.
116. Boks, M.P., et al., *The relationship of DNA methylation with age, gender and genotype in twins and healthy controls.*, in *PLoS ONE*. 2009. p. e6767.
117. Adkins, R.M., et al., *Racial differences in gene-specific DNA methylation levels are present at birth.*, in *Birth Defects Res. Part A Clin. Mol. Teratol.* 2011. p. 728-736.
118. Adkins, R.M., et al., *Parental ages and levels of DNA methylation in the newborn are correlated.* 2011. p. 47.

119. Cheng, C.K.-W., et al., *Complete blood count reference interval diagrams derived from NHANES III: stratification by age, sex, and race.* 2004. p. 42-53.
120. Hsieh, M.M., et al., *Prevalence of neutropenia in the U.S. population: age, sex, smoking status, and ethnic differences.*, in *Ann. Intern. Med.* 2007. p. 486-492.
121. Yousefi, P., et al., *Estimation of blood cellular heterogeneity in newborns and children for epigenome-wide association studies.*, in *Blood Cellular Heterogeneity in Newborns and Children*. 2015. p. 751-758.
122. Peters, T.J., et al., *De novo identification of differentially methylated regions in the human genome.*, in *Epigenetics & Chromatin*. 2015. p. 6.
123. Cotton, A.M., et al., *Landscape of DNA methylation on the X chromosome reflects CpG density, functional chromatin state and X-chromosome inactivation.*, in *Hum Mol Genet*. 2015. p. 1528-1539.
124. Joo, J.E., et al., *Human active X-specific DNA methylation events showing stability across time and tissues.*, in *Eur. J. Hum. Genet.* 2014. p. 1376-1381.
125. Sharp, A.J., et al., *DNA methylation profiles of human active and inactive X chromosomes.*, in *Genome Res*. 2011. p. 1592-1600.
126. Peters, T.J. and M.J. Buckley, eds. *DMRcate: Illumina 450K methylation array apatial analysis methods. R package version 1.2.0.*
127. Aravin, A.A., G.J. Hannon, and J. Brennecke, *The Piwi-piRNA pathway provides an adaptive defense in the transposon arms race.*, in *Science*. 2007. p. 761-764.
128. Siddiqi, S., M. Terry, and I. Matushansky, *Hiwi mediated tumorigenesis is associated with DNA hypermethylation.*, in *PLoS ONE*. 2012. p. e33711.
129. Avner, P. and E. Heard, *X-chromosome inactivation: counting, choice and initiation.*, in *Nat Rev Genet*. 2001. p. 59-67.
130. Xu, H., et al., *Sex-biased methylome and transcriptome in human prefrontal cortex.*, in *Hum Mol Genet*. 2014. p. 1260-1270.
131. Hall, E., et al., *Sex differences in the genome-wide DNA methylation pattern and impact on gene expression, microRNA levels and insulin secretion in human pancreatic islets.*, in *Genome Biol.* 2014. p. 522.
132. De Bellis, M.D., et al., *Sex differences in brain maturation during childhood and adolescence.*, in *Cereb. Cortex*. 2001. p. 552-557.
133. Kim, Y.-K., et al., *Association between norepinephrine transporter gene (SLC6A2) polymorphisms and suicide in patients with major depressive disorder*, in *Journal of Affective Disorders*. 2014, Elsevier. p. 127-132.
134. Thakur, G.A., et al., *Comprehensive Phenotype/Genotype Analyses of the Norepinephrine Transporter Gene (SLC6A2) in ADHD: Relation to Maternal Smoking during Pregnancy*, in *PLoS ONE*. 2012. p. e49616-23.
135. Buttenschøn, H.N., et al., *The norepinephrine transporter gene is a candidate gene for panic disorder*, in *J Neural Transm*. 2011. p. 969-976.
136. Leblond, C.S., et al., *Genetic and Functional Analyses of SHANK2 Mutations Suggest a Multiple Hit Model of Autism Spectrum Disorders*, in *PLoS Genet*. 2012. p. e1002521-17.
137. Peça, J., et al., *Shank3 mutant mice display autistic-like behaviours and striatal dysfunction*, in *Nature*. 2011, Nature Publishing Group. p. 437-442.
138. Wilson, C.A. and D.C. Davies, *The control of sexual differentiation of the reproductive system and brain*, in *Reproduction*. 2007. p. 331-359.

139. Babenko, O., I. Kovalchuk, and G.A.S. Metz, *Stress-induced Perinatal and Transgenerational Epigenetic Programming of Brain Development and Mental Health.*, in *Neurosci Biobehav Rev*. 2014. p. 70-91.
140. Yousefi, P., et al., *Considerations for normalization of DNA methylation data by Illumina 450K BeadChip assay in population studies.* Epigenetics, 2013. **8**(11): p. 1141-52.
141. (2013), R.C.T., ed. *R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL: http://www.R-project.org/.* R Foundation for Statistical Computing.
142. Simpkin, A.J., et al., *Longitudinal analysis of DNA methylation associated with birth weight and gestational age.* Hum Mol Genet, 2015. **24**(13): p. 3752-63.
143. Hochberg, Y. and Y. Benjamini, *More powerful procedures for multiple significance testing.*, in *Statist. Med.* 1990. p. 811-818.
144. Smyth, G.K., *Limma: linear models for microarray data.*, in *Bioinformatics and Computational Biology Solutions Using R and Bioconductor*, R. Gentleman, et al., Editors. 2005, Springer Science & Business Media: New York, NY. p. 397–420.
145. Satterthwaite, F.E., *An approximate distribution of estimates of variance components.*, in *Biometrics*. 1946. p. 110-114.
146. Huang, D.W., B.T. Sherman, and R.A. Lempicki, *Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources.*, in *Nat Protoc*. 2009. p. 44-57.
147. Huang, D.W., B.T. Sherman, and R.A. Lempicki, *Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists.*, in *Nucleic Acids Res*. 2009. p. 1-13.
148. Wang, J., et al., *WEB-based GEne SeT AnaLysis Toolkit (WebGestalt): update 2013.*, in *Nucleic Acids Res*. 2013. p. W77-83.
149. Kamburov, A., et al., *ConsensusPathDB: toward a more complete picture of cell biology.*, in *Nucleic Acids Res*. 2011. p. D712-7.
150. Supek, F., et al., *REVIGO summarizes and visualizes long lists of gene ontology terms.*, in *PLoS ONE*. 2011. p. e21800.
151. Besis, A. and C. Samara, *Polybrominated diphenyl ethers (PBDEs) in the indoor and outdoor environments – A review on occurrence and human exposure*, in *Environmental Pollution*. 2012. p. 217-229.
152. Hale, R.C., et al., *Polybrominated diphenyl ether flame retardants in the North American environment.*, in *Environ Int*. 2003. p. 771-779.
153. Bradman, A., et al., *Factors associated with serum polybrominated diphenyl ether (PBDE) levels among school-age children in the CHAMACOS cohort.*, in *Environ. Sci. Technol.* 2012. p. 7373-7381.
154. Windham, G.C., et al., *Body burdens of brominated flame retardants and other persistent organo-halogenated compounds and their descriptors in US girls*, in *Environmental Research*. 2010. p. 251-257.
155. Eskenazi, B., et al., *A comparison of PBDE serum concentrations in Mexican and Mexican-American children living in California.*, in *Environ Health Persp*. 2011. p. 1442-1448.
156. Rose, M., et al., *PBDEs in 2-5 year-old children from California and associations with diet and indoor environment.*, in *Environ. Sci. Technol.* 2010. p. 2648-2653.

157. Trudel, D., et al., *Total consumer exposure to polybrominated diphenyl ethers in North America and Europe.*, in *Environ. Sci. Technol.* 2011, American Chemical Society. p. 2391-2397.
158. Wu, X.M., et al., *Polybrominated diphenyl ether serum concentrations in a Californian population of children, their parents, and older adults: an exposure assessment study.* Environ Health, 2015. **14**: p. 23.
159. Chen, A., et al., *Prenatal Polybrominated Diphenyl Ether Exposures and Neurodevelopment in US Children through 5 Years of Age: The HOME Study*, in *Environ Health*. 2014.
160. Eskenazi, B., et al., *In utero and childhood polybrominated diphenyl ether (PBDE) exposures and neurodevelopment in the CHAMACOS study.*, in *Environ Health Persp*. 2013. p. 257-262.
161. Erkin-Cakmak, A., et al., *In Utero and Childhood Polybrominated Diphenyl Ether Exposures and Body Mass at Age 7 Years: The CHAMACOS Study.*, in *Environ Health Persp*. 2015. p. 636-642.
162. Chevrier, J., et al., *Polybrominated diphenyl ether (PBDE) flame retardants and thyroid hormone during pregnancy.*, in *Environ Health Persp*. 2010. p. 1444-1449.
163. Bakulski, K.M. and M.D. Fallin, *Epigenetic epidemiology: promises for public health research.* 2014. p. 171-183.
164. Tamashiro, K.L.K. and T.H. Moran, *Perinatal environment and its influences on metabolic programming of offspring.*, in *Physiol Behav*. 2010. p. 560-566.
165. Itoh, H., et al., *Association between serum organochlorines and global methylation level of leukocyte DNA among Japanese women: a cross-sectional study.*, in *Sci. Total Environ.* 2014. p. 603-609.
166. Kim, K.-Y., et al., *Association of low-dose exposure to persistent organic pollutants with global DNA hypomethylation in healthy Koreans.*, in *Environ Health Persp*. 2010. p. 370-374.
167. Lind, L., et al., *Global DNA hypermethylation is associated with high serum levels of persistent organic pollutants in an elderly population.*, in *Environ Int*. 2013. p. 456-461.
168. Shutoh, Y., et al., *Low dose effects of dichlorodiphenyltrichloroethane (DDT) on gene transcription and DNA methylation in the hypothalamus of young male rats: implication of hormesis-like effects.*, in *J Toxicol Sci*. 2009. p. 469-482.
169. Skinner, M.K., et al., *Ancestral dichlorodiphenyltrichloroethane (DDT) exposure promotes epigenetic transgenerational inheritance of obesity.*, in *BMC Med*. 2013. p. 228.
170. Byun, H.-M., et al., *Epigenetic effects of low perinatal doses of flame retardant BDE-47 on mitochondrial and nuclear genes in rat offspring.*, in *Toxicology*. 2015. p. 152-159.
171. Huen, K., et al., *Effects of age, sex, and persistent organic pollutants on DNA methylation in children.*, in *DNA Methylation in Children*. 2014. p. 209-222.
172. Lubin, J.H., et al., *Epidemiologic evaluation of measurement data in the presence of detection limits.*, in *Environ Health Persp*. 2004, National Institute of Environmental Health Science. p. 1691-1696.
173. Castorina, R., et al., *Determinants of serum polybrominated diphenyl ether (PBDE) levels among pregnant women in the CHAMACOS cohort.*, in *Environ. Sci. Technol.* 2011. p. 6553-6560.

174. Verner, M.-A., et al., *Increasing sample size in prospective birth cohorts: back-extrapolating prenatal levels of persistent organic pollutants in newly enrolled children.*, in *Environ. Sci. Technol.* 2015. p. 3940-3948.
175. Vandiver, A.R., et al., *Age and sun exposure-related widespread genomic blocks of hypomethylation in nonmalignant skin.*, in *Genome Biol.* 2015. p. 80.
176. Watson, C.T., et al., *Genome-wide DNA methylation profiling in the superior temporal gyrus reveals epigenetic signatures associated with Alzheimer's disease.*, in *Genome Medicine*. 2016, BioMed Central. p. 5.
177. Yang, I.V., et al., *DNA methylation and childhood asthma in the inner city.*, in *J. Allergy Clin. Immunol.* 2015. p. 69-80.
178. Liu, Y., et al., *GeMes, Clusters of DNA Methylation under Genetic Control, Can Inform Genetic and Epigenetic Analysis of Disease.*, in *Am J Hum Genet*. 2014. p. 485-495.
179. Shoemaker, R., et al., *Allele-specific methylation is prevalent and is contributed by CpG-SNPs in the human genome.*, in *Genome Res*. 2010. p. 883-889.
180. Dolzhenko, E. and A.D. Smith, *Using beta-binomial regression for high-precision differential methylation analysis in multifactor whole-genome bisulfite sequencing experiments.*, in *Bmc Bioinformatics*. 2014. p. 215.
181. Jaffe, A.E., et al., *Bump hunting to identify differentially methylated regions in epigenetic epidemiology studies*, in *Int J Epidemiol*. 2012. p. 200-209.
182. Sen, A., et al., *Lead exposure induces changes in 5-hydroxymethylcytosine clusters in CpG islands in human embryonic stem cells and umbilical cord blood.*, in *Epigenetics*. 2015. p. 607-621.
183. Yuan, T., et al., *An integrative multi-scale analysis of the dynamic DNA methylation landscape in aging.*, in *PLoS Genet*. 2015. p. e1004996.
184. Kappil, M.A., et al., *Placental Expression Profile of Imprinted Genes Impacts Birth Weight.*, in *Epigenetics*. 2015. p. 0-8.
185. Salas, L.A., et al., *DNA methylation levels and long-term trihalomethane exposure in drinking water: an epigenome-wide association study.*, in *Epigenetics*. 2015. p. 650-661.
186. Aschebrook-Kilfoy, B., et al., *Polybrominated Diphenyl Ethers and Thyroid Cancer Risk in the Prostate, Colorectal, Lung, and Ovarian Cancer Screening Trial Cohort.*, in *Am J Epidemiol*. 2015. p. 883-888.
187. Kerr, J.S. and C.H. Wilson, *Nuclear receptor-binding protein 1: a novel tumour suppressor and pseudokinase.*, in *Biochem. Soc. Trans.* 2013. p. 1055-1060.
188. Ruiz, C., et al., *High NRBP1 expression in prostate cancer is linked with poor clinical outcomes and increased cancer cell growth.*, in *Prostate*. 2012. p. 1678-1687.
189. Wei, H., et al., *NRBP1 is downregulated in breast cancer and NRBP1 overexpression inhibits cancer cell proliferation through Wnt/beta-catenin signaling pathway.* Onco Targets Ther, 2015. **8**: p. 3721-30.
190. St Pourcain, B., et al., *Association between a high-risk autism locus on 5p14 and social communication spectrum phenotypes in the general population*, in *Am J Psychiatry*. 2010. p. 1364-1372.
191. Chojnicka, I., et al., *Analysis of four genes involved in the neurodevelopment shows association of rs4307059 polymorphism in the cadherin 9/10 region with completed suicide.*, in *Neuropsychobiology*. 2012. p. 134-140.

192. DeGeer, J., et al., *Hsc70 chaperone activity underlies Trio GEF function in axon growth and guidance induced by netrin-1.*, in *J. Cell Biol.* 2015, Rockefeller Univ Press. p. 817-832.
193. Podjaski, C., et al., *Netrin 1 regulates blood-brain barrier function and neuroinflammation.*, in *Brain*. 2015. p. 1598-1612.
194. Lesnick, T.G., et al., *Beyond Parkinson disease: amyotrophic lateral sclerosis and the axon guidance pathway.* PLoS One, 2008. **3**(1): p. e1449.
195. Lin, L., et al., *Axon guidance and synaptic maintenance: preclinical markers for neurodegenerative disease and therapeutics.*, in *Trends Neurosci.* 2009. p. 142-149.
196. Su, A.I., et al., *A gene atlas of the mouse and human protein-encoding transcriptomes.*, in *Proc Natl Acad Sci U S A.* 2004, National Acad Sciences. p. 6062-6067.
197. Wu, C., et al., *BioGPS: an extensible and customizable portal for querying and organizing gene annotation resources.*, in *Genome Biol.* 2009, BioMed Central Ltd. p. R130.
198. Hsieh, H.L., et al., *S100 protein translocation in response to extracellular S100 is mediated by receptor for advanced glycation endproducts in human endothelial cells.* Biochem Biophys Res Commun, 2004. **316**(3): p. 949-59.
199. Marenholz, I., C.W. Heizmann, and G. Fritz, *S100 proteins in mouse and man: from evolution to function and pathology (including an update of the nomenclature),* in *Biochemical and Biophysical Research Communications*. 2004. p. 1111-1122.
200. Dishaw, L.V., et al., *Exposures, mechanisms, and impacts of endocrine-active flame retardants.*, in *Curr Opin Pharmacol*. 2014. p. 125-133.
201. Fowles, J.R., et al., *Immunologic and endocrine effects of the flame-retardant pentabromodiphenyl ether (DE-71) in C57BL/6J mice.*, in *Toxicology*. 1994. p. 49-61.
202. Stoker, T.E., et al., *Assessment of DE-71, a commercial polybrominated diphenyl ether (PBDE) mixture, in the EDSP male and female pubertal protocols.*, in *Toxicol Sci*. 2004. p. 144-155.
203. Zhou, T., et al., *Developmental exposure to brominated diphenyl ethers results in thyroid hormone disruption.*, in *Toxicol Sci*. 2002. p. 105-116.
204. Makey, C.M., et al., *Polybrominated Diphenyl Ether Exposure and Thyroid Function Tests in North American Adults,* in *Environ Health Persp*. 2015. p. 1-26.
205. Zota, A.R., et al., *Polybrominated diphenyl ethers, hydroxylated polybrominated diphenyl ethers, and measures of thyroid function in second trimester pregnant women in California.*, in *Environ. Sci. Technol.* 2011, American Chemical Society. p. 7896-7905.
206. Chevrier, J., et al., *Prenatal exposure to polybrominated diphenyl ether flame retardants and neonatal thyroid-stimulating hormone levels in the CHAMACOS study.*, in *Am J Epidemiol*. 2011, Oxford University Press. p. 1166-1174.
207. Li, L., et al., *Quantitative structure-property relationships for octanol-water partition coefficients of polybrominated diphenyl ethers.*, in *Chemosphere*. 2008. p. 1602-1606.
208. Wong, F., I.T. Cousins, and M. Macleod, *Bounding uncertainties in intrinsic human elimination half-lives and intake of polybrominated diphenyl ethers in the North American population.*, in *Environ Int*. 2013. p. 168-174.
209. Sjodin, A., et al., *Serum concentrations of polybrominated diphenyl ethers (PBDEs) and polybrominated biphenyl (PBB) in the United States population: 2003–2004,* in *… science & technology*. 2008.

210. Houseman, E.A., et al., *DNA Methylation in Whole Blood: Uses and Challenges.*, in *Curr Environ Health Rep*. 2015. p. 145-154.
211. Wang, X., et al., *Obesity related methylation changes in DNA of peripheral blood leukocytes.*, in *BMC Med*. 2010. p. 87.
212. Lendle, S.D., M.S. Subbaraman, and M.J. van der Laan, *Identification and efficient estimation of the natural direct effect among the untreated.* Biometrics, 2013. **69**(2): p. 310-7.
213. Jiang, Z. and T.J. VanderWeele, *When Is the Difference Method Conservative for Assessing Mediation?* Am J Epidemiol, 2015. **182**(2): p. 105-8.
214. van der Laan, M.J. and S. Rose, *Targeted Learning: Causal Inference for Observational and Experimental Data*. 2011: Springer New York.