DNA METHYLATION AND THE DEVELOPMENTAL ORIGINS OF HEALTH AND DISEASE:

Epigenome-Wide Associations in Canadian Birth Cohorts

Randa Stringer

Supervisor: Dr. Guillaume Pare

Committee: Dr. Sonia Anand & Dr. Joseph Beyene

TABLE OF CONTENTS

CHAPTER 1: INTRODUCTION

1.1 Developmental Origins of Health and Disease (DOHaD)

The Developmental Origins of Health and Disease (DOHaD) is a field that has become prominent in recent years for its potential role in the prevention of non-communicable diseases (NCDs) such as diabetes, cancer, and asthma/allergy. Increasing evidence has demonstrated that early life exposures, including those that occur *in utero,* may have significant and life-long effects on an individual's health and predisposition to disease. This process has come to be known as 'fetal programming', where the prenatal environmental exposures experienced leads to an 'adaptive' physiological response in the fetus.

The DOHaD field was introduced in the 90's with the description of the 'thrifty gene' hypothesis. Hales and Barker proposed in 1992 that inadequate nutrition in early development impaired endocrine pancreatic function and increased susceptibility to type 2 diabetes mellitus (T2DM).[1] Specifically, they suggested that poor nutrition *in utero* and in early life cause the growing fetus (often with low birth weight) to develop mechanisms to conserve and utilize nutrients; thus, the 'thrifty phenotype' hypothesis was born. This concept of fetal programming, in which environmental influences have a persistent effect on the growing fetus' structural and functional development, has been further characterized by the increasing body of DOHaD research.[2]

A key series of studies in the field involved outcomes related to the Dutch Hunger Winter. In the spring and winter of 1944, German occupation and train strikes combined to cause famine across social classes in the western region of The Netherlands during which daily caloric intake was only 400 – 800 calories per person. Women who were pregnant at the time and their offspring were thus exposed to prenatal undernutrition, and these individuals have been studied since the tragedy of the Hunger Winter to observe its effects on their long-term health.[3] They have shown increased risk of glucose intolerance, heart disease, and other chronic illnesses well into adult life, providing compelling evidence for fetal programming based on this severe prenatal caloric restriction.[4]

As the prevalence of NCDs increases, knowledge of how they develop and what we may do to prevent them is becoming more necessary to individual and population health. Insights into the earliest aetiologies of these diseases provide an opportunity for long-lasting intervention that may ease societal burden while improving health for both pregnant mothers and their offspring.

1.2 Gestational Diabetes Mellitus (GDM)

The original description of the 'thrifty phenotype' was based on observations of increased type 2 diabetes in low birth weight populations, and issues of maternal and offspring dysglycemia have continued to occupy a prominent space in DOHaD research.[1] Gestational diabetes mellitus (GDM) has drawn particular focus, as this dysglycemic condition can have significant short- and long-term effects on both the mother and offspring. It has been estimated to affect 2 – 12% of pregnancies, and rates are increasing with the rising prevalence of obesity.[5] During the prenatal period GDM increases risk of respiratory distress, macrosomia, cardiac malformations, Caesarian section, birth trauma, preterm birth, and pre-eclampsia/eclampsia; mothers who develop GDM during their pregnancy later experience a higher rate of metabolic syndrome and T2DM as well.[6] Offspring exposed to GDM show increased glucose dysregulation in the long-term, including increased levels of cardiovascular disease (CVD), hypertension, and T2DM.[7] Cohort studies involving mothers and infants experiencing GDM provide an invaluable resource in elucidating the basis of fetal programming.

Like T2DM, GDM is characterized by insulin resistance, and can be diagnosed when glucose intolerance first occurs in pregnancy.[7] The oral glucose tolerance test (OGTT) is routinely administered between the 24th and 28th week of pregnancy in either a 2- or 1-step method with a 50g or 75g glucose challenge, respectively. Glucose is measured at baseline (fasting), 1, and 2 hours, and a single elevated value is diagnostic for GDM. The cut-off values set by the International Association of Diabetes and Pregnancy Study Groups (IADPSG) and commonly used in Canada are, for a 2-step OGTT, glucose levels of ≥ 5.3 mmol/L at fasting, ≥ 10.6 mmol/L at 1h, and ≥ 9.0 mmol/L at 2 hours. Slightly lower cut-offs for a 1-step OGTT are ≥ 5.1 mmol/L, ≥ 10.0 mmol/L, and ≥ 8.5 mmol/L at fasting, 1 hour, and 2 hours, respectively.

However, these values may not be optimal internationally and for all ethnicities. South Asian populations in particular face a high rate of GDM, estimated at up to 24% of all pregnancies.[8] While the 'thrifty phenotype' was initially described as such because it was based on undernourished and low birth weight infants, studies in South Asian populations expanded the concept to show an increased incidence of T2DM in adults who were short at birth and had a high ponderal index, indicating a U-shaped risk pattern for birth weight.[9] The Pune Maternal Nutrition Study further characterized what they described as the 'thin-fat Indian baby', demonstrating that babies born to rural Indian mothers compared to white Caucasian women in Southampton, UK, were smaller in all body measurements but preserved body fat as measured by skinfold thicknesses.[10] The authors suggest that this may continue postnatally and predispose to insulin resistance, adding further complexity to the DOHaD relationships and mechanisms potentially at play. More recent study of South Asian women in the Born in Bradford cohort has established that optimal diagnostic cut-offs may also vary based on ethnicity, and identified values of ≥ 5.2 mmol/L and ≥ 7.2 mmol/L at fasting and 2 hours which they suggest using in South Asian populations.

1.3 DNA Methylation & Epigenetics

While the evidence for fetal programming in exposures such as dysglycemia is substantial, a key question in DOHaD is how exactly these prenatal environments can affect the fetus in ways that persist into adulthood. One of major mechanisms theorized to underly this process is epigenetic modification. Epigenetics refers to genetic changes that do not affect the genetic code but modify gene expression in heritable, transmissible ways. These changes include histone modifications, microRNAs (miRNAs), and the best characterized mechanism, DNA methylation. Histone modifications involve the addition of chemical groups (ex. acetylation, methylation, phosphorylation) to specific amino acids in the histone proteins around which DNA strands wrap. Factors such as the size and charge of the added group(s) can alter chromatin structure by causing expansion or condensation of DNA packaging. This in turn affects the ability of transcription factors and other proteins to access the necessary sequence(s) to initiate gene expression. miRNAs are small, non-coding RNAs (ncRNAs) which contain sequence that is complimentary to one or more transcriptional products. By binding to messenger RNA (mRNA) transcripts, miRNAs reduce translation of their complimentary targets and thereby inhibit protein production.

DNA methylation directly affects the nuclear DNA through the addition or subtraction of a methyl group to the 5' carbon of a cytosine base. This modification is maintained through DNA replication by the action of DNA methyltransferases (DNMTs) and is therefore heritable across cell division and can be maintained long-term. Methylation occurs at CpG dinucleotides, where a cytosine and guanine are adjacent. Overall these dinucleotides are underrepresented in the human genome, but those present are generally concentrated in 'islands' often located upstream of gene promoter regions. CpG islands tend to have consistent methylation across sites and contribute to genetic regulation. Methylation levels near gene promoters may affect protein binding and influence gene expression, with increased methylation generally leading to decreased expression and vice versa.

In combination, these epigenetic mechanisms play a crucial role in cellular differentiation and environmental interaction. During development stem cells undergo dramatic epigenetic changes as part of the process of both stimulating and maintaining differentiation.[11] Similarly, the relative plasticity of the epigenome (compared to the genetic code) allows cells to respond to environmental stimuli such as stress and toxins in a productive (gene expression changes) and transmissible manner.[12] DNA methylation in offspring can be affected not only by the parental epigenomes but also by the fetus' *in utero* experience, making this a prime candidate mechanism by which environmental exposures could result in long-term fetal programming.

Unsurprisingly, GDM has remained a key exposure of interest. Targeted studies have often focused on the *IGF2*/*H19* locus, where imprinted genes are known to show methylation changes when programming the metabolic profile in early life.[13] Methylation at the *IGF2/H19* locus was associated with low birth weight and intrauterine hyperglycemia in a GDM model.[13] *IGF2/H19* methylation appears to modulate fetal growth and may be crucial in metabolic programming for late onset obesity.[14] At the genome-wide level, recent EWASs have demonstrated genome-wide changes to offspring methylation associated with exposure to GDM.[15] Other prenatal exposures and childhood health outcomes have also shown associations with differential DNA methylation.[16–18]

The primary method for assessing DNA methylation involves treating DNA with sodium bisulfite. Unlike DNA replication within the cell, expansion using the polymerase chain reaction (PCR) does not maintain DNA methylation patterns. Genomic DNA is therefore treated with sodium bisulfite, which converts all unmethylated cytosine bases to uracil. 5-methylcytosine is unaffected by this treatment, so following amplification of converted DNA any remaining cytosine bases are known sites of methylation. Unmethylated cytosines will instead be sequenced as thymine bases. Chip-based technology incorporates this method with probe-based assays to provide genome-wide coverage of DNA methylation. The Illumina 27K, 450K, and EPIC BeadChip arrays assess methylation at > 27 000, > 450 000, and > 850 000 sites across the genome, respectively.

Site-specific DNA methylation is generated using two probe types on the 450K and EPIC models, which necessitates further statistical adjustment to normalize. Relative red and green intensities are converted to β values as a representation of methylation. The methylated (M) and unmethylated (U) channels are used to calculate $\beta = M/(M+U+\alpha)$, where α is an offset value. β values range from 0 (no methylation) to 1.0 (fully methylated), and particularly in a heterogeneous tissue like blood, most sites will show partial methylation as the β value represents a composite measure from all cells involved.

1.4 Blood Composition & Methylation

Blood is often sampled for use in both GWASs and EWASs, but poses significant challenges in accurately assessing epigenetic markers. While genomic sequence is expected to remain consistent across cell types, DNA methylation and other epigenetic processes may show dramatic variation between cell populations. Although blood is a heterogeneous tissue, blood-based methylation methods will simply report a single β value per site which represents the mean value across all cell types within the sample. Not correcting for inter-individual variations in cellular composition may therefore yield biased results driven or influenced by blood composition rather than the variable of interest.

Solutions for this issue have focused on deriving estimated cellular composition from the overall methylation values observed using two primary methods: reference-free, and reference-based. Reference-free algorithms use statistical models to represent underlying composition; for example, the ReFACTor algorithm generates principal components that can act as covariates to account for cell composition.[19] In contrast, reference-based methods depend on reference datasets detailing methylation patterns of purified blood cells. The Houseman et al. (2012) method uses 100 significant probes with the greatest magnitude of effect (50 positive and 50 negative) for each cell type and combines these data with user-inputted datasets to estimate the relative proportions of named cells: B, CD4$^+$ T, and CD8$^+$ T cells, eosinophils, granulocytes, monocytes, neutrophils, and NK cells.[20]

Studies using cord blood samples have faced additional hurdles as the differences in cellular composition between cord and peripheral blood necessitate separate reference panels. Bakulski et al. (2016) developed a reference-based method for calculating cord blood cellular composition based on the Houseman algorithm which uses 100 significant probes with the greatest magnitude of effect (regardless of direction) for each cell type and provides estimates for B, CD4$^+$ T, and CD8$^+$ T cells, granulocytes, monocytes, neutrophils, and nucleated red blood cells (nRBCs).[21] The reference panel incorporates a smaller number of individual samples than those referenced for peripheral blood and may generate less accurate estimates. However, despite the limitations, both methods provide a novel option for estimating cellular composition in blood without the need for cytometry.

1.5 Conclusions

The study of genome-wide DNA methylation in cord blood is an essential experimental model in characterizing the process of fetal programming. Evidence suggests that prenatal exposures may modify infant DNA methylation, patterns which may be maintained long-term and influence lifelong health. This is a key mechanism of interest in the DOHaD field, but there remain many difficulties in assessing the evidence available thus far.

In contrast to GWASs, the results of which are compiled in numerous databases, EWASs remain largely independent with little apparent effort to review and replicate findings. There are many published EWAS papers, but finding all those relating to a particular outcome remains challenging. Furthermore, attempts at replication are inconsistent, and some studies (particularly those published soon after the release of the 450K BeadChip) fail to make sufficient statistical adjustments to their datasets especially regarding blood cell deconvolution. This thesis will therefore provide an overview and replication of existing evidence in the field while further exploring the nature of the relationship between dysglycemia and infant DNA methylation.

CHAPTER 2: DNA Methylation Changes in Cord Blood and the Developmental Origins of Health and Disease – a Review and Replication Study

2.1 Introduction

The Developmental Origins of Health and Disease (DOHaD) is an area of study that has become increasingly popular in recent years. It is now evident that early life exposures, including those happening *in utero*, have significant effects on the development of later health conditions in children and adults. This may occur through fetal programming, whereby the environment experienced *in utero* may 'program' a fetus for the expected environment outside the womb in a manner that may have a significant impact on health. It is suspected that epigenetic mechanisms, in particular DNA methylation, may mediate the process of fetal programming. Alterations in DNA methylation have been associated with a variety of prenatal exposures and childhood health outcomes, including: gestational diabetes, maternal smoking, gestational age, and asthma.[16–18,22] It has been well established that the prenatal environment can modify the fetal epigenome, which may lead to long-term effects.

While DNA methylation has been consistently implicated in the DOHaD model, there has been a lack of consistency in the way this problem has been addressed. Due in part to technological constraints, many early analyses were targeted to specific loci already known to be involved in the outcome of interest. The development of the Illumina HumanMethylation27 BeadChip (27K) and the subsequent 450K and Epic models have allowed true epigenome-wide association studies (EWASs) to be conducted. This has provided an unprecedented scope and consistent site notation to the field which allows for replication and validation of previous findings.

Perhaps in part because of the relative novelty of the genome-wide technology, there is currently a lack of consensus around findings in many areas. A PubMed search for ‘epigenome-wide association study’ generates almost 600 results, but without consistency and replication it is very difficult to assess the merit of this growing body of evidence. We thus propose to search and review the literature surrounding several DOHaD areas of interest and conduct a replication study in two birth cohorts. This will serve to 1) gather the available evidence for the role of DNA methylation in fetal programming by maternal dysglycemia, maternal pre-pregnancy BMI, diet during pregnancy, maternal smoking, and gestational age; 2) identify any loci implicated consistently across these studies to better characterize the underlying processes, and 3) help to demonstrate how replicable these DNA methylation findings may be and suggest causes for any lack of reproducibility.

2.2 Methods

2.2.1 Literature Search Strategy and Study Selection

A generalized search was conducted in the PubMed database in order to identify EWASs conducted in birth cohorts and related to the DOHaD paradigm. In concordance with the outcomes of interest in the NutriGen data, we restricted our search to five exposures: prenatal nutrition (especially dietary patterns), maternal smoking, maternal or infant dysglycemia (especially GDM), gestational age, and maternal pre-pregnancy BMI. The following search was used: "cord blood" AND "DNA methylation" AND "pregnancy" AND ("diet" OR "nutrition" OR "dietary pattern" OR "diabetes" OR "dysglycemia" OR "glucose" OR "insulin" OR "smoking" OR "gestational age" OR "weight" OR "BMI"). All returned abstracts were then screened for our inclusion and exclusion criteria, with some studies moving on for full text evaluation and possible inclusion in our review.

In addition to the requirement that EWASs be related to one of our five exposures of interest, we also restricted included studies to those reporting a genome-wide significant association in cord blood. Targeted studies, those that reported statistically insignificant results, or EWASs in peripheral blood or placental tissue were therefore excluded, among others. Our search returned a total of 151 results which were screened at the abstract level. Of these, 42 full text articles were accessed for closer inspection, and a final 15 studies were identified for inclusion in our review.

2.2.2 Replication in the Literature

For three of our exposures of interest (gestational age, dysglycemia, and smoking), we identified more than one published EWAS paper reporting genome-wide significant DNA methylation changes in cord blood. We therefore extracted the significant sites reported in each of these studies in order to assess the status of the literature for each exposure and whether any sites had been independently replicated across papers. We therefore first looked for any sites identified as significant in 2 or more EWASs of the same exposure, also considering whether the direction of effect was consistent between studies. Next, significant loci for each outcome were assessed for their proximity to one another; loci showing 2 or more significant sites < 50kb, < 100kb, and < 200kb apart were identified. Finally, all significant sites across all four represented outcomes (BMI, GDM, gestational age, and smoking) were compared to isolate any sites significantly associated with more than one of our exposures of interest.

2.2.3 NutriGen Study, Cohorts, & Measures

Data were generated as part of the NutriGen study, an alliance of four Canadian birth cohorts seeking to better understand the impact of maternal and infant nutrition on infant and child health and disease. The Family Atherosclerosis Monitoring In Early life (FAMILY) cohort is designed to study determinants of CVD and is comprised of 859 mothers/901 infants, primarily white Caucasian, from Southwestern Ontario. They were recruited between 2004 and 2009 with long-term (5+ year) follow-up. The Canadian Healthy Infant Longitudinal Development (CHILD) study is a longitudinal birth cohort study based out of four Canadian centres. 3 600 mother/child pairs were recruited between 2008 and 2012 with the primary goal of investigating environmental and genetic determinants of allergic disorders. The South Asian birth Cohort (START) began recruiting in 2011 in urban Canada as well as rural and urban Bangalore in order to study environmental, genetic, and epigenetic influences on adiposity, growth, and cardio-metabolic factors in a birth cohort of South Asian women. Finally, the Aboriginal Birth Cohort (ABC) enrolled pregnant mothers from the Six Nations Reserve in Ontario. Beginning in 2012 women were recruited to study the determinants of cardiometabolic health and type 2 diabetes (T2DM) in an Aboriginal population.

The NutriGen study combines participants from all four of these cohorts to investigate nutrition and environmental determinants of childhood health, with a focus on genetic, epigenetic, and microbiome contributions. All cohorts administered a food frequency questionnaire (FFQ) to assess prenatal diet and collected anthropometric and other health measures of the mother and infant throughout pregnancy and early childhood. These variables have been harmonized across cohorts for comparisons. Cord blood was also collected at birth for genetic and cardiometabolic analysis.

Many of the metrics collected by the NutriGen cohorts are utilized in our epigenetic investigations. FFQ data were explored with principal component (PC) analysis to derive individual scores representing dietary patterns; vegetarianism was represented as a dietary pattern in these models, and was also tested as a unique binary yes/no variable. Individual nutrient information regarding fatty acid intake was used, including: polyunsaturated fatty acids (PUFAs), saturated fats, and the ratio of PUFAs to saturated fats (P:S). Two diet quality scores, the modified Alternative Healthy Eating Index (mAHEI) and the DOHaD score developed by the NutriGen group, were calculated and tested as well. Dietary variables were adjusted for total energy consumption when applicable. In START, OGTT results were available (as well as the harmonized GDM variable in both START and CHILD), and AUC glucose tested as another measure of maternal glycemic control. It was modeled as both a continuous variable and a binary variable based on a cut-off of 835. The OGTT results also allowed us to conclude participants' GDM status using the previously described cut-offs developed in the Born in Bradford cohort for South Asian women (Farrar et al., 2015). Maternal smoking and pre-pregnancy BMI were available in CHILD only (START had no smokers).

To further study the role of genetics and epigenetics in the relationship between prenatal exposures and offspring health, most participants (mothers and infants) were assessed for genome-wide genotyping using the Illumina HumanCoreExome BeadChip. A subset of infant in START and CHILD, approximately 500 from each, were also assayed for genome-wide DNA methylation. Because CHILD is a multi-ethnic cohort (unlike START, which exclusively recruited South Asian women, the samples chosen for methylation analysis were distributed across the ethnicities represented.

2.2.4 Genome-Wide Methylation Assay

Cord blood samples were collected upon delivery in each cohort and processed for DNA extraction using standard protocols. Bisulfite conversion was used to prepare DNA samples for chip-based assay. Samples were hybridized to the Illumina HumanMethylation450K BeadChip array, designed for genome-wide DNA methylation assessment. The 450K chip contains > 485 000 sites covering > 96% of RefSeq genes, with an average (but variable) 17 probes per gene. It is a two-colour array that also employs two types of probes for measurement to improve coverage. In loci assessed with the Infinium I assay, there are two probes per site: a ‘methylated’ probe corresponding to a cytosine extension, and an ‘unmethylated’ probe with a thymine extension. The relative intensities of these two probes is used to calculate methylation at these loci. In the Infinium II design, a single probe is used, which can be extended with either a ‘methylated’ or ‘unmethylated’ base. The relative intensities generated by each of the extensions provides an estimate of methylation. The Infinium II design is more suited to many areas of the genome and increases the total number of sites that can be included for assay; this probe type therefore represents the majority of the sites on the 450K.

A total of 512 START and 511 CHILD cord blood samples were selected from their respective cohorts and randomized across arrays for methylation assessment. Illumina iScan software was used to read intensities and generate and export idat files to R for pre-processing and quality control.

2.2.5 Data Pre-Processing and Quality Control

Raw data from iScan were imported into R version 3.2.0 with the *minfi* package, which was used for all subsequent quality control and pre-processing.[23] Samples from the START and CHILD cohorts were processed separately. Sample quality was assessed in each cohort first based on missingness criteria; any samples with a proportion of failed probes > 0.01 was removed from analysis (a total of 2 samples in START and 14 in CHILD). The getSex function in *minfi* was also used to estimate biological sex in each sample based on the methylation patterns of the X and Y chromosomes. This was compared to the reported sex and inconsistent samples were removed. In total 5 sample were removed from START and 7 from CHILD based on these criteria. A final sample of 506 individuals in START and 511 in CHILD remained.

Probe quality was determined with missingness criteria based on probe failure in > 0.05 of samples. Any probes exceeding this threshold were excluded: 756 from START, 634 from CHILD. In addition, certain underlying probe sequences have been demonstrated to influence perceived methylation status. Probes that contain a single nucleotide polymorphism (SNP) at or near the site of interrogation may bind inconsistently depending on genotype. There are also many probes that, in part due to the difficulty of working with bisulfite-converted DNA, have been shown to hybridize to more than one genomic location. This could also lead to spurious methylation estimates. For this reason, all probes known to contain a SNP (70 889) or demonstrate cross-reactivity (29 233) were also removed from analysis. A final dataset of 393 400 probes in START and 393 449 probes in CHILD remained following these quality control measures. Table 1 summarizes the quality control measures applied and the samples/sites lost at each level.

Table 1. Summary of samples and probes excluded at each stage of quality control.

Samples			Probes		
	START	**CHILD**		**START**	**CHILD**
Initial	512	511	**Initial**	> 485 000	
Sex Check	5	7	**Failed**	756	634
Missingness	2	14	**Polymorphic**	70 889	
Final	506	491	**Cross-Reactive**	29 233	
			Final	393 400	393 449

The two-type probe design of the 450K chip presents an obstacle in data processing, as the probe types generate two independent peaks that must be unified. The subset within-array normalization (SWAN) method was used to normalize the START data following quality control to reduce differences in beta value distribution between probe types.[24] Technical differences such as when the chip was run can also affect the distribution of beta values across samples. The comBat algorithm in the *sva* package, which uses surrogate variables, was applied to correct for these potential batch effects based on chip.[25] The reference-free ReFACTor algorithm was used to determine the top 7 principal components of cellular composition, which were incorporated into our statistical models.[19]

A dataset based on regions of DNA methylation was also generated using the cpgCollapse function in *minfi*. Briefly, this method combines any sites less than 500 bp apart with a total width no greater than 1500 bp into a single region. This average of nearby beta values increases statistical power while better reflecting the underlying biology of DNA methylation patterns.

2.2.6 Epigenome-Wide Association

We tested a variety of exposures and outcomes for association with cord blood DNA methylation, focusing on areas similar to our literature review: diet (including dietary patterns, macronutrient intake, diet quality scores, gestational age, GDM, and smoking during pregnancy (CHILD only). GDM and dysglycemia outcomes were assessed for association in our regional datasets as well as the standard site-by-site data. Statistical analysis was applied only to the white European subset ($N = 295$) of the CHILD population, as this was the only ethnic group with enough individuals available for study. Table 2 outlines the variables tested for association with DNA methylation in both the START and CHILD cohorts.

Table 2. Variables tested for epigenome-wide association in the START and CHILD cohorts

Cohort	Diet	Dysglycemia	Other
START	Patterns - PC-based - Vegetarian Nutrients - eaPUFAs - ea P:S Scores - mAHEI - DOHaD	GDM - Harmonized - SA cut-offs AUC glucose - Continuous Binary	GestAge
CHILD	Patterns - PC-based - Vegetarian Nutrients - eaPUFAs - ea P:S Scores - mAHEI - DOHaD	GDM - Harmonized	GestAge Maternal Smoking

Multivariable linear regression models were used to test for association between DNA methylation β values at each site and our continuous variables; multivariable logistic regression models were used for binary outcomes. Metrics known or observed to influence DNA methylation were incorporated as covariates in our models: maternal age, infant sex, gestational age, study centre, processing time (CHILD only), smoking (CHILD only), and cellular composition. Statistical significance was assessed using the Bonferroni multiple testing correction with thresholds of $p < 1.27 \times 10^{-7}$ and $p < 2.76 \times 10^{-7}$ for the site-by-site and regional datasets, respectively. False discovery rate-adjusted p values were also calculated, with significance set at FDR-adjusted $p < 0.05$. All association testing took place in START and CHILD separately, generating independent results for each cohort.

2.2.7 Replication Study

Where possible given available data, we used the START and CHILD cohorts to conduct a targeted replication analysis based on the findings of our literature review. Linear and logistic regression models were constructed for each variable as they were in our EWAS analyses, but only those sites identified in the literature as significantly associated with a given outcome were tested. Data for gestational age and GDM were available in both START and CHILD, while smoking and pre-pregnancy BMI were assessed only in CHILD. Bonferroni-adjusted p-value thresholds were set for each model. Table 3 summarizes the traits and sites tested in each cohort as well as the significance thresholds applied.

Table 3. Replication models tested in START and CHILD for each outcome

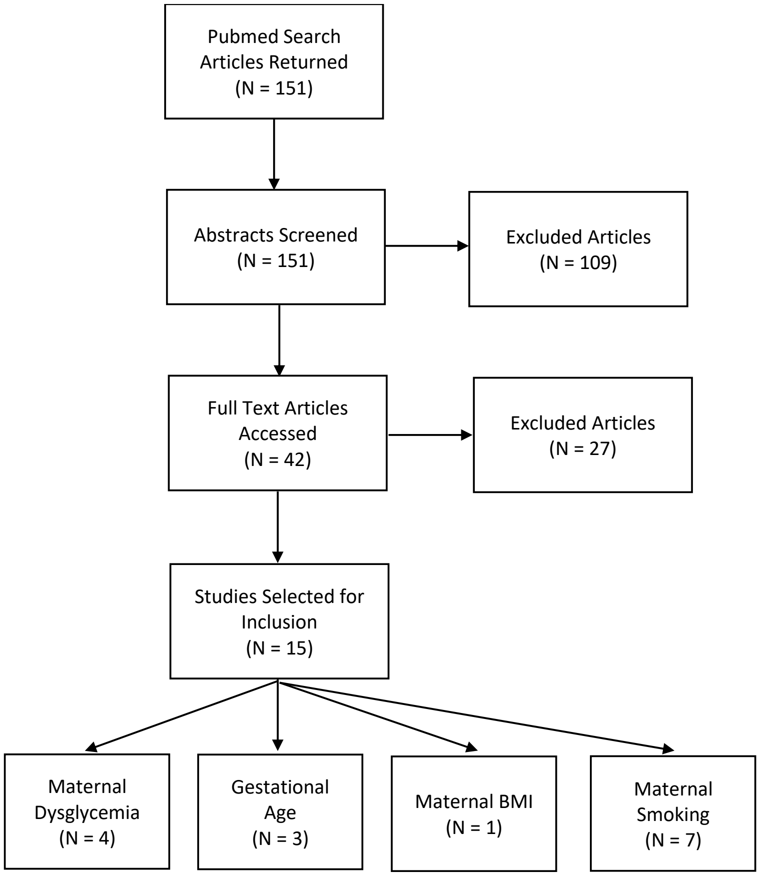
Outcome	Sites in Literature	Available in CHILD	Available in START	Adjusted threshold
GDM	307	216	216	< 0.00023
GestAge	309	278	279	< 0.00018
Smoking	161	94	NA	< 0.00053
BMI	1	1	NA	< 0.05

2.3 Results

2.3.1 Literature Search & Review

Our literature search returned a total of 151 studies to be screened for our inclusion and exclusion criteria. Abstract screening eliminated 109 studies, leaving 42 remaining for full-text access. This deeper review identified a final 15 papers meeting our criteria. Four of our five areas of interest were represented by at least one study; we did not find any papers reporting genome-wide significant effects of prenatal diet on DNA methylation in cord blood. The 15 included papers are distributed across our remaining exposures: 1 for maternal pre-pregnancy BMI, 3 for gestational age, 4 for maternal dysglycemia, and 7 for maternal smoking. Figure 1 details the literature review and study screening process.

Figure 1. Search and screening process.


Pubmed Search Articles Returned (N = 151)
Abstracts Screened (N = 151)
Excluded Articles (N = 109)
Full Text Articles Accessed (N = 42)
Excluded Articles (N = 27)
Studies Selected for Inclusion (N = 15)
Maternal Dysglycemia (N = 4)
Gestational Age (N = 3)
Maternal BMI (N = 1)
Maternal Smoking (N = 7)

2.3.2 Replication in the Literature

Replication of results within existing literature was inconsistent across outcomes but was identified at multiple sites for both gestational age and maternal smoking. The three included studies focused on gestational age reported a total of 310 significant sites, 9 of which were identified by more than one paper. One site (cg16536918) in the vasopressin (*AVP*) gene was replicated by all 3 papers examined. The 7 studies of maternal smoking reported 142 unique sites, of which 35 were replicated. Two of the 7 maternal smoking studies identified sites with no replication within the literature, but the remaining 5 papers all demonstrated overlapping results, including one site (cg16536918) in the aryl hydrocarbon receptor repressor (*AHRR*) gene replicated by 5/5 of these studies. In all replicated sites (a total of 9 for gestational age and 35 for maternal smoking), the reported direction of effect (increased or decreased methylation) was consistent in all instances of significant association for that site. Figures 2 and Table 4 illustrate the replicated sites observed within the literature for association with gestational age and maternal smoking, respectively.

Figure 2. Overview of all sites associated with gestational age. **Green** sites showed increased methylation, **red** sites were decreased.

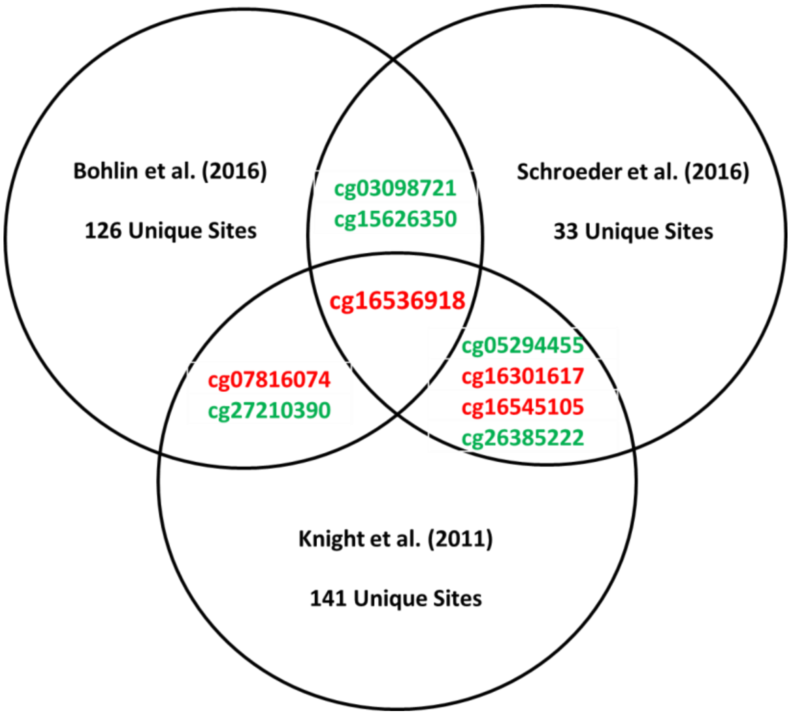
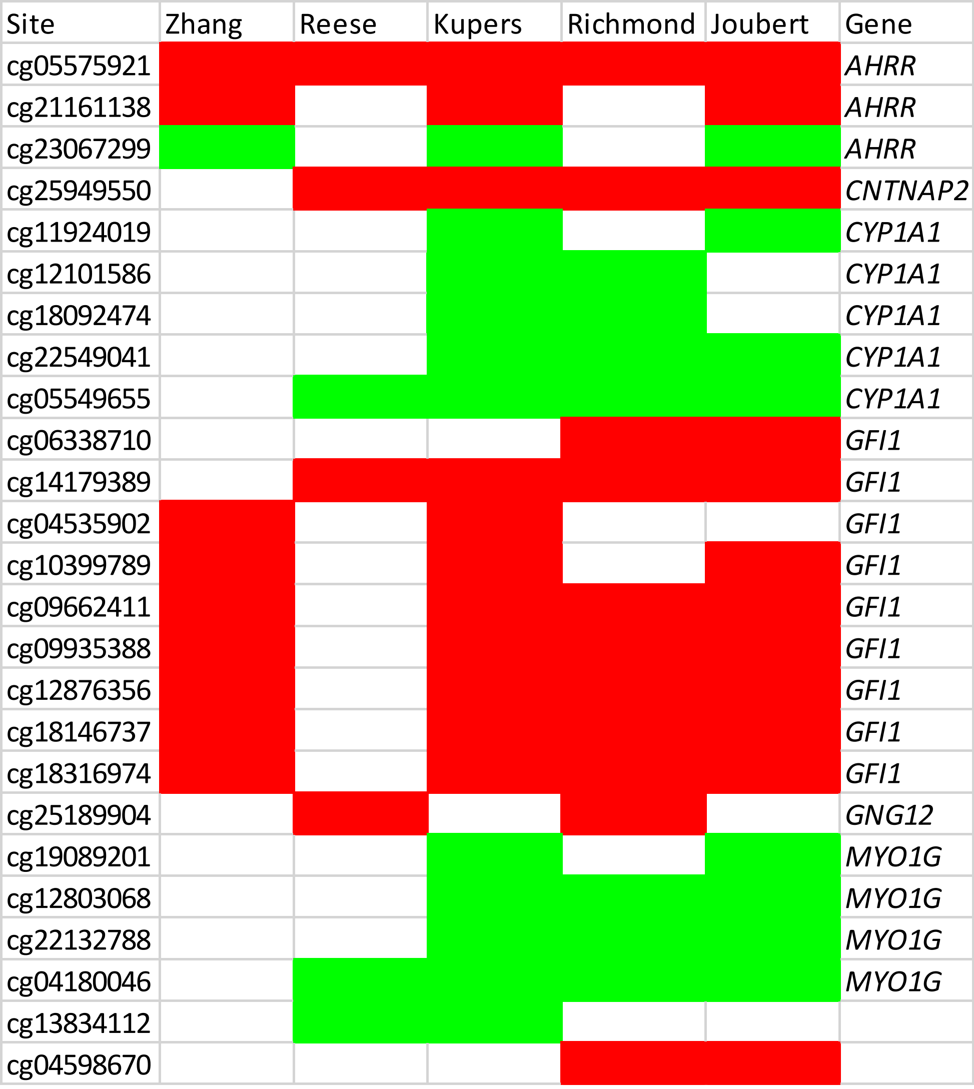


Table 4. Significant sites and direction of effect duplicated across studies. **Green** sites showed increased methylation; **red** sites were decreased.



Site	Zhang	Reese	Kupers	Richmond	Joubert	Gene
cg05575921	🟥	🟥	🟥	🟥	🟥	*AHRR*
cg21161138	🟥		🟥		🟥	*AHRR*
cg23067299	🟩		🟩		🟩	*AHRR*
cg25949550		🟥	🟥	🟥	🟥	*CNTNAP2*
cg11924019			🟩		🟩	*CYP1A1*
cg12101586			🟩	🟩		*CYP1A1*
cg18092474			🟩	🟩		*CYP1A1*
cg22549041			🟩	🟩	🟩	*CYP1A1*
cg05549655		🟩	🟩	🟩	🟩	*CYP1A1*
cg06338710				🟥	🟥	*GFI1*
cg14179389		🟥	🟥	🟥	🟥	*GFI1*
cg04535902	🟥		🟥			*GFI1*
cg10399789	🟥		🟥		🟥	*GFI1*
cg09662411	🟥		🟥	🟥	🟥	*GFI1*
cg09935388	🟥		🟥	🟥	🟥	*GFI1*
cg12876356	🟥		🟥	🟥	🟥	*GFI1*
cg18146737	🟥		🟥	🟥	🟥	*GFI1*
cg18316974	🟥		🟥	🟥	🟥	*GFI1*
cg25189904		🟥		🟥		*GNG12*
cg19089201			🟩		🟩	*MYO1G*
cg12803068			🟩	🟩	🟩	*MYO1G*
cg22132788			🟩	🟩	🟩	*MYO1G*
cg04180046		🟩	🟩	🟩	🟩	*MYO1G*
cg13834112		🟩	🟩			
cg04598670				🟥	🟥	

The 4 papers focused on GDM were also assessed for replication, but although a total of 307 sites were reported to be significantly associated with GDM, none of these sites was replicated in the literature. The single study of pre-pregnancy BMI identified only 1 significantly associated site which of course showed no replication. Table 5 summarizes the replicated sites for each outcome based on our review of the literature.

Table 5. Sites demonstrating replication in the literature for each outcome.

Outcome	Number of Studies	Total Sig. Sites	Replicated in Literature	Replicated Sites	Genes
GestAge	3	310	9	cg03098721 cg07816074 cg15626350 cg16536918 cg27210390 cg05294455 cg16301617 cg16545105 cg26385222	*TTLL7* *SH3TC1* *ESR1* *AVP* *TOM1L1* *MYL4* *TMC6* *CRHBP* *HCA112*
Smoking	7	142	35	cg05575921 cg21161138 cg23067299 cg25949550 cg05549655 cg11924019 cg12101586 cg18092474 cg22549041 cg04535902 cg06338710 cg09662411 cg09935388 cg10399789 cg12876356 cg14179389 cg18146737 cg18316974 cg25189904 cg04180046 cg12803068 cg19089201 cg22132788 cg04598670 cg13834112	*AHRR* *AHRR* *AHRR* *CNTNAP2* *CYP1A1* *CYP1A1* *CYP1A1* *CYP1A1* *CYP1A1* *GFI1* *GFI1* *GFI1* *GFI1* *GFI1* *GFI1* *GFI1* *GFI1* *GFI1* *GNG12* *MYO1G* *MYO1G* *MYO1G* *MYO1G*
GDM	4	307	0	NA	NA
BMI	1	1	NA	NA	NA

Although some sites were directly replicated within the literature for both the gestational age and smoking outcomes, this may underestimate the actual overlap in findings if multiple papers have reported sites located near each other. We therefore looked for clusters of significant sites identified in different papers but localized within 50, 100, and 200 kb. For gestational age, a total of 12 regions containing nearby sites from different studies were identified (Table 6). Among results of the smoking studies, 5 clusters were identified, several of which contained sites that were directly replicated in addition to those nearby but reported in only a single paper (Table 7). Most of the regions identified for both outcomes had a total genomic span of < 50 kb, which may be a close enough proximity to suggest that perhaps these different site readings are signalling the same locus.

Table 6. Gestational age sites in proximity.

		Total	Total				
Chr	Gene	Sites	Distance	Site	Location	Distance	Papers
5	CRHBP	2	112	cg21842274	76248637	112	Shroeder
				cg16545105	76248749	NA	Knight, Schroeder
6	HIST1H3E	2	131	cg26092675	26225258	131	Bohlin
				cg07922606	26225389	NA	Knight
6	NA	3	7000	cg13959344	32901642	6597	Bohlin
	HLA-DMB			cg17022232	32908239	403	Bohlin
	HLA-DMB			cg00575744	32908642	NA	Knight
10	VENTX	2	69291	cg19875532	135052004	69291	Bohlin

	TUBGCP2; ZNF511			cg15856055	135121295	NA	Knight
11	NEAT1	3	132608	cg11352146	65192550	100544	Bohlin
	SCYL1			cg22417398	65293094	32064	Shroeder
	LTBP3			cg08965235	65325158		Bohlin
12	TENC1	2	2002	cg06311778	53441582	2002	Knight
				cg17094065	53443584	NA	Bohlin
12	NCOR2	3	1211	cg04347477	125002007	467	Bohlin
				cg22580512	125002474	744	Knight
				cg22820108	125003218	NA	Knight
14	TGFB3	2	180	cg03395898	76448011	180	Shroeder
				cg16187883	76448191	NA	Bohlin
16	LOC100128788; SRRM2	3	131889	cg03507326	2801952	46845	Bohlin
	TESSP1			cg19403023	2848797	85044	Knight
	FLYWCH2			cg12741488	2933841	NA	Knight
16	ADCY7	2	82	cg06897661	50322074	82	Bohlin
				cg23580000	50322156	NA	Knight
20	MRPS26; GNRH2	4	39435	cg26060255	3025968	26256	Knight
	OXT			cg26267561	3052224	13119	Shroeder
	AVP			cg25551168	3065343	60	Shroeder

	AVP			cg16536918	3065403	NA	Bohlin,Knight, Schroeder
21	HLCS	2	43208	cg21081878		43208	Bohlin
	DSCR6			cg12564962		NA	Shroeder

Table 7. Smoking sites in proximity.

		Total	Total				
Chr	Gene	Sites	Distance	Site	Location	Distance	Papers
1	GNG12	2	18	cg25189904	68299493	18	Replicated
				cg26764244	68299511	NA	Reese
1	GFI1	9	2293	cg10399789	92945668	464	Replicated
				cg09662411	92946132	55	Replicated
				cg06338710	92946187	513	Replicated
				cg18146737	92946700	125	Replicated
				cg12876356	92946825	210	Replicated
				cg18316974	92947035	297	Replicated
				cg04535902	92947332	256	Replicated
				cg09935388	92947588	373	Replicated
				cg14179389	92947961	NA	Replicated
5	AHRR	9	82454	cg01970407	323320	587	Kupers
				cg23067299	323907	62	Replicated
				cg08606254	323969	44478	Zhang
				cg03991871	368447	396	Joubert

				cg11902777	368843	4535	Richmond
				cg05575921	373378	19542	Replicated
				cg14817490	392920	6440	Kupers
				cg21161138	399360	6414	Replicated
				cg22937882	405774	NA	Kupers
7	MYO1G	4	632	cg19089201	45002287	199	Replicated
				cg22132788	45002486	250	Replicated
				cg04180046	45002736	183	Replicated
				cg12803068	45002919	NA	Replicated
15	CYP1A1	6	1378	cg23680900	75017924	1219	Kupers
				cg05549655	75019143	60	Replicated
				cg12101586	75019203	48	Replicated
				cg22549041	75019251	32	Replicated
				cg11924019	75019283	19	Replicated
				cg18092474	75019302	NA	Replicated

No site significantly associated with GDM was replicated within the literature we reviewed. However, some of the sites reported by different papers are located near each other, including two pairs (cg14088574 and cg02990567, cg08440349 and 84486704) which are less than 50KB apart. There are also 3 reported sites on chromosome 16, each from a different study, and all within 200KB of each other. While none of these nearby sites map to the same genes, their proximity does suggest that certain loci may be more strongly implicated than others. Table 8 shows the pairs of nearby sites, while Table 9 illustrates the potential cluster on chromosome 16.

Table 7. GDM sites in proximity

Chr	Site	Location	Study	Site	Location	Study	Distance
6	cg14088574	33234976	Haertle	cg02990567	33266961	Kang	31985
10	cg11818589	134800741	Weng	cg06355908	134897731	Finer	96990
15	cg13794888	75917792	Kang	cg06717289	75978121	Finer	60329
19	cg11449134	51897791	Haertle	cg26605406	52074318	Kang	176527
20	cg03467235	21003839	Weng	cg23695133	21087075	Kang	83236
16	cg26828643	88802820	Haertle	cg08136432	88902276	Weng	99456

Table 8. GDM cluster on chromosome 16

Chr	Site	Location	Distance	Study
16	cg02219997	84328104	158600	Kang
16	cg08440349	84486704	33317	Haertle
16	cg05208607	84520021	NA	Weng

Finally, the results for each outcome (GDM, gestational age, smoking, and BMI) were compared. A single site, cg11864574, was reportedly negatively associated with gestational age and positively associated with maternal smoking.[18,26] This site is located in the body of the sperm associated antigen 6 (*SPAG6*) gene. Although *SPAG6* has not been the focus of many studies, it was found to be differentially methylated in the promoter region as well as differentially expressed in non-small cell lung cancers.[27]

2.3.3 *Population Characteristics*

Table 9 summarizes the population characteristics of the participants from START and CHILD included in our EWAS and replication analyses. Following quality control, pre-processing, and restriction to white European participants in CHILD, a total of 491 START and 295 CHILD samples moved forward.

Table 9. Population characteristics of START and CHILD participants

	START	CHILD
Sample Size	491	295
Ethnicity	South Asian	White European
Maternal Age (years, μ ± SD)	30.9 ± 3.9	32.7 ± 4.4
Infant Sex (% Female)	51.9	46.1
Gestational Age (weeks, μ ± SD)	39.2 ± 1.3	39.5 ± 1.3
Birth Weight (g, μ ± SD)	3265 ± 454.5	3498.6 ± 485.9
GDM (%/N)	13.2/65	5.1/15
Smoking (%/N)	0/0	7.8/23

2.3.4 *Epigenome-Wide Associations*

None of our models investigating the effects of diet, dysglycemia, or maternal smoking identified any sites associated at a genome-wide significant level. However, gestational age was significantly associated with DNA methylation at 1044 and 1560 sites in START and CHILD, respectively. The top 10 sites most significantly associated with gestational age in each cohort are annotated in Table 10.

Table 10. Top 10 sites most significantly associated with gestational age in START and CHILD

Site	P Value	Chr	Location	Gene
START				
cg04347477	6.42E-22	12	125002007	NCOR2
cg17133774	3.89E-21	1	6198667	CHD5
cg11932158	3.58E-19	3	155422129	PLCH1
cg18623216	4.93E-18	3	155421970	PLCH1
cg16103712	7.70E-18	8	99023869	MATN2
cg00220721	8.80E-18	11	36422443	PRR5L
cg12713583	1.01E-17	19	940724	ARID3A
cg08412913	1.21E-17	16	85429522	
cg06870470	1.94E-17	19	11315767	DOCK6
cg03048432	2.72E-17	14	51290751	NIN
CHILD				
cg18623216	1.42E-24	3	155421970	PLCH1

cg16103712	2.42E-24	8	99023869	MATN2
cg11932158	8.52E-24	3	155422129	PLCH1
cg12713583	1.85E-23	19	940724	ARID3A
cg08817867	5.13E-23	17	19656554	
cg20334115	6.10E-21	1	226107899	PYCR2
cg17133774	1.19E-20	1	6198667	CHD5
cg02001279	3.28E-20	19	940967	ARID3A
cg12697139	1.91E-18	1	209571889	
cg23009780	1.72E-17	2	96774492	

2.3.5 Replication in NutriGen Cohorts

We conducted a replication study of our literature findings in the START and CHILD cohorts. The significant sites for each outcome identified through our literature review were used in a targeted replication analysis. GDM and gestational age were tested in both START and CHILD, while smoking and pre-pregnancy BMI replication was confined to CHILD due to data availability. No sites with reported association with GDM or pre-pregnancy BMI were significant in our replication models. Of the 94 sites associated with maternal smoking and available in our CHILD dataset, 8 reached Bonferroni-level significance (Table 11). All 8 of these sites had been previously replicated in the literature according to our review.

Table 11. Maternal smoking sites replicated in CHILD

	P Value	CHR	Location	Gene
cg05549655	3.77E-06	15	75019143	CYP1A1
cg11924019	2.23E-05	15	75019283	CYP1A1
cg22549041	3.55E-05	15	75019251	CYP1A1
cg23067299	4.75E-05	5	323907	AHRR
cg22132788	7.90E-05	7	45002486	MYO1G
cg18092474	0.000122	15	75019302	CYP1A1
cg12803068	0.000278	7	45002919	MYO1G
cg12101586	0.000409	15	75019203	CYP1A1

A total of 279 and 278 sites from the literature and available in CHILD and START, respectively, were targeted for replication of their association with gestational age. Both models generated statistically significant results, with 81 sites replicating in START and 81 in CHILD. Moreover, 54 of these sites were replicated in both cohorts. Each cohort also replicated 7/9 sites previously replicated in our literature review; only one of these sites (cg15626350) failed to replicated in either cohort. Table 12 summarizes the design and results of all six replication models tested.

Table 12. Summary of replication study results for outcomes in START and CHILD

Outcome	Literature Sig. Sites	Available in CHILD	CHILD Sig. Sites	Available in START	START Sig. Sites
GDM	307	216	0	216	0
GestAge	309	278	87	279	81
Smoking	161	94	8	NA	NA
BMI	1	1	0	NA	NA

2.4 Discussion

The recent proliferation of the epigenome-wide association study has generated a wealth of available evidence for the role of DNA methylation in the DOHaD paradigm. Despite many outcomes being examined by multiple studies, it is difficult to elucidate from a basic search what the most pertinent findings are and whether those have been replicated across studies. Our own EWAS analyses failed to generate genome-wide significant results for almost all variables, an issue that many studies in our screening also faced. We thus conducted a literature search, review, and replication study to determine the significant EWAS findings in the field related to diet, maternal dysglycemia, maternal BMI, gestational age, and maternal smoking.

Our search revealed that some of these relationships have been characterized far more successfully than others. We were unable to find an EWAS in cord blood that identified a significant relationship between DNA methylation and maternal diet at the genome-wide level. While diet is a notoriously difficult variable to measure and work with, this deficit combined with our inability to reach genome-wide significance in our own analyses suggests that there may be minimal effects of maternal diet on methylation, or that these effects are so small they cannot yet be identified with current techniques and sample sizes. The body of research for maternal pre-pregnancy BMI was similarly sparse, with a single study identifying one significant site in

ZCCHC10. In contrast, both gestational age and maternal smoking were investigated by multiple studies and showed consistency across studies and the ability to replicate findings.

Gestational age was the only variable that reached genome-wide significance in our EWAS analysis, with > 1000 significant sites identified in both the South Asian START and white European CHILD cohorts. The systematic review also located three additional studies looking at gestational age that were considered in our replication analysis. A total of 10 sites were identified in at least 2/3 of these studies, with one site in *AVP* showing consistent significance in all three papers. This site also reached genome-wide significance in our original EWAS. A list of 279 important sites was compiled from the results of our systematic review and run as a targeted replication in START and CHILD, with 81/279 replicating at Bonferonni significance.

The successful replication combined with the repeated identification of specific sites suggests a strong and regulated relationship between gestational age and DNA methylation. *AVP* codes for arginine vasopressin, which along with oxytocin regulates uterine contractions and impacts the timing of delivery. Other delivery-related genes such as estrogen receptor 1 (*ESR1*) and the corticotropin releasing hormone binding protein (*CRHBP*) were also consistently identified across studies.[28] Gestational age thus may influence the DNA methylation of important pathways involved in childbirth in a consistent and replicable manner.

As with gestational age, there was consistency observed across studies for maternal smoking, even when the outcome was measured differently (ex. Self-reported smoking vs. measured cotinine levels). In particular, the site cg05575921 in *AHRR* was identified in 5/7 papers, and two other sites in this gene were also found in 3/7. This aryl hydrocarbon receptor repressor plays an important role in detoxifying compounds from tobacco smoke through the aryl hydrocarbon receptor (ArH) signaling pathway.[17] Multiple sites in *CYP1A1*, also involved in this pathway, were identified across studies as well. This consistent implication of a crucial pathway in tobacco metabolism suggests that the effects being seen are in fact a direct result of the tobacco exposure. In fact, the same genes have demonstrated differential methylation in cases of adult smoking.

Questions remain despite the consistency of findings. The study by Reese et al. (2017) attempted to generate a methylation 'score' to use as a clinical biomarker of sustained maternal smoking.[26] However, a tool such as this may still be imperfect based on our current understanding. Studies disagree on the dose-response relationship, although most seem to identify sustained smoking as key rather than smoking discontinued early in pregnancy. Preliminary results from an African American cohort suggest that sites may replicate across ethnicities, but more research is needed to ensure any 'smoking signature' develops is applicable outside of a white European population. Finally, there is some evidence for sex-specific effects on the fetus that may need to be addressed.[29] Despite these remaining questions, the consistency observed across many studies on smoking is very encouraging.

Despite finding four studies examining the relationship between gestational diabetes and infant methylation, there was far less consistency observed across studies for this outcome than for gestational age or maternal smoking. No site was identified in more than one paper, nor were the same genes implicated repeatedly. We did find some evidence that broader loci may be involved; three sites across three papers were within a 200KB range on chromosome 16, and there were two pairs of sites identified within 50KB of each other. All the sites observed failed to replicate within our cohort.

Similarly, maternal pre-pregnancy BMI was associated with a significant change in only one site in the single paper published. While this is a comment on the lack of available evidence as much as anything else, it does suggest that these more diffuse exposures (dysglycemia, obesity) may have a more subtle and difficult to elucidate effect on the developing epigenome.

Taken together, these results illuminate some interesting trends in EWAS data. First, it is gratifying to observe that for some outcomes it is indeed possible to identify consistent changes in DNA methylation. It is perhaps unsurprising that this occurred for gestational age and maternal smoking. Age has repeatedly been seen to have a strong relationship with DNA methylation, so much so that 'clocks' have been developed to determine an adult's biological age using their methylation patterns. Two of the papers included in our systematic review were developing a similar clock for predicting gestational age. Given that gestational age is both an 'exposure' a trait intrinsic to the fetus itself, it makes sense that there are consistent observable changes in cord blood methylation based on this outcome. Interestingly, the sites most useful in predicting gestational age are almost entirely different from those used in adult methylation 'clocks', underscoring the importance of studying fetal and infant methylation directly rather than drawing conclusions from adult samples.[18] Maternal smoking, while still a comparatively 'external' exposure, is also a very strong variable to work with as well; tobacco smoke is a known carcinogen with documented negative outcomes in pregnancy.

Changes caused by exposures such as GDM or pre-pregnancy BMI seem to be far more difficult to characterize. It is understandable that the effects on a fetus growing in a dysglycemic, obesogenic, or inflammatory environment may be less extreme and more diffuse than those resulting from exposure to a toxin. The frequently duplicated sites for gestational age and smoking were located in *AVP* and *ARHH*, respectively, both genes intricately involved in mediating their associated exposures. The effects of a dysglycemic uterine environment may not affect a single process this dramatically. There may also be an issue of power in these studies. If the effects are consistent, but small in magnitude and involved in various processes scattered across the genome, a small sample size may be simply unable to detect their presence accurately. The largest GDM study identified here was approximately 300 total cases and controls. In contrast, larger GWAS analyses may use hundreds of thousands of participants to elucidate small effect sizes.

In conclusion, we were able to systematically review the evidence for DNA methylation in DOHaD and replicate some findings in our cohorts. Gestational age and sustained maternal smoking both show consistent alterations to cord blood methylation across multiple studies. Studies of GDM have identified genome-wide significant changes, but these have failed to be duplicated by other groups in independent EWAS analyses. These results demonstrate the promise of DNA methylation in explaining the process of fetal programming, while showing that there is still much to be done to better characterize the relationship with many types of exposures.

CHAPTER 3: Evaluating the role of cellular composition in cord blood GIR

3.1 Introduction

The rising prevalence of non-communicable diseases such as cardiovascular disease (CVD) and diabetes poses a significant challenge in population health. Increasingly, early life experiences, including *in utero* exposures, are being shown to have an important effect on these long-term health outcomes. These observations have led to the concept of fetal programming, in which the environment experienced by a fetus during gestation may affect its development and disease risk in later stages of life. An increasing body of literature around the Developmental Origins of Health and Disease (DOHaD) has demonstrated the importance of understanding these early events to manage health across a lifetime.

3.1.1 Maternal Dysglycemia & Gestational Diabetes

One gestational exposure of significant interest in the DOHaD community is gestational diabetes (GDM), as this appears to have short- and long-term effects on both maternal and child health. GDM is identified based on evidence of glucose intolerance with a first onset during pregnancy.[7] This is generally evaluated based on an oral glucose tolerance test (OGTT) administered between weeks 24 and 28 of gestation.[30]

The prevalence of GDM is variable, but it has been estimated at 10% in the USA, 5.4% in Europe, and 7.24% in a large French study.[7,31,32] Six different criteria applied to the South Asian women in the Born in Bradford cohort generated an estimated prevalence between 4% and 24%. Risk factors include increased maternal age and obesity/BMI.[30,33] This poses a significant health risk to both the pregnant women and their children in the short and long term. Billionnet et al. (2017) studied GDM in a cohort of 716 152 births in France and found an increase in risk of preterm birth, Caesarian section, pre-eclampsia/eclampsia, macrosomia, respiratory distress, birth trauma, and cardiac malformations.[32] Beyond the perinatal period, mothers who experience GDM are more likely to develop metabolic syndrome and/or type 2 diabetes (T2DM) after their pregnancy.[6]

Children of GDM mothers experience fetal programming based on this exposure that leads to lifelong health consequences. In particular, they are at higher risk for glucose dysregulation in their own lives; they experience a higher rate of cardiovascular disease (CVD), hypertension, and T2DM.[7] GDM is therefore a key factor in the developmental origins of health and disease at the individual and population level. A better understanding of the means by which *in utero* exposure programs long-term health factors may lead to more effective interventions in GDM and its associated morbidities.

3.1.2 DOHaD & Epigenetics

One of the most promising mechanisms thought to produce the phenomenon of fetal programming is epigenetic modification, which causes altered gene expression without affecting the underlying genetic code. Epigenetic marks are both modifiable and heritable, allowing for the possibility of an early alteration to persist through cell generations to adulthood. The most characterized epigenetic mechanism is DNA methylation, in which a methyl group is added to the fifth cytosine in a CpG dinucleotide (pair of adjacent cytosine and guanine bases. DNA methylation is crucial for gene regulation; in general, an increase in DNA methylation at a gene promoter leads to decreased expression of that gene, and vice versa.

Numerous studies have demonstrated an association between exposure to a dysglycemic *in utero* environment and altered DNA methylation patterns in placental and infant tissues. Of particular interest is the *IGF2*/*H19* locus, an imprinted region involved in metabolic programming.

Differential methylation and gene expression have been observed at this locus in cord blood samples exposed to intrauterine hyperglycemia.[13] Epigenome-wide association studies (EWASs) have also identified a variety of genes with potentially altered methylation patterns. Cardenas et al. (2018) examined placental tissue and observed differential methylation at 7 sites over four genes, correlating to altered expression in three of these.[34] Similar studies in cord blood have together identified > 100 potential sites of differential methylation in GDM cases.[15,35,36] However, the same sites are not replicated across studies and EWAS results are often different than targeted analyses. It seems likely that DNA methylation patterns are therefore playing an important role in long-term programming of the fetal genome upon exposure to GDM, but more investigation must be done to identify consistent and replicable effects.

3.1.3 Fetal Insulin Sensitivity

In addition to identifying GDM-induced changes in cord blood, it is important to characterize alterations that directly relate to glucose homeostasis in the infant. Measures of fetal insulin sensitivity can help describe an infant's metabolism and may also be associated with changes in DNA methylation. Insulin sensitivity at birth can be assessed in cord blood samples, and is often quantified using circulating concentrations of proinsulin, as well as the glucose-to-insulin ratio (GIR). Alterations in both measures have been observed in cases of maternal dysglycemia.[37]

Cord blood GIR has been associated with numerous gestational exposures and metabolism-related infant metrics. In a study of infants of dysglycemic compared to euglycemic pregnancies, Luo et al. (2010) demonstrated a significantly reduced GIR in association with both high OGTT blood glucose levels and GDM.[37] Gesteiro et al. (2011) found a reduction in GIR in infants born to mothers with impaired glucose tolerance (IGT) compared to their normal maternal glucose tolerance (NGT) counterparts.[38] Sahasrabuddhe et al. (2013) used a composite outcome measure of 'complicated pregnancy', which included cases such as pregnancy induced hypertension (PIH), thyroid dysfunction, and GDM. This outcome was also associated with decreased GIR in cord blood.[39] Preterm birth has also been shown to positively correlate with GIR and other measures of insulin sensitivity.[40]

Other studies have examined the relationship between GIR and metabolism-related measures in cord blood. Luo et al. (2013) found that the concentration of cord blood leptin, but not adiponectin, was associated with the GIR as well as proinsulin level. They suggested that this could be a mechanism by which a predisposition to obesity and insulin resistance could be transmitted from mother to infant.[41] Another study by Zhao et al. (2014) looked at arachidonic acid (AA) and docosahexaenoic acid (DHA) levels in cord blood. These fatty acids are important in maintaining pancreatic beta-cell function and structure and were hypothesized to be involved in fetal insulin sensitivity. Although AA showed no correlation, DHA concentrations were lower in the offspring of GDM mothers compared to non-diabetics and were associated with lower fetal insulin sensitivity (both GIR and proinsulin concentration). The authors suggested that these reduced DHA levels may be involved in the perinatal programming leading to increased type 2 diabetes (T2DM) susceptibility.[42]

3.1.4 Conclusion

In order to better characterize the relationship between maternal dysglycemia, fetal insulin sensitivity, and long-term programming, we propose to examine the role of DNA methylation. The primary goal is to determine the association between genome-wide DNA methylation and the GIR in infant cord blood. Changes in methylation found to be associated with GIR may provide clues as to what biological processes are being affected by maternal dysglycemia and how these translate into lifelong predispositions.

3.2 Methods

3.2.1 Study Population & Outcome Measures

Data were generated as part of the NutriGen study, which aims to understand the impact of maternal nutrition and exposures in long-term infant and child health. Of particular interest is the role of genetics and epigenetics in these relationships. This study combines four diverse Canadian birth cohorts: the Family Atherosclerosis Monitoring In Early life (FAMILY) cohort, the Canadian Health Infant Longitudinal Development (CHILD), the South Asian birth Cohort (START), and the Aboriginal Birth Cohort (ABC). Measures of maternal and infant health were collected and harmonized across cohorts for analysis. A subset of approximately 500 samples each from START and CHILD were also assessed for genome-wide DNA methylation.

Additionally, cord blood samples in START were assessed for levels of glucose and insulin. A composite variable was created using the ratio of measured glucose:insulin. Because this yielded small values, they were multiplied by ten for use in models. The final variable, (glucose:insulin) x 10, is our GIR.

Follow-up analysis utilized a subset of samples from the Steroids In caRdiac Surgery Trial (SIRS Trial). This is a multicentre, international, randomized controlled trial to investigate the effect of perioperative steroid administration on death and MI in patients undergoing cardiac surgery requiring cardiopulmonary bypass. Peripheral blood from adult patients was collected and a subset were assessed for genome-wide DNA methylation. We accessed these data following standard preprocessing and quality control measures, leaving a remaining sample of 466 individuals.

3.2.2 Genome-Wide DNA Methylation

Cord blood samples were processed and DNA was extracted using standard protocols. In preparation for the methylation assay, samples underwent sodium bisulfite treatment. Briefly, this process converts unmethylated cytosine bases to uracil, leaving methylated cytosines unchanged and allowing for the relative quantification of DNA methylation at a given site.

Samples were assayed with the Illumina HumanMethylation450K BeadChip, which measures DNA methylation at > 485 000 sites genome-wide covering > 96% of RefSeq genes. A total of 512 START and 511 CHILD samples were randomized across arrays for methylation assessment. Intensities were read with the Illumina iScan and idat files exported to R for pre-processing and quality control.

3.2.3 Data Processing and Quality Control

Raw iScan data were imported into the R version 3.2.0 which was used for all pre-processing, quality control, and downstream statistical analysis. First, idat files were loaded into the R/Bioconductor *minfi* package. Any samples showing > 1% of probes failing detection were removed. Genetic sex was estimated using the getSex function and compared to phenotypically reported sex; any mismatched samples were also removed. Next, probes were assessed for missingness, and those showing failure in > 1% of samples were excluded. Probes known to be cross-reactive or to contain SNPs were removed as well. A total of 506 START and 491 CHILD samples remained following the application of all quality control measures, with a final 393 400 and 393 449 sites, respectively (Table 13). Data were normalized with subset within-array normalization (SWAN) and exported from *minfi* as beta (β) values.[24] Batch effects were corrected for by chip using the ComBat algorithm in the *sva* package.[25] Because data in the START cohort were used in an EWAS analysis, we also created a regional START dataset in *minfi* with the cpgCollapse function, generating regions no wider than 1500 bp with all sites < 500 bp apart. Analysis in CHILD samples was restricted to white European individuals to ensure a large enough sample to avoid confounding.

Underlying cellular composition was estimated in processed START, CHILD, and SIRS samples using several methods. The reference-free ReFACTor algorithm was used to generate the first 7 principal components (PCs) representing cellular composition. The estimateCellCounts function in the *minfi* package also incorporates the Houseman et al. (2012) and Bakulski et al. (2016) methods for estimating proportions of relative cell types in peripheral blood and cord blood, respectively.[20,21] Cell counts were estimated based on peripheral blood algorithms in SIRS and cord blood in START and CHILD.

Table 13. Summary of samples and probes excluded at each stage of quality control.

Samples Probes

	START	**CHILD**		**START**	**CHILD**
Initial	512	511	**Initial**	> 485 000	
Sex Check	5	7	**Failed**	756	634
Missingness	2	14	**Polymorphic**	70 889	
Final	506	491	**Cross-Reactive**	29 233	
			Final	393 400	393 449

3.2.4 Epigenome-Wide Association

An EWAS was conducted in START to determine the relationship between DNA methylation patterns and GIR in the infant cord blood. Statistical analysis was performed using R version 3.2.0. A multivariable linear regression model was used to test for association and was adjusted for maternal age, infant sex, gestational age, study centre, and the top 7 principal components of cellular composition as determined by the ReFACTor algorithm.[19] Statistical significance was set using the Bonferroni correction for multiple testing at $p < 1.27 \times 10^{-7}$ and $p < 2.76 \times 10^{-7}$ for the site-by-site and regional datasets, respectively.

Part 2: Cellular Composition

3.2.5 Verifying Associations

Several methods were used to validate the statistically significant findings. First, the multivariable linear regression model was repeated with the GIR variable replaced with a permutation for ten iterations. The GIR was then winsorized, log transformed, and quantile normalized, and the regression analysis redone using each version of the variable. For best results the quantile normalized GIR variable (QGIR) was used in all downstream analyses.

3.2.6 Identifying Independent Signals

A forward stepwise regression model was used to test the independence of the top signals in the QGIR association. Based on Bonferroni correction the threshold for significance was set at $p < 1.27 \times 10^{-7}$. In the first iteration, the most significant site from our original association analysis was incorporated into the regression model as a covariate. The most significant site from these results was then added to the model, and so on until no significant associations remained. These and all subsequent models included the first 5 principal components of ancestry (derived from START genotyping data) as covariates in addition to those previously listed.

The top 5 sites from the QGIR analysis were also assessed for correlation with each other. Correlation between each possible pairing of the top 5 sites was calculated.

3.2.7 Pathway Analysis

To investigate underlying biological relationships that may be driving the observed associations, we conducted pathway analysis. This method attempts to characterize biological pathway that are statistically overrepresented among the more significant sites identified. Slight modifications were necessary to accommodate methylation data as the tools are primarily designed for genotyping results. First the InCroMap tool was applied, using both the Kyoto Encyclopedia of Genes and Genomes (KEGG) and Gene Ontology (GO) annotations. The Data-driven Expression Prioritized Integration for Complex Traits (DEPICT) method was also used, which combines numerous annotation sets.

3.2.8 Methylation Risk Score

Next, a risk score was constructed for both START, CHILD, and SIRS samples using the top associations from the regression analysis. First, the site-by-site and regional regressions for QGIR were repeated excluding all cellular composition estimates as covariates to identify the top 10 most significant regions. The lead site within each region was defined as the one reaching the strongest significance in the site-by-site model. The risk score was calculated as the sum of the product of the beta value and effect size at each of these 10 sites over every sample.

To test which cell types may be involved in the observed relationship, this risk score was used in a subsequent linear regression model with the estimated underlying cell counts as the predicted variable. This association was conducted in START and CHILD (cord blood) as well as the SIRS methylation samples (adult peripheral blood). In CHILD, 184 samples underwent a complete blood count (CBC), and our methylation risk score was therefore also tested for association in this subset with directly quantified cellular proportions.

3.2.9 Blueprint Epigenome Project

To further investigate the role of individual cell types, data were obtained from Blueprint Epigenome. This project aims to generate reference epigenomes for healthy and malignant haematopoieitic cells. Their data included 64 venous blood and 46 cord blood samples. We accessed genome-wide DNA methylation data for all available cell types for both cord and venous blood. In cord blood, two each of the following were used: neutrophils, monocytes, dendritic cells, B cells, and CD8T cells, as well as one macrophage. For venous blood we accessed two each of: monocytes, macrophages, dendritic cells, neutrophils, NK cells, B cells, CD8T cells, and CD4T cells. All methylation data were imported in the form of BigWig files and processed using the WiggleTools package. Beta values for the ten sites used in the methylation risk score were exported from each sample. and assessed across cell types to determine which cell(s) showed an inverse methylation pattern compared to the others.

3.3 Results

Part 1: GIR and Methylation Association

3.3.1 Study Population

Table 14 summarizes the population characteristics observed in the mothers and infants from START and the white Europeans subset of CHILD.

	START	CHILD (European)
Sample Size	491	295
Maternal Age (years, µ ± SD)	30.9 ± 3.9	32.7 ± 4.4
Infant Sex (% Female)	51.9	46.1
Gestational Age (weeks, µ ± SD)	39.2 ± 1.3	39.5 ± 1.3
Birth Weight (g, µ ± SD)	3265 ± 454.5	3498.6 ± 485.9
GDM (%)	13.2	5.1
Smoking (%)	NA	7.8

Table 14. Population characteristics in START and CHILD methylation samples.

3.3.2 Epigenome-Wide Association

Our initial EWAS, conducted using our untransformed GIR variable, led to genome-wide significant findings in both the site-by-site and regional datasets. A total of 13 significant sites and 8 regions were identified. Table 15 shows the significant sites and their associated genes. Overall, both site and regional models had 4 significant genes in common: *DNAJB6*, *CXXC5*, *AP3D1*, and *ABI3*.

Site	Estimate	P Value	Chr.	Position	Gene	Location
cg06858263	-0.00099	3.54E-09	7	157148509	DNAJB6	5'UTR
cg13707793	-0.00106	2.51E-08	5	139045301	CXXC5	5'UTR
cg16214653	-0.00176	2.82E-08	15	100048500		

cg08087047	-0.00101	4.23E-08	17	72461209	CD300A	TSS1500
cg09159050	-0.00105	4.51E-08	1	1563500	MIB2	Body
cg12811871	-0.00184	5.09E-08	4	2322078	ZFYVE28	Body
cg25892537	-0.00091	5.35E-08	4	6611128	MAN2B2	Body
cg05875239	-0.0021	5.92E-08	14	99787559		
cg11133963	-0.00123	6.68E-08	17	47297512	ABI3	Body
cg05656688	-0.00145	9.59E-08	1	25254088	RUNX3	Body
cg04392554	-0.00174	1.00E-07	22	46685472	TTC38	Body
cg01830256	-0.00119	1.13E-07	14	105861940	PACS2	3'UTR
cg11349093	-0.00109	1.19E-07	19	2112885	AP3D1	Body

Table 15. Sites showing genome-wide significant association with untransformed GIR.

Part 2: Cellular Composition

3.3.3 Verifying Association

In ten permutations no statistical significance or evidence of association was observed. Transforming the GIR variable increased the number of significant sites, and analysis moving forward used the QGIR. Compared to the original 13 significant sites, a total of 190 reached significance in the QGIR regression model (including 12/13 previously associated sites). The top 15 most significant associations are detailed in Table 16. QGIR was also shown to be significantly associated with > 60 regions. These results suggest a true association with our outcome rather than a spurious finding due to underlying data distribution or other issues.

Site	Estimate	P Value	Chr	Location	Gene
cg13917614	-0.02434	7.97E-13	17	40125660	CNP
cg03538296	-0.01669	3.85E-12	1	15392433	KIAA1026
cg12811871	-0.01925	5.01E-12	4	2322078	ZFYVE28
cg15220605	-0.01944	9.16E-12	17	80393595	HEXDC
cg18247172	-0.01658	1.17E-11	15	91370233	
cg00910503	-0.01595	1.33E-11	17	80393666	HEXDC
cg01291375	-0.01318	1.49E-11	6	170753313	
cg11349093	-0.01142	1.84E-11	19	2112885	AP3D1
cg06706159	-0.02199	2.95E-11	19	18260350	MAST3
cg26355072	-0.0154	2.96E-11	5	141674679	
cg24137511	-0.0173	4.34E-11	19	18260330	MAST3
cg15636859	-0.014	6.31E-11	20	55982844	RBM38
cg03831847	-0.01663	6.80E-11	16	88832485	FAM38A
cg09115713	-0.01752	1.09E-10	16	88832476	FAM38A
cg09159050	-0.01064	1.09E-10	1	1563500	MIB2

Table 16. Top 15 sites most significantly associated with QGIR.

3.3.4 Identifying Independent Signals

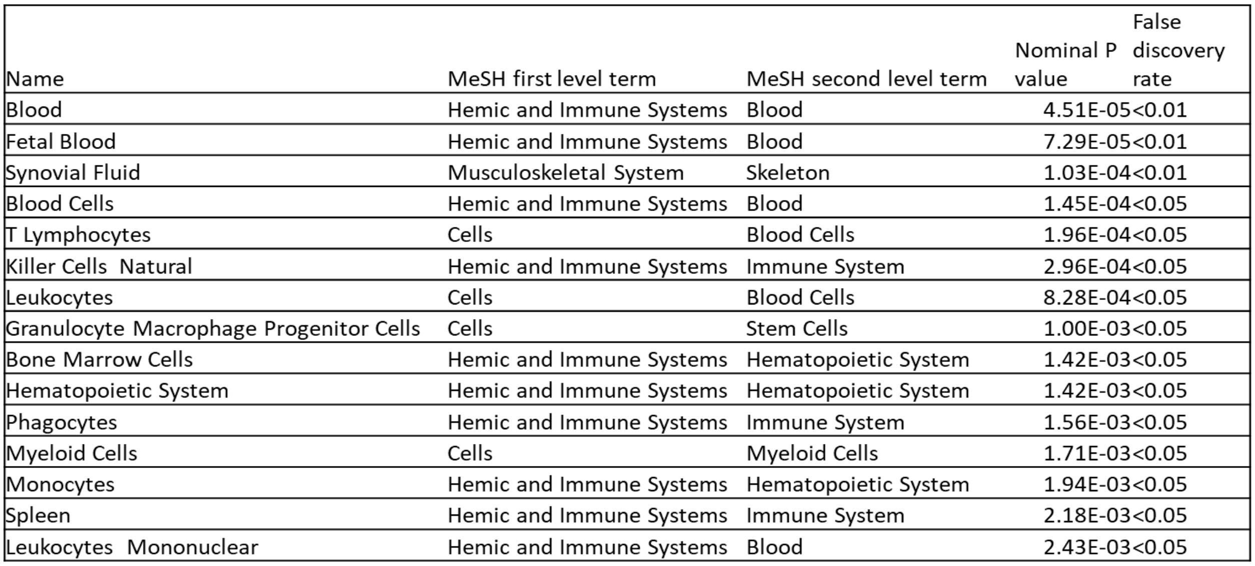
In both the site-by-site and regional datasets, significance was eliminated after just one iteration of the stepwise regression model; the incorporation of only the top DMP/DMR was enough to remove the observed association. This suggests that rather than being a set of independent signals, the significant sites/regions are in fact related in some way despite falling on a variety of chromosomes across the genome. Beta values at the five top sites were also seen to correlate strongly to each other (Table 17). There is no obvious structural explanation underlying these findings, which suggests instead that the observed associations may be driven by a common biological relationship among the genes identified.

Site	cg03538296	cg13917614	cg12811871	cg15220605	cg18247172
cg03538296	1.00	0.79	0.82	0.78	0.7
cg13917614	0.79	1.00	0.9	0.85	0.86
cg12811871	0.82	0.9	1.00	0.85	0.83
cg15220605	0.78	0.85	0.85	1.00	0.82
cg18247172	0.7	0.86	0.83	0.82	1.00

Table 17. Pearson correlation coefficients (r) for the 5 most significantly associated sites.

3.3.5 Pathway Analysis

Pathway analysis allowed us to characterize what biological processes may be involved in our results. The InCroMap tool generated non-specific pathways that were inconsistent with those identified by DEPICT. In contrast, DEPICT repeatedly implicated pathways involved in ‘Hemic and Immune Systems’ (Table 18), which suggests a role for processes such as haematopoiesis. This further implies that underlying cellular heterogeneity not accounted for by the ReFACTor PCs may be driving the association. Cellular composition was therefore derived using the Bakulski et al. (2016) method, generating estimates for each cell type and allowing for further investigation into which hemic cell(s) may be involved in the observed relationship.[21]

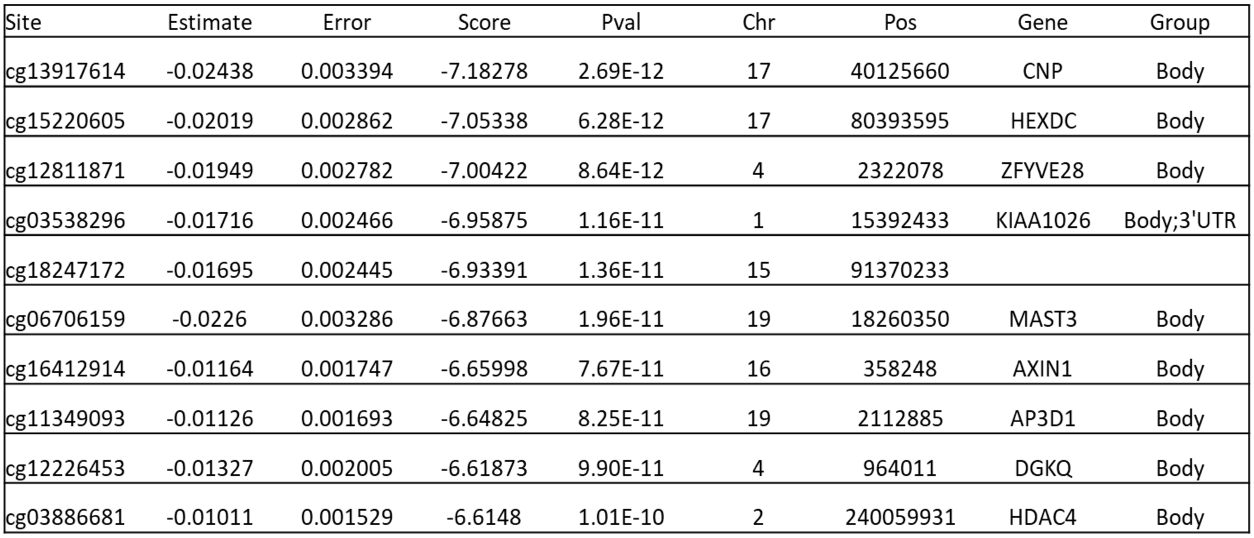


Name	MeSH first level term	MeSH second level term	Nominal P value	False discovery rate
Blood	Hemic and Immune Systems	Blood	4.51E-05	<0.01
Fetal Blood	Hemic and Immune Systems	Blood	7.29E-05	<0.01
Synovial Fluid	Musculoskeletal System	Skeleton	1.03E-04	<0.01
Blood Cells	Hemic and Immune Systems	Blood	1.45E-04	<0.05
T Lymphocytes	Cells	Blood Cells	1.96E-04	<0.05
Killer Cells Natural	Hemic and Immune Systems	Immune System	2.96E-04	<0.05
Leukocytes	Cells	Blood Cells	8.28E-04	<0.05
Granulocyte Macrophage Progenitor Cells	Cells	Stem Cells	1.00E-03	<0.05
Bone Marrow Cells	Hemic and Immune Systems	Hematopoietic System	1.42E-03	<0.05
Hematopoietic System	Hemic and Immune Systems	Hematopoietic System	1.42E-03	<0.05
Phagocytes	Hemic and Immune Systems	Immune System	1.56E-03	<0.05
Myeloid Cells	Cells	Myeloid Cells	1.71E-03	<0.05
Monocytes	Hemic and Immune Systems	Hematopoietic System	1.94E-03	<0.05
Spleen	Hemic and Immune Systems	Immune System	2.18E-03	<0.05
Leukocytes Mononuclear	Hemic and Immune Systems	Blood	2.43E-03	<0.05

Table 18. Most significantly enriched pathways identified by DEPICT.

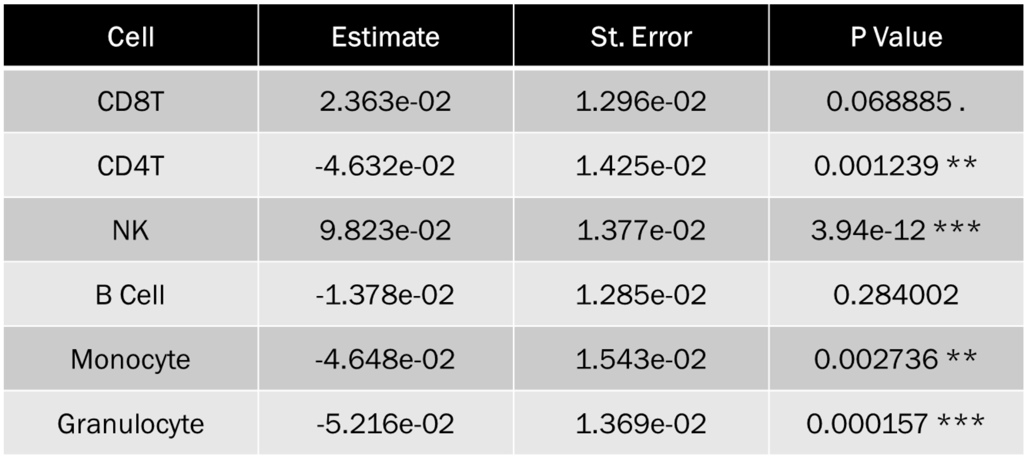
3.3.6 Methylation Risk Score

The methylation risk score was constructed based on the sites in Table 19 and assessed for its association with QGIR in START as well as measures of cellular composition in CHILD, START, and SIRS. Results were significant in all these models ($p < 0.05$). The START and CHILD results suggest that lymphocytes are playing an important role in modulating the risk score, but it is unclear which cell population(s) may be responsible. The SIRS data, deconvoluted using a more robust reference panel, show clearer effects the directionality of the relationships (Table 20). This indicates a potential role for natural killer (NK) cells in driving the observed association.



Site	Estimate	Error	Score	Pval	Chr	Pos	Gene	Group
cg13917614	-0.02438	0.003394	-7.18278	2.69E-12	17	40125660	CNP	Body
cg15220605	-0.02019	0.002862	-7.05338	6.28E-12	17	80393595	HEXDC	Body
cg12811871	-0.01949	0.002782	-7.00422	8.64E-12	4	2322078	ZFYVE28	Body
cg03538296	-0.01716	0.002466	-6.95875	1.16E-11	1	15392433	KIAA1026	Body;3'UTR
cg18247172	-0.01695	0.002445	-6.93391	1.36E-11	15	91370233		
cg06706159	-0.0226	0.003286	-6.87663	1.96E-11	19	18260350	MAST3	Body
cg16412914	-0.01164	0.001747	-6.65998	7.67E-11	16	358248	AXIN1	Body
cg11349093	-0.01126	0.001693	-6.64825	8.25E-11	19	2112885	AP3D1	Body
cg12226453	-0.01327	0.002005	-6.61873	9.90E-11	4	964011	DGKQ	Body
cg03886681	-0.01011	0.001529	-6.6148	1.01E-10	2	240059931	HDAC4	Body

Table 19. Lead sites and estimates for the 10 regions most significantly associated with QGIR and used to generate methylation risk scores in START, CHILD, and SIRS.



Cell	Estimate	St. Error	P Value
CD8T	2.363e-02	1.296e-02	0.068885 .
CD4T	-4.632e-02	1.425e-02	0.001239 **
NK	9.823e-02	1.377e-02	3.94e-12 ***
B Cell	-1.378e-02	1.285e-02	0.284002
Monocyte	-4.648e-02	1.543e-02	0.002736 **
Granulocyte	-5.216e-02	1.369e-02	0.000157 ***

Table 20. Association between methylation risk score and cell composition in SIRS

3.3.7 Blueprint Epigenome

DNA methylation in the individual cell lines accessed from the Blueprint Epigenome Project support the hypothesis that natural killer cells are driving our results. Although the small number of samples prohibited statistical analysis, all sites consistently showed inverted methylation patterns in NK cells as compared to other blood cells in both cord and venous blood samples (Table 21).

Site	Estimate	Neutr	Mono	DC	Macro	B Cell	CD8T	NK
cg13917614	-0.0244	0.962	0.962	0.855	0.932	1	0.949	0
cg15220605	-0.0202	0.948	1	0.839	0.978	1	1	0.533
cg12811871	-0.0195	0.949	0.912	0.806	0.967	0.94	1	0.053
cg03538296	-0.0172	0.958	0.955	0.773	0.946	0.955	0.944	0
cg18247172	-0.0170	0.97	0.965	0.88	0.867	1	0.833	0.147
cg06706159	-0.0226	0.91	0.948	0.906	0.907	0.96	0.929	0.034
cg16412914	-0.0116	0.933	0.959	0.857	0.912	1	0.968	0.174
cg11349093	-0.0113	0.947	0.986	0.863	0.969	0.949	1	0.136
cg12226453	-0.0133	0.866	0.971	0.77	0.954	0.981	0.739	0.074
cg03886681	-0.0101	1	0.958	0.9	0.923	0.867	1	0.075

Table 21. Methylation at risk score sites in individual cell lines

Most compelling is the observation that all 10 sites used to generate the risk score demonstrated a negative association with the QGIR outcome, but trend toward hypermethylation (approaching $\beta = 1.0$) in all non-NK blood cells. NK cells, however, are the only ones showing hypomethylation (approaching $\beta = 0$), strongly suggesting that they are responsible for the observed direction of effect. This inverse pattern is also consistent at the two positively associated sites assessed. Overall, these findings provide convincing evidence that NK cells are playing a key role in the relationship between cord blood QGIR and DNA methylation.

3.4 Discussion

In this study we sought to characterize the relationship between cord blood GIR and DNA methylation. We found that GIR is significantly associated with 190 CpG sites and > 60 regions, but that despite their distribution across the genome these sites are all correlated with each other rather than representing independent signals. The identification of ‘Hemic and Immune System’ pathways with the DEPICT pathway analysis tool was the first suggestion that this relationship may be driven by underlying cellular composition. By utilizing a methylation risk score and data from the Blueprint Epigenome Project, we were able to identify natural killer (NK) cells as the population most strongly influencing our results.

3.4.1 Natural Killer (NK) Cells

Natural killer (NK) cells were first described in the 1970s and have since become characterized as a key component of the innate immune system.[43] Approximately 15% of circulating lymphocytes are NK cells, and most knowledge of NK cells comes from the study of peripheral blood.[44] Originally two subsets of NK cells were identified: CD56^{dim} (which have potent cytotoxic activity and are the dominant subset in peripheral blood and lung tissue) and CD56^{bright} (cytokine-producing but with poor cytotoxic ability) cells.[43] However, this categorization vastly underestimates the diversity amongst NK cells, which is in turn related to their function(s). A study of monozygotic twins and unrelated donors using mass spectrometry estimated 6 000 – 30 000 phenotypic populations of NK cells in an individual.[45] Cytomegalovirus exposure can also cause clonal-like expansion of NK cells, significantly modifying their DNA methylation patterns and yielding cells more similar to toxic CD8^{+} T cells than to typical NK cells.[46] The variety of NK cell phenotypes beyond the initial subset definitions is beginning to become apparent.

NK cells are best known for their role in non-self response, in particular for killing malignant cells of hematopoietic origin as well as virus-infected.[43,44,46,47] They also produce and respond to many cytokines and chemokines, and interact with a both immune and non-immune cells.[46] In response to cytokine production by nearby cells, NK cells release further cytokines and chemokines and thereby potentiate their responsiveness to cellular targets.[46] They are also regulated through a complex system of activating and inhibitory receptors that sense ligands on surrounding cells. Inhibitory receptors often target HLA class 1 ligands, enabling NK cells to target cells without self-HLA class 1 expression in a process known as ‘missing self’ recognition.[46]

Horowitz et al. (2013) noted that inhibitory receptor expression is driven largely by genetics, while environmental factors are most influential on activation receptor expression.[45] Recent findings have also identified infection-induced adaptations in NK cells, suggesting that they may have durable, memory-like responses in addition to their role in innate immunity.[43]

A successful pregnancy is also dependent on uterine NK (uNK) cells. In early gestation, interaction between NK cells and dendritic cells (DCs) is crucial in developing maternal tolerance and beginning angiogenesis.[48] uNK cells produce the IL-10 necessary for DC interaction which allows for enhanced angiogenesis and placental development; uterine artery remodeling also requires proper NK cell function.[49]

3.4.2 NK Cells & Obesity

Obesity has been demonstrated to have a significant effect on NK cell phenotype and function. O’Rourke et al. (2013) identified an increased activation profile in NK cells derived from adipose tissue compared to peripheral blood.[50] They later suggested that NK cells may be able to regulate adipose tissue macrophages and influence insulin resistance. Other studies have supported both these claims. Viel et al. (2017) compared peripheral blood NK cells between obese and non-obese individuals and identified an activated phenotype associated with obesity and characterized by elevated CD69 and granzyme B along with decreased CD16. They also noted a trend toward increased numbers of NK cells with increasing BMI.[51] Bahr et al. (2018) observed no significant changes in total NK cells between normal-weight and obese individuals but did identify changes in NK cell subsets. They found that obese subjects showed increased numbers of CD56^{bright} (low cytotoxicity) NK cells and a decrease in CD56^{dim} (high cytotoxicity) cells.[52]

This dysregulation of NK cells may have a direct effect on key obesity-related health outcomes. NK cell uptake of lipids from the environment may interfere with mTOR-PPAR pathways and cause metabolic paralysis, thereby interfering with the cytotoxic process and efficient tumour killing.[53] This may be one mechanism by which the incidence of certain cancers increases in obese populations. The previously noted tendency of NK cells in adipocytes to recruit macrophages may also lead to excess IL-1β production by these macrophages and resulting insulin resistance and T2DM.[54] Obesity may also interfere with proper uNK cell function. Perdu et al. (2016) identified a significant reduction of uNK cell numbers in obese compared to lean women as well as impaired uterine artery remodeling. uNK cells in obese women also overexpressed decorin, which limited trophoblast survival and inhibited placental development.[49]

3.4.3 NK Cells & Atherosclerosis

NK cells may also be involved in another disease of chronic inflammation: atherosclerosis. They have been observed within atherosclerotic plaques in humans, and may in fact be recruited to the site by chemokines known to be present in these lesions. Monocyte chemoattractant protein-1 (MCP-1) and fractalkine (CX3CL1) are both found in atherosclerotic lesions and are known to be a chemoattractant to NK cells and to induce NK cell migration and activation, respectively.[55] Within plaques they are mostly found within tissues adjacent to the necrotic core as well as in shoulder regions.[47] Higher circulating NK cell levels have also observed in patients with severe atherosclerotic disease.[55] In particular, increased activation of NK cells by various methods may increase atherosclerosis and high-risk plaque development.[56] However, most evidence in humans is limited to observation and association; murine models have attempted to elucidate a causal role for NK cells in atherosclerosis.

In beige mice (which carry a mutated Lyst gene causing a loss of cytolytic NK cell function) with an LDLR$^{-/-}$ background, a potential atheroprotective effect of NK cells was observed that may be cytokine-mediated.[57] In contrast, LDLR$^{-/-}$ mice with Ly49A transgenic bone marrow modeling NK cell deficiency showed reduced lesions, suggesting that NK cells may be pro-atherogenic.[58] Selathurai et al. (2014) used a combination of gain- and loss-of-function models in ApoE$^{-/-}$ mice on a high-fat diet and determined that NK cells are indeed atherogenic and that they contribute to necrotic core expansion through perforin and granzyme B production.[59] However, all of these studies suffered from being unable to attribute the results exclusively to NK cells, as their models also affect other cytotoxic lymphocytes. More recently, Nour-Eldine et al. (2018) utilized several novel mouse models and demonstrated no direct effect of NK cells on atherosclerosis but suggested that they may play a role in cases of systemic NK-cell overactivation.[60] In particular, they provided evidence that the proatherogenic effects of anti-asialoganglioside M1 antiserum treatment noted by studies such as that by Selathurai et al. (2014) is due to effects on CD8^{+} T and NKT cells.

Overall, a role for NK cells in atherosclerosis seems promising, but the evidence is as yet inconclusive and in most cases indirect. It has also been suggested that crosstalk between NK and DC cells may exacerbate atherosclerosis, perhaps through IFN-γ release and inflammatory incitement.[47] Much remains to be elucidated regarding NK cells in this disease model and their interaction with other players such as macrophages and DCs.

3.4.4 Conclusions

Our findings, in combination with numerous other observations of disturbed NK cell levels in inflammatory disease, suggest that perhaps the long-term effects of maternal dysglycemia and insulin sensitivity are modulated at the tissue level by moderating cellular composition. This counters the general assumption that epigenetic mechanisms such as DNA methylation are the primary mediator in the process of fetal programming. Instead, methylation differences may be the result of changes to the underlying cell populations, as observed here with NK cells.

A major limitation of this study is our inability to directly quantify NK cells. Almost all of our cellular composition estimates were derived from methylation data rather than being directly assessed in blood. While lymphocytes can be reported as part of a CBC, this is only a proportional measure relative to other cell types and does not provide an independent NK variable. It may be possible to use an ELISA-based assay to target NK-specific proteins (such as *CD335* or *S1PR5*) in order to develop a method of measuring NK cells from venous blood samples. This would allow for quantification in large sample sizes to generate direct association analyses with inflammatory and CVD-related outcomes.

Future investigations will also focus on the genetics underlying cellular composition. The deconvolution methods developed for DNA methylation arrays can be used in publicly available datasets to derive cellular composition. Where genotyping data are also available, genome-wide association analyses may be conducted in order to identify what genes and processes are most implicated in determining the proportion of various blood cell types. This may provide greater insight into the mechanisms by which cellular composition may be modified in gestational exposure paradigms.

CHAPTER 4: DISCUSSION

4.1 Study Overview

In these studies, we attempted to achieve a better understanding of the role of DNA methylation in the mediation of fetal programming. By conducting a systematic review and replication study, we were able to gather the available EWAS evidence related to numerous outcomes into a cohesive whole. This demonstrated that some exposures are well characterized; for example, differential methylation associated with maternal smoking and gestational age has been reported in multiple papers with notable consistency across studies regarding the significant loci identified. Other relationships have sparser evidence available. While GDM was the subject of several papers, sample sizes were generally small and results failed to replicate across studies. Literature regarding DNA methylation changes associated with maternal pre-pregnancy BMI and prenatal diet is negligible. Overall, despite promising results in several exposures, significant gaps in the field of epigenetics and DOHaD remain and what evidence is available is poorly organized and often inconsistent.

Following our review of the literature, we were able to utilize our NutriGen dataset in a replication study. We replicated many loci associated with both gestational age and smoking; however, the sites reportedly related to GDM exposure and the single locus associated with pre-pregnancy BMI failed to replicate in our models. These findings are in line with the results of our literature review and the ability of other cohorts to identify significance.

To better characterize the relationship between dysglycemia and DNA methylation we then conducted an epigenome-wide association analysis using fetal insulin sensitivity (measured as

GIR) as our outcome. Contrary to our previous models, clear genome-wide significance was achieved at numerous CpG sites, but these signals were all found to associate with one another through their involvement with hematological pathways. We were able to identify changes in underlying cellular composition within blood samples as the origin of the observed differences in methylation, and determined that NK cells are likely driving these associations. These changes in NK cell count could be responsible for many of the long-term inflammatory effects of a dysglycemic environment. While some exposures, such as maternal smoking, clearly cause direct changes in DNA methylation patterns, the results of this investigation suggest that we must also consider the possibility that fetal programming may be mediated at the tissue level by altering the relative levels of different cell types. This is particularly relevant to studies using heterogeneous tissues such as blood in their investigations and certainly merits further research to elucidate its role in DOHaD. It also suggests that earlier EWAS studies that did not account for cellular heterogeneity may require re-analysis and interpretation, and we should be mindful of the potential for false positive results generated by underlying cellular proportions.

4.2 Tissue Modification & Neutrophil-to-Lymphocyte Ratio (NLR)

The results of our study of insulin sensitivity and DNA methylation in cord blood suggest that the predicted role of epigenetic modification in DOHaD models may need to be reconsidered. Prenatal exposures are generally hypothesized to cause persistent alterations in offspring DNA methylation, which would then cause downstream changes in gene expression and influence long-term phenotype development. While this model may hold true for some exposures (the consistency of results relating to maternal smoking, for example, suggests this may be the mechanism), our findings lead us to believe that it may be an incorrect or incomplete description of fetal programming in other settings. Instead, alterations may occur at the tissue level, which then leads to observed changes in DNA methylation. It is important to note that tissue-level changes could be occurring in tissues other than blood, which would be difficult to identify using a blood-based assays. In the case of insulin sensitivity, we inferred altered NK cell counts and therefore overall changes in the cellular composition of the cord blood samples. Modifications such as this could have significant effect on immune function and inflammation, both key processes in the development of NCDs.

Of particular relevance in the case of dysglycemia and insulin sensitivity is measurement of the neutrophil-to-lymphocyte ratio (NLR). The NLR is constructed from the complete blood count (CBC), a routine clinical measure, and appears to have independent utility in assessing prognosis of NCDs including cancer.[61] NLR is an important indicator of subclinical inflammation in other chronic conditions as well; it can aid in risk stratification of patients with coronary artery disease (CAD) and may be a predictive marker of insulin resistance in T2DM.[62,63] NLR is significantly increased in both diabetic and prediabetic populations and may have a strong predictive role in dysglycemia.[64]

Increased NLR is associated with elevated HbA1c in individuals with T2DM.[65] It may also predict complications, as NLR shows association with numerous measures of diabetic comorbidities, including carotid artery intima-media thickness and diabetic peripheral neuropathy.[66,67] While NLR does not appear to be a predictive marker of GDM, elevation indicating subclinical inflammation may still be an important measure in pregnancy.[68]

By inducing even small changes in the proportions of blood cell populations, *in utero* exposures could cause magnified alterations to the ratio of these cell types and dramatically influence markers such as the NLR. Our studies of insulin sensitivity indicate that GDM may modify the quantity of NK cells in cord blood; this suggests that perhaps changes to NK cell populations, reflected in the denominator (lymphocytes) of the NLR may be driving some of its predictive power in dysglycemic environments. While few studies have examined alterations in cellular composition, Hadartis et al. (2016) did examine cord blood hematopoietic stem and progenitor cell (HSPC) populations in GDM vs normoglycemic pregnancies and identified increased proportions of $CD34^{+}CD45^{dim}$ cells. They suggested that this may be the result of increased fetal stem cell mobilization in cases of GDM but cautioned that their conclusions were limited by a small (N = 87) sample size.[69] In combination with our findings, it seems probable that GDM causes changes to underlying cellular composition in infants that may be related to their long-term health.

4.3 Conclusions

Systematic review of EWAS literature in the DOHaD field revealed an area of research with a great deal of promise but little current organization or consistency. Unlike GWASs, which use well-established terminology and are catalogued in various databases, EWASs related to DOHaD can prove difficult to find and comparing between studies may be impeded by methodological differences. Despite this, several prenatal exposures are linked to specific and replicable changes in cord blood DNA methylation, providing compelling evidence for the role of epigenetics in fetal programming.

In cases of dysglycemia, our results provide novel evidence that even when differential DNA methylation is observed, this may be the result of tissue-level modifications rather than genetic regulation. Preliminary results suggest that NK cells in particular may have a key influence on insulin sensitivity and the predictive power of the NLR. However, further study must be undertaken to better characterize these relationships. Study of individual cell populations through cytometry, rather than composite measures of overall blood methylation, may help to elucidate the way NK cells effect change in immunological processes upon dysglycemic exposure.

The ability to use genome-wide methylation data and deconvolution algorithms to accurately estimate cellular proportions also provides exciting methodology to better understand the genetic basis of blood composition. Large sample sizes that may be too cumbersome for cytometric techniques can now have their relative cell proportions quantified using methylation arrays. In study populations that measure both genome-wide genotype and DNA methylation, a GWAS could be conducted to search for associations between an individual's genetics and their cellular composition. This could be an invaluable tool for understanding the genetic basis of measures like the NLR and their role in long-term health. Methylation-derived NLR (mdNLR) has already been calculated by several studies and investigated for association with rheumatoid arthritis and several cancers.[70–73] Exciting new techniques such as this will undoubtedly only increase our understanding of the complicated but fascinating process of fetal programming and its relationship with epigenetics.

5.0 REFERENCES

1. Hales, C. N. & Barker, D. J. P. Type 2 (non-insulin-dependent) diabetes mellitus: the thrifty phenotype hypothesis. *Int. J. Epidemiol.* **42**, 1215–1222 (2013).
2. Yajnik, C. S. Commentary: Thrifty phenotype: 20 years later. *Int. J. Epidemiol.* **42**, 1227–1229 (2013).
3. Schulz, L. C. The Dutch hunger winter and the developmental origins of health and disease. *Proc. Natl. Acad. Sci. U. S. A.* **107**, 16757–16758 (2010).
4. Roseboom, T., de Rooij, S. & Painter, R. The Dutch famine and its long-term consequences for adult health. *Early Hum. Dev.* **82**, 485–491 (2006).
5. Weng, X. *et al.* Genome-wide DNA methylation profiling in infants born to gestational diabetes mellitus. *Diabetes Res. Clin. Pract.* **142**, 10–18 (2018).
6. Durnwald, C. Gestational diabetes: Linking epidemiology, excessive gestational weight gain, adverse pregnancy outcomes, and future metabolic syndrome. *Semin. Perinatol.* **39**, 254–258 (2015).
7. Monteiro, L. J., Norman, J. E., Rice, G. E. & Illanes, S. E. Fetal programming and gestational diabetes mellitus. *Placenta* **48**, S54–S60 (2016).
8. Farrar, D. *et al.* Association between hyperglycaemia and adverse perinatal outcomes in south Asian and white British women: Analysis of data from the Born in Bradford cohort. *Lancet Diabetes Endocrinol.* **3**, 795–804 (2015).
9. Fall, C. H. D. *et al.* Size at birth, maternal weight, and type 2 diabetes in South India. *Diabet. Med.* **15**, 220–227 (1998).
10. Yajnik, C. S. *et al.* Neonatal anthropometry: The thin-fat Indian baby. The Pune maternal nutrition study. *Int. J. Obes.* **27**, 173–180 (2003).

11. Khavari, D. A., Sen, G. L. & Rinn, J. L. DNA methylation and epigenetic control of cellular differentiation. *Cell Cycle* **9**, 3880–3883 (2010).

12. Moosavi, A. & Ardekani, A. M. Role of epigenetics in biology and human diseases. *Iran. Biomed. J.* **20**, 246–258 (2016).

13. Su, R. *et al.* Alteration in expression and methylation of IGF2/H19 in placenta and umbilical cord blood are associated with macrosomia exposed to intrauterine hyperglycemia. *PLoS One* **11**, (2016).

14. St-Pierre, J. *et al.* IGF2 DNA methylation is a modulator of newborn's fetal growth and development. *Epigenetics* **7**, 1125–1132 (2012).

15. Haertle, L. *et al.* Epigenetic signatures of gestational diabetes mellitus on cord blood methylation. *Clin. Epigenetics* **9**, 1–11 (2017).

16. Finer, S. *et al.* Maternal gestational diabetes is associated with genome-wide DNA methylation variation in placenta and cord blood of exposed offspring. *Hum. Mol. Genet.* **24**, 3021–3029 (2014).

17. Joubert, B. R. *et al.* 450K epigenome-wide scan identifies differential DNA methylation in newborns related to maternal smoking during pregnancy. *Environ. Health Perspect.* **120**, 1425–1431 (2012).

18. Bohlin, J. *et al.* Prediction of gestational age based on genome-wide differentially methylated regions. *Genome Biol.* **17**, 1–9 (2016).

19. Rahmani, E. *et al.* Sparse PCA corrects for cell type heterogeneity in epigenome-wide association studies. *Nat. Methods* **13**, 443–445 (2016).

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

21. Bakulski, K. M. *et al.* DNA methylation of cord blood cell types: Applications for mixed cell birth studies. *Epigenetics* **11**, 354–362 (2016).

22. Gunawardhana, L. P. *et al.* Differential DNA methylation profiles of infants exposed to maternal asthma during pregnancy. *Pediatr. Pulmonol.* **49**, 852–862 (2014).

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

24. Maksimovic, J., Gordon, L. & Oshlack, A. SWAN: Subset-quantile within array normalization for illumina infinium HumanMethylation450 BeadChips. *Genome Biol* **13**, R44 (2012).

25. Leek, J. T., Johnson, W. E., Parker, H. S., Jaffe, A. E. & Storey, J. D. The SVA package for removing batch effects and other unwanted variation in high-throughput experiments. *Bioinformatics* **28**, 882–883 (2012).

26. Reese, S. E. *et al.* DNA methylation score as a biomarker in newborns for sustained maternal smoking during pregnancy. *Environ. Health Perspect.* **125**, 760–766 (2017).

27. Altenberger, C. *et al.* SPAG6 and L1TD1 are transcriptionally regulated by DNA methylation in non-small cell lung cancers. *Mol. Cancer* **16**, 1–12 (2017).

28. Schroeder, J. W. *et al.* Neonatal DNA methylation patterns associate with gestational age. *Epigenetics* **6**, 1498–1504 (2011).

29. Zhang, B. *et al.* Maternal smoking during pregnancy and cord blood DNA methylation: new insight on sex differences and effect modification by maternal folate levels. *Epigenetics* **13**, 505–518 (2018).

30. Feig, D. S. *et al.* Clinical Practice Guidelines Diabetes and Pregnancy Diabetes Canada

Clinical Practice Guidelines Expert Committee. *Can. J. Diabetes* **42**, s255–s282 (2018).

31. Eades, C. E., Cameron, D. M. & Evans, J. M. M. Prevalence of gestational diabetes mellitus in Europe: A meta-analysis. *Diabetes Res. Clin. Pract.* **129**, 173–181 (2017).

32. Billionnet, C. *et al.* Gestational diabetes and adverse perinatal outcomes from 716,152 births in France in 2012. *Diabetologia* **60**, 636–644 (2017).

33. Martin, K. E., Grivell, R. M., Yelland, L. N. & Dodd, J. M. The influence of maternal BMI and gestational diabetes on pregnancy outcome. *Diabetes Res. Clin. Pract.* **108**, 508–513 (2015).

34. Cardenas, A. *et al.* Placental DNA methylation adaptation to maternal glycemic response in pregnancy. *Diabetes* **67**, 1673–1683 (2018).

35. Chen, P. *et al.* Differential methylation of genes in individuals exposed to maternal diabetes in utero. *Diabetologia* **60**, 645–655 (2017).

36. Kang, J., Lee, C. N., Li, H. Y., Hsu, K. H. & Lin, S. Y. Genome-wide DNA methylation variation in maternal and cord blood of gestational diabetes population. *Diabetes Res. Clin. Pract.* **132**, 127–136 (2017).

37. Luo, Z. C. *et al.* Maternal glucose tolerance in pregnancy affects fetal insulin sensitivity. *Diabetes Care* **33**, 2055–2061 (2010).

38. Gesteiro, E., Bastida, S. & Sánchez Muniz, F. J. Effects of maternal glucose tolerance, pregnancy diet quality and neonatal insulinemia upon insulin resistance/sensitivity biomarkers in normoweight neonates. *Nutr. Hosp.* **26**, 1447–1455 (2011).

39. Sahasrabuddh, A., Pitale, S., Raje, D. & Sagdeo, M. M. Cord blood levels of insulin and glucose in full term pregnancies. *J. Assoc. Physicians India* **61**, 378–382 (2013).

40. Ahmad, A. *et al.* Indices of glucose homeostasis in cord blood in term and preterm

newborns. *JCRPE J. Clin. Res. Pediatr. Endocrinol.* **8**, 270–275 (2016).

41. Luo, Z. C. *et al.* Maternal and fetal leptin, adiponectin levels and associations with fetal insulin sensitivity. *Obesity* **21**, 210–216 (2013).

42. Zhao, J. P. *et al.* Circulating docosahexaenoic acid levels are associated with fetal insulin sensitivity. *PLoS One* **9**, 1–7 (2014).

43. Zitti, B. & Bryceson, Y. T. Natural killer cells in inflammation and autoimmunity. *Cytokine Growth Factor Rev.* **42**, 37–46 (2018).

44. Cooper, M. A., Fehniger, T. A., Caligiuri, M. A., Cooper, M. A. & Caligiuri, M. A. Cooper et al., 2001 - Cópia. **22**, 633–640 (2001).

45. Horowitz, A. *et al.* Natural Killer cell Diversity Revealed By Mass Cytometry. *Sci. Transl. Med.* **5**, (2014).

46. Björkström, N. K., Ljunggren, H. G. & Michaëlsson, J. Emerging insights into natural killer cells in human peripheral tissues. *Nat. Rev. Immunol.* **16**, 310–320 (2016).

47. Parisi, L. *et al.* Natural Killer Cells in the Orchestration of Chronic Inflammatory Diseases. *J. Immunol. Res.* **2017**, (2017).

48. Blois, S. M. *et al.* NK cell-derived IL-10 is critical for DC-NK cell dialogue at the maternal-fetal interface. *Sci. Rep.* **7**, 1–9 (2017).

49. Perdu, S. *et al.* Maternal obesity drives functional alterations in uterine NK cells. *JCI Insight* **1**, 1–21 (2016).

50. O’Rourke, R. W., Gaston, G. D., Meyer, K. A., White, A. E. & Marks, D. L. Adipose tissue NK cells manifest an activated phenotype in human obesity. *Metabolism.* **62**, 1557–1561 (2013).

51. Viel, S. *et al.* Alteration of Natural Killer cell phenotype and function in obese

individuals. *Clin. Immunol.* **177**, 12–17 (2017).

52. Bähr, I. *et al.* Impaired natural killer cell subset phenotypes in human obesity. *Immunol. Res.* **66**, 234–244 (2018).

53. Michelet, X. *et al.* Metabolic reprogramming of natural killer cells in obesity limits antitumor responses. *Nat. Immunol.* **19**, 1330–1340 (2018).

54. O'Shea, D. & Hogan, A. E. Dysregulation of natural killer cells in obesity. *Cancers (Basel).* **11**, 1–12 (2019).

55. Bonaccorsi, I. *et al.* Natural killer cells in the innate immunity network of atherosclerosis. *Immunol. Lett.* **168**, 51–57 (2015).

56. Kyaw, T., Tipping, P., Toh, B. H. & Bobik, A. Killer cells in atherosclerosis. *Eur. J. Pharmacol.* **816**, 67–75 (2017).

57. Schiller, N. K., Boisvert, W. A. & Curtiss, L. K. Inflammation in atherosclerosis: Lesion formation in LDL receptor-deficient mice with perforin and Lystbeige mutations. *Arterioscler. Thromb. Vasc. Biol.* **22**, 1341–1346 (2002).

58. Whitman, S. C., Rateri, D. L., Szilvassy, S. J., Yokoyama, W. & Daugherty, A. Depletion of natural killer cell function decreases atherosclerosis in low-density lipoprotein receptor null mice. *Arterioscler. Thromb. Vasc. Biol.* **24**, 1049–1054 (2004).

59. Selathurai, A. *et al.* Natural killer (NK) cells augment atherosclerosis by cytotoxic-dependent mechanisms. *Cardiovasc. Res.* **102**, 128–137 (2014).

60. Nour-Eldine, W. *et al.* Genetic Depletion or Hyperresponsiveness of Natural Killer Cells Do Not Affect Atherosclerosis Development. *Circ. Res.* **122**, 47–57 (2018).

61. Faria, S. S. *et al.* The neutrophil-to-lymphocyte ratio: A narrative review. *Ecancermedicalscience* **10**, 1–12 (2016).

62. Bhat, T. *et al.* Neutrophil to lymphocyte ratio and cardiovascular diseases: A review. *Expert Rev. Cardiovasc. Ther.* **11**, 55–59 (2013).

63. Lou, M. *et al.* Relationship between neutrophil-lymphocyte ratio and insulin resistance in newly diagnosed type 2 diabetes mellitus patients. *BMC Endocr. Disord.* **15**, 4–9 (2015).

64. Mertoglu, C. & Gunay, M. Neutrophil-Lymphocyte ratio and Platelet-Lymphocyte ratio as useful predictive markers of prediabetes and diabetes mellitus. *Diabetes Metab. Syndr. Clin. Res. Rev.* **11**, S127–S131 (2017).

65. Sefil, F. *et al.* Investigation of neutrophil lymphocyte ratio and blood glucose regulation in patients with type 2 diabetes mellitus. *J. Int. Med. Res.* **42**, 581–588 (2014).

66. Li, X. *et al.* High neutrophil-to-lymphocyte ratio is associated with increased carotid artery intima-media thickness in type 2 diabetes. *J. Diabetes Investig.* **8**, 101–107 (2017).

67. Liu, S. *et al.* Neutrophil-to-lymphocyte ratio is associated with diabetic peripheral neuropathy in type 2 diabetes patients. *Diabetes Res. Clin. Pract.* **130**, 90–97 (2017).

68. Sargın, M. A. *et al.* Neutrophil-to-lymphocyte and platelet-to-lymphocyte ratios: Are they useful for predicting gestational diabetes mellitus during pregnancy? *Ther. Clin. Risk Manag.* **12**, 657–665 (2016).

69. Hadarits, O. *et al.* Increased proportion of hematopoietic stem and progenitor cell population in cord blood of neonates born to mothers with gestational diabetes mellitus. *Stem Cells Dev.* **25**, 13–17 (2016).

70. Ambatipudi, S. *et al.* Assessing the role of DNA methylation-derived neutrophil-to-lymphocyte ratio in rheumatoid arthritis. *J. Immunol. Res.* **2018**, (2018).

71. Koestler, D. C. *et al.* DNA methylation-derived neutrophil-tolymphocyte ratio: An epigenetic tool to explore cancer inflammation and outcomes. *Cancer Epidemiol.*

Biomarkers Prev. **26**, 328–338 (2017).

72. Grieshober, L. *et al.* Methylation-derived Neutrophil-to-Lymphocyte Ratio and Lung Cancer Risk in Heavy Smokers. *Cancer Prev. Res.* **11**, 727–734 (2018).

73. Arroyo, V. M. *et al.* Pilot study of DNA methylation-derived neutrophil-to-lymphocyte ratio and survival in pediatric medulloblastoma. *Cancer Epidemiol.* **59**, 71–74 (2019).