CONTRIBUTIONS OF INTRINSIC AND ENVIRONMENTAL FACTORS TO EARLY LIFE DNA METHYLATION

by

Sarah Jessica Goodman

B.Sc., Queens University, 2012

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

in

THE FACULTY OF GRADUATE AND POSTDOCTORAL STUDIES (Medical Genetics)

THE UNIVERSITY OF BRITISH COLUMBIA (Vancouver)

September 2019

© Sarah Jessica Goodman, 2019
The following individuals certify that they have read, and recommend to the Faculty of Graduate and Postdoctoral Studies for acceptance, the dissertation entitled:

Contributions of intrinsic and environmental factors to early life DNA methylation

submitted by Sarah J. Goodman in partial fulfillment of the requirements for
the degree of Doctor of Philosophy
in Medical Genetics

Examing Committee:

Dr. Michael S. Kobor
Supervisor

Dr. Wendy P. Robinson
Supervisory Committee Member

Dr. Robert McMaster
Supervisory Committee Member

Dr. Jan M. Friedman
University Examiner

Dr. Ruth Eckstein Grunau
University Examiner

Additional Supervisory Committee Members:

Dr. Stuart E. Turvey
Supervisory Committee Member
Abstract

Many early experiences and exposures are known to cause health disparities later in life, suggesting that they are somehow ‘biologically embedded’. The mechanisms underlying ‘biological embedding’ are currently not well understood. However, emerging evidence has implicated a potential contribution from epigenetic modifications, such as DNA methylation (DNAm), which has been shown to associate with early life experiences of low socioeconomic status. Additional related experiences have also been connected with DNAm, including parental stress, childhood maltreatment or deprivation, and maternal mental health problems during the perinatal period. The relationship between early life experience and epigenetics is complicated by internal psychological and physiological factors, as well as genetic variation, which can account for 20% to 80% of inter-individual epigenetic differences. As well, stress reactivity and temperament are predictive of how a child may interact with his or her environment and can affect how such exposures are internalized. Thus, the main objectives of my dissertation were to elucidate the relationships between childhood environment, DNAm, genetic variation, and behaviour, to understand how these systems influence one another. Using matched DNAm profiles from blood and buccal tissue from a cross-sectional cohort of Canadian children, I uncovered tissue-specific and -shared DNAm signatures in order to glean the utility of accessible tissues in epigenetic association studies. In a longitudinal cohort, I tested the hypothesis that indicators of children’s early internal, biological and behavioural responses to stressful challenges are linked to stable patterns of DNAm later in life; I found relationships between biobehavioural response propensities in early life and patterns of DNAm in DLX5 and IGF2 genes at ages 15 and 18. Finally, I examined the epigenetic correlates of familial socioeconomic status in matched childhood peripheral blood and dried neonatal blood spot samples, allowing
me to assess the DNAm pattern over time. Together these findings build upon our current understanding of the role of DNAm in biological embedding and more broadly, the field of social epigenetics.
Lay Summary

Childhood is a formative time, during which experiences and environments can alter developmental outcomes. Familial socioeconomic status (SES) in early life, for example, is a powerful determinant of health across the life course. However, this relationship is complicated by individual factors including genetic variation and stress reactivity. The biological mechanism by which all these components interact and ultimately “get under the skin” to impact long term health and wellbeing is not well understood. However, emerging evidence has implicated a contribution of DNA methylation (DNAm), a molecular marker that regulates gene activity. Early life exposures including low SES, parental stress, and childhood maltreatment or deprivation have all been linked to unique DNAm patterns. The research presented here, supports and broadens our understanding of how both internal and external factors associate with DNAm with the goal of informing targeted, multilevel interventions capable of abating societal inequalities in health.
Preface

All data chapters in this thesis (Chapters 3-5) are presented in manuscript format, as they are currently published (Chapters 3 and 4) or in preparation for submission (Chapter 5).

Portions of Chapter 1 (Introduction) were copied or adapted from a previously published review article:


Specifically, subsection 1.6, titled “DNA methylation over the life course”, contains excerpts of this review article which was jointly written and edited Dr. M. Jones and myself in Dr. M. Kobor’s lab. The remainder of this chapter is original and unpublished, and written by myself.

Chapter 2 is original, unpublished work summarizing methodology, which I wrote.

A version of Chapter 3 has been published as:


This study was conducted in Dr. M. Kobor’s lab. Dr. S. Islam and I contributed equally to this work with regards to study design, statistical analysis, and manuscript writing. Drs. J. MacIsaac and L. McEwen assisted in running DNA methylation arrays. The primary investigators of the GECKO and C3ARE cohorts are Dr. W.T. Boyce and Dr. R. Barr, respectively.

A version of Chapter 4 has been published as:


Dr. M. Essex, the principal investigator of the cohort. Dr. D. Roubinov, Dr. M. Kobor, Dr. M. Essex, Dr. W.T. Boyce and I all contributed intellectually to study design and hypothesis generation. L. Lam, S. Nuemann and Dr. M. Park ran the DNA methylation arrays. I ran the statistical analysis, performed the pyrosequencing in Dr. M. Kobor’s lab. I wrote the manuscript with Dr. D. Roubinov. Dr. M. Kobor and Dr. W.T. Boyce made editorial contributions.

Chapter 5 is original work conducted in Dr. M. Kobor’s lab, in collaboration with Dr. W.T. Boyce, the principal investigator of the cohort. I ran the DNA methylation and genotyping arrays with assistance from Dr. J. MacIsaac and Dr. G. Leung. I also ran all statistical analysis and wrote the manuscript.

Chapter 6 (Conclusions) is original, unpublished work, which I wrote.
All research articles noted above are available under the terms of the Creative Commons Attribution License (CC BY), which permits distribution and reproduction in any medium, given proper citation.

Ethics approval for sample collection and epigenetic analysis was obtained separately for each cohort (GECKO, C3ARE, and WSFW) from the University of British Columbia, Children and Women’s Hospital Ethics board (Certificates H07-01317 (C3ARE) and H07-02773 (GECKO, WSFW)). Written informed consent was obtained from a parent or legal guardian, and assent was obtained from each child before study participation.
Table of Contents

Abstract ........................................................................................................................................ iii
Lay Summary .................................................................................................................................. v
Preface ........................................................................................................................................... vi
Table of Contents ........................................................................................................................... ix
List of Tables .................................................................................................................................... xv
List of Figures ................................................................................................................................. xvi
List of Abbreviations ..................................................................................................................... xviii
Acknowledgements ....................................................................................................................... xxi
Dedication ......................................................................................................................................... xxii

Chapter 1: Introduction ..................................................................................................................... 1

1.1 Developmental origins of health and disease through developmental plasticity ........... 1
1.2 The impact of early life socioeconomic status and social environment on health and
development ............................................................................................................................... 3
1.3 Differential susceptibility to early life environments ............................................................. 6
  1.3.1 Gene by environment interactions in the context of early life environments .......... 7
1.4 Epigenetics ............................................................................................................................... 8
1.5 DNA methylation ..................................................................................................................... 8
  1.5.1 DNA methylation involvement in gene regulation ...................................................... 11
  1.5.2 Genetic background contributes to DNA methylation variation .............................. 12
  1.5.3 Tissue specific DNA methylation and cellular differentiation ................................. 13
1.6 DNA methylation over the life course .................................................................................. 13
1.7 Evidence of DNA methylation alterations by SES-related early environmental exposures

1.7.1 Animal studies of epigenetics and early exposures ........................................ 16
1.7.2 Human studies of epigenetics and early exposures ........................................ 17

1.8 Synthesis and thesis objectives ........................................................................... 20

Chapter 2: Common methods and materials .............................................................. 22

2.1 Introduction to epigenome-wide association studies ........................................ 22

2.2 Study samples ..................................................................................................... 23
2.2.1 Wisconsin Study of Family and Work .............................................................. 23
2.2.2 Gene Expression Collaborative Kids Only ...................................................... 25

2.3 DNA extraction and sodium bisulfite conversion .............................................. 26

2.4 Illumina DNA methylation microarrays ............................................................. 27
2.4.1 Illumina Infinium HumanMethylation27 BeadChip ......................................... 28
2.4.2 Illumina Infinium HumanMethylation450 BeadChip ....................................... 28

2.5 DNA methylation data processing pipeline ....................................................... 29
2.5.1 Prepossessing and quality control ................................................................. 29
2.5.2 Normalization .................................................................................................. 31
2.5.3 Cell-type correction of 450K DNA methylation data ...................................... 31
2.5.4 Batch correction ............................................................................................. 32

2.6 Data analysis ....................................................................................................... 33
2.6.1 Data reduction .................................................................................................. 33
2.6.2 Statistical analysis of differential DNAm ..................................................... 33
Chapter 3: Integration of DNA methylation patterns and genetic variation in human pediatric tissues help inform EWAS design and interpretation

3.1 Introduction

3.2 Materials and methods

3.2.1 Study cohorts and tissue samples

3.2.2 DNA isolation and DNA methylation arrays

3.2.3 DNA methylation array data quality control and normalization

3.2.4 Cell-type correction of DNA methylation data

3.2.5 Assessment of cross-tissue correlation, tissue-specific variability and tissue-specific differences in DNA methylation data

3.2.6 SNP genotyping arrays

3.2.7 Preprocessing of SNP genotyping data and PCA analyses for genetic ancestry

3.2.8 Cis-mQTL analyses

3.2.9 Representation of identified sites in published EWAS findings

3.3 Results

3.3.1 Study cohorts and DNAm data processing

3.3.2 BEC DNAm had greater inter-individual variability than PBMC DNAm

3.3.3 Variable CpGs were more highly correlated between tissues

3.3.4 Genetic variation contributed to tissue concordance

3.3.5 Tissue-specific differential DNAm was consistent across cohorts

3.3.6 cis-mQTLs were present in previously published EWAS findings

3.4 Discussion
<table>
<thead>
<tr>
<th>Section</th>
</tr>
</thead>
<tbody>
<tr>
<td>5.2</td>
</tr>
<tr>
<td>Methods and materials</td>
</tr>
<tr>
<td>5.2.1</td>
</tr>
<tr>
<td>Study sample</td>
</tr>
<tr>
<td>5.2.2</td>
</tr>
<tr>
<td>SES measures</td>
</tr>
<tr>
<td>5.2.3</td>
</tr>
<tr>
<td>DNA methylation data</td>
</tr>
<tr>
<td>5.2.4</td>
</tr>
<tr>
<td>Genotyping data generation and use in inferring population structure</td>
</tr>
<tr>
<td>5.2.5</td>
</tr>
<tr>
<td>Statistical analysis of DNAm in PBMCs</td>
</tr>
<tr>
<td>5.2.6</td>
</tr>
<tr>
<td>Gene ontology</td>
</tr>
<tr>
<td>5.2.7</td>
</tr>
<tr>
<td>Identification of differentially methylated regions</td>
</tr>
<tr>
<td>5.2.8</td>
</tr>
<tr>
<td>mQTL analysis</td>
</tr>
<tr>
<td>5.2.9</td>
</tr>
<tr>
<td>Statistical analysis in dried blood spots</td>
</tr>
<tr>
<td>5.3</td>
</tr>
<tr>
<td>Results</td>
</tr>
<tr>
<td>5.3.1</td>
</tr>
<tr>
<td>Objective and subjective SES measures in the GECKO cohort</td>
</tr>
<tr>
<td>5.3.2</td>
</tr>
<tr>
<td>Subjective SES and objective SES tended to associate with different CpGs</td>
</tr>
<tr>
<td>5.3.3</td>
</tr>
<tr>
<td>Loci associated with objective SES contained more underlying DMRs than those associated with subjective SES</td>
</tr>
<tr>
<td>5.3.4</td>
</tr>
<tr>
<td>Few SES-associated CpGs were also driven by genetic variation</td>
</tr>
<tr>
<td>5.3.5</td>
</tr>
<tr>
<td>Assessing DNAm in neonatal dried blood spots</td>
</tr>
<tr>
<td>5.4</td>
</tr>
<tr>
<td>Discussion</td>
</tr>
<tr>
<td>5.4.1</td>
</tr>
<tr>
<td>Limitations</td>
</tr>
<tr>
<td><strong>Chapter 6: Conclusions</strong></td>
</tr>
<tr>
<td>6.1</td>
</tr>
<tr>
<td>Summary of dissertation results</td>
</tr>
<tr>
<td>6.2</td>
</tr>
<tr>
<td>Limitations</td>
</tr>
<tr>
<td>6.3</td>
</tr>
<tr>
<td>Future directions</td>
</tr>
</tbody>
</table>
References ............................................................................................................. 148

Appendices ........................................................................................................... 179

Appendix A Supplementary materials for chapter 3 ....................................................... 179
   A.1 Supplementary figures ................................................................................. 179
   A.2 Supplementary tables ................................................................................. 184

Appendix B Supplementary materials for chapter 4 ....................................................... 187
   B.1 Supplementary figures ................................................................................. 187
   B.2 Supplementary tables ................................................................................. 193

Appendix C Supplementary materials for chapter 5 ....................................................... 194
   C.1 Supplementary figures ................................................................................. 194
   C.2 Supplementary tables ................................................................................. 195
List of Tables

Table 3.1 Sample characteristics for C3ARE and GECKO cohorts ........................................... 48

Table 4.1 Mental health, temperament and ANS traits collected over seven years and included in the analysis .................................................................................................................................. 74

Table 5.1 Sample characteristics for full GECKO cohorts and subset presented in this study... 109

Table 5.2 CpGs associated with both PosCan and CompSES (|Δβ| > 5% and p-value < 5x10⁻⁴) ........................................................................................................................................... 121
List of Figures

Figure 3.1 BEC DNAm was consistently more variable than PBMC DNAm at the genome-wide and probe-wise level. .................................................................51

Figure 3.2 Variable CpGs were more highly correlated between tissues. .........................54

Figure 3.3 Independently validated cis-mQTL were more likely to be shared across tissues than expected by chance. ........................................................................................................57

Figure 3.4 Tissue-specific differential DNA methylation was consistent across cohorts........59

Figure 3.5 Overlap and representation of identified CpGs in previously published pediatric EWAS findings.................................................................62

Figure 4.1 Results of principal component analysis revealed biobehavioural reactivity as biologically driven composite measure.................................................................85

Figure 4.2 Schematic of DLX5, IGF2, MYO16 and PRUNE2 genes, which each contained more than 2 high or medium confidence CpGs significantly associated with Biobehavioural Inhibition/Disinhibition.................................................................88

Figure 4.3 Correlations between DNA methylation at age 15 and Biobehavioural Inhibition/Disinhibition in males and females.................................................................91

Figure 4.4 DLX5 and IGF2 DNA methylation remained significantly associated with Biobehavioural Inhibition/Disinhibition at age 18. .................................................................93

Figure 4.5 CpG stability across three years and pyrosequencing of DLX5 gene to verify DNA methylation findings. ........................................................................................................96

Figure 5.1 Relations between sample characteristics and SES measures in full GECKO cohort were maintained in subset used in DNAm analysis.................................................................118
Figure 5.2 Genome-wide linear regressions showed limited numbers of CpGs associated with each SES variable. ................................................................. 121
Figure 5.3 Limited enrichment of genomic features found in CpGs associated with CompSES or PosCan. .................................................................................. 123
Figure 5.4 CompSES found to be associated with more DMRs than PosCan. ......................... 125
Figure 5.5 Little overlap between CpGs associated with PosCan or CompSES and mQTL-associated CpGs. ................................................................. 126
Figure 5.6 CpGs associated with SES in both PBMCs and DBSs showed similar effect sized across time points................................................................. 129
List of Abbreviations

27K array – Illumina Infinium HumanMethylation27 BeadChip array
450K array – Illumina Infinium HumanMethylation450 BeadChip array
|Δβ| – Absolute delta beta
ANS – Autonomic nervous system
BEC – Buccal epithelial cells
BID – Biobehavioural inhibition/disinhibition
bp – base pair
C – Celsius
C3ARE – Cleaning, Carrying, Changing, Attending, Reading and Expressing
CBC/Diff – Complete blood count with differential
CGI – CpG island
CompSES – Composite socioeconomic status
CpG – Cytosine-guanine dinucleotide
DBS – Dried blood spot
DMNT – DNA methyltransferase
DMR – Differentially methylated region
DNA – Deoxyribonucleic acid
DNAm – DNA methylation
DOHaD – Developmental Origins of Health and Disease
ESC – Embryonic stem cell
EWAS – Epigenome-wide association study
FDR – False discovery rate
GECKO – Gene Expression Collaborative Kids Only

GO – Gene ontology

GR – Glucocorticoid receptor

GSR – Gene score resample

GWAS – Genome-wide association study

GxE – Gene by environment interactions

HPA – Hypothalamic-pituitary-adrenal

kb – Kilobase

mQTL – Methylation quantitative trait locus

mQTLs – Methylation quantitative trait loci

mRNA – Messenger ribonucleic acid

NIMH – National Institutes of Mental Health

ORA – Over-representation analysis

PBMC – Peripheral blood mononuclear cell

PCR – Polymerase chain reaction

PosCan – Position in Canada

RNA – Ribonucleic acid

SES – Socioeconomic status

SNP – Single nucleotide polymorphism

TET – Ten-eleven translocation methyleytosine dioxygenase

TSS – Transcriptional start site

UCSC – University of California, Santa Cruz Genome Browser

VNTR - Variable number tandem repeat
WSFW – Wisconsin Study of Family and Work
Acknowledgements

I am very grateful to have been mentored by many tremendous researchers who educated and supported me throughout these past seven years. To my supervisor, Dr. Michael Kobor, thank you for guidance and encouragement on all of my research endeavours and for allowing me to be a part of your amazing team. To my collaborators, Dr. Tom Boyce, Dr. Marilyn Essex, Dr. Nicki Bush and Dr. Danielle Roubinov, I am so fortunate to have had the opportunity to work with you and learn from you. This research would not have been possible without your passion and knowledge. Many thanks to my advisory committee, Dr. Wendy Robinson, Dr. Rob McMaster, and Dr. Stuart Turvey for your thoughtful feedback and insight, and for calming my nerves during committee meetings.

To my incredible lab mates, friends, and an amazing cohort of MedGen students, thank you for being my problem solvers, confidantes, and my Vancouver family (especially Maria, Magda, and Gian). Special thanks to Cheryl Bishop for your helping hand, endless support, and enthusiasm.

Last but not the least, I would like to thank my parents and sisters for never allowing me to go more than a few days without a phone call to remind me that I wasn’t really that far from home and you were expecting me back soon.
Dedication

To my inspiring and selfless mother.
Chapter 1: Introduction

1.1 Developmental origins of health and disease through developmental plasticity

Early epidemiological studies showing geographical relationships between rates of infant mortality and adult coronary heart disease collectively gave rise to the developmental origins of health and disease (DOHaD) hypothesis; this paradigm posits that environmental factors during fetal development and infancy contribute to chronic disease susceptibility (1). Fetal mortality in the UK, commonly attributed to low birthweight, was geographically linked to rates of cardiovascular disease; this indicated that prenatal environment, namely nutrition, contributed to later cardiovascular health (1). Findings from Norway, Finland and the US showed that poor living conditions, such as recurrent exposure to infection, and resultant infant mortality rates were also geographically associated with the occurrence of mortality from heart disease (1-3). These findings established the contribution of prenatal and neonatal environment to health outcomes across the lifespan and the field of DOHaD research, which has since expanded to include the study of childhood or “early life” environment (4).

In early development, a critical window of time exists during which biological systems are altered by and thereby adapt to environmental exposures and influence long-term health, and this phenomenon of environmental programming is a central tenet of the DOHaD approach (5). Animal studies of short nutritional alterations provide overwhelming evidence of developmental programming and consequential life-long health outcomes (5). Overfeeding infant baboons leads to obesity which only arises in adulthood (6). In rat studies both pre- and postnatal nutrition have been shown to influence adult health (5). Male rats born to mothers fed a low protein diet during pregnancy but suckled by mothers with a normal diet have significantly shorter life spans than
controls (5); however, another study showed that rats were permanently smaller after suckling mothers fed a low protein, but were normal sized when mother were fed the same diet during pregnancy (7). In sum, animal studies have revealed prenatal nutrition altering brain development, learning, and behaviour in addition to metabolic health outcomes.

In humans, longitudinal findings from adults exposed in utero to the Dutch Hunger Winter include increased risk of obesity, abnormal lipid profiles, cardiovascular disease, and affective disorders (8). These outcomes differ based on the timing of exposure; those exposed only during early gestation had normal birth weights, while those exposed at late gestation had reduced birth weights (8). As such, windows of environmental sensitivity or critical periods are thought to differ for different biological systems. Exposures beyond nutrition, which have been frequently studied and reproducibly linked to later health outcomes include BPA, mercury, air pollution and particulate matter, lead, organophosphates (9). These and other environmental exposures have been associated with a wide range noncommunicable diseases including obesity, diabetes, hypertension, asthma and allergy, immune and autoimmune diseases, neurodevelopmental and neurodegenerative diseases, infertility, cancers, and mental health disorders (10-12). Less tangible exposures such as various causes of and exposures to stress, have also been examined in relation to developmental outcomes (13). Offspring of pregnant mothers exposed to stress show altered endocrine, immune and neurobehavioural outcomes as compared to controls, as well as decreased birth weight and decreased gestational age at birth (14-17). Therefore, during fetal development and early life, the body and brain are vulnerable to the surrounding environment, both positive and negative, which makes this time frame especially pertinent to human health and disease research.
1.2 The impact of early life socioeconomic status and social environment on health and development

Socioeconomic Status (SES) is a complex, multifaceted measure of the economic or social standing or class of an individual or group. It is often measured as a combination of education, income and occupation and, importantly, is one the most potent environmental determinants of human health, with well-established influences on morbidity and mortality across industrialized populations (18). Across the life span adverse health outcomes increase gradually from the most to the least privileged populations, creating an observable gradient (18). Adverse health outcomes to low SES include increased incidence of heart disease, asthma, obesity and diabetes in a graded fashion even when access to health care, ethnicity and health risk behaviours are all taken into account (19,20). Within these associations, SES in childhood has proven to be as strong a predictor of health outcomes as adult SES (21-23). As such, SES researchers have argued for the importance of the “life course approach”, which examines SES across the life course to better understand the relationship between SES and variations in health (21,24). Most simply put, early life SES strongly predicts SES in adult life, with advantage or disadvantage often accumulating within one’s life (25). Beyond this association, early life SES can impact many aspects of development including physical, social-emotional and cognitive development and concurrent health outcomes which may beget later health outcomes (19,26,27).

One prominent theory addressing why the outcomes of early-life exposures sometimes manifest later in life was proposed by Gluckman and Hanson (2004) (28). They posit that in childhood, alterations in response to the environment may be grouped into three categories based on representing: 1) immediately beneficial and adaptive responses to certain environments; 2) the negative outcomes or damage from exposure to certain environments; 3) predictive adaptive
responses which “confer advantage by establishing metabolic physiology appropriate for the postnatal environment predicted to exist” (28). In the last scenario, such alterations set the child or infant on a specific development trajectory that would only be adaptive if environments later in life mirror those in early life. Such developmental programming can lead to both beneficial and detrimental ‘latent effects’, outcomes which are not concurrent with the environment but rather manifest later in life.

Environments altered by SES affect all aspects of life. Children from low-SES families, in comparison to more advantaged peers, are more likely to live in poor-quality housing with inadequate water, heating, and sewage (29). These homes tend to be noisier, crowded, and unsafe (29,30). This is also seen in the schools they attend and the neighborhoods in which children reside (29). Healthy food is more likely to be scarce in low-SES households, due to high costs and limited availability in local markets (29,31). The neighborhoods, in general, may be less safe with higher crimes rates (29). These differences represent persistent misfortunes of children growing up in low-SES families and unfortunately, maltreatment and disadvantage co-occur more often than expected by chance alone, meaning that children who are abused or neglected may also be more likely fall below the population average with regards to material resources (19,29). However, a sizable minority of individuals show resiliency to their surroundings, maintaining their mental and physical health despite adversity (32,33). There are psychological adaptations that may act as buffers; these include the child’s perspective of his/her SES, his/her temperament and cognitive abilities (32). Family and school setting can also offset misfortune by providing a warm, supportive environment for children to thrive (32). Finally, early interventions may diminish the risks associated with low-SES environments in childhood, highlighting the necessity of understanding the mechanisms by which SES gradients influence health (34).
Although the evidence of early environment shaping one’s health is overwhelming, the molecular mechanisms through which environmental exposures are “biologically embedded”, thereby leaving lasting biological residues, remain largely unknown. Nevertheless, numerous studies have linked SES-related health problems to dysregulation of inflammatory pathways (35-38). More specifically, low SES consistently associates with a wide-range of pro-inflammatory profiles including elevated levels of circulating C-reactive protein and interleukin-6, as well as increasing cortisol output, which is indicative of accumulative “wear and tear” of stress-response pathways due to chronic stress (35,39). As well, these findings have been supported by analyses of genomic expression profiles of individuals from low and high SES backgrounds (40,41). Resulting immunological differences have the capacity to influence a wide-range of physical and mental health outcomes. Mild, chronic inflammation, such as that associated with low-SES, contributes to the pathogenesis of atherosclerosis, obesity, cancer, chronic obstructive pulmonary disease, asthma and neurodegenerative disease (42).

Inflammation linked to SES can also be associated with mental health outcomes. Individuals with major depressive disorder consistently display a pro-inflammatory profile, especially in cytokines (36,43). Cytokines are small proteins involved in cell signaling to mediate the innate immune response. Cytokines may contribute to depression by influencing hypothalamic-pituitary-adrenal (HPA) axis function, as well as crossing the blood-brain barrier and affecting neurotransmitter activity (43). Dysregulation of inflammatory pathways undoubtedly couples early life SES to later disease occurrence, however other biological mechanisms, such as genetic variation and temperament, i.e. individual, inborn differences in one’s reactions to environmental stimuli, can also contribute to these outcomes.
1.3 Differential susceptibility to early life environments

Not all children raised in impoverished or adverse environments develop the common negative health outcomes observed at population levels. Similarly, some children will not meet the expected developmental trajectories despite being afforded privileged upbringings. Individual factors, which can be summarized as differential susceptibility to the environment, can moderate or buffer early life experiences of SES and therefore alter health outcomes (32,44,45). Many factors contribute to one’s overall sensitivity or resiliency to environmental exposures (33). Differential susceptibility is closely related to levels of autonomic nervous system (ANS) reactivity in response to stressful events, i.e. the physiological measures of the “flight or fight” response. Temperament, which can be defined as characteristics that emerge shortly after birth and govern a child’s responses to new or different situations, is another common plasticity or susceptibility marker (44). Temperament has been shown to buffer the effects of parenting style and quality of child-care on psychological outcomes such as externalizing behaviour (44). ANS reactivity and the construct of temperament are closely connected, as a child who has a more fearful, inhibited temperament also typically displays stronger ANS reactivity (46,47). Furthermore, individuals’ perceptions of the environment also mediate their outcomes to negative experiences (48,49); Subjective factors (e.g., feelings of loneliness) can bear a greater influence on health outcomes, such as lowered transcription of anti-inflammatory genes, elevated systolic blood pressure, and mortality risk in the elderly, than objective measures of social isolation (50-52). Thus, an individual’s internal milieu can provide insight into neurodevelopment and potential health outcomes in the face of adversity and stress (44).
1.3.1 Gene by environment interactions in the context of early life environments

Gene environment (GxE) interactions, whereby genes outcomes to environmental exposures arise in an allele-dependent manner, can underlie differences in environmental sensitivity and resiliency. Numerous genes, many of which are related to stress-response pathways and/or neuronal circuitry, have been shown to interact with early life environments to alter health outcomes. For example, the gene encoding the serotonin transporter, \textit{SLC6A4}, contains an upstream variable number tandem repeat (VNTR), 5-HTTLPR, with two common alleles. The likelihood of developing mental health problems following adverse experiences in childhood such as maltreatment, stress, and sexual abuse has been consistently associated to 5-HTTLPR genotype frequencies (53). Reported associated health outcomes include but are not limited to depressive symptoms, anxiety, PTSD, and suicidal ideation (53). Other genes often cited as displaying GxE interactions with early life adversity, include \textit{BDNF} (Brain derived neurotrophic factor), \textit{COMT} (catechol-O-methyltransferase), \textit{DRD4} (dopamine receptor D4), \textit{NR3C1} (glucocorticoid receptor), and \textit{MAOA} (monoamine oxidase A) (54-57). Work from the Bucharest Early Intervention Project uncovered a protective effect of the \textit{COMT} met allele (val158met polymorphism) against depressive symptoms in children who remained longer in institutions (58). Not surprisingly, the functions of many GxE genes are critical to neurodevelopment and thus contribute to inter-individual variation in temperament and reactivity (59). Therefore, the relationship between genotype, reactivity/temperament and differential susceptibility to early life adversity is complex but critical to predicting which specific children are most at risk.
1.4 Epigenetics

Epigenetics is a relatively nascent and expanding field investigating the nexus of genome and environment; more specifically, epigenetics is commonly defined as modifications to DNA or DNA packaging that are potentially transmissible to daughter cells and that do not involve changes to the DNA sequence (60,61). It encompasses the idea of “cellular memory” or molecular changes that persist after the original stimuli has ceased (61). This definition is quite close to the original use of “epigenetics”, as described by Conrad Waddington in 1942 – “the branch of biology which studies the causal interactions between genes and their products, which bring the phenotype into being” (62). We now know that epigenetic mechanisms are responsible for the structural organization and packaging of DNA; by altering DNA accessibility, gene expression programs may be orchestrated to drive cellular processes. However, it has been argued that temporal and tissue-specific differences in gene expression are more appropriately described as “transcriptional regulation” and do not fall into the realm of epigenetics (61). Epigenetic marks include to acetylation, phosphorylation, and methylation of histone proteins and as well as other post-translational modifications; covalently bound modifications to DNA, such as DNA methylation (DNAm); non-canonical histone variants; and non-coding RNAs.

1.5 DNA methylation

DNAm refers to the covalent attachment of a methyl group primarily to the 5’carbon of cytosine bases found most often at cytosine-phosphate-guanine dinucleotides (CpGs) (63). DNAm occurring at non-CG dinucleotides is relatively infrequent, found predominantly in fetal tissue, specifically fetal brain tissue, and in low but measurable levels in adult neuronal tissue (64,65). The process of DNAm is catalyzed by enzymes called DNA methyltransferases
(DNMT), which transfer methyl groups from a methyl donor, S-Adenosyl methionine, to unmethylated DNA. There are three functional DNMTs, which are highly conserved across mammals: DNMT3A, DNMT3B, DNMT1. DNMT3A and DNMT3B are de novo methyltransferases and function through the interaction of a regulatory factor DNMT3L with specific histone modification, followed by recruitment the methyltransferase (66). DNMT1 is the maintenance methyltransferase responsible for DNAm at hemi-methylated sites during cell division (67). CpG dinucleotides are palindromic and therefore both DNA strands are methylated at CpGs (63). As such, during cell division, if the parental strand is methylated; it can be used a template for methylation of the newly synthesized strand by DNMT1, thus replicating the DNAm state in a semiconservative fashion (68). However, human cancer cells lacking DNMT1 were able to maintain methylation following division; this suggests de novo methyltransferases play a role in maintenance of DNAm. DNMT3A and DNMT3B are required both for normal fetal development, while loss of DNMT1 is embryonic lethal (68,69).

DNA demethylation is currently believed to occur both passively and actively; the former can take place when DNAm is not maintained following DNA replication and subsequently becomes “diluted” out (70). Active demethylation is likely a function of the ten-eleven translocation methylcytosine dioxygenase (TET) family of enzymes, which catalyze the first step of the demethylation process through the oxidation of the 5-methylcytosine (71). This reaction produces hydroxymethylation, which is both an intermediary step in demethylation and an important epigenetic marker involved in development and brain function (65,72). To that end, hydroxymethylation marks are most prevalent in progenitor cells and neurons (72); however the exact function of this mark and its role gene regulation and chromatin modification is not well understood (72).
The distribution of DNAm along the genome is non-random and highly dependent on the underlying genomic landscape, partly because CpGs themselves are not evenly distributed across the genome (63). CpGs are under-represented in the genome but are enriched in some regions, which are referred to as CpG islands (CGIs). Transcription factor binding is thought to play a role in both CpG density in the DNA sequence, as well as DNAm status (73). With regards to the latter, transcription factors, in part, establish genome-wide DNAm states. This phenomenon is especially evident in development, during which binding sites of cell-specific transcription factors exhibit little to no DNAm (74). Transcription factors bound to DNA, especially those with CGs in their binding motifs, likely prevent methyl sites from being added to the DNA sequence. CpGs found outside of gene regulatory regions are less likely to be bound by transcription factors (73,75). Evolutionarily, these sites were more likely to be methylated and therefore, less stable and more prone to mutation (73). As such, transcription factor binding is believed to have contributed to the current patterning of CpG density in the genome and the existence of CpG islands (CGI).

CGIs are often associated with gene promoters and are commonly defined as regions spanning 200 base pairs or longer and with at least 50% guanine or cytosine content and 0.6 observed/expected ratio of CpGs (76,77). Despite the high density of CpG dinucleotides, CGI are less likely methylated compared to non-island CpGs. One exception to this pattern is CpGs found at repetitive elements, such as Alu and LINE-1, which are highly methylated in order to repress transposition (78,79). Regions immediately surrounding CGIs are referred to as “shores”, followed by “shelves”. Specifically, CpG shores are found within 2kb outside of the CGI boundaries and CpG shelves are between 2-4kb outside of the CpG island boundaries (80,81). Beyond CGI status, CpGs are can be described in terms of their location relative to genes and
other genomic features, such as transcription start sites (TSS) or gene enhancers. These classifications prove useful for in silico analyses of DNAm patterns and inferring downstream effects.

1.5.1 DNA methylation involvement in gene regulation

Nearly 70% of gene promoters are near a CGI, and this proximity results in a strong relationship between promoter DNAm and gene expression (76,82). Generally, high levels of DNAm at promoter-associated CGIs are associated with low levels of expression, and vice versa (83). However, the association of DNAm and gene activity is complex. A negative correlation between DNAm levels and gene expression is not upheld when comparing expression and DNAm levels for a specific gene across individuals (84-87). As well, recent findings suggest the causation may be reversed and gene expression levels influence levels of DNAm, not vice versa (85). Furthermore, DNAm marks found outside promoters show less consistent associations with gene expression (85). For example, in CGI shores, higher levels of DNAm have been associated with higher levels of gene expression (63,88).

Detailed molecular analysis has provided a likely framework through which DNAm can inhibit gene expression via the recruitment of methylated DNA-binding proteins, which in turn recruit chromatin modifying enzymes to package the DNA such that it is inaccessible to transcription machinery (63). While this relationship between DNAm and gene transcription is supported by a large body of research, whether DNAm is a causal mark or acts in maintaining other repressive signals is currently unknown. Recently, emerging evidence suggested, via the artificial induction of DNAm at gene promoters, that DNAm alone cannot repress transcription (89). However, reanalysis of these data suggested that forcible DNAm does repress transcription
and is associated with reduced H3K4me3, a histone mark associated with active transcription (89,90). Such conflicting findings are representative of our current understanding of the specific role of DNAm in gene regulation.

1.5.2 Genetic background contributes to DNA methylation variation

The underlying DNA sequence can influence DNAm. Reminiscent of work examining expression quantitative trait loci, sites of linkage between allelic variation and gene expression, multiple studies have identified SNPs associated DNAm at specific CpGs, known as methylation quantitative trait loci (mQTLs) (83,85). This relationship was originally uncovered investigating differences in DNAm between ethnicities, as well as higher than expected heredity of DNAm patterns (83,91,92). CpGs under genetic influence are estimated to make up anywhere from 20 to 80% of all CpGs (83,93,94). One mechanism through which genetic variation alters DNAm is whereby a SNP disrupts the specific sequence recognized by a DNA binding protein responsible for creating boundaries between methylated and unmethylated genomic regions (85). As variation at this SNP may be associated with methylation patterns at certain CpGs, it would constitute an mQTL. With respect to the interplay between SNPs, gene expression and DNAm, SNPs have been shown to most commonly alter methylation and expression independently, as opposed to a SNP altering gene expression first, followed by gene expression affecting DNAm or a SNP altering DNAm first, followed by DNAm affecting gene expression (85). These interactions may be mediated by transcription factor binding, which alters gene expression and DNAm simultaneously. Regardless of the mechanism by which these relationships function, it is crucial to consider genetic variation in the analysis of DNAm.
1.5.3 **Tissue specific DNA methylation and cellular differentiation**

A well understood role of DNAm occurs during embryonic and fetal development, wherein it regulates cell differentiation, conferring tissue-specific identity that is stable and mitotically heritable (67). As such, DNAm displays tissue- and cell-specific patterns (95-97). In fact, tissue of origin is one of the largest determinants of DNAm variation in healthy individuals, accounting for greater variation than genetic background (96-99). Much of what is known about the role of epigenetic mechanisms during embryogenesis has been elucidated through the study of mouse embryonic stem cells (ESCs). During differentiation, DNAm is required to silence pluripotency factors. For example, the promoters of genes associated with pluripotency, such as *Oct4* and *Nanog* are hypermethylated and silenced (67). As well, DNAm acts to upregulate markers associated with germ-layer specificity; in ESCs lacking DNAm, differentiation is inhibited.

Further knowledge of the role of DNAm in tissue specificity has been gleaned from studying the human hematopoietic system. During hematopoietic stem cell differentiation, DNAm gains and losses at specific regions of the genome are believed to maintain the differentiation marks required for cellular identity (100). Myeloerythroid and lymphoid lineages display lineage-specific epigenetic differences, which further develop as the cells mature (101-103).

1.6 **DNA methylation over the life course**

DNAm is most dynamic during fetal development, when epigenetic patterns play an integral part in the complex processes of embryogenesis (64,104). Highlighting the dynamic nature of DNAm, levels in neonatal blood are lower than that observed at most other points
during the lifespan (105,106). After birth, average DNAm levels increase in blood throughout the first year of life (105,107). These changes occur preferentially at CpG island shores and shelves, enhancers, and promoters lacking CpG islands (108). In both blood and buccal epithelial cells, DNAm between monozygotic twins has been shown to become more divergent in the first year of life (105,109). After the first year, median global DNAm levels are relatively stable, with the exception of certain regions that frequently gain DNAm (105). The first few years of life have been extensively studied; however, there are relatively few reports of changes in DNAm throughout later childhood and adolescence. Studies that examined this period of human development have reported that DNAm levels increase rapidly and then stabilize by adulthood in both brain and blood (65,110).

DNAm in respect to the aging process, from adulthood to advance age, has also been extensively studied. Overall levels of DNAm remain stable in blood, while inter-individual variability increases over that time (111,112). Post adulthood, many studies have found a mean decrease in blood DNAm with increasing age (112-119). These changes are less likely to occur in promoters and more likely to be observed in enhancers (118). Regions that gain DNAm with age are enriched for CGI, while non-CGIs tend to lose DNAm with age (112,115,116,119,120). In sum, sites that show overall low DNAm, such as promoter-associated CGIs, tend to increase methylation with age, while those with high DNAm such as intergenic non-island CpGs tend to lose methylation with age. As most CpGs are located outside of CGIs and are highly methylated, this translates to an overall loss of DNAm in later life as well as a tendency for DNAm levels to shift toward the global mean DNAm with increased age (76,77,82,112,117,121). Taken together, these findings suggest that DNAm shows reduced stringency in maintenance over the lifespan.
Despite a gain in DNAm in early life and gradual loss in later life across the genome, these changes are not symmetrical. They differ in two major ways: (i) the rate of change is much higher in early life than later life, and (ii) the genomic locations of the changes are quite different. In early life, DNAm is gained globally, but more at island shores and intergenic regions, while in later life, DNAm is lost globally, but still gained at islands and shores (108,110,122). Overall, most tissues fit with the pattern of increase in average DNAm early in life, with a gradual decrease later in life (123,124). For example, many studies have shown that brain regions follow this pattern, showing rapid changes in DNAm in the early life period and then slowing gradually over the lifespan (65,116,125).

In addition to general patterns of DNAm change with age, DNAm levels at specific sites in the genome, which are highly associated with age, can be used to predict chronological age (112,116,117,119,126). These sites underlie the concept of the ‘epigenetic clock’. These epigenetic clock sites have been found both within a specific tissue and across tissues, and have been shown to be much more concordant across tissues than gene expression changes across tissues with age (112,116,117,119).

1.7 Evidence of DNA methylation alterations by SES-related early environmental exposures

Complementing its role in cell type specification, DNAm is also emerging as a mechanism of cellular memory of past exposures (61). Although DNAm is mitotically stable, transmitted from mother to daughter cells, across an individual’s life time it may change following specific environmental exposures (127). The paradigm of environmentally-induced DNAm alterations has recently been applied to early life SES (45,128). In the following sections
I will explore the current findings in both mouse models of early life adversity and human research.

1.7.1 Animal studies of epigenetics and early exposures

Numerous animal studies established causal links between stress exposures altering epigenetic markers and subsequent health or behavioural outcomes. While SES itself cannot be modelled in animal studies, chronic or acute stress exposure represents one of the most powerful aspects of SES-related early life experiences. To that end, low levels of maternal licking/grooming and arch-back nursing behaviour have been associated with decreased DNAm at the glucocorticoid receptor (GR) gene in cross-fostered offspring (129). Outcomes to the offspring also included decreased expression of the GR gene, Nr3c1, and increased HPA axis responsivity and anxious behaviour. Moreover, the molecular signatures and stress response phenotype were reversible in adulthood with central infusion (into brain ventricles) of a histone deacetylase inhibitor. The study of maternal care in rodents has also provided evidence of DNAm moderating the expression of estrogen receptor alpha, Esr1, in the hypothalamus (130). Offspring raised by low licking/grooming dams had decreased expression the Esr1 gene and reduced neural activation; this caused mature female offspring to provide less licking/grooming to their own offspring. Gene expression was associated with differential DNAm promotor region of Esr1 (130). Together these findings provide strong causal evidence of the role epigenetics in translating early life experiences into long-term physical outcomes.

Beyond nursing behaviour, many different animal models of early life stress, such social isolation, exposure to aggressive social interactions, and environmental manipulations, have been used to investigate the role of epigenetics as a biological mechanism of the long-term effects
Early postnatal stress in mice lead to a range of behavioural outcomes, including increased stress reactivity and addiction, altered gene expression profile, and DNAm changes depending on the stress paradigm and its duration (36). For example, stress induced by infant-mother-separation in mice resulted in hyperactivity of the HPA axis, specifically hypersecretion corticosterone, increased arginine vasopressin (Avp) expression in the hypothalamus and loss of DNAm at an enhancer of *Avp* (131). Importantly, AVP functions in stress response, namely through stimulatory effects on the HPA axis. Behavioural changes, loss of DNAm and altered *Avp* expression all persisted beyond the 10 day exposure to maternal separation (131). Finally, differential susceptibility in response to environmental stress has also been modeled in mice. One study found that adult mice exposed to 10 days of a standard “social defeat” protocol used induce stress to model stress-related disorders such as anxiety, displayed either behavioural susceptibility or resiliency, as determined by subsequently avoiding or interacting with unfamiliar mice, respectively (132). Only susceptible mice displayed loss of DNAm and increased expression at the *Crh* gene. Like AVP, CRH (corticotrophin releasing hormone) is a neuropeptide involved in the HPA axis. Animal models have proven especially valuable in understanding the mechanisms of social epigenetics and have causally implicated DNAm in “biological embedding” of early life stress by allowing for the longitudinal assessment of DNAm.

### 1.7.2 Human studies of epigenetics and early exposures

In addition to extant research in animal models, human research has offered evidence of the role of environmental exposures in DNAm patterns through association studies. By necessity, most research on humans is correlational rather than causal and many uses peripheral tissue to
assess DNAm, as postmortem brain samples are limited and many human tissues require invasive sampling methods to obtain. Currently, the most widely reproduced association between DNAm and an environmental exposure is that of cigarette smoke exposure and DNAm in the aryl hydrocarbon receptor (\textit{AHRR}) gene (133-136). \textit{AHRR} is involved in the aryl hydrocarbon receptor signaling pathway which responds to toxic compounds in polluted air and tobacco smoke. Decreased DNAm in the gene body of \textit{AHRR}, particularly at cg05575921, has reproducibly correlated with smoke exposure in many independent cohorts, including current smokers, individuals exposed to secondhand cigarette smoke, and infants exposed prenatally to maternal smoking, in tissues including lymphoblasts, PMBCs, monocytes, and cord blood (133-136).

The same methodologies used in EWASs of physical environmental exposures have been applied to early life stress and familial SES. Such studies have strengthened and expanded our understanding of SES as a social determinant of health. As previously described, the negative health effects of SES are better explained by childhood SES, as compared to adulthood SES; this paradigm has been recently observed in DNAm. In a cohort of adults from either low or high SES backgrounds, based on parents’ job prestige in early life, who were balanced for current SES, DNAm was associated with early life SES irrespective of concurrent SES; concurrent SES was not associated with DNAm (87). Furthermore, pre-stimulation DNAm differences in leukocytes were predictive their cytokine responses when stimulated through the Toll-like receptor pathways (87). As such, one could speculate that this specific DNAm pattern connects past SES exposures to later immune response, thus connecting early life experiences to health outcomes in adulthood. DNAm research has also broadened our understanding of differential susceptibility to early life stress and adversity.
Individuals with a “sensitivity allele” in the *FKBP5* gene, which causes increased induction of the gene, and who experienced childhood trauma were found to have loss of DNA methylation (DNAm) at *FKBP5* and an increased risk of developing stress-related psychiatric disorders in adulthood, including post-traumatic stress disorder, as compared to individuals with childhood trauma but no copies of the sensitivity allele (137). *FKBP5* is a regulator of the stress hormone system and demethylation at this gene is followed by increased stress-dependent gene transcription and ultimately by a long-term dysregulation of the stress hormone system and a global effect on areas of the brain associated with stress regulation(138). Notably, this study also implicates dysregulation of stress response pathways as a mechanism by which adversity can lead to disease. Beyond these studies, a wide range of research on DNA methylation and exposures to early life psychosocial adversity or SES exists. DNA methylation appears to be associated with being reared in institutional environments (139,140), childhood maltreatment or deprivation (141-143); maternal mental health problems during the perinatal period (144,145) and parental adversity and stress in either infancy or the preschool (146,147).

The take home message from our current understanding of social epigenetics is that complex social experiences, such as early life SES, are intertwined with physical exposures and subjective experience and any or all of these factors may play a role in DNA methylation (DNAm) alterations. Furthermore, the causation and directionality of this relationship is not interpretable in cross-sectional studies and outcomes are complicated by individuals’ genetic backgrounds and temperaments, dictating their differential susceptibility. This is underscored by a genome-wide study of genotype, DNA methylation and various *in utero* exposures, including maternal smoking, maternal depression, maternal BMI and birth order, in which the vast majority of high variability DNA methylation
sites of were best modelled by an interaction of genotype and exposure, as compared to genotype alone or exposure alone (148).

1.8 Synthesis and thesis objectives

Epigenetic processes, such as DNAm, represent a likely mechanism of biological embedding that can extend to any biological system. Hypothetically, epigenetic mechanisms could enable the body to mount conditionally adaptive changes to metabolic processes, the endocrine system, neuroregulatory function, etc. depending on the specific environmental onslaught and the tissue type in which the epigenetic change occurs. Therefore, assaying DNAm in the context of “early life” environment can bring to light aberrations underlying development, learning and behavioural differences, and later health outcomes. In the context of this dissertation, the term “early life” is used broadly to refer to a period of time starting at conception and ending at puberty. The relationship between early life experience and epigenetics is complicated by internal psychological and physiological factors, as well as genetic variation. As well, stress reactivity and temperament are predictive of how a child may interact with his or her environment and can affect how such exposures are internalized.

In sum, the early life environmental exposure and intrinsic, behavioural differences may imprint on the DNA methylome at genes related to stress response pathways and/or neurodevelopment, thus underlying the gradient of health outcomes associated with one’s childhood socioeconomic status. Focusing my research around this hypothesis, the goal of this dissertation was to build foundational knowledge of molecular mechanisms of “biological embedding” to ultimately inform targeted, multilevel interventions capable of abating lifelong societal inequalities in health and human development. Given early state social epigenetics
research, many obstacles exist limiting the interpretability and generalizability of SES-related EWAS findings. For examples, epigenetic studies often use accessible, surrogate tissues in place of primary tissues of interest. Given the vast differences in DNAm patterns between cell types and tissues, the functionality of epigenetic findings in surrogate tissue are not fully understood. Thus, the main objectives of my dissertation were to elucidate the relationships between childhood environment, DNAm and genetic variation, and behaviour, to understand how these systems influence one another but also to assess the utility of such findings. These objectives were divided into the following three research projects:

I. Evaluate the concordance of DNAm in matched blood and buccal tissues from the pediatric age range to determine benefits and disadvantages of using these accessible tissues, as well as uncover tissue-specific DNAm patterns.

II. Examine associations of early, internal differences in biobehavioural responses to later epigenetic modifications, expecting to uncover significant relationships between biobehavioural response propensities and patterns of DNAm, similar to previously documented associations between early life environments and DNAm.

III. Examine the epigenetic correlates of early-life socioeconomic status in PBMCs and BECs collected at ages 8-10 and dried blood spots collected at birth to understand the tissue-specificity and timing of SES-related DNAm patterns.
Chapter 2: Common methods and materials

2.1 Introduction to epigenome-wide association studies

Epigenome-wide association studies (EWASs), aim to discover DNA methylation loci or CpGs that associate with a phenotype or exposure of interest. EWASs, like genome-wide association studies (GWASs), can discover associations with no *a priori* hypotheses of the genes involved (124,149,150). As well, EWASs cannot specify causal relationships and therefore may uncover DNAm patterns which are mechanistically involved or biomarkers(94,149,150).

Unfortunately, due to multiple test correction and small samples sizes, EWASs can be plagued by high type I error rate (false negatives). Furthermore, failure to account for technical confounders, population complexity, and other confounders in the data leads to a large number of type II errors (false positive results). Unlike GWASs, power calculations, aimed at estimating appropriate sample sizes, are not common. A small number of publications have provide suggestions and guidelines for applying power calculations to EWASs (151,152). However, calculating cohort-based EWASs have proven challenging due to the difficulties in predicting the required inputs for power calculations, such as predicted effect size and number of significantly associated CpGs. Running under-powered EWASs has most likely contributed to the infrequency of replicability of findings across independent cohorts due to high type I error rates. With the steep rise in EWAS publications, methods being employed to ensure the quality of reported EWAS findings limit type I and II error rates include but are not limited to batch correction, cell-type correction, accounting for ethnicity or genetic background, and independent replication (94,149,150,152). In this chapter, I outline the common methods and materials used in chapters 3 to 5 of this dissertation.
2.2 Study samples

Two main cohorts were analyzed in this dissertation - Wisconsin Study of Families and Work (WSFW) and Gene Expression Collaborative Kids Only (GECKO). WSFW samples are analyzed in chapter 4 and GECKO is used in chapters 3 and 5. In chapter 3, the C3ARE cohort is also assessed; all processing and analyses of the C3ARE cohort were performed by S. Islam and is not described here. Please see chapter 3 for sample information on the C3ARE cohort (147,153). All experimental procedures were conducted in accordance to institutional review board policies at the University of British Columbia (Certificates H07-01317 (C3ARE) and H07-02773 (GECKO, WSFW)). Written informed consent was obtained from a parent or legal guardian and assent was obtained from each child before study participation.

2.2.1 Wisconsin Study of Family and Work

Wisconsin Study of Families and Work (WSFW) is a longitudinal study that was originally established to examine the effects of maternity leave on families. Dr. Marilyn J. Essex was the principal investigator of the project, which began in 1994 after concluding an initial phase, for which Dr. Janet S. Hyde was the principal investigator (154).

A total of 570 pregnant women, 18 years or older and living with the biological father, were recruited for participation from their physicians’ clinics (154). Seventy eight percent of the sample came from the Milwaukee, Wisconsin and twenty-two percent from the Madison, Wisconsin. Participants were required to be between 21 and 25 weeks pregnant during the recruitment process, which occurred between June 1990 and September 1991. As well, the
mother was required to be employed or a homemaker; if the latter was reported, the father was required to be employed.

Of the sample of 570 women, 93% self-identified as “White” and 7% reported a minority ethnic heritage (Indian/Alaskan, Asian/Pacific Islands, African-American, Hispanic, or White Hispanic); 223 were pregnant with their first child, 95% were married. The median age of mothers was 29 (20-43); the median educational experience was college graduating (with experience ranging from no completion of high school to post-graduate degree); the median family income was $45,000 USD (range $7,500 to $200,000 USD).

Following birth, participants partook in 15 assessments over the course of 18 years, which included mail-out interviews, in-person interviews, questionnaires, or home visits to collect observational data. For certain assessments, fathers, care-givers, teachers and children all participated in interviews and questionnaires. The analyses presented here are based on a subsample of children, referred to as the “MacArthur 120”. During grade 1 (in 1998), these 120 children (73 girls, 47 boys), their parents and teachers completed additional assessments, for the purpose of evaluating the MacArthur Assessment Battery for Middle Childhood (155). These children were recruited based on scoring either high (upper 20% percentile on internalizing and/or externalizing symptoms) or low on mental health symptomology in the MacArthur Health and Behavior Questionnaire (HBQ) completed in kindergarten. Therefore, unlike the greater sample, the MacArthur 120 was designed to be balanced for high and low pre-syndromal mental health symptoms. The subsample did not differ significantly from the larger sample in family income, parental education, or marital status.
Buccal swabs were collected at ages 15 and 18 using the Isohelix Buccal Swabs (Cell Projects Ltd., Kent, UK) and stabilized with Isohelix Dri-Capsules for storage at -80° C prior to DNA extraction (156).

2.2.2 Gene Expression Collaborative Kids Only

GECKO is a cross-sectional study of epigenetic profiles, SES and family adversity in approximately 400 children living in socioeconomically diverse neighborhoods from the Vancouver lower mainland. This project was designed to observe how neurological and genomic factors interact with early life experiences to alter developmental outcomes. It combines population level data regarding neighborhood-level socioeconomic and school performance/educational information, and individual health, genetics, epigenetics and stress reactivity in order to generate a “whole systems” understanding of the early life experience. Dr. W. Thomas Boyce was the principal investigator and Drs. Michael Kobor, Tim Oberlander and Jelena Obradovic were co-investigators.

The initial round of data collection took place between December 2009 and November 2011. During this time a travelling van, i.e. a mobile laboratory, outfitted with all required instrumentation travelled to the participants houses to perform approximately two hours of laboratory stress tests. Specifically, electroencephalography (EEG) and echocardiography ECG measures, were recorded, as well as Galvonic skin response and respiratory rate. These measures assessed autonomic nervous system (ANS) reactivity and prefrontal cortex (PFC) activity during the stress protocol. Saliva was collected at 4 time points across the assessment for the purpose of measuring adreno-cortisol levels. The protocol itself involved answering questionnaires given by
the research associate, following by a visual oddball detection task, an executive functioning task and the Trier Social Stress Task.

During the protocol, parents answered questionnaires aimed at assessing socioemotional vulnerability, adversity, financial stress, conflict tactics, life events, child health and development, education and household income.

Participants provided buccal swabs and saliva for DNA analysis as part of the mobile lab protocols; the latter was used only for genotyping and never for DNAm analysis. Buccal samples were preserved, as per Isohelix protocol, with proteinase k and “Is” solutions; the samples could then be stored at room temperature without DNA degradation. Saliva samples were preserved using DNA Genotek Oragene collection kits. For a subset of participants, peripheral whole blood was drawn between 6 months and 2 years after the initial visit using Vacutainer® CPT™ Cell Preparation Tubes (Becton, Dickinson and Company, NJ, USA) and PBMCs were isolated following centrifugation, washing and resuspension into R10 media (Sigma-Aldrich, MO, USA), as previously described (35). PBMC pellets were frozen and stored at -80°C until DNA extraction. As well, in subset of participants, I obtained neonatal dried blood spot collected from heel pricks using Whatman Protein Saver cards by the BC Newborn Screening Program.

2.3 DNA extraction and sodium bisulfite conversion

Genomic DNA from stabilized buccal samples was isolated using Isohelix Buccal DNA Isolation Kits (Cell Projects Ltd., Kent, UK) and was purified and concentrated using DNA Clean & Concentrator (Zymo Research, CA, USA). Genomic DNA was extracted from PBMC pellets using the DNeasy kit (Qiagen, MD, USA). DNA yield and purity were assessed using a Nanodrop ND-1000 (Thermo Fisher Scientific, MA, USA). Genomic DNA was extracted from
DBSs using the GenSolve DNA Recovery Kit, GVR100 (IntegenX), followed by purification using the QiAmp DNA Micro Kit (Qiagen). Finally, genomic DBS DNA is concentrated using Microcon DNA Fast Flow Devices, MRCF0R100 (Millipore).

Bisulfite conversion of all genomic DNA (750 ng) was performed using the Zymo Research EZ DNA Methylation Kit (Zymo Research, CA, USA), as per manufacturer’s instructions. Bisulfite conversion translates DNA methylation into sequence differences by a deamination process. Unmethylated cytosines become uracils, and then thymines following amplification, while methylated cytosines are protected from deamination and remain thymine (157). Thus, percent methylation (ranging from 0 to 1) of a single CpG in a cell population may be quantitated by methods similar to those used for DNA polymorphisms.

2.4 Illumina DNA methylation microarrays

There are many technologies available to assay genome-wide DNAm, including microarrays and next-generation sequencing (158), which are typically performed after sodium bisulfite treatment of genomic DNA. However, Illumina DNAm microarrays are one of the most commonly used in human DNAm EWASs (149,150,159). Advantages of this technology include low cost per sample, single-nucleotide resolution and high reproducibility (80,160). Disadvantages of DNAm microarrays include low coverage of the methylome and batch effects. Over the last 10 years, the DNAm microarrays offered by Illumina have greatly improved coverage. In this dissertation, I used two iterations of this Illumina’s microarray, Infinium HumanMethylation27 BeadChip (27K array) and Infinium HumanMethylation450 BeadChip (450K array), based on what was commercially available at the time (80,161). Both assays required the same general workflow(161):
1. 160ng of each bisulfite converted DNA sample was denatured and neutralized.
2. Samples were isothermally amplified overnight, resulting in whole genome amplification.
3. Amplified sample were fragmented, then precipitated and resuspended in preparation for hybridization.
4. Samples were hybridized to microarray beadchips and incubated overnight.
5. Beadchips were washed of any unhybridized DNA.
6. Beadchips underwent single base pair extension of probes using hybridized DNA as template and were then stained with fluorescent labels.
7. Beadchips were scanned, producing IDAT files consisting of light intensity reading that could be converted to beta values.

2.4.1 Illumina Infinium HumanMethylation27 BeadChip

Genomic BEC DNA collected at age 15 from WSFW participants was assayed using the 27K array. This microarray platform covers 27,578 CpGs across 13,500 promoters, as well as some gene bodies. The 27K array contains only Infinium type I probes (161). For this technology, two unique probes exist for each targeted locus. One probe targets methylated DNA and the other targets unmethylated DNA.

2.4.2 Illumina Infinium HumanMethylation450 BeadChip

Genomic BEC DNA collected at age 18 from WSFW participants, as well as all DNA collected from the GECKO cohort was run on the 450K array. The 450K array contains over 485,000 methylation sites, covering over 99% of RefSeq genes (80). Unlike the 27k array, probes are distributed across the promoter, 5'UTR, first exon, gene body, and 3'UTR. In addition
to covering 96% of CpG islands, island shores and the regions flanking them are also covered. The 450k array contains two types of probes. Infinium type I probes, described above, are used on the 450k array for high CpG density regions. The majority of probes, however, are Infinium type II probes (80). Type II probes differ by using a single probe for both methylated and unmethylated alleles of the CpG. Following hybridization of the DNA, a single base extension using auto-fluorescing nucleotides is used to determine the methylation status of the CpG. Type I probes use an all-or-nothing assumption, i.e. all CpGs within the sequence of methylated probes are assumed to be methylated and all CpGs within the sequence of unmethylated probes are assumed to be unmethylated. However, type II probes are degenerate, containing all possible combinations of methylated and unmethylated underlying CpGs.

2.5 DNA methylation data processing pipeline

Prior to analysis of the DNA methylation array data, it must be assessed for quality and data normalized to remove background noise and technical effects (162). The processing pipeline is not standardized and there are steps which can be included or excluded depending on the study design. As well, different methods were used to normalize the 27k array data and the 450k array data that were based on the standards in the field at the time of processing. As such, each data chapter contains a detailed methods section of the processing and normalization methods used. The general pipeline one follows for processing 450K array data sets is outlined below.

2.5.1 Prepossessing and quality control

Raw intensity values from the DNA methylation arrays were imported into Illumina GenomeStudio V2011.1 software and subjected to initial quality control checks for array
staining, extension and bisulfite conversion followed by color correction and background adjustment using control probes contained in the 450K array(163). The data were then exported from GenomeStudio as beta values (β) which represent the estimated DNAm level based on a ratio of intensities between methylated and unmethylated alleles, with beta values ranging from 0 (unmethylated) to 1 (fully methylated). DNAm was also reported as a percent, which is calculated by multiplying the beta value by 100. Subsequent processing and analyses were performed in R (http://www.r-project.org).

Sample filtering is performed using multiple methods. Each sample was visually assessed for a typical bimodal distribution of methylation values; clustered by sex probes to check for agreement with assigned sex and confirmed that less than 5% of probes perform poorly (metrics described below) (163). Failing any of these quality control measures merits removal of sample. Additionally, R functions, such as removeOutlier, were implemented to identify and remove outlying samples based on sample clustering(164).

Following sample filtering, probe filtering was performed based on two metrics: 1) probes with detection p-values greater than 0.01, indicating that the signal is not significantly greater than background signal, were removed, and 2) probes with missing beta values, for which less than three beads contributed to the signal, were removed (163). In smaller cohorts, if a probe fails either metric in a single sample, this merited removal; in larger cohorts, the probe must fail in a given percentage of samples to be removed. For example, in a cohort of 100 samples, the limit may be set at 5%, in which case a probe must fail either metric in more than 5 samples in order to be removed.

Additional probe removal was based on re-annotation of the 450K array (165). Polymorphic probes, probes at which the assayed CpG is polymorphic, were filtered out of the
dataset, as well as probes that have nonspecific in silico binding to the sex chromosomes were also removed (165). Finally, 65 control SNP probes were removed and in cohorts containing both male and female participants, probes mapping to the sex chromosomes were removed, as males and females cannot be directly compared and to limit the confounding effects of sex.

2.5.2 Normalization

Following preprocessing and quality control of the array data, normalization is required to minimize or remove background noise, as well as variability attributed to technical differences. One notable contributor to noise, is the two different technologies used on the 450K array uses, type I and type II probes (166). In the cohorts presented in this dissertation, two normalization methods were used, Beta Mixture Quantile dilation (BMIQ) on GECKO and C3ARE and Subset Within-Array Normalization (SWAN) on WSFW age 18 data (166,167). The WSFW age 15 data, run on the 27k array were quantile normalized (168). Details of these normalization methods are provided in the data chapters to which they are pertain.

2.5.3 Cell-type correction of 450K DNA methylation data

Following normalization, cell-type correction may be applied to the dataset if DNAm is measured in a heterogeneous tissue, such as PBMCs or BECs. The latter is composed of BECs and leukocytes from saliva. In this dissertation, PBMC cell type proportions were estimated using a standard blood deconvolution method (169,170). No published deconvolution method exists for BECs. To account for buccal epithelial cell versus white blood cell proportions in BECS, in chapters 3 and 5, I used saliva-based deconvolution method (171). This method was used because saliva and BEC samples are composed of the same cells in different proportions.
and it also allowed me to investigate potential blood contamination (171). Both methods output estimated cell proportions for each sample, which were used to normalize cellular heterogeneity across individuals using a linear regression (172).

### 2.5.4 Batch correction

Finally, a batch correction technique, such as ComBat, can be employed to adjust for variation attributed batch effects (173). Common causes of batch effects in DNAm arrays can include:

1. **Plate** – samples are randomized onto 96 well plates. On these plates, samples undergo genome-wide amplification, fragmentation, precipitation and resuspension and are then pipetted onto the microarray chips.

2. **Chip** – each 450K array chip contains 12 samples. The DNA is hybridized to the beadchip overnight and then undergoes single base-pair extension and staining and imaging.

3. **Chip row** – On the chip, samples are arranged in two columns of six rows. During the extension and staining steps, reagents are pipetted onto the top of the chip and allowed to flow down.

In the ComBat package, the user can input which “batch” he/she wants removed as well any covariates of interest which should be protected. However, if samples are unbalanced, which is more likely to occur in small sample sizes, Combat is likely not appropriate and batches must be accounted for using a different method (174). For this reason, Combat was not used in all data chapters.
2.6 Data analysis

Prior to analysis, beta values were log-transformed to M-values, which are less biologically intuitive but are more statistically valid for differential analysis because transformation minimizes heteroscedasticity (163, 175). EWAS analyses do not follow pre-set pipeline, but rather depend upon the specific research questions. Therefore, different statistical tests were employed in different data chapters. Here, I outline common techniques used throughout this dissertation.

2.6.1 Data reduction

In some instances, not all CpGs assayed on a microarray are pertinent to the planned analysis; for example, sex chromosomes may be removed in cohorts containing males and females, as they are not directly comparable between sexes and therefore any stratification of the data by sex can confound interpretation. Data reduction techniques based on inter-individual variability of CpGs allows for a reduction in the multiple test burden, as well as the removal of CpGs that are not of biological interest (152). Therefore, I calculated inter-individual variability of each CpG, based on the range between the 10th percentile and 90th percentile beta values for each site (176). I chose to use this calculation as it captures variability across the bulk of samples while being robust to outliers (176). This calculation does not rely on the assumption that CpGs are normally distributed.

2.6.2 Statistical analysis of differential DNAm

In this dissertation, differential DNAm was assessed using a number of statistical tests. Choice of test was dependent on the study design and the available meta data. Due to the
relationship between genetic variation and DNAm variation, ethnicity/ancestry and relatedness should be considered in DNAm analyses (91,177). For all cohorts, ethnicity/population structure was estimated by the Illumina PsychChip array, an exome genotyping array. Sex and age are also commonly, but not always, included as covariates in EWAS analyses (149). For example, in chapter 4, I used a simple correlation between the variable of interest and DNAm at each variable CpG, while in chapter 5, a linear regression was applied to the data (178). In chapter 4, phenotypic variables, including ethnicity, age, and SES were either independent of the explanatory variable or homogenous across the cohort, meriting a correlation. By comparison, both ethnicity and sex were associated with the variable of interest in the GECKO cohort and therefore, were included as covariates in a linear regression. Finally, the tests performed in Chapter 3 were paired because they were performed on matched tissues. For example, to test differential methylation between two matched PBMCs tissues, a Wilcoxon signed-rank test was performed.

When reporting significance either nominal p-values were reported or false discovery rate (FDR) corrected p-values. Multiple test correction was performed using the Benjamini-Hochberg method, which controls FDR (179). This method is less stringent than a method such as Bonferroni correction which controls the familywise error rate. As well as using a significance threshold, associated CpGs are also filtered using a biological threshold. Across this dissertation, CpGs were required to meet a minimum effect size, specifically an absolute delta beta > 5% (|Δβ| > 5%). The |Δβ|, or effect size, was calculated for ordinal or continuous variables using a linear model to estimate the difference in beta values across the variable of interest. Specifically, a regression line is fit to the data at each CpG and change in DNAm across the regression line is
calculated. For dichotomous variables, $|\Delta \beta|$ is calculated as the difference in mean DNAm between the two levels.
Chapter 3: Integration of DNA methylation patterns and genetic variation in human pediatric tissues help inform EWAS design and interpretation

3.1 Introduction

Epigenome-wide association studies (EWASs) are becoming increasingly popular, in part due to their potential to enhance our understanding of the determinants of health and disease, including potential early life embedding of experiences and exposures and their association with later life outcomes (94,150,180-184). The term ‘epigenetics’, describes mitotically heritable modifications of DNA and its regulatory components, including chromatin and non-coding RNA, that potentially modulate cellular states or fate through gene expression changes, without changing the DNA sequence itself (61,185,186). DNA methylation (DNAm), which involves the covalent attachment of a methyl group to a cytosine primarily at cytosine-phosphate-guanine (CpG) dinucleotides, is the most well studied chromatin mark in human populations due to its relative stability and ease of measurement on quantitative array-based methods (187,188). To date, EWASs have identified differential DNAm across a broad range of contexts including disease states, genetic background and environmental exposures, thereby providing evidence for the potential contribution of DNAm in mediating gene-by-environment (GxE) interactions (94,189,190).

Given that tissue specificity is an integral feature of epigenetic profiles, as different tissues and cell types acquire distinct epigenomes during differentiation, the selection of tissue source is a key consideration in the careful design and interpretation of EWAS analyses (95,191,192). The collection of a disease-relevant, target tissue allows for the direct assessment
of epigenetic associations that may be implicated in the underlying phenotypic or disease biology. In certain cases, readily accessible peripheral tissues may represent the target tissue; for example, use of PBMCs for the investigation of DNAm associations to immune or inflammatory phenotypes (99,193-195). However, in many cases, the target tissue, such as brain, muscle, adipose tissue, among others, may be impossible or very difficult to collect from living individuals or at sufficient quality for analysis from postmortem samples (181). Easily accessible peripheral tissues are therefore often used in human epigenetic studies for biomarker discovery in lieu of target tissues that are difficult to collect. This is particularly relevant to pediatric cohorts in which biopsy specimens with invasive collection procedures or postmortem samples are less common than in adult populations. As such, more readily accessible tissues with minimally invasive collection procedures, such as cord blood, saliva, buccal epithelium cells (BECs) or peripheral blood mononuclear cells (PBMCs), are widely used tissue source materials for early life EWASs. The use of pediatric tissues in DNAm analyses is further complicated by the fact that widespread alterations occur in tissue-specific DNAm patterns during development, therefore conferring additional complexity in the selection of appropriate source material for early life DNAm studies (96,196). Furthermore, changes in cell composition within a tissue are a source of potential confound in EWAS, as shown for a number of DNAm associations, including changes during development and certain environmental exposures such as smoking (94,197-201).

Currently, two major focal points in human epigenetic research are to elucidate the tissue specificity of DNAm patterns with respect to individual CpGs as well as assess inter-individual variation within a single tissue (96,202-204). At a population level, a number of studies have examined the concordance of DNAm patterns across multiple tissues (99,203-208). Findings
have shown that beyond tissue-specific differences in absolute DNAm measures, inter-individual DNAm variability also varies by tissue type (99,205). For example, previous work by our group has shown that BECs have greater DNAm variability over matched PBMCs at both the genome-wide level and at individual CpGs (99). Moreover, CpG sites with higher DNAm variability tend to be more correlated between matched tissues (203-205,208). Although these results provide important insights into the comparability of DNAm measures across matched tissues, the analyses to date have been conducted in adult tissues, thereby limiting their relevance to DNAm profiles from pediatric samples. As previous studies have demonstrated that developmental changes in blood DNAm patterns tend to be more pronounced and occur more rapidly in childhood, the examination of DNAm concordance and variability in pediatric tissues represents an important and currently missing step in our understanding of EWAS associations from pediatric peripheral tissues (96,196).

Genetic variation represents an additional contributor to DNAm patterns in tissues, with genetic influences accounting for nearly 20-80% of DNAm variance within a tissue (83,93,209-212). Methylation quantitative trait loci (mQTLs), sites at which DNAm is associated with genetic variation, are present across the genome and are often consistent across tissues, ancestral populations and developmental stage (92,177,213,214). Notably, genetically influenced sites of inter-individual DNAm variation, which can co-occur across tissues, may be biologically informative. For example, allele-specific DNAm of the FK605 binding protein 5 (FKBP5) gene, which has been associated with risk of developing stress-related psychiatric disorders, responds to glucocorticoid stimulation in a similar way in peripheral blood cells and neuronal progenitor cells (215). Within a particular tissue, such as blood, mQTLs often are stable across development (177,216). Moreover, approximately 75% of the inter-individual regional DNAm variance within
a single tissue can be best described by GxE models (148). As such, delineating the contribution of genetic influences to tissue-specific DNAm may help clarify the interpretation of EWAS associations.

Given that early life development brings about sizable changes to DNAm patterns, it is important to examine DNAm variability and concordance between peripheral tissues, as well as genetic influences on early life DNAm patterns, in childhood (96,110). To this end, we used matched PBMC and BEC samples, two commonly used peripheral tissues in EWAS, from two independent early life cohorts in order to identify a) differences in inter-individual variability and concordance of DNAm between these tissues and b) genetic contributions to these patterns at the site-specific level. Our results showed that genome-wide DNAm variability differed between tissues, with BECs exhibiting greater inter-individual DNAm variability over PBMCs. Moreover, we found that highly variable CpGs were more likely to be positively correlated between matched tissues and enriched for DNAm sites under genetic influence. Finally, we demonstrated the relevance of our findings to EWAS analysis by categorizing DNAm associations that were previously identified in pediatric BECs and peripheral blood. Collectively, these findings highlighted a number of potential insights and considerations for the appropriate design and interpretation of EWAS analyses performed in commonly used peripheral tissues of pediatric samples.

3.2 Materials and methods
3.2.1 Study cohorts and tissue samples

Matched tissues were obtained from a subset of two separate pediatric cohorts. Specifically, a subset of samples from the previously described C3ARE (Cleaning, Carrying, Changing, Attending, Reading and Expressing) cohort were collected from 16 individuals (8 females; 50%) aged 3-5 years (age range: 3.6-4.2 years (BEC) and 4.5-5.2 years (PBMC)) from Vancouver, British Columbia (147). The GECKO cohort samples (Gene Expression Collaborative Kids Only) comprised of 79 individuals (36 females; 46%) aged 6-13 years (age range: 6-11 years (BEC) and 7-13 years (PBMC)) also from Vancouver, British Columbia. Birth dates were not available for all GECKO participants; age in years was recorded at the BEC sample collections. In both cohorts, the majority of BEC samples were collected at the first visit and PBMCs were collected at a later date. In the C3ARE cohort, follow-up visits ranged from 7 days to 1.5 years, with three pairs of matched BECs and PBMCs being collected on the same day. In the GECKO cohort, the follow-up visits at which peripheral blood was collected ranged from 6 months to 2.3 years after the initial visit. All experimental procedures were conducted in accordance to institutional review board policies at the University of British Columbia (Certificates H07-01317 (C3ARE) and H07-02773 (GECKO). Written informed consent was obtained from a parent or legal guardian and assent was obtained from each child before study participation. For both cohorts, BECs were collected using the Isohelix Buccal Swabs (Cell Projects Ltd., Kent, UK) and stabilized with Isohelix Dri-Capsules for storage at room temperature prior to DNA extraction, as previously described (217). Whole blood was collected into Vacutainer® CPT™ Cell Preparation Tubes (Becton, Dickinson and Company, NJ, USA) and PBMCs were isolated following centrifugation, washing and resuspension into R10 media.
(Sigma-Aldrich, MO, USA), as previously described (218). PBMC pellets were frozen and stored at -80°C until DNA extraction.

3.2.2 DNA isolation and DNA methylation arrays

Genomic DNA from stabilized buccal samples was isolated using Isohelix Buccal DNA Isolation Kits (Cell Projects Ltd., Kent, UK) and was purified and concentrated using DNA Clean & Concentrator (Zymo Research, CA, USA). Genomic DNA was extracted from PBMC pellets using the DNeasy kit (Qiagen, MD, USA). DNA yield and purity were assessed using a Nanodrop ND-1000 (Thermo Fisher Scientific, MA, USA). Bisulfite conversion of DNA (750 ng) was performed using the Zymo Research EZ DNA Methylation Kit (Zymo Research, CA, USA). Samples were subsequently randomized and 160 ng of bisulfite-converted DNA was applied to the Illumina Infinium HumanMethylation450K Beadchip (450K) array, as per manufacturer’s protocols (Illumina, CA, USA) (80).

3.2.3 DNA methylation array data quality control and normalization

Data from each cohort were analyzed separately. Specifically, raw intensity values from the DNAm arrays were imported into Illumina GenomeStudio V2011.1 software and subjected to initial quality control checks for array staining, extension and bisulfite conversion followed by color correction and background adjustment using control probes contained on the 450K array. Data were exported from GenomeStudio as beta values, which represent the estimated DNAm level based on a ratio of intensities between methylated and unmethylated alleles, such that beta values range from 0 (unmethylated) to 1 (fully methylated). Subsequent processing and analysis were performed in R Version 3.2.1 (http://www.r-project.org). Profiles from 65 probes targeting single nucleotide polymorphisms (SNPs) were used to ensure matched tissue samples originated.
from the same individual. The 65 SNP probes were subsequently filtered out of the dataset. Since the cohorts were not equally matched for sex, we removed sex chromosome probes (11,648) from both datasets. Additional probe filtering was performed in which poor performing probes including those with detection \( p \)-values greater than 0.01 or probes with missing beta values in more than 2% of samples were removed (14,400 C3ARE, 13,374 GECKO). Re-annotation of the Illumina 450K array was used to filter probes that are known to be polymorphic at the target CpG. Probes which have non-specific \textit{in silico} binding to the sex chromosomes were also removed (219). Final probe count after quality control probe filtering was 429,494 probes for C3ARE and 430,581 probes for GECKO. Following quality control processing, quantro determined quantile normalization to be inappropriate as the global DNAm distributions between the two distinct tissues were highly differential (220). Beta Mixture Quantile dilation (BMIQ) normalization was performed to remove differences between Type I and Type II probes on the 450K array, yielding normalized DNAm (166).

3.2.4 Cell-type correction of DNA methylation data

The effects of cellular heterogeneity on DNAm measures were removed from PBMC and BEC samples in both cohorts. Specifically, blood cell type proportions were estimated for the PBMC samples using the established Houseman blood deconvolution method (169,170). This blood deconvolution algorithm has been previously used in pediatric blood DNAm profiles where it was shown to perform reasonably well (197). To test whether this was indeed also true in our GECKO and C3ARE samples, we assessed the appropriateness of the Houseman probeset panel in our pediatric blood samples compared to adult blood profiles (169,221). We downloaded the original adult blood DNAm dataset (Reinius) on which the Houseman method
was trained (Accession # GSE35069) and filtered to 500 probes used in the algorithm that were common across all GECKO, C3ARE (following preprocessing) and Reinius samples (222). Given that this Houseman signature comprises 600 statistically-related probes, 500 of which passed quality control in both GECKO and C3ARE, we chose to use two commonly used analytical approaches, principal component analysis (PCA) and hierarchical clustering, to determine the relationship of methylation states between cohorts in the data. PCA showed an overlap of child and adult PBMC profiles in the two top-ranking PCs (accounting for 98% of the DNAm variance of the Houseman probeset panel) and similarly, adult samples did not cluster separately from child samples in the hierarchical clustering analysis. Collectively, these findings suggested that DNAm at CpGs used in the Houseman deconvolution signature were similar between adult and child blood samples (Supplementary Figure 3.1). Given that no cell deconvolution algorithm for buccal tissues exists and that buccal swabs, like saliva, are predominantly composed of BECs and leukocytes, we used a saliva-based deconvolution method which was designed to predict these cell types from underlying DNAm patterns (171,223,224). Predicted cell proportions from both PBMC and BEC tissues were used to normalize cellular heterogeneity within each tissue using a regression-based strategy (Supplementary Figure 3.2) (172). PCA was subsequently used to confirm that the correlation of estimated cell-type proportions to DNAm variance within a tissue were minimal in the corrected 450K datasets (data not shown).
3.2.5 **Assessment of cross-tissue correlation, tissue-specific variability and tissue-specific differences in DNA methylation data**

Prior to subsequent DNAm analyses, the corrected 450K datasets were filtered down to overlapping probes (419,507) between the GECKO and C3ARE cohorts. Probe-wise cross-tissue Spearman’s correlations were calculated on beta values between the matched PBMC and BEC tissues. Inter-individual variability of each CpG was calculated as the range between the 10th and 90th percentile beta values for each CpG, referred to as “reference range” (176). This method captures variability across the bulk of samples while being largely robust to outlier samples.

In order to assess sample size-related differences in our DNAm analyses between GECKO and C3ARE, we performed 100 trials of Monte Carlo simulations. Specifically, we randomly subsampled the GECKO cohort to the equivalent size as the C3ARE cohort (n = 16 individuals) 100 times and re-ran the cross-tissue correlations and reference range calculations on the subsamples. We reported the average correlation coefficients, p-values and reference ranges from the 100 trials, which we refer to as “GECKOsub.”

Paired Wilcoxon signed-rank tests were used to compare global differences in reference range between matched BEC and PBMC samples. Fligner-Killeen tests were used to compare probewise variability differences in each of the cohorts. Using previously published methods, we aimed to identify informative sites between BECs and PBMCs, which we defined as CpGs that are both variable across individuals and highly correlated between both tissues (225). To identify informative sites, we first subset each cohort down to CpGs with a reference range greater than 0.10 in both tissues. We subsequently ran a beta mixture model on Spearman correlation rho values generating two Gaussian distributions, which separated out a group of highly concordant
CpGs (Supplementary Figure 3.3). The Spearman rho distributions in this set of highly correlated CpGs was used to define a threshold correlation coefficient, the cutoff being two standard deviations lower than the mean of the distribution. In the GECKO cohort rho > 0.47 was determined as the threshold and in the C3ARE cohort, rho > 0.32 was determined as the threshold. We also set a minimum reference range of 0.05 in both tissues to exclude CpGs with little inter-individual variation.

Finally, we identified CpGs that were differentially methylated between tissues by running Wilcoxon signed-rank tests across all probes in the C3ARE, GECKO and GECKOsub datasets. For all tests, the resulting p-values were adjusted using the Benjamini-Hochberg (BH) false discovery rate (FDR) method (179). CpGs which passed an FDR < 0.05 and an effect size threshold, absolute delta beta > 5% (|∆β| > 5%), independently in all three datasets, C3ARE, GECKO and GECKOsub, were classified as “differential sites”.

3.2.6 SNP genotyping arrays

In the GECKO cohort, DNA for genotyping was collected from saliva samples of 63 individuals using the Oragene OG-500 DNA all-in-one system as per manufacturer’s protocol (DNA Genotek Inc, ON, Canada). In the C3ARE cohort, genomic DNA for genotyping was obtained from PBMC samples as described above. Genotyping data was measured at 588,454 SNP sites using the Illumina Infinium PsychChip BeadChip (PsychChip), as per manufacturer’s protocols (Illumina, CA, USA). Content for the PsychChip includes 264,909 proven tag SNPs found on the Infinium Core-24 BeadChip, 244,593 markers from the Infinium HumanCoreExome BeadChip, and 50,000 additional markers associated with common psychiatric disorders.
3.2.7 Preprocessing of SNP genotyping data and PCA analyses for genetic ancestry

Quality control pre-preprocessing of Illumina Infinium PsychChip data was performed separately for each cohort according to recommended guidelines (226). Specifically, SNPs with a low 10th percentile GenCall score or with a low average GenCall score were filtered out. Additionally, SNP probes located on mitochondrial DNA, on sex chromosomes or without chromosome labels were removed. After probe filtering, final SNP probe counts for the C3ARE and GECKO datasets were 550,200 and 547,662, respectively. To test for difference in genetic ancestry between the two cohorts, we ran all samples in PCA, using the 542,699 SNPs called for every individual in both processed datasets. Genetic ancestry was not found to differ significantly between the cohorts (Supplementary Figure 3.4), as determined by Wilcoxon ranked sum test of GECKO versus C3ARE in PC1 scores ($p = 0.8$) and PC2 scores ($p = 0.4$). Therefore, genetic ancestry was not considered in further analyses.

3.2.8 Cis-mQTL analyses

We ran cis-mQTL analyses in each cohort separately, using GECKO as the discovery cohort and C3ARE as the validation cohort. In the GECKO cohort, PsychChip data were filtered after quality control to remove any SNP probes containing missing values in 5% of all samples, leaving 560,770 SNPs. In addition, SNPs with a minor allele frequency less than 5% or not in Hardy-Weinberg equilibrium were removed. Remaining SNPs (249,835) were then numerically coded as 1, 2, or 3, for correlational analyses. Therefore, all SNPs used in mQTL analyses were directly measured on array, rather than generated through imputation. CpGs with a reference range of less than 5% were removed from mQTL analysis; this was performed separately in each
tissue, leaving 131,706 CpGs in PBMCs and 210,784 CpGs in BECs. Finally, SNP-CpG pairs less than 5 kb apart were tested as for mQTL associations using Spearman correlations. We selected a 5kb window as previous mQTL analyses using whole genome bisulfite sequencing data reported that associations between SNP–CpG pairs are more likely to be causal within a 5 kb window (85,148,227-229). In GECKO, a total of 165,591 unique SNP-CpG pairs in PBMC and 261,739 unique SNP-CpG pairs in BEC were interrogated for associations between DNAm and allelic variation; this included 145,222 SNP-CpG pairs tested in both tissues. Pairs with FDR ≤ 0.05 and DNAm change per allele ≥ 2.5% were designated as cis-mQTL candidates and followed up for validation in the C3ARE cohort (230). For validation testing in the C3ARE samples, SNP-CpG pairs were further filtered to exclude those with SNPs that were a) not present in the filtered C3ARE PsychChip data or b) monomorphic or had less than 2 heterozygotes in the C3ARE samples. The mQTL analyses were repeated in the C3ARE data. SNP-CpG pairs with FDR ≤ 0.05 and DNAm change per allele ≥ 2.5% were designated as validated cis-mQTLs and followed up in subsequent analyses. All genotyping and DNAm data were analyzed using the human assembly GRCH37 (hg19) genome build. All SNPs are reported on the (+) strand, according to standard practices in the field.

3.2.9 Representation of identified sites in published EWAS findings

In order to relate our results to published EWAS findings performed in pediatric cohorts, we selected five published studies that used the 450K array to measure DNAm profiles in pediatric BECs or peripheral blood. Specifically, these studies examined DNAm variation associated with puberty, aging in early life, childhood psychotic symptoms, fetal alcohol spectrum disorder and autism spectrum disorder (217,231-234). For each study, we downloaded
the list of probes reported as significant and matched these probes to sites, which we identified as: 1) informative sites, 2) differential sites and/or 3) cis-mQTL-associated CpGs. For one study, in which differentially methylated regions (DMRs) were reported, we downloaded the dataset (Accession # GSE50759) and extracted individual probes underlying the DMRs (234).

3.3 Results

3.3.1 Study cohorts and DNAm data processing

To explore the tissue-specific DNAm patterns of pediatric PBMCs and BECs, we used subsets from two independent human cohorts, GECKO and C3ARE, both of which contained matched tissue samples from healthy children from the Lower Mainland Vancouver area. In GECKO, individuals ranged in age from 6 to 11 years at time of BEC collection (median = 8.8) and 7 to 13 years at time of PBMC collection (median = 10.3). Of the GECKO study sample (n = 79), 46% were female (n = 36). In C3ARE (n = 16), individuals ranged in age from 3 to 5 years at time of BEC collection (median = 4.5) and 4 to 5 years at time of PBMC collection (median = 5.1) and 50% were female (n = 8) (Table 3.1).

<table>
<thead>
<tr>
<th>Characteristics</th>
<th>C3ARE</th>
<th>GECKO</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age Range (years) at BEC collection (mean)</td>
<td>3.7-5.8 (4.5)</td>
<td>6-11 (8.8)</td>
</tr>
<tr>
<td>Age Range (years) at PBMC collection (mean)</td>
<td>4.2-5.9 (5.1)</td>
<td>7-13 (10.3)</td>
</tr>
<tr>
<td>------------------------------------------</td>
<td>---------------</td>
<td>-------------</td>
</tr>
<tr>
<td>Sex</td>
<td>n = 16 total (50% F)</td>
<td>n = 79 total (46% F)</td>
</tr>
</tbody>
</table>

DNAm data, as measured across ~485,000 CpGs by the Illumina 450K array, were filtered down to overlapping 419,507 sites which passed independent quality control measures in both cohorts. Each 450K dataset was normalized to remove probe type differences and adjusted for cell-type heterogeneity in each tissue using established bioinformatic correction methods (166,169-171). Genetic variants were measured genome-wide using the Illumina Infinium PsychChip. Following probe filtering for low-quality probes, 550,200 and 547,662 SNP probes remained for analysis in C3ARE and GECKO, respectively. We used these corrected DNAm and genotyping data of matched PBMC and BEC samples from both cohorts to assess inter-individual DNAm variability, DNAm concordance across tissues and genetic influence on DNAm, in order to gain insight into DNAm variation in these commonly used pediatric peripheral tissues.

### 3.3.2 BEC DNAm had greater inter-individual variability than PBMC DNAm

As inter-individual DNAm variability within a tissue likely relates to the potential effect sizes that are detectable in EWAS analyses, we were interested in assessing tissue-specific DNAm variability. To this end, we first interrogated the global differences in inter-individual DNAm variability between PBMC and BEC samples, following *in silico* correction for cell type differences in each tissue. We used reference range as a measure of DNAm variability as opposed to absolute range in order to minimize potential skewing by outlier values and non-normal DNAm values at individual CpGs, as previously described (225,235). Within each
cohort, BEC DNAm had a significantly greater reference range than PBMC DNAm (Figure 3.1A; Wilcoxon signed-rank test, all \( p \)-values = 2.2 x 10\(^{-16} \)). In GECKO, the median reference range, measured in beta values, was 1.9\% higher in BECs (5.2\%) than in PBMCs (3.3\%). Similarly, in C3ARE, the median reference range was 1.6\% higher in BECs (3.6\%) than in PBMCs (2.0\%). The difference in reference range was not dependent on sample size, as demonstrated by the consistency between GECKO and GECKOsub, the GECKO cohort randomly subsampled to the sample size of C3ARE (\( n = 16 \)) 100 times (Figure 3.1A). In addition, tissue-specific differences in DNAm variability were observed at individual CpGs, as determined by a Fligner-Killeen test, a non-parametric test measuring homogeneity of variances between two groups. In GECKO, 217,091 probes had significantly greater variability in BEC at FDR \( \leq 0.05 \), while only 32,350 probes were more variable in PBMC. Similarly, in the C3ARE cohort, 127,472 probes had greater variability in BECs (FDR \( \leq 0.05 \)) and 8,183 probes in PBMCs (FDR \( \leq 0.05 \); Figure 3.1B). This consistent difference in variability between BECs and PBMCs was best represented by cg10852045, cg14245471 and cg1855901 (Figure 3.1C). Collectively, 85\% of C3ARE probes (108,498) with greater variability in BEC were also found in the GECKO cohort to have greater BEC variability. These 108,498 CpGs were enriched for sites with high inter-individual BEC variability in both cohorts (10,000 permutations, \( p \)-value < 1 x 10\(^{-4} \)). As well, 84\% of C3ARE probes (6,840) with greater variability in PBMCs, were also more variable in PBMCs in the GECKO cohort; similarly, this subset was enriched for CpGs with high PBMC variability in both cohorts (10,000 permutations, \( p \)-value < 1 x 10\(^{-4} \)). These findings suggested that BEC DNAm was consistently more variable than PBMC DNAm across both cohorts, in line with previous analyses using adult tissues (99).
Figure 3.1 BEC DNAm was consistently more variable than PBMC DNAm at the genome-wide and probe-wise level.

A) Distribution of reference range in C3ARE, GECKO and GECKOsub, showing a significantly great variability in BEC vs. PBMC (Wilcoxon p < 2.2x10^{-16} in each cohort). B) Scatterplot of PBMC versus BEC reference range in each cohort. C) Three examples of CpGs with the greatest reference range difference between tissues. Individuals from the GECKO cohort are shown in red and individuals from C3ARE are shown in blue.
Apart from tissue-specific differences in reference range, we also observed a cohort-specific difference in DNA methylation (DNAm) variability. Specifically, CpGs in GECKO had a significantly greater median reference range than C3ARE CpGs in both tissues (Wilcoxon rank sum test, \( p \)-values = 2.2 x 10^{-16}). In BECs, the median reference range was 1.6% higher in GECKO than C3ARE and in PBMCs, it was greater by 1.3%. This difference remained significant when GECKOsub was used in lieu of GECKO (BEC difference = 1.2%, PBMC difference = 1.1%, Wilcoxon rank sum test, \( p \)-values = 2.2 x 10^{-16}), suggesting that these cohort-specific DNA methylation variability differences occurred irrespective of sample size and may be related to age-associated increases in DNA methylation variability, as previously described (236-241) or an unaccounted for difference in sampling.

### 3.3.3 Variable CpGs were more highly correlated between tissues

Taking advantage of the matched tissue design of our cohorts, we evaluated whether DNA methylation variation in one tissue reflected DNA methylation variation in the other. We performed probe-wise Spearman’s correlations between paired BEC and PBMC samples for the C3ARE, GECKO and GECKOsub datasets, respectively (Supplementary Figure 3.5). Using multiple reference range thresholds to capture increasingly variable CpGs, as previously described, we observed progressively greater enrichment of highly positively correlated CpGs, irrespective of sample size (Figure 3.2A and Supplementary Table 3.1) (225). This suggested that, broadly, CpGs with greater variability were more likely to be correlated between these tissues than less variable CpGs.

We next sought to investigate DNA methylation variability and concordance at individual CpGs. Specifically, we aimed to identify “informative sites”, which we defined as CpGs that are both variable across individuals and highly correlated between BECs and PBMCs, using a previously
described method (225). Such CpGs may be predictive of PBMC DNAm when measured in BECs or vice versa. To be classified as informative, i.e. variable and concordant, a CpG was required to have a reference range ≥ 5% in both tissues and meet the minimum correlation coefficient between tissues of 0.47 in GECKO samples and 0.32 in C3ARE samples, as determined by a beta mixture model run on highly variable CpGs in each cohort. Overlapping CpGs that met these criteria in both cohorts resulted in a set of 8,140 informative sites. Of note, we observed a greater than expected by chance overlap (3682 out of 8140 sites, 45%, 10,000 permutations, p < 1x10^{-4}) between our set of informative sites and informative CpGs previously identified between matched samples from adult brain and blood tissues (225). Visualization of our six most correlated informative sites revealed continuous distributions of positively correlated DNAm values between the tissues, as expected (Figure 3.2B). However, the most variable informative sites exhibited discrete distributions with 2 to 3 distinct clusters, rather than a typical continuous distribution, suggesting that these CpGs may be enriched for CpGs which are likely under genetic influence (Figure 3.2B) (204).
Figure 3.2 Variable CpGs were more highly correlated between tissues.
A) Density distribution plots of Spearman’s correlation rho between matched PBMCs and BECs across C3ARE, GECKO and GECKOsub datasets showing progressively greater enrichment of highly positively correlated CpGs at increasing reference range thresholds. Reference range thresholds were set along a sliding scale with cut-offs at 0, 0.05, 0.1, 0.2 and 0.5 (depicted by gradient of green lines). B) Scatterplots of BEC DNAm versus PBMC DNAm for a representative set of informative sites (defined as CpGs that are both variable across individuals and highly correlated between BECs and PBMCs). Top-ranking correlated informative sites (shown in the left two columns) exhibited continuous distributions. In contrast, top-ranking variable informative sites (shown in the right two columns) exhibited discrete distributions, suggesting that these CpGs may be under genetic influence. C3ARE samples are shown in blue while GECKO samples are shown in red.
3.3.4 Genetic variation contributed to tissue concordance

In order to determine the influence of local genetic variation on inter-individual DNAm variability and concordance of DNAm signal across matched peripheral tissues, we identified cis-mQTLs in both BEC and PBMC samples, respectively. Briefly, CpGs were filtered by DNAm variability (reference range $\geq 0.05$) in their respective tissues and were correlated against all SNPs within a 5kb window, a window size previously demonstrated to enrich for mQTLs that are more likely to be functionally linked to proximal CpGs (85,148,229). As the GECKO cohort had a larger sample size as compared to C3ARE and was therefore more adequately powered for cis-mQTL detection, the GECKO samples were used as the discovery cohort. A total of 16,880 and 18,245 significant cis-mQTLs were identified in GECKO PBMCs and BECs, respectively (FDR $\leq 0.05$ and DNAm change per allele $\geq 2.5\%$), with 6,359 mQTLs in common between tissues (Figure 3.3A). These mQTLs were selected for validation testing in C3ARE.

After quality control processing and variability filtering of the C3ARE DNAm and genotyping data, 16,138 and 17,563 SNP-CpG pairs could be tested for validation in PBMCs and BECs, respectively (mQTLs that were not tested for validation lacked genetic variability in the C3ARE cohort). This resulted in a total of 1,871 PBMC-specific, 3,705 BEC-specific and 1,097 shared-tissue validated cis-mQTLs (FDR $\leq 0.05$ and DNAm change per allele $\geq 2.5\%$), which exhibited highly consistent effect sizes between GECKO and C3ARE cohorts (Spearman rho = 0.92, $p = 2.2 \times 10^{-16}$) (Figure 3.3A and 3.3B). The overlap between validated cis-mQTLs between tissues was greater than expected by chance (10,000 permutations, $p$-value $< 1 \times 10^{-4}$) (Figure 3.3A and Supplementary Figure 3.6). This suggested that genetic influences contributed to co-variation between tissues. Finally, we found a significant overlap of our 1871 PBMC-specific and 1097 shared-tissue cis-mQTLs with previously published mQTL hits from whole
blood samples of 7-year-old children in the AIRES cohort (1810 out of 2968 sites, 61%, 10,000 permutations, $p < 1 \times 10^{-4}$), further supporting our mQTL findings (148).
Figure 3.3 Independently validated cis-mQTL were more likely to be shared across tissues than expected by chance.
A) Stacked bar plot representing number of cis-mQTLs identified in GECKO discovery cohort (shown in blue) and number of cis-mQTLs validated in C3ARE cohort (shown in red) in either BECs, PBMCs or shared across both tissues. (B) Scatterplot of DNAm change per allele in GECKO versus C3ARE across all validated cis-mQTLs shows mQTL effect sizes (measured as DNAm change per allele) were highly consistent across cohorts (BEC-specific, PBMC-specific and shared-tissue mQTLs shown in different colours). C) Boxplots of genotype versus DNAm for representative examples of a shared-tissue (top left), BEC-specific (top right) and a PBMC-specific (bottom) validated cis-mQTLs. C3ARE samples are shown in blue while GECKO samples are shown in red.
We next sought to characterize our validated *cis*-mQTLs by their genomic localization and functional features. Firstly, the 4,980 unique CpGs associated with the validated *cis*-mQTLs showed a greater than expected by chance enrichment in intergenic regions and were depleted in intragenic and north shelf regions (2-4 kb upstream of CpG islands) (Supplementary Figure 3.7A, FDR ≤ 0.05). In particular, both the CpGs associated with tissue-specific *cis*-mQTLs and the CpGs associated with shared-tissue *cis*-mQTLs were significantly enriched at intergenic and intragenic regions and showed significant depletion at promoters and CpG islands, where DNAm levels tend to be low and there is limited inter-individual variation (Figure 3.7B and C, FDR ≤ 0.05). However, tissue-specific mQTL-associated CpGs exhibited significant enrichment at south shelf regions (2-4 kb downstream of CpG islands) whereas shared-tissue mQTL-associated CpGs were significantly enriched in north shores (0-2 kb upstream of CpG islands) but depleted in north shelf regions (Supplementary Figure 3.7B and C, FDR ≤ 0.05). In addition, we found that CpGs associated with shared-tissue *cis*-mQTLs exhibited a greater than expected by chance enrichment of informative CpGs (687 out of 812 unique CpGs in shared-tissue *cis*-mQTLs, 85%, 10,000 permutations, p < 1x10^-4), further substantiating that site-specific DNAm correlation between tissues are influenced, in part, by genetic variation (Supplementary Figure 3.8).

### 3.3.5 Tissue-specific differential DNAm was consistent across cohorts

Taking further advantage of our matched tissue design, we subsequently assessed differential DNAm between PBMCs and BECs at individual CpGs for both cohorts. In the GECKO samples, 36% of CpGs (150,647) were differentially methylated between matched BECs and PBMCs (Wilcoxon signed rank test; FDR ≤ 0.05 and |Δβ| ≥ 5%). The number of
significant differentially methylated sites was not greatly affected by sample size differences as GECKOsub had similar findings with 36% of sites exhibiting differential DNAm (149,094 CpGs, with 148,767 sites overlapping with GECKO). Similarly, in C3ARE, 38% of CpGs (157,992) were significantly differentially methylated (Wilcoxon signed rank test; FDR ≤ 0.05 and |Δβ| ≥ 0.05). The overwhelming majority of these CpGs (139,662) were differentially methylated in the same direction in GECKO, GECKOsub and C3ARE (Figure 3.4). Of these sites, 102,203 (73%) had greater average DNAm in PBMCs and 37,459 (27%) had greater average DNAm in BECs. This corresponded with a greater median DNAm across all PBMC probes (68%, 68%) as compared to all BEC probes (47%, 50%) in both C3ARE and GECKO, respectively.

Figure 3.4 Tissue-specific differential DNA methylation was consistent across cohorts.
Volcano plots of differential methylation analysis (run using a paired Wilcoxon signed rank test) between BEC and PBMC tissues for C3ARE, GECKO and GECKOsub datasets. Vertical lines represent an effect size threshold of > 0.05 for absolute mean difference between tissues (BEC - PBMC) and the horizontal line represents the nominal p-value corresponding to an FDR < 0.05 in each cohort. CpGs in dark purple met the effect size and significance cut-offs independently in all three datasets (139,662 CpGs). GECKO -log p-values were ~5X greater than that of GECKOsub and C3ARE likely due to sample size differences between datasets (n=79, n=16, n=16, respectively); y-axes were left unstandardized to display trends within each cohort.
3.3.6  **cis-mQTLs were present in previously published EWAS findings**

To provide a granular categorization of CpGs measured on the 450K array, we overlapped CpGs that were identified as a) informative (i.e., variable across individuals and correlated between BECs and PBMCs) (8,140), b) differentially methylated between matched tissues (139,662), or c) under genetic influence (4,980; i.e. number of unique CpGs associated with validated *cis*-mQTLs) across both GECKO and C3ARE cohorts. Of all CpGs associated with *cis*-mQTLs, 17.7% were informative and 76.2% were differentially methylated (Figure 3.5A). However, in CpGs associated with cross-tissue *cis*-mQTLs (812 unique CpGs in total), 84.6% were informative and 58.8% were differentially methylated.

We then applied this categorization scheme to previously reported EWAS findings performed in pediatric BEC or PBMC tissues to provide an example of how the classification of CpGs can aid in the interpretation of such studies. We selected five published studies that used the 450K array in pediatric BECs or peripheral blood to assess DNAm variation associated with puberty, aging in early life, childhood psychotic symptoms, fetal alcohol spectrum disorder and autism spectrum disorder (217,231-234). By implementing our CpG classification scheme on their respective list of significant EWAS hits, we found that *cis*-mQTLs, as identified here, accounted for 0.02-13.5% of significant CpGs reported in these five studies. Differentially methylated CpGs comprised the most represented type of CpG across all 5 studies with only one study demonstrating an overlap of 24.3% with our identified informative sites (Figure 3.5B and Supplementary Table 3.2) (234). This suggested that the majority of DNAm associations identified in these EWASs were likely specific to peripheral blood or BECs, rather than shared across tissues. Finally, we tabulated our CpGs classifications across all 419,507 DNAm probes.
assessed in our study in order to serve as a resource for researchers wishing to compare their own EWAS results (Additional file 11: Supplementary File 1). Collectively, these findings reveal the importance of considering DNAm variability and concordance between tissues, as well as genetic influences on these patterns, when interrogating and interpreting EWAS findings from pediatric peripheral tissues.
Figure 3.5 Overlap and representation of identified CpGs in previously published pediatric EWAS findings. 
A) Venn diagram of CpGs identified as informative, differentially methylated between tissues, or underlying our set of validated cis-mQTLs. Scatterplots display three representative CpGs from the pairwise intersections between categories. B) Stacked bar plot showing proportion of CpGs of each defined category represented in significant CpGs of various pediatric EWAS publications in BECs or PBMCs. (All = all categories; Differential = differentially methylated between tissues; Informative = informative CpG; Inform + Diff = informative and differential; mQTL = CpG associated with mQTL; mQTL+Diff = mQTL-associated CpG and differential; mQTL+Inform = mQTL-associated CpG and informative; None = not in any of the listed categories).
3.4 Discussion

In this study, we comprehensively compared genome-wide DNAm in BECs and PBMCs using matched samples from two independent pediatric cohorts. Moreover, we leveraged the strength of paired DNAm and genotyping profiles to define cis-mQTLs across the genome and assess the influence of local genetic variation on DNAm variability and tissue concordance. Our findings showed that at the genomic and site-specific level, BECs had greater inter-individual DNAm variability over PBMCs, with highly variable CpGs more likely to be positively correlated between the matched tissues. In our subsequent cis-mQTL analyses, we observed distinct genetic influences on tissue-specific DNAm and confirmed that a sizeable proportion of shared DNAm patterns between tissues resulted from allelic variation. Finally, we provided a classification framework for the post-hoc examination of EWAS associations and examined the representation of our categorized CpGs in published EWAS findings performed in pediatric BECs and PBMCs.

Our findings highlighted extensive differences in DNAm patterns between tissues and thus the importance of tissue selection when designing an EWAS. To a large extent, EWAS tissue selection in early life cohorts is guided by two factors. Firstly, ease of collection is particularly important in this age range and may restrict tissue availability. Buccal swabs are less invasive than intravenous puncture, and the latter contributes to participation refusal in pediatric cohorts (242). Secondly, the relevance of the tissue to the phenotype or exposure being tested represents an important consideration for all EWAS analyses, irrespective of age. As peripheral blood represents a circulating tissue with broad immune and inflammatory functions, it might be more relevant to a wider range of health phenotypes than BECs. However, another hypothesis posits that tissues that arise from the same germ layer are more epigenetically similar and thus
might be a preferred choice for surrogate tissue selection (243). For example, in comparison to blood, it has been proposed that BEC DNAm may more closely reflect brain DNAm than blood DNAm, as both derive from the ectodermal germ layer (93,234). Adding to the complexity of this issue, we found that BEC DNAm had significantly greater inter-individual variability than PBMC DNAm at the genome-wide level and at the site-specific level, a finding consistent with adult BECs and PBMCs (99). Having a higher proportion of variable CpGs might be desirable for EWAS analyses as testing any tissue with little inter-individual DNAm variation would naturally limit effect sizes. From this perspective, BECs might represent a more appropriate choice of peripheral tissue for population-based epigenetic studies over PBMCs. However, it is worth noting that while we did correct for cellular heterogeneity in both tissues using bioinformatic deconvolution approaches, the higher proportion of variable CpGs in BECs may, to some extent, be attributed to the increased diversity of cell types or residual cellular heterogeneity in BECs over PBMCs (i.e. epithelial and hematopoietic in the former and entirely hematopoietic in the latter) (224).

Taking advantage of our matched sample design, we were able to rigorously interrogate the extent of correlation between DNAm signatures of BECs and PBMCs. CpGs with greater variability were more likely to be correlated between matched tissues, as best exemplified by the 8,140 informative sites we identified. These may aid in the inference of unmeasured PBMC or BEC DNAm (when the other tissue is measured) as well as for prioritization of sites for cross-tissue replication. In the latter case, cross-tissue replication typically involves the generation of candidate gene lists in accessible tissues for validation in less available tissues, such as post-mortem samples, an approach which can boost confidence in identified associations (244-246). There was a substantial overlap (45%) between our informative sites and those previously
published in matched adult blood and brain tissues (225). However, we found only 1.9% of total measured CpGs to be informative by our measures and thresholds as compared to 9.7% found in the previous analyses of adult samples from our laboratory (225). These quantitative differences might be due to a number of reasons, with the most likely being that the blood-brain informative sites were identified using a single cohort while our blood-buccal informative sites were filtered down to sites that were common across both GECKO and C3ARE cohorts; other explanations may be methodological (i.e. slight differences in analytical thresholds derived from empirical testing), or biological (i.e. blood may be more epigenetically similar to brain tissue than to BECs, resulting in more informative sites). An in-depth analysis of such cross-tissue comparisons between pediatric and adult samples, ideally by means of longitudinal sampling of DNAm, may help elucidate such sources of tissue variation across the lifespan.

Integration of genetic and epigenetic information may further clarify the relative contribution of genetic and environmental factors on inter-individual DNAm variability. We found that genetic variation contributed to both inter-individual DNAm variation within a tissue, as well as common DNAm variation between tissues. This is in general agreement with previous findings that show that many – but not all - mQTLs have consistent effects across tissues and human populations and are generally depleted in genomic regions which tend to have low DNAm variability such as promoters and CpG island but enriched in more variable intergenic and intragenic regions (91,177,213,216,247). It is currently unclear why we observed more BEC-specific mQTLs in our matched design as compared to PBMC-specific or cross tissue mQTLs. The most likely explanation is that BECs contained more validated cis-mQTLs due to greater inter-individual DNAm variability. It is also tempting to speculate that allelic variation contributes more strongly to DNAm in BECs over PBMCs, because blood DNAm might be
more plastic and responsive due to the role of blood cells in the immune system (248-250). For example, changes in genome-wide transcriptional programs and DNA methylation (DNAm) profiles are observed in response to an inflammatory stimulus in blood leukocytes, which could be incongruent with a high degree of fixed, genetically-driven DNAm patterns in these cells (248-250). In a more complicated paradigm, DNAm variation may be best explained by the interaction of both genetic and environmental factors (GxE interactions), as previously demonstrated in blood-based DNAm profiles (137,148).

As touched upon in several recent reviews, genetic contribution to DNAm might be more prominent in shaping the DNA methylome than initially anticipated, and thus affect the analysis and interpretation of Epigenome-wide Association Studies (EWAS) findings (94,251). To illustrate this, we tested for the presence of our categorized CpGs in published EWAS findings. Notably, we found that while most identified EWAS associations may be distinct to the tissue in which they were examined, in some instances, these associations may be reflected across multiple tissues and/or under genetic influence. For example, we observed CpGs associated with autism spectrum disorder to contain the highest proportion of cis-mQTLs. While there might be a number of reasons for this, it is possible that the proportion of genetically-influenced CpGs found in an EWAS may be proportional to the heritability of the phenotype under examination, although such hypotheses will require rigorous testing in large cohorts across a diverse spectrum of phenotypes with and without heritable contributions. Furthermore, it is difficult to discern whether having a high proportion of mQTLs in EWAS analyses is favourable or not. Previous work has shown, the majority of variably methylated regions are best described by an interaction of both genetic and environmental factors (148,246). Emerging findings from neonatal blood samples have additionally shown that the bulk of variable DNAm sites are best accounted for by either additive
(G+E) or interaction (GxE) models, suggesting that environmental influences on DNAm may be further delineated with the inclusion of genotype information (148,252). As such, any mQTL-associated CpGs found in an EWAS may offer alternate interpretations to phenotypic associations with DNAm and would require further investigation for potential gene-environment effects.

It is worth noting that our study had a few inherent limitations. Firstly, in both GECKO and C3ARE cohorts, PBMCs were collected from individuals at a slightly later time point than BECs, resulting in an age-related difference (0 – 1.5 years for C3ARE; 0.5 – 2.3 years for GECKO) between matched tissues, which may have affected analyses of DNAm variability. However, we anticipate that age-related differences in DNAm variability are relatively small compared to tissue-specific differences as our findings are consistent with previous work performed on age-matched tissues in adults (99). Another limitation was the relatively small sample size of our cohorts, which may have inflated type II error rates. We also chose to not assess distal genetic effects on DNAm (ie trans-mQTLs) due to the increased multiple testing burden, but rather prioritized cis-mQTLs as previous work has suggested these may be more functionally linked to nearby CpGs (85,148,229). As well, previous work in blood has shown that the proportion of DNAm variance explained by trans-mQTLs is much lower than that of cis-mQTLs (216). For these reasons, we examined SNPs that were directly measured and not imputed, as performed in other pediatric mQTL analyses, within a 5 kb window (148,166). As a result, we likely underestimated the number of mQTLs present in our tissues. Future work using large cohorts will be required to clarify the contribution of distal genetic variants to DNAm in other peripheral tissues. In addition, our mQTL findings were limited by the coverage of the 450K array, which interrogates less than 2% of all DNAm sites across the genome, although this
includes 94% of all mapped CpGs islands. As such, it is generally biased towards CpG-dense promoter regions, which typically have limited inter-individual and inter-tissue variation (80,88,96,253). Finally, while we found the Houseman blood deconvolution method to perform well in our cohorts, evidence of substantial DNAm changes across the lifespan, especially during early childhood, necessitates the refinement of cell deconvolution methods, including adjusting for age, to allow for more nuanced estimation of cell types in early life (96,110,241).

The work here presents a comprehensive assessment of local genetic influences on DNAm in matched BECs and PBMCs, as well as a characterization of DNAm variability and concordance between paired pediatric tissues. Moreover, our results highlight a number of possible considerations for EWAS analyses, including the potential enrichment of mQTL findings following pre-filtering to variable CpGs to reduce multiple test barriers and possible strategies to facilitate in-depth curation of EWAS hits. Such post-hoc examination of significant differentially methylated CpGs will hopefully support the interpretation of EWAS findings and aid in the prioritization of candidate associations for functional validation.
Chapter 4: Children’s biobehavioural reactivity to challenge predicts DNA methylation in adolescence and emerging adulthood

4.1 Introduction

From the earliest moments of life, children’s health and development are shaped by the qualities of their environmental contexts. Processes termed “biological embedding” elucidate the possible mechanisms of such relations and describe how exposures to environmental adversity get “under the skin” to influence critical biological pathways affecting health across the lifespan (19,127). Epigenetic processes represent one model of biological embedding and have been increasingly recognized as a potential link between stressful childhood environments and later health outcomes (127,254). DNA methylation (DNAm) patterns associated with environment or experience are also influenced by other factors such as individual health behaviours, differences in temperament, and disease states (255).

DNAm is the most studied epigenetic modification in human populations and consists of a methyl group addition to the 5’ cytosine of CpG dinucleotides (CpGs). Once believed to be a gene silencing epigenetic mark, DNAm is context- and location-specific and has been linked to increased, decreased, and unchanged gene activity (63,85,88). The complex mechanisms by which DNAm can alter gene activity include inhibiting or enhancing transcription factor binding to DNA, recruiting enzymes to alter histone modifications, and altering splice sites, among others (63,256). DNAm is most dynamic during fetal development when epigenetic patterns play an integral part in the complex processes of embryogenesis (64,104) and rates of change generally stabilize in adulthood, (110). However, adolescence is also understood to be a time of
increased methylome alterations (65,110), though studies of DNAm changes during this developmental period are more scarce compared to those conducted in early childhood and later adulthood (241).

A growing body of research has revealed associations between exposures to early life environmental and psychosocial adversity and DNAm in accessible tissues such as buccal epithelial cells (BECs), saliva and peripheral blood (for an excellent review of the epigenetics patterns of traumatic stress, see Vinkers et al. (2015)) (257). For example, children reared in institutional environments show increased DNAm among many genes in peripheral blood mononuclear cells (PBMCs) and whole blood, as compared to children reared by biological parents (139,140). DNAm measured in tissues including PBMCs, saliva, and BECs also appears to be associated with early life experiences of low socioeconomic status (87), childhood maltreatment or deprivation (137,141-143) and maternal mental health problems during the perinatal period (144,145). In a prior study of this cohort conducted by our research team, exposure to maternal stress in infancy and childhood was associated with differential DNAm among offspring in mid-adolescence (146). Paternal stress in childhood was also associated with DNAm changes in mid-adolescence among female offspring only.

Beyond the influence of adverse early environmental experiences, epigenetic patterns may also be shaped by intra-individual biology. Genetic variation, for example, is a strong predictor of DNAm patterns (91,211,258). Allelic variation may alter individual susceptibility to adverse social and environmental conditions leading differences in DNAm (190). For example, an allelic variant of the FKBP5 stress-response gene altered whether adults who experienced childhood abuse or trauma also exhibited loss of DNAm at this gene (137). In addition to allelic
variation, factors indexing an individual’s internal psychological and physiological state may also be associated with patterns of DNAm (259,260).

Empirical studies examining the association between individual-level phenotypic factors and epigenetic differences are scarce. In two papers, measures of physiological reactivity to stress during infancy and childhood were associated with DNAm of BECs and placental cells (259,260), and physical aggression in early life has been shown to predict differential patterns of DNAm in T cells in adulthood (128,261). Recent research in a group of young rhesus macaques also showed anxious temperament to be associated with differentially methylated loci in the central nucleus of the amygdala (262).

A limited number of studies have thus examined associations among discrete biological and behavioural stress response parameters, psychological health, and epigenetic modifications. The inherent coupling involved in “mind-body relations,” however, suggests that a more comprehensive understanding might be gleaned from synthesizing the interrelations among individuals’ internal, individual-level biological and behavioural qualities into an integrated factor that could be examined for associations with DNAm. To this end, the current study derived measures of children’s biobehavioural response predispositions from the shared variation among temperamental traits, presyndromal behavioural symptoms, and autonomic reactivity to stressful laboratory challenges. Temperament has been defined as “constitutionally based individual differences in reactivity and self-regulation in the domains of affect, activity, and attention” (263). Such differences have established biological underpinnings and are known to influence children’s physiological and behavioural responsivity to environmental conditions (264-266). Both temperament and stress reactivity are related to the development of later forms of psychopathology and may act as antecedent, subclinical precursors, or endophenotypes (267-
When taken together, an integrative measure of temperamental traits, presyndromal mental health symptoms, and biological reactivity might plausibly provide a more powerful indicator of a child’s internal biobehavioural response predispositions than if those domains are explored independently. To my knowledge, no extant research has examined relations between biobehavioural responses and DNAm over time.

In light of previous research, the present study examined prospective associations between early, internal differences in biobehavioural responses and later epigenetic modifications across two time points within a sample of 55 individuals from the Wisconsin Study of Families and Work (WSFW). This developmentally-oriented, longitudinal research project established a birth cohort from which data on child temperament, autonomic reactivity, mental health and DNA were collected at multiple time points from prenatal life to age 18. I anticipated significant relations between childhood biobehavioural response propensities (i.e., internal factors) and adolescent patterns of DNAm, paralleling prior work performed in my lab (and that of other investigators) documenting linkages between stressful life experiences (i.e., external, environmental factors) and DNAm. I examined the relations between early life biobehavioural measures and DNAm at two time points, 15 years and 18 years. Utilizing these rich longitudinal data, I ran an additional analysis of the temporal stability of DNAm, examining whether such stability is required for the longitudinal persistence of biobehavioural associations.

### 4.2 Materials and methods
4.2.1 Study sample

Participants in the current study were drawn from a WSFW subsample (n = 120) of children, parents, and teachers (154). Children were selected for that subsample to provide a balanced representation of high and low reported mental health symptoms (155). The present analyses are based on a subset of 55 children who had complete data on temperament, mental health symptomatology, and ANS reactivity in the infancy, preschool, and kindergarten periods, and who provided DNA samples at ages 15 and 18 years (Table 1). Mann–Whitney U tests indicated no significant differences on any biological or behavioural measure between the 55 children in the present analysis and the larger WSFW subsample from which they were drawn (p > 0.05 at each variable). Of the 55 individuals, 19 were male and 36 were female. Mean family income measured at 12 months postpartum and preschool (4.5 years old) was $51,480 (median = 47,000) and $63,220 (median = 56,000), respectively. Six children were of non-Caucasian minority status. All children entered primary school in the same school year (in 1996). Ethics approval for the WSFW was obtained from the University of Wisconsin-Madison Institutional Review Board and informed consent was obtained from all participants.

4.2.2 Temperament, ANS, and mental health measures

A summary of all measures collected at each time point, organized by construct, can be found in Table 4.1.
Table 4.1 Mental health, temperament and ANS traits collected over seven years and included in the analysis

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Variable</th>
<th>Reporter</th>
<th>Age/time point</th>
<th>Instrument</th>
<th>Measure</th>
<th>Min.*</th>
<th>Median*</th>
<th>Max.*</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>ANS response</strong></td>
<td>OG1-HR</td>
<td>observational</td>
<td>grade 1</td>
<td>ANS stress reactivity</td>
<td>Heart rate reactivity (slope)</td>
<td>-1.30</td>
<td>-0.23</td>
<td>2.85</td>
</tr>
<tr>
<td></td>
<td>OG1-PEP</td>
<td>observational</td>
<td>grade 1</td>
<td>ANS stress reactivity</td>
<td>Pre-ejection period reactivity (slope)</td>
<td>-2.34</td>
<td>0.41</td>
<td>1.13</td>
</tr>
<tr>
<td></td>
<td>OG1-RSA</td>
<td>observational</td>
<td>grade 1</td>
<td>ANS stress reactivity</td>
<td>Respiratory sinus arrhythmia reactivity (slope)</td>
<td>-2.31</td>
<td>0.29</td>
<td>1.20</td>
</tr>
<tr>
<td></td>
<td>OG1-MAP</td>
<td>observational</td>
<td>grade 1</td>
<td>ANS stress reactivity</td>
<td>Mean arterial pressure reactivity (slope)</td>
<td>-1.42</td>
<td>-0.22</td>
<td>1.97</td>
</tr>
<tr>
<td><strong>Temperament</strong></td>
<td>MI-AN</td>
<td>mother</td>
<td>12 months</td>
<td>IBQ</td>
<td>Approach negativity (activity level, distress to limitations)</td>
<td>-2.14</td>
<td>-0.02</td>
<td>1.99</td>
</tr>
<tr>
<td></td>
<td>MI-WN</td>
<td>mother</td>
<td>avg. 3.5 &amp; 4.5 years</td>
<td>CBQ</td>
<td>Withdrawal negativity (distress to novelty, startle)</td>
<td>-2.09</td>
<td>0.00</td>
<td>3.06</td>
</tr>
<tr>
<td></td>
<td>MP-WN</td>
<td>observational</td>
<td>avg. 3.5 &amp; 4.5 years</td>
<td>CBQ</td>
<td>Withdrawal negativity (distress to novelty, startle)</td>
<td>-2.36</td>
<td>0.03</td>
<td>1.74</td>
</tr>
<tr>
<td></td>
<td>OP-WN</td>
<td>observational</td>
<td>4.5 years</td>
<td>LabTAB</td>
<td>Withdrawal negativity (distress to novelty, startle)</td>
<td>-2.31</td>
<td>-0.10</td>
<td>1.85</td>
</tr>
<tr>
<td></td>
<td>OG1-WN</td>
<td>observational</td>
<td>grade 1</td>
<td>LabTAB</td>
<td>Withdrawal negativity (distress to novelty, startle)</td>
<td>-2.58</td>
<td>-0.08</td>
<td>2.01</td>
</tr>
<tr>
<td></td>
<td>MP-ANG</td>
<td>mother</td>
<td>avg. 3.5 &amp; 4.5 years</td>
<td>CBQ</td>
<td>Approach negativity (anger)</td>
<td>-2.48</td>
<td>0.10</td>
<td>1.87</td>
</tr>
<tr>
<td></td>
<td>OP-ANG</td>
<td>observational</td>
<td>4.5 years</td>
<td>LabTAB</td>
<td>Approach negativity (anger)</td>
<td>-1.67</td>
<td>0.14</td>
<td>2.34</td>
</tr>
<tr>
<td></td>
<td>OG1-ANG</td>
<td>observational</td>
<td>grade 1</td>
<td>LabTAB</td>
<td>Approach negativity (anger)</td>
<td>-1.99</td>
<td>-0.09</td>
<td>1.81</td>
</tr>
<tr>
<td><strong>Mental health symptom</strong></td>
<td>MK-INT</td>
<td>mother</td>
<td>12 months</td>
<td>HBQ</td>
<td>Internalizing (depression, separation anxiety, overanxious)</td>
<td>-1.34</td>
<td>-0.19</td>
<td>2.82</td>
</tr>
<tr>
<td></td>
<td>TK-INT</td>
<td>kindergarten</td>
<td></td>
<td></td>
<td>Externalizing (oppositional, conduct, overt aggression)</td>
<td>-0.85</td>
<td>-0.43</td>
<td>3.61</td>
</tr>
<tr>
<td></td>
<td>MK-EXT</td>
<td>mother</td>
<td>12 months</td>
<td>LabTAB</td>
<td>Externalizing (oppositional, conduct, overt aggression)</td>
<td>-1.67</td>
<td>-0.10</td>
<td>2.59</td>
</tr>
<tr>
<td></td>
<td>TK-EXT</td>
<td>teacher</td>
<td>12 months</td>
<td>LabTAB</td>
<td>Externalizing (oppositional, conduct, overt aggression)</td>
<td>-0.57</td>
<td>-0.45</td>
<td>3.85</td>
</tr>
</tbody>
</table>

Note: M = mother-report, T = teacher-report, O = observed; I = infancy, P = preschool, G1 = grade 1; AN = approach negativity, WN = withdrawal negativity, ANG = anger, INT = internalizing symptoms, EXT = externalizing symptoms, HR = heart rate, PEP = pre-ejection period, RSA = respiratory sinus arrhythmia, MAP = mean arterial pressure. * Reporting the standardized descriptives that were used for PCA
During in-home assessments completed during first grade, children participated in a 15-minute standardized, developmentally-appropriate stress reactivity protocol (271). Briefly, the protocol consisted of challenges across social (interview with the child), cognitive (digit recall task), sensory (a taste identification task), and emotional (a fear- and sadness-eliciting movie clip) domains (271). Measurements of ANS activity, including heart rate (HR), pre-ejection period (PEP), respiratory sinus arrhythmia (RSA), and mean arterial pressure (MAP) were assessed via electrocardiography and impedance cardiography throughout the protocol (155,271). Autonomic reactivity was indexed as increases in HR and MAP and decreases in PEP (reflecting sympathetic activation) and RSA (reflecting parasympathetic withdrawal), relative to resting levels (155). All specific reactivity measures were computed as the slope of the ANS measure reactivity regressed on time. Positive slopes on HR and MAP and negative slopes on RSA and PEP all indicated up-regulation in general ANS arousal (155).

Temperament was assessed via both maternal report and observational coding methods, capturing the specific domains of temperamental negativity that have been theoretically and empirically related to children’s physiological reactivity (272). First, mothers reported on the Infant Behavior Questionnaire (IBQ (273)) at age 1 year and a modified version of the Child Behavior Questionnaire (CBQ (274)) at ages 3.5 and 4.5 years. Both instruments have been widely validated for use in their respective target populations (273-276).

Observational measures of children’s temperament were also collected using the Laboratory Temperament Assessment Battery (LabTAB) administered in a standardized fashion during home assessments at 4.5 years and in grade one. The LabTAB is comprised of 12 emotion-eliciting behavioural episodes that simulate everyday situations (e.g., a social interaction with an unfamiliar adult, waiting for a signal before eating a snack, and using fine
motor skills at a toy workbench). These situations were used to evoke affective reactions within three domains: negative affectivity, positive affectivity, and behavioural control-regulation. All LabTAB assessments were videotaped and rated by two independent reviewers who coded facial, vocal, motor, behavioural and postural responses (277). To remain consistent with the temperament domains assessed by maternal report on the IBQ and CBQ, only coded behavioural observations of approach negativity and withdrawal negativity, elicited during negative affectivity episodes, were included in the present analyses. Maternal report and laboratory-based observations of temperament and behaviour that were not expected to relate to children’s physiological reactivity were excluded. For additional details of LabTAB episodes, administration, and coding (see Luby et al., 2002). Both LabTab- and questionnaire-derived temperamental measures have been shown reliable and internally consistent in this sample (277).

Children’s presyndromal, internalizing and externalizing symptoms were assessed in kindergarten using maternal and teacher reports on subscales from the MacArthur Health and Behavior Questionnaire (HBQ) (278,279). The HBQ internalizing and externalizing subscales are well-validated measures of emotion regulation and reactivity difficulties relevant to the present analyses of children’s biobehavioural reactivity (e.g., sadness, withdrawal, irritability, anger) (280,281). The validity and reliability of these measures have been previously established in this sample (281).

4.2.3 DNA methylation

BECs were collected from participants at ages 15 and 18 years using MasterAmp Buccal Swabs (Epicentre Biotechnologies) and were stored at -80° C. Genomic DNA was extracted from buccal swabs using Buccal DNA Isolation Kits (Isohelix Ltd), then purified and concentrated
with DNA Clean & Concentrator kits (Zymo Research). DNA quality was assessed by a NanoDrop ND-1000 (Thermo Scientific). 750ng of genomic DNA underwent bisulfite conversion using the EZ DNA Methylation Kit (Zymo Research).

Bisulfite converted DNA was treated according to established protocols (Illumina) in preparation for loading onto microarrays. DNA from buccal swabs collected at age 15 years was assayed using the Infinium HumanMethylation27 BeadChip (27K array); previous findings on these data can be found in Essex et al. (2011). DNA from samples collected at age 18 was assayed using the next generation of this technology, the Infinium HumanMethylation450 BeadChip (450K array).

For samples collected at age 15, 160 ng of bisulfite-converted DNA was whole-genome amplified, fragmented, and hybridized onto the 27K array. The 27K array analyzes DNA at 27,578 CpGs, primarily at DNA sequences that map onto gene promoter regions. Raw DNA data from the scanned microarrays are available in the gene expression omnibus (GEO) database under the accession number GSE25892 at: https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE25892 (146).

For samples collected at age 18, 160 ng of bisulfite-converted DNA was whole-genome amplified, fragmented, and hybridized onto the 450K array. The 450K array covers over 485,000 CpGs, representing 99% of all RefSeq genes, and includes ~90% of the CpGs that are on the 27K array. Microchips with hybridized DNA, data were input into Illumina’s Genome Studio software.

For the 27K array, DNA data were background-adjusted and quantile-normalized, as previously described (146). Probes underwent rigorous quality control processes, were assessed for detection p-value, number of underlying probes, and signal levels in each subject and
replaced with “NA” in the subject of interest if they failed at any metric (total of 4567 CpG measurements). One sample was removed based on poor quality data (>10% probes with a detection p-value > 0.05), leaving 109 samples, and one probe was removed due to poor quality data (detection p-value > 0.05 in more than 10% of individuals). The remaining samples had between 0 and 194 “NA” values (mean = 27.19, median = 13), which were replaced with imputed values using the “impute.knn” function in the R package “impute” (282). Probes on the X chromosome were removed as these CpGs differ by sex. Finally, probes were removed based on lack of inter-individual variability based on existing data reduction methods. First, probes in which β values across all individuals were < 0.05 or > 0.95 were excluded (9,346) (283). Second, analyses were restricted to individuals with complete biobehavioural measures only (n=55) and any probes with a range < 0.05 β, as calculated between individuals within the 10th and 90th percentile were removed (8,309). This left 9,922 probes for analysis (176). Although this variability cut-off was stringent most probes measured on the 27K array are located within promoter CpG islands, which are the most highly stable CpGs in the genome (88). The 9,922 CpGs were annotated to 6,583 unique genes; 2,212 genes contained two or more variable probes. Finally, the “detectOutlier” function in the lumi package was applied and no outliers were identified (n=55) (164).

DNAm data from the 450K array were background-adjusted and color-corrected in Genome Studio (Illumina). Three outlier samples were detected using the “detectOutlier” and removed, leaving 52 individuals; this method assumes a single cluster of samples and removes samples which are greater than two-times the median distance from the center of the cluster and was applied following data visualization due to visible outlying samples (164). A total of 4,314 probes deemed low quality were removed. Additional probes that were removed included those
that mapped onto sex chromosomes (1, 216), assayed single nucleotide polymorphisms (SNPs; 64), and were found to cross-hybridize to sex or autosomal chromosomes (37, 909) or SNPs (19, 999), based on a previous annotation (165). Data were then normalized across samples using quantile-normalization, followed by normalization of probe-type differences using Subset-quantile Within Array Normalization (SWAN) (167). An empirical Bayes method (ComBat) was applied to correct for effects associated the separation of samples into batches during the microarray experiments, specifically into 96-well plates, into microarray chips holding 12 samples per chip, and into 6 rows present on each chip (173).

Of note, the DNAm data collected at ages 15 and 18 were run at different times and on different technologies (the 27k array and the 450k array, respectively), and these data were treated differently during preprocessing. Specifically, the 450k array contains two probe types (type I and type II), which have different dynamic ranges and therefore different \( \beta \) value distributions; this was corrected for using SWAN. The 450k array was also corrected for batch effects using ComBat due to technical variability. The 27k array contains only type I probes and batch effects as measured by plate, chip and row were not significantly correlated with the variables of interest (all \( p \)-values > 0.05), nor were they correlated with the first principal component of the DNAm array data, which accounted for over 90% of total variability. Therefore, neither SWAN nor ComBat were used on the 27k array at age 15. For consistency, CpG sites assayed on both the 27k array and the 450k array were mapped to genes using the Price annotation (165), while the Illumina HumanMethylation27 Manifest File was used for CpG assayed only on the 27 array.

At each CpG assayed on the microarray platforms, a \( \beta \) value ranging from 0 (completely unmethylated) to 1 (100% methylated) was calculated for each sample using the signal intensity
from the scan. β values of all samples from the 27K array and 450K array were log transformed into M-values prior to analysis to adjust for the heteroscedastic nature of β values (175). All results are reported in β values to facilitate biological interpretation.

4.2.4 Statistical analysis

Principal component analysis (PCA) is a multivariate technique designed to reduce the dimensionality of a large set of non-independent variables, while retaining important sources of variation in the dataset (284). In the present analysis, PCA was used to derive biological and behavioural response propensities that reflect the shared, underlying regulatory processes common across the measures of temperament, ANS reactivity, and internalizing/externalizing symptoms (Supplementary Figure 1). PCA provided an effective analytic alternative to conducting separate analyses of the relations between the biological/behavioural measures and DNAm, which would have been statistically and conceptually problematic due to the increased rate of Type I error and the interrelatedness of the reactivity measures. PCA was run on the 16 scaled variables derived from measures of temperament, ANS, and internalizing/externalizing symptoms (Table 1), producing 16 principal components (PCs) that represent mathematically unique but conceptually overlapping aspects of the child’s biological and behavioural response propensities. Each PC (hereafter referred to as a biobehavioural reactivity factor) represents an independent axis of variation among the data, driven by different combinations of weights from the original measures. The first three biobehavioural reactivity factors were chosen for DNAm analysis based on examination of the scree plot, which indicates a clear “break” between the third and fourth component (285).
A number of covariates were evaluated for potential inclusion in the analyses. The sample was homogenous in age (all participants entered preschool within the same school year) and ethnicity (87% Caucasian), precluding the need to control for these variables. Family income, measured at 9 months postpartum and kindergarten, were also excluded as covariates, because they were not correlated with any of the three PCs (all \( p \)-values > 0.05). The effect of sex was examined post hoc.

The cell composition of buccal swabs can vary between individuals, altering DNAm patterns, and was therefore also considered for inclusion as a covariate. The percentages of underlying BECs and leukocytes were calculated from the DNAm profiles measured at age 15 using a cell deconvolution algorithm trained on BECs and saliva samples (171). Using this tool, the proportions of BECs in the age 15 cohort was estimated to range from 81-96% (mean 87.7%). However, this proportion was not correlated with the measures of interest (\( p \)-values >0.05) and was also excluded from further analysis.

Associations between biobehavioural reactivity and DNAm

Associations between the biobehavioural reactivity factors and DNAm at age 15 were examined from the 27K array. The sample of 9,922 variable CpGs was tested against the three biobehavioural reactivity factors using Spearman rank order correlations. \( P \)-values were corrected using the Benjamini-Hochberg method to estimate false discovery rates (FDRs), which limits the expected proportion of false positives and therefore reduce the number of Type I errors (179). The magnitude of change in DNAm across individuals, termed \( \Delta \beta \), was calculated using the slope of the regression line and reported for each differentially methylated CpG as a measure of effect size (87). CpGs that were significantly associated at a FDR corrected \( p \)-value of 0.05 or smaller and had an absolute delta beta > 5% (\(|\Delta \beta| > 5\%\)) were reported as high confidence.
differentially methylated CpGs (146). CpGs with a FDR corrected $p$-value between 0.05 and 0.2 and an absolute $|\Delta \beta| > 5\%$ were reported as medium confidence differentially methylated (146).

To test whether associations identified at age 15 persisted at age 18, I examined associations between the biobehavioural reactivity factors and DNAm measured at age 18 using the 450K array. Because age was perfectly confounded with both microarray platform and processing methods, I chose not to compare directly $\beta$ values at ages 15 and 18 and rather repeated the correlations using DNAm assessed at age 18. This tested whether variation across individuals was also associated with biobehavioural reactivity at the later time point. As described in more detail below, I focused on genes with multiple high and medium confidence CpGs found at age 15. Again, $p$-values were corrected using the Benjamini-Hochberg methods to control for FDR. As an additional test of the strength of the associations, correlation coefficients between the PCA-derived biobehavioural reactivity factors and differentially methylated CpGs were permuted 100 times, generating a null distribution, and compared to the true correlation coefficients.

We also tested stability in DNAm from age 15 to age 18 using mixed effects models that evaluated the prediction of DNAm at age 18 from DNAm at age 15. This analysis was performed on a subset of the CpGs significantly associated with biobehavioural reactivity at age 15. As such, biobehavioural reactivity was included as a covariate.

Gene ontology (GO) analysis was performed using the software ErmineJ (286,287). The 9,922 CpG sites used in the analysis were annotated to genes as previously described, in order to generate a complete gene list or “background” from which to test for enrichment (288). Analysis was then performed using precision-recall and the following parameters: use the best scoring
replicate, include only “Biological Process” related GO terms, minimum gene set size of 5, maximum gene set size of 100, and test the effect of multifunctional genes.

4.2.5 Pyrosequencing experiments

Pyrosequencing experiments were performed in order to confirm significant associations between children’s biobehavioural reactivity factors and DNAm in the DLX5 gene found at both age 15 and age 18. A pyrosequencing assay was designed to examine the DLX5 3’ CpG island; this assay spanned approximately 200 base pairs and included 5 CpGs, two of which are covered by the 450K array (cg12041387, cg08835113); cg12041387 is also covered by the 27k array. Of the 55 child samples used in this analysis, only 42 had enough remaining genomic DNA from samples collected at age 18 to conduct experiments. All reactions were run on a PyroMark Q96 MD Pyrosequencer, following the manufacturer’s protocol. All CpG loci passed Pyro Q-CpG software quality control. Primer sequences used for DNA amplification and pyrosequencing are available upon request.

4.3 Results

4.3.1 Principal components analyses of biobehavioural reactivity

Following analyses, the first three principal components were examined to derive and understand the primary factors that described children’s biobehavioural reactivity. The first PC explained approximately 18% of the total variation, while the second and third PCs explained 13.5% and 11.5%, respectively (Figure 4.1a). Maternal-report of children’s temperament and behaviour loaded strongly onto the first PC, while teacher-report and observational measures of children’s biobehavioural reactivity did not
(Figure 4.1b). Thus, the first PC distinguished mothers’ broad perspectives on children’s functioning from measures of children’s reactivity in more context- and stressor-specific settings. The third PC differentiated HR reactivity (an indicator of the activity of both the parasympathetic and sympathetic branches of the ANS) from RSA (a measure of parasympathetic activity only; Figure 4.1b). My review of the PCA results indicated that both the first and third principal components reflected variance attributable to method- and/or reporter-based differences, and thus I hypothesized that these components would not associate with DNA methylation.

### 4.3.2 Associations between biobehavioural reactivity factors and DNAm at age 15

As expected, when correlated with DNAm at all 9,922 CpGs measured at age 15, PC1 and PC3 displayed uniform $p$-value distributions, suggestive of a null distribution, and neither PC was significantly correlated with any individual CpG after FDR correction (Figure 4.1c).
Figure 4.1 Results of principal component analysis revealed biobehavioural reactivity as biologically driven composite measure.

(a) Plotting the percent variability of the principal components (PCs) showed a flatter distribution than what is typically expected when running PCA on psychological variables. (b) Loadings of original variables onto PC1, PC2 and PC3 (from left to right). (c) p-value distributions of genome-wide correlations between DNA methylation and PC 1 to 3 (left to right) (n= 55).
The second PC represented individual differences in children’s biobehavioural inhibition and disinhibition, ascertained across teacher-, maternal-, and laboratory-based observational measures (Figure 4.1b). Observed measures of withdrawal negativity and internalizing symptoms loaded positively onto PC2, whereas anger and externalizing symptoms loaded negatively. Measures of autonomic reactivity also showed moderate loadings on PC2: HR loaded positively, while PEP, MAP and RSA loaded negatively. Thus, PC2 integrated both behavioural and biological response characteristics—across reporters, contexts, and stressors—distinguishing inhibited children (higher scores) from disinhibited children (lower scores) and is hereafter referred to as Biobehavioural Inhibition/Disinhibition (BID). Correlations between BID and DNAm showed a left-skewed $p$-value distribution that deviated from the distribution that would be expected by chance, suggesting an association with age 15 DNAm (Figure 4.1c). Examining this association more closely, I found that BID was significantly associated with 12 CpGs at an FDR cut-off of 0.05 (“high confidence” CpGs) and an additional 81 CpGs were associated with at an FDR between 0.05 and 0.2 (“medium confidence” CpGs). Therefore, this multi-method, multi-reporter composite trait of early life biobehavioural reactivity showed an observable and statistically significant DNAm signature. To check the robustness of my findings, a linear regression was also run on all 9922 CpGs, using BID as the explanatory variable and including sex and minority status as covariates. The significant CpGs were largely stable across statistical tests (Supplementary Figure 4.2). Given the lack of covariate effects and my interest in the basic bivariate associations, I focus on the results of the Spearman correlations.

Next, genomic locations of the 93 high confidence and medium confidence CpGs differentially methylated by BID were examined. These CpGs mapped to multiple genes, including the imprinted genes, GNAS complex locus ($GNAS$) and insulin-like growth factor 2
(IGF2), and genes related to neurotransmitter secretion, including vesicle-associated membrane protein 5 (VAMPS) and otoferlin (OTOF). However, after running gene ontology analysis on genes ranked by p-value of associated CpGs, I did not find that genes containing differentially methylated CpGs could be classified by shared biological processes.

All genes contained only a single differentially methylated, high or medium confidence CpG, with the exception of four: DLX5 (distal-less homeobox 5), IGF2, MYO16 (myosin XVI), and PRUNE2 (prune homolog 2) (Figure 4.2). Nine CpGs mapped to the DLX5 gene, located within the coding region of the gene between 1.5 and 4kb downstream of the transcription start site. Five were high confidence CpGs (FDR < 0.05) and four were medium confidence (FDR 0.05-0.2). The nine CpGs were contiguous except for one interrupting CpG that fell just outside of the medium confidence threshold (cg02101486, FDR p-value < 0.209). These CpGs thus constitute a differentially methylated region (DMR), a region in which multiple, adjacent CpGs exhibit associations with BID. All 9 differentially methylated CpGs within DLX5 correlated negatively with BID scores, with Spearman’s rank order correlation coefficients (rhos) ranging between -0.60 and -0.44. The nine high and medium confidence CpG sites possessed large DNAm differences observed across individuals with the highest and lowest BID scores; the |Δβ| ranged from 9%-38% (149).
Figure 4.2 Schematic of DLX5, IGF2, MYO16 and PRUNE2 genes, which each contained more than 2 high or medium confidence CpGs significantly associated with Biobehavioral Inhibition/Disinhibition.
Green bars in gene schematic represent CpG Islands and gray lines or boxes represent genomic locations of CpGs plotted. Scatter plots of individual CpG DNA methylation are colored by sex (males = blue, females = pink) (n= 55).
IGF2, PRUNE2 and MYO16 each contained two probes significantly associated with BID. IGF2 contained one high confidence CpG (cg11005826, FDR p-value < 0.04, rho= -0.52, Δβ=15%) and one medium confidence CpG (cg21237591, FDR p-value < 0.13, rho= -0.45, Δβ=9%), which were negatively associated with BID and were located 3620bp and 169bp upstream of the transcript start site, respectively. PRUNE2 contained two medium confidence CpG sites which were positively correlated with BID scores (cg11880010, FDR p-value < 0.09, rho= 0.47, Δβ=18%; cg19282250, FDR p-value < 0.15, rho= 0.42, Δβ=15%), located 9620 and 9507bp upstream of the transcript start site, respectively. Finally, MYO16 contained two medium confidence CpGs. One CpG, located 452bp upstream of the transcription start site, was positively correlated with BID (cg14396117, FDR p-value < 0.08, rho= 0.48, Δβ=10%) and one CpG, located 4bp downstream of the transcription start site was negatively correlated (cg18946226, FDR p-value < 0.13, rho= -0.44, Δβ=7%).

To ensure that effect sizes of the high and medium confidence sites were not being inflated by individuals with extreme DNAm values, correlations with BID were recalculated after a 90% winsorization (5% was modified from each tail) of all medium and high confidence CpGs. Results remained largely unchanged and the mean difference between p-values generated before and after winsorization was 6.66x10^{-6} (median= -6.84x10^{-7}, 1st quartile = -1.67x10^{-5}, 3rd quartile = 1.63x10^{-5}). Prior to winsorization nominal p-values ranged from 1.06x10^{-6} to 1.97x10^{-3} and after winsorization nominal p-values ranged from 9.52x10^{-7} to 2.1x10^{-3}. Similarly, |Δβ| differences were minor; effect sizes changed by -4.05x10^{-3} on average after winsorization (median= -1.71x10^{-3}).
To add an additional level of rigor to the analyses and to test that these correlations did not occur by chance, I permuted the BID scores 100 times and correlated the permuted scores with the 15 high and medium confidence CpGs mapping to *DLX5*, *IGF2*, *PRUNE2* and *MYO16* (Supplementary Figure 4.3). With the exception of cg14396117, all of the true correlation coefficients were significantly greater than correlations expected by chance (*p*-value < 0.01). The correlation coefficient for cg14396117, located in *MYO16*, fell outside the 99th percentile of the null distribution (*p*-value < 0.02). Therefore, these findings were unlikely to be spurious, but rather reflected significant associations between DNAm of *DLX5* and childhood reactivity, as measured by BID.

### 4.3.3 Examination of sex differences in correlations between biobehavioural reactivity factors and DNA methylation at age 15

Given previously observed differences between sexes in DNAm, temperament, and mental health symptoms, correlations between the 93 BID-associated CpGs were re-examined separately within males (n=19) and females (n=36) (289,290). In general, the correlation coefficients (rhos) generated in females were similar to the coefficients generated when including both sexes in correlations (mean absolute difference in rho = 0.07) (Figure 4.3 top panel); however, the correlation coefficients generated in the males differed more strongly (mean absolute difference in rho = 0.13). To test whether this sex difference was driven by sample size, females were subsampled 100 times down to the sample size of males (n = 19) and correlations were rerun. The correlations were predominantly stable (mean absolute difference in rho = 0.02) (Supplementary Figure 4.4), suggesting that the difference in sexes was not entirely driven by sample size.
Additionally, I directly compared the DNAm values of males to females and assessed whether high and medium confidence CpGs differed by sex. A Mann–Whitney U test was run on all 9,922 CpGs sites to test for CpGs differentially methylated by sex, revealing 192 such sites (FDR p-value < 0.05), including five CpGs significantly associated with BID (cg01796228, cg09340639, cg09565688, cg11005826, cg15731815). These mapped to the genes LIFR, FCRL1, CLEC12B, IGF2, and RNF207. In all five CpGs differentially methylated by sex, associations with BID remained significant in females (all p-values < 0.05) but were no longer significant in males (all p-values > 0.05) (Figure 4.3 bottom panel).

Figure 4.3 Correlations between DNA methylation at age 15 and Biobehavioural Inhibition/Disinhibition in males and females.
(top) Spearman’s correlation coefficients of 93 high and medium confidence CpGs, calculated in full cohort (gray circle, n=55), females only (pink square, n=36) and males only (blue triangle, n=19). (bottom) 5 CpGs which were associated with Biobehavioural Inhibition/Disinhibition and differentially methylated by sex. In all CpGs, correlations remained significant in females only but lost significance in males (n = 55).
We then examined potential sex differences in biobehavioural reactivity-DNAm relation within the \textit{DLX5} gene only. All correlations with CpGs (9 CpGs) remained significant in females, with correlation coefficients ranging from -0.38 to -0.63 (\textit{p}-value < 0.05). In males, two CpGs, cg12041387 and cg11500797, remained significant, with correlation coefficients of -0.52 and -0.49, respectively (\textit{p}-value < 0.05). In sum, only five of the 93 CpGs reported (cg01796228, cg09340639, cg09565688, cg11005826, cg15731815) were differentially methylated by sex, suggesting that DNAm patterns linked to BID are not likely driven by sex-specific differences.

4.3.4 Persistence of associations between BID and DNAm at age 18

Given the strong associations between BID scores and multigenic DNAm at age 15 years, I hypothesized that those same associations would hold three years later, when participants were 18 years old. Using the 450K data at age 18, I examined all variable CpGs annotated to the four genes of interest to take advantage of the added coverage of the 450K array, as compared to the 27k array. I examined correlations between BID scores and 39 CpGs in \textit{DLX5}, 35 in \textit{IGF2}, 9 in \textit{PRUNE2}, and 37 in \textit{MYO16}; this included all CpGs in those genes found to be significant at age 15.

A total of 15 probes at age 18 were found to be significant after FDR correction for the 120 tests (FDR \textit{p}-value < 0.05 and $|\Delta \beta| > 5\%$) (Figure 4.5). As before, these differentially methylated CpGs were permuted 100 times to create null distributions, and all observed correlation coefficients either fell outside of the 97\% percentile of the null distribution (Supplementary Figure 4.4).
Figure 4.4 DLX5 and IGF2 DNA methylation remained significantly associated with Biobehavioral Inhibition/Disinhibition at age 18.

(A) DNA methylation at age 18 in 15 probes located upstream and within the DLX5 gene. (B) DNA methylation at age 18 in 4 probes located upstream of IGF2 gene. Panel titles represent distance from transcription start site, followed by CpG ID. Correlations found to be significant at age 18 are labelled with a red box; Correlations found to be significant at both age are labelled with a red/black box; all remaining CpGs were found to be significant at age 15 only. Males are plotted in blue; females are plotted in pink (n=52).
The significant probes included 13 *DLX5* CpGs, including two of the nine probes found significant at age 15 (cg12041387 (rho = -0.45), cg00503840 (rho = -0.39)). An additional four CpGs that were significant at age 15 had a *p*-value < 0.05 but did not pass FDR correction at age 18 (Spearman’s rho = -0.34 – -0.31, |Δβ| = 10 – 12%). There were an additional 11 significant CpGs at age 18 that were assayed only on the 450K array and thus not tested at age 15. These were located within the same region reported at age 15, as well as upstream of the transcription start site (600-1700bp). Again, effect sizes were notable and ranged from 10 – 21%.

The remaining two age 18 probes significantly associated with BID were *IGF2* CpGs (cg02425416 and cg11701022) with correlation coefficients of -0.40 and -0.47. These were not measured at age 15. Two *IGF2* CpGs measured at age 15 that were significantly associated with BID (cg11005826 and cg21237591) were no longer significant at age 18.

### 4.3.5 Longitudinal stability in DNA methylation

Taking further advantage of the longitudinal DNA methylation data, I directly examined its stability between the age 15 and 18. First, I compared the inter-individual ranges of 17039 CpG sites run on both platforms, after removing low quality CpGs and those at which β values across all individuals were <0.05 or > 0.95. A few sites differed substantially, an expected finding, given that samples were taken three years apart. However, the average difference in inter-individual variation neared zero, suggesting no systematic difference across the study sample due to either age or microarray platform (median = -0.011, mean = -0.022) (data not shown).

We then examined the 15 CpGs found in *DLX5, IGF2, PRUNE2* or *MYO16*, that were significantly associated with BID at age 15, to ask whether these sites would reflect the global trend in stability seen above. Using a linear mixed effects model, DNA methylation at age 15 was tested as
the explanatory variable, predicting DNAm at age 18, and BID was included as a covariate. These regression models were significant for 13 of the 15 CpGs (FDR corrected p-value < 0.05), with DNAm at age 15 explaining 8-24% of the variation in DNAm at age 18 (Supplementary Table 4.1). Although DNAm patterns at ages 15 and 18 were significantly associated, median DNAm in the 15 CpGs changed from -18% to +11%, with DNAm decreasing in 13 of the 15 CpGs three years later (figure 5a). These findings suggested a lack of stability of DNAm from age 15 to 18.

4.3.6 Pyrosequencing experiments verified associations between Biobehavioural Inhibition/Disinhibition and DLX5 DNA methylation

Finally, pyrosequencing experiments were conducted to verify the DNAm patterns in DLX5 detected on the microarray platforms. All 5 probes in the pyrosequencing assay were negatively correlated with BID (Figure 5b), with correlation coefficients ranging from -0.47 to -0.52. However, DNAm levels assessed by pyrosequencing were consistently 5-15% lower than in the 450K array in the two CpGs measured in both platforms (Figure 5c). Serial dilutions of artificially methylated and unmethylated samples run on pyrosequencing confirmed that the assay was not biased; this indicated that the array probes at these loci may exhibit a preference for binding methylated DNA over unmethylated DNA.
Figure 4.5 CpG stability across three years and pyrosequencing of DLX5 gene to verify DNA methylation findings.

(A) Changes in DNA methylation in 15 CpGs measured across both ages. Red/black boxes indicate CpGs significantly associated with Biobehavioral Inhibition/Disinhibition at both age; all remaining CpGs were found to be significant at age 15 only (B) 5 CpGs assayed by pyrosequencing DNA collected at age 18 were associated with Inhibition/Disinhibition scores. Two CpGs are identified by their 450K IDs, the remainder were not assayed by the 450K array. (C) Correlations of DNA methylation values at cg12041387 and cg08835113 generated by pyrosequencing and the 450K array (n= 42).

Pearson correlation coefficients ranged from 0.84 to 0.86. However, pyrosequencing generated consistently lower DNA methylation values (n= 42).
4.4 Discussion

Current research on the epigenetic alterations that accompany early adversity exposures has led to a growing set of studies supporting external, environmental influences on DNA methylation (DNAm). Much of this research has examined the relation between early childhood stressors on DNAm (45,127). However, variations in physiological and psychological functioning have also been shown to correlate with DNAm (259,261,262) and may reflect the linkage of a child’s internal reactivity to the environment through epigenetic signatures. Few empirical studies have tested this latter supposition, however, and those that have are mainly cross-sectional in nature. I sought to address these limitations by testing the hypothesis that children’s biological and behavioural response propensities would also be related to DNAm measured later in life. I also examined the persistence of relations between Biobehavioural Inhibition/Disinhibition (BID) and DNAm at age 15 and age 18 and the stability of DNAm at these CpGs.

Multiple-reporter and multi-method measures of early childhood temperament, behaviour, and ANS reactivity were input into a PCA examining early life biobehavioural reactivity factors. Three components explained a significant percentage of the overall variance in temperament, mental health symptoms, and ANS reactivity and were further examined for their associations with DNAm at age 15. Two components, PC1 and PC3, were not associated with DNAm patterns at age 15, likely because they reflected method- and reporter-based variance rather than trait-like differences in children’s biobehavioural reactivity. The second principal component (PC2), however—BID—reflected the intersection of observed temperamental withdrawal, anger, autonomic reactivity, and internalizing/externalizing symptoms and was
found to have both a broad DNAm signature across many genes and a particularly strong association with multiple sites within *DLX5* and *IGF2* genes.

Elevated levels of children’s biobehavioural *disinhibition* (approach negativity, anger, externalizing symptoms) were associated with significantly higher DNAm in *DLX5* and *IGF2* at age 15 and at age 18 years. Conversely, those with greater childhood *inhibition* (fear, withdrawal negativity, internalizing symptoms, and heart rate response) showed lower *DLX5* and *IGF2* methylation (Supplementary Figure 6). While median DNAm within these CpGs changed over the course of 3 years, the association between children’s early biobehavioural reactivity and *DLX5* and *IGF2* methylation was maintained from adolescence into young adulthood, a period marked by significant changes across various domains of psychological development (291). This illustrates the paradoxically fixed yet dynamic nature of DNAm. Despite changes across the lifespan, there may be enduring patterns of DNAm in certain genes over time, particularly those associated with early biological and behavioural reactivity.

Given the tissue of origin in the current study (BECs), I can only suggest that these DNAm patterns represent a biomarker of behavioural reactivity. However, Gene-Tissue Expression (GTEx) RNA sequencing data (292) indicate that *DLX5* is expressed in skin tissue and some brain regions, and it is possible that the magnitude of inter-individual DNAm differences in *DLX5* reflects actual differences gene transcription levels, affecting downstream biological processes (149). *IGF2* is not expressed at significant levels in either skin tissue or brain regions, though this does not negate a potential for DNAm differences to be a biomarker of biobehavioural reactivity. Importantly, results of present research are preliminary, and validation of these findings in a larger sample, with more repeated measures of DNAm and reactivity, are needed.
DLX5 is homeobox gene involved in neuron, craniofacial and bone development. Its protein regulates glutamic acid decarboxylases involved in the synthesis of gamma-aminobutyric acid (GABA), the chief inhibitory neurotransmitter in GABAergic neurons (293). If the differences in DNA methylation (DNAm) observed in this cohort correspond to regulatory epigenetic patterns in brain tissue, such differences could produce altered levels of GABA, corresponding to inhibited or disinhibited behavioural proclivities. Although GO analysis of all CpGs ranked by p-value did not find any significant enrichment of GABA-related genes, these results cannot fully rule out the involvement of GABA circuitry. Future research should explore the underlying pathways between reactivity-DNA methylation associations.

DLX5 is also highly expressed in osteoblasts during embryogenesis and plays an important role in craniofacial development. Commensurate with my finding of temperament and behavioural response-associated differences in DLX5 methylation, aspects of temperament have been previously linked to the bizygomatic width of facial structure, i.e., the ratio of the facial diameter across the cheekbones to the vertical height of the head. Specifically, four-month-old infants who showed propensities to biobehavioural reactivity had smaller bizygomatic widths (i.e., narrower faces) at 14 and 21 months than infants who were less reactive (294). Although the various functions of the DLX5 gene suggest that the protein may act as a regulator of the early development of inhibition and reactivity, further research is needed to confirm the relation between inhibition and DLX5 DNA methylation.

IGF2 is expressed in many fetal tissues and encodes a growth factor that primarily acts to promote overall growth during gestation via cell differentiation and proliferation. Unlike the DLX5 gene, the function of DNA methylation patterns in this gene have been well-studied. IGF2 is imprinted, i.e. shows parent-of-origin specific expression, and expressed only from the paternal
allele (295). Aberrations in imprinting of this gene, as well as other nearby genes, are associated with Beckwith-Wiedemann and Silver-Russell syndromes—two congenital, growth-affecting conditions—as well as a number of cancers (296-300). While there are no published clinical features of Beckwith-Wiedemann relating to temperament, parents of affected children often describe them as more tenacious than their siblings (R. Weksberg, personal communication, December 2017). Previous studies of IGF2 methylation have also found associations with ADHD symptoms and prenatal maternal anxiety (301,302). Given that the BID measure was comprised of measures of internalizing and externalizing symptoms, these previous associations between IGF2 methylation and mental health are commensurate with the identified relation between DNAm at this gene and BID.

The focus of this study was on early-life internal, individual differences in biobehavioural reactivity in order to extend research that has largely focused on relations between DNAm and external, environmental adversities. However, the developmental influences of such internal and external factors do not operate in isolation of each other, as demonstrated by existing literature on environmental correlates of IGF2 methylation. In a previous study of this cohort, associations were observed between early parental stress and DNAm in adolescence (see Essex et al. 2011). BID was not associated with DNAm levels in CpGs found related to parental stress; however, early parental stress was significantly associated with CpGs in IGF2 and DLX5, although not those identified here. In sum, results of the present study provide a strong foundation on which future research can further explore relations among external environmental influences, internal biobehavioural factors, and DNAm patterns.
4.4.1 Limitations

Results of the present study must be weighed in light of several limitations. First, the current study’s assessment of DNAm at ages 15 and 18 years suggests at least short-term persistence of associations with biobehavioural reactivity, but I am unable to infer when differences in DNAm may have arisen in development. It is possible that the observed DNAm pattern was established during early embryogenesis in the ectodermal germ layer, in response to allelic differences and fetal exposures. If this were true, then both BECs and neurons, with their common ectodermal origins, might be expected to exhibit comparable patterns of DNAm. While it is possible that DNAm in tissues other than BECs played a causal role in the development of differing levels of biobehavioural reactivity, DNAm and reactivity may be related in at least two other ways. DNAm patterns could result from temperamental response predispositions leaving a chemical mark on the epigenome, or a third factor, such as genetics, could have affected both reactivity and DNAm. Regarding the latter, I did not test the CpGs of interest for the influence of genetic variability although such CpGs, termed methylation quantitative trait loci, are common in the genome (148,258). The relatively small and homogenous nature of the sample precluded extensive testing of the influence of race and sex on relations between BID and DNAm, though the results were retained across models that controlled for the effects of covariates. Future longitudinal studies with larger, more heterogeneous samples are needed to build on the present findings and address more complex questions of antecedence, causality, and modifying factors.

Due to ethical and other conspicuous prohibitions on acquiring samples of brain, I studied DNAm of BECs collected from oral swabs, rather than DNAm in neural tissues, where biobehavioural patterns of response originate. DNAm is tissue-specific in nature, and there are clearly many differences among the epigenetic marks measured in buccal epithelium and brain
tissue. Current research in epigenetics is elucidating the level of DNAm concordance between brain and tissues/liquids commonly collected for DNA analysis, including saliva, BECs, and white blood cells (97,171,225). Although such research is ongoing, emergent literature suggests that DNAm located in specific genomic regions, such as CpG islands within gene coding regions, may be more highly conserved across tissues and therefore potentially informative of brain DNAm patterns (225,303).

Our results reveal age-related DNAm changes from age 15 to 18 in DLX5, IGF2, PRUNE2 or MYO16, consistent with prior research identifying increased DNAm changes in brain tissue and blood from childhood to adolescence as compared to more minimal changes that occur across adulthood (65,110). However, given the confounding of DNAm processing technology with age in this study, I cannot rule out that changes in DNAm between age 15 and age 18 resulted from differences between the 27K array and 450K technologies and differences in data processing and normalization. Additionally, differences in the cell composition of buccal samples between age 15 and age 18 may have contributed. A different study design will be required to test broadly how age is reflected in DNAm changes. Finally, to test my hypotheses related to longitudinal stability, I conducted, by necessity, a more constrained analysis of DNAm at age 18. Future analyses will explore the full range of data collected at 18 in relation to both early environmental and biobehavioural reactivity factors.

These limitations notwithstanding, the present study’s integration of pertinent information from different biological arenas (e.g., epigenetics, autonomic physiology) with psychological constructs, is consistent with a new strategy in mental health research proposed by the National Institute of Mental Health (NIMH). The Research Domain Criteria (RdoC) framework guides mental health research to recognize the broader biological and psychological
contexts of health and behaviour and to understand the phenotypic variability within clinical disorders (304,305). As such, advancing investigations into the relations among epigenetic differences, childhood behaviour problems, and patterns of stress responsivity could support earlier identification of inauspicious developmental and mental health outcomes.

In conclusion, this study revealed strong, prospective associations of observational measures of childhood inhibition/disinhibition with patterns of DNAm in BECs harvested at both mid-adolescence and early adulthood. Though current analyses do not allow for firm inferences of antecedence and causality, such associations focus attention upon possible linkages between inhibition/disinhibition dimensions of children’s temperament and behaviour and the DLX5 and IGF2 genes that have diverse developmental and regulatory functions. These findings offer provisional evidence for a developmental, epigenetic biomarker of internal biobehavioural response predilections. As with many other developmental processes, epigenetic modifications integrate the complex interactions of environmental context and constitutional biology, providing insight into developmental trajectories and long-term health outcomes.
Chapter 5: Assessing early life patterns of DNA methylation with concurrent objective and subjective socioeconomic status

5.1 Introduction

Socioeconomic status (SES) is a powerful determinant of health across the life span and has been linked to chronic diseases such as diabetes, cardiovascular and respiratory disease, and premature mortality (20,306). Low SES experienced in infancy and childhood is especially potent, exhibiting graded associations with stunted physical, social-emotional and cognitive development, in addition to latent effects, such as higher rates of morbidity and mortality (19,38,307). These associations suggest that social experiences are “biologically embedded” and leave a physical residue or imprint on the body, which has wide-reaching effects on long-term health and well-being (308,309). Currently, we are only beginning to understand the mechanisms underlying this relationship, and have yet to elucidate which aspects of SES are causal, the timing of embedding, and the biological systems involved.

Recent research into the systems underlying embedding has uncovered two, related occurrences: upregulation of inflammation and dysregulation of stress response pathways (310). In both adults and children, low SES is associated with chronic activation of the hypothalamic-pituitary-adrenocortical axis, including higher cortisol output and greater expression of the glucocorticoid receptor and inflammatory cytokines (311-313). However, the persistence of such changes, even after an individual experiences an upward trajectory in SES, indicates a more fundamental biological pathway is involved (87,314).
DNA methylation (DNAm), which typically involves the covalent attachment of a methyl group to the cytosine of a cytosine-guanine dinucleotide (CpG), is the most commonly studied epigenetic mark in humans (63,188). DNAm is paradoxically stable and plastic, in that it is mitotically heritable, maintaining patterns attributed to histological differences and genetic variation, but is also to some extent malleable by environmental exposures (94,190,315). However, epidemiological studies of DNAm patterns across human populations are complicated by the respective relationships between tissue specificity, genetic variation, ethnicity, age and DNAm patterning (150). Tissue from which the DNA is extracted is one of the strongest determinants of DNAm (95,192,316). During differentiation, cells acquire unique epigenomes which contribute to – and are affected by – unique transcription profiles and cellular identity (67,192). As such, genome-wide DNAm patterns from one tissue and different individuals are more closely related than DNA patterns from one individual and multiple tissues (95,97,316). Moreover, blood cell composition may differ by SES or related phenotypes such as birth weight, or adoption status (139,317).

Genetic variation is currently thought to account for 20-80% of DNAm variation (94,209-211). At the genome-wide level, individuals who share ethnicity or ancestry, will have more similar DNAm patterns, in large part likely due to underlying genetic variation (91,318,319). At the CpG level, inter-individuals DNAm patterns have been associated with nearby or even more distant single nucleotide polymorphisms (SNPs); SNPs at which genetic variation is associated with DNAm are referred to as methylation quantitative trait loci (mQTLs) (91,177,213). With respect to the relationship between mQTLs and gene expression, most SNPs in the genome are predicted to influence DNAm and gene expression independently, rather than demonstrate a linear relationship (85,320). Finally, age has been strongly correlated with DNAm, such that
DNAm can be used to accurately predict chronological age (241,321). The role of DNAm in aging is not well understood, but in early life, pre- and post-natal, DNAm is hypothesized to act as a critical regulator of development (65,67,322). Notably, prenatal environmental exposures have been consistently associated with DNAm changes in childhood and may mediate related long-term health outcomes (148,323,324). As early life DNAm patterns are especially dynamic, they may be more susceptible to alterations from environmental exposures (96,110,148). These features, as well as its fundamental role in regulating gene activity, make DNAm a plausible mechanism of biological embedding (314,315,325-327).

Multiple studies have demonstrated that DNAm is reliably associated with early life social class, irrespective of current SES, as well as related to inflammatory signaling, transcriptional profiles (35,87,314,327). Although in some cohorts, associations between early life SES and epigenetic patterns measured later in life depend on concurrent SES (328). Various components of SES including family income, parental education, and family psychosocial adversity have also been associated with concurrent, differential DNAm and CpGs were enriched for genes related to immune function and development (325). Taken together these findings suggest that DNAm can be linked to different components of SES, measured at times in the life course, and call for studies probing which underlying aspects the SES experience, such as education or financial standing, are most predictive of DNAm.

Experiences often stratified by SES range from those at the family level, including maltreatment or abuse, and extend to neighborhood and city characteristics, dictating quality of housing, proximity to violence, etc. (29). Extensive research has attempted to tease apart how singular exposures and experiences relate to and affect health, but the underlying relationships are often obscured by the extensive variation in SES-related experiences. Beyond this, the same
experience may be perceived and internalized differently by two individuals. This raises the question of how subjective vs. objective measures of SES compare in their relationships to later or concurrent health outcomes (49,329). Subjective measures of SES, such as an individual’s perception of their social rank, is reliably associated with health outcomes across many different populations, even when corrected for objective measures of SES (49,330,331). While previous studies have shown a significant but non-redundant relationship of traditional indicators of SES and subjective rankings of SES, the latter may encompass a broader picture of an individual’s experience, and therefore may be more closely related to his/her stress levels (332). In a sample of healthy Caucasian women, subjective social status was more consistently and strongly related to psychological functioning and health-related factors, including cortisol habituation to repeated stress, as compared to objective SES indicators (49). Increased subjective SES has also been associated with lower susceptibility to upper respiratory infection while controlling for objective SES (329). These findings suggest that psychological perceptions of social status may be an important contributor to the SES-health gradient.

In children specifically, interactions with parents act as an important buffer of SES-related experiences in early life, mitigating a child’s experience of SES-related stress and disadvantage in the home (311,333). For example, children with social support from the parent significantly reduced the cortisol stress response, as compared to support from a stranger (334). Moreover, adults reared in low-SES households who experienced high maternal warmth have decreased pro-inflammatory signaling compared to adults from low-SES households who experienced low maternal warmth (311). Given that parental interactions act to “buffer” stress and adversity in children, this raises the question of how parents’ subjective experiences of SES,
informed by stress and their own environmental sensitivity, relate to their children’s biological outcomes.

Based on extent research, I hypothesized that DNAm correlates of familial material resources, as measured by objective SES, would differ from DNAm correlates of parents’ subjective social standing, in a similar fashion to other biological of outcomes. In the present study, I interrogated DNAm in a cohort of children aged 6-11 years old and compared concurrent DNAm from peripheral blood mononuclear cells (PBMCs) and objective and subjective SES, finding unique DNAm associations in each SES measure despite their interrelatedness. Using matched neonatal dried blood spots (DBSs), I also addressed the question of whether these correlates existed at birth. Finally, I performed a targeted analysis of the CpGs correlated with SES, testing for mQTLs to explore if relationships between DNAm and SES, like other environmental exposures, are influenced by genetic variation. Taken together, these findings provide preliminary evidence of non-redundant DNAm loci associating with objective vs. subjective SES, highlighting the importance of SES measure selection in epidemiological studies and the potential role of epigenetics in biological embedding.

5.2 Methods and materials

5.2.1 Study sample

Participants in the current study were recruited as part of the Gene Expression Collaborative for Kids Only (GECKO), a cross-sectional study of epigenetic profiles, SES and family adversity. Participants ranged from 6 to 11 years old at time of first visit, 49% were females (197 of 402), and were living in the Greater Vancouver Area (Table 5.1). The present analyses are
based on a subset of 63 who provided both peripheral blood and saliva samples. Blood was collected between 6 months to 2.3 years after the first visit using Vacutainer® CPT™ Cell Preparation Tubes (Becton, Dickinson and Company, NJ, USA) and PBMCs were isolated within four hours of collection, as previously described (35). Isolated PBMC pellets were stored at -80°C prior to DNA extraction. Saliva samples were collected and preserved in Oragene collection kits (DNA Genotek Inc., ON, Canada). From a subset of 37 participants, I obtained neonatal DBSs collected from heel pricks using Whatman® Protein Saver cards (Sigma-Aldrich, MO, USA). Blood spots were collected and stored by the BC Newborn Screening Program.

Ethics approval for GECKO was obtained from the University of British Columbia Institutional Review Board and informed consent was obtained from all participants or guardians.

Table 5.1 Sample characteristics for full GECKO cohorts and subset presented in this study

<table>
<thead>
<tr>
<th></th>
<th>Full cohort</th>
<th>Subset</th>
</tr>
</thead>
<tbody>
<tr>
<td>sample size</td>
<td>402</td>
<td>63</td>
</tr>
<tr>
<td>% female</td>
<td>49% (197)</td>
<td>43% (26)</td>
</tr>
<tr>
<td>age at first visit</td>
<td>6 - 11; med = 9</td>
<td>6 - 11; med = 9</td>
</tr>
<tr>
<td>parent respondent</td>
<td>350 mother, 52 father</td>
<td>61 mother, 2 father</td>
</tr>
<tr>
<td>household income ($ CAD)</td>
<td>&lt;5,000 - ≥200,000; med = 75,000 - 99,999</td>
<td>&lt;5,000 - ≥200,000; med = 100,000-149,999</td>
</tr>
<tr>
<td>net worth ($ CAD)</td>
<td>&lt;500 - ≥500,000; med = 50,000-99,999</td>
<td>&lt;500 - ≥500,000; med = 50,000-99,999</td>
</tr>
<tr>
<td>highest education (years)</td>
<td>highschool (10) - graduate school (≥20); med= college (16)</td>
<td>highschool (11) - graduate school (≥20); med= college (16)</td>
</tr>
<tr>
<td>SES</td>
<td>2.1 - 1.57; med = 0.04</td>
<td>1.76 - 1.57; med = 0.1</td>
</tr>
<tr>
<td>Position in Canada</td>
<td>1 - 10; med = 7</td>
<td>3 - 10; med = 7</td>
</tr>
</tbody>
</table>

Note: med = median
5.2.2 SES measures

Socioeconomic information was collected during the first visit via a parent questionnaire regarding home life, income and educational background. Of the 63 participants used in the present analysis, 61 mothers and 2 fathers completed the questionnaire. A composite measure of objective SES was created by taking the average of three measures after transforming them to z-scores: household income, top education level of the two the parents, and net worth.

To acquire household income information, the respondents were asked to record gross household income from the past year. Eleven options were provided which ranged from below $5,000 to $200,000 or greater and increased in $5,000 intervals. To determine net worth, respondents were asked to record how much cash would be gained after liquidation of their assets. Finally, respondents were asked to record highest grade (in years) in school completed by either parent. Answers ranged from 11 years, which corresponded to high school, to greater than 20 years (graduate school). These measures were scaled and centered, creating z-scores, prior to averaging to create a composite measure of SES.

To obtain a measure of subjective SES, The MacArthur Scale of Subjective Social Status was used (331,335). This social ranking instrument is a simple picture of a ladder with 10 steps and has been widely validated across different populations (335,336). Here, the parent respondent was asked to indicate on the ladder where his/her immediate family stands in his/her country (Canada) in terms of income, education, and occupation. Notably, the subset used in this analysis did not represent the full range of economic diversity of the GECKO cohort, as individuals with the lowest economic measures tended not to consent to providing blood samples.
5.2.3 DNA methylation data

Genomic DNA was extracted from PBMC pellets using the Dneasy kit (Qiagen, MD, USA). DNA yield and purity were assessed using a Nanodrop ND-1000 (Thermo Fisher Scientific, MA, USA). For DBSs, The GenSolve DNA Recovery Kit, GVR100 (IntegenX, CA, USA) was used to extract genomic DNA, followed by purification using QiAmp DNA Micro Kit (Qiagen), concentration using Microcon DNA Fast Flow Devices, MRCF0R100 (Millipore) and quantification using a Nanodrop ND-1000. Next 750ng from each PBMC and DBS genomic DNA sample underwent bisulfite conversion using EZ DNAm Kits (Zymo Research) and 160ng of bisulfite-convered DNA was applied onto the Illumina Infinium HumanMethylation450 BeadChip (450K array) (Illumina Inc, CA, USA) per manufacturer’s instructions. PBMC DNA samples were run on 8 chips in one batch and DBS DNA samples were run on 5 chips in one batch. Both were scanned were scanned on an Illumina HiScan.

Following scanning of the microarrays, DNA methylation data were imported in Illumina’s GenomeStudio software for background-subtraction and color-correction; all subsequent processing steps and analyses were performed in R Version 3.2.1 (http://www.r-project.org). In R, probes were first assessed for quality control and genomic location. Of the original 485,577 probes, 73,502 were removed because they had missing beta values in at least 5% of samples; were not detected above background in at least 5% of samples (detection p-value > 0.01); mapped to SNPs; mapped the X/Y chromosomes; were polymorphic for SNPs; or were measured by cross-hybridizing probes (165). Following quality control, 412,075 probes remained in the dataset, which was then normalized using Beta Mixture Quantile dilation (BMIQ) to remove differences in dynamic range between Type I and Type II probes (166). PBMC samples were corrected for blood cell type proportion using the Houseman deconvolution.
method (169), which has been used in pediatric blood samples, including this cohort (153,197).
The algorithm outputs the relative predicted proportions of the six cell types (CD8 T cells, CD4 T cells, natural killer cells, B cells, monocytes and granulocytes), which were then used to regress out variability attributable to inter-individual differences in cell proportion (153). As granulocytes are not present in PBMCs they are treated as a “negative control” of the extraction procedure. After correction, the Houseman method was re-run on the DNAm which estimated cell proportions consistent across all samples (data not shown). The neonatal DBSs were not corrected for cell type heterogeneity for two reasons. First, neonatal peripheral blood contains nucleated red blood cells that would likely decrease the efficacy of a prediction method trained on adult blood cells. Secondly, CpGs tested in DBSs only included those found to be associated with SES in PBMCs irrespective of cell proportions. Data were corrected for microarray chip in ComBat to remove batch effects (173). Prior to analysis data were filtered for inter-individual variability in the PBMC samples. For each CpG measured in PBMCs the range between beta values at the 10\textsuperscript{th} and 90\textsuperscript{th} percentile was calculated and those with a range greater than 5\% were included in the analysis (176). Following this step, 107,295 CpGs remained.

5.2.4 Genotyping data generation and use in inferring population structure

DNA from saliva samples was extracted by ethanol precipitation and was purified and concentrated using DNA Clean & Concentrator (Zymo Research). Next, 200 ng of genomic DNA were amplified and hybridized to the Illumina PsychChip SNP genotyping array (PsychChip) (Illumina Inc, San Diego, CA) per manufacturer’s instructions. Chips were scanned and genotype data were imported into Illumina’s GenomeStudio using default parameters. Genotype calls were generated using Illumina cluster definitions with a Gencall threshold of
Genotyping data were then exported into R Version 3.2.1 for probe filtering. Each probe was assessed for genomic location and quality. Probes mapping to sex chromosomes, mitochondria, or no chromosome were removed; for quality control, probes with a low average Gencall scores (< 0.1) or an unexpected distribution (10\textsuperscript{th} percentile < 0.2) were removed. After these steps, 547,477 SNPs remained and were filtered for high pairwise linkage disequilibrium and high variability using PLINK software; this was performed to remove genomic patterns that could mask greater patterns related to population structure (337).

Next, the 223,089 SNPs that remained following SNP pruning were run through principal component analysis (PCA) as a means of data reduction. The first two principal components, which accounted for a 65.1% of the total variation (62.9% and 2.2%, respectively) and had standard deviations greater than 1, were used in analysis as measures of population structure in the GECKO cohort.

Of note, self-reported ethnicity was collected but over 100 reporters identified their children as “Canadian”. Most of these individuals clustered with individuals of European descent at PC1 and PC2 of the genotyping data, but many individuals reported to be “Canadian” clustered most closely with individuals of East Asian ancestry or multiethnic ancestry (Supplementary Figure 5.1). As such, these individuals were removed from comparisons between ethnicity and SES factors in section 5.3.1.

### 5.2.5 Statistical analysis of DNAm in PBMCs

First, two linear models were run on each of the 107,295 variables CpGs, using either objective or subjective SES as the explanatory and sex and PC1 and PC2 from the genotyping data as covariates (178). As age was not correlated with either SES outcome, it was omitted from
the linear model, to preserve statistical power. Associations with an uncorrected \( p \)-value < \( 5 \times 10^{-4} \) and an effect size threshold of absolute delta beta > 5% (\(|\Delta \beta| \geq 5\%\)) were followed-up for downstream analysis. This threshold was chosen as no sites met strict criteria for significance, i.e. FDR-adjusted \( p \)-value < 0.05, but as this was a pilot study, lower thresholds were used to identify candidate sites (152). These sites, associated with objective or subjective SES, were tested for enrichment of underlying genomic features as compared to a “null” expected count of CpGs in each region from 1,000 iterations (205).

5.2.6 Gene ontology

Gene ontology (GO) analysis was performed using the software ErmineJ (287). The 412,075 CpG, which passed quality control and normalization in PBMCs, were annotated to genes as previous described (165), generating a 450K GO annotation background. This background included sites removed prior to analysis due to low inter-individual variability, as these sites can also considered as not related to the measures of interest and therefore should be tested against in enrichment analysis. Enrichment for GO terms was then performed using over-representation analysis. The following ErmineJ parameters were used: include only “Biological Process” related GO terms, minimum gene set size of 5, maximum gene set size of 100, and test the effect of multifunctional genes. Reported \( p \)-values were corrected for both FDR and multifunctionality of GO terms.

5.2.7 Identification of differentially methylated regions

As multiple studies have reported that functionally relevant findings are likely to be associated with genomic regions rather than a single CpG, I chose to assess the genomic
locations of SES-associated CpGs as to identify underlying differentially methylated regions (DMRs) (81,338-340). Reported DMRs were required to contain at least two SES-associated CpGs, as determined by the previous linear regression, which could be no more than 1kb apart. CpGs in DMR were not required to be contiguous as discordant CpGs have be discovered within DMRs, specifically in cancer, as well as in Chapter 4 of this dissertation (341). A window of 1kb was selected as co-methylation has been found to exist most strongly within a distance of 1 kb (342). As well, methylation haplotype blocks identified from whole genome bisulfite sequencing data were all less than 1500bp, with 99.98% 1000bp or smaller in size (64).

5.2.8 mQTL analysis

For analysis of genetic influence on SES-associated CpGs, a list of mQTLs previously uncovered in the GECKO cohort were used (See chapter 3). Briefly, using 74 GECKO PBMC samples, SNPs measured on the PsychChip data were numerically coded, as 1, 2, or 3, and correlated with any variable CpG within 5kb using Spearman correlations. Pairs with an FDR corrected \( p \)-value \( \leq 0.05 \) and DNAm change per allele \( \geq 2.5\% \) in the discovery cohort (GECKO) were tested in the C3ARE cohort for validation. A total of 2,857 pairs of associated CpG-SNPs validated and were overlapped with SES-associated CpGs. SES-associated CpGs were also interrogated against mQTLs discovered in the Accessible Resource for Integrative Epigenomic Studies (ARIES cohort) (216). Blood samples collected from over 1000 children at age 7 years were used to generate these data, which were accessed through the mQTL database (http://www.mqtldb.org/). All mQTLs meeting a \( p \)-value of \( 1 \times 10^{-7} \) were downloaded, trans-mQTLs were removed and an effect size threshold of \( \geq 2.5\% \) DNAm change per allele was implemented, leaving 2,838 mQTLs associated with 28 unique CpGs for comparison.
5.2.9 Statistical analysis in dried blood spots

CpGs found to be correlated with objective or subjective SES in early life PBMC samples, were analyzed in 37 neonatal DBS samples to assess if the associations were present at birth. Of these 37 individuals with DBS samples, 32 were matched to the PBMC samples; however, all 37 samples were used to avoid further reduction of the small sample size. Identical linear models to those run in PBMC DNAm data were used for the DBS DNAm data and CpGs reported as potentially associated with SES had an uncorrected $p$-value < 0.05 and an absolute $|\Delta \beta| > 5\%$. Benjamini-Hochberg corrected $p$-values were also reported (179). Following these analyses, samples were reduced to those with matched PBMC samples (n=32) to compare effect sizes across both time points, i.e. in both PBMCs and DBSs.

5.3 Results

5.3.1 Objective and subjective SES measures in the GECKO cohort

To understand the relationship between objective and subjective SES in GECKO, a cross-sectional cohort of children aged 6-11 years, I began by analyzing these two measures in relation to each other and in relation to sex, age, and ethnicity/population structure (Figure 5.1). In this cohort, objective SES (hereafter referred to as a CompSES, an abbreviation of composite SES) was calculated as an average of three parent-reported measures: highest education completed by either parent, household income, and net worth. Subjective SES was measured using a parent-report ranking his/her social and economic position in Canada, on a scale from 1 to 10 (hereafter referred to as a PosCan) (331). Correlations were performed in both the full cohort (n=402,
CompSES was not available for 22 individuals) and the subset of children who provided peripheral blood samples from which PBMCs were isolated (n = 63, CompSES was not available for 2 individuals). A summary of demographic information and SES measures can be found in Table 5.1.
Relations between sample characteristics and SES measures in full GECKO cohort were maintained in subset used in DNAm analysis. Correlations between CompSES, PosCan, age, sex and ethnicity (principal components 1 and 2 of genotyping array) in full cohort (A and B; n=402, CompSES was not available for 22 individuals) and subset used in DNAm analysis (C and D; n = 63, CompSES was not available for 2 individuals). Colours and values in correlation heatmaps (A and C) represent Spearman correlation coefficients.
We uncovered a significant association between CompSES and PosCan in both the full cohort and the subsample (r=0.52 and r=0.6, respectively, p-values < 0.05). In the full cohort, neither CompSES, nor PosCan was associated with sex or age, however in the subsample of families whose children provided blood samples, CompSES was associated with sex (r=0.29, p-value<0.05). As such, sex was included as a covariate in later DNAm analyses.

As DNAm reflects in part genetic ancestry, I built a molecular genetic classification of ancestry by applying PCA to data from a genotyping array (91,318). Of the full cohort, 304 individuals provided saliva samples, from which the extracted DNA was run on the Illumina PsychChip genotyping array to assess population structure. Using principal component analysis, population structure was defined by component 1 (PC1) and component 2 (PC2) (Supplementary Figure 5.1). In both the full cohort (n=304) and the subsample (n=63), PC2 was significantly correlated with PosCan (p-values < 0.05), but not CompSES. Of note, PC2 mainly separated individuals who self-identified as East Asian or Southeast Asian from Caucasian, suggesting a potential trend in the reporting of subjective SES, consistent with previous findings of subjective SES differing by demographic group (336). However, a t-test of PosCan in all Caucasian (n=61) vs. Asian and Southeast Asian individuals (n=64) did not reveal significant differences between these groups (p-value=0.12). Taken together, these results indicated that while the full GECKO cohort was generally balanced for sex, age, and ethnicity across the SES gradient, the subset of individuals who gave blood samples were slightly more skewed with regard to these variables. These findings were incorporated into the design of the subsequent epigenetic analysis intended to elucidate the DNAm correlates of early-life objective and subjective SES.
5.3.2 Subjective SES and objective SES tended to associate with different CpGs

DNAm extracted from the PBMCs of 63 participants was measured across ~485,000 CpGs genome-wide using the Illumina Infinium HumanMethylation450 BeadChip (450K, Illumina Inc, San Diego, CA). After filtering for non-variable probes and correcting for cell-type heterogeneity, 107,295 CpGs remained and were included an analysis of DNAm correlates of objective and subjective SES. Notably, predicted cell type proportions were not significantly correlated with PosCan, nor CompSES (data not shown).

Following data processing, I ran two linear regression models, using either PosCan (subjective) or CompSES (objective) as the explanatory variable, and included sex and PC1 and PC2 of the genotyping data as covariates. In cognizance of the small sample size, I used a liberal \( p \)-value threshold of \( 5 \times 10^{-4} \) when reporting SES-associated CpGs (152). In PosCan, 98 CpGs met the criteria for associated loci (\( p \)-value < \( 5 \times 10^{-4} \), \(|\Delta \beta| > 5\% \)) and in CompSES, 63 CpGs met these criteria (Figure 5.2; Supplementary Table 5.1); no sites were significant following FDR correction, at a Benjamini-Hochberg corrected \( p \)-value of 0.05. A total of six CpGs were associated with both PosCan and CompSES (Table 5.2) and while this overlap was significantly greater than what was expected by chance based on 20,000 Monte Carlo simulations (\( p \)-value < \( 2 \times 10^{-4} \)), it was lower than anticipated based on the strong association between PosCan and CompSES.
Figure 5.2 Genome-wide linear regressions showed limited numbers of CpGs associated with each SES variable. Associations between PosCan (left) and CompSES (right) with DNAm at 107, 295 variable CpGs. Volcano plots of each CpG showing effect size (db) vs. p-value. Points plotted in blue are CpGs found significant at an $|\Delta \beta| > 5\%$ (vertical lines) and $p$-value $< 5 \times 10^{-4}$ (horizontal lines).

Table 5.2 CpGs associated with both PosCan and CompSES ($|\Delta \beta| > 5\%$ and $p$-value $< 5 \times 10^{-4}$)

<table>
<thead>
<tr>
<th>CpG</th>
<th>CompSES p-value</th>
<th>CompSES $\Delta \beta$</th>
<th>PosCan p-value</th>
<th>PosCan $\Delta \beta$</th>
<th>Chromosome</th>
</tr>
</thead>
<tbody>
<tr>
<td>cg02556042</td>
<td>2.59E-04</td>
<td>-0.101</td>
<td>3.78E-04</td>
<td>-0.105</td>
<td>12</td>
</tr>
<tr>
<td>cg02966332</td>
<td>4.50E-04</td>
<td>-0.059</td>
<td>2.60E-04</td>
<td>-0.064</td>
<td>20</td>
</tr>
<tr>
<td>cg06650914</td>
<td>2.29E-04</td>
<td>-0.053</td>
<td>1.18E-04</td>
<td>-0.060</td>
<td>18</td>
</tr>
<tr>
<td>cg07033961</td>
<td>1.56E-04</td>
<td>0.059</td>
<td>4.96E-04</td>
<td>0.057</td>
<td>5</td>
</tr>
<tr>
<td>cg18808777</td>
<td>2.05E-05</td>
<td>0.089</td>
<td>2.48E-04</td>
<td>0.083</td>
<td>6</td>
</tr>
<tr>
<td>cg20320656</td>
<td>2.07E-04</td>
<td>-0.125</td>
<td>4.17E-04</td>
<td>-0.118</td>
<td>12</td>
</tr>
</tbody>
</table>

Further examination of the CpGs associated with both PosCan and CompSES trended toward an inverse relationship between DNAm and SES. For CompSES 48 of 63 CpGs exhibited
a negative correlation and for PosCan, 80 of 98 CpGs exhibited a negative correlation. While one might speculate that lower SES has a repressive effect of gene activity via increased DNAm, the location of these CpGs relative to genes suggested otherwise. Both sets of associated CpGs were tested for enrichment in gene features and CpG islands relative to all CpGs that were analyzed (107,295 CpGs; Figure 5.3) (205). Using 1000 permutations of randomly selected CpGs, PosCan-associated sites were found to be significantly depleted at gene promoters, the regions at which the inverse relationship between DNAm and gene activity is strongest (FDR-correction p-value <0.05). By comparison, CompSES-associated CpGs did not significantly differ in their genomic locations from the “background” set of CpGs. Finally, I performed Gene Ontology analysis using ErmineJ to identify any enrichment of genes in shared biological pathways. For both PosCan and CompSES, no biological pathways were significant after correction for multiple testing and multifunctionality.
Figure 5.3 Limited enrichment of genomic features found in CpGs associated with CompSES or PosCan. Representation of enrichment or depletion of various genomic features mapping to A) 98 CpGs associated with Poscan and B) 63 CpGs associated with CompSES. Height of bars denote the fold-change between actual CpG count in each genomic region and the mean count of randomly selected CpGs in that same genomic feature, based on 1000 iterations. Error bars show standard error (* denotes significant enrichment or depletion at at Benjamini-Hochberg corrected p-value < 0.05) (S = South; N = North).
5.3.3 Loci associated with objective SES contained more underlying DMRs than those associated with subjective SES

Given that the reported CpGs had increased likelihood for false positives due to the relaxed $p$-value threshold, I employed additional methods to improve the probability of reporting true associations. To that end, I took advantage of co-methylation (i.e. the dependent nature of DNAm at neighbouring CpG sites) and used a region-specific approach (342,343). Starting with the 98 PosCan-associated CpGs and 63 CompSES-associated CpGs, I identified differentially methylated regions (DMRs) containing two or more sites within 1kb of each other. This uncovered four CompSES-associated DMRs and two PosCan-associated DMRs (Figure 5.4, Supplementary Table 5.2). Of the four CompSES-associated DMRs, the largest was composed of six CpGs, all within approximately 300bp of each other and located on the island shore of a promoter CpG island upstream of ASCL2, achaete-scute family bHLH transcription factor 2. As well, one CompSES-associated DMRs, composed of two CpGs within the HCP5 gene, contained one CpG (cg18808777) which was one of the six CpGs found to associate with both CompSES and PosCan. By contrast, both PosCan-associated DMRs were composed of two CpGs and the single DMR that mapped to a gene was intragenic to the HOXA6 gene, encoding homeobox protein Hox-A6. While I used a liberal threshold to define associated CpGs (uncorrected $p$-value $< 5 \times 10^{-4}$), the finding of DMRs associated with CompSES or PosCan strengthened the likelihood of these loci reflecting true associations.
Figure 5.4 CompSES found to be associated with more DMRs than PosCan.
DMRs (differentially methylated regions) associated with CompSES (A) or PosCan (B). DMRs were identified as loci containing 2 or more CpGs associated with SES found no greater than 1kb apart (nominal $p$-value $< 5 \times 10^{-4}$ and $|\Delta \beta| > 5\%$). Represented are the four DMRs associated with CompSES, mapping to $ASCL2$, intergenic, $CRYGD$ and $HCP5$, respectively, and two DMRs associated with PosCan, mapping to $HOXA6$ and intergenic.
5.3.4 Few SES-associated CpGs were also driven by genetic variation

Given the emerging insight that DNAm patterns in variable CpGs are best modelled by an interaction between genetic variation and environment, I assessed the genetic contribution to the SES-associated CpGs (148). First, I utilized existing associated CpG-SNPs previously discovered in GECKO PBMC samples and independently validated (see Chapter 3) (153). I intersected these mQTL-associated CpGs with those associated with Poscan or CompSES. Six of the 98 PosCan-associated CpGs were also associated with SNPs, i.e. mQTLs, and three of the 63 CompSES-associated CpGs were also associated with mQTLs (Figure 5.4). Similarly, I compared the CpG sites to mQTLs discovered in the ARIES cohort, a more appropriately sized cohort for mQTL discovery, and found 28 of the 98 PosCan CpGs were associated with ARIES mQTLs, while just one of the 63 CompSES-associated CpGs was also associated with ARIES mQTLs (216).

Figure 5.5 Little overlap between CpGs associated with PosCan or CompSES and mQTL-associated CpGs. Overlap of SES-associated CpGs with previously reported mQTL-associated CpGs, identified in the GECKO cohort.
5.3.5 Assessing DNAm in neonatal dried blood spots

Given that both pre- and postnatal SES environments have been associated with health outcomes, the timing of the biological embedding of SES is difficult to predict. As such, the SES-associated DNAm patterns in the GECKO cohort may have been established prenatally. I used 37 DBS DNAm samples collected at birth to test whether the DNAm patterns observed in PBMCs from childhood were also present at birth, with the caveat that SES measures were collected up to 11 years postnatally.

CpGs previously associated with SES, CompSES and PosCan, in PBMCs were tested for validation in the DBSs using identical linear models to those run on the PBMC DNAm data. Of the 63 CpGs associated with CompSES in PBMCs, 12 had a \( p \)-value < 0.05 and \( |\Delta \beta| > 5\% \) in DBSs (Supplementary table 3). However, only a single CpG remained significant following FDR correction for the 63 tests run (cg14087413; FDR-corrected \( p \)-value < 0.05). Similarly, 14 of 97 CpGs tested in DBSs were nominally associated with PosCan (Supplementary table 3); five of which remained significant following multiple test correction (cg03609398, cg14453935, cg15282632, cg16443667, cg27034935; all FDR-corrected \( p \)-values < 0.05). Despite differences in age, sample size and cell-composition, the effect sizes in PBMCs and DBSs at these 26 sites were strikingly similar (Figure 5.5a). All CpGs validated in DBSs were uniquely associated with either PosCan or CompSES and no two CpGs mapped to the same gene. However, five CpGs were previously identified as belonging to SES-associated DMRs (cg24353535, cg25429719, cg23129930, cg02315096, cg18808777) (Figure 5.5b). Cg18808777, located in the \( HCP5 \) gene, was associated with both PosCan and CompSES in PBMCs; located within a CompSES-associated DMR; and associated with PosCan in DBS DNAm. Replicating associations using neonatal DBSs indicated that a portion of SES-associated DNAm patterns were present at birth.
and provided an additional level of support for the relationship between DNAm and both CompSES and PosCan.
Figure 5.6 CpGs associated with SES in both PBMCs and DBSs showed similar effect sized across time points.

Associations between CompSES and PosCan with matched DNA methylation (DNAm) from neonatal dried blood spots (DBSs). A) Effect sizes (Δβ) of corresponding CpGs measured in PBMCs in childhood and DBSs at birth (n=32). Blue points represent CpGs associated with either PosCan or CompSES at both time points (p-value < 0.05); grey points are associated with SES at childhood but not at birth. B) DNA methylation at birth at CpGs located within DMRs, previously identified from childhood samples, associated with CompSES (above) or PosCan (below) (n=37).
5.4 Discussion

Biological embedding of SES and adversity in childhood results in life long health disparities and there is mounting evidence that DNAm patterns can reflects this phenomenon (19,87,314,315). However, SES is a complex construct and the various methods used define or measure SES, such as subjective vs. objective, affect its association with health outcomes (49,329,332). Hypothetically, epigenetic correlates of SES may also differ by objective vs. subjective measures. This study represented the intersection of previous findings, using parents’ perceived social standing to predict DNAm patterns in children and compared these findings to those generated with more traditional measures of SES.

Testing the DNAm patterns of variable CpGs from PBMCs in childhood, I found surprisingly little overlap between CpGs associated with each measure, given the strong correlation of between the measures themselves. While there were a large number of CpGs that met the initial $p$-value and effect size thresholds, only two DMRs, each containing two CpGs, were found within PosCan-associated CpGs. By comparison, the CompSES-associated CpGs contained four underlying DMRs, the largest of which was comprised of six CpGs, approximately 1kb upstream of *ASCL2* (Achaete-scute complex homolog 2). Of note, *ASCL2* is most strongly expressed in the placenta, functioning in trophoblast differentiation, but is also expressed in tissues of the gastrointestinal tract and in skin (344). Leveraging the available genotyping data, I found limited genetic influence on SES-associated CpGs at the fairly coarse level of granularity offered by the approach. However, there was a striking difference in number of ARIES mQTLs associated with PosCan vs. CompSES, 28 CpGs and one CpG, respectively. Finally, I found a subset of CpGs which were validated in neonatal DBSs despite that at birth,
individuals having only been “exposed” to SES in utero. Taken together, these results reflect unique patterns of association between DNAm and objective and subjective SES.

One main objective of this study was to test whether subjective SES was more strongly related to DNAm than objective SES. Previous findings have indeed suggested that subjective SES may better predict a variety of health outcomes as compared to traditional, objective economic measures. This is likely due the role of perceived social standing in chronic stress. As well financial and overall parents’ stress in childhood have been linked to discipline styles and responses, and has also been associated with differentially DNAm in the child (146,345). Neither PosCan, nor CompSES stood out as being more strongly linked to DNAm than the other and there was no considerable overlap between the associating CpGs. This finding supports previous research findings of objective and subjective SES representing intersecting but unique facets of the SES experience (329,346). As well, previous research of DNAm correlates of family adversity, parent education levels and income also found little overlap between CpGs associated with each variable (325). In the GECKO cohort, the subjective SES measure, PosCan, was reported by a parent and therefore may be more closely related to the child’s experience of his/her parent’s stress and well-being, than CompSES. Essex et al. (2011) linked exposure to parents’ stress in infancy and preschool to DNAm patterns in adolescents, 10-15 years later. The CompSES CpGs overlapped with one gene previously reported gene in Essex et al., IGFAS; while the PosCan CpGs overlapped with three previously reported genes, ARSG, MOBP, and TBX1. The replication of these genes in independent cohorts, assessing related environmental exposures, is promising and warrants further investigation of these candidate biomarkers of early exposure to stress in caregivers.
Incorporating genetic variation with the DNAm findings also highlighted the difference in the epigenetic correlates of PosCan and CompSES. As previously noted, PosCan was likely more strongly correlated to parental stress and mental health, which have genetic components, than CompSES (49,347). As well, research into DNAm changes in psychiatric disorders has identified mQTLs as contributing to the epigenetic variation associated with diagnosis (138,348,349). Taken together, it is tempting to speculate that the PosCan-associated DNAm “signature” contains more mQTL-associated CpGs than the CompSES-associated signature, which may be related to the heritable, genetic components of mental health. Although gene ontology analysis on these CpG sites did not garner significant results, multiple genes containing CpGs associated with PosCan were related to brain function and mental health; these include \textit{CUX2} (cut-like homeobox 2), \textit{GRIK1} (glutamate receptor, ionotropic, kainite 1) and \textit{MOBP} (myelin associated oligodendrocyte basic protein) (350-353).

Finally, taking advantage of the matched, longitudinal samples, I was able to test the replicability of the SES-associated CpGs in neonatal DBSs. Effect sizes across these time points were strikingly similar, suggesting that the correlation between SES and DNAm was stable over time and across different tissue compositions, whole blood versus PBMCs, at least for some CpGs. In addition to upholding the results of the PBMC analysis, these findings supported the knowledge that SES experienced \textit{in utero}, i.e. social and economic experience of a pregnant mother, has long term outcomes on the child and underscore the value of DNAm as a biomarker of these experiences (354,355).
5.4.1 Limitations

The sample size was small, especially relative to the size of the whole GECKO cohort, which led me to use lenient statistical cut-offs, increasing the rate of false positives in the SES-associated CpGs I reported. Due to the sample size, I also chose to analyze males and females together rather independently, which may have obscured sex-specific responses. These findings, therefore, should be tested replication in an independent cohort to determine whether or not the associations are replicable. The approach of forgoing multiple test correction for validation using an independent cohort multiple, has been recently proposed and utilized (152,245). However, I believe the additional analyses run on these sites somewhat offset these concerns and provided a biological perspective to increase confidence. After using an initial $p$-value threshold to report SES-associated sites, I then further interrogated these sites for underlying DMRs and for their replicable associations over time using neonatal samples.

We chose not to run a full-scale EWAS on the DBS DNAm data due to a number of limitations. Firstly, I did not have SES measures collected concurrent to when the neonatal DBSs were collected, therefore these analyses assume that social and economic standing of these families did not greatly change. Ideally, it would have been preferable to have concurrent measures to be able to treat the data as longitudinal. Secondly, while the PBMCs were corrected for cell type heterogeneity across individuals, the DBSs were not. Blood cell type proportions have been previously shown to associate with SES status, which is likely related to the pro-inflammatory effects of low SES (87,328); therefore, differential methylation in the DBSs might have been driven by differences in cell type heterogeneity. However, all CpGs were differentially methylated by SES in the DBSs, were also associated with SES in the PBMCs irrespective of cell type proportion. Therefore, it is unlikely that findings in the DBSs were
entirely driven by cell type. As such, I felt the limitations of the study design were mitigated by my careful selection of analytic methods.

In addition to DNAm, early life SES has been associated with other molecular characteristics such as blood cell proportions, transcription levels of genes in inflammatory pathways, and circulating hormones (87,328,356). Of note, I found that neither CompSES, nor PosCan, was associated with PBMC cell proportion as predicted by a validated deconvolution method (169,170). While previous publications have reported such associations between early life SES and blood cell proportions, other factors may have obscured such relations within this cohort. For example, sex and ancestry have all been shown to alter blood cell counts in healthy individuals (357,358). As well, differences in blood cell proportion may be a latent effect which only arises in adulthood, similar to many SES-associated health outcomes. I was not able to test SES against blood cell proportions at birth as current \textit{in silico} cell type predictions are less reliable in newborn blood, in part due to the presence of nucleated red blood cells. Although, it is a plausible hypothesis that the cell composition of neonatal whole blood be receptive to prenatal, SES-related exposures. Finally, based on the correlative nature of these analyses I cannot infer causation, but rather evidence of an association between early life SES and DNAm.

The data presented here provide early evidence of the unique means through which subjective and objective SES in early life may be biologically embedded. Potential epigenetic sites of biological embedding uncovered in this study warrant further investigation to understand the timing of DNAm patterning; to tease apart what specific environment or exposure; and to determine causality. Moreover, I highlighted the importance of integrating genetic and epigenetic data, finding a number of SES-associated CpGs under some level of genetic influence. DNAm represents a potential stable biomarker of childhood stress and adversity, which could inform an
individual’s susceptibility to SES, developmental trajectories and risk for future negative health outcomes.
Chapter 6: Conclusions

6.1 Summary of dissertation results

DNA methylation (DNAm) has become a popular tool in epidemiology to link environmental exposures or complex traits to a quantitative and often meaningful biological outcome. However, the more I learn about DNAm, the more complex the picture becomes. The path from exposure to DNAm alteration, to ultimate health outcome is often not linear. Timing of the exposure, tissue of origin and its cellular composition, as well as genetic variation, and likewise ethnicity, are all related to DNAm patterns and therefore may confound or obscure epigenome-wide association study (EWAS) observations (94,148,359-361). As well, EWASs do not elucidate causation, and relationships may be reversed in that phenotype precedes DNAm aberrations (94). With this in mind, this dissertation began with an investigation into the contributions of genetic variation and tissue of origin on the DNA methylome. These findings informed the design and analysis of the following two chapters, which focused on uncovering loci differentially methylated by internal characteristics and environmental exposures.

Tissue of origin is one of the largest contributors to DNAm variation, so much so that based on hierarchical clustering, two DNAm profiles from different individuals but the same tissue show greater similarity than two DNAm profiles from the same individual but different tissues (95,97,316). This has prompted increasing interest in the study of tissue-specific DNAm variation. In chapter 3, I presented a systematic comparison of genome-wide DNAm patterns between matched pediatric buccal epithelial cells (BECs) and peripheral blood mononuclear cells (PBMCs). My co-first author, S. Islam and I found that BECs had greater inter-individual DNAm variability than PBMCs and that highly variable CpGs were more likely to be positively
correlated between the tissues. CpGs that were both variable and correlated were enriched for methylation quantitative trait loci (mQTLs), i.e. sites under genetic influence, suggesting that a substantial proportion of the DNAm variation common between tissues can be attributed to genetic variation. Current estimates of the genetic influence on DNAm variation ranges from 20-80% (83,93,94). This finding raises questions regarding what mechanisms dictate whether mQTLs are tissue specific or shared between multiple tissues. Plotting mQTLs specific to one tissue often revealed consistently high methylation patterns at the linked CpG across all individuals in the other tissue. Based on this observation, one could speculate that high DNAm, when located in CpGs islands at gene promoters, may be involved in repression of gene activity and this role may override genetic contribution to DNAm variation. Supporting this postulation, CpGs with the largest DNAm differences between three cell types isolated from cord blood, were found to be enriched for eQTMs, expression quantitative trait methylation loci, i.e. CpGs with significant pairwise correlation with gene expression at a single gene (85,320).

At a genome-wide level, the prevailing relationship between SNPs, gene expression and DNAm demonstrates that genetic variation influences DNAm and gene expression independently, with no linear relationship between the three (85,320). As well, a small proportion of transcriptome-wide RNA levels have been correlated epigenetic changes shown to be independent of genetic variation (211); although, directionality cannot be assumed and transcriptional levels may be altering DNAm. One potential future evaluation of this relationship could assess whether genes expressed in multiple tissues, such as housekeeping genes, are be more likely to contain cross-tissue mQTLs than genes that dictate tissue or cell specificity. A positive finding would support the function of DNAm “repressing” the relationship between
genetic variation and expression, known as expression quantitative trait loci (eQTL), in tissue-specific genes.

In this chapter, I also categorized each CpG assayed on the Infinium HumanMethylation450K (450K) array, as being 1) associated with an mQTL, 2) differentially methylated between tissues, and 3) concordant between tissues. I assessed published EWASs for representation of each category. This analysis uncovered a number of mQTLs within published EWAS “hits” suggesting these analyses may be inadvertently uncovering genetic variants associated with their traits of interest. A high percentage of CpGs associated with autism spectrum disorder were also associated with SNPs, i.e. mQTLs, as compared to early-life aging, onset of puberty, childhood psychotic disorder, and fetal alcohol spectrum disorder. As such, I hypothesize that heritability of the trait of interest may dictate the discovery of mQTLs in EWAS results. Although different analytical approaches could plausibly be responsible for the differences in the underlying mQTLs. Regardless, for EWASs run on environmental exposures or experiences, identifying mQTLs may still be of interest. For example, CpGs associated with genetic variation in cohorts stratified for environmental exposures may be indicators of gene by environment interactions. Such CpGs, which are associated with both the exposure and genotype, may be located in genes that confer environmental sensitivity or resilience.

In Chapter 5, I ran a conventional EWAS on socioeconomic status (SES) in PBMC samples from the GECKO cohort. Childhood experiences of SES can impact developmental trajectories and lead to health disparities later in life, suggesting that they become ‘biologically embedded’ (19). However, SES is a broad, multifactorial construct that can be measured in many ways and at multiple levels – the societal level, the community or neighborhood level, or the individual level – depending on the objective of the research (29). Here, I characterized SES in
two ways: 1) an objective, composite measure of household income, education level and net worth and 2) a subjective measure of parent respondents’ social and economic rank in comparison to the rest of Canada. Previous research of subjective SES, suggests that the latter may represent a more encompassing view of life experiences and, in some scenarios, can better predict health outcomes than objective SES measures \((306,329)\). Therefore, by comparing the EWAS results of objective vs. subjective SES, I examined whether this outcome also occurs with respect to DNA methylation (DNAm).

I ran a linear model across all variable CpGs using either objective SES, \(CompSES\), or subjective SES, \(PosCan\), as the explanatory variable. After reducing the associated CpGs to those found within differentially methylated regions (DMRs), \(CompSES\) showed a stronger association with DNAm. Specifically, I found four DMRs, containing a total of 14 CpGs, to be associated with \(CompSES\). By comparison, \(PosCan\) was associated with two DMRs, each containing two CpGs. I further analyzed the SES-associated CpGs, rerunning the same model on DNAm from dried blood spots (DBSs) collected at birth. Of the 98 \(PosCan\)-associated CpGs found in PBMCs, 14 were also associated in DBSs and of 64 \(CompSES\)-associated CpGs, 12 CpGs remained associated in DBSs. Despite, DBSs remaining uncorrected for cell type composition, the effect sizes of the CpGs significant in both tissues, measured at birth and in early childhood, were strikingly similar.

To clarify whether the SES-DNAm associations observed in both early childhood and at birth were influenced by genetic variation, I intersected these CpGs with mQTL-associated CpGs (identified in the previous chapter). Surprisingly, only two stably associated CpGs were also associated with mQTLs. I also overlapped these sites with mQTLs found in the ARIES cohort finding approximately one third of \(PosCan\) CpGs were ARIES mQTLs, while just one
CompSES-associated CpGs was also associated with ARIES mQTLs. Overall, this provided support that the stable association found between SES and DNAm was not strongly driven by underlying genetic differences within the cohort. An obvious limitation to this analysis was the lack of SES information from birth, although family SES tends to be stable throughout childhood (362).

As outlined in the introduction and tested in Chapter 5, a growing body of research has documented associations between childhood environments and DNA methylation, highlighting epigenetic processes as potential mechanisms through which early external contexts influence health across the life course. In Chapter 4, I tested a complementary hypothesis: indicators of children’s early internal, biological and behavioural responses to stressful challenges may be linked to stable patterns of DNAm later in life. Using the longitudinal WSFW cohort, I combined temperament, mental health and autonomic reactivity traits measured across infancy and childhood to generate a broad, composite measure of inhibition/disinhibition. This measure, termed biobehavioural inhibition/disinhibition (BID), separated inhibited children with strong withdrawal negativity and heart rate response to tests designed to elicit ANS activity from disinhibited children with higher anger and externalizing symptoms. Importantly, this composite biobehavioural reactivity measure was correlated with a number of DNAm sites measured from buccal epithelial cells (BECs) at age 15. This included multiple CpGs in each of the following genes: DLX5, IGF2, MYO16, and PRUNE2.

Building on this finding, I tested whether the associations in these four genes, measured at age 15, persisted in DNAm measured at age 18 in the same individuals. Both DLX5 and IGF2 contained CpGs whose DNAm was correlated with biobehavioural reactivity at both ages despite these sites not displaying stability across the three years. Therefore, both age and early life
biobehavioural reactivity are reflected in DNAm in BECs at the sites. At a conceptual level, this result is in line with a common finding of EWASs, i.e. that DNAm variability in any given CpG may be best explained by multiple variables (148,359-361).

One common finding between chapters 4 and 5 was the location of associated CpGs in imprinted genes, which are genes that are differentially expressed depending on parent of origin and are often associated with differentially methylation at an imprinting control region; the gene promoters of imprinted genes often also display intermediate DNAm levels due to monoallelic methylation dependent on the parent of origin. In the GECKO cohort, multiple CpGs in ASCL2 were associated with objective SES and in WSFW, multiple CpGs in IGF2 were associated with biobehavioural reactivity. These associations are likely the result of different mechanisms as SES is an environmental factor and biobehavioural reactivity is an independent, “internal” factor.

Imprinted genes are important for fetal growth and development and therefore are tightly regulated (363). However, the findings presented here support and broaden a theory postulating that DNAm at imprinted genes are more susceptible to environmentally induced epigenetic changes (364,365). Previous studies of both “internal” and “external” factors including early-life, maternal socioeconomic status and neurobehavioural measures have been associated with inter-individual differences in DNAm at imprinted genes (366-368). In future, such patterns should be assessed with considerations for representation of imprinted genes on the 450K array; whether reported CpGs are part of imprinting-specific differentially methylated regions and are truly mono-allelically methylated; and if the proximal genes are imprinted in the measured tissue.
6.2 Limitations

Unlike disease states or chemical exposures, social determinants of health and psychological/behavioural measures in healthy individuals often have a much smaller signal in the epigenome (369). These relatively small effect sizes, paired with the multiple testing problem that plagues most –omics research and the typically larger effects of age, sex, and ethnicity on DNAm, make identifying true associations in challenging. As no cohort is ideal, there were a number of limitations to these study designs, many of which are common amongst human EWASs (94). EWAS results, like GWAS findings, cannot conclude causation, only correlation. As yet, it is unknown what function, if any, DNAm alterations play in “embedding” SES. In studies such as the one performed in chapter 5, reverse causation is not possible, i.e. DNAm cannot influence SES, however there are multiple alternative hypotheses beyond direct “biological embedding” or cellular reprogramming. For example, DNAm patterns may be indicative of certain genetic variants which have been associated with education levels and social-class mobility (370). Another possibility is that the developmental changes that accompany low SES may somehow result in epigenetic alterations rather than the exposure in itself. With the study of internal traits or phenotype, the possibilities become greater. In chapter 4, I discussed the numerous pathways through which DNAm and biobehavioural reactivity may become associated. Not being able to infer a pathway can make interpreting EWAS findings challenging. One way to elucidate the relationship between the trait of interest and DNAm is to gather longitudinal data. This is especially important to DoHAD studies in which the timing of exposure may directly alter the outcome. However, longitudinal information in epidemiological studies cannot clarify the relationship between DNAm and prenatal exposures, inborn traits, or
phenotypes with a strong genetic component including biobehavioural reactivity. Such clarification will likely require animal models or cell culture.

Two common issues arise with tissue selection in human cohorts: 1) accessibility of tissue of interest, and 2) cell composition of tissue. With respect the former, surrogate tissues are often used in lieu of inaccessible, primary tissues. In chapter 4, I associated BEC DNAm with behavioural measures, although brain tissue would have been ideal. Due to the tissue-specific nature DNAm patterns, which I interrogated in Chapter 3, use of surrogate tissues can make interpretation and generalizability of results difficult. Multiple publications have assessed the relationships between brain DNAm and DNAm from peripheral tissues, such as cord blood, peripheral blood and saliva, as well as generated tools to help infer brain DNAm (171,204,225). However, this research has also revealed a stark difference in DNAm patterns between brain regions (97,204,225). Nonetheless, finding an association in DNAm from a surrogate tissue can still be of great value in its use as a biomarker. As a biomarker, DNAm differences can aid in determining disease risk, disease detection or past exposures, irrespective of biological mechanism (369). Therefore, depending on the objective of the research, the use of surrogate tissues or peripheral tissues is not necessarily hindered by the fact that DNAm is so strongly driven by tissue of origin.

Cell composition within a tissue is also major confounding factor for epigenetic studies. Multiple studies have shown that previous epigenetic associations were either confounded with or driven by cell composition (197,371). For example, if the underlying proportions of blood cells change with age, then comparing DNAm across ages could indicate a change in DNAm where there is none (372). Many solutions have arisen to try to address this problem, including using sorted cells and incorporating cell counts into statistical models. In blood, many in silico
cell prediction methods have been generated to address the issue, including those which use the methylomes of sorted cells as references and so-called “reference-free” methods which estimate underlying linear components representing cell types (169,170,373). These methods are not always in agreement with regards to the predicted proportions and selecting the “best” one can be challenging given the algebraic knowledge required to dissect the underlying methodology. Nonetheless, researchers are currently applying these methods to tissues beyond peripheral blood cells (171,374). In chapters 3 and 5, I ran cell-type correction on BECs and PBMCs prior to running analyses, while in chapter 4, I chose to leave BECs uncorrected for underlying cell composition because BEC cell proportions were not correlated the biobehavioural measures. Both of these methods were and are imperfect. Both methods would likely fail to identify signatures that are present in a cell type that constitutes only a small fraction of the tissue as such signatures would be obscured by signals from dominant cell types unless measured in isolated cells. Although, cells present in such low proportions in accessible tissues may not be practical choices for the identification of biomarkers. As no ideal method currently exists for addressing the underlying cell composition of any given tissue, this is a common limitation of DNAm analysis and the subject of many debates within the epigenetics community.

Beyond the need to address or correct for cell composition in EWAS design, cell composition is at the center of another discussion, i.e. whether environmental exposures do in fact trigger cellular reprogramming, as measured by DNAm, and ultimately change phenotypic outcomes. An emerging theory has posited that differential DNAm found in EWASs can be attributed to changes in subtypes of cells (94). More specifically, the current model of reprogramming describes mature cells “responding” to environmental exposures with altered DNAm patterns, whereas this new model, termed polycreadism, suggests that these differences
could arise due to differentiating cells responding to perturbations and altering cell fate. If true, polycereodism would require new means of defining cell types and subtypes beyond the current classification using histology and morphology, such as utilizing single-cell technologies to confirm a mosaic population of cells previously thought to be homogeneous (183). Although this is an interesting theory and has many implications for understanding the mechanisms underlying social epigenetics, the final association between environment and DNAm, or phenotype and DNAm would still stand. To that end, current implications for this theory based on the findings presented in this dissertation are difficult to surmise. For example, results in chapter 5 support DNAm as a means of biological embedding of SES. If these DNAm differences were a result of SES altering cell fate rather than DNAm patterns of a mature cell, all hypothesizes regarding biological embedding would still hold true. Further research is required to resolve these somewhat competing theories and the concept of polycereodism may alter how epigeneticists design studies and interpret findings.

6.3 Future directions

Knowing the common normative range for a specific CpG, in a given the tissue and at a given age, as well as potential effectors of variability, such as nearby genetic variation or specific environmental exposures, is the ultimate goal of the research presented in this dissertation. If pursued over the life course, this would essentially create something of a “growth chart” of DNAm across the life span and in situations in which aberrant DNAm precedes or quantifies negative health outcomes, could facilitate faster interventions by healthcare providers, whether it be additional monitoring or diagnosis and treatment. Currently work in genomics is aided by a plethora of online resources and repositories of genetic data, in which one could look up gene
function, common variants, expression patterns across tissues, phenotypic-associations, etc. Hopefully in the future such resources will be available for epigenetic marks, including DNAm. DNAm has already been shown to provide accurate information on the pathogenicity on variants of unknown significance, as well as diagnostic efficacy in disorders with overlapping phenotypes (375,376). Beyond this, DNAm may provide information regarding the sensitivity of individuals to environments and exposures for use in individualized treatments or interventions, following the trend of personalized medicine.

Future changes to epigenetics research, nearer than the creation of a DNAm growth chart, will hopefully address the issues in methodology used for EWASs. Epigenetics is a relatively nascent field, which means that the technologies and statistical methods are rapidly changing as more knowledge is gained about DNAm itself. Not only is the treatment of cell composition a matter of ongoing discussion, there is also no general consensus on processing and normalization methods, statistical testing, and most polarizing, interpretation of EWAS findings. The latter has spurred debates within the epigenetics field, often through the use of social media platforms such as Twitter or personal blogs, to evaluate and critique the latest publications. This open dialogue, from an outspoken community has arisen out of a desire to raise the standards of practice in the field. However, it is conceivable that it has also diminished the merit of epigenetic epidemiology in the eyes of outside researchers and scientists. Personally, I believe the community would benefit from more research into and rigorous testing of the methodologies used in EWAS analyses. Rather than generating new algorithms and R packages, assessing those currently in use, through common performance metrics, would be greatly beneficial. This also brings into question whether the publication process has been effective, given the common occurrence of new publications receiving negative feedback for inappropriate methods or over-interpretation of
findings. As more scientists are becoming experts in DNAm, hopefully this will translate into more critical reviewers. In the meantime, researchers are using pre-print servers, such as bioRxiv, and open source code, to gain critical feedback from the epigenetic community prior to publication.
References


35. Miller GE, Chen E, Chen E, Fok AK, Fok AK, et al. Low early-life social class leaves a biological residue manifested by decreased glucocorticoid and increased proinflammatory signaling. Proc Natl Acad Sci USA [Internet]. 2009 Aug 25;106(34):14716–21.


51. Hawkley LC, Masi CM, Berry JD, Cacioppo JT. Loneliness is a unique predictor of age-related differences in systolic blood pressure. Psychology and Aging. 2006;21(1):152–64.


152


<table>
<thead>
<tr>
<th>No.</th>
<th>Author(s)</th>
<th>Title</th>
<th>Journal</th>
<th>Year</th>
</tr>
</thead>
<tbody>
<tr>
<td>185</td>
<td>Henikoff S, Greally JM</td>
<td>Epigenetics, cellular memory and gene regulation</td>
<td>Curr Biol</td>
<td>2016;26</td>
</tr>
<tr>
<td>186</td>
<td>Bird A</td>
<td>Perceptions of epigenetics</td>
<td>Nature</td>
<td>2007;447</td>
</tr>
<tr>
<td>187</td>
<td>Jones PA</td>
<td>Functions of DNA methylation: islands, start sites, gene bodies and beyond</td>
<td>Nat Rev Genet</td>
<td>2012;13</td>
</tr>
<tr>
<td>189</td>
<td>Jirtle RL, Skinner MK</td>
<td>Environmental epigenomics and disease susceptibility</td>
<td>Nat Rev Genet</td>
<td>2007;8</td>
</tr>
<tr>
<td>190</td>
<td>Meaney MJ</td>
<td>Epigenetics and the biological definition of gene x environment interactions</td>
<td>Child Development</td>
<td>2010 Jan;81(1):41–79</td>
</tr>
<tr>
<td>191</td>
<td>Yuen RKC, Jiang R, Peñaherrera MS, McFadden DE, Robinson WP</td>
<td>Genome-wide mapping of imprinted differentially methylated regions by DNA methylation profiling of human placentas from triploidies</td>
<td>Epigenet Chromatin</td>
<td>2011;4</td>
</tr>
<tr>
<td>192</td>
<td>Laird PW</td>
<td>Principles and challenges of genomewide DNA methylation analysis</td>
<td>Nat Rev Genet</td>
<td>2010;11</td>
</tr>
<tr>
<td>193</td>
<td>Michels KB, Binder AM, Dedeurwaerder S, Epstein CB, Greally JM, Gut I</td>
<td>Recommendations for the design and analysis of epigenome-wide association studies</td>
<td>Nat Methods</td>
<td>2013;10</td>
</tr>
<tr>
<td>195</td>
<td>Hatchwell E, Greally JM</td>
<td>The potential role of epigenomic dysregulation in complex human disease</td>
<td>Trends Genet</td>
<td>2007;23</td>
</tr>
<tr>
<td>196</td>
<td>Alisch RS, Barwick BG, Chopra P, Myrick LK, Satten GA, Conneely KN</td>
<td>Age-associated DNA methylation in pediatric populations</td>
<td>Genome Res</td>
<td>2012;22</td>
</tr>
<tr>
<td>197</td>
<td>Jaffe AE, Irizarry RA</td>
<td>Accounting for cellular heterogeneity is critical in epigenome-wide association studies</td>
<td>Genome Biol [Internet]</td>
<td>2014;15(2):R31</td>
</tr>
</tbody>
</table>


218. Miller GE, Chen E, Fok AK, Walker H, Lim A, Nicholls EF. Low early-life social class leaves a biological residue manifested by decreased glucocorticoid and increased proinflammatory signaling. Proc Natl Acad Sci USA. 2009;106.


323. Tobi EW, Slocker RC, Luijk R, Dekkers KF, Stein AD, Xu KM, et al. DNA methylation as a mediator of the association between prenatal adversity and risk factors for metabolic


Appendices

Appendix A Supplementary materials for chapter 3

A.1 Supplementary figures

Supplemental Figure 3.1 Cell type correction.
A) Predicted proportions of cell types in PBMCs for both datasets (Mono = monocytes; Gran = Granulocytes). B) Predicted proportions of cell types in BECs for both datasets.
Supplementary Figure 3.2 Identifying correlation coefficient thresholds for informative CpGs.
Beta mixture modeling on Spearman correlation rho values between matched BECs and PBMCs for A) GECKO and B) C3ARE cohorts. The bimodal distribution of Spearman rho values indicated two underlying populations of CpGs, a set of uncorrelated CpGs (shown in red) and a set of right-skewed highly positively correlated CpGs (shown in green). Correlation coefficient threshold for informative CpGs were determined at 2 standard deviations minus the mean of the green Gaussian distribution (GECKO rho = 0.47; C3ARE rho = 0.32).
Supplementary Figure 3.3 Similarities in population structure of GECKO and C3ARE. Principal component analysis of PsychChip genotyping profiles (542,699 SNPs) for C3ARE (shown in blue) and GECKO (shown in red) revealed that genetic ancestry did not differ significantly between the cohorts as determined by Wilcoxon ranked sum test of GECKO versus C3ARE in PC1 scores ($p = 0.8$) and PC2 scores ($p = 0.4$).

Supplementary Figure 3.4 Distribution of tissue correlations. Density distribution of Spearman’s correlation coefficient (Rho) across 419, 507 CpGs in matched BEC and PBMC tissues for GECKO, GECKOsub, GECKOsub Averaged (mean of 100 trials of GECKOsub) and C3ARE datasets.
Supplementary Figure 3.5 *cis*-mQTL identification. Overlap of *cis*-mQTLs identified in matched tissues of both C3ARE and GECKO cohorts, respectively.
Supplementary Figure 6. *cis*-mQTL enrichment in genomic features and informative CpGs. A) Representation of 4,980 CpGs underlying validated *cis*-mQTLs across various genomic features. Bars show the fold-change between CpG count in each genomic region and the mean count of randomly selected CpGs in that same genomic feature, from 10,000 iterations. Error bars show standard error. (* denotes significant enrichment or depletion at FDR ≤ 0.05) (S = South; N = North). B) A) Stacked bar plot representing overlap of identified informative sites associated with BEC-specific, PBMC-specific and shared-tissue validated *cis*-mQTLs.
## A.2 Supplementary tables

### Supplementary Table 3.1

<table>
<thead>
<tr>
<th></th>
<th>C3ARE</th>
<th>PBMC Reference Range</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0</td>
<td>≥ 0.05</td>
</tr>
<tr>
<td><strong>Positive</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Correlation</strong></td>
<td><strong>Rho</strong></td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>419,507</td>
<td>64,204</td>
</tr>
<tr>
<td>≥ 0.3</td>
<td>74,137</td>
<td>24,542</td>
</tr>
<tr>
<td>≥ 0.6</td>
<td>10,476</td>
<td>6,489</td>
</tr>
<tr>
<td>≥ 0.9</td>
<td>45</td>
<td>45</td>
</tr>
<tr>
<td><strong>Negative</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Correlation</strong></td>
<td><strong>Rho</strong></td>
<td></td>
</tr>
<tr>
<td>≤ -0.3</td>
<td>51,436</td>
<td>4,866</td>
</tr>
<tr>
<td>≤ -0.6</td>
<td>3,300</td>
<td>327</td>
</tr>
<tr>
<td>≤ -0.9</td>
<td>3</td>
<td>0</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>GECKO</th>
<th>PBMC Reference Range</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0</td>
<td>≥ 0.05</td>
</tr>
<tr>
<td><strong>Positive</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Correlation</strong></td>
<td><strong>Rho</strong></td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>419,507</td>
<td>131,227</td>
</tr>
<tr>
<td>≥ 0.3</td>
<td>333,22</td>
<td>29,158</td>
</tr>
<tr>
<td>≥ 0.6</td>
<td>6,285</td>
<td>6,174</td>
</tr>
<tr>
<td>≥ 0.9</td>
<td>41</td>
<td>41</td>
</tr>
<tr>
<td><strong>Negative</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Correlation</strong></td>
<td><strong>Rho</strong></td>
<td></td>
</tr>
<tr>
<td>≤ -0.3</td>
<td>1,557</td>
<td>331</td>
</tr>
<tr>
<td>≤ -0.6</td>
<td>8</td>
<td>8</td>
</tr>
<tr>
<td>≤ -0.9</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>GECKOsub</th>
<th>PBMC Reference Range</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0</td>
<td>≥ 0.05</td>
</tr>
<tr>
<td><strong>Positive</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Correlation</strong></td>
<td><strong>Rho</strong></td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>419,507</td>
<td>115,404</td>
</tr>
<tr>
<td>≥ 0.3</td>
<td>30,615</td>
<td>26,385</td>
</tr>
<tr>
<td>≥ 0.6</td>
<td>5,252</td>
<td>5,172</td>
</tr>
<tr>
<td>≥ 0.9</td>
<td>11</td>
<td>11</td>
</tr>
<tr>
<td>-------</td>
<td>-----</td>
<td>-----</td>
</tr>
<tr>
<td><strong>Negative Correlation Rho</strong></td>
<td>≤ -0.3</td>
<td>1,357</td>
</tr>
<tr>
<td></td>
<td>≤ -0.6</td>
<td>5</td>
</tr>
<tr>
<td></td>
<td>≤ -0.9</td>
<td>0</td>
</tr>
</tbody>
</table>
## Supplementary Table 3.2

<table>
<thead>
<tr>
<th>Type of Site</th>
<th># of Sites in Almstrup et al. 2017 (%)</th>
<th># of Sites in Berko et al. 2014 (%)</th>
<th># of Sites in Fisher et al. 2015 (%)</th>
<th># of Sites in Portales et al. 2016 (%)</th>
<th># of Sites in Xu et al. 2017 (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>All categories</td>
<td>0 (0%)</td>
<td>0 (0%)</td>
<td>1 (0.4%)</td>
<td>3 (0.5%)</td>
<td>33 (0.3%)</td>
</tr>
<tr>
<td>Differential</td>
<td>67 (71.3%)</td>
<td>27 (36.5%)</td>
<td>42 (16.7%)</td>
<td>347 (52.7%)</td>
<td>6629 (68.3%)</td>
</tr>
<tr>
<td>Informative</td>
<td>1 (1.1%)</td>
<td>18 (24.3%)</td>
<td>2 (0.8%)</td>
<td>9 (1.4%)</td>
<td>63 (0.6%)</td>
</tr>
<tr>
<td>Informative &amp; Differential</td>
<td>4 (0.04%)</td>
<td>6 (8.1%)</td>
<td>2 (0.8%)</td>
<td>21 (3.2%)</td>
<td>166 (1.7%)</td>
</tr>
<tr>
<td>mQTL CpG</td>
<td>2 (0.02%)</td>
<td>0 (0%)</td>
<td>0 (0%)</td>
<td>6 (0.9%)</td>
<td>45 (0.5%)</td>
</tr>
<tr>
<td>mQTL &amp; Differential</td>
<td>0 (0%)</td>
<td>10 (13.5%)</td>
<td>2 (0.8%)</td>
<td>18 (2.7%)</td>
<td>168 (1.7%)</td>
</tr>
<tr>
<td>mQTL &amp; Informative</td>
<td>0 (0%)</td>
<td>0 (0%)</td>
<td>1 (0.4%)</td>
<td>2 (0.3%)</td>
<td>9 (0.09%)</td>
</tr>
<tr>
<td>None</td>
<td>20 (21.3%)</td>
<td>13 (17.6%)</td>
<td>202 (80.2%)</td>
<td>252 (0.4%)</td>
<td>2591 (26.7%)</td>
</tr>
<tr>
<td>Total Reported</td>
<td>94 (100%)</td>
<td>74 (100%)</td>
<td>252 (100%)</td>
<td>658 (100%)</td>
<td>9704 (100%)</td>
</tr>
</tbody>
</table>
Appendix B Supplementary materials for chapter 4

B.1 Supplementary figures

Supplementary Figure 4.1 Associations between temperament, mental health and ANS measures supported preexisting relationships between these traits. Correlation matrix of pairwise spearman correlations run on 16 original traits ($n=55$).
Supplementary Figure 4.2 Comparison of FDR corrected \( p \)-values and effect sizes (\( |\Delta \beta| \)) from Spearman correlations and linear regressions. In left panel, Spearman \( q \)-values vs. \( q \)-values produced by linear regression covaried by sex and minority status. In right panel, \( |\Delta \beta| \) from Spearman correlation analysis vs. linear regression. Of the 93 high and medium confidence CpGs, 40 were retained at a \( q \)-value < 0.2; this included 11 of the 12 original high confidence CpGs and 29 of 81 medium confidence. Horizontal lines represent thresholds used for medium confidence sites (\( q \)-value < 0.2, absolute \( |\Delta \beta| \) > 5%). Sites are colored by their significance in Spearman analysis (red = high confidence, blue = medium confidence, grey = not significant) (n= 55).
Supplementary Figure 4.3 Associations between Inhibition/Disinhibition scores and DLX5 differentially methylated CpGs passed permutation tests at age 15. Inhibition/Disinhibition scores were randomly assigned 100 times and correlated with DNA methylation of the respective differentially methylated CpGs at age 15. True coefficients fell outside of the 99 percentile of the null distributions suggesting that these associations were not spurious. Colors represent to which gene the CpG is mapped: purple = DLX5, red = IGF2, blue = MYO16, green = PRUNE2 (n = 55).
Supplementary Figure 4.4 Correlations between DNA methylation at age 15 and Biobehavioural Inhibition/Disinhibition in males, females, subsampled females. Difference in magnitude in correlation coefficient between males and females was not greatly altered by sample size in high and medium confidence CpGs. Full cohort (gray circle, n=55), females only (pink square, n=36) and males only (blue triangle, n=19). Average of subsampled females to n=19 and running correlations 100 times in black squares.
Supplementary Figure 4.5 Associations between Inhibition/Disinhibition scores and DLX5 differentially methylated CpGs passed permutation tests at age 18. Inhibition/Disinhibition scores were randomly assigned 100 times and correlated with DNA methylation of the respective differentially methylated CpGs at age 18. True coefficients fell outside of the 97 percentile of the null distributions suggesting that these associations were not spurious. Colors represent to which gene the CpG is mapped: purple = DLX5, red = IGF2 (n= 52).
Supplementary Figure 4.6 Associations between measures underlying Inhibition/Disinhibition and differentially methylated CpGs in DLX5 and IGF2. Observational anger measured at grade 1 was the most strongly correlated with DNA methylation at both genes.
## B.2 Supplementary tables

### Supplementary Table 4.1

<table>
<thead>
<tr>
<th>site</th>
<th>r2</th>
<th>adjusted r2</th>
<th>p-value</th>
<th>FDR p-value</th>
<th>gene name</th>
<th>Distance to TSS (bp)</th>
<th>significant at age 18</th>
</tr>
</thead>
<tbody>
<tr>
<td>cg24115040</td>
<td>0.26705472</td>
<td>0.23713859</td>
<td>0.00020298</td>
<td>0.00304475</td>
<td>DLX5</td>
<td>1768</td>
<td>no</td>
</tr>
<tr>
<td>cg11500797</td>
<td>0.24492205</td>
<td>0.21410254</td>
<td>0.00211411</td>
<td>0.01585586</td>
<td>DLX5</td>
<td>2020</td>
<td>nominal</td>
</tr>
<tr>
<td>cg18946226</td>
<td>0.23227322</td>
<td>0.20093743</td>
<td>0.00333816</td>
<td>0.01669078</td>
<td>MYO16</td>
<td>4</td>
<td>no</td>
</tr>
<tr>
<td>cg00503840</td>
<td>0.25395101</td>
<td>0.22350003</td>
<td>0.01088797</td>
<td>0.02557812</td>
<td>DLX5</td>
<td>3634</td>
<td>FDR</td>
</tr>
<tr>
<td>cg11005826</td>
<td>0.17095002</td>
<td>0.13711125</td>
<td>0.01193646</td>
<td>0.02557812</td>
<td>IGF2</td>
<td>-3620</td>
<td>no</td>
</tr>
<tr>
<td>cg13462129</td>
<td>0.2096587</td>
<td>0.17739987</td>
<td>0.00824102</td>
<td>0.02557812</td>
<td>DLX5</td>
<td>2862</td>
<td>nominal</td>
</tr>
<tr>
<td>cg27016494</td>
<td>0.20067186</td>
<td>0.16804622</td>
<td>0.00986032</td>
<td>0.02557812</td>
<td>DLX5</td>
<td>2557</td>
<td>nominal</td>
</tr>
<tr>
<td>cg19282250</td>
<td>0.18840758</td>
<td>0.15528136</td>
<td>0.01402684</td>
<td>0.02630033</td>
<td>PRUNE2</td>
<td>-9507</td>
<td>nominal</td>
</tr>
<tr>
<td>cg20080624</td>
<td>0.20769858</td>
<td>0.17535975</td>
<td>0.01860574</td>
<td>0.03100956</td>
<td>DLX5</td>
<td>3032</td>
<td>nominal</td>
</tr>
<tr>
<td>cg12041387</td>
<td>0.26227119</td>
<td>0.23215981</td>
<td>0.02260787</td>
<td>0.03391181</td>
<td>DLX5</td>
<td>3972</td>
<td>FDR</td>
</tr>
<tr>
<td>cg14396117</td>
<td>0.15362305</td>
<td>0.11907705</td>
<td>0.02515218</td>
<td>0.03429843</td>
<td>MYO16</td>
<td>-452</td>
<td>no</td>
</tr>
<tr>
<td>cg08878323</td>
<td>0.12021293</td>
<td>0.08430326</td>
<td>0.02750238</td>
<td>0.03437798</td>
<td>DLX5</td>
<td>1679</td>
<td>no</td>
</tr>
<tr>
<td>cg18873386</td>
<td>0.10884631</td>
<td>0.07247269</td>
<td>0.03491743</td>
<td>0.04028935</td>
<td>DLX5</td>
<td>2228</td>
<td>no</td>
</tr>
<tr>
<td>cg11880010</td>
<td>0.1058737</td>
<td>0.06937875</td>
<td>0.046848</td>
<td>0.05019428</td>
<td>PRUNE2</td>
<td>-9629</td>
<td>no</td>
</tr>
<tr>
<td>cg21237591</td>
<td>0.10259212</td>
<td>0.06596322</td>
<td>0.13349063</td>
<td>0.13349063</td>
<td>IGF2</td>
<td>-169</td>
<td>no</td>
</tr>
</tbody>
</table>
Appendix C Supplementary materials for chapter 5

C.1 Supplementary figures

Supplementary Figure 5.1 Self-reported ethnicity compared to first and second principal component, PC1 and PC2, of genotyping data. PC1 and PC2 accounted for 62.9% and 2.2%, respectively, of variation in the genotyping typing data (n=304). Individuals how were reported as Canadian (black points) did not cluster with a single group.
C.2 Supplementary tables

Supplementary Table 5.1

<p>| Association | CpG          | p-value     | Delta beta (|Δβ|)     |
|-------------|--------------|-------------|------------|
| PosCan      | cg14156792   | 0.00010115  | 0.21364137 |
| PosCan      | cg04121631   | 0.00047624  | 0.09693551 |
| PosCan      | cg13798679   | 0.00017467  | -0.0567288 |
| PosCan      | cg19412669   | 0.00032285  | -0.0694238 |
| PosCan      | cg24375409   | 0.00040432  | -0.1561574 |
| PosCan      | cg23800435   | 0.00042675  | -0.0700827 |
| PosCan      | cg07813265   | 0.00046666  | -0.0653389 |
| PosCan      | cg10759651   | 0.00044754  | -0.0650657 |
| PosCan      | cg15642380   | 0.00042281  | -0.0523463 |
| PosCan      | cg14264194   | 0.00024619  | -0.0891523 |
| PosCan      | cg16962115   | 0.00014267  | 0.06193473 |
| PosCan      | cg07700062   | 0.00011868  | -0.0696864 |
| PosCan      | cg21032567   | 0.00026929  | -0.1224142 |
| PosCan      | cg04944784   | 0.00043091  | -0.0641391 |
| PosCan      | cg04036593   | 0.0002826   | -0.0726475 |
| PosCan      | cg13099839   | 5.23E-05    | -0.0515544 |
| PosCan      | cg02643782   | 0.00035128  | -0.0501024 |
| PosCan      | cg02750792   | 0.00031068  | -0.1353819 |
| PosCan      | cg17641710   | 0.00013531  | -0.0620558 |
| PosCan      | cg12615535   | 0.00032064  | -0.0582703 |
| PosCan      | cg26095395   | 0.00023092  | -0.0861033 |
| PosCan      | cg19934294   | 0.00038678  | -0.0527305 |
| PosCan      | cg20080590   | 0.00017811  | 0.0604394  |
| PosCan      | cg22725685   | 4.13E-05    | -0.0571288 |
| PosCan      | cg09526685   | 0.00017105  | -0.2373299 |
| PosCan      | cg07033961   | 0.0004648   | 0.05767375 |
| PosCan      | cg07593523   | 0.00017744  | 0.12077722 |
| PosCan      | cg22179564   | 0.00027546  | -0.115591  |
| PosCan      | cg27034935   | 0.00033082  | -0.1268946 |
| PosCan      | cg04467589   | 0.00043884  | 0.07443024 |
| PosCan      | cg18808777   | 0.00024927  | 0.08343909 |
| PosCan      | cg12231373   | 0.00015062  | -0.0712878 |
| PosCan      | cg23317857   | 0.000299    | -0.0667741 |
| PosCan      | cg17117243   | 8.03E-05    | -0.0664841 |</p>
<table>
<thead>
<tr>
<th>PosCan</th>
<th>cg20706496</th>
<th>0.00048304</th>
<th>0.0776308</th>
</tr>
</thead>
<tbody>
<tr>
<td>PosCan</td>
<td>cg26247093</td>
<td>0.00044294</td>
<td>0.05752354</td>
</tr>
<tr>
<td>PosCan</td>
<td>cg23129930</td>
<td>0.00014753</td>
<td>-0.0841488</td>
</tr>
<tr>
<td>PosCan</td>
<td>cg05928186</td>
<td>0.00023347</td>
<td>-0.0747499</td>
</tr>
<tr>
<td>PosCan</td>
<td>cg11850943</td>
<td>0.00023262</td>
<td>0.10052228</td>
</tr>
<tr>
<td>PosCan</td>
<td>cg13477780</td>
<td>0.00037471</td>
<td>-0.061358</td>
</tr>
<tr>
<td>PosCan</td>
<td>cg05928186</td>
<td>0.00033659</td>
<td>-0.0670469</td>
</tr>
<tr>
<td>PosCan</td>
<td>cg01545109</td>
<td>0.00037998</td>
<td>0.06239001</td>
</tr>
<tr>
<td>PosCan</td>
<td>cg12978800</td>
<td>0.00044255</td>
<td>-0.0821842</td>
</tr>
<tr>
<td>PosCan</td>
<td>cg07910560</td>
<td>0.00013986</td>
<td>-0.0749372</td>
</tr>
<tr>
<td>PosCan</td>
<td>cg14316231</td>
<td>0.00024185</td>
<td>-0.0589439</td>
</tr>
<tr>
<td>PosCan</td>
<td>cg17450425</td>
<td>0.00024747</td>
<td>0.15829943</td>
</tr>
<tr>
<td>PosCan</td>
<td>cg04332163</td>
<td>0.00046729</td>
<td>-0.0841953</td>
</tr>
<tr>
<td>PosCan</td>
<td>cg24756403</td>
<td>0.00013329</td>
<td>-0.0647096</td>
</tr>
<tr>
<td>PosCan</td>
<td>cg19054360</td>
<td>0.00017341</td>
<td>-0.0725608</td>
</tr>
<tr>
<td>PosCan</td>
<td>cg07692929</td>
<td>0.00010954</td>
<td>-0.0520654</td>
</tr>
<tr>
<td>PosCan</td>
<td>cg15920975</td>
<td>0.00037333</td>
<td>-0.0752409</td>
</tr>
<tr>
<td>PosCan</td>
<td>cg16118839</td>
<td>0.00018664</td>
<td>-0.0595886</td>
</tr>
<tr>
<td>PosCan</td>
<td>cg18031880</td>
<td>0.00018902</td>
<td>-0.0983282</td>
</tr>
<tr>
<td>PosCan</td>
<td>cg19704853</td>
<td>0.00035763</td>
<td>-0.0506376</td>
</tr>
<tr>
<td>PosCan</td>
<td>cg01158527</td>
<td>0.00043732</td>
<td>-0.0786045</td>
</tr>
<tr>
<td>PosCan</td>
<td>cg15282632</td>
<td>0.00013413</td>
<td>-0.098587</td>
</tr>
<tr>
<td>PosCan</td>
<td>cg20320656</td>
<td>0.00042416</td>
<td>-0.1181435</td>
</tr>
<tr>
<td>PosCan</td>
<td>cg01290565</td>
<td>5.46E-05</td>
<td>-0.0867748</td>
</tr>
<tr>
<td>PosCan</td>
<td>cg02556042</td>
<td>0.00037104</td>
<td>-0.1055742</td>
</tr>
<tr>
<td>PosCan</td>
<td>cg04217706</td>
<td>0.00017364</td>
<td>-0.0730008</td>
</tr>
<tr>
<td>PosCan</td>
<td>cg03885975</td>
<td>0.00041433</td>
<td>-0.0935023</td>
</tr>
<tr>
<td>PosCan</td>
<td>cg22124136</td>
<td>0.00041323</td>
<td>-0.069707</td>
</tr>
<tr>
<td>PosCan</td>
<td>cg09842196</td>
<td>0.00035373</td>
<td>-0.0626533</td>
</tr>
<tr>
<td>PosCan</td>
<td>cg26027442</td>
<td>0.00025074</td>
<td>-0.0502048</td>
</tr>
<tr>
<td>PosCan</td>
<td>cg13427748</td>
<td>0.00012653</td>
<td>0.05918896</td>
</tr>
<tr>
<td>PosCan</td>
<td>cg00611495</td>
<td>0.00020907</td>
<td>-0.0671208</td>
</tr>
<tr>
<td>PosCan</td>
<td>cg04022912</td>
<td>4.35E-05</td>
<td>-0.0601433</td>
</tr>
<tr>
<td>PosCan</td>
<td>cg04337176</td>
<td>7.50E-05</td>
<td>0.06980444</td>
</tr>
<tr>
<td>PosCan</td>
<td>cg08634357</td>
<td>0.00034218</td>
<td>-0.0804004</td>
</tr>
<tr>
<td>PosCan</td>
<td>cg13710086</td>
<td>7.30E-05</td>
<td>-0.0909291</td>
</tr>
<tr>
<td>PosCan</td>
<td>cg01443020</td>
<td>0.00049335</td>
<td>-0.1321152</td>
</tr>
<tr>
<td>PosCan</td>
<td>cg</td>
<td>Value</td>
<td>Log2FoldChange</td>
</tr>
<tr>
<td>------------</td>
<td>--------</td>
<td>---------</td>
<td>----------------</td>
</tr>
<tr>
<td>PosCan</td>
<td>cg20520147</td>
<td>0.00024052</td>
<td>0.05538793</td>
</tr>
<tr>
<td>PosCan</td>
<td>cg26933683</td>
<td>0.00033785</td>
<td>-0.0614819</td>
</tr>
<tr>
<td>PosCan</td>
<td>cg13093111</td>
<td>9.23E-05</td>
<td>-0.0618555</td>
</tr>
<tr>
<td>PosCan</td>
<td>cg25867973</td>
<td>0.00031598</td>
<td>-0.0563275</td>
</tr>
<tr>
<td>PosCan</td>
<td>cg16387467</td>
<td>0.00035251</td>
<td>-0.0603054</td>
</tr>
<tr>
<td>PosCan</td>
<td>cg06650914</td>
<td>0.00011353</td>
<td>-0.0604465</td>
</tr>
<tr>
<td>PosCan</td>
<td>cg14453935</td>
<td>0.00042824</td>
<td>-0.1140602</td>
</tr>
<tr>
<td>PosCan</td>
<td>cg06744585</td>
<td>0.00044713</td>
<td>-0.0897358</td>
</tr>
<tr>
<td>PosCan</td>
<td>cg16521040</td>
<td>0.00012418</td>
<td>-0.0610724</td>
</tr>
<tr>
<td>PosCan</td>
<td>cg21448033</td>
<td>0.00011843</td>
<td>-0.0519864</td>
</tr>
<tr>
<td>PosCan</td>
<td>cg19498228</td>
<td>5.78E-05</td>
<td>-0.0727495</td>
</tr>
<tr>
<td>PosCan</td>
<td>cg09265161</td>
<td>6.00E-05</td>
<td>0.08819041</td>
</tr>
<tr>
<td>PosCan</td>
<td>cg00587465</td>
<td>0.00033157</td>
<td>-0.0794221</td>
</tr>
<tr>
<td>PosCan</td>
<td>cg24083746</td>
<td>0.00014622</td>
<td>-0.067266</td>
</tr>
<tr>
<td>PosCan</td>
<td>cg17611686</td>
<td>0.00045125</td>
<td>-0.0797195</td>
</tr>
<tr>
<td>PosCan</td>
<td>cg00481227</td>
<td>0.00041518</td>
<td>-0.0758613</td>
</tr>
<tr>
<td>PosCan</td>
<td>cg01906944</td>
<td>0.00025278</td>
<td>-0.0670976</td>
</tr>
<tr>
<td>PosCan</td>
<td>cg03199996</td>
<td>0.0002659</td>
<td>-0.0634823</td>
</tr>
<tr>
<td>PosCan</td>
<td>cg16443667</td>
<td>0.00025983</td>
<td>-0.096753</td>
</tr>
<tr>
<td>PosCan</td>
<td>cg20652954</td>
<td>3.29E-05</td>
<td>-0.0783382</td>
</tr>
<tr>
<td>PosCan</td>
<td>cg02966332</td>
<td>0.00024204</td>
<td>-0.0645007</td>
</tr>
<tr>
<td>PosCan</td>
<td>cg01622416</td>
<td>0.00021334</td>
<td>0.06907193</td>
</tr>
<tr>
<td>PosCan</td>
<td>cg17449254</td>
<td>0.00027277</td>
<td>-0.0880016</td>
</tr>
<tr>
<td>PosCan</td>
<td>cg05298458</td>
<td>0.00045953</td>
<td>-0.0771892</td>
</tr>
<tr>
<td>PosCan</td>
<td>cg10941566</td>
<td>0.00022258</td>
<td>-0.0706103</td>
</tr>
<tr>
<td>PosCan</td>
<td>cg14335069</td>
<td>1.08E-05</td>
<td>-0.0647062</td>
</tr>
<tr>
<td>CompSES</td>
<td>cg09895325</td>
<td>0.00014561</td>
<td>0.0893803</td>
</tr>
<tr>
<td>CompSES</td>
<td>cg18936040</td>
<td>0.00018443</td>
<td>-0.0784756</td>
</tr>
<tr>
<td>CompSES</td>
<td>cg06936779</td>
<td>0.00042399</td>
<td>-0.0629242</td>
</tr>
<tr>
<td>CompSES</td>
<td>cg09671955</td>
<td>6.86E-05</td>
<td>0.07365604</td>
</tr>
<tr>
<td>CompSES</td>
<td>cg20707323</td>
<td>0.00038393</td>
<td>-0.0520781</td>
</tr>
<tr>
<td>CompSES</td>
<td>cg23933289</td>
<td>2.76E-05</td>
<td>0.05123025</td>
</tr>
<tr>
<td>CompSES</td>
<td>cg22622057</td>
<td>0.00040571</td>
<td>-0.0689931</td>
</tr>
<tr>
<td>CompSES</td>
<td>cg07919760</td>
<td>0.0004145</td>
<td>0.05684397</td>
</tr>
<tr>
<td>CompSES</td>
<td>cg04231677</td>
<td>0.00023307</td>
<td>-0.0668022</td>
</tr>
<tr>
<td>CompSES</td>
<td>cg00493617</td>
<td>0.00021664</td>
<td>-0.0559326</td>
</tr>
<tr>
<td>CompSES</td>
<td>cg25356886</td>
<td>0.00015271</td>
<td>-0.0710399</td>
</tr>
<tr>
<td>CompSES</td>
<td>cg25429719</td>
<td>0.00037047</td>
<td>-0.1274762</td>
</tr>
<tr>
<td>CompSES</td>
<td>cg</td>
<td>q value</td>
<td>beta</td>
</tr>
<tr>
<td>---------</td>
<td>----</td>
<td>-----------</td>
<td>----------</td>
</tr>
<tr>
<td>CompSES</td>
<td>c08217411</td>
<td>0.00030288</td>
<td>-0.0718378</td>
</tr>
<tr>
<td>CompSES</td>
<td>cg04216051</td>
<td>0.00028257</td>
<td>0.0687372</td>
</tr>
<tr>
<td>CompSES</td>
<td>cg23855365</td>
<td>0.00019923</td>
<td>-0.0669263</td>
</tr>
<tr>
<td>CompSES</td>
<td>cg11519983</td>
<td>0.00020843</td>
<td>0.05959503</td>
</tr>
<tr>
<td>CompSES</td>
<td>cg07033961</td>
<td>0.00020843</td>
<td>0.05959503</td>
</tr>
<tr>
<td>CompSES</td>
<td>cg23855365</td>
<td>0.00019923</td>
<td>-0.0669263</td>
</tr>
<tr>
<td>CompSES</td>
<td>cg11519983</td>
<td>0.00020843</td>
<td>0.05959503</td>
</tr>
<tr>
<td>CompSES</td>
<td>cg14677130</td>
<td>0.00014382</td>
<td>-0.0740824</td>
</tr>
<tr>
<td>CompSES</td>
<td>cg25843003</td>
<td>0.00017644</td>
<td>0.05902741</td>
</tr>
<tr>
<td>CompSES</td>
<td>cg18808777</td>
<td>2.03E-05</td>
<td>0.08900093</td>
</tr>
<tr>
<td>CompSES</td>
<td>cg04452713</td>
<td>0.00037745</td>
<td>-0.054961</td>
</tr>
<tr>
<td>CompSES</td>
<td>cg26325286</td>
<td>0.0004448</td>
<td>-0.1087012</td>
</tr>
<tr>
<td>CompSES</td>
<td>cg15970595</td>
<td>2.94E-05</td>
<td>-0.0706644</td>
</tr>
<tr>
<td>CompSES</td>
<td>cg24617008</td>
<td>0.0002581</td>
<td>0.05015504</td>
</tr>
<tr>
<td>CompSES</td>
<td>cg14087413</td>
<td>0.00046576</td>
<td>-0.0561771</td>
</tr>
<tr>
<td>CompSES</td>
<td>cg03330490</td>
<td>0.00027947</td>
<td>0.05268399</td>
</tr>
<tr>
<td>CompSES</td>
<td>cg1664405</td>
<td>1.21E-05</td>
<td>-0.0550979</td>
</tr>
<tr>
<td>CompSES</td>
<td>cg19572051</td>
<td>0.0004547</td>
<td>-0.2256547</td>
</tr>
<tr>
<td>CompSES</td>
<td>cg02425416</td>
<td>0.00022209</td>
<td>-0.0772744</td>
</tr>
<tr>
<td>CompSES</td>
<td>cg03306615</td>
<td>0.00014973</td>
<td>-0.1374105</td>
</tr>
<tr>
<td>CompSES</td>
<td>cg26051413</td>
<td>0.00022444</td>
<td>-0.1700336</td>
</tr>
<tr>
<td>CompSES</td>
<td>cg13762320</td>
<td>0.00042667</td>
<td>-0.0904743</td>
</tr>
<tr>
<td>CompSES</td>
<td>cg11644479</td>
<td>0.00024021</td>
<td>-0.1147157</td>
</tr>
<tr>
<td>CompSES</td>
<td>cg24335353</td>
<td>9.76E-05</td>
<td>-0.1179523</td>
</tr>
<tr>
<td>CompSES</td>
<td>cg09394785</td>
<td>0.00021661</td>
<td>-0.0753936</td>
</tr>
<tr>
<td>CompSES</td>
<td>cg08577913</td>
<td>0.00033554</td>
<td>-0.1244275</td>
</tr>
<tr>
<td>CompSES</td>
<td>cg23745839</td>
<td>0.00039035</td>
<td>-0.0598093</td>
</tr>
<tr>
<td>CompSES</td>
<td>cg22040889</td>
<td>7.97E-05</td>
<td>-0.1038312</td>
</tr>
<tr>
<td>CompSES</td>
<td>cg15392457</td>
<td>0.00019981</td>
<td>-0.0555871</td>
</tr>
<tr>
<td>CompSES</td>
<td>cg07451370</td>
<td>0.00032614</td>
<td>-0.0674942</td>
</tr>
<tr>
<td>CompSES</td>
<td>cg09274587</td>
<td>0.00025504</td>
<td>-0.0608646</td>
</tr>
<tr>
<td>CompSES</td>
<td>cg18071532</td>
<td>3.20E-05</td>
<td>-0.0906912</td>
</tr>
<tr>
<td>CompSES</td>
<td>cg20320656</td>
<td>0.00021098</td>
<td>-0.1248849</td>
</tr>
<tr>
<td>CompSES</td>
<td>cg02556042</td>
<td>0.0002555</td>
<td>-0.1014616</td>
</tr>
<tr>
<td>CompSES</td>
<td>cg21892295</td>
<td>0.00018245</td>
<td>0.13881512</td>
</tr>
<tr>
<td>CompSES</td>
<td>cg09460553</td>
<td>0.0004271</td>
<td>-0.1398706</td>
</tr>
<tr>
<td>CompSES</td>
<td>cg02315096</td>
<td>0.00012557</td>
<td>-0.158518</td>
</tr>
<tr>
<td>CompSES</td>
<td>cg18865445</td>
<td>0.00028788</td>
<td>-0.1540926</td>
</tr>
<tr>
<td>CompSES</td>
<td>cg17797229</td>
<td>0.0003987</td>
<td>-0.1929233</td>
</tr>
<tr>
<td>---------</td>
<td>------------</td>
<td>------------</td>
<td>------------</td>
</tr>
<tr>
<td>CompSES</td>
<td>cg16761754</td>
<td>1.42E-05</td>
<td>0.21773072</td>
</tr>
<tr>
<td>CompSES</td>
<td>cg13563725</td>
<td>0.00039187</td>
<td>-0.0510089</td>
</tr>
<tr>
<td>CompSES</td>
<td>cg22713356</td>
<td>0.00020384</td>
<td>-0.1161049</td>
</tr>
<tr>
<td>CompSES</td>
<td>cg26826117</td>
<td>0.00028432</td>
<td>0.0874998</td>
</tr>
<tr>
<td>CompSES</td>
<td>cg07933656</td>
<td>6.58E-05</td>
<td>-0.1860448</td>
</tr>
<tr>
<td>CompSES</td>
<td>cg06650914</td>
<td>0.00022659</td>
<td>-0.0529829</td>
</tr>
<tr>
<td>CompSES</td>
<td>cg22761077</td>
<td>0.00022981</td>
<td>-0.0510489</td>
</tr>
<tr>
<td>CompSES</td>
<td>cg07425555</td>
<td>0.00026551</td>
<td>-0.0768959</td>
</tr>
<tr>
<td>CompSES</td>
<td>cg08440266</td>
<td>0.00011204</td>
<td>-0.1103966</td>
</tr>
<tr>
<td>CompSES</td>
<td>cg01817521</td>
<td>0.00032099</td>
<td>0.05556459</td>
</tr>
<tr>
<td>CompSES</td>
<td>cg02966332</td>
<td>0.00043331</td>
<td>-0.0587546</td>
</tr>
<tr>
<td>CompSES</td>
<td>cg06066601</td>
<td>0.00015817</td>
<td>-0.0915649</td>
</tr>
</tbody>
</table>
Supplementary Table 5.2

| DMR | Association | CpG     | \( p \)-value | Delta beta (\(|\Delta \beta|\)) | Chrom. | UCSC gene name | UCSC gene region |
|-----|-------------|---------|---------------|-------------------------------|--------|----------------|------------------|
| 1   | CompSES     | cg25356886 | 0.00015666    | -0.0709474                    | 2      | CRYGD          | Body             |
|     |             | cg25429719 | 0.00038003    | -0.1273029                    | 2      | CRYGD          | TSS200           |
| 2   | CompSES     | cg25843003 | 0.00017652    | 0.05901498                    | 6      | HCP5           | 3'UTR            |
|     |             | cg18808777 | 2.05E-05      | 0.08997175                    | 6      | HCP5           | 3'UTR            |
| 3   | CompSES     | cg03306615 | 0.00014735    | -0.1375116                    | 11     | ASCL2          | TSS1500          |
|     |             | cg26051413 | 0.00021987    | -0.1702083                    | 11     | ASCL2          | TSS1500          |
|     |             | cg13762320 | 0.00042162    | -0.0905445                    | 11     | ASCL2          | TSS1500          |
|     |             | cg11644479 | 0.00023712    | -0.1148043                    | 11     | ASCL2          | TSS1500          |
|     |             | cg24353535 | 9.66E-05      | -0.1190188                    | 11     | ASCL2          | TSS1500          |
| 4   | CompSES     | cg09394785 | 0.00021844    | -0.0753635                    | 11     | ASCL2          | TSS1500          |
|     |             | cg09460553 | 0.00041728    | -0.1399785                    | 13     |                |                  |
|     |             | cg02315096 | 0.00012539    | -0.1585108                    | 13     |                |                  |
|     |             | cg18865445 | 0.00028475    | -0.1541414                    | 13     |                |                  |
|     |             | cg17797229 | 0.00038613    | -0.1931293                    | 13     |                |                  |
| 5   | PosCan      | cg23129930 | 0.00015588    | -0.0838737                    | 7      | HOXA6          | 1stExon          |
|     |             | cg05928186 | 0.00024912    | -0.0744607                    | 7      | HOXA6          | 1stExon          |
| 6   | PosCan      | cg13710086 | 7.3006E-05    | -0.0909291                    | 17     |                |                  |
|     |             | cg01443020 | 0.00049335    | -0.1321152                    | 17     |                |                  |
## Supplementary Table 5.3

<table>
<thead>
<tr>
<th>Association</th>
<th>CpG</th>
<th>p-value</th>
<th>FDR p-value</th>
<th>Delta beta ((\Delta\beta))</th>
</tr>
</thead>
<tbody>
<tr>
<td>CompSES</td>
<td>cg02315096</td>
<td>0.0136646</td>
<td>0.145755682</td>
<td>-0.172215</td>
</tr>
<tr>
<td>CompSES</td>
<td>cg02556042</td>
<td>0.02623619</td>
<td>0.167911635</td>
<td>-0.129417</td>
</tr>
<tr>
<td>CompSES</td>
<td>cg03330490</td>
<td>0.00403811</td>
<td>0.098718093</td>
<td>0.05830829</td>
</tr>
<tr>
<td>CompSES</td>
<td>cg07451370</td>
<td>0.02196388</td>
<td>0.156184047</td>
<td>-0.0866381</td>
</tr>
<tr>
<td>CompSES</td>
<td>cg14087413</td>
<td>7.39E-05</td>
<td>0.004728329</td>
<td>-0.1885805</td>
</tr>
<tr>
<td>CompSES</td>
<td>cg16761754</td>
<td>0.0167502</td>
<td>0.153144701</td>
<td>0.22322337</td>
</tr>
<tr>
<td>CompSES</td>
<td>cg21721566</td>
<td>0.0415096</td>
<td>0.241510396</td>
<td>-0.1507618</td>
</tr>
<tr>
<td>CompSES</td>
<td>cg21892295</td>
<td>0.00616988</td>
<td>0.098718093</td>
<td>0.19760927</td>
</tr>
<tr>
<td>CompSES</td>
<td>cg22713356</td>
<td>0.00532333</td>
<td>0.098718093</td>
<td>-0.070068</td>
</tr>
<tr>
<td>CompSES</td>
<td>cg23745839</td>
<td>0.02045977</td>
<td>0.156184047</td>
<td>-0.0605778</td>
</tr>
<tr>
<td>CompSES</td>
<td>cg24353535</td>
<td>0.04644109</td>
<td>0.2476858</td>
<td>-0.095346</td>
</tr>
<tr>
<td>CompSES</td>
<td>cg25429719</td>
<td>0.01268284</td>
<td>0.145755682</td>
<td>-0.1436036</td>
</tr>
<tr>
<td>PosCan</td>
<td>cg03609398</td>
<td>0.00210781</td>
<td>0.040891504</td>
<td>0.17576249</td>
</tr>
<tr>
<td>PosCan</td>
<td>cg04467589</td>
<td>0.03214918</td>
<td>0.271474505</td>
<td>0.06417462</td>
</tr>
<tr>
<td>PosCan</td>
<td>cg06744585</td>
<td>0.01558784</td>
<td>0.18900253</td>
<td>-0.0585618</td>
</tr>
<tr>
<td>PosCan</td>
<td>cg07910560</td>
<td>0.03756607</td>
<td>0.271474505</td>
<td>-0.0594183</td>
</tr>
<tr>
<td>PosCan</td>
<td>cg11850943</td>
<td>0.04955304</td>
<td>0.271474505</td>
<td>0.06798295</td>
</tr>
<tr>
<td>PosCan</td>
<td>cg14453935</td>
<td>0.00039672</td>
<td>0.019240826</td>
<td>-0.1360803</td>
</tr>
<tr>
<td>PosCan</td>
<td>cg15282632</td>
<td>0.0019624</td>
<td>0.040891504</td>
<td>-0.1373493</td>
</tr>
<tr>
<td>PosCan</td>
<td>cg16443667</td>
<td>0.00133221</td>
<td>0.040891504</td>
<td>-0.0524907</td>
</tr>
<tr>
<td>PosCan</td>
<td>cg17492954</td>
<td>0.04441765</td>
<td>0.271474505</td>
<td>-0.0519896</td>
</tr>
<tr>
<td>PosCan</td>
<td>cg18031880</td>
<td>0.00673665</td>
<td>0.10890913</td>
<td>-0.123132</td>
</tr>
<tr>
<td>PosCan</td>
<td>cg18808777</td>
<td>0.02171587</td>
<td>0.234048826</td>
<td>0.10168585</td>
</tr>
<tr>
<td>PosCan</td>
<td>cg23129930</td>
<td>0.035589</td>
<td>0.271474505</td>
<td>-0.0729138</td>
</tr>
<tr>
<td>PosCan</td>
<td>cg24375409</td>
<td>0.04422148</td>
<td>0.271474505</td>
<td>-0.0824201</td>
</tr>
<tr>
<td>PosCan</td>
<td>cg27034935</td>
<td>3.15E-05</td>
<td>0.003059637</td>
<td>-0.1474872</td>
</tr>
</tbody>
</table>