DNA methylation patterns derived from fetal vulnerability to maternal smoking relate to future child outcomes

by

Jane WY Ng

A thesis submitted in partial fulfillment of the requirements for the degree of

Doctor of Philosophy
Department of Physiology
University of Alberta

Scientific Abstract

Maternal smoking during pregnancy (MSP) has an independent and causal effect on fetal health outcomes. Through accumulating epidemiologic and experimental evidence, our understanding of the breadth and duration of health effects of this toxic exposure is expanding. MSP has been linked to the etiology of many non-fatal, non-communicable common complex diseases (CCDs) such as depression and anxiety, poor cognitive performance, asthma, cardiovascular disease, diabetes and obesity. Despite the potency and prevalence of this exposure around the world, the mechanisms mediating these effects on human health are still unknown.

While numerous studies hypothesize that MSP dysregulates fetal developmental programming through epigenetic modifications such as DNA methylation (DNAm), there is yet not a single clinically useful epigenetic marker for CCDs. This is despite an explosion of human cohorts and animal models studying DNAm since about the early- to mid-2000's. This failure is juxtaposed with the success and rapid advancement of epigenetic markers and therapies in cancers of multiple forms within a similar period.

Perhaps one of the greatest barriers to clinical translation is the "gap" between genes, epigenetics and phenotype in complex traits. A frequent finding in association studies of CCDs is that many individuals may have shared phenotypic traits, but at best weakly share individual environmental risk factors or genetic/epigenetic markers. This gap further widens with factors such as varying intronic genetic mutations, phenotype heterogeneity and complex gene x environment (GxE) interactions.

In this thesis, we consider that mis-assignment of the contribution of genetic and environmental factors relevant to any given individual can lead to false conclusions regarding its effect on observed disease and/or epigenetic manifestations. Moreover, we reason that epigenetic differences persistent and potent enough to underlie the pathogenesis of CCDs must shift the mechanics of regulation across the genome. Thus, we speculate that DNAm related to CCD must alter chromosomal activity by orchestrating changes in DNA interactions that are stably maintained and have regulatory consequences on multiple genes through modification of their physical contact with chromosomal and other nuclear structures. Based on these premises, we explore context-based mapping of two entities: 1) individual-level risk profiling based on vulnerability to exposure rather than exposure alone and 2) DNAm profiling based on genome-wide patterns rather than single feature differences. In this way, we seek the *relevance* of a given MSP-related signal by

couching it within its clinical-level and genome-level context in order to visualize and adjudicate its relation to health.

We apply this context-dependent analytic approach to population-based data from the UK pregnancy cohort, Avon Longitudinal Study of Parents and Children (ALSPAC). These children have wide-ranging vulnerability to MSP and heterogeneous physical and mental outcomes – an ideal situation to model common epigenetic pathways among diverse GxE contexts for complex traits. We recruit specific multidimensional data analysis methods to extract a small number of DNAm patterns found in cord blood that are representative of fetal vulnerability to MSP. The biological coherence of these patterns is supported by three main findings. First, patterns are enriched for sites of chromosomal regulation at both genic and intronic regions. Second, specific patterns are shared among phenotypically similar children throughout childhood suggesting a common epigenetic shift underlying their physical and mental developmental trajectories. Third, most patterns persisted in blood collected in middle childhood and adolescence despite random, technical and physiologic methylation changes expected over time. This supports a robust relation to stable phenotypic effects starting from fetushood. We observed replication of these vulnerability patterns in cord DNAm data in an independent cohort (Generation R, Netherlands). These patterns were derived directly from ALSPAC with no additional clinical data from Generation R. Yet, these “template-based” DNAm patterns related similarly to later childhood phenotype within Generation R as in ALSPAC.

The novelty of this work lies in its use of context-based patterns of risk and epigenetic differences to provide a more detailed map of complex trait architecture. As envisioned by biologists like Conrad Waddington over a half century ago, the overlap of such maps - rather than unimodal data points - may provide deeper and potentially more accurate insights into the molecular underpinnings of complex diseases.

Preface

This thesis is an original work by Jane Ng. No part of this thesis has been previously published.

This research uses data from the Avon Longitudinal Study of Parents and Children (ALSPAC) cohort. Ethical approval was granted from the ALSPAC Law and Ethics Committee and the local Research Ethics Committee in accordance with the guidelines of the Declaration of Helsinki

Dr. Bei Jiang at the University of Alberta was involved with concept development for analysis described in Section 2.4.3 and 2.5.2.

Some of the research conducted for this thesis forms part of an international research collaboration with Associate Professor Janine F. Felix at the Erasmus University Medical Centre Rotterdam. Dr. Felix conducted the analysis described in this work on Generation R cohort DNA methylation data described in Section 3.8.

Dedication

For my family,

who have shared me with and shared with me my endeavours. You rascals are my heart and soul.

So the universe is not quite as you thought it was. You'd better rearrange your beliefs, then.

Because you certainly can't rearrange the universe.

-Isaac Asimov

Acknowledgements

This thesis would never have seen the light of day without the unwavering support of Prof. David Olson. His enthusiasm, hard work and boundless energy still awe me - even nearly a decade after our first meeting. Our first meeting was while we were both on different continents via a grainy skype video call a few months after having my first child. Dave, I could not have wished for a wiser or kinder mentor and hero in my life. Thank you.

I would never have dreamed I could turn this passion into research in epigenetics without Prof. Caroline Relton. With her guidance and laboratory (then at Newcastle University,) she armed with knowledge, resources and skills that would have been otherwise completely inaccessible to me. Caroline, your work and stewardship will always be an inspiration to this field and myself.

To Prof. Greg Funk and Prof. Po-Yin Cheung, thank you for sticking with me through thick and thin on this long road. Your encouragement lifted me up when I was sure I should just go back to my day job. As well, I would like to thank Prof. Peter Mitchell for his support during my fledgling thesis years. I am deeply grateful to Dr. Bei Jiang and Prof. George Davey Smith for their provocative and challenging questioning of my thought processes and assumptions. As well, thank you to Prof. Gerlinde Metz, Prof. Igor Kolvachuk and Dr. Sheena Fang for their generous support (aka rescuing) whenever I needed.

I am so very thankful to the families of the ALSPAC study, the midwives for their help in recruiting them, and the whole ALSPAC team, which includes interviewers, computer and laboratory technicians, clerical workers, volunteers, managers, receptionists and nurses. Many thanks to the ALSPAC researchers that administer and enhance this precious data resource. In particular, I would like to thank Dr. Kate Northstone, Dr. Laura Howe, Dr. Nicole Warrington, Prof. Tom Gaunt, Dr. Sue Ring, Dr. Hannah Elliott and Dr. Kate Potter for their guidance.

I am deeply grateful to Dr. Janine F. Felix and the Generation R Study Group for the immeasurable value of testing in an independent cohort.

I am fortunate to have received generous funding from the Biotechnology and Biological Sciences Research Council (UK Research and Innovation), Newcastle University, Genomic & Epigenomic Complex Disease Epidemiology programme (European Commission Community Research and Development Information Service), Molly Towell Perinatal Research Foundation, and the Women and Children's Research Institute (latter two both from the University of Alberta). Without this support, I would not have been able to take this time or my family on this PhD journey.

Table of Contents

List of Tables

List of Figures

Abbreviations

3C	Chromosome conformation capture
AHRR	Aryl-hydrocarbon receptors repressor gene
ALSPAC	Avon Longitudinal Study of Parents and Children
ANOVA	Analysis of variance
ARIES	Accessible Resource for Integrated Epigenomics Studies
BCD	Bisulphite-converted DNA
BH	Benjamini and Hochberg
BMI	Body mass index
BOP	Blocks of probes
CCD	Non-fatal, non-communicable common complex disease
cg	Prefix for Illumina Infinium ® CpG loci identification
CpG	Cytosine-phosphate-guanine (dinucleotide)
CTCF	CCCTC-Binding Factor
CTD	Comparative Toxicogenomics Database
CYP1A1	Cytochrome P450 family 1 member A1
DMR	Differentially methylated regions
DNAm	DNA methylation
DOHaD	Developmental origins of health and disease hypothesis
EWAS	Epigenome wide association study
FDR	False discovery rate
FTO	Fat mass and obesity-associated gene
GECKO	Groningen Expert Center for Kids with Obesity
GFI1	Growth factor independent 1 transcriptional repressor gene
GxE	Gene x environment interaction
IV	Instrumental variable
kb	Kilobases (referring to DNA base pairs)
MANOVA	Multivariate analysis of variance
ML	Machine learning
MoBa	Norwegian Mother and Child Cohort
moloc	Multiple-trait-colocalization
MSEP	Mean Square Error of Prediction
MSP	Maternal smoking in pregnancy
p>>n	Number of variables greatly exceeds number of observations/subjects
PAI	Promoter anchored chromatin interaction
PCA	Principal component analysis
PLS	Partial least squares
QTL	Quantitative trait locus
R^2	Coefficient of determination
RF	Random forest
RNA-seq	RNA-sequencing
SNP	Single nucleotide polymorphism
TAD	Topologically associated domain
TF	Transcription factor
TSS	Transcription start site
WISC	Wechsler Intelligence Scale for Children-III ^{UK}
WGBS	Whole genome bisulphite sequencing

Chapter 1 Introduction

1.1 Background

1.1.1 Developmental origins of health and disease hypothesis and common disease – a convergence of multiple dimensions

Common complex diseases (CCDs) describe a group of disorders that are non-fatal and non-communicable. However, 15 million people per year die from CCDs, rendering it the greatest cause of mortality across the world (Smith, Taylor F., Maccani, & Knopik, 2012). Sadly, over 85% of these deaths are under age 70. Moreover, CCDs cause the greatest burden of premature and chronic disability. Examples of diseases include depression (Williams *et al.*, 1998), cognitive disorders, (Forray & Foster, 2015; Heinonen *et al.*, 2011), asthma (Gilliland *et al.*, 2001), cardiac disease (Leybovitz-Haleluya *et al.*, 2018) and obesity and diabetes (Li, L. *et al.*, 2016; Rogers, 2019). Various lines of evidence point to early-life as the etiologic origins of many CCDs (Rogers, 2019; Roseboom, T., de Rooij, & Painter, 2006). In other words, they may result from the gene and environment interactions (GxE) that alter the developmental trajectory of health starting from fetal life.

In recent years, many have suggested broadening GxE to refer to genetic x non-genetic factor interactions that lead to a phenotype (Smith, Martyn T., McHale, & de la Rosa, 2019). In part, this is motivated by the low rate of heritability (10% or less) uncovered through genome wide association studies (GWAS) for the vast majority of CCDs (Zhang, Y., Qi, Park, & Chatterjee, 2018). Even when considering non-disease traits, non-genetic factors must be at least half of the equation in estimating phenotype (Polderman *et al.*, 2015). Regardless, the most commonly used framework for understanding these interactions for CCD is known as the Developmental Origins of Health and Disease hypothesis (DOHaD) (see Barouki, Gluckman, Grandjean, Hanson, & Heindel (2012) for an excellent review). Around the 1980s, the DOHaD concept was popularized by David Barker and colleagues (Barker & Osmond, 1995). Its foundation lies in the proposed link between low birth weight caused by fetal undernourishment to increased risk of stroke and coronary heart disease in adulthood. Subsequently, this concept has garnered further support from decades of observational human studies, most notably from the UK, northern Europe, the US and India (Barouki *et al.*, 2012).

The DOHaD framework posits that *in utero* environments offer a "long term forecast" for the fetus indicating whether the postnatal environment will be adverse or protective. In response, the fetus adapts structurally and functionally in order to best survive the predicted postnatal conditions. In other words, the fetus undergoes *biological programming.* A common misconception is that the DOHaD framework links adverse early environments to poor health. Instead, it posits that poor health is linked to a mismatch between the prenatal forecast and the actual postnatal environment (Hales & Barker, 1992). Among various studies, those from the Dutch Famine Birth Cohort study are the most well-known to demonstrate this concept. During a 16-month period in World War II, embargoes in combination with a harsh winter restricted food supplies such that adult rations provided only between 400-800 calories per day (Roseboom, T. *et al.*, 2006). Individuals born to mothers exposed to this famine expected a calorie insecure environment postnatally but instead experienced resource-rich postnatal environments. This disparity is believed to account for this group's higher rates of cancer, cardiovascular disease, type II diabetes (Roseboom, T. *et al.*, 2006), obesity (Roseboom, Tessa J. *et al.*, 2001), schizophrenia (Bale *et al.*, 2010; Susser & Lin, 1992). Not only this, researcher observed a third dimension of time to be important: the gestational period of exposure to famine modified the relation to certain types of adult disease traits (Roseboom, T. *et al.*, 2006). For example, lower birth weight and impaired glucose tolerance in adulthood related to famine exposure from mid- to late-gestation, whereas early gestation exposure is related to a more atherogenic serum lipid profile, higher BMI, increased stress responsiveness and lower self-rated health in adulthood.

From the DOHaD framework, several life course models such as the cumulative effects (a.k.a. accumulation of risk) model and pathway model have developed in the field of epidemiology (Boyce & Ellis, 2005; Ellis, Essex, & Boyce, 2005). However, a common thread over time between the various models is that employing predictors that focus on the direct exposure alone will miss the greater context of an individual's risk of disease. A more accurate predictor should capture the "net" force of the exposure (Figure 1). In this way, we can view health as a "vector sum" of the magnitude and timing of interactions between the environment and genes.

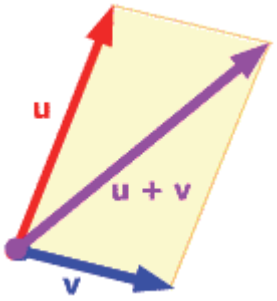


Figure 1: "Net" effect of GXE interactions using an algebra analogy. Vector u and v represent the influence on health of two predictors. Vector "u + v" is the net influence on health. Source: http://thejuniverse.org/PUBLIC/LinearAlgebra/LOLA/index.html.

Seeking to reflect reality more accurately, we consider that individuals are subject to various harmful and protective gene-environment interactions that shape their biological programming in early-life. If we are the products of such influences, then integrating multiple sources of information to capture the net effect may offer a better estimate of the overall impact on an individual's health trajectory (Figure 2).

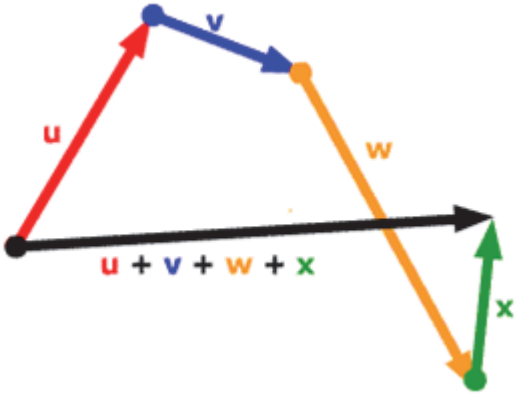


Figure 2: A more comprehensive predictor may combines multiple data sources to refine the estimation of the net effect of GXE interactions. Source: http://thejuniverse.org/PUBLIC/LinearAlgebra/LOLA/index.html

I draw upon this mathematical concept of vectors to evoke the multidimensionality of health. For graphical purposes, it is easiest to visualize this "net" effect of GXE as two-dimensional (2D) and linear vectors. However, this is very likely an immense and potentially misleading over-simplification. There is at least the third dimension of time and fourth dimension of "dose" of each factor. As well, drawing these as directional vectors assumes that the influence on health

by a given factor has the following properties: 1) absolutely known and accurately measured quantities, 2) remains the same over time 3) remains the same regardless of interactions with other factors, 4) is linear and 5) homogeneously affects all individuals in a population.

The intangibility of multidimensional health effects perhaps is more analogous to the famous Bohr's electron cloud model (Bohr, 1913). First described in 1913, it remains the most accepted model of the atom. In this model, estimates of electron location predict atomic chemical properties. The location of electrons is expressed in probabilities that are not equally distributed around the atom and influenced constantly by interactions with other forces. Thus, to estimate the probability of human vulnerability to disease due to genetic and non-genetic interactions may also require more sophisticated modeling than linear pathways. If humanity can model and thus harness the usefulness of even the smallest particles such as electrons, we can hope to similarly embrace the challenge of modeling omnipotent GxE influences to be clinically useful for the health of the population from birth to old age.

1.1.2 Barriers to translation – the gap between biologic markers and common complex disease

To understand the molecular underpinnings of CCD, there has been an explosion of GxE studies in the past decade using numerous biological markers (Bookman *et al.*, 2011; Ober & Vercelli, 2011; Welter *et al.*, 2013). However, yet a single test can provide patient-specific risk of complex disease development that is applicable to the general population.

Despite the failure of this area in making the clinical translation leap, the theoretical rationale and potentially immense health and economic impact globally continues to stoke research interest. Perhaps one of the greatest barriers to translation is the gap between genes, biologic markers and phenotype in CCDs (Manolio *et al.*, 2009; Petronis, 2010; Relton, C. L. & Davey Smith, 2010). Factors such as genetic variants and pleiotropy, phenotype heterogeneity and direct and indirect GxE interactions further widen this gap. Mis-assignment of the contribution of genetic and environmental factors can lead to false conclusions regarding effect on phenotype. Based on this premise, this thesis explores shifting the view of individual risk classification and biologic profiles from separate and independent entities to that of relative predisposition using multi-factor context based features. In order to unravel this context, it is critical to acknowledge the dimensions of time, exposure and phenotype.

1.1.2.1 Time dimension

When speaking of “origins in early-life”, one typically refers to events occurring before adulthood. This has given researchers the flexibility to refer to one or more windows in which their environmental factor of study is likely to have an effect on the individual. As such, this could refer to childhood, gestation, pre-conception and/or even the time of exposure to one’s ancestors (Kyle & Pichard, 2006; Roseboom, T. *et al.*, 2006). As mentioned, the DOHaD often refers to developmental programming influenced by exposures in early-life that may render an individual better or worse adapted to face later exposures in their environment (Barker & Osmond, 1995; Hales & Barker, 1992; McMillen & Robinson, 2005) . The major point is that both observational and experimental data have shown the effect of timing of exposure to be clinically relevant (Bosch *et al.*, 2012; Class, Lichtenstein, Långström, & D'onofrio, 2011; Roseboom, T. *et al.*, 2006). Thus, data from a single time point may have limited ability to predict health trajectories.

1.1.2.2 Exposure dimension

In 2005, Christopher Wild coined the phrase “exposome” to describe environmental influences (including social, behavioural and chemical) over a life course (Wild, 2005). He borrowed the suffix *–ome* to link this area to the other “omic” disciplines in biology. In contrast to the genome, which is static and can be precisely measured, the exposome is dynamic over time and often difficult to objectively and accurately measure. Whether it be exposure to famine, environmental pollutants or stress from war, abuse or poverty, scientists struggle with misclassification and/or imprecision of exposure estimates (Manrai *et al.*, 2017). Advances in areas such as pharmacokinetic models, smartphone-based sensors, geolocation technologies and self-reported questionnaire methodologies have all improved exposome assessment (Turner *et al.*, 2017). However, as discussed above, multiple repeat time point assessments add critical value to understanding common disease.

1.1.2.3 Phenotype dimension

Like the exposome, the phenome is a vast area involving organism-wide, high dimensional profiling of phenotypic traits. This discipline has expanded exponentially since its introduction in evolutionary biology a half century ago (Houle, Govindaraju, & Omholt, 2010). Also like the exposome, it is has an important time dimension where phenotype can vary from one time to the next and the rate of change can vary depending on life cycle stage. In addition, it has the added complexity of cellular heterogeneity. Increasingly, phenome and exposome fields are overlapping. For instance, exposome researchers are actively seeking the use of various phenome technologies as exposure biomarkers to improve accuracy and precision (Turner *et al.*, 2017). Of course, this is with caution to the possibility of reverse causality – did the omic difference result from the exposure or is the result of the disease? In addition, we realize better now that certain traits lead to specific health behaviours, further blurring the lines between phenome and exposome.

Viewing these three dimensions together, we appreciate that none can be extricated from another nor can be perfectly captured. Thus, the separation of these dimensions is not only artificial but also potentially misleading. Systems biology approaches aim to coalesce these dimensions into a "multi-omic" space that encompasses the various molecular omics (e.g. genome, epigenome, transcriptome, proteome, metabolome, and/or microbiome levels,) as well as the exposome and phenome. This multi-omic space acknowledges that complex systems are more than the sum of their parts and therefore components cannot be viewed individually but in context of its network relations (Martino, Ben-Othman, Harbeson, & Bosco, 2019). Arguably, medicine has enjoyed great success in many areas such as infection and cancer where accurate and effective treatment is led by individual and/or categorical patient characterization. However, we argue that most CCDs defy clear-cut definition due the effect of context on whether an individual will or will not succumb to disease (Figure 3).

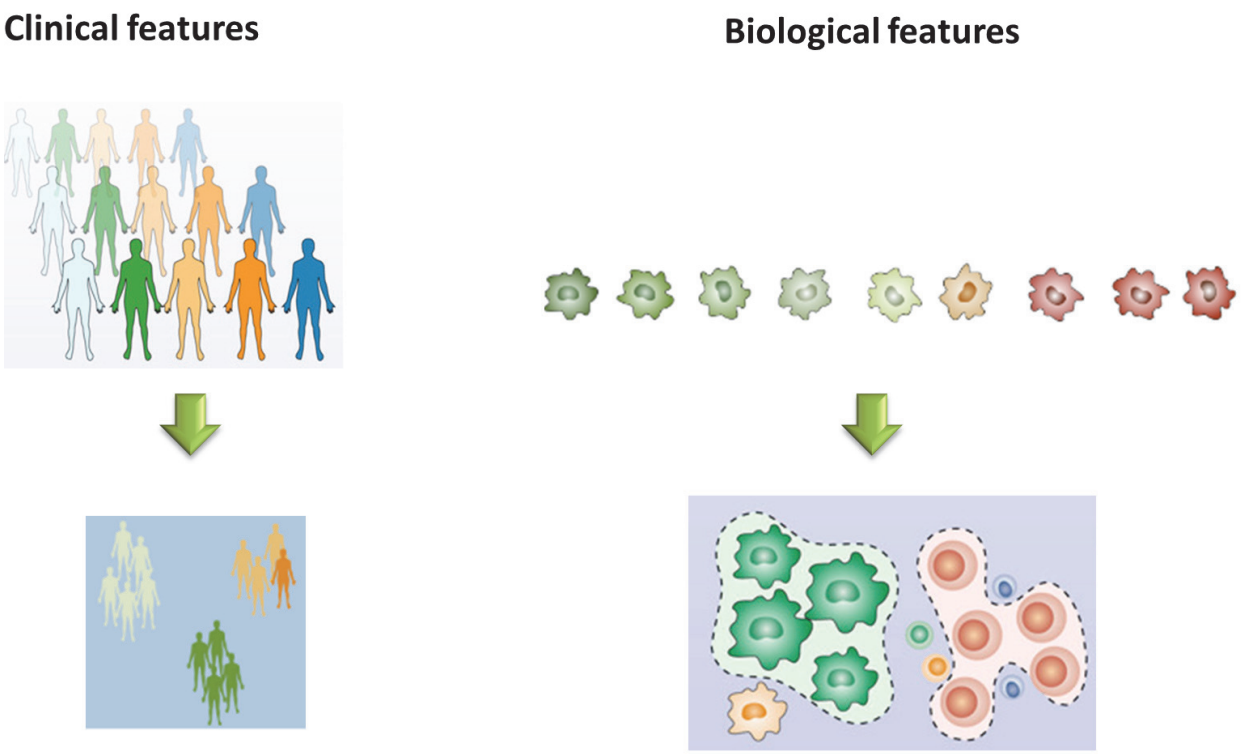


Figure 3: Schematic of traditional mapping paradigm for diseases amendable to categorical description. Disease definition for entities where clinical features, (e.g. timing, exposure,) and/or biological features can clearly inform medical management. Examples of such disease include cancer and infectious diseases. In these situations, features have clear definitions of typical versus atypical categories and these categories provide clinically relevant in that they inform diagnosis, prognosis and/or therapeutic target, presumably because the categories correlate strongly with the underlying molecular pathology.

By weaving together multiple sources of information into a context of disease predisposition over time, rather than prediction using any single piece of information by itself, one may be able to attenuate misdirection by misclassification, imprecision or irrelevance to the clinical question under study. This fits well with current trends in precision medicine and shows signs of success for CCDs (see Figure 4 of a recent example of real life application of multi-data source precision medicine.) Bringing this back to the GxE paradigm of DOHaD, we posit that each individual can be positioned in a multidimensional matrix of genetic and non-genetic factors that shape his/her health trajectory.

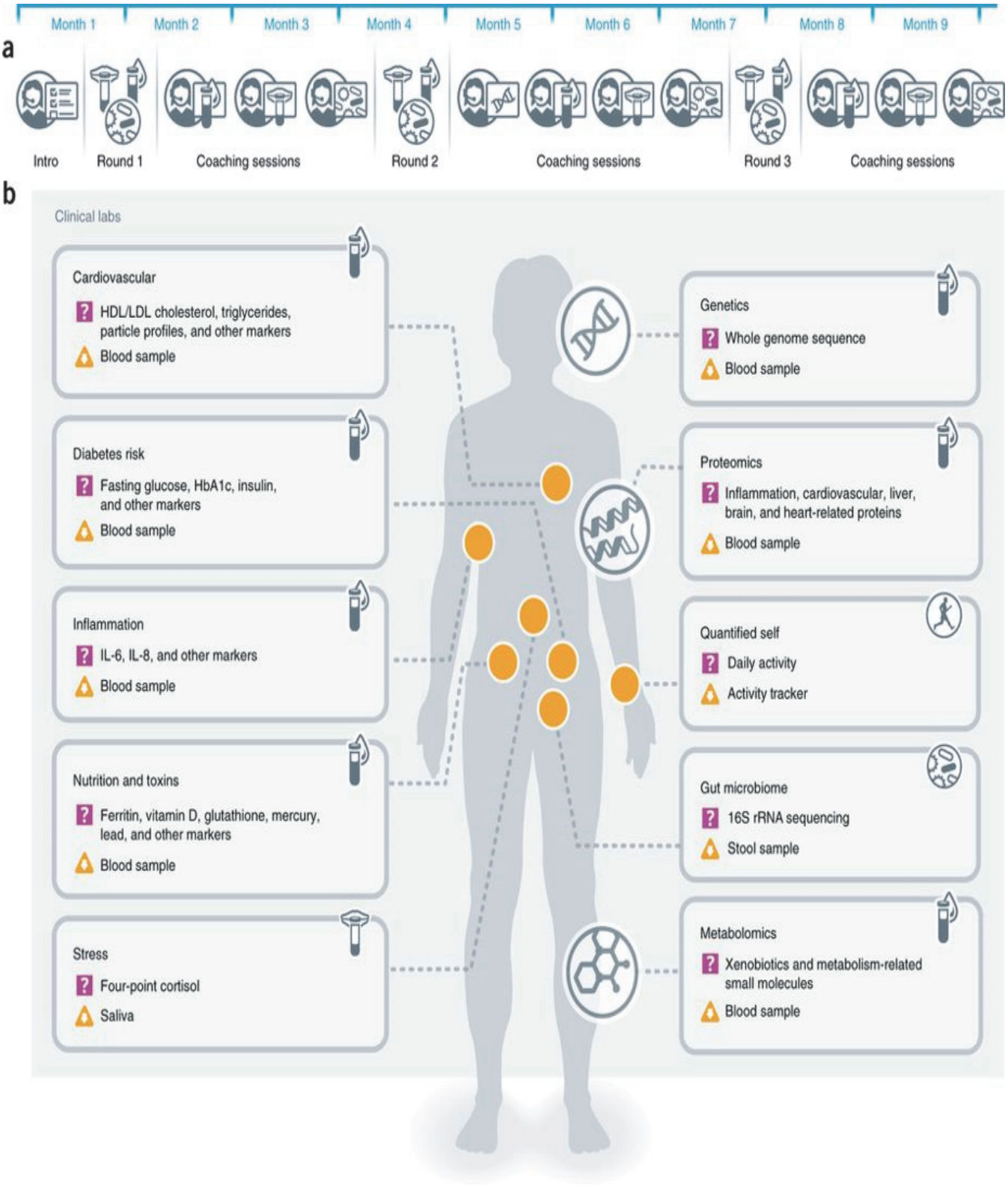


Figure 4: Example of real-life multiomic approach. The P100 study collected a dense data cloud of multiomic data for 108 individuals for 9 months. Actionable results were provided to subjects along with customized

behavioural coaching. Many subjects demonstrated improved clinical biomarkers (e.g. for diabetes and cardiovascular risk factors) during the study. Image from (Price, N. D. *et al.*, 2017).

We borrow from systems biology and apply a context-based approach to a widely used model of DOHaD: adverse effects of maternal smoking during pregnancy (MSP) and early and late childhood physical and mental outcomes. We employ epigenetic, anthropometric, questionnaire, and linked public repository data at multiple time point to develop an interconnected and multidimensional view of health trajectory.

1.2 Epigenetics and common complex disease

1.2.1 Overview of epigenetics

Over a half century ago, Conrad Waddington described a new field he called epigenetics. He defined it as "the branch of biology which studies the causal interactions between genes and their products which bring the phenotype into being" (Waddington, 1957). He used the metaphor of an "epigenetic landscape" where a given cell (depicted as a ball in the figure) is poised to take various paths to different cell fates (Figure 5).

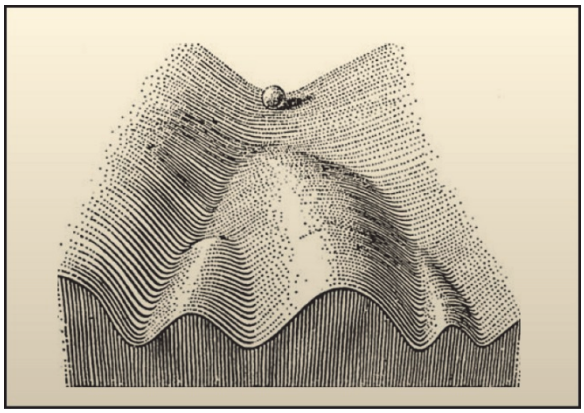


Figure 5: The "epigenetic landscape". This model was originally used to describe cellular decision-making during differentiation and development from an embryologic standpoint. A cell is represented by the ball and the landscape forms paths leading to different cell fates. Figure reprinted from (Waddington, 1957).

On a molecular level, epigenetic changes were initially defined as chemical modifications that alter the physical coiling and looping structure of DNA (Figure 6).

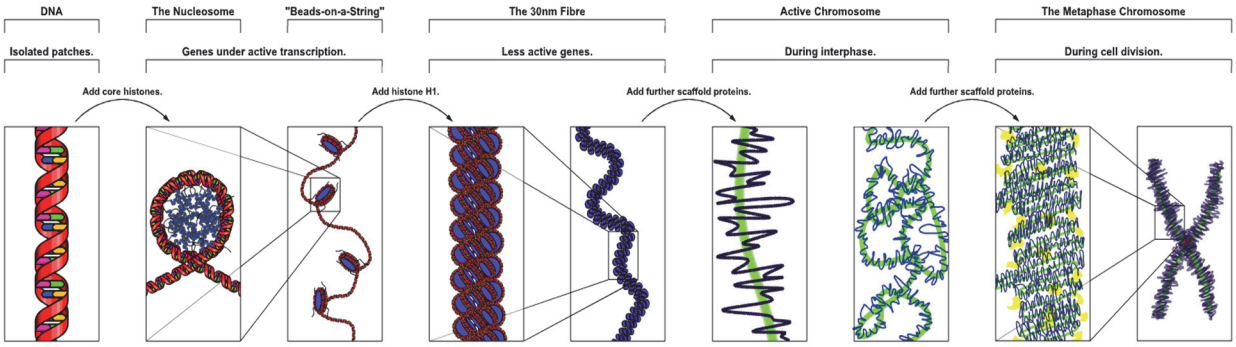


Figure 6: Successive stages of chromosome compaction depend on the introduction of additional proteins. The differential packaging of genomic regions leads to physically different accessibility to the various components of regulatory and transcriptional machinery. Relatively accessible areas of chromatin are referred to as "open" whereas those tightly associated with various proteins as part of higher-order chromatin assembly are called "closed". Source: Richard Wheeler at en.wikipedia.

While the famous landscape metaphor is visually powerful, this description paints a more analytically practical image of these structural modifications: "Patterns of activation and silencing, known as the epigenome, exist across all the genes in a cell." (Lamb, N., 2007) These patterns are intimately connected to chromosomal function and ultimately, the flow of information to and from DNA to cellular phenotype. It is clear that epigenetic factors are a critical part of normal and necessary biologic processes, such as X-inactivation, cellular differentiation and genomic imprinting. It is one among many molecular mechanisms that exert force on the cellular and nuclear environment to influence DNA function (

Figure 7). However, its research fascination springs from its posited role in continuously coordinating adaptation by relaying information from the external environment to influence chromosomal changes. Thus, epigenetics could be seen as patterns of environmental imprints that modify genome structure which in turn contribute to how external influences "get under the skin" to alter health (Boyce & Ellis, 2005; Ellis, Essex, & Boyce, 2005; Pluess & Belsky, 2011).

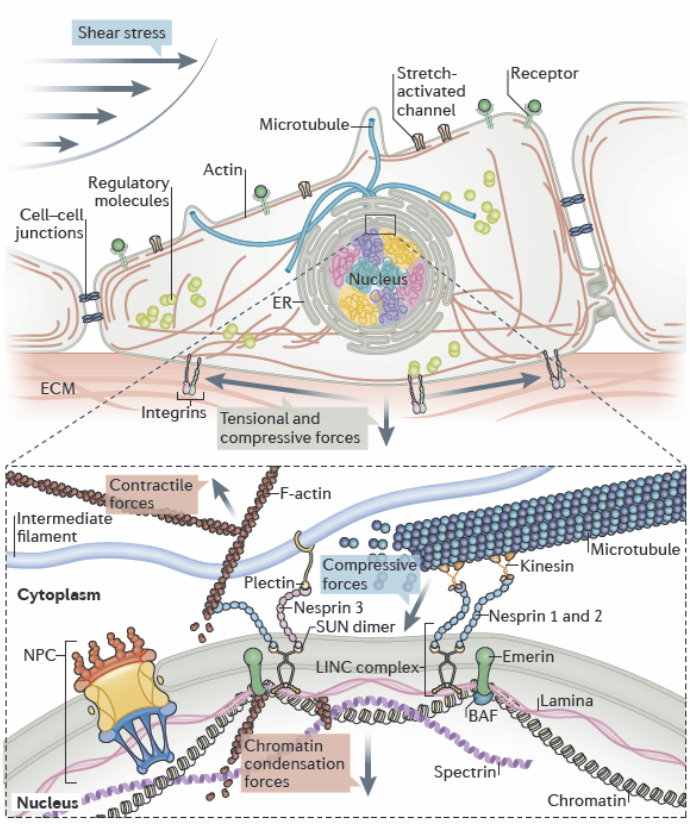


Figure 7: The forces of DNA interactions in eukaryotic cells. The cellular microenvironment consists of both physical and chemical signals which are transduced by various protein assemblies that connect the cell surface, cytoplasm, nuclear envelope, and nuclear compartments. Image from Annunziato (2008).

The ever changing local and global architecture of chromatin emerges from various layers of cross talk between epigenetic mechanisms such as DNA methylation, nucleosome positioning (modulated by ATP-dependent chromatin remodeling machines), histone modifications, small RNAs, non-coding RNAs and topographical location within the nucleus that coordinates gene regulation and transcription product features (Fischle, Wang, & Allis, 2003; Geiman & Robertson, 2002). Studies exploiting the 3-D visualization of gene regulation indicate that "cis effects" are just the tip of the iceberg of chromosome regulation: interactions between epigenetic mechanisms have been shown to effect whole domains of chromatin or even a whole chromosome (Fischle *et al.*, 2003). Where it once thought that euchromatin is transcriptionally active compared to inert and tightly coiled heterochromatin, it is now clear that even heterochromatin has conformationally flexible and thus active or poised domains (Tchasovnikarova & Kingston, 2018). Besides exerting influence on DNA shape, epigenetic marks direct nuclear location of chromatin during interphase. Epigenetic signals orchestrate a concert of DNA configurations with clusters of transcription-related factors – together described

as a “transcription factory” (Figure 8) (Cook, P. R. & Marenduzzo, 2018; Uhler & Shivashankar, 2017).

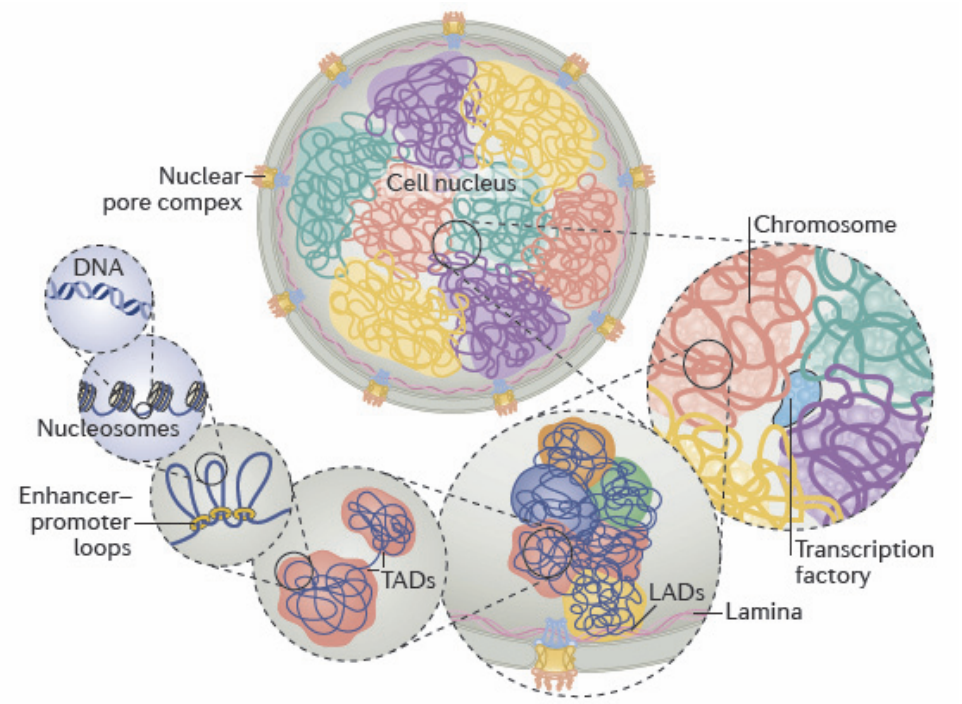


Figure 8: The 3D organization of chromatin is non-random and links nuclear morphology, chromosome organization and gene expression in a manner that tends to be evolutionarily conserved within cell types. Using fluorescence in situ hybridization and various chromosome conformation capture techniques, studies have demonstrated the interaction of chromosomes with various distinct nuclear territories that lead to differential associations with transcription factories, chromosome conformations and inter-chromosomal intermingling. Image from Uhler & Shivashankar (2017). TAD – topologically associated domain. LAD – lamina associated domain.

Akin to a manufacturing line, the tools needed to perform a task (e.g. transcription, replication, repair, etc.,) are localized in high concentration at certain sites instead of being randomly scattered. By the law of mass action, this high concentration drives efficient processing. In this way, epigenetic marks coordinate the co-localization of foci of transcription machinery and receptive DNA conformations to promote efficient RNA production. For example, initiation requires the collision of a RNA polymerase with an accessible promoter region of a gene. As well, effective transcription requires the DNA helix to rotate both laterally and rotationally, manoeuvres which also require epigenetic mechanisms. Thus, epigenetic patterns shape gene expression through multiple means (e.g. guiding shape, nuclear location and possibly movement of DNA) that interactively alter physical dynamics at various sites (Fischle *et al.*, 2003; Uhler &

Shivashankar, 2017). Current research suggests that the majority of dynamic transcription occurs in topologically associated domains (TADs – see Figure 8), which range in size from 500 to 1000 kilobases (kb) (Mishra & Hawkins, 2017; Yu *et al.*, 2017).

Epigenetic modifications sit at the crossroads between these higher order interactions discussed above and the direct interactions with gene expression machinery. Currently, a commonly referenced definition of epigenetics originates from the NIH Epigenomics Roadmap Project initiative, which states, "Epigenetics refers to both heritable changes in gene activity and expression (in the progeny of cells or of individuals) and also stable, long-term alterations in the transcriptional potential of a cell that are not necessarily heritable" (Roadmap *et al.*, 2015). Using this definition, four categories can describe epigenetic mechanisms: 1) DNA modifications 2) chromatin modifications 3) non-coding RNAs that are involved in transcription regulation and transcript stability and 4) RNA modifications that can affect mechanisms such as splicing, transport, and stability of transcription products but also protein associations involved in chromatin regulation (see Figure 9).

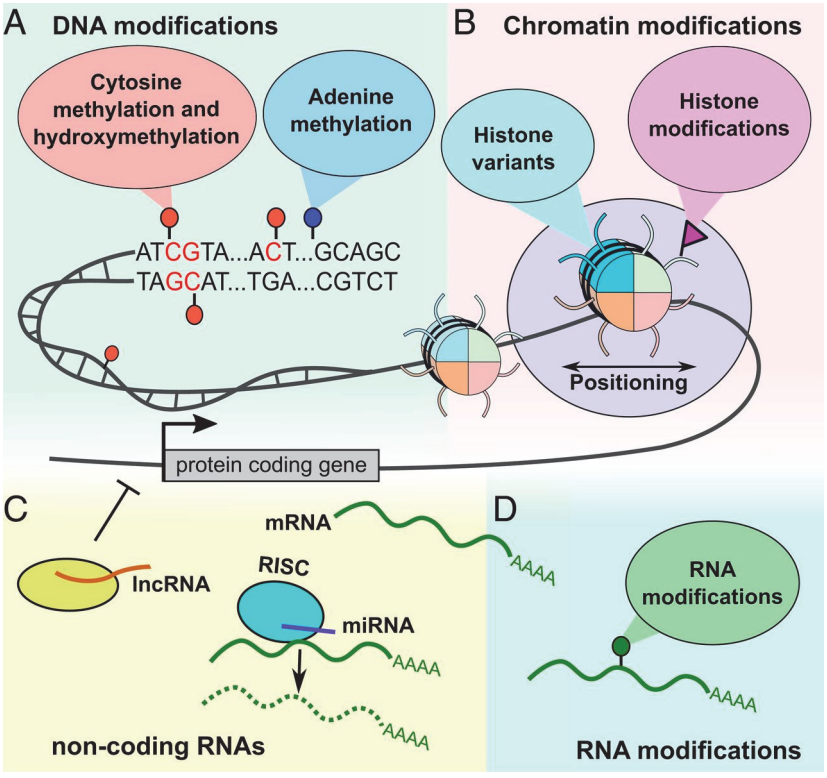


Figure 9: Classes of epigenetic mechanisms. Using a broad definition of epigenetics, the types of mechanisms could be divided into four classes of alterations that share the characteristic of being able to stably alter the transcriptional potential of a cell. Image from (Aristizabal *et al.*, 2019).

Of all epigenetic modifications, DNAm is the most widely studied. It involves methylation or hydroxymethylation at the five ‑ position of cytosines in cytosine – phosphate - guanine (CpG) dinucleotide sites (unless otherwise specified, DNAm in this work will refer only to methylation as hydroxymethylation is currently far less commonly assayed in human and animal studies.) Early in DNAm studies, CpG dense regions (known as islands) and gene promoters were the primary research focus. This was driven by the observation that hypermethylation of CpG islands generally lead to transcription repression (Bird, 1992). However, it is now clear that patterns of methylation along a DNA strand in CpG sparse regions and at enhancers, introns, exons and intergenic regions have varying and context dependent effects on gene regulation, making DNAm a far more complex entity than previously thought. Though we are far from fully understanding this epigenetic mechanism, it is obvious that metrics that are gene-centric (e.g. distance from transcription start sites (TSS) and promoters) or based on CpG density are restricted to describing the one-dimensional (1-D) linear sequence of DNA. These metrics bear variable correlation to the *in vivo* "appearance" of methylation in context of a dynamic, 3-D region of chromatin and thus may have limited functional relevance.

Clearly, this field is still at its infancy and we are only just beginning to see the forest for the trees. Nevertheless, we must resist adopting traditional views of processes as in signal transduction cascades that were functionally captured with largely forward flowing pathways onto downstream effects. As we shift our perspective from the linear DNA sequence and local epigenetic marks, we see that the functional substrate of gene regulation may be chromatin and nuclear organization itself. In this spirit, researchers have been scrambling to use 1-D epigenetic data to predict the 3-D form and therefore true functionality of the genome in a given cell (Di Pierro, Cheng, Aiden, Wolynes, & Onuchic, 2018). Specifically, several groups are developing high-throughput methods to predict how DNAm affects DNA shape and impact on DNA-protein interactions (Lazarovici *et al.*, 2013; Rao *et al.*, 2018). However, most epigenome wide studies (EWASs) in humans use DNAm data from microarray chips. To date, the most commonly used chip in EWAS is the Illumina HumanMethylation 450 K BeadChip® (450K chip) which covers around 450,000 methylation sites across the genome (Min, Hemani, Davey Smith, Relton, & Suderman, 2018; Sandoval *et al.*, 2011). Though data from the 450K chip has to date had little correlation with DNA shape directly, it has been instrumental in the initial foray into DNAm changes in various tissues across multiple CCDs. Several international consortia have collated 450K chip data to understand better this data in context of disease and variability from

technical artifacts, tissue type and population differences (Adams *et al.*, 2012; Roadmap *et al.*, 2015).

Not only is DNAm highly relevant to function, it is attractive in terms of feasibility. DNAm as measured through chips like the 450K chip are relatively inexpensive, technically robust, and high throughput (meaning can be upscaled to population wide testing,) compared with techniques like whole-genome bisulphite sequencing (WGBS). This latter technique is the current "gold standard" of DNAm quantification and employs random fragmentation of the genome followed by bisulphite sequencing to obtain complete genome CpG coverage. However, this comes at the cost of a large amount of sequencing (Suzuki, M. *et al.*, 2018). In contrast, the 450K chip only assays sites curated by the manufacturer. All Illumina DNA methylation assays use the same library to label genomic CpG loci facilitating cross-platform comparison and annotation, (identifiers all have the prefix *cg*.) The 450K chip is particularly attractive as it requires only a small amount of sample, remains stable in various storage conditions and has a well-standardized protocol (Forest *et al.*, 2018; Sandoval *et al.*, 2011). This is in contrast to assays for WGBS or RNA sequencing for example. This lends DNAm assays for use in large human cohorts for research purposes or a clinical test in the general population. Though the most common tissue used in studies to date is venous blood, successful and reliable DNAm data extraction in large cohort studies arise from sources such as buccal cells from mouth swabs (Forest *et al.*, 2018), archived newborn blood spot cards used for public health screening (Joo *et al.*, 2013), urine (O'Reilly *et al.*, 2019) and skin (Zhou, F. *et al.*, 2016).. The use of these less invasive tissue sources in clinical applications is being actively explored (Forest *et al.*, 2018).

At this point in technologic advancement, the study of CCD on a population level will require better utilization of 1-D epigenetic data to infer 3-D changes that are relevant to cell function. Given the current paucity of knowledge, we posit that the viewing of patterns across the genome may help avoid making false assumptions regarding what represents functionally relevant DNAm changes. Using patterns may be akin to using every other piece of a jigsaw puzzle to surmise the appearance of the whole image. Though it is still inadequate, we posit that the alternative of focusing on linear-based views of DNAm risks the bias of only collecting puzzle pieces of certain colour or shape – it is unclear if and unclear what may be missing. We posit that genome-wide pattern finding is an important avenue to explore to attenuate this risk.

1.2.2 Overview of common complex disease and precision medicine

As the name implies, the diseases so far discussed are both “common” and “complex”. Common in that these diseases are prevalent across race, sex and social class. The “complex” term typically refers to the cause being multi-factorial (usually a mixture of genetic and non-genetic factors) or has multiple single causes that lead to one shared disease description (Bookman *et al.*, 2011; Lewis *et al.*, 2007). This term could arguably also refer to the fact that these diseases often affect multiple organ systems or cause multiple morbidities. To date, research focus is predominantly trained on metabolic syndrome, (encompassing traits of insulin resistance, dyslipidemia, obesity, hypertension, etc.,) cardiovascular disease, mood disorders, cancer and neurologic disease, (e.g. Alzheimer disease and autism) (Buchanan, Weiss, & Fullerton, 2006). These groups of diseases receive immense research investment given their societal cost due to high prevalence, mortality and/or morbidity. The actual cost in terms of financial burden as well as loss of work force productivity and quality of life is compounded by the chronicity of these diseases and increasingly early age of onset, in addition to the commonly and long-observed inter-generational component within families (Murray & Lopez, 1997). Moreover, once clinically detectable, these diseases already tend to be intractable to amelioration and/or cure once. Thus, the most cost-effective and efficient public health battle strategy against CCDs would be pre-disease detection combined with patient- and disease-specific prevention. This is the major goal of precision medicine and drives a multi-billion dollar biomarker development market (Akhmetov & Bubnov, 2015).

Today, one of the great frustrations for both the patient and physician is it is unknown who will ultimately be affected by a given risk factor. Disease-risk estimates are derived from populations heterogeneous in their mixtures of vulnerable and resilient individuals and subject to forces such as mortality selection (i.e. dying from another disease related or not to the exposure) and thus has very limited applicability to any single individual. So, even though one can advise that smokers have as much as a 30% greater chance of death from prostate cancer than non-smokers (Huncharek, Haddock, Reid, & Kupelnick, 2010), the chances for the smoker in front of you is completely unknown. In addition, the rigours of the scientific method may inadvertently contribute to the lack of credence of medical community when risk and disease links are questioned or even disproved (see

Table 1 for a list of what are now considered equivocal disease associations (Buchanan *et al.*, 2006).

Table 1: Examples of disease associations with markedly inconsistent published results. From Buchanan *et al.*, (2006).

Table of irreproducible results?
Hormone replacement therapy and heart disease
Hormone replacement therapy and cancer
Stress and stomach ulcers
Annual physical checkups and disease prevention
Behavioural disorders and their cause
Diagnostic mammography and cancer prevention
Breast self-exam and cancer prevention
Echinacea and colds
Vitamin C and colds
Baby aspirin and heart disease prevention
Dietary salt and hypertension
Dietary fat and heart disease
Dietary calcium and bone strength
Obesity and disease
Dietary fibre and colon cancer
The food pyramid and nutrient RDAs
Cholesterol and heart disease
Homocysteine and heart disease
Inflammation and heart disease
Olive oil and breast cancer
Fidgeting and obesity
Sun and cancer
Mercury and autism
Obstetric practice and schizophrenia
Mothering patterns and schizophrenia
Anything else and schizophrenia
Red wine (but not white, and not grape juice) and heart disease
Syphilis and genes
Mothering patterns and autism
Breast feeding and asthma
Bottle feeding and asthma
Anything and asthma
Power transformers and leukaemia
Nuclear power plants and leukaemia
Cell phones and brain tumours
Vitamin antioxidants and cancer, aging
HMOs and reduced health care cost
HMOs and healthier Americans
Genes and you name it!

It is clear that medicine has struggled in the realm of CCD to deliver precise and early diagnoses and treatment. Currently, preventative measures are often either delivered on a community level or based on patient risk factor profiles. The former method attempts to reduce risk on a population level. Examples include public messages advocating for better diet, exercise and warnings to avoid toxins such as smoking, alcohol and drugs. These measures are common and affordable, but have low efficacy (Ekpu & Brown, 2015; Howells, Musaddaq, McKay, & Majeed, 2016). The latter method targets patient-specific risk behaviours. There is some limited success of such interventions, but often only available in trial settings or in small subsets of populations. To date, cost renders none feasible for population-wide public health implementation (Cnattingius, 2004). Nevertheless, ongoing development of these measures are driven by accumulating data that patient-tailored management of complex disease is more effective than general interventions in terms of effective changes in health behaviour and/or health status (Bennett *et al.*, 2010; Strecher, Wang, Derry, Wildenhaus, & Johnson, 2002). Thus, the ability to measure a patient's specific predisposition to an illness may not only be important diagnostically, but for preventive intervention in areas such as patient adherence. Mobile and online health applications exploit this aspect of patient psychology by providing individualized health counselling and monitoring. A meta-analysis of studies providing genetic testing for complex disease concurs with this trend, even when no possible prevention or cure exists (Frieser, Wilson, & Vrieze, 2018).

For these reasons, biomarkers hold theoretical and practical promise of precision medicine to stem the alarming rising tide of complex disease seen across the globe. Currently, nearly all clinical biomarkers used in the general population rely on some degree of organ dysfunction. For example, the clinical categorization "diabetic" is simply marking the passage of the individual from below to above threshold on a spectrum of glucose tolerance. To render biomarkers capable of detecting the pre-disease status, we require a better understanding of 1) the full spectrum of disease vulnerability and progression on a patient specific level and 2) to do so before overt cellular dysfunction (Martino *et al.*, 2019).

Another important aspect of complex disease is that a patient may present initially with only one organ dysfunction, but his/her risk of multi-organ involvement or co-morbidities is high. The dependency between the physiologic dysfunction of one system and the likelihood of developing dysfunction in another system is well recognized, but poorly understood. This knowledge gap increases disease heterogeneity and undermines efforts to understand and prevent the multi-organ spread of CCDs (Bookman *et al.*, 2011; Sanavia, Aiolli, Da San Martino, Bisognin, & Di

Camillo, 2012). Uncovering the common molecular threads between physiologic systems in disease evolution may distill multisystem dysfunction into common denominators that could hold the key to global organ recovery.

Precision medicine seeks to provide individualized diagnosis and treatment that is pre-clinical and disease modifying. To succeed in this endeavour for complex diseases, the field must find molecular signatures linked to disease etiologies that exist on different genetic and environmental backgrounds (Dover, 2009; Martino *et al.*, 2019). Based on this information, one could offer patient-specific management that has better compliance than general advice and would better target finite health resources. Additionally, what if the health care provider could also tell if your loved ones like your children or partner were similarly affected? This would help synergize family-based adherence that has far more powerful clinic effect than, for example, one family member dieting while other members eat potato chips (Trivedi & Asch, 2019). Last, what if one could track and feedback the health status, while still in the pre-clinical stage? This may help sustain compliance and tailor the intensity, duration and/or type of intervention. These data could also be used to better design and evaluate interventions (Akhmetov & Bubnov, 2015). As well, historically and currently health resources focus on the most seriously ill both in terms of acute and long-term care for medications and/or utilization of health care personnel and medical facilities. For individuals suffering at these extremes of disease, it is an unfortunate reality that the efficacy of treatment and gains in quality of life and workforce strength is the absolute lowest. It is also at these extremes that the likelihood of developing subsequent complications and co-morbidities increase. Thus, the power of personal data in modifying individual and inter-generational health could be exponential through the synergy of enhanced patient-level (Mirowsky & Ross, 2015) and medical-level (Price, N. D. *et al.*, 2017) efficacy, as well as cost-effectiveness.

Related to the last point, targeting health resources to only those at risk of disease has other non-health related benefits. For instance, it is a fact worldwide that cost-based policies are the most effective public health intervention (Ekpu & Brown, 2015). As an example, a 10% increase in tobacco tax reduces smoking prevalence by up to 8%. However, increased taxation is associated with increased cigarette smuggling, theft, counterfeiting, and tax evasion, as well as being unpopular with powerful tobacco industry lobbyists. Moreover, governments benefit from tobacco industry activities, (ranging from crop production, processing, marketing, distribution, etc.,) through tax revenues, employment opportunities and economic stimulation. More morbidly, premature death saves governments money on expenses such as senior benefits,

disability support and pensions. Thus, patient-specific allocation of resources makes strong economic sense.

Risks abound across the population and it is impossible and possibly unhelpful with finite resources and changing government agendas to attempt to provide intervention to all individuals. Patient-specific and pre-disease detection is required to target the most vulnerable individuals for optimal treatment efficacy and economic efficiency. In the next section, we consider epigenetics as a potentially viable test of individual-specific vulnerability to exposures related to CCDs.

1.2.3 DOHaD, DNAm and common complex disease

Using data from the Dutch Hunger Winter, researchers have found distinct difference in DNA methylation 60 years later in subjects exposed to early versus mid and late gestational exposure to famine (Tobi *et al.*, 2015). Again, this points back to the importance of timing in accurately interrogating the underlying mechanisms relating early-life exposome and later life phenome. As well, several lines of epidemiologic and experimental evidence converge to support three specific lifecourse time points of intervention: the in utero, postnatal to infancy, and peri-pubertal periods (Murgatroyd & Spengler, 2011; West-Eberhard, Mary Jane, 2003). These time points optimize the balance of heightened epigenetic plasticity and adaptation to novel challenges to extract maximal gain in human "biological capital" (Barouki, Gluckman, Grandjean, Hanson, & Heindel, 2012; Burdge & Lillycrop, 2010; Godfrey, Costello, & Lillycrop, 2016). Moreover, environment sensitive epigenetic changes during these periods are long lasting. Can epigenetics be the game-changer for our understanding of CCD? To consider this question, we consider how precision medicine, CCD and epigenetic changes can be juxtaposed biologically and methodologically for possible integration.

Buchanan and colleagues summarized a powerful metaphor of CCD using an hourglass (Figure 10). Patient phenomes, genomes and exposomes may vary widely and may flow one from the other in complex pathways. For example, the flows may be multi-directional, time-dynamic, overlapping, and/or stochastic. Despite the breadth of these factors, this concept theorizes that a given complex disease will converge through a common channel of molecular derangement, akin to the hourglass neck that connects far broader bases. If DNAm can detect the net result of

genetic and non-genetic interactions on chromatin poise, it would be an ideal place to look for critical common molecular mechanisms. As discussed in Section 1.1.2: Barriers to translation – the gap between biologic markers and common complex disease, it reduces our reliance on exposome and phenome measures that unlikely wholly capture disease vulnerability and are difficult to accurately quantify in the first place. DNAm may enhance measurement accuracy by shifting away from traditional clinical risk algorithms that coerce internally heterogeneous variables into pure categories. In other words, clinical risk "calculators" rely heavily on entities such as sex, race, social context, smoking, alcohol use, premature birth, etc., as categorical or even dichotomous values. These values often form the basis of binary bifurcations of decision trees or can shift the "normal" reference range of a biomarker. Can concepts such as race or social status be measured or even understood as having homogeneous causes or effects? The use of a "continuous" epigenetic gauge of GxE may not only be more accurate, but also more clinically and biologically realistic.

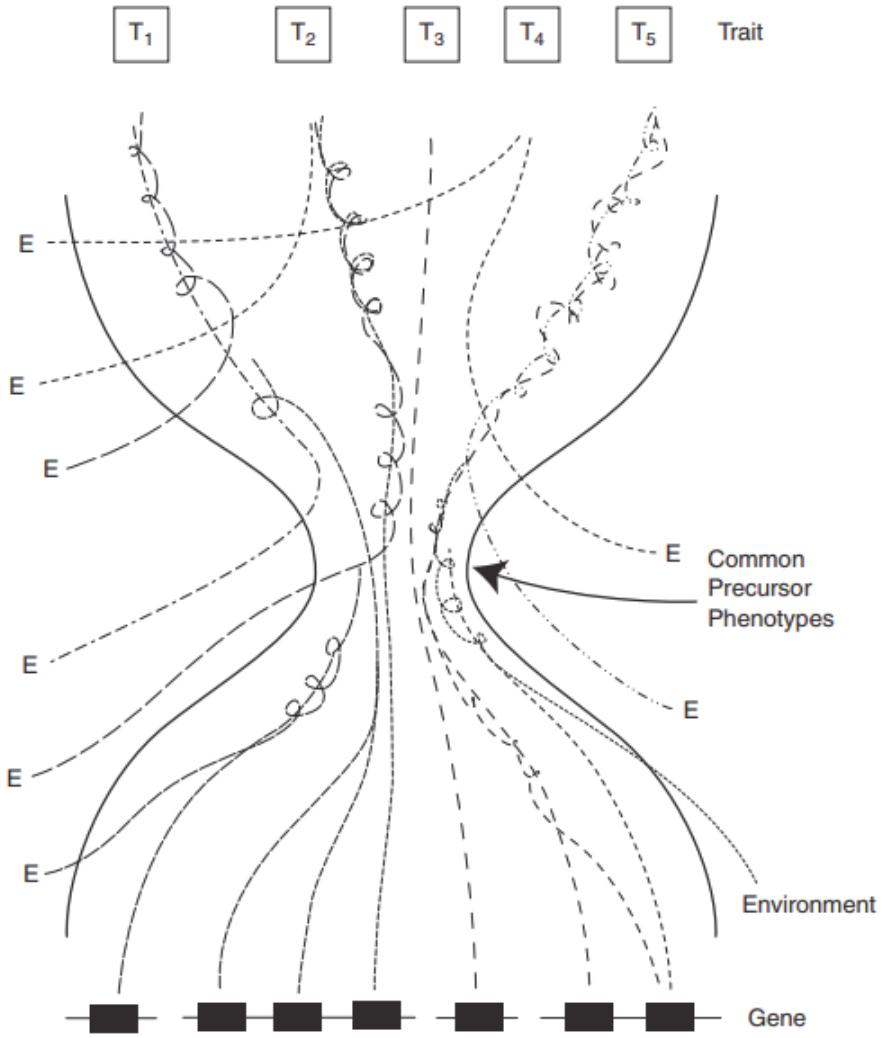


Figure 10: Hourglass metaphor for the genetic and non-genetic interactions in complex disease. E = environmental factors, T = phenotypic traits. This thesis explores the use of DNAm patterns as one of "Common Precursor Phenotypes" at the neck of the hourglass. Regarding the lines between Gene, Environment and Trait, note the lack of arrows and the twisting of lines. This represents the unknown directionality and complexity of these interactions (Buchanan *et al.*, 2006).

Dysregulated DNAm could represent a downstream, upstream or ripple-life effect of a biological cascade set off by a stressful insult. On a genome wide scale, it can give a live profile that can be matched to catalogued profiles with similar insults and comorbidities. The power of this profile or epigenetic phenotype is its fluidity: it is unrestrained by bounds of normal versus abnormal. It is simply a map that can be overlapped with other trait maps. It does not require an already dysfunctional biological pathway to "create" an abnormal paucity or accumulation of transcripts, proteins, lipids, sugars, etc., in order to provide actionable information. In this way,

transcriptomic, proteomic and metabolomic markers may be considered more “downstream” and thus later indicators of cellular dysfunction than epigenetic phenotypes.

Clinicians demarcate transition from health to disease using diagnostic tests whose sensitivity and specificity heavily depends on what reference range is used. In contrast, epigenetic profiles can be mapped together with risk factors, clinical findings and/or biomarkers to describe the patient’s unique propensity for disease. Since there is no reliance on cellular dysfunction, this has the potential to detect the disease precursor state. This widening of the effective preventative window is critical to precision medicine. This is because as this window closes, the palliative window opens. By changing the clock of disease, clinicians may finally have the chance to offer rescue rather than band-aid solutions.

Perhaps we currently miss the mark for disease prediction from lack of accounting for all the forces that impact health trajectories. Though it is a daunting task, it may be the key to better understanding the mechanical underpinnings of disease and health development. In the context of epigenetic processes within the cell, we could view these net forces as altering the propensity towards certain cell fates. Lappalainen and Greally propose a refinement of Waddington’s epigenetic landscape metaphor to suggest that reprogramming could “deepen” a furrow in that landscape, thereby raising the likelihood of cells entering that phenotypic channel as seen in Figure 11 (Lappalainen & Greally, 2017). This results in a shift in the propensity of a given phenotype to appear. In this work, we use patterns in DNAm across the genome as a proxy of cellular poise resulting from net genetic and non-genetic forces. We ultimately seek patterns that identify furrows deepened by MSP that lead to increased propensity to poorer health outcomes.

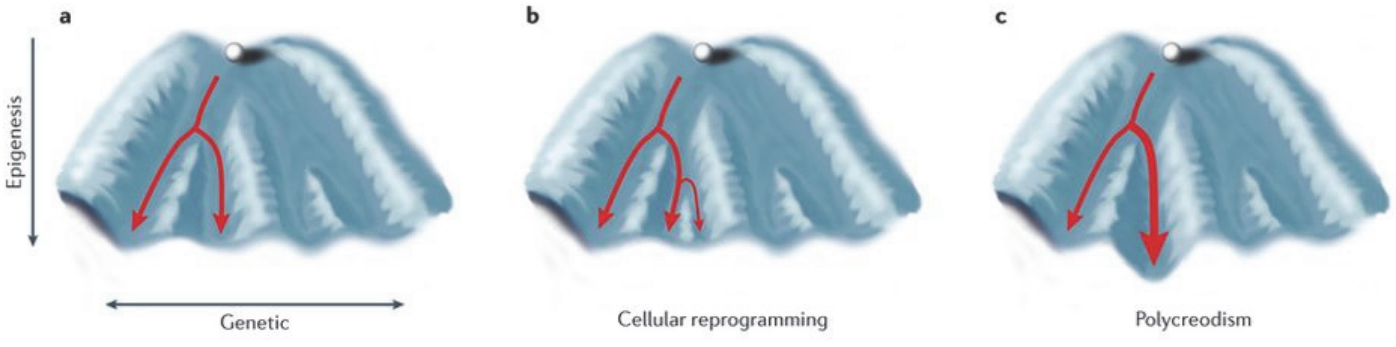


Figure 11: Net genetic and non-genetic forces alter the epigenetic landscape and cellular poise. In a), the cell represented by the ball has equal propensity to follow one of two channels leading to two different fates. Both b and c illustrate two different possible effects of reprogramming. In b), the reprogramming results in a bifurcation that results in the emergence of a subpopulation of cells within the same lineage but with minor phenotypic differences. This leads to mosaicism within that lineage, thus expanding the phenotypic variability of that cell type. However, in c), the reprogramming deepens one of the channels, thus increasing the propensity of that phenotype appearing. In other words, the reprogramming changes the relative proportions of the two lineages. Both b) and c) illustrate two models by which reprogramming could have phenotypic consequences. Image from Lappalainen & Greally (2017).

1.3 Thesis scope

1.3.1 Maternal smoking in pregnancy as a epigenetic model of common complex disease

Several groups have investigated maternal smoking during pregnancy (MSP) as an important human model to connect the DOHaD hypothesis to the etiology of CCDs (Agrawal *et al.*, 2010; Pickett, Wood, Adamson, D'Souza, & Wakschlag, 2008a; Suter, M. A., Anders, & Aagaard, 2013; Suter, M., Abramovici, & Aagaard-Tillery, 2010). Tobacco is the most prevalent substance used during the periconceptual and gestational periods in North America and Europe (Cnattingius, 2004; Cook, J. L. *et al.*, 2017; Rodriguez & Smith, 2019). It is associated with adverse fetal and infant outcomes, including poor fetal growth, prematurity, ischemia-hypoxia, sudden infant death syndrome, respiratory disease in early and later life, cardiovascular morbidity, increased adiposity, multiple cognitive and behavioural morbidities and delinquency (Cnattingius, 2004; Forray & Foster, 2015; Leybovitz-Haleluya *et al.*, 2018; Li, L. *et al.*, 2016; Wakschlag, Pickett, Cook Jr, Benowitz, & Leventhal, 2002). It is widely believed that tobacco exposure dysregulates fetal developmental programming through epigenetic modifications. In this context, DNAm is among if not the most intensely interrogated potential molecular mediator linking MSP with adverse health outcomes over the past decade (Choukrallah *et al.*, 2018; Knopik, Marceau, Bidwell, & Rolan, 2019).

Despite this fervent interest, there is yet not a single clinically useful epigenetic marker for MSP. This failure is juxtaposed with the success and rapid advancement of epigenetic markers and therapies in cancers such as in leukemia, lymphoma, myeloma (Kelly & Issa, 2017) and potentially some solid tumours (Linnekamp *et al.*, 2017) within a similar period (Issa, 2007).

We review briefly various streams of research that support the study of MSP-sensitive DNAm changes in context of the DOHaD hypothesis and child development. First, it is biologically reasonable for MSP to affect fetal DNAm. Cigarette smoke contains a large number of chemicals, such as carcinogens, nicotine, and carbon monoxide that easily pass the placenta from the mother to fetus (Jauniaux & Burton, 2007). These chemicals can modify DNAm through various mechanisms, such as by causing DNA damage, inducing fetal hypoxia, altering DNA-binding factors or directly disrupting methylation machinery through altered substrate or cofactor availability (Lee, K. W. & Pausova, 2013; Toledo-Rodriguez *et al.*, 2010). There exists a high overlap between MSP-associated differential DNAm in children and that with current smoking in mothers and other adults in various studies (Knopik *et al.*, 2019; Suter, M. *et al.*, 2010; Suter, M. *et al.*, 2011). This has led scientists to question whether MSP-related DNAm differences in children actually reflect epigenetic inheritance rather than the direct effect of intra-uterine MSP exposure. To investigate this question, Joubert *et al.* (2014) used a large pregnancy cohort, the Norwegian Mother and Child Cohort Study (MoBa), to compare the effects on newborn DNAm between paternal smoking, grandmaternal smoking (i.e. mother's exposure to MSP) and maternal smoking (categorized by four classes: never smoker, stopped before pregnancy, stopped before 18 weeks gestation and smoked through 18 weeks gestation) (Joubert *et al.*, 2014). This comparison is relevant as the ovum and sperm that created the newborn are exposed to smoking in the grandmother or father, respectively. As well, DNAm in sperm is likely affected by smoking (Jenkins *et al.*, 2017). Joubert and colleagues found that maternal smoking past 18 weeks gestation was the only significant association to newborn methylation compared to the never smoker group. The authors interpreted this to indicate that DNAm differences in the children are due to direct intra-uterine MSP effects of sustained smoking rather than being inherited. A later study observed a similar finding in another large and similarly designed observation cohort, the Avon Longitudinal Study of Parents and Children (ALSPAC) cohort (Richmond *et al.*, 2015).

Second, MSP has the potential to induce DNAm changes that are present in all somatic cells. This is because early embryogenesis is a highly sensitive and plastic period of development. At this formative stage, stem cells undergo global demethylation and subsequent reestablishment of DNAm patterns (Smith, Zachary D. *et al.*, 2012). These newly established patterns then are propagated from the stem cells to all subsequent somatic cell lineages. For example, MSP is related to differential DNAm in fetal lung (Chhabra *et al.*, 2014), fetal brain (Chatterton *et al.*, 2017) and placenta (Maccani, Koestler, Houseman, Marsit, & Kelsey, 2013; Smith, Taylor F. *et al.*, 2012), as well as in blood samples from newborns (Joubert *et al.*, 2016; Markunas *et al.*,

2014; Miyake *et al.*, 2018) and older children (Lee, K. W. *et al.*, 2014; Richmond *et al.*, 2015). This would be in keeping with the multi-organ effects of MSP on child physical and mental development. As well, the top molecular hits from MSP studies in human and animal studies implicates a wide range of developmental and growth biological processes. These include gene targets such as aryl-hydrocarbon receptors repressor (*AHRR* – involved in cell growth and differentiation regulation), growth factor independent 1 transcriptional repressor (*GFI1* – silences gene promoters in diverse tissues but especially in the hematopoietic system), fat mass and obesity-associated (*FTO* – physiologic function unclear but has strong association with body mass index, obesity risk, and type 2 diabetes across various ethnicities,) and cytochrome P450 family 1 member A1 (*CYP1A1* – involved in drug detoxification and lipid family biosynthesis) (Joubert *et al.*, 2012; Joubert *et al.*, 2016; Lee, K. W. *et al.*, 2014; Markunas *et al.*, 2014; Richmond *et al.*, 2015). Together, MSP-related DNAm likely has global effects on fetal development, rendering it an ideal candidate to mediate the early biological programming of multi-system traits. Thus, MSP is a well-suited model for the study of DOHaD and the etiology of complex, chronic and multi-system diseases.

Third, MSP-related DNAm changes may persist throughout the life course. DNA methyltransferases are enzymes that can copy DNAm from parent to daughter cells during cell division and are responsible for maintaining DNAm patterns in post-differentiated cells (Lee, K. W. & Pausova, 2013; Petronis, 2010). Several independent cohorts demonstrate the stability of MSP-related differential DNAm from early and late childhood (Joubert *et al.*, 2016; Lee, K. W. *et al.*, 2014; Richmond *et al.*, 2015) and early adult hood and midlife (Tehranifar *et al.*, 2018; Wiklund *et al.*, 2018) . This feature is a particular advantage in DOHaD research compared to biomolecules like RNA. For example, the gene expression changes associated with adult smoking revert by >50% one year after quitting. Ten years after quitting, the reversion rate is > 85%. In contrast, the reversion rate after one year of quitting ranges from 17 to 33% and differences remain detectable even 40 years after smoking cessation (Tsai, Spector, & Bell, 2012). The relative stability of DNAm compared to other biomolecules may speak to its role in complex chromatin regulating mechanisms that are less easily dismantled. This DNAm "memory" may also lend greater sensitivity of the nucleus to mobilize should a subsequent exposure event occur. While the answer remains a mystery, the persistence of specific DNAm changes renders it a good candidate to inform if not mediate the effects of early-life exposures to later life.

Fourth and as alluded to above, MSP-related DNAm changes can be sensitive to exposure changes. While it appears largely stable, DNAm at certain sites loci appears reversible. For example, using the same cohort as the discovery cohort in this study, Richmond *et al.* found that differential DNAm related to MSP persisted from birth to age 7 and 17, (*AHRR, MYO1G, CYP1A1, CNTNAP2*) while others appeared "reversible", (*GFI1, KLF13, ATP9A*). Similarly, Miyake and colleagues found differential methylation sites between children exposed to MSP that was sustained throughout pregnancy versus smoking cessation early in pregnancy (Miyake *et al.*, 2018). The study by Joubert and colleagues using the MoBa cohort could be similarly interpreted (Joubert *et al.*, 2014). One interpretation of these two well-designed studies is that sustained versus sustained smoking have distinct epigenetic effects. Alternatively, it may suggest that smoking cessation may reverse the DNAm response to early gestational MSP exposure. This would mirror adult studies showing reversibility of DNAm changes after smoking cessation (Tsai *et al.*, 2012). Whether these studies imply environmental sensitivity or reversibility of DNAm, the apparent plasticity of DNAm to MSP exposure makes it a stronger candidate not only as a biomarker but also potentially as a therapeutic target.

MSP is the most common toxin exposure in childhood in both the developed and developing countries (Rodriguez & Smith, 2019). Its effects are linked to lifelong and broad consequences for the child. And regrettably, the exposure is completely preventable. The responsiveness yet stability of DNAm changes, as well as its biologic mechanism of action, render DNAm a strong candidate as a molecular marker and mediator in the DOHaD model. As well, the relative low cost and technical robustness of DNAm microarrays compared to other biomarkers makes it feasible for human study in large cohorts, a critical bottle-neck in the study of CCD. Thus, the study of DNAm patterns underlying MSP-related disease may enhance our understanding of common molecular pathways in the DOHaD context and bring us closer to patient-specific diagnosis and management of complex diseases affecting all ages across the globe.

1.3.2 Current evidence linking maternal smoking and offspring complex disease to DNA methylation

Besides cancer, metabolic syndrome[1] and neuropsychiatric outcomes have received the most research attention among CCDs related to MSP. This is likely due to their rapidly rising burden

[1] a constellation of traits including glucose intolerance, abnormal cholesterol and lipid metabolism, hypertension, and overweight

both in terms of quality of life and societal costs (Biederman, Monuteaux, Faraone, & Mick, 2009; Li, L. *et al.*, 2016; Taal *et al.*, 2013; Wiklund *et al.*, 2018). Metabolic syndrome is a canonical example of the DOHaD paradigm. As discussed in Section Developmental origins of health and disease hypothesis and common disease – a convergence of multiple dimensions, adverse pregnancy conditions relate to poor fetal growth, (as evidenced by low birth weight,) a likely precursor to abnormal childhood fat accrual and adulthood metabolic syndrome traits (Drake & Walker, 2004; Roseboom, Tessa J. *et al.*, 2001; Suter, M. A. *et al.*, 2013). Using a modern cohort of Dutch children with 450K data, Küpers and colleagues demonstrated that GFI1 hypermethylation in cord blood mediated the relation between MSP and low birth weight in a meta-analysis of three independent European descent cohorts, Groningen Expert Center for Kids with Obesity (GECKO) , ALSPAC (UK) and Generation R (GenR) (Kupers *et al.*, 2015). They showed that differential methylation at three loci at the *GF1* gene accounted for 12-19% of the 202 g lower birth weight seen in MSP exposed infants. Murphy *et al.* also examined birth weight, but in a multi-ethnic birth cohort in the United States (Murphy *et al.*, 2012). This study specifically examined DNAm using pyrosequencing at two imprinted genes, Insulin-like Growth Factor 2 (*IGF2*) and *H19*. They found that IGF2 differential methylation accounted for 21% of the proportion low birth weight in male infants. Another prospective pregnancy study found methylation at cg25685359 in cord blood was positively associated with MSP and negatively associated with birth weight. This locus is associated with the miRNA let-7b host gene (LET7BHG) which is implicated in adipocyte differentiation and insulin signaling in both animal and human models. In fact, a study in pre-menopausal women showed an 8-fold decrease in blood LET7BHG miRNA levels after a 12 month intervention to reduce dietary glycemic load, the only marker in the study to show such a dramatic change (McCann *et al.*, 2013). Recently, researchers using 450K data from adolescents in the Raine Study found that two out of 23 CpG sites significantly associated with MSP were also linked to cardiometabolic measures in adolescence (Rauschert *et al.*, 2019). These two sites were found in the FTO and CYP1A1 regions. As well, this group did not find that other smoking exposures, (paternal smoking during pregnancy, childhood exposure to second hand smoke (SHS) or personal smoking of the adolescent,) affected DNAm directly nor altered the relation between MSP and methylation. This is consistent with the findings by Richmond and colleagues who also found no effect of paternal smoking during pregnancy on the relation between MSP and newborn DNAm in ALSPAC (Lee, K. W. *et al.*, 2014; Richmond *et al.*, 2015).

To support the role of intra-uterine MSP effects on birth weight rather than other maternal factors on birth weight, researchers use the placenta as a proxy of the fetal experience of

environmental exposures (Maccani *et al.*, 2013). A recent meta-analysis of seven birth cohorts in the US, Europe and Australia using 450K data from placental tissue found four CpG loci that related to MSP were also associated with birth weight z-scores (Everson *et al.*, 2019). Three of these four sites are located near known birth weight related SNPs, (*LEKR1, WBP1L* and *EDC3*).

There are also a number of studies demonstrating MSP-related DNAm differences to neurologic development (Chatterton *et al.*, 2017; Toledo-Rodriguez *et al.*, 2010). Because of the obvious challenges in sampling brain tissue, human studies are mostly limited to using blood DNAm as a proxy. Post-mortem samples provide guidance on the use of such surrogates, with studies showing a relative concordance of DNAm between peripheral blood and brain tissues at specific CpG sites (Edgar, Jones, Meaney, Turecki, & Kobor, 2017; Hannon, Lunnon, Schalkwyk, & Mill, 2015). As well, research benefits from several successful animal models. For example, a mouse model of ADHD and prenatal nicotine exposure found that global DNA hypomethylation in striatal and frontal cortical cells was related to altered cortico-striatal neurotransmitter-related signaling. This clinically correlated with enhanced nicotine preference and ADHD-like psychopathology (Buck *et al.*, 2019). Moreover, this effect persisted in the "grandchildren" (i.e. F2 generation). This and similar models have expanded our understanding of multi-generational, molecular and pharmacologic aspect of neuropsychiatric pathology in humans (Petronis, 2010). Using DNAm in blood as a surrogate for that in brain, a recent expansive meta-analysis encompassing 2821 human subjects used genetic instrumental variable (IV) analysis to infer a causal relation between MSP-related differential DNAm and psychiatric morbidity in later life (Wiklund *et al.*, 2018).

These groups and others have used various study designs and statistical techniques to strengthen causal inferences supporting the mediating role of DNAm in these diseases. To date, EWAS has identified literally hundreds of statistically significant differentially methylated sites or regions in various tissues related to MSP (Knopik *et al.*, 2019). Most studies focus on gene-centric findings, such as changes located at or around SNPs like *AHRR* and *CYP1A1*, two of the most consistently identified hits relating to offspring blood DNAm and MSP (Lee, K. W. *et al.*, 2014). Despite prospective and longitudinal cohorts, intense pooling and collaboration of international resources and samples, and a cost-effective and technically robust means of assessing DNAm across multiple populations and sample types (Joehanes *et al.*, 2016), there is yet no the leap to clinical translation. The remainder of this chapter discusses concepts and examples borrowed from various research realms that specifically target this field towards clinical applications.

1.4 Mapping individuals to the risk context of MSP

Today, modern cohorts amass immense multi-omic data. We are better able to profile human disease than ever before. This enables more fine-grained study of the manifold biologic and environmental interactions that establish an individual's poise between healthy and diseased states. This idea of individual-specific vulnerability to health risks borrows from evolutionary-developmental biology theory (West-Eberhard, M. J., 2003). This concept has fostered several theoretical frameworks since the mid-1990s. Among these, "biologic sensitivity to context" (Boyce & Ellis, 2005) and "differential susceptibility" (Belsky & Pluess, 2009) are prominent. These concepts share roots with the DOHaD hypothesis. In our study, we employ this concept to visualize a spectrum of individuals ranging from "typical" i.e. manifested outcomes are as expected given risk profile to "atypical" i.e. demonstrate poor outcomes despite low risk or good outcomes given high risk factors (Boyce & Ellis, 2005). In Figure 12, the outcome is a proxy of fetal development, birth weight and MSP is the risk. By comparing the four categories at the bottom, we can better disentangle pathways that are 1) MSP-related processes, 2) processes related to birth weight but not due to MSP or 3) processes related to MSP but not the outcome. For example, we could compare the typical-vulnerable and atypical-vulnerable cases. Differential DNAm that overlaps in both these groups would be unrelated to smoking and more likely related to processes that cause low birth weight. By using such comparisons akin to counterfactual analysis, we may better characterize the fetuses and/or children that would succumb to MSP adverse effects and hopefully target the molecular underpinnings underlying this vulnerability.

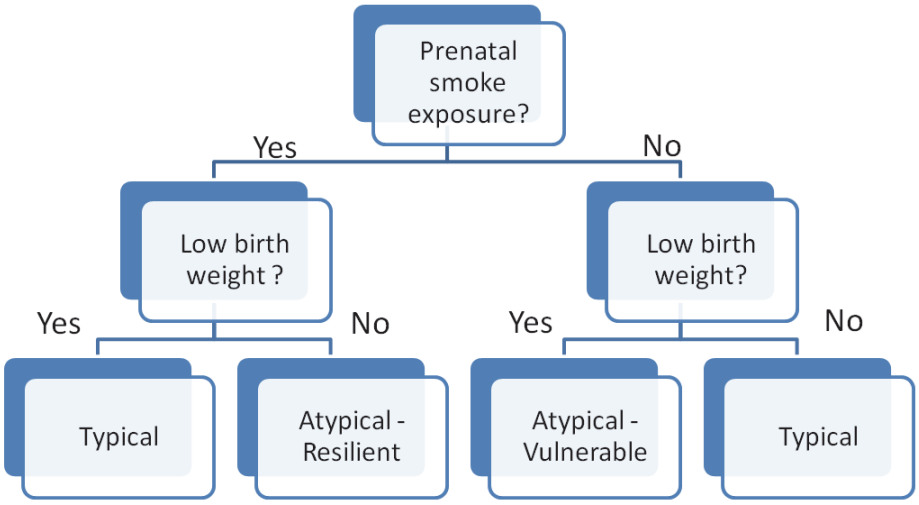


Figure 12: Typical and atypical risk-phenotype association. This figure uses the example of the effect of MSP on fetal growth. Though this depiction suggests a linear spectrum of between typical and atypical, individual vulnerability exists on a multi-dimensional grid that represents the joint contribution of factors that may be protective or harmful depending on context.

However, these four typical-atypical categories may underrepresent other smoke exposures and genetic, environmental and familial factors that are relevant to MSP-related fetal programming. We speculate that including these additional factors may provide different clinical insight and perspective. By moving from a two-dimensional to a multi-dimensional characterization of each subject's risk factors, we shift from a linear spectrum to a potentially richer topography of vulnerability. The benefit of a multidimensional view of the impact of MSP is backed by the clinical relevance of GxE MSP effects on child outcomes (Agrawal *et al.*, 2010), but also the known distinct effects of genetic, non-genetic and GxE effects on DNAm in cord blood (Czamara *et al.*, 2019). While using greater than two variables is less conducive to the "clean" dissection of counterfactual comparisons, it offers a view of the interactions that may confer relative vulnerability or resilience *depending on context* as would occur in reality. Put another way, we propose to profile MSP-vulnerability using a composite of multiple variables.

There is an additional and related potential benefit to the use of multiple variables to form an MSP composite measure. Our overall objective is to overlap a subject's MSP-related risk with his/her epigenetic 450K data. The "curse" of high dimensional data is classically known as the phenomenon where data from thousands of variables (*p)* are collected on a relatively small sample of subjects (*n)*, (a situation often symbolized as $p >> n$.) Because subjects are few, an outlier subject can have a powerful (and likely non-generalizable) influence on model selection and the inferences drawn. We further aggravate outlier influence if we misclassify MSP-related

risk. Outliers are defined as data that lies distant to other observations. They can result from true data variance, measurement error, or recording error. As we add multiple "layers" to a composite measure, we potentially attenuate the influence of subjects that may be considered "outlying" based on one measure, but when combined with other measures, can be better placed in context of a landscape of vulnerability. For instance, it is fortunate that in a population only a small number of children are exposed to high and continuous levels of smoking as a fetus. If we used maternal smoking in pregnancy alone, this small group of children would powerfully influence the data model built using the high dimensional epigenetic data. There are numerous means of dealing with outliers, including removing them, removing and then re-introducing them in sensitivity analysis, data transformation, changing models (e.g. switching from a linear to quadratic model assuming the outlier is a valid observation) and various methods of "down-weighting" the observation so that it contributes less to the overall model (Debruyne, Höppner, Serneels, & Verdonck, 2017). The discovery and management of outliers is a field unto itself. However, the most important message in outlier literature is that one should be very curious as to the "why" a data point is labelled an outlier in the first place. A commonly quoted statement to remind us of this lesson is below:

"So unexpected was the hole that for several years computers analyzing ozone data had systematically thrown out the readings that should have pointed to its growth."

— New Scientist 31st March 1988

There exists an implicit danger in assuming any not obviously "wrong" observation is an outlier. One could systematically remove potentially valuable information. When considering the issue of outliers, we posit that the use of a composite will improve our capture of signals present in high dimensional by both avoiding data loss and attenuating the polarizing influence of subjects who cluster at extremes of certain variables but not necessarily all variables. By viewing subjects using multi-view versus single-view perspectives, we "break up" these clusters and reduce the chances that subjects can unduly influence model behaviour based on one feature alone.

We also consider the increased error due to mis-reporting of smoking status. Systematic underreporting of smoking in pregnant women is a known source of error (Brand *et al.*, 2019; Dietz *et al.*, 2011; Shipton *et al.*, 2009; Valeri *et al.*, 2017). Mothers may either falsely deny or underestimate daily usage, likely due to social stigma and personal guilt regarding substance

use in pregnancy. In addition to the random error surrounding any variable measurement, this error aggravates misclassification and introduces bias that may be a critical cause of misleading and inconsistent findings between cohorts, (for a specific example in MSP research and methylation, see results of (Kupers *et al.*, 2015) versus (Valeri *et al.*, 2017). Put together, we believe mapping individuals to their risk context may improve overall accuracy in identifying DNAm differences likely to have a biological basis of MSP vulnerability by attenuating reliance on (and therefore the error influence of) any given variable.

We also consider another source of error: selection bias. Pregnant smokers are more likely to be poor, young, unmarried and have health risks such as poor nutrition, and psychiatric disorders such as depression, anxiety, substance use and attention-deficit hyperactivity disorder (for overviews, see Rodriguez & Smith, 2019 and Rodriguez-Bernal *et al.*, 2010.) Thus, there is unequal distribution of potential confounders between risk categories. For instance, Fang *et al.* found no difference between MSP exposed versus non exposed infants in terms of developmental outcome or birth weight until selection bias, (estimated using a propensity score,) was included in the models (Fang *et al.*, 2010). In another example, Fertig analysed data from multiple cohorts to investigate the nearly doubled risk of low birth weight in the year 2000 compared to 1958. Her analysis estimates that as much as half the association is due to selection bias (Fertig, 2010). Clearly, selection bias in MSP research is a critical source of unreliable estimates that can have ramifications not only on biological research but also on public health decisions. Numerous methods exist to mathematically correct or account for selection bias such as IV analysis, Heckman correction, fixed effects and propensity score matching. However, many of these methods require confounding covariates to abide by statistical assumptions such as linearity, normality, or additive effects. As well, the quality of the correction is affected by the confounder data available in the dataset and if the method can reliably accommodate the number and types of data (e.g. continuous, ordinal or nominal.) Last, certain techniques are only able to study selection bias in specific contexts. For example, the fixed effects approach used in Abrevaya (2006) requires measured variation in maternal smoking over subsequent pregnancies. In general, the inherent nature of techniques to counter selection bias is to account for what is unmeasured and/or unknown. Instead, we attempt to better describe what we do know in hopes that multi-source data (e.g. over time different time points, origins, or methods of data collection) are less likely prone to be all similarly biased and thus offer a more accurate subject characterization.

In addition, we believe that “mapping” of vulnerability may be particularly relevant in a MSP model of DOHaD. For instance, approximately 20% of low birth weight is attributable to MSP in European ancestry population studies (Cnattingius, 2004; Kramer, 1987). This rate is lower in Black populations. This means in a given population that many more infants have average or even high birth weight despite high reported MSP In fact, it is estimated that about half of the estimated effect size of MSP on birth weight is actually due to unobserved variables that are related to MSP and are not MSP itself (Abrevaya, 2006; Fertig, 2010). Also interesting is that while the level of toxins like tar and nicotine inhaled from cigarettes have decreased since the 1950’s, the association between MSP and low birth weight reported in more recent cohorts is actually stronger than in older cohorts (Fertig, 2010; Kline, Stein, & Hutzler, 1987; Wehby *et al.*, 2011). Whether this is true or due to artifacts like selection bias is unclear (Fertig, 2010). However, what is clear is that no simple relation exists between MSP and infant outcomes. Thus, it is even more imperative to avoid oversimplifying MSP related risk if we seek its underlying pathogenic mechanisms.

Last, mapping provides a natural response to the “mismatch” hypothesis posed by DOHaD and related life course models. Mapping avoids assuming which prenatal to postnatal influences are considered “matched” or “mis-matched” in terms of predicting long-term health. We aim to generate a composite profile that provides a comprehensive view of various MSP-related influences stemming from pre-/post-natal sources. The influences traverse gestational (e.g. maternal smoking in pregnancy, grandmaternal smoking, fetal growth, etc.) and postnatal (e.g. paternal smoking, household smoking, etc.) effects. Another advantage particularly relevant to epigenetic research is that is difficult in observational human studies to dissect apart genetic versus non-genetic, and similarly transgenerational versus intergenerational, influences given the action of shared genes and shared environments across generations (Horsthemke, Bernhard, 2018). For instance, maternal genetic predilection for tobacco addiction and tobacco toxin xenobiotic metabolism could have both genetic and non-genetic influences on the fetus. As well, secondary epimutations from a certain insult may mimic transgenerational epigenetic inheritance (Horsthemke, B., 2006). However, such epimutations actually arise from a DNA sequence change in a neighbouring gene area that affects the methylation and transcription of the gene of interest. While the epimutation occurs in the F0 generation and may appear to persist in the F3 generation without further exposure to the certain insult, it is due to genetic rather than epigenetic transmission. A composite profile avoids these research design conundrums by avoiding assumptions of how one effect interacts with others and how that impacts health outcome.

To this end, this thesis will also explore transforming the 2-D spectrum expressed in Figure 12 into a multidimensional topography by triangulating genetic and non-genetic MSP factors with impact on fetal growth.

1.5 Mapping individual epigenetic data to genome wide patterns

We also explore a context-based view of epigenetic profiles. We consider that each individual expresses a mixture of patterns of DNAm across the genome, each pattern arising from the net effects of genetic and non-genetic influences. This study of patterns provides at least three important biological contexts and several statistical advantages.

1.5.1 Biological advantages

First, we posit that DNAm patterns may enhance biological context by giving clues regarding chromatin function. Numerous studies have uncovered patterns of gene and histone based epigenetic modifications that modulate chromatin structure that translate into altered gene regulation (Fortin & Hansen, 2015; Huang, Marco, Pinello, & Yuan, 2015; Wu, Y. *et al.*, 2019; Zhu *et al.*, 2016). The driving hypothesis behind this interest is that co-varying sites of epigenetic variance are functionally linked by their physical 3-D proximity to other epigenetic marks (e.g. histone modifications) and localization to chromatin/nuclear structures (Zhang, L. *et al.*, 2017). On a molecular level, the fact that patterns emerge is unsurprising. At its most basic level, active chromatin must accommodate transcription machinery and silent chromatin must wrap and twist into fibres and coils. Thus, epigenetic patterns could be thought of as a "connect-the-dots"drawing– but we need to understand the rules in order to visualize the chromatin structure underlying the dots.

As discussed in Section 1.2.1: Overview of epigenetics, we must find a way to reconcile relevant changes to chromatin function and structure with the 1-D DNAm information that is currently the most feasible source of epigenetic data that can be extracted on a population-scale as required in CCD research.. . However, the real-time functions of DNAm range from directing nuclear traffic to orchestrating the 3-D conformation of DNA strands (Feng *et al.*, 2006; Lay *et al.*, 2015; Liu, S. *et al.*, 2018; Price, M. E. *et al.*, 2013; Rao *et al.*, 2018; Raviram *et al.*, 2016; Schoft *et al.*, 2009; Tajbakhsh, 2011; Xu, C. & Corces, 2018; Zhang, L. *et al.*, 2017; Zhu *et al.*, 2016). This fascinating realm of research reminds us of how difficult it is to fully envision with linear 1-D

metrics alone the relation between methylation and transcriptional output. Whether using individual DNAm loci, gene-centric annotation or linear DNA sequence distances, it is clear we do not fully appreciate the what, where and how differential DNAm actually affects cell fate.

Currently, there are large-scale efforts to map chromatin state. This mapping strives to characterize the genome as regions of probable transcriptional activity that is more finely grained than merely open (or active) versus closed (or inactive) regions. These maps are genome wide and integrate information regarding histone variants, chromatin modifications, (e.g. post-translational methylation and/or acetylation of nucleosome proteins,) and/or data from chromatin immunoprecipitation assays for example. These “marks” on the chromatin help infer the *degree* of openness of that region of DNA to transcriptional machinery. Interestingly, non-coding regions harbour a vast number of key marks linked to differential chromatin activity. Moreover, data support that these chromatin features are a critical mechanism in exposure-mediated shifts in gene expression and overall cell function (Schvartzman, Thompson, & Finley, 2018).

In addition, the advancement of 3-D chromosome structure assays has enabled the generation of chromatin interactome maps. This intriguing body of research uses high-resolution analysis of 3-D chromatin structure to visualize DNA as fibres that are woven into a hierarchy of loops that are tethered by anchor points attached to various nuclear membrane structures, nuclear regulatory factors and other chromatin areas (Figure 13). The working paradigm of these chromatin interactome maps is that the relation between gene expression and chromatin context is due to the physical “tethering” of chromosome regions to elements such as promoters, enhancers and transcription machinery. It is believed that most dynamic gene expression regulation occurs within more transcriptional active regions called topologically associated domains (TADs) (Mishra & Hawkins, 2017) as seen in Figure 8 and Figure 13.

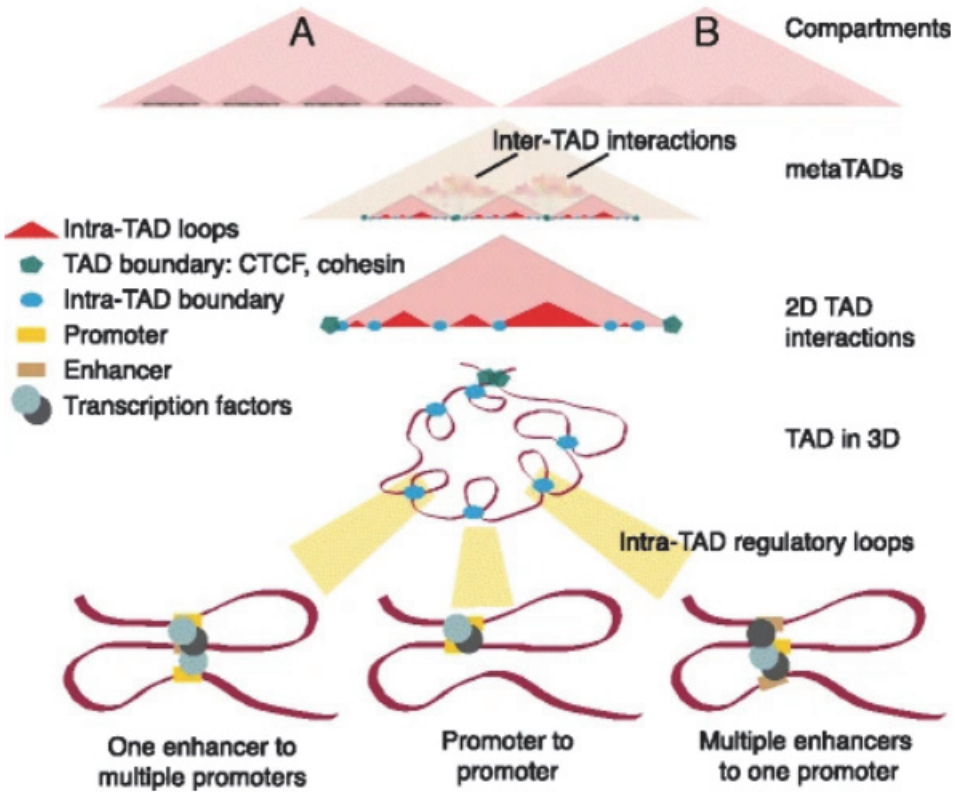


Figure 13: Schematic of hierarchical chromatin organization in both 2-D and 3-D views. Top two images show 2-D view of chromosomes based on TAD location. Bottom two images show 3-D view of chromatin focusing on within TAD regions. Bottom image shows examples of anchoring of chromatin loops to proximate DNA regions.

With this intricate weaving of chromatin fibres across various nuclear structures, it is no wonder that studies have shown less than 10% of putative target genes from chromatin interactome studies map to the "closest" regulatory region by linear genomic distance (Jung *et al.*, 2019; Wu, Y. *et al.*, 2020). As well, studies show that the context-specific role of DNAm on chromatin changes is what functionally effects gene expression (Collings & Anderson, 2017; Lay *et al.*, 2015). Such maps have mechanistically linked DNAm on sites from within a few kilobases of one another to as “far” as on different chromosomes.

The work of areas like chromatin interactome studies shows that epigenetic changes function in physically related clusters. This correlates well with clinical associations that show (with the exception of some cancer-related DNAm aberrations) concordant changes in groups of CpG sites are more likely to affect phenotype than single CpG site differences (Portela & Esteller, 2010). It is clear even from sequence based (as opposed to chromosome structure based)

analysis that differential methylation at one site may have indirect effects on proximal and distal genes. In their study comparing cord blood of children of smoking versus non-smoking mothers, Bauer *et al.* found that 93.2% of intragenic enhancers overlapping with DMRs targeted at least one gene outside the local host gene (Bauer *et al.*, 2016). Moreover, about a third of these did not even interact with the local gene. Mothers' blood samples demonstrated a similar finding.

Despite these findings, most EWAS to date continue to use statistical association to exposure or outcome to filter out relevant genomic sites or use sequence- or gene-based clustering. This may risk destroying the information embedded in the functional clustering of methylation at certain sites. The task of interpreting 450K DNAm data with our current state of knowledge may be likened to building a 450,000 piece jigsaw puzzle without knowing what picture to expect. Imposing assumptions on DNAm sites without understanding their role in the topography of epigenetic regulation of chromatin activity may be akin to discarding most pieces and/or forcing certain pieces together in the cases of filtering and clustering, respectively. Instead, we adopt a pattern finding approach to find DNAm sites that *en bloc* relate to vulnerability profiles. We hypothesize that these patterns may reflect a specific chromatin context that predisposes to a certain cell fate. Contrary to traditional studies, Bauer and colleagues used a cohort of mothers and children to perform a functionally rooted profiling involving RNA-seq, whole genome bisulphite sequencing and ChIP-seq (Bauer *et al.*, 2016). In this way, they characterized genome wide gene expression, DNAm and histone modifications in mothers and in children at birth, age 1 and 4 years. This group found that 82% of the MSP-related DMRs found at birth persisted at age 1 and 4 years. Interestingly, they observed that MSP exposed children had more transitions from repressed to active chromatin states than non-exposed children, and that the opposite was true for non-smoking versus smoking mothers. In addition, such transitions were more pronounced near MSP-related DMRs. This offers strong evidence of a link between MSP, differential DNA and chromatin dynamics (Bauer *et al.*, 2016).

Second, we posit that patterns will more effectively capture the net impact of protective and risk factors that simultaneously influence an individual at a given time. In the context of epigenetic processes within the cell, we could view these net forces on biological programming as altering the propensity towards certain cell fates. As seen in Figure 11, Lapplalainen and colleagues expanded Waddington's epigenetic landscape metaphor to suggest that reprogramming could "deepen" a furrow in that landscape, thereby raising the likelihood of cells entering that phenotypic channel (Lappalainen & Greally, 2017).

Using the lens of epigenome-wide poise within the DOHaD framework, we aim to capture an individual's cumulative MSP-related biological programming and adaptation to current conditions. As such, we view these patterns as the sum of active biological processes that together represent a degree of tissue-based poise predisposing to a given phenotype. This departs from traditional biomarker development that, in order to quantify a value of relative risk or dose-response effect, removes individual-context in exchange for individual-independent categories.

Third, we posit that DNAm patterns can provide a disease-based context of underlying molecular mechanisms. In genetics, pleiotropy is the phenomenon where a single gene contributes to more than one unrelated phenotypic trait. Data suggests this widely exists among CCDs. If an isolated genomic area can exhibit such pleiotropy, then it stands to reason that this may also be true of epigenetic differences. Understanding pleiotropic effects may be critical to uncovering the common biological mechanisms underlying and connecting the multiple traits in the constellation of a given complex disease entity (Vattikuti, Guo, & Chow, 2012; Yang, C., Li, Wang, Chung, & Zhao, 2015). For instance, medicine previously considered the traits of hypertension, central obesity, dyslipidemia and glucose intolerance as separate entities. The recognition that these traits are biologically linked in what is today called metabolic syndrome has changed the paradigm of medical management and research efforts. We speculate that a pattern of epigenetic differences may exhibit a more predictable relation to underlying pathogenic mechanisms that may be shared among traits. Thus, instead of trying to trace traits to their molecular common denominators, patterns may be an easier means of tracing molecular commonalities to their related constellation of traits.

1.5.2 Statistical advantages

Besides providing a different analytic lens, pattern finding has a number of statistical advantages. First, pattern finding across hundreds of subjects attenuates the risk of spurious results due to probes which are unequally affected by technical artifacts or batch effect. For example, most EWAS population studies to date used mixed cell type tissue. The problem arises when epigenetic forces are not equivalent for all probes in all cell types within a tissue. This leads to a very significant degree of variability simply due to cell type discrepancy that can and has been mistaken as disease specific markers (Jaffe & Irizarry, 2014). The patterns we propose to extract from blood represent the landscape of DNAm variability across a mixture of cell types of varying proportions. Blood cell populations vary physiologically with age throughout childhood. While various means of correcting for cell type heterogeneity have been proposed,

the main advantage of using patterns which involve numerous DNAm sites is that it relies less heavily on any single site that may or may not be affected by either physiologic or pathogenic processes at a given stage of a child's development. Replacing the single probe view with a context based view diffuses the chance error of concentrating on a problematic probe.

Second, error can be introduced by genotype-dependent methylation. This is a problem when a genetic variant may be associated to disease status – but also alters DNAm. This would lead to an association between DNA methylation levels and disease that is not necessarily causal. Genotype is known to affect methylation levels in both cis and even trans locations. For example, single nucleotide polymorphisms (SNPs) can directly or indirectly alter methylation (Zhi *et al.*, 2013; Zhou, D. *et al.*, 2015; Zhou, W., Laird, & Shen, 2017). In the former case, the actual site of methylation can be reduced by half or totally in the heterozygous or homozygous state, respectively. In the latter case, methylation can be altered indirectly via a change in the TF binding site which then alters the local level of DNA methylation (Gutierrez-Arcelus *et al.*, 2013) or histone modifications (Lappalainen & Greally, 2017). Taken together, effect of genotype variability also argues against using traditional gene-centric EWAS approaches as previously discussed. A global view of the methylation landscape reduces the likelihood of targeting genotype rather than environment related DNAm differences. Similarly, this may also lessen the variability seen between ethnicities and thus broaden the general applicability of findings.

Third, pattern finding seeks to distill information from thousands of sites into a number of meaningful features. This reduces the likelihood of type 1 errors that arise from performing multiple individual statistical tests (Anderson, Burnham, & Thompson, 2000; Vacha-Haase & Thompson, 2004). Type 1 errors refer to the likelihood of rejecting the null hypothesis erroneously, (e.g. stating there is a difference, relation or effect when there truly is none.) Multivariate techniques such as pattern finding conduct comparisons of variables simultaneously rather than through multiple separate tests.

Fourth, patterns may help confront issues due to low effect size. The study of epigenetic features linked to CCDs faces the difficult challenge of comparing phenotypes that are linked to small mean differences in between subjects, (typically in the 5% range as opposed in the 30% and above range for cancer phenotypes) (Bacalini *et al.*, 2015; Teschendorff, Andrew E. & Relton, 2018). In addition, few targets are expected to have detectable differential function. The study byBauer *et al.* illustrates this point where the paucity of differentially expressed genes

impelled the authors to instead target the downstream pathways targeted by DMRs. Using RNA-seq to compare genome wide gene expression in children of smoking versus non-smoking mothers, this group found very few genes passed the multiple testing threshold, (the fairly standard 10% false discovery rate (FDR) with the Benjamini and Hochberg (BH) correction used in EWAS.) In the specific context of DNAm, Meissner and colleagues recently reported in a study of 30 human cell and tissue types that only ~20% of CpGs are differentially methylated (Roadmap *et al.*, 2015). Of those, the majority are cell type specific differences and likely disease unrelated. With such small effect sizes and relatively large sources of data noise, numerous studies have attempted to optimize detection of differential methylation by grouping multiple sites together using various criteria, typically based on array-specific architecture, biological annotation or localization of CpG sites relative to one another (Teschendorff, Andrew E. & Relton, 2018). This follows the current paradigm in GWAS of complex diseases which suggest that a large portion of phenotypic variability can be accounted for by multiple common genetic variants with small effects (Vattikuti *et al.*, 2012). Pattern seeking follows this line of thought given that DNAm variability at a single site likely has a small effect on phenotype, but that clustering these small effects can enhance detection. For instance, Bacalini and colleagues (2015) proposed a multivariate method using 450K data that grouped sites by CpG density and proximity to a gene. They tested this method in a large meta-analysis and found that it could detect significant differences between samples in regions with very low DNAm variability between probes. They also found that these differences would be non-significant if tested individually using univariate methods (Bacalini *et al.*, 2015). To make things worse, more intrinsic DNAm variability at individual DNA sites is found in samples exposed to cigarette smoking (Jenkins *et al.*, 2017; Petronis, 2010; Vazsonyi & Belliston, 2006). Researchers speculate this is caused by increased heterogeneous and/or stochastic events cause by cigarette toxins (Petronis, 2010). While better understanding of such events would greatly improve analysis of epigenetic alterations, at this point it remains a black box. What is known is that such high variability further challenges statistical power to distinguish health status using a DNAm biomarker. We posit that intra-subject (in other words, subject specific) information can attenuate this problem. Intra-subject DNAm differences extracted by multivariate techniques can express deviation from a given pattern and is less dependent on inter-subject variability at a single probe. In other words, patterns may help better visualize how similar or dissimilar an individual is to a given genome wide DNAm configuration rather than comparing similarity between individuals.

Related to low-effect size at individual CpG sites is another possible source of bias related to the 450K chip design. Silva-Martínez and colleagues argue that the risk of probe density-driven false-negatives is increased by using differences in methylation levels between disease and healthy controls (Silva-Martínez, Zaina, & Lund, 2017). Using 13 case-control disease studies, this group found a near perfect linear relation between probe density and frequency of differentially methylated genes based on beta methylation differences between diseased and normal samples. Moreover, they demonstrate that accounting for probe density provides more pathobiologically relevant hits compared to using beta methylation differences alone. This claim was based on relatively greater gene function category enrichment, overlap with expression data (both in number and degree), and disease or tissue specificity. Using univariate techniques and multiple testing, it is to be expected by chance that more dense areas will be more likely selected. However, pattern based techniques avoid multiple testing and therefore are less susceptible to distribution biases.

In summary, we propose that mapping patterns of genome wide DNAm provide statistical advantages as well as facilitate biological interpretation (Figure 14). Epigenetic studies of complex diseases struggle with low effect size, high variability of DNAm at single sites and multiple testing combined with merely emerging knowledge of the effect of genotype, stochastic effects, technical artefacts and the true biological relations between differential DNAm and exposures and/or outcomes. We consider the specific gains in exploring a small number of pattern-based indicators of differential DNAm. It will reduce the number of hypothesis tests and may attenuate the effects of high inter-subject data variability. It may better model biological effects by offering a glimpse of genome-wide chromatin poise as well as offering a gene-agnostic view that minimizes the impact of genotype-related DNAm and/or disease risk influences. As such, it would refine signal precision. Put together, these points could substantially increase the power to detect differential methylation and the discovery of clinically useful biomarkers (Klein & Hebestreit, 2016).

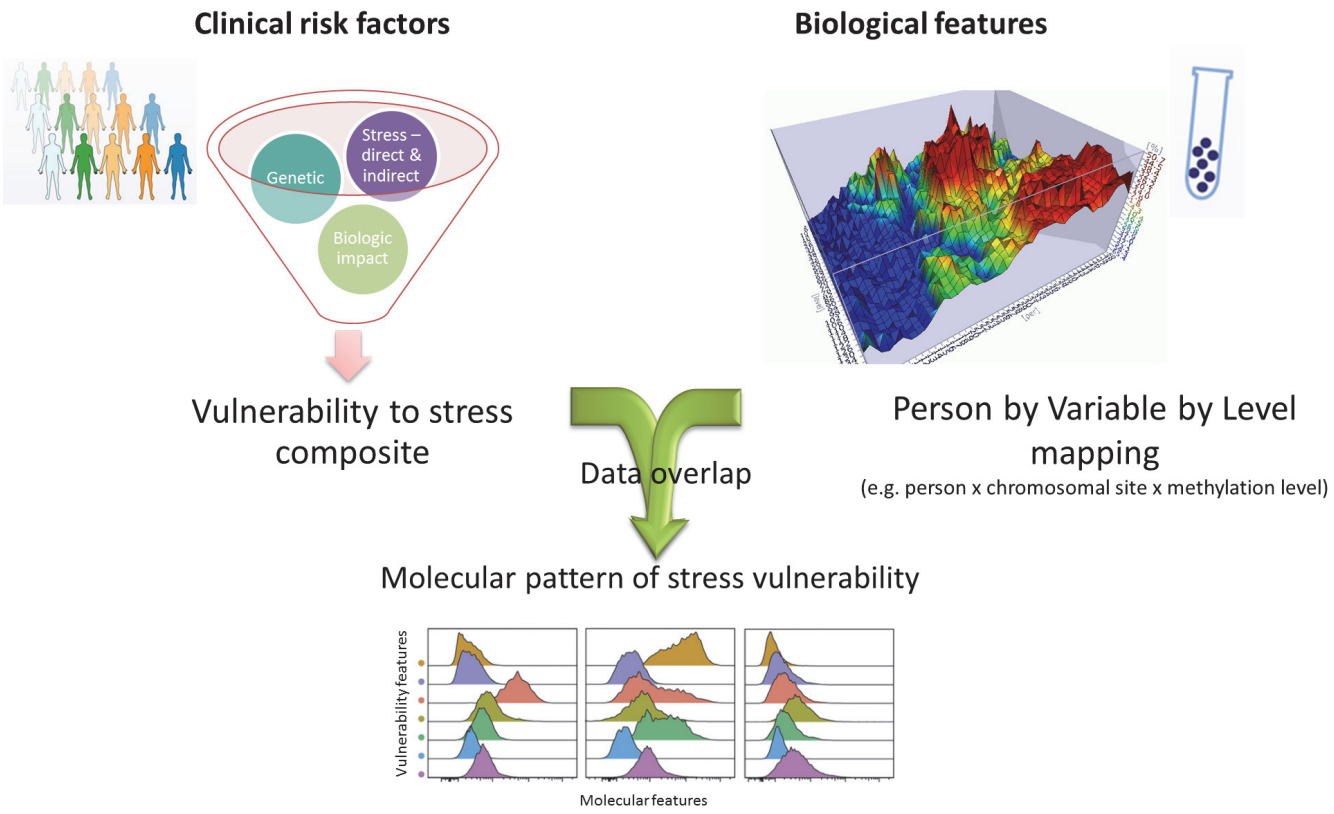


Figure 14: Mapping paradigm for subtle or multi-factorial diseases. This employs a precision medicine based characterization of individuals based on overlapping clinical and biological data to map vulnerability to CCD. Unlike Figure 3, the goal is to identify patterns of features related to disease vulnerability. Figure 3 describes a biomarker paradigm ideal for post overt disease onset while this figure describes positioning each individual on a multidimensional landscape of disease predisposition.

1.6 Thesis Outline

1.6.1 Rationale

Over the past decade, an explosion of studies has left little doubt that MSP leaves an indelible imprint on child health and the epigenome. Despite this fervent interest, there is yet not a single clinically useful epigenetic biomarker for this prevalent and potent exposure. Omics research, particularly in phenomics and exposomics, have enabled the development of rich data landscapes with improved accuracy and precision in positioning individuals in terms of their risks and outcomes. Coupled with molecular advances that enable high-throughput, genome wide interrogation of epigenetic modifications like DNAm, we now have an opportunity to better overlay multiple layers of information to generate a multidimensional rendering of human health.

This work integrates multiclass prospective data from questionnaires, public education repositories, clinical measures and DNA methylation microarrays. The proposed methodology aims to provide a context-based view of risk that is not only more accurate but attenuates both misclassification and data loss through avoidance of discretization. As well, it uses pattern finding methods to describe DNAm differences rather than traditional EWAS methods that focus on categorical comparisons at multiple genomic sites or user defined regions. In medicine, diagnostic accuracy improves when combining patient endorsed symptoms, clinician detected signs and ancillary investigations. While including this much information can often include "red herrings", the overall pattern of illness cannot be obtained without it. Moreover, CCDs likely involve multiple pathways and thus are likely best described as a combination of multiple factors as opposed to a few.

Thus, the primary motivation of this thesis is to test if and what differences in genome-wide DNAm patterns characterize fetal susceptibility to MSP. If MSP-related DNAm patterns exist, how do these differences relate to future health trajectories and regulation of chromosomal activity? As discussed in Section 1.2.3: DOHaD, DNAm and common complex disease with the hourglass analogy, one of the challenges of studying and clinically managing CCD is that it may be broad and heterogeneous in its risk factors and manifestations. Not all individuals with a complex disease have the same risk factors. Similarly, not all individuals with a complex disease will manifest all traits. Thus, we use continuous and "all-inclusive" ranges of risk and various multiple outcomes to best capture the two wide-ranging reservoirs of the complex disease hourglass.

1.6.2 Hypotheses

There is a landscape of genetic and non-genetic factors that shape health trajectories. This thesis seeks to explore this landscape with the following hypotheses:

1. Individuals can be topographically positioned relative to other subjects on a multi-dimensional map of exposomic factors. This "vulnerability map" encompasses both familial and environmental factors in the form of family history and maternal report of MSP, as well as an indication of the degree of impact these factors had on the individual. In this way, we hypothesize this map can point to individuals relatively more vulnerable or more resilient to MSP-related risks in terms of their fetal development.

2. We can use the positioning on this map as “bait” to find DNAm patterns that are related to varying degrees of MSP vulnerability. Based on the observation that MSP has persistent and broad reaching effects on child health, we posit the following inter-related corollaries regarding the relevant DNAm patterns:

a) They will be present in predominantly impact active chromatin domains – repressed chromatin domains are less open to cellular and extracellular signaling and thus less likely to participate in environment-sensitive changes.

b) They will have pervasive system effects: the effect of maternal smoking in pregnancy is known to be related to multi-system dysregulation in the exposed child. Thus, we expect the area(s) of chromatin to plausibly dysregulate function across multiple pathways from the molecular to tissue level.

i. Impact areas annotated to diverse tissue and/or biological functions: we expect the area(s) of chromatin to plausibly dysregulate multiple pathways, which may already be known to be implicated in functions such as growth, proliferation, and inflammation. Most importantly, these disrupted areas will form reproducible patterns that will affect multiple organs that impact physical, cognitive and psychological growth.

ii. We expect the pervasiveness of these effects to be across both sexes, acknowledging the likely sex based susceptibility to maternal smoking based on various epidemiologic (Suzuki, K. *et al.*, 2011; Zaren, Lindmark, & Bakketeig, 2000) and epigenetic (Murphy *et al.*, 2012; Zhang, B. *et al.*, 2018) studies. However, this first pass work will focus on autosomal chromosomes only. While the effect of MSP very likely has interactions with sex chromosomes, the far more limited pathways likely implicated with the latter limits the informative versus complicating consequences. For this reason, sex chromosomes are excluded from analysis.

3. The DNAm patterns will impact chromatin function in a stable manner. We expect that epigenetic changes capable of altering phenotype through cellular reprogramming will involve interactions with other epigenetic mechanisms that will reinforce and propagate these effects through multiple rounds of cell turnover.

a) Biological stability: For example, these changes will likely implicate mechanisms such as histone modification, miRNA-related regulation, and changes in both extronic and intronic areas, etc.. As such, we expect the patterns to be widely spread throughout the

genome and affect multiple functional areas outside of traditional regulatory elements like TSSs. This marks a specific shift away from gene-centric views of DNAm differences.

b) Temporal stability: following the above, if a DNAm pattern mediates phenotype-relevant cellular reprogramming over the subject's life course, it follows that it will persist through time such that impacted DNAm domains can be identified in samples throughout childhood.

4. The DNAm patterns will have universal effects across populations. We hypothesize that using subject vulnerability to MSP will potentially identify core changes to cellular programming that affect human health trajectories over time. Core changes should be independent of genetic and non-genetic disease susceptibility forces. Thus, the changes will be reproducible in truly independent cohorts. While cohorts will use the same or similar DNA methylation microarrays to ensure similar chromosomal coverage, other data variables such as the population, (e.g. racial mixture, exposures, demographics,) timing of sample collection, storage, extraction, and statistical processing of the data do not necessarily need to be uniform in order to capture the similar areas of impacted DNA.

1.6.3 Objectives

This work represents an entirely novel approach to viewing the interplay of genetic and non-genetic factors related to MSP and DNAm. It explores the use of multidimensional data to better define individual predisposition to MSP-related health differences. We venture that improved estimation of disease susceptibility from early-life exposures will better identify rational DNAm targets that relate to vulnerable health trajectories in childhood, and thus prevent more harmful and/or irreversible pathology in adulthood. To test this methodology and associated hypotheses, the specific objectives of this thesis were as follows below and presented in Figure 15:

1. Mapping infants to their MSP-related vulnerability
2. Mapping MSP-related vulnerability to DNAm patterns
3. Exploring clinical relevance of DNAm patterns
4. Exploring functional relevance of DNAm patterns
5. Exploring the replicability of DNAm patterns over time
6. Exploring the replicability of DNAm patterns in a different population

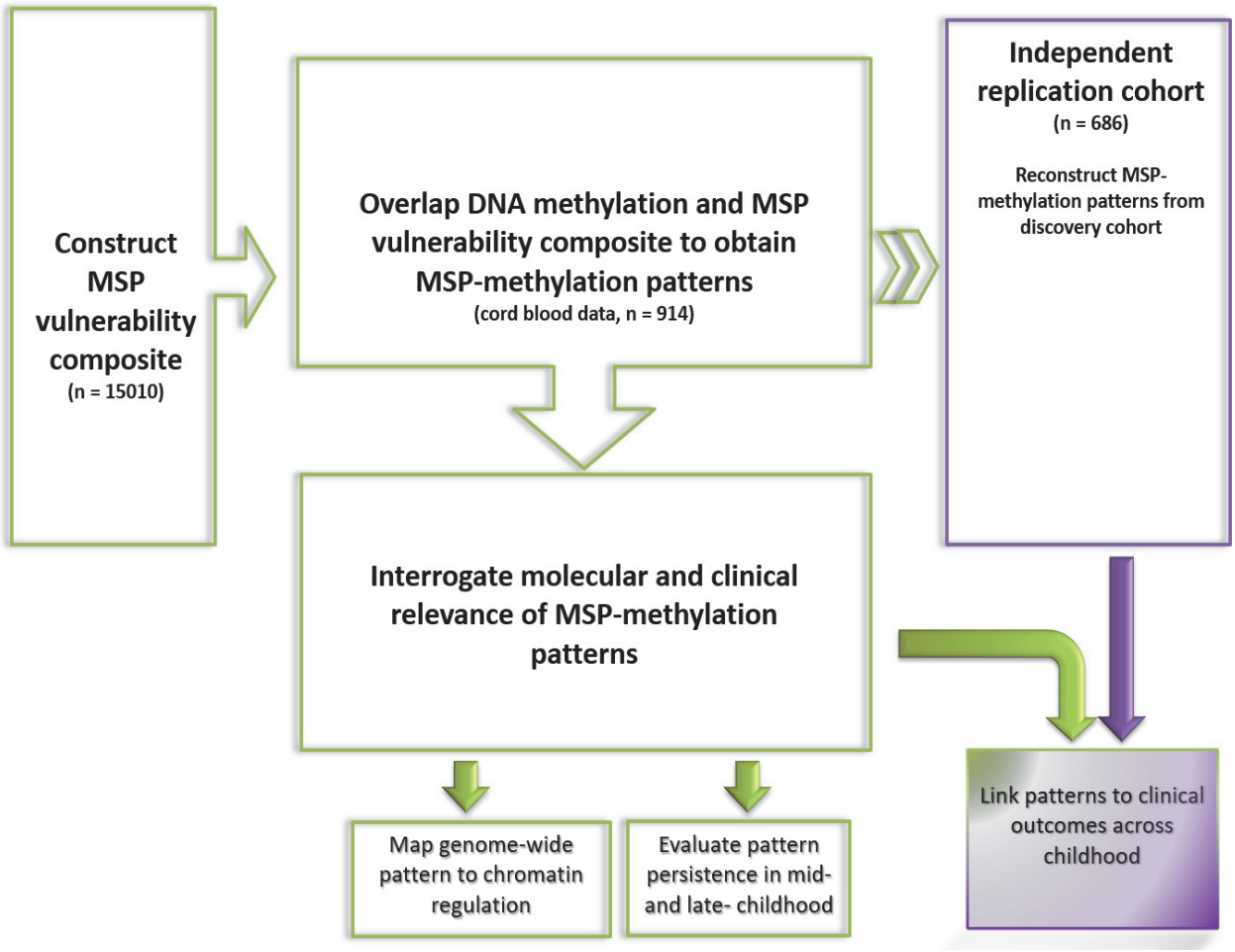


Figure 15: Schematic of thesis objectives. Green – data from discovery cohort. Purple – data from replication cohort. For clarity: Factor analysis (MSP composite) results referred to as "dimensions", PLS (DNA methylation) results referred to as "components".

Chapter 2 Methods

2.1 Data source

2.1.1 Discovery cohort

ALSPAC collected data on a total of 15,454 pregnancies resident in the former county of Avon, United Kingdom (UK), with expected dates of delivery 1 April 1991 to 31 December 1992.(Golding, 1990) This resulted in 15,589 known foetuses. The study website contains

details of all the data that is available through a fully searchable data dictionary[2]. Written informed consent was obtained for all ALSPAC participants. Ethical approval was granted from the ALSPAC Law and Ethics Committee and the local Research Ethics Committee in accordance with the guidelines of the Declaration of Helsinki.

Table 2 shows a summary of variables available including the source of the data and the approximate timing of collection.

Table 2: Schematic of ALSPAC variables showing source and timing.

Time		Fetal Period				Postnatal period															
		8 weeks	12 weeks	18 weeks	32 weeks	Birth	8 weeks	8 months	21 months	33 months	47 months	61 months	73 months	7 years	8 years	9 years	10 years	11 years	12 years	13 years	15 years
NHS STORK maternity database		x	x	x	x	x															
Questionnaire about child environment and health	Mother-reported	x	x	x	x		x	x	x	x	x	x	x	x	x	x	x			x	
	Child-reported													x	x	x	x	x	x	x	x
School results from national database											x	x	x		x					x	
DNA methylation data collected						x								x							x
Child biophysiologic testing	Includes blood tests, exercise tests, X-ray, etc..													x				x		x	x

2.1.1.1 Exposure data

Mothers completed multiple self-administered questionnaires. At about 18 and 32 weeks gestation, maternal smoking was assessed used the following questions to assess smoking: "*Did you smoke regularly at any of the following times in the last 9 months: 1) Before pregnancy, 2) first 3 months of pregnancy and 3) last 2 weeks*". Mothers endorsing smoking

[2] (http://www.bris.ac.uk/alspac/researchers/data-access/data-dictionary)

had the options of cigarettes, cigars, pipe or other. Data was recoded such that any answer but "No" or "I don't know" was coded as yes. Mothers also answered questions regarding smoking of the maternal grandparents (including grandmaternal smoking during the mother's gestation), the mother's partner and other household members. SDP is related to genetic variants and offers the possibility of conducting IV analysis using SNPs to strengthen causal inferences between exposure related methylation and outcome (for example, see use of Mendelian randomisation in (Wiklund *et al.*, 2018). However, we are underpowered to conduct such analysis in this single cohort (Richmond *et al.*, 2015). We transformed birth weight into a z-score (labeled as bwzscore) adjusted for sex and gestational age using a multi-ethnic, multi-country reference available through the R package hbgd (Villar *et al.*, 2014).

2.1.1.2 Outcome data

The following variables were selected due to their theoretic relevance to exposure to MSP as well as whether serial data are available.

Behaviour – ratings using the Strengths and Difficulties Questionnaire (SDQ) provided by the mother at ages 81, 115 and 140 months and by the teacher at 120 and 156 months. Subscale scores were pro-rated if one or two items were missing.

Academic performance – Scores from Standard Assessment Tests (SATs) administrated by the UK Department of Education were linked to ALSPAC subjects for Key Stages 1 to 3, (corresponding to ages 5-7, 8-11 and 12-14 years, respectively.) As in previous research using this cohort data, we used raw scores in all models (Booth et al., 2014; Meadows, Herrick, Feiler, & ALSPAC Study Team, 2007).

Neurodevelopment - Assessed at age four years and age eight using researcher administered measures: the Wechsler Pre-school and Primary Scale of Intelligence ^{UK} (WPPSI) and Wechsler Intelligence Scale for Children-III ^{UK} (WISC), respectively. Full details of the tests, scoring and inter-rater reliability described in (Taylor, C. M., Kordas, Golding, & Emond, 2017). In summary, subtest scores for both IQ tests were calculated to create a Verbal IQ and Performance IQ (as well as a Total IQ score reflecting a combination of both.) Child development before age four assessed using maternal questionnaire based on the Denver Developmental Screening Test at four time points: 6 months, 18 months, 30 months and 42 months (Iles-Caven, Golding, Gregory, Emond, & Taylor, 2016).

Anthropometric – height, weight, body circumferences and blood pressure were measured throughout infancy and in 1-2 year intervals throughout childhood starting at age 7. Body composition estimated using dual-energy x-ray absorptiometry (DEXA) at age 9, 11 and 13 years. Values were converted to internal z-scores (sex specific) using data from the full ALSPAC cohort. Individual growth trajectories calculated using multilevel models with random effects are available from the ALSPAC data repository for all subjects with two or more measures. Internally derived z-score for deviation from average is labeled as zwres0. Internally derived z-score for deviation from average change in weight/height between birth and 3 months, 3 and 12 months and 12-36 months of age are labelled as zwres1/zhres1, zwres2/zhres2 and zwres3/zhres3, respectively.

2.1.2 ARIES DNAm data

A convenience sample from ALSPAC had DNAm data collected from blood samples at birth (cord blood) and around ages 7 and 15 years, henceforth referred to as Accessible Resource for Integrated Epigenomics Studies (ARIES) data. DNA extraction, bisulphite conversion and DNAm data measurement using the Illumina Infinium HumanMethylation450 BeadChip (450K beadchip) as well as semi-random sample distribution across chips (to reduce batch effects), quality control and subject mismatch checks using genotype probes/sex-match as previously described (Relton, Caroline L. *et al.*, 2015).

Using the 450K beadchip, DNAm data are expressed as percentages ranging from 0 to 100% at each methylation site, known as a beta value. This value reflects the percentage of the sample that is methylated at that specific probe site Figure 16. It does not reflect partial methylation as methylation is an all or none event at a given locus on each DNA strand. That means that in non-gamete cells, methylation at a site can only be 0%, 50%, or 100% within a single cell.

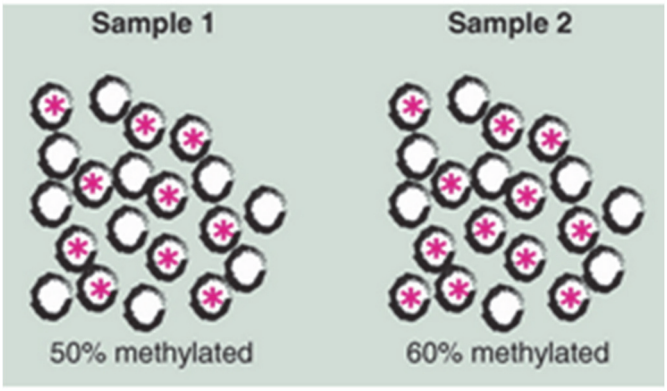


Figure 16: Schematic of mixture of cell with either methylation or no methylation at a single CpG site. The different proportion of cells with methylation results in beta values that could theoretically range from 0-100% on the 450K chip. Adapted from Holbrook, Huang, Barton, Saffery, & Lillycrop (2017).

Other literature has argued for the use of M-values, the logit transformation of beta values. This is a worthwhile consideration for linear analysis as M-values are more homoscedastic. However, several studies have shown that beta values are unlikely to cause statistical conclusion problems and therefore conversion to M values does not appear necessary and may even lead to the "creation" of outliers (Teschendorff, Andrew E. & Relton, 2018). With this understanding, we proceed using beta values.

Initial data analysis was performed using raw beta values from ARIES. Subsequently, analysis was updated using Version 2 data release (Spring 2016) where data are provided preprocessed on site at the University of Bristol. ARIES pre-processed data methylation data (background correction and subset quantile normalization performed within each time point) in R (version 3.0.1) with the wateRmelon package (Pidsley *et al.*, 2013) using the algorithm created by Touleimat & Tost (2012) to reduce the non-biological differences between probes. The beadchip probes were filtered by detection *p*-values and variability. Data points with low signal:noise ratio (detection $p > 0.01$) or with methylated or unmethylated read counts of 0 were also excluded. This left 417832 probes in the cord blood ARIES data from the ~485 000 cytosines tested on the 450K beadchip. Previous work has shown the DNAm measured with this chip can be confounded by cross-hybridization with other genomic sites and proximity to SNPs (Chen, Y. A. *et al.*, 2013). We used the DMRcate R package to filter these ~48 000 unreliable probes (Peters *et al.*, 2015). This left 377 460 probes. We performed discovery analysis of relevant DNAm patterns from all 914 subjects with cord blood data from ARIES.

2.1.2.1 Filtering of probes with low variability

A common first step before dimensionality reduction is to remove data with low variability as these data are unlikely to contribute to methodologically distinguishable processes. As well, testing only higher variability probes can reduce multiple test correction penalties (Edgar *et al.,* 2017).

As previously described by Edgar *et al* (2017), we filtered probes using a threshold of 5% range in change in beta values. Specifically, we retained only sites with DNAm values in the cord blood cohort that varies by at least 5% when taking the difference between the top and bottom methylated values between the 10^{th} and 90^{th} percentile range. This filtered about 200000 probes, leaving 185466 probes for further analysis.

2.1.2.2 Noise

We aim to use as much data as possible to distinguish true relations between variables of interest and unmeasured and/or unknown sources of noise (Teschendorff, Andrew E. & Relton, 2018). While PCA and ICA may have limited utility in identifying MSP-related patterns, it has been explored specifically in DNAm to capture variability related to noise. As such, we employed the use of the R package reFACTor (Rahmani *et al.*, 2016) and normFact (Teschendorff, Andrew, Renard, & Absil, 2014) to estimate the variability related to cell type and batch, respectively.

(a) Cell type composition

Regarding cell type correction, we use data driven methods as the ARIES database does not contain cell composition information. For cord blood DNAm data, we estimated cell composition using the meffil R package (Min *et al.*, 2018). This package has several cord blood cell type references. One of the most commonly used references in cord blood DNAm studies is that from (Bakulski *et al.*, 2016). However, this resulted in a large number of negative cell count estimates (both with and without filtering as described above.) We compared these results with the other references available in meffil. Among these, the gse68456 reference provided the least number of negative estimates. This reference uses a more "stringent" cell sorting protocol in that it excludes erythroid lineage-specific markers (de Goede *et al.*, 2015). Accordingly, we found very different estimated counts particularly with nucleated red blood cells (nRBCs) between the two references, as well as between the two nRBC estimates and their relation with DNAm components (data not shown but available upon request.) Furthermore, this reference has also been previously employed with ARIES data (Timms *et al.*, 2019). As such,

we proceeded with this gse68456 reference in our analysis. The cord blood reference uses seven cell types. As in Houseman, Molitor, & Marsit (2014), negative proportion estimates are corrected *post hoc* to zero.

The reFACTor package is specifically designed to perform reference-free cell type heterogeneity adjustment using PCA. It is an unsupervised method in that no information about exposure or phenotype is entered. We used this reference-free method as there is no appropriate age-specific blood cell composition references for ages 7 and 15. Using reFACTor, we trialed setting k to values 5, 6 and 7 to represent the common cell types sorted by flow cytometry (CD14+ monocytes, CD19+ B cells, CD4+ helper T cells, CD56+ NK cells, CD8+ cytotoxic T cells, eosinophils and neutrophils). We found similar correlations (that had $p < 0.05$) between DNAm patterns and the reFACTor components regardless between these three k values (see Appendix C.) As such, we set k=5 to represent the major cell types found in healthy individuals outside the newborn period.

(b) Batch effect

In the ARIES dataset, bisulphite-converted DNA (BCD) plate is known to be the strongest batch effect, even compared to covariates such as physiologic artifacts like sex or blood cell count artifacts or technical artifacts related to the microarray chip, laboratory conditions, etc. (Joubert, Bonnie R. *et al.*, 2016). We proceeded to use spatiotemporal ICA factorization (Teschendorff, Andrew *et al.*, 2014) to remove this batch effect on data processed to this point. This is a matrix factorization based technique to correct batch effects. This method is commonly applied in biology to model covariates in studies of differentially expressed genes. It assumes that rank-one components can represent gene by sample differences. This would apply in cases where a clear artefact (such as BCD plate) exists that facilitates recovery of batch-related components. As well, this method well matches our rationale that DNAm signals of biological interest are intermixed with noise. This ICA method leaves behind a “cleaned” data matrix that theoretically only has variability specifically due to that batch effect removed. Though this risks leaving behind variability related to other technical artefacts, we spare removing variability intertwined with biological signal (Renard & Absil, 2017). Thus, this ICA method is in keeping with our attempt to carefully “unearth” the underlying and subtle MSP-related vulnerability patterns in cord blood.

To implement the normFact R function (provided directly first author of Renard & Absil, 2017), the user must set the α parameter (a value bound by 0 to 1) that represents the extremes of two

ICA factorization assumptions: independence among genes versus independence among samples, respectively. An alpha of 0.5 would thus represent a perfect trade-off between both of these options. After discussion, the author recommended to set alpha to zero. Figure 17 shows the normFact results on cord DNAm with the relation between the ICA component and each BCD number. As suggested by the authors, BCD components with $R^2 > 0.5$ were removed. Thus, we removed eight BCD-related components.

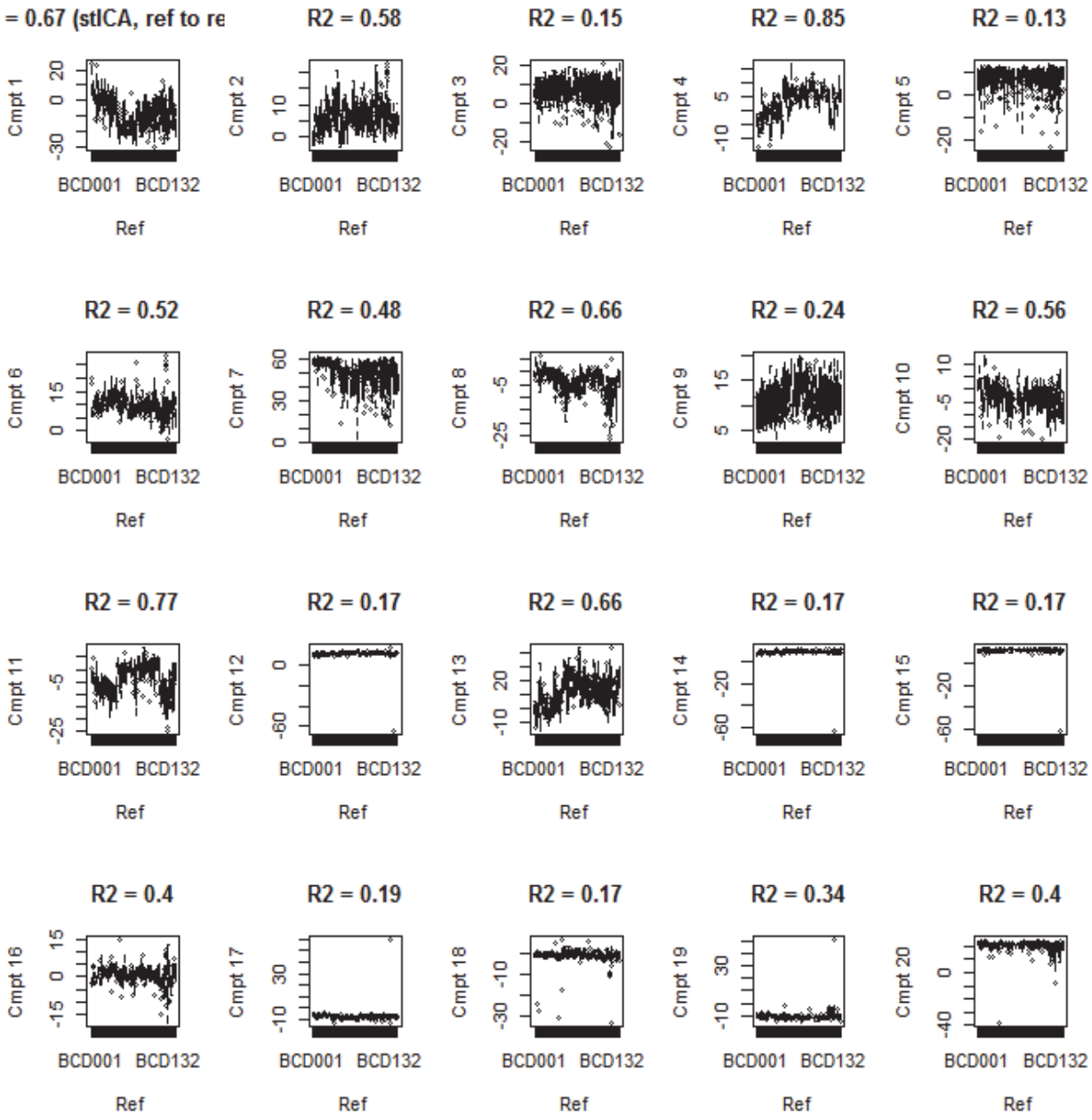


Figure 17: BCD plate related components obtained using normFact R function (spatio-temporal independent component analysis, alpha = 0) on cord ARIES DNAm data. Components with R2 > 0.5 are removed.

(c) Subject sex

Sex is a biological variable known to mediate the effect of exposures on health outcomes. For example, studies have observed the differential effect of maternal smoking on child outcomes based on child sex, (see Figure 18 for an example of this phenomenon seen in childhood BMI trajectories.)

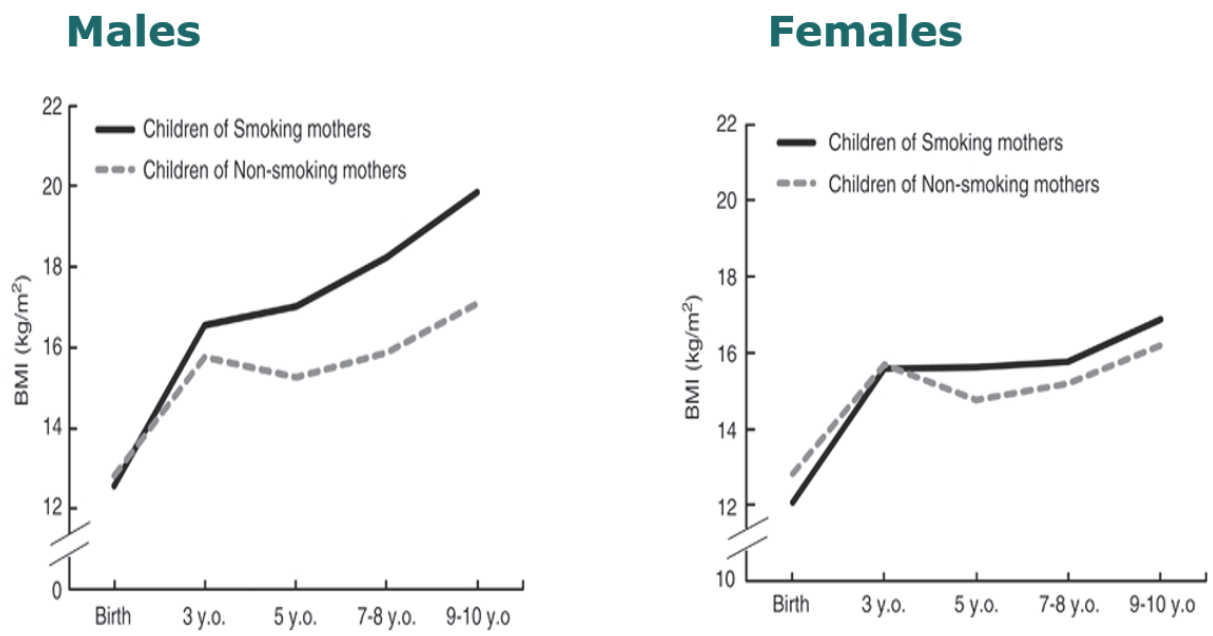


Figure 18: Example of differential effect of maternal smoking on risk of obesity based on child sex. Suzuki , Int J Obesity, 2010

One can consider two means to deal with this very powerful influence: 1) analyse the two sexes separately, 2) attempt to statistically control for sex during outcome/predictive analysis and 3) try to account for variability due to sex before outcome/predictive analysis. The effect of subject sex on health is an entire topic unto itself. However, a cursory examination of these 3 possible means to account for this factor reveals some clear limitations. The first option has the lowest risk that any opposite effects observed by sex would not "dilute" the total observed effect. However, this option also decreases the study power but limiting the sample size usually by half. The second option is frequently used in medical studies. It assumes that there is a linear and direct and/or interactive relation between sex and the outcome that can be accounted for such that if sex played an important role, the effect of the predictors would be attenuated once sex was included in the model. These assumptions may or may not be true.

The third has become increasingly popular in biological high dimensional data analysis. A canonical example is the use of SVA to try and characterize the variability due to the covariates such as sex. This variability is represented as data-specific variable unto itself i.e. it is not left as a binary "female versus male" variable but an expression of the specific pattern of sex in the observed data.

While there is likely no single correct answer for how best to deal with the influence of sex, it is clear that it is an important variable to consider if we are to better understand the foundation of molecular disruptions caused by exposures that lead to cellular dysfunction. In this thesis, we posit that there exist underlying disruptions that are sex-nonspecific. However, that does not preclude that sex likely interacts with such disruptions to alter the manifested phenotype. As well, there are likely important sex-specific disruptions that occur due to a given exposure. However, the size of our cohort would make separate analysis by sex very underpowered. As well, our search for MSP-related patterns starts at birth, before the emergence of secondary sexual characteristics which significantly expand the phenotype divergence between sexes. As such, we believe our assumption that non sex specific changes identified at birth can have important short and long term effects on both sexes is reasonable. Moving forward, we attempt to identify DNAm patterns that appear strongly influenced by sex. While these patterns may contain important information, we choose to exclude such patterns at this time and to focus on other patterns that have a sex-ambivalent effect. As well, as suggested in the 3rd option, we include DNAm variability due to sex as a predictive variable. In that way, sex "competes" with other DNAm patterns in terms of relevance to predictors. If sex related variability ranks among the other patterns, that may suggest direct or interactive molecular effects on outcome.

2.1.2.3 Collinearity

As well, single-probe based analysis is less reflective of the biological reality of DNAm mediated gene regulation. Differential methylation related to disease or aging largely demonstrate spatial correlation, usually within 500 base pairs but even beyond. These patterns likely represent the physical occupation of epigenetic machinery like endonucleases and cofactors on the DNA strand (Teschendorff, Andrew E. & Relton, 2018). Statistically, this represents collinearity. This means that one could linearly predict the value of methylation at one site from that of another with a non-trivial degree of accuracy. Pattern recognition is ideal to deal with this statistical property, whereas this is a challenge for traditional parametric techniques such as ANOVA.

2.1.3 Replication cohort

The Generation R Study (GenR) is a population-based prospective pregnancy cohort study. It included 9778 women and their children, born between April 2002 and January 2006. This cohort collected DNA methylation data from 1396 cord blood samples using the Illumina 450K Infinium BeadChip. Besides detailed pregnancy data, this cohort collected substantial offspring data, including anthropometrics at birth, in infancy, and at ages 6, 10 and 13 years. Full background and design of this cohort have been previously described in detail (Kooijman *et al.*,

2016a). Briefly, obstetric records of mothers were retrieved from hospitals and midwife practices to obtain pregnancy outcomes such as infant sex and birth weight. Information about child growth (length (height), weight, head circumference) was collected at each visit to the routine child health centers in the study area using standardized procedures and at the research center at 6, 10 and 13 years of age. We used DNAm data collected from the cord blood of a subgroup sample of GenR consisting of a total of 969 children of European descent as described in Joubert *et al* (2016). Briefly, DNA extracted (using the salting-out method). Quality control of analyzed samples was performed using standardized criteria. Probes with a single nucleotide polymorphism in the single base extension site with a frequency of > 1% in the GoNLv4 reference panel (Genome of the Netherlands Consortium, 2014) were excluded, as were probes with non-optimal binding (non-mapping or mapping multiple times to either the normal or the bisulphite-converted genome) (Bonder *et al.*, 2014), resulting in the exclusion of 49,564 probes, leaving a total of 436,013 probes in the analysis. We ran DASES normalization using a pipeline adapted from that developed by Touleimat & Tost (2012). DASES normalization includes background adjustment, between-array normalization applied to type I and type II probes separately, and dye bias correction applied to type I and type II probes separately and is based on the DASEN method described by Pidsley et al, but adds the dye bias correction, which is not included in DASEN (Pidsley *et al.*, 2013). We then overlapped these sites with those representing the ARIES cord DNAm patterns, followed by low variance probe filtering as described in Section 2.1.2.1. However, this overlap followed by filtering at the 5% change threshold resulted in a remaining set of 88807 probes. This was unsurprising as the variance threshold employs quartiles that would reasonably shift “inwards” after the probe matching with ARIES DNAm patterns. To account for this and better match the dimension size of the two cohorts, we relaxed the threshold to 2% that left a final set of 173565 probes, (a 3% threshold left 148285 probes and a 1% threshold left 175060 probes.) We then proceeded with batch effect removal as described for ARIES data in Section 2.1.2.1.

2.2 Mapping individual data

The first three methods represent the bulk of this thesis. The first two objectives of mapping MSP-related vulnerability and DNAm share the common need to condense a relatively high number of predictors (e.g. exposome data or DNAm microarray sites) into a smaller number of representative variables that are “concentrated” in their relation to biological mechanisms related to MSP. The third objective seeks to relate these new representative variables to the newborn’s

future childhood outcomes. However, all three objectives involve pattern finding. At this juncture, we should mention the concept of overfitting which will have important implications in general for pattern finding but particularly in context of high numbers of predictors.

2.2.1 Overfitting

To understand overfitting, we must first establish the primary goal of pattern finding: obtaining among competing models the solution that best describes the data overall. Overfitting describes the situation when a method and/or high dimensionality of the data allows a degree of flexibility such that the data can be "made" to fit the question posed by the researcher. Moreover, the resulting solution fits a certain sample of data very well, but cannot be generalized to other data of the same phenomenon.

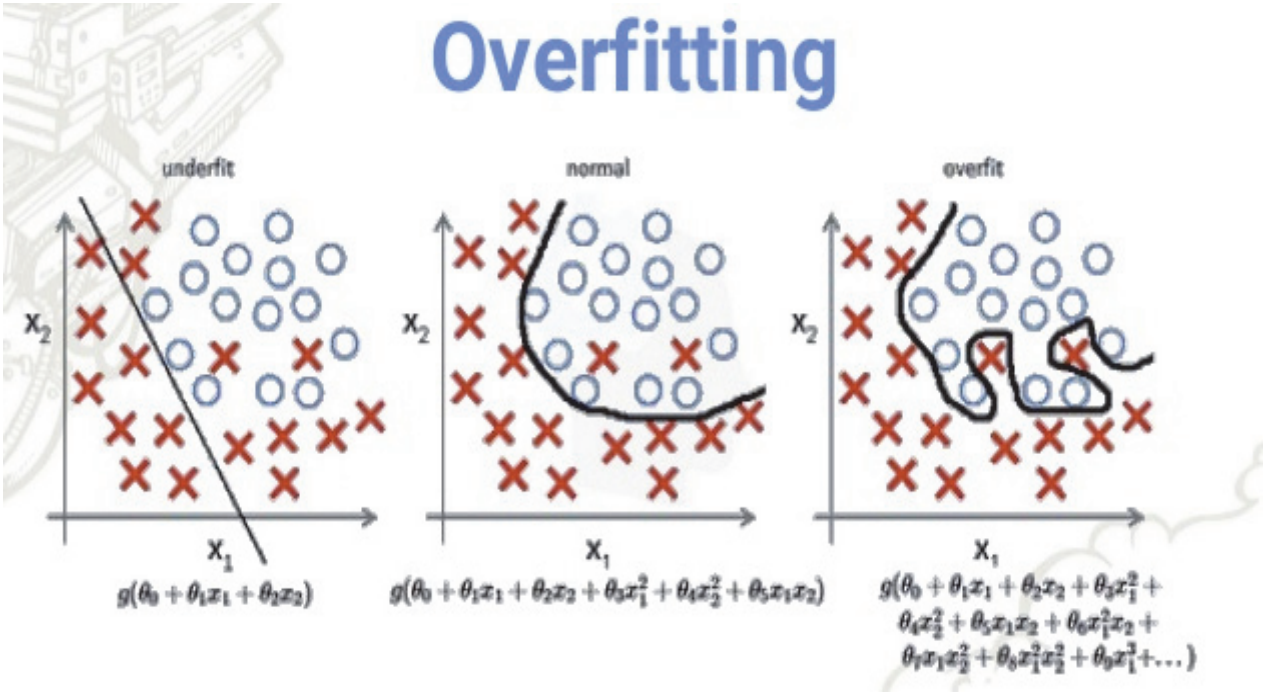


Figure 19: Schematic of overfitting in pattern finding. Image from wiki.org.

In many disciplines (including medicine), the philosophical principle of Occam's Razor is taught to imply that simpler solutions are "better". In its original form, it states "Entities should not be multiplied beyond necessity." The right most image in Figure 19 may seem to imply that more complex solutions are "wrong" – however, the solution is inappropriate if it cannot be generalized to new data. It has also been argued that studies demonstrate more examples of

the success of simpler versus more complex solutions merely because of the types of problems researchers ask (Duda, Hart, & Stork, 2001). As humans, we may naturally ignore problems that only have complex solutions. With this in mind, this section discusses overfitting with the aim to find the best fit to our problem and not necessarily the simplest classifier.

To combat overfitting the data, there are several methods including regularization, penalization, minimization of description length, etc. The comparison of these continue to be hot topics and far beyond the scope of this thesis. However, it is agreed that there is no universal means to avoid overfitting (Wolpert & Macready, 1997). The choice of method is problem dependent. In other words, overfitting avoidance can have different effects in one context compared to another and can lead to artificially worse performance in some cases (Duda *et al.*, 2001). The success of a solution depends on the match between the problem and the pattern finding technique – not the overfitting method imposed upon it.

We provide a cursory overview of broad concepts to combat overfitting. First, many researchers propose using as large a sample as possible. By feeding more data into the pattern finding method, the method may better learn to find the signal of interest. This practice is supported by the often poorer performance when low sample sizes are used. However, if the additional data contains a lot of noise, this method can actually worsen generalizability. As well, obtaining more data can be infeasible.

A second broad category of techniques involves resampling. These methods are the most popular among omics research. Appearing in many forms, the basic principle is to use multiple samples in an attempt to better estimate statistics by comparing/collating the model's performance on “unseen” data. Examples include cross-validation (CV), k-fold CV, repeated k-fold CV, Monte Carlo CV, bootstrap, boosting, Jackknife, and learning with queries. However, a clear minimum of prediction errors may not be obtained with resampling, making model selection difficult.

A third category is penalization. This group is based on the idea that high dimensional data may only contain a sparse set of relevant information. Thus, overfitting can occur more readily if a model is more complex due to the inclusion of irrelevant information. As such, this group of methods penalize complexity that does not improve a given model’s performance. This is accomplished by controlling parameters in the model that affect model complexity. Often, these parameters cannot be directly estimated from the data and are supplied by the user. This typically involves a search procedure to find the most generalizable tuning parameters, such

as a grid search, gradient search or random search. To make tuning systematic and reproducible, many pattern finding techniques now have built-in optimization procedures. However, tuning in general is computationally expensive. As well, it can be dataset specific so re-tuning may be required when applied to other data such as for validation in an independent cohort. Penalization methods include regularization (for example, leading to sparsity) and pruning. Last, a model can be more prone to overfitting due to the presence of data disturbances due to randomness, outliers and noise. Ideally, the researcher is able to remove these disturbances before analysis. While it is difficult mathematically to extract randomness from data, many methods exist to denoise and to detect and remove outliers. However, as alluded to in Section 1.4: Mapping individuals to the risk context of MSP in the discussion of outliers, removing this unwanted variability from the data could also inadvertently remove important information. Therefore, many researchers now advocate for methods that are robust to variability originating from these entities either due to their match with the type of data and/or study design and research question. We consider both approaches in our methods described below.

2.2.2 Other statistical challenges

In addition to these general concerns with overfitting, we also face unique challenges posed by our specific dataset. In Objective 1, our goal is to represent a multifactorial entity using data that has both random and non-random missingness. This new vulnerability variable must represent degrees of risk of MSP-related effects despite the fact that there are relatively few instances of MSP exposure compared to non-exposure. Moreover, this vulnerability variable must be clinical sensible in order to act as appropriate "bait" to capture important DNAm patterns. In Objective 2, we seek to represent important DNAm variability in high dimensional data that likely has a very low signal to noise ratio, both due to the nature of DNAm and of complex traits. The following sections describe these challenges and our proposed solutions.

2.3 Mapping clinical data

2.3.1 Mapping individuals with unidimensional and/or categorical data

As discussed in Section 1.3.1, many studies have demonstrated clinical outcome or biological differences between children based on MSP. Such studies employ various methods using self-reported maternal smoking to classify this exposure. Some are simply binary (nonsmokers versus other), others attempt to quantify daily dosage (e.g. cigarettes per day,) and others use

smoking in various gestational periods. Using the latter method in the ALSPAC cohort, there is adequate maternal data to categorize MSP as in Figure 20.

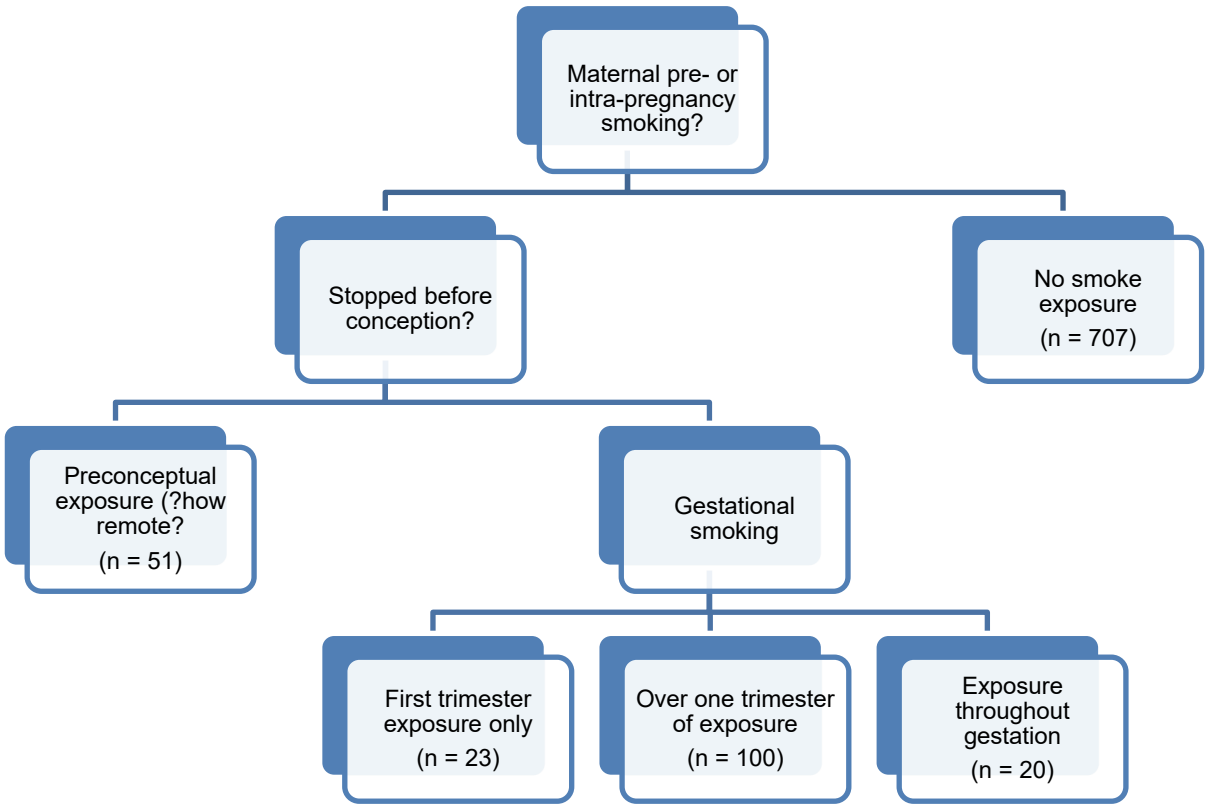
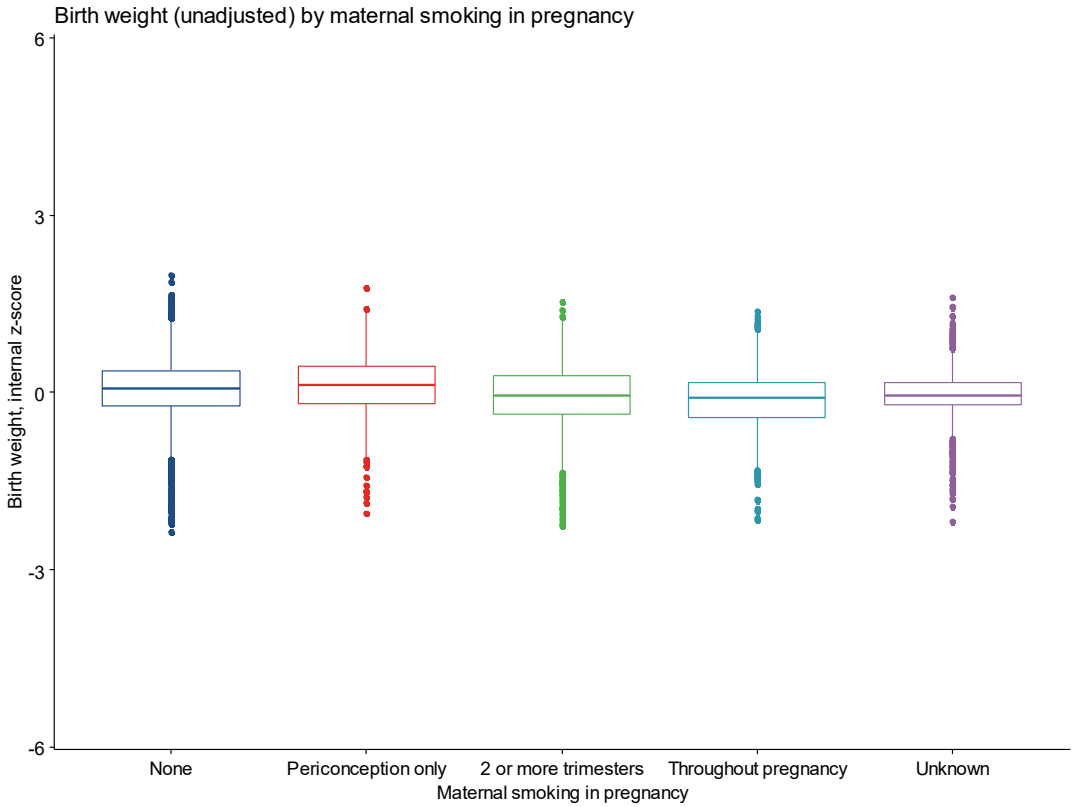


Figure 20: Maternal smoking - classification by gestational period. Numbers in brackets represent class size within the mother-infant dyads with infant cord DNAm data available, (total 914 dyads, 13 of which had missing or irregular coding of maternal smoking in pregnancy.)

Given the "successes" of these other reports, what evidence do we have that a different characterization system will improve our understanding of the impact of MSP on offspring? To investigate this question, we will compare this traditional MSP characterization with our other proposed methods. As discussed in Section 1.4: Mapping individuals to the risk context of MSP, we are curious about using counterfactual typical-atypical contrasts to better understand exposure-related susceptibility. This requires a proxy of vulnerability of the specific fetus to the exposure. One of the clinical proxies for fetal growth and development is birth weight, leading to linked MSP-birth weight categories shown in Figure 12. To explore how such a scheme would look in our data, we can plot the birth weight of children in each to explore the impact of this exposure (Figure 21).



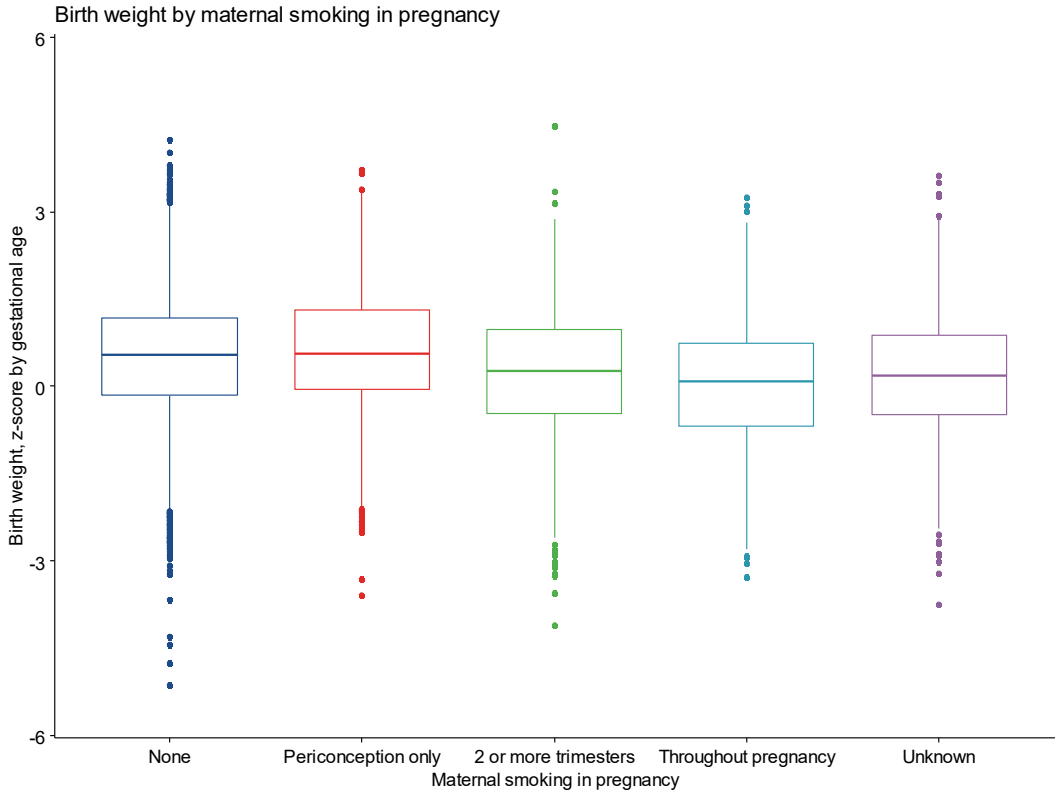


Figure 21: Boxplot of birth weight by maternal smoking classification. Top plot: Birth weight by z-score (internal to ALSPAC.) Bottom plot: Birth weight by z-score based on national references for gestational age.

Based on MSP, there remains a high level of variability on the effect on the child using outcomes like birth weight. Though there is a significant difference between the groups ($p < 0.001$), there is considerable "spread" seen in the range of weights. This is more pronounced once birth weight is corrected for gestational age using national reference values. This is particularly notable in the non-smoking group, likely reflecting the many other factors that dictate the size of a baby at a given stage of development.

This possibility motivates us to go beyond MSP and birth weight alone to capture fetal programming. As discussed in Section 1.4 Mapping individuals to the risk context of MSP, we wonder if we can better discern between MSP versus non-MSP-related risk by combining genetic and non-genetic factors. This leads us to further propose combining reports of MSP with family smoking history, social factors, and pregnancy factors in conjunction with birth weight. The combination of multiple variables to represent the cumulative prenatal environment particularly in the context of the DOHaD hypothesis is not new. This environment is complex and considered impossible to observe or measure directly (Bollen, Noble, & Adair, 2013; Camerota & Bollen, 2016; Camerota & Willoughby, 2019). Camerota and colleagues use the term "favorable fetal growth conditions" (FFGC) to describe "an abstract variable that encompasses all of the environmental, genetic, and epigenetic factors that program prenatal development." (Camerota & Bollen, 2016) These authors and others modeled FFGC as a latent variable using structural equation models. Other researchers combine multiple variables using a summative approach in which individual predictors are discretized (typically dichotomized) such that they are "unidirectional" i.e. the researcher determines which direction is considered a positive or negative influence and assigns points that are summed (Laucht, Esser, & Schmidt, 1997; Silveira *et al.*, 2017; Wade, Madigan, Akbari, & Jenkins, 2015). Table 3 is a typical example of such a method (Silveira *et al.*, 2017). Silveira *et al.* found that their cumulative score was a better predictor of several neurodevelopmental outcomes than birth outcomes or single predictors in the discovery cohort, the Maternal Adversity, Vulnerability and Neurodevelopment (MAVAN) cohort in Canada. They replicated this in an ethnically distinct cohort with different data variables, Growing Up in Singapore Towards Healthy Outcomes (GUSTO). These findings with robust replication point to the validity and utility of cumulative early-life environmental variables that relate to later health outcomes in the DOHaD paradigm.

Table 3: Typical example of summative index approach to creating a cumulative prenatal environment variable.

MAVAN	GUSTO
• Presence of chronic disease during pregnancy (diabetes, hypertension, asthma, current or resolved), current severe vomiting, vaginal spotting or bleeding during the past 4–6 weeks, current anemia/constipation/blood in stool, current vaginal/cervical/urinary tract infection/diarrhea	• Presence of chronic disease during pregnancy (diabetes, hypertension, current severe vomiting, vaginal spotting or bleeding during the past 4–6 weeks, current anemia)
• Birth size percentile below 10th percentile or above 90th percentile	• Birth size percentile below 10th percentile or above 90th percentile
• Gestational age ≤37 weeks	• Gestational age ≤37 weeks
• Household total gross income <$30,000/year	• Household total gross income <$1999/month
• Lack of money score above 9	• Smoking during pregnancy
• Presence of domestic violence or sexual abuse during pregnancy	• Maternal mental health at Week 26 (presence of BDI ≥ 14, EPDS ≥ 93)
• Marital strain score >2.9	
• Smoking during pregnancy	
• Pregnancy anxiety >1.95	
• Prenatal depression score ≥22	

Note: The presence of each component (described in each bullet) yielded 1 point, and the scores represent the summation of points.

2.3.2 Combining multiclass data from multiple sources

The major challenge posed by integrating exposome data is how to harmonize and aggregate multiclass data from multiple sources. While there are many techniques to represent the combined effect of quantitative variables, methods to combine that and qualitative data are far fewer. In the following, we consider only those methods capable of combining both quantitative and qualitative data types.

2.3.2.1 *Integrating variables - Methodologic assumptions regarding variable relations*

When exploring relations between variables, we have already discussed the importance of considering which variables are relevant and among those, whether there is a direct, inverse or even non-linear relation of one variable relative to the other. As well, what is the strength of that relation? In other words, one requires the variable weighting. Weighting includes a numeric value and an ordinality that captures the variable's importance and its direction of effect on the specific outcome interest.

The summative scoring method has the advantage of flexibility in assuring that selected variables always contribute to the global score. In this method, variables must always be

discretized and transformed to an ordinal scale. The ordinal value is multiplied by the weight and then all variables are summed. Weighting typically is distributed relative to all variables and so usually the sum of all weights equals 100%. If all variables are considered of similar importance, then the weight of each could be considered equal to 1 divided by the number of variables. Weight assignment may also be based on previous knowledge. However, there is not always data available to guide assignment in which cases researcher-defined parameters must be used. In either case, each study is influenced by these choices making comparison of results between studies more challenging. This is a problem that exists not only in biomedical research but in other areas where mapping is required (Le Clec'h *et al.*, 2016; Ragland, 1992; Taylor, J. M. & Yu, 2002) Scoring also has the advantage of easy interpretability when scores are very high or very low. However, scores in the moderate range provide little information to distinguish the different possible causes in each individual. Another way to view this limitation is that the score does not provide any information on the inter-relations between variables.

Another strong limitation of this method is that the weight of a variable is specific for a given outcome, meaning that a new weight is required when studying the same score for a different outcome. This highly complicates the design of experiments using more than one outcome. As well, it creates further sources of variation when comparing results from different studies. Therefore, there are several reasons why the summative scoring method may have restricted applicability and validation in CCD that affect the general population and likely involve multiple affected traits. The issue of weighting also brings up the issue of weighting similar information from different sources. Is the mother's perception of her partner smoking to be weighted the same or differently than the partner's self-report? Which one "matters" more to the fetus?

Instead of a summative score, we could categorize individuals based on key qualitative and quantitative variables. As discussed in Section 1.4: Mapping individuals to the risk context of MSP, we could visualize individuals on a spectrum of MSP exposure versus effect on fetal growth. In this context using birth weight as a proxy of fetal growth, we could analyse the characteristics of "typical" individuals who demonstrate outcomes as expected given their risks versus "atypical" individuals i.e. those that demonstrate low birth weight despite low risk or high birth weight given high risk factors. In this scenario, the advantages is a clear demarcation between typical and atypical subjects which allows the discovery of similarities based on exposure-related poor outcomes, rather than effects that may be related to the exposure but do not appear to effect outcome (often referred to as the bystander effect). This would maximize the observed impact on clinical outcomes and support causal inferences.

An additional advantage is that weighting is unnecessary. This decreases the variability if the approach is applied in another population. However, the typical-atypical categorization approach shares some similar drawbacks to summative scoring. Both require discretization of variable values.

In general, discretization is also known to cause data loss and even provide biased results. In various clinical and simulation scenarios, categorization cut-off levels can falsely alter apparent associations with outcomes (Ragland, 1992; Royston, Altman, & Sauerbrei, 2006; Selvin, 1996; Taylor, J. M. & Yu, 2002). Thresholds have population and experimenter specific biases, even if based on previous literature. As well, thresholds risk misclassification and data loss – both which can dilute or inflate the observed effect of an exposure in an unknown manner (Ragland, 1992; Taylor, J. M. & Yu, 2002) This known phenomenon has resulted in the often ignored advice to avoid discretizing continuous outcomes (Royston *et al.*, 2006; Van Belle, Fisher, Heagerty, & Lumley, 2004). Furthermore, neither method accounts for inter-relations between variables.

Another disadvantage that represents the double-edged sword phenomenon of contrasting typical versus atypical individuals is that it tends to work best at the extremes of exposure and outcome. This best accentuates the contrast exposure-related outcomes versus bystander effects. However, this excludes subjects in the in-between of exposure and outcome. The analysis thus suffers from reduced number of subjects and subsequent statistical power loss. Moreover, this leads to concerns with applicability in the general population. For the study of CCD, most individuals are not at the extremes of the population. As well, optimal extraction of typical versus atypical usually involves 2 to 3 variables, in contrast to summative scoring where the researcher has the theoretic ability to include as many variables as deemed relevant in the score.

2.3.2.2 *Objective variable selection and mapping based on similarity of MSP vulnerability*

The above methods are limited primarily by the need for user-defined parameters for selection, discretization and relation between the variables of interest. While these parameters may be amenable to the study of diseases with extreme and/or more clear-cut risks and phenotypes

(e.g. the relation between smoking and race on prostate cancer) they may ill suit the study of CCDs with wide-ranging risk and phenotype across the population.

For this reason, we explored data reduction techniques that seek to represent the similarity and dissimilarity between individuals with a select number of dimensions that do not require researcher imposed parameters. These include techniques such a confirmatory factor analysis (Winchester, Sullivan, Roberts, Bryce, & Granger, 2018) and principal component analysis (PCA) (Duda *et al.*, 2001). This latter family of methodology has successfully linked these cumulative early-life variables to later child outcomes.

PCA is a widely used method for multivariate analysis of multidimensional data (Duda *et al.*, 2001). It is often described as an "unsupervised" learning algorithm meaning that no previous knowledge of the data under study informs the analysis (Pagès, 2014). PCA and related techniques finds an alternative set of coordinates to represent the data more simply. More specifically, PCA uses linear transformation of variables to generate *N* dimensional vectors, a new coordinate system where all vectors are orthogonal to each other. PCA can be understood as a mathematical means to express all variables through a small number of principal components (PCs). These PCs account for the maximum variance in the data. In a PCA model, the data (X) is decomposed into two vectors (the scores (T) and the loadings (P) where $X = TP^{T} + E$. E is the residual (i.e. the information not captured by the multiplication of the scores and the loadings.)

The true power of PC scores is revealed through plotting on the dimensional vectors: this allows the observation of groupings of subjects who are most similar based on the given vectors. In addition, plotting loadings reveals the most important variable features in the vector.

The methods described in the section above required the researcher to select which variables would be used in the score. In PCA, the researcher does not select the variables used in each dimension. Each dimensional vector is differentially influenced by some or all of the original variables. In addition, the researcher does not need to assign the directionality of relations between variables. As well, the relation between variables can vary in different dimensions. In this way, this method does not exclude possible inter-relations between multiple variables that may vary in different individuals. For instance, higher birth weight may or may not associate with MSP depending on the presence of grandmaternal smoking during pregnancy (i.e. maternal exposure as a fetus to smoking.) This adds richness to the characterization of vulnerability that can be accomplished through machine learning that is not possible with only research-defined parameters.

There are also model-based clustering of mixed data (e.g. R packages available include Rmixmod, VarSelLCM, ClustOfVar, MixAll, etc.) These approaches often combine variable selection along with model fitting using maximum likelihood estimates to achieve the best performance (Fop & Murphy, 2018). The overall goal is to separate individuals into clusters where within cluster variance is minimal and/or between cluster variance is maximal. In this way, clusters represent distinct patterns and members belong exclusively to one cluster. We attempted to use the package MixAll (Iovleff, 2019). We trialed various cluster numbers from 2-6. While a good proportion of subjects were assigned to each cluster (10% or more), it was difficult to interpret the clusters in terms of relative variable contribution (data available upon request.) As such, it rendered parameter selection ambivalent. We also attempted to use this method with the R package VarSelLCM (Marbac & Sedki, 2019). However, the variable selection only found the continuous variable (birth weight) relevant and therefore the clusters contained no information regarding the categorical variables. Given that we have a strong theoretic backing for the relevance of these MSP-related variables (Section 1.4: Mapping individuals to the risk context of MSP), this algorithm was unsatisfactory. We decided to not further pursue clustering as it did not appear a good fit with our design rationale and data.

2.3.2.3 *Factor analysis in MSP vulnerability composite construction*

In contrast to cluster analysis that aims to separate individuals, factor analysis focuses on variables. Factor analysis seeks the inter-relations between these variables that suggest they are causally linked or at least driven by a shared but unmeasured factor (Pagès, 2004). For example, one cannot directly measure mood. However, one can rate the level of fear of large groups, worries about health, feelings of dread, etc. Together, high ratings are more likely with anxiety traits. Thus factor analysis may combine these measures into a factor representing anxiety. Similarly, measures of sadness, lack of enjoyment, and low motivation would be higher in people with more depression and thus these might be combined to create a depression factor. In this work, the aim is to find factors that represent certain commonalities underlying the variables that connect or separate individuals in terms of their vulnerability to MSP and MSP-related risks. Thus, each individual is assigned his/her own individual profile expressed as to what degree he or she can be described by a given factor. Importantly, individuals may overlap in terms of their similarity to a given factor. Arguably, cluster analysis is more readily interpretable as individuals are assigned to mutually exclusively categories unlike factor analysis where individuals are ascribed a combination of factors.

A consequence of factor analysis is dimension reduction. Since the combination of two or more variables will be replaced by a factor, this methodology has the statistical advantage of minimizing Type 1 error due to multiple comparisons. As well, we hypothesize that using a MSP composite that combines multiple sources of information may provide a finer grained and more relevant measure to the specific individual under study as well as attenuating the risk of relying on a single quantity of MSP with an unknown level of measurement error.

We used the Factor Analysis for Mixed Data (FAMD) method in the FactoMineR R package which employs PCA and PCA based methods to describe the similarity between individuals using both continuous and categorical variable types (Husson, Josse, Le, Mazet, & Husson, 2016). Specifically, it uses PCA for the former and multiple correspondence analysis for the latter variable type. FAMD provides information on the inter-relations between variables that is entirely data driven and produces continuous values representative of both quantitative (i.e. birth weight) and qualitative (i.e. environmental smoke exposure) variables. As such, each dimension reflects information contributed by varying degrees by all variables. In other words, each dimension represents a pattern that arises from a specific configuration of variables driven by the relation between those variables.

Before analysis, FAMD requires normalization (centering and scaling) of the data elements prior to their aggregation. This is an attractive feature for data arising from multiple sources as each source is treated separate from the other. Otherwise, this method requires little manipulation of the data in that it does not require discretization, weighting or directionality assignment. For example, MSP can remain a nominal variable with no positive or negative assumptions imposed. Consequently, it is not restricted to have a specific ordinality in relation to infant birth weight within a specific context. The weighting of each variable in a given dimension is entirely driven by the data. This reduces researcher influence on assignment of meaning of the variables and thus undue influences on final results. This is particularly important given that our driving motivation is to obtain the “net” influence of diverse and interacting MSP-related factors. Assignment of meaning insinuates that the researcher knows the exact nature of the inter-relations between variables that exist in all individuals. When this is not true, PCA based techniques can be used to objectively select variables used to map similarities between individuals (or the subject of interest) (Northstone *et al.*, 2013; Rahmani *et al.*, 2016). This matches our objective to use infants with similar MSP-related vulnerability as bait to capture common DNAm differences that may represent health diatheses. Last, another feature of PCA-

based methods is that the extracted composites are orthogonal to each other, attenuating multicollinearity issues among predictors.

Another advantage of the FactoMineR package is the built-in multiple imputation function to deal with missing values. Before conducting FAMD, we performed data imputation for missing data using the MIFAMD function. We used the regularised iterative algorithm for imputation to avoid overfitting as advised by the authors.[3]

2.3.3 MSP vulnerability - Variables of interest

Considering previous studies and the non-genetic and genetic factors employed (see Table 3 for examples,) we explored factors such as smoke exposure during pregnancy, maternal second hand smoke exposure and familial smoking history, as well as pregnancy variables such as gestational weight gain, diabetes and hypertension. For familial smoking, we included the history of both the grandfather and grandmother. Regarding the latter, we included both grandmaternal history of ever smoking and smoking during the gestation of the subject's mother. The mother's exposure as a fetus not only potentially represents a genetic predisposition to MSP exposure, it may also pose a transgenerational consideration as this is a direct exposure of the ovum that ultimately forms the subject (Figure 22).

[3] We also tested another popular PCA based method to analyse mixed data: PCAmixdata (R package). We trialed both the PCAmix and MFAmix (multiple factor analysis with mixed data within a dataset) functions. Most components generated with this method were related to MSP. Despite entry of different numbers and combinations of variables, the maternal reported MSP remained the strongest influence, (data available upon request.) Of note, PCAmixdata uses generalized singular value decomposition compared to the factor analysis method in FactoMineR. Further exploration of why this lead to different results in our dataset is unfortunately beyond the scope of this thesis.

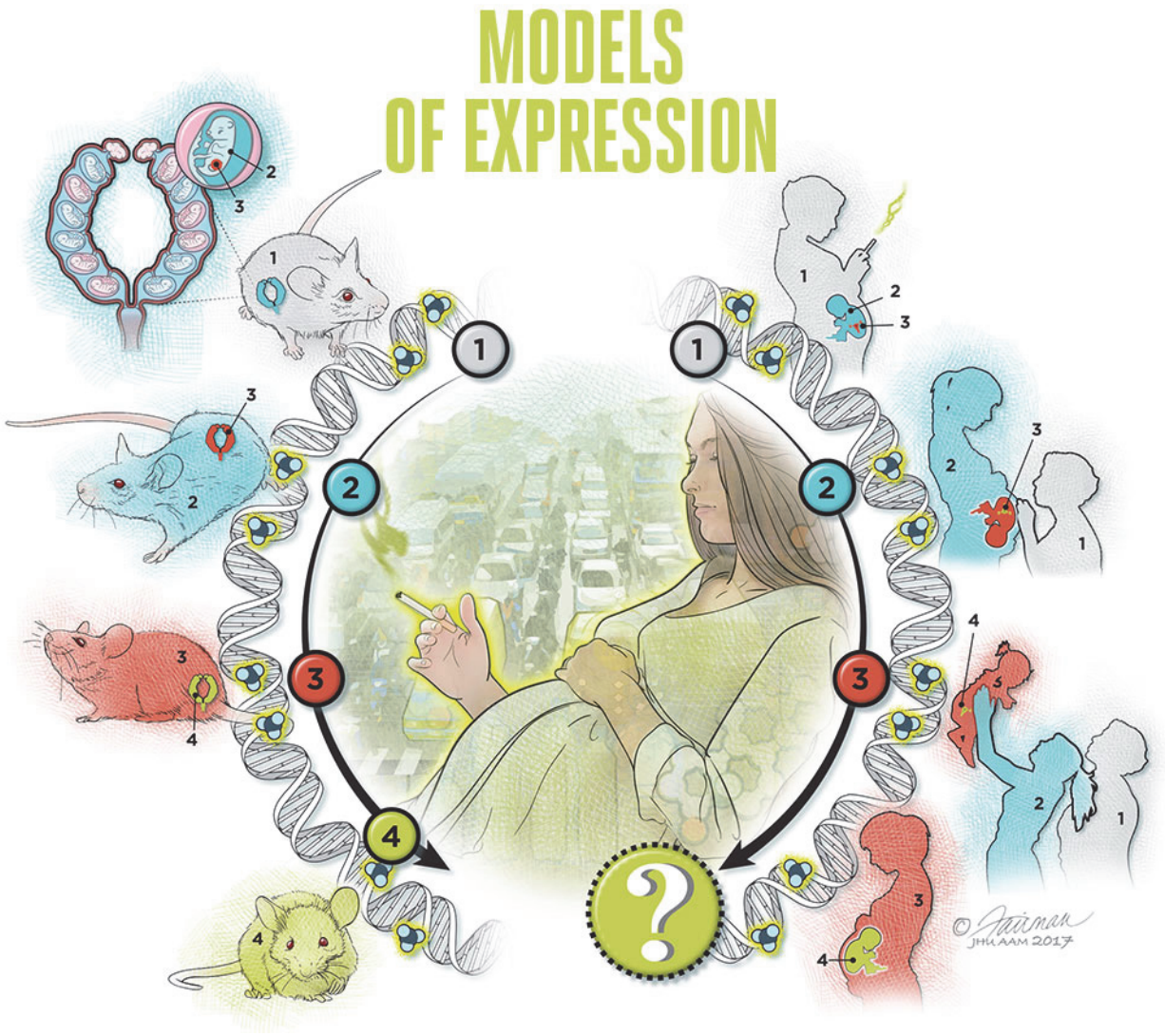


Figure 22: Genetic effects of smoking across generations. (Image from https://magazine.jhsph.edu/2017/fall/features/lasting-legacy-epigenetics-and-prenatal-environmental-exposures-studies/)

2.3.4 Summary of composite analysis and construction

Variables were centered and scaled before analysis. We predicted missing values based on similarity between individuals and relation between known variables using MIFAMD function. We performed tuning using the estim_ncpFAMD function to obtain the number of components used to predict missing data (result being ncp = 2). The number of imputed datasets was increased to 50 (nboot=50) from the default of 20 as per the author provided vignette; otherwise the function

default settings were used. Thus, the final dataset was the culmination of 50 imputed sets. We then used the FAMD function to conduct the composite analysis. To avoid confusion with the results of later analysis, the components of this analysis will henceforth be referred to as "dimensions".

2.4 Mapping the epigenome to visualize vulnerability related profiles

As aforementioned, most work in DNAm and complex disease has largely focused on the use of single sites independently related to outcomes or single sites conglomerated together. The latter has more frequently been taking the form of summative score creation – likely bolstered by the increasing success of genetic (i.e. allelic) scores as health predictors (for a recent example of the summative DNAm score approach being used in ALSPAC, see (Reed, Suderman, Relton, Davis, & Hemani, 2020). As discussed in Section 2.3.2.1: Integrating variables - Methodologic assumptions regarding variable relations, unknown methodological biases may arise from user-defined parameters for selection of DNAm sites and the direction, intensity, interactions and context (linear versus non-linear) of its relation between the variable(s) of interest.

As an alternative to linear methods, data mining and machine learning methods are better suited to detect multiple interactions, even non-linear ones. Such methods include multifactor dimensionality reduction, artificial neural network and statistical epistasis network (Moore, Asselbergs, & Williams, 2010). These methods are increasingly employed in omic studies as multi-way interactions are very likely the biological reality in complex diseases (Loucoubar *et al.*, 2017). However, there is a heavy computational cost when calculating all 2-way combinations of variables, let alone 3- or 4-way interactions. This computational burden is multiplied several hundreds of thousands fold in the case of microarrays like the 450K chip, rendering such analytic methods untenable on the whole chip. Approaches to reduce computational load include selecting only sites related to exposure or outcome, (this is the most common method in EWAS thus far,) filtering variables such as by removing "redundant" variables (e.g. methods like Spatially Uniform RelieF have been applied to GWAS (Greene, Penrod, Kiralis, & Moore, 2009),) or subset specific analysis, (e.g. separately analyzing probes belonging to similar annotations based on pathway analysis as reviewed in (Yu *et al.*, 2017). However, such data reduction methods carry the same risk of imposing known and unknown assumptions on the importance and interaction of variables in relation to phenotype (see discussion in Section 1.5: Mapping individual epigenetic data to genome wide patterns.)

Mapping DNAm using pattern recognition attempts to attenuate these problems. As well, it fits our hypothesis that MSP-related changes in DNAm will occur in a manner that will change the poise of multiple chromosome areas as opposed to modifying directly the transcription regulation of a handful of genes. Moreover, exploring patterns may be one way of overcoming our current ignorance of both the quantitative relation between MSP and DNAm changes or how those changes interact with the "neighbouring" DNAm context, either within a few kilobases or perhaps even on another chromosome. Regardless of how MSP is measured, (e.g. using self-report or cotinine levels), several studies suggest that a simple linear dose-response relation may not exist (Bauer *et al.*, 2016; Gao, L. *et al.*, 2018). At the molecular level, it is well established there is complex and poorly understood relation between DNA methylation and its functional impact on DNA accessibility and subsequent RNA transcription (for example, see consortia report from international analysis of 111 reference epigenomes in (Roadmap *et al.*, 2015).) Using data-driven pattern finding to derive a context based view of DNAm, we aim to extract patterns that depend on both the relation to MSP risk and relative DNAm levels across genomic sites on the microarray chip. As such, there are no assumptions from the researcher how a given exposure level relates to how or where DNAm shifts occur across the genome. While many epigenetic studies adopt the transcriptomic view of characterizing differential DNAm in terms of hypo- or hypermethylation, this assumes that the decrease or increase in methylation can alone and directly account for changes in the variable of interest. This may be reasonable in transcriptomic studies given that RNA expression of a gene has a relatively direct effect on its downstream molecular pathway(s). In contrast, DNAm occurs in both coding and non-coding regions and thus has a relatively more indirect but more wide-spread effect on regions of chromosomal activity. Pattern finding allows the researcher to avoid imposing this direct hypo/hypermethylation view and instead be able to freely explore the "what, where, how" of DNAm in the DOHaD framework.

2.4.1 Explore theories re: multidimensional data usage

Considering our goal of pattern recognition, we sought a multidimensional data mining technique that would best balance between the exploitation of genome-wide analysis of DNAm and the complex challenges posed by high-dimensional data, especially those unique to 450K data. As such, we sought a method that would fulfill the following features:

1. A distinguishing feature of our analysis is the simultaneous consideration of all sites that pass quality control parameters. We considered this to be an important criterion given the infancy of our current understanding of methylation and its effect on phenotype. We believe the

elimination of sites poses high and unpredicatable risk of eliminating useful information. In order to optimize the strength of high dimensions, we exercise extreme caution in eliminating or making unfounded assumptions regarding variables to correctly extract the most relevant information possible.

2. It will be computationally taxing to satisfy the above condition. It is for this reason that very intense filtering usually occurs before any data analysis. However, in order to explore data thoroughly, the selected method must use both time and computing resource feasible methods. This criterion is also highly related to criterion #1 because the use of high dimensional data have a major drawback – the analysis is prone to over-fitting. There are two primary means to combat over-fitting: variable selection and cross validation. Both these strategies, particularly the latter, require intense computer resources and time in order to test multiple iterations of data to select the best models.

3. Referring to Hypothesis 3b) and 4), the selected method must enable comparison between and within individuals over time, as well as between populations. As such, we require computational feasibility to upscale to other high dimensional and high throughput data, but also a method robust enough to an additional layer of data noise introduced when comparing different data sets with varying biological and technical sources of variability.

2.4.2 Pattern finding in high dimensional data

Data mining involves the recovery of data patterns that originate from signals that exist amidst data noise. This objective is usually divided into two branches: Unsupervised and supervised. Unsupervised refers to the use of the observed data set alone to extrapolate the signal source. In contrast, supervised refers to the experimenter introducing additional data to guide the recovery of relevant signals. Obviously, this experimenter-informed analysis has the benefit of better targeting the results towards the biological question of interest. However, this advantage is a double-edged sword: the major drawback of supervised analysis is greater risk of "overfitting" as previously discussed. We further discuss the pros and cons of various techniques below.

2.4.2.1 Unsupervised pattern finding

As discussed above in Section 2.3.2.2: Objective variable selection and mapping based on similarity of MSP vulnerability, PCA is a canonical example of unsupervised pattern finding.

PCA aims to capture the greatest data variability. In other words, it extracts components focused on the strongest signals. In microarray data, such strong signals can arise from sources like sex, cell type heterogeneity and technical artifacts. In addition, p>>n datasets are prone to false associations that arise at random that may be equal or actually stronger than true associations (Wang, Miller, & Clarke, 2008). As well, owing to the "curse of high dimensionality", outliers can become fortified by the additional data to become even more powerful influences compared to low-dimensionality settings as discussed in Section 1.4: *Mapping individuals to the risk context of MSP.*

However, we are specifically seeking a signal that represents a common but likely subtle difference in DNAm that links heterogeneous individuals along a shared MSP-related health trajectory. Moreover, the signal we seek is likely weaker than signals from noise. This may seem counter to traditional GWAS (or even EWAS) that seek the genomic loci that demonstrates the most significant signal (i.e. the smallest *p*-value after multiple hypothesis testing.) However, for DNAm related to CCD, the strongest signal is not necessarily our goal.

For this reason, we turned to explore the use of independent component analysis (ICA). Like PCA, it is an unsupervised data decomposition approach that has the benefit of extracting easily interpretable components. ICA is known as a "blind source separation" technique. It is based on the idea that the variability seen in data is actually the overlap of various sources of signal each with its own probability distribution. In other words, the data can be represented as a linear combination of statistically independent components. This is a particularly popular technique in the presence of large data noise where each isolated component can be checked for their correlation to the phenotype/exposure of interest versus known sources of noise. We trialed two R packages, JADE and fastICA. We found the components derived from ICA had limited relation to future child outcomes, (data available upon request.)

We speculate that unsupervised techniques like PCA and ICA may favour extracting unrelated signals, (such as strong known and unknown sources of confounding,) relative to subtle CCD signals that may be weaker (i.e. represent little data variability.) Alternatively, it may be difficult for unsupervised techniques to extract less organized signals. For instance, a single pathogenic process may involve various biological processes that overlap among those directly related to the stress and those that are merely bystanders. Therefore, a single pathogenic event may appear to emit signal from multiple sources. In these cases, it may require the guidance of clinical or other biologic data to tease apart signals related to a given trait from ubiquitous

bystanders like general inflammation. As such, it would seem that unsupervised approaches poorly fit our hypothetical framework to distill MSP-related DNAm patterns.

2.4.2.2 Supervised pattern finding

Supervised pattern finding in current DNAm literature can be distance based, pathway based or interdependence based (Teschendorff, Andrew E. & Relton, 2018). We briefly discuss these broad categories below.

Distance based methods

Distance based methods consider contiguity in DNA sequence distance and almost always in relation to the distance to gene or promoter. Several statistical tools are available to perform this analysis (examples of commonly used R packages include Min fi (Aryee *et al.*, 2014), ELMER (Silva *et al.*, 2019) and RnBeads (Müller *et al.*, 2019). However, an estimated four fifths of human genome transcription involves non-protein coding RNA (Kapranov *et al.*, 2007). As well, noncoding sequences are implicated in chromatin organization and transcription (Dhanasekaran, Kumari, & Kanduri, 2013). Thus, the likelihood that important smoking-related chromatin changes involve non coding intergenic spaces is high. This poses a number of problems. First, the 450K chip preferentially surveys protein coding genes and their promoters, as well as CpG rich regions. Second, these non-coding RNAs are often transcribed with coding sequences meaning the important areas of chromatin disruption may overlap both genic and intergenic regions (Hubé & Francastel, 2018). Third, different genomic regions have widely varying probe coverage in terms of distribution and density. For example, CpG-poor regions (not CpG-rich promoters, gene bodies, intragenic regions) have far lower probe density and higher probe sparsity than CpG-rich regions (Sandoval *et al.*, 2011). Fourth, methylation of probes within 500bp has high correlation (Geeleher *et al.*, 2013). Fifth, the distance to a regulatory element may have non-causal associations with nearby gene expression (Bauer *et al.*, 2016; Vives-Usano *et al.*, 2020; Xu, H., Zhang, Yi, Plewczynski, & Li, 2020). For example, only about 40% of enhancers regulate their nearest genes. This is believed to be a result of the mechanism of action: enhancers associate with promoters to upregulate transcription by forming chromatin loops that physically approximate genic regions to transcription factories (Margueron & Reinberg, 2010). This elaborate orchestration of chromatin structures involves TF complexes, cohesin extrusion and various other known and unknown mechanisms (Chambeyron & Bickmore, 2004; Chen, T. & Dent, 2014; Kazakevych, Sayols, Messner, Krienke, & Soshnikova, 2017). Thus, patterns arising from information such as distance to gene regulatory regions may risk being functionally non-contributory at our current state of knowledge.

Thus, it could be misleading to analyse all probes equally when they are unequally represented on the chip and unequally correlated to functional molecular activity. To address these issues, several techniques to analyse differentially methylated regions (DMRs) rather than single sites have been proposed. However, the majority of these again perform a pooled univariate analysis of single CpG probes (Butcher & Beck, 2015; Peters *et al.*, 2015), meaning individual comparison of probes is still being performed. Bacalini *et al.* proposed a solution whereby relatively lower density and thus less represented probes are analysed through single probe comparisons while higher density/better represented probes are analysed through region based comparison (Bacalini *et al.*, 2015). Based on the probe location relative to CpG islands and genes, this group divided all probes into 4 classes thought to have different epigenetic functions on gene expression and chromatin structure. They refer to these classes as "blocks of probes" (BOP). This group demonstrated that such a method increased the intersection of age-related DNAm differences identified in a meta-analysis of 3 separate 450K datasets relative to single-probe based ANOVA.

The approach by Bacalini and colleagues considers the inherent design of the array *before* analysis and modifies the analytic method accordingly. By accounting for the potentially differential effect of DNAm on chromatin activity based on location, the BOP method adjusts whether single versus region based analysis is performed. However, it remains that this approach relies on sequence and gene-centric annotation. Currently, we have very limited understanding of how the epigenome functionally flows to and from proximal (e.g. transcription products) and distal (e.g. chromatin loops,) levels. As well, no distance-based method can fully account for the 3-D proximity of a methylation site relative to another site or regulatory region to date. Thus, it can be risky to base analysis on current knowledge of how DNAm single sites or regions behave based on location or inferred function. Until integrative analysis of chromatin architecture, chromatin modifications and RNA expression can better annotate the inferred effect of DNAm on chromatin behaviour, this bias may either inflate or attenuate the importance of differential DNAm. For this reason, massive international efforts continue to juxtapose 3-D chromatin data with DNAm data as well as other genetic variants and gene regulatory marks (Stunnenberg *et al.*, 2016).

Pathway based methods

Pathway based methods are so called because they overlap the biological data of interest, (e.g. DNAm data,) with pathway databases that typically intersect molecular (e.g. transcriptome,

proteome, genome, etc.,) and clinical (e.g. disease associations, cell type, ethnicity, etc.,) features. Gene Set Enrichment Analysis (GSEA) is among the most popular of these methods. GSEA aims to find if certain DNAm site candidates are overrepresented more than would be expected by chance among sets of genes related by molecular and/or clinical features. Examples of available tools and approaches include (Pina, Pinto, Feijo, & Becker, 2005; Sofer *et al.*, 2012; Subramanian *et al.*, 2005; Tiong & Yeang, 2019; Trajkovski, Lavrac, & Tolar, 2008).

Despite using pathway data, GSEA does not exploit links between genes within a given pathway, such as their positions and roles and the directions and types of the signals transmitted from one gene to another. In contrast, topology-based pathway analysis is designed to integrate this information. In effect, they attempt to create "wiring diagrams". This can be more insightful when exploring downstream effects or understanding underlying mechanisms (Presson *et al.*, 2008). Since the development of topology-based pathway analysis, there has been an explosion of available tools, for example PathNet (Dutta, Wallqvist, & Reifman, 2012), GEPAT (Weniger, Engelmann, & Schultz, 2007) and NEO (Aten, Fuller, Lusis, & Horvath, 2008).

The drawback of any pathway analysis is that it requires a ranked list of candidate sites usually determined by calculating the delta-change between exposure and phenotype categories. This returns us to the same issues previously discussed in Section 2.3.2.1: *Integrating variables - Methodologic assumptions regarding variable relations* regarding researcher-imposed thresholds and discretization. Even within the transcriptomics field where pathway analysis was born, the fold-change[4] threshold is known to cause bias and inconsistency among data sets (Nguyen, Shafi, Nguyen, & Draghici, 2019; Yu *et al.*, 2017). The delta-change based ranking also presents a specific issue due to the design of the 450K chip. Using data from 13 human DNAm studies, Silva-Martinez and colleagues found that genes with a lower density of CpG sites have a higher false negative rate when testing for enrichment in disease- or tissue-specific function gene ontology categories (Silva-Martínez *et al.*, 2017). The authors conclude this is due to a link between likelihood of detecting a delta-change difference and the CpG density of a gene on the 450K chip that is a technical property and not biological. Moreover, the complex and yet relatively unknown relation between DNAm and biological function makes annotation based analysis less suitable than that used for expression microarrays (Harper, Peters, & Gamble, 2013; Huang da, Sherman, & Lempicki, 2009; Nguyen *et al.*, 2019; Subramanian *et al.*, 2005). Furthermore, most databases are disease-biased, especially to cancer (Huang da *et al.*,

[4] changes in expression are generally greater in expression studies thus the metric is usually fold-change rather than delta-change

2009; Nguyen *et al.*, 2019). This renders drawing inferences problematic especially for more subtle or pre-morbid health states such as in CCD.

Interdependence based methods

Interdependence based pattern recognition attempts to visualize DNAm in relation to the signal of interest without any *a priori* assumptions about where or to what degree DNAm differences will occur. A fascinating body of research using high-resolution analysis of chromatin structure and gene expression has mapped the context-specific role of DNAm on functional chromatin state (Collings & Anderson, 2017; Lay *et al.*, 2015). Without *a priori* restrictions of "where to look" on the 450K chip, we may be better able to detect DNAm changes that may be subtle or appear functionally unrelated but actually are part of a larger change in chromatin structure and functions. This global view may also help uncover the convergence of areas of differential methylation that may otherwise seem unrelated by previously reported features like physical genomic location, predicted role in gene expression or protein interactions or previously identified association in clinical disease.

Given that the observed DNAm data have a non-normal distribution, we focused on pattern recognition techniques that do not assume a specific data distribution, which is the case of methods such as Fisher's linear discriminant analysis. One of the most flexible classes among pattern finding methodologies is machine learning (ML). This class is popular in data mining research and includes methods such as random forests (RFs), artificial neural networks and support vector machines. It can be distinguished from classical estimation and classification statistical methods in that its modeling centres on an optimisation problem. This shift typically involves computationally intense algorithms; a shift enabled by now easily accessible high performance cluster (HPC) computer resources. We consider the "cost-accuracy trade-off" of using these newly emerging techniques. Computational cost is a critical limiting step in mining very high dimensional data sets. For example, many ML methods are optimized to analyse more than two datasets with a few hundred to thousand variables (Gao, C. *et al.*, 2018). After removal of low quality probes, our data set still has over 200,000 DNAm data points for each of over 900 subjects. Then each subject has data from three different ages. This equates to about 1.2 million data points per subject. Despite the use of HPCs, such analysis required days to weeks when they were trialed and some did not even complete in that time, (e.g. support vector machine using R Package e1071 by Meyer & Wien, 2015). As well, our primary objective was to overlap the phenotypic maps of individuals using biologic (i.e. DNAm) and clinical (i.e. MSP related vulnerability) data. Considering

again the cost-accuracy ratio, one may consider it “overkill” to overlap only two datasets using machine learning. As such, we experimented with four commonly used pattern recognition tool that have favourable cost-accuracy profiles when handling two datasets: tensor factorization (Hore *et al.*, 2016), canonical correlation analysis (mixomics) and partial least squares (PLS) (Gromski *et al.*, 2015). Among these, only the latter extracted results with relations to both MSP-related data and child outcome data. In the interest of space, we focus on this method moving forward in this work.

The goal of PLS analysis is to find the relation between two or more datasets, (in our case, between the MSP composite and DNAm sites. The two most common categories: 1) the symmetric form optimizes the covariance between datasets and 2) the asymmetric form aims to predict one dataset from the other(s). PLS is popular in various fields typically for classification purposes, such as in biochemical profiling in chemometrics, food authentication in quality control and tissue pathology and forensic evidence in medical science (Boulesteix & Strimmer, 2007; Durif *et al.*, 2018). It is robust to missing data, non-normality of residuals, noise and collinearity in context of adequate sample size (Wold *et al.,* 1983) and efficient for use in p>>n cases (Dennis & Forzani, 2018). Compared to other machine learning tools, PLS has the critical advantage of relatively easy interpretability. Given the exploratory nature of our investigation, interpretability is a major issue as our current understanding of relevant versus irrelevant (i.e. biologically uninteresting or confounding) patterns in DNAm is limited. As such, PLS was selected among other machine learning methods as the best fit considering our hypothesis and design rationale, data load, and feasibility based on computing resources.

PLS can be thought of as a regression extension of the previously discussed method, PCA. Their shared conceptual basis enhances the relative ease of use and interpretability of PLS. While PCA attempts to describe the maximum variation in the observed data, PLS aims to maximise the overlap of covariance projections between the data and the outcome. In other words, PCA returns features caused by the attribute with the biggest variance. In contrast, PLS returns features caused by the property under investigation.

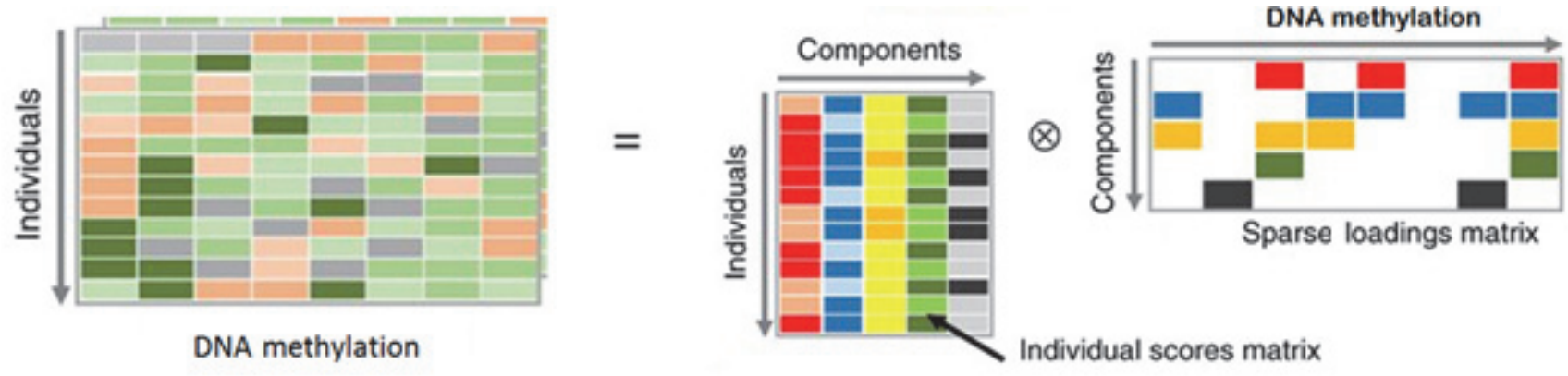


Figure 23: Graphical representation of decomposition of data into a two sparser matrices. Image adapted from Hore *et al.* (2016). We aim to use PLS to represent MSP related variability within DNAm data using a small number of components.

The output of PLS is to represent sources of variability within the high dimensional DNAm microarray data using a small number of components (Figure 23). PLS decomposes the DNAm data on the left of the equal sign of Figure 23 into the two matrices on the right. Referring to the matrices on the right of the equal sign, each consists of 2 dimensions that describe the relative contribution of the individual (left) or DNAm site (right) to that component, respectively. In our application, the goal of PLS is to generate components that represent a consistent pattern of DNAm at certain genomic sites that are related to MSP variability among subjects. We posit that these "vulnerability-informed" DNAm patterns as expressed as PLS components represent biologically meaningful differences in epigenetic poise.

As aforementioned, interpretability was a major consideration in selecting PLS. This is best exemplified using partial least squares discriminant analysis (PLS-DA) which is frequently used for classification of samples, (e.g. healthy versus diseased.) This is an extension of PLS which essentially adds two more procedures. The first is recoding continuous variables into categorical variables (*i.e.* ordinal or nominal). This involves the use of dummy coding (*i.e.* coding each variables as a combination of +1 and 0). The second is discriminant analysis to conduct prediction modeling. After PLS-DA decomposition, one can easily visualize the discrimination ability of the algorithm by observing how individuals cluster based on their DNAm scores.

There are a number of PLS packages available in R. Among those with applications in high dimensional biological data, these include the pls package (specifically the Canonical Powered-PLS (CPPLS) function) (Mevik & Cederkvist, 2004), the mixOmics package (specifically the PLS and PLSDA functions) (Cao-Lei, Elgbeili, Szyf, Laplante, & King, 2019) and the sgPLS package (Liquet, de Micheaux, Hejblum, & Thiebaut, 2016). The former employs an asymmetric PLS form while the latter two employs a symmetric form. We first explored the pls package as it is among the first R packages to offer the PLS algorithm as described by Svante Wold, the pioneer of nonlinear PLS modeling (Wold, Martens, & Wold, 1983). It allows parallelization which greatly assists in time spent during cross validation. It is very fast and easy to use. We used both the PLS and CPPLS functions, the latter of which appeared particularly applicable to our application as it allows the entry of multiclass data. It aims to better optimize the latent variables in context of high data noise by replacing covariance between the DNAm data and MSP data with canonical correlation in PLS methodology. This would theoretically sharpen the focus on relevant predictors. Previous work has used CPPLS for biomarker selection (Mehmood *et al.*, 2012). However, in the context of our DNAm data, this is not truly variable selection in that we could not find the loci that most strongly contributed to each pattern. The mixOmics package was very intuitive to use. However, the tuning and performance evaluation functions were very computing intensive, taking over 7 days using parallel computing on multiple CPUs. The sgPLS package offer tuning and performance functions that had less intense parallel computing resources.

2.4.2.3 Comparability

Pattern finding also aids in comparisons over time. The first step in PLS is to centre and scale values. Therefore, scores are always relative to differences in DNAm relative to other sites at the same time point. As such, comparison of scores at different times is truly a comparison of the relative pattern shift over time, rather than any changes of any single sites over time that may or may not be due to a biological signal of interest.

Pattern finding with PLS also facilitates comparisons among different data sets. Recall in Objective 3b and Objective 4 that we specifically aim to test the stability of our DNA signals over time within the same cohort and also across another population. Using the DNAm loadings (Figure 23), one has a "template" with which to re-create the patterns generated in one data set in a separate data set. Statistically, testing for temporal stability within the ARIES cohort and replicability in the GenR cohort are means of internal and external validation, respectively. Clinically, if the patterns are relevant to subject outcomes in more than one data set, that would lend support to the biological meaningfulness of DNAm at those specific genomic sites. In our case, that could mean children sharing DNAm patterns share similar developmental pathways related to common early-life exposures.

2.4.3 Objective variable selection and mapping based on MSP vulnerability

Just like in PCA, the loadings of the components identified by PLSDA provide information on the DNAm sites that form the basis for differentiation between components. Thus, as discussed in Section 2.3.2.2: Objective variable selection and mapping based on similarity of MSP vulnerability, variable selection in DNAm is again data driven. After testing multiple algorithms, we selected that found in the the sgPLS R package (Liquet *et al.*, 2016) based on feasibility (data resource requirements) for tuning and performance assessment.

This package offers the study of the relation between an omics dataset and a multivariate phenotype with simultaneous variable selection in a one-step strategy. This method was motivated by the increased biological relevance revealed by multi-dataset integrative approaches compared to analysis of data sets singly or in tandem (Liquet *et al.*, 2016; Loucoubar *et al.*, 2017).

Using the sparse PLS (sPLS) function, the sgPLS package performs variable selection using L_1 penalizations on the loading vectors of both the MSP and DNAm matrices (conceptually equivalent to the *T* and *P* vectors described for PCA (Section 2.3.2.2: Objective variable selection and mapping based on similarity of MSP vulnerability) and offers intuitive graphic functions that are provided in the mixOmics packaged discussed in the previous section.

The sparse PLS function (sPLS) in this package was easy to implement using our data. The package did not have parallelization capability at the time of this analysis. However, it still was faster to tune compared to the mixOmics package but far slower than the pls package. As well, this

method appears less prone to over-fitting as it conducts variable selection using cross validation to establish component-wise thresholds.

Using data from an transcriptomic and immune marker clinical trial, the authors of the sgPLS package compared the performance of their method to least absolute shrinkage and selection operator (LASSO) penalized regression methods, another very popular high dimensional method with variable selection that also uses cross validation for parameter selection. Results were comparable (data available upon request.) We did not pursue the use of LASSO or ridge regression as it has been shown in DNAm data to select sites with weak functional relations, particularly in mixed cell type tissues such as in our dataset (Zhong, Kim, Zhi, & Cui, 2019).

2.4.4 Summary of DNAm analysis

We posit that patterns reflect changes in DNA shape and thus function. As such, this work aims to identify DNAm patterns that relate to early-life exposures that alter chromatin activity. We use the MSP composite to bait DNAm pattern finding in cord blood because various human studies suggest that MSP-sensitive differential DNAm is already present at birth (Bauer *et al.*, 2016; de Vocht, Simpkin, Richmond, Relton, & Tilling, 2015; Joubert *et al.*, 2012; Joubert, Bonnie R. *et al.*, 2016; Miyake *et al.*, 2018). Furthermore, differential DNAm at birth has functional links later in life such as correlation with expression of target genes, phenotype and histone modifications (Bauer *et al.*, 2016).

We used the PLSDA function from the mixOmics package for categorical MSP variables and the sPLS function from the sgPLS package for the continuous MSP composite. To avoid confusion, we henceforth refer to PLS results as “components” and FAMD results as “dimensions”. PLS is prone to overestimate the accuracy of classification. To attenuate this issue, we conducted tuning to find the optimal parameters when possible. This included optimisation by M-fold CV and test grids to compare performance using various ranges of number of DNAm components, number of MSP dimensions and number of DNAm sites to retain. For components, we varied the number from one to 50 in increments of 10. We used the following performance metrics: predictive residual sum of squares (PRESS), R^2, mean squared error of prediction (MSEP) and proportion of DNA methylation variability captured. For MSP dimensions, we systematically trialled the number from two to 10 dimensions. In order to tune the number of DNAm variables to retain, we used the tuning.sPLS.X function in the sgPLS package that computes the MSEP of the PLS model. We used 10-fold CV and a grid of keepX values (representing number of CpG sites to retain) of 500, 1000, 2000, 5000 and 10000. For the majority of model, keepX optimized at a value of 1000. Previous work with PLS suggests that 5- to 10-fold CV offers a good balance between performance and computational efficiency (Mevik & Cederkvist, 2004). We performed training with 75% of total sample selected with the dplyr R package. This selection was random with the caveat that training was only performed on subjects with complete MSP data, which consisted of 901 of the 914 subjects with DNAm data.

As discussed in Section 2.4.2.3, we will extract PLS components from other data based on the PLS model from cord DNAm. To do so, we use the *predict* function of the sgPLS package. Similar to any other regression "prediction" technique, this function could be described roughly to calculate the component scores of the new data set based on the regression coefficients of each CpG site, (from the loadings matrix of Figure 23) projected onto the DNAm data matrix. Full mathematical details are in (Gonzalez, Cao, Davis, & Dejean, 2012; Liquet, de Micheaux, Hejblum, & Thiebaut, 2016).

2.5 Mapping methylation patterns to explore molecular mechanisms related to child outcomes

2.5.1 Considerations when mapping high dimensional data in complex traits

As discussed, EWASs have uncovered a great number of differentially methylated sites associated with disease, yet in context of our limited knowledge of the architecture of epigenetic changes, genetic background and disease manifestation, we cannot accurately predict disease risk from this wealth of epigenetic information. This is particularly challenging for complex disease which is fraught with influences such as genetic heterogeneity, epistasis (gene-gene interactions) and GxE that include multiple mechanisms besides DNA methylation. Traditional methods that have been used to analyze the epigenetic-disease associations are usually linear or comparative based, e.g. regression (commonly linear or logistic,) chi-square, etc., or variants of such tests. These methods test for main effects one variable at a time and cannot account for interactions when considering the relation between DNAm and exposure/phenotype (Moore *et al.*, 2010).

Another important consideration is that atypical outcome distributions challenge linear modeling. While there are strategies to "force" the variables to fit linear regression assumptions (e.g. transformation, binning, etc.,) there are often cases where this is simply inadequate. For example, Figure 24 shows the distribution of the responses to the Strengths and Difficulties Questionnaire (SDQ). This is a well-known measure of child behavior and was measured by both parents and teachers at up to six different time points in ALSPAC.

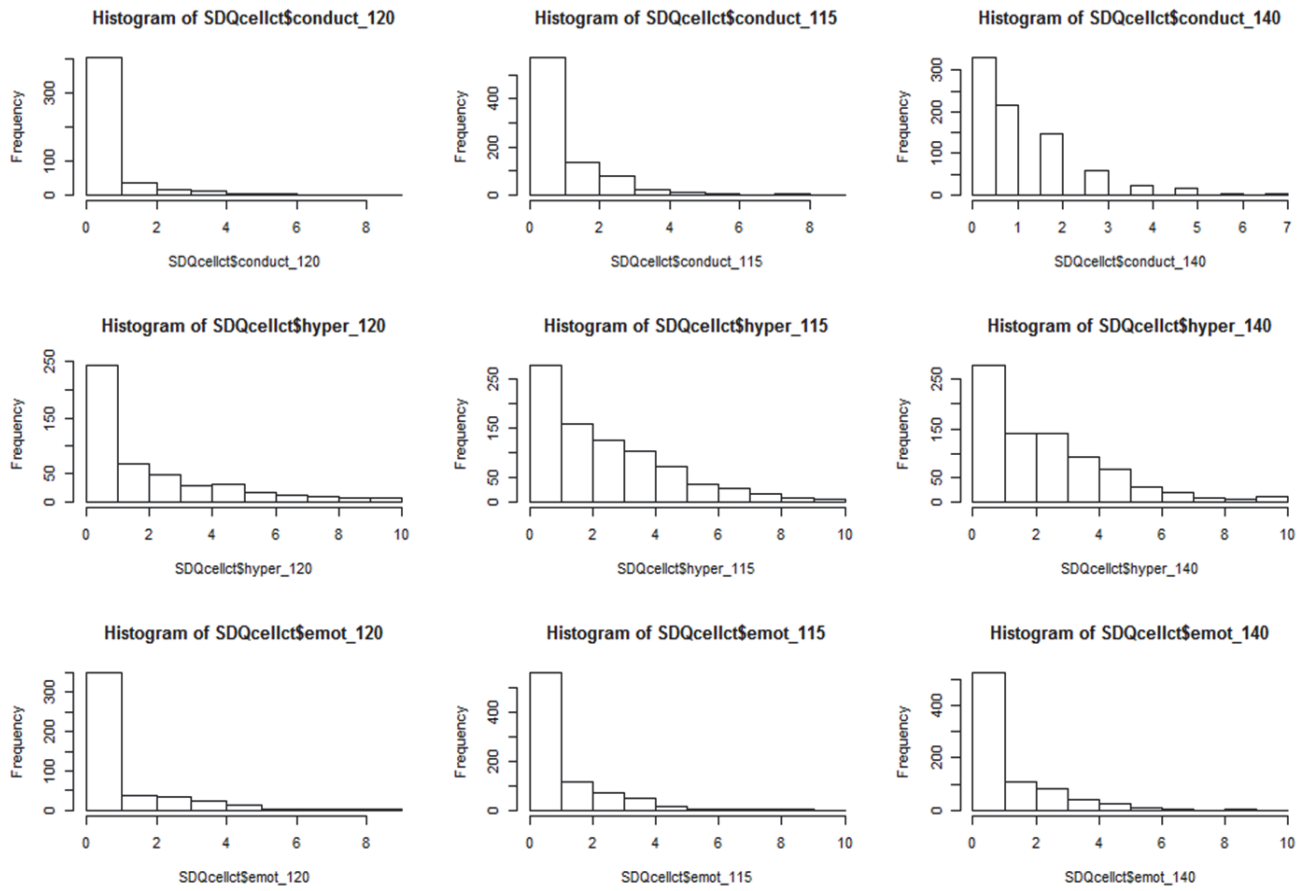


Figure 24: Distribution of scores on Strengths and Difficulties Questionnaire.

We trialed various type of transformations, including square root, cube root, and log (for the latter, we added a constant value of 1 to avoid undefined values from log(0).) We also tried the Box-Cox transformation (using the MASS R package.) There was no uniform transformation that normalized this data. We trialed different regression models for count data, including negative binomial, zero-inflated, and poisson. There was poor model fit in all trials.

In the interest of accounting for non-linear and non-main effect relations, as well as employing a uniform approach regardless of the distribution of the outcome, we moved from exploring regression based techniques to machine learning methods that could accommodate both these challenges.

2.5.2 Random forest selection in exploratory studies

Clearly, one must compromise between representing molecular reality and computational cost, as well as non-violation of statistical assumptions. Random forest (RF) learning is considered well-suited to achieve this balance in investigating the etiology of CCD using p>>n data (Degenhardt, Seifert, & Szymczak, 2019; Diaz-Uriarte & Alvarez de Andres, 2006; Kursa, Miron Bartosz, 2014; Li, J., Tran, & Siwabessy, 2016; Qi, 2012). Studies have shown that RF can outperform univariate tests when the number of relevant candidates is far smaller than irrelevant ones (Degenhardt *et al.*, 2019; Kursa, Miron Bartosz, 2014). As well, sensitivity to detect small marginal effects, particularly in context of noisy data (Lunetta, Hayward, Segal, & Van Eerdewegh, 2004; Scornet, Biau, & Vert, 2015). Moreover, univariate tests have no power when main effects are absent. An example of this

situation occurs in GWAS studies when genetic heterogeneity and SNP-SNP interactions result in no main effect (Lunetta *et al.*, 2004; Winham *et al.*, 2012). Genetic heterogeneity is particularly relevant in the complex diseases, where there are likely multiple pathways to acquire a trait that involve different subsets of genes/regions. Because RF can detect interactions (by collecting such insights in multiple different trees), it can outperform univariate testing in such situations (Breiman, 2001; Diaz-Uriarte & Alvarez de Andres, 2006; Qi, 2012; Rai, 2017). Importantly, RF can incorporate these higher-order interactions between the predictors yet is computationally efficient.

RF is a machine learning algorithm that can be used for ranking exploratory variables in high-dimensional data by their relative importance in predicting an outcome (Breiman, 2001). It falls under the family of tree-based methods that use recursive partitioning to conduct regression/classification, with subsequent collation of results from an ensemble of randomized samples. This family of methods are non-parametric and require no *a priori* assumptions regarding the relation between the predictors and outcome. This advantage is increasingly exploited in genome wide association studies when the genetic architecture of a trait is relatively unknown (Breiman, 2001; Degenhardt *et al.*, 2019; Genuer, Poggi, & Tuleau-Malot, 2015; Kursa, Miron Bartosz, 2014). A decision tree as seen in Figure 25 is constructed by step-by-step splits of the data using a sequence of hierarchical Boolean questions (e.g. testing whether Xi ≤ θj is true, where θj is a threshold value. For regression, this threshold value is the local average of the outcome values at each split.) These splits are called nodes. Current algorithms can comb through thousands of these questions efficiently. As such, it has become useful when screening large numbers of candidate predictors (e.g. candidate genes or methylation sites.)

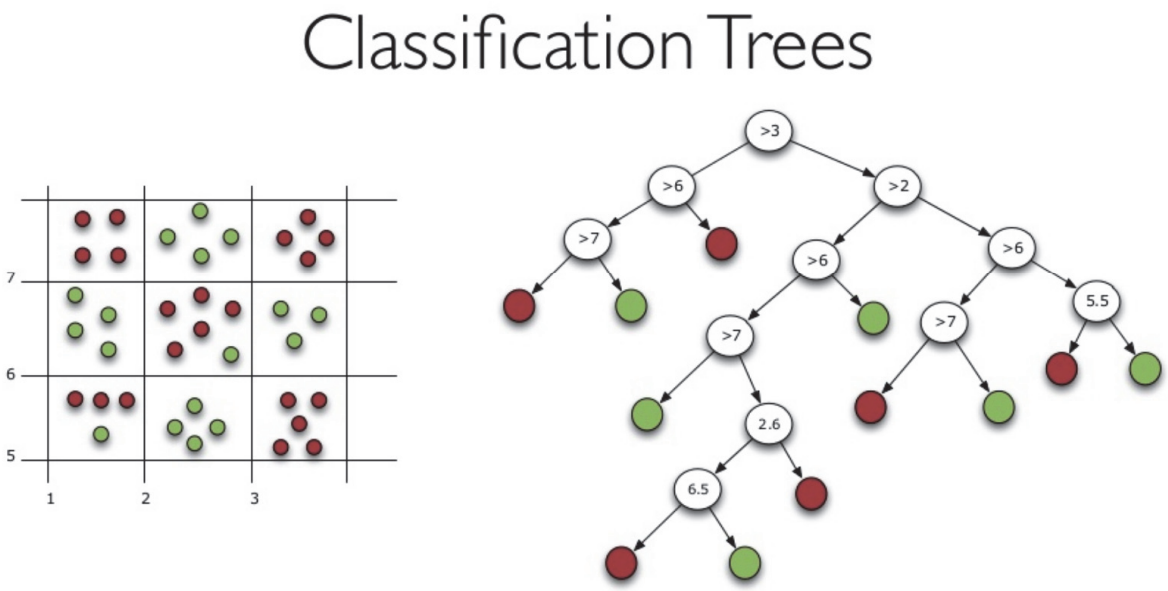


Figure 25: Schematic of classification tree "sorting" of observations. The grid on the left is drawn from the tree on the right. The axes reflect the different variable values that can be seen at each node (branch) of the tree. (Image source: https://www.solver.com/classification-tree.)

A single decision tree built as described will fit the data perfectly. If one altered the sampling of observations or the predictors entered into the model, then that likely alters the selections at each node (i.e. the answers to the Boolean questions) and thus the construction of an entirely different tree. As discussed in Section 2.2.1, this is called overfitting – where the model fits too well the specific set of data. While each tree that uses a different subset of observations or predictors could reveal a different relation and potentially important relation among predictors and outcomes, how does one know which tree to trust? Random forests offers a solution as an ensemble machine learning method that combines multiple trees as well as several different randomization methods to "vote" for the best model (Breiman, 2001) (Figure 26). By collating the results of hundreds to thousands of trees, RF is able to also provide a numerical estimate of variable importance.

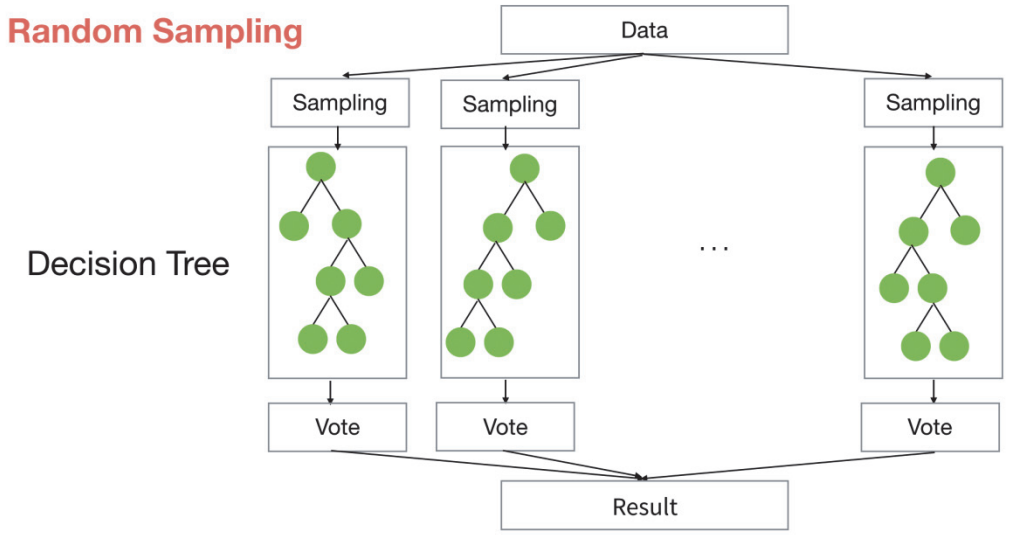


Figure 26: Random forest uses an ensemble of multiple decision trees. (Image source: https://towardsdatascience.com/random-forest-learning-essential-understanding-1ca856a963cb)

In this way, RF is a means of collecting insights from many different models but never relying on one model too heavily. It accomplishes this task through several features. First, RF ensures that all predictors have a similar chance of being selected by limiting the number of predictors that can be split at each decision node. As well, each tree accesses a random subset of predictors. In this way, not all the same predictors are grabbed at a given node at each tree. Second, RF grows each tree independently on its own bootstrap sample with replacement of the original observations. The "with replacement" means that in each sample, some subjects are represented multiple times, while others are left out. The left out subjects, (called "out-of-bag" or OOB,) are a built-in validation mechanism that can be used as the test sample to estimate performance. These features of recursive partitioning and local model fitting of potentially extreme values within independent decision trees make RF relatively robust to outliers, a major concern in heavily data dependent models (Karimpour-Fard, Epperson, & Hunter, 2015).

2.5.3 Random forest disadvantages

Despite the aforementioned advantages, there remain drawbacks of the RF approach. First, it remains more computationally intensive than univariate techniques. Second, no missing data are allowed. This is a critical issue in human cohorts compared to experimental models. Last, the interpretability compared with conventional regression analyses is more difficult, earning it a reputation of being a "black box" method. While regression models provide easy to understand metrics for predictors based on regression coefficients, the importance of a predictor as estimated by RF contains its complex interaction structure with all other predictors included (Qi, 2012). Consequently, it is difficult to estimate the direct, indirect and interactive associations between a predictor and outcome from RF models. However, this is a rapidly expanding field and better techniques to render RF more interpretable and efficient continue to evolve.

Based on our hypothesis, the DNAm patterns found in cord blood represent early differences in epigenetic programming. As discussed, previous research on MSP largely sought differentially methylated sites and regions and used effect size based calculations to determine clinical relevance. So far, the effect sizes or variability attributed to DNAm differences observed are typically low (Knopik *et al.*, 2019). This is despite the use of the extremes of MSP exposures in a number of studies in order to maximize effect. In contrast, our work considers the use of relatively subtle differences in positioning of MSP-related vulnerability to identify patterns. As well, it is arguable whether there is a linear relation between DNAm differences and various outcomes. Given these points, the utility of using effect-size based models is doubtful in the context of our hypotheses and data set.

2.5.4 Contrast with biomarker discovery

Both molecular mechanism and biomarker discovery studies involve feature selection to obtain the important players in disease pathology. Therefore, their technologies are closely related to each other. However, there is a subtle but critical difference in their respective goals regarding feature selection. The objective of biomarker discovery is to find a small set of biomarkers (e.g. genes or proteins) to achieve good prediction accuracies, usually by excluding redundant or "weak" predictors. This allows the development of more affordable and efficient diagnostic tests. In contrast, the goal of mechanistic exploration is to better understand the complete pathologic processes underlying disease. Thus, the aim is to uncover all the important players rather than building the best predictive model (Qi, 2012).

In machine learning, this contrast is captured by the difference between the "minimal-optimal" and "all relevant" problem. The former describes the phenomenon where accuracy is impaired when the number of variables is higher than optimal (Kursa, M., Rudnicki,W., 2010; Li, J. *et al.*, 2016). The solution to this problem is to find a small number of variables that provides the best possible prediction. This is helpful when looking for one or a small number of biomarkers that can best

classify disease status. In comparison, the latter problem aims to identify all predictors which relate to the outcome in at least one circumstance. In other words, accuracy is not the only criteria upon which a predictor is determined as important or not.

The Boruta algorithm is one of the most well-known algorithms in RF that attempts to solve the "all relevant" problem. It uses an entity called the "shadow variable" to gauge whether a variable is related to the outcome by more than chance. The shadow variable is a new variable based on existing variable values but is randomly shuffled such that it should bear no correlation to the outcome. It generates shadow variables with each RF tree and then collates the importance of this variable. Last, it compares the importance of this variable with the other "real" variables using a binomal test to statistically evaluate the likelihood of whether a given variable is more useful that the shadow variable. By having multiple shuffled copies of the shadow variable, Boruta is able to construct a distribution to compare the number of times the shadow variable performed better or worse than the other "real" variables (Figure 27). Boruta provides a robust estimate of the variable importance because it is designed to distinguish between variables that appear related to the outcome simply by chance (Kursa, Miron Bartosz, 2014). This is because the shadow variable can only be related to the outcome due to data stochasticity and therefore acts as an external reference of what may appear important due to chance. This gives an alternative and data independent criteria for selecting important variables besides prediction accuracy. As summarized by the authors, "adding randomness to the system and collecting results from the ensemble of randomized samples one can reduce the misleading impact of random fluctuations and correlations."

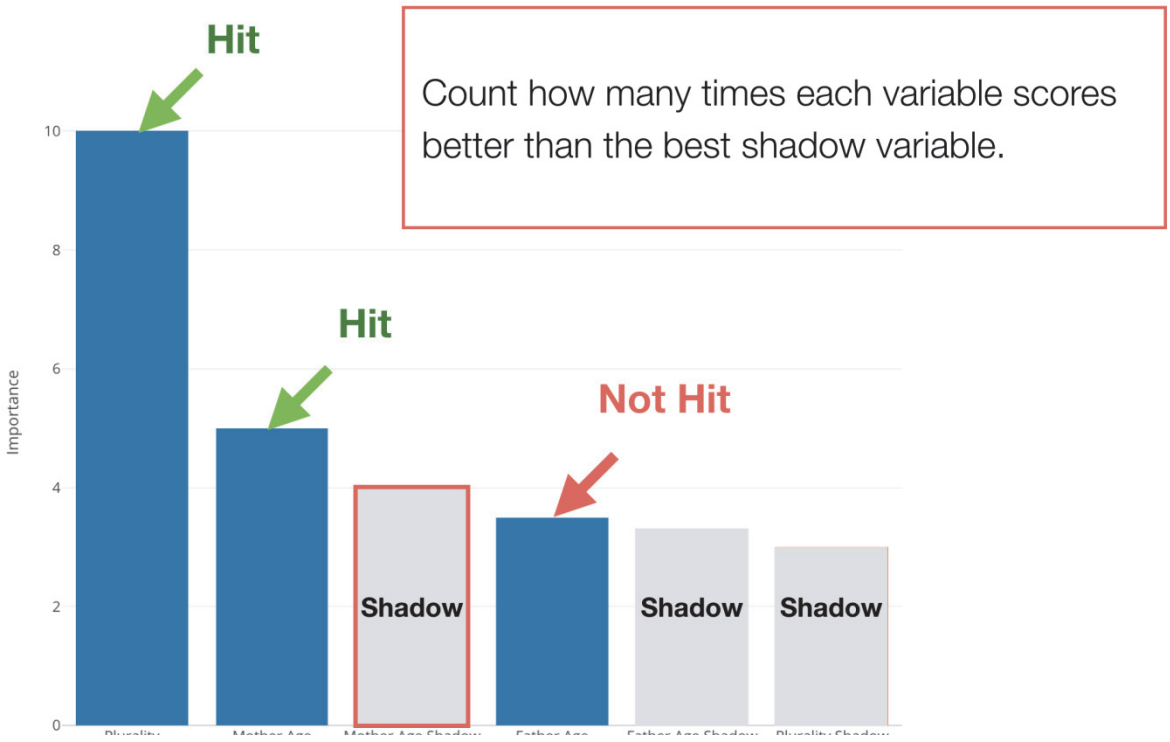


Figure 27: Example of the shadow variable and its relative importance compared to "real" variables. Variables ranking in importance lower than the shadow are no more likely related to the outcome than by chance.

2.5.5 Summary of child outcome analysis

We explore the relation between methylome patterns with child outcomes using RF family machine learning methods to explicitly model both main effects and interactions. Given our hypothesis that domains of DNA methylation interact with other genetic and non-genetic factors in biological networks underlying phenotype, this method fits well with our study aims. Among association analysis methods, it is relatively robust to major concerns that plague p >> n datasets in human cohorts, including overfitting and outlier sensitivity.

2.5.5.1 *Detecting relevant variables – Random forest analysis with Boruta pre-selection*

We apply the same two-step RF strategy as recommended in (Rai, 2017). These authors used simulated and real-life data to systematically compare pipeline and tuning parameters on high dimensional data. We briefly describe the process in the following. As well, some of the testing and tuning results will be referred to here instead of in the results section to streamline contents.

In Step 1, we used the Boruta algorithm (implemented using the R package of the same name performed with maximum 1000 iterations) to obtain all relevant predictors, including the covariates sex, socioeconomic indicators, and blood cell type estimates. We increased the *ntree* value to 1000 given previous work suggesting that high dimensional datasets may encounter unstable importance

scores and false negatives when *ntree* is inadequate (Kursa, Miron Bartosz, 2014). Component recurrence may support a more robust association with outcome as each model places all variables, including sex, social factors and relevant covariates in "competition" for relevance. Thus, components that are consistently selected across outcomes after thousands of iterations and re-samplings and are not related to covariates are considered for further examination.

In Step 2, we then evaluate the model using the RF the caret R package. We only use the variables selected by Boruta. This reduces computational load and has been found to increase model performance as well as attenuate overfitting (Li, J. *et al.*, 2016; Rai, 2017). We use the train function (set.seed(100), *ntree* = 1000) to conduct RF and resampling with 5 fold cross validation with 3 repeats. We used mean square error (MSE) and coefficient of determination (R^2) metrics to compare models with and without Boruta pre-selection of variables (Li, J. *et al.*, 2016).

2.5.5.2 ***Covariates***

In all models, we include infant sex and social factors (we selected maternal education and paternal social status). For anthropometric models, we also include control variables known to influence the outcome. For example, previous research has already demonstrated DNAm differences are related to body composition, both before and concurrent to DNAm sampling (see (Agha *et al.*, 2016; Cao-Lei *et al.*, 2019) for recent examples). As well, weight velocity in infancy is known to influence peri-pubertal fat mass, waist circumference and weight. We wish to avoid focusing on associations between the DNAm patterns and future outcomes that may be due to the tendency of certain phenotypes to demonstrate continuity as we age. Thus, the anthropometric models tend to have several more covariates than other outcome models. This is performed to hopefully distinguish DNAm differences that are due to MSP-related vulnerability. We also employ the randomForestSRC package to visualize results using partial dependence plots.

2.5.5.3 ***Tuning***

RFs are known to typically perform well with little tuning of parameters. That said, there are two main parameters that may be tuned. The first parameter we will look at is number of variables randomly selected at each split, (called *mtry* in caret). Consistent with previous biologic applications of RF, most notably by Breiman (2001), there is a minimal effect of *mtry* over a wide range of values surrounding the square root of the number of predictor variables (e.g. SNPs, methylation probes, etc.) (Kursa, M., Rudnicki,W., 2010).

The second parameter is the number of trees, (called *ntree* in caret.) Most literature focuses on prediction using *ntree* between 100–1000. Based on simulations using high dimensional data, (e.g. SNP data) to estimate variable importance where true associations may be far outnumbered by

those due to noise, one observes that more trees are needed to achieve stable estimates (Degenhardt *et al.*, 2019; Diaz-Uriarte & Alvarez de Andres, 2006). Consistent with previous evaluations of RF on both real life and simulated data, *ntree* was set to 1000 both for feature selection with Boruta and model estimation with randomForest (Kursa, M., Rudnicki,W., 2010; Lunetta *et al.*, 2004). This number of trees is about 10 times higher than defaults. We aimed for these settings to be sensitive enough to detect true predictors with small effects and interactions yet remain computationally feasible (Winham *et al.*, 2012). We evaluated changes to tuning parameters using the caret R package.

We also explored the effect of data partitioning between randomly selected training and testing sets (random seed = 100). We compared performance using R^2 and root mean square error (RMSE). A better model is implied by higher R^2 and lower RMSE. The default parameters were good if not the best options as seen in other studies (Diaz-Uriarte & Alvarez de Andres, 2006; Li, J. *et al.*, 2016; Rai, 2017) however the differences among changed settings was marginal (data available upon request) so we used default settings for the sake of reproducibility in all analysis.

2.5.5.4 ***Model stability***

We further checked the stability of the results by rerunning the analysis twice with different random seeds as in (van der Meer *et al.*, 2017). In addition, we compared the relative ranking of relevant predictors by Boruta and by RF after Boruta selection.

Further, to assess whether our findings were due to specific and unknown properties of our data and the Boruta algorithm, we compared models using RF alone without Boruta. We used variance explained and error rate on OOB samples as performance metrics. We performed sensitivity analysis by comparing results with and without data imputation.

2.6 Mapping methylation patterns to explore molecular relevance

One of the biggest challenges in the study of complex diseases and epigenetics is how to link clinical and molecular relevance. Depending on the data set size and analytic method employed, dozens to hundreds of “significant” DNAm sites can be found related to infant exposure to MSP. This has been repeatedly shown in numerous cohorts around the world from various ethnic groups, (for reviews, see (Joubert *et al.*, 2016; Knopik *et al.*, 2019; Taal *et al.*, 2013).) What is the relation, if any, between the significant CpG “hits” and how do these interact with molecular mechanisms that can plausibly lead to altered phenotype?

Most studies borrow from the gene expression field and perform pathway analysis. This method combines large molecular interaction databases and statistical testing to link hits with annotated gene products, regulatory mechanisms, biological process and/or correlations with clinical disease. This testing is known to be prone for false positive bias and researchers have developed numerous correction and resampling techniques to overcome this problem (for examples, see (Geeleher *et al.*, 2013; Silva-Martínez *et al.*, 2017; Yu *et al.*, 2017). This is particularly an issue with the 450K chip given it was designed to over-represent genic sites with known or inferred relevance to disease (especially cancer) and key biological processes.

Given the costs of animal models and multi-omic human studies, the intermediate means to interrogate functionality in the case of CCD and epigenetics remains actively explored. In the meantime, in order to better understand the possible molecular implications of the DNAm patterns suggested by the components, we venture to consider the overlap between our DNAm patterns and chromatin activity and topology, through both experimentally derived and computationally inferred means.

2.6.1 Mapping patterns to chromatin activity

The global collaborative efforts of the NIH Roadmap Epigenomics Mapping Consortium (referred to as the NIH Roadmap hereafter) have created an immense reference data set for various types of epigenetic data in various tissues and species (Roadmap *et al.*, 2015). The NIH Roadmap defined 15 chromatin states based on genome-wide histone modification patterns and CCCTC-Binding Factor (CTCF – also known as 11 Zinc Finger Transcriptional Repressor) binding patterns (Table 4). By combining these patterns, they have demonstrated that these states are associated with differential DNA methylation and accessibility (see

Figure 7) (Chen, T. & Dent, 2014; Geiman & Robertson, 2002; Rye *et al.*, 2014; Tchasovnikarova & Kingston, 2018; Zhang, L. *et al.*, 2017; Zhu *et al.*, 2016). For example, certain states are related with relative chromatin activity because they are promoter associated, (e.g. H3K4me1, H3K9ac, H3K4me3, and H3K27ac.) Other states are linked to inactivity as repressive marks (e.g. H3K27me3) are widespread in those regions.

Table 4: Chromatin state definitions and abbreviations. Image from Roadmap *et al.* (2015).

STATE NO.	MNEMONIC	DESCRIPTION
1	TssA	Active TSS
2	TssAFlnk	Flanking Active TSS
3	TxFlnk	Transcr. at gene 5' and 3'
4	Tx	Strong transcription
5	TxWk	Weak transcription
6	EnhG	Genic enhancers
7	Enh	Enhancers
8	ZNF/Rpts	ZNF genes & repeats
9	Het	Heterochromatin
10	TssBiv	Bivalent/Poised TSS
11	BivFlnk	Flanking Bivalent TSS/Enh
12	EnhBiv	Bivalent Enhancer
13	ReprPC	Repressed PolyComb
14	ReprPCWk	Weak Repressed PolyComb
15	Quies	Quiescent/Low

Using this repository, we mapped each components' representative CpG sites to their chromosomal location and its predicted chromatin states. To avoid bias due to usage of a specific cell reference, we separately conducted analysis using 3 different cell types: Primary B and T cells from cord blood (reference E031 and E033, respectively) and primary mononuclear cells from peripheral blood (reference E062.)

2.6.2 Mapping methylation to chromatin topology

As discussed in Section 1.5: Mapping individual epigenetic data to genome wide patterns and as shown in Figure 8 and Figure 13, chromosomal architecture is hierarchically organized in an elaborate looping structure that mediates gene expression by decreasing or increasing likelihood of physical contact between distantly located *cis*-regulatory elements and transcription factories. Recently, decreasing costs and increasingly high throughput of 3D chromosome analysis technologies have improved chromatin interactome maps in terms of resolution, genome coverage and tissue and species coverage (Jung *et al.*, 2019; Mishra & Hawkins, 2017).

To illustrate the utility of chromatin interactome data in linking disease to genetic data, we now show an example of how a small disruption can dramatically alter the physical shape and therefore function of chromatin. Using SNP data from 26 human populations and combining this with 3-D

chromatin structure data from a human cell line, Sadowski and colleagues demonstrated that the deletion shown in Figure 28 leads to significant changes in transcription of multiple genes, including an increase in Protein Phosphatase 2 Regulatory Subunit B (PPP2R3C) expression (Sadowski *et al.*, 2019). This gene is associated with hematologic and reproductive diseases. Moreover, the expression changes occur even though the deletion is entirely in the intronic (i.e. non-coding) region of the KIAA0391 gene. The authors posit that the deletion increases PPP2R3C transcription by disrupting a chromatin anchor point, thus changing the loop structure and "freeing" the PPP2R3C promoter to make contact with enhancers (see third row, Figure 28).

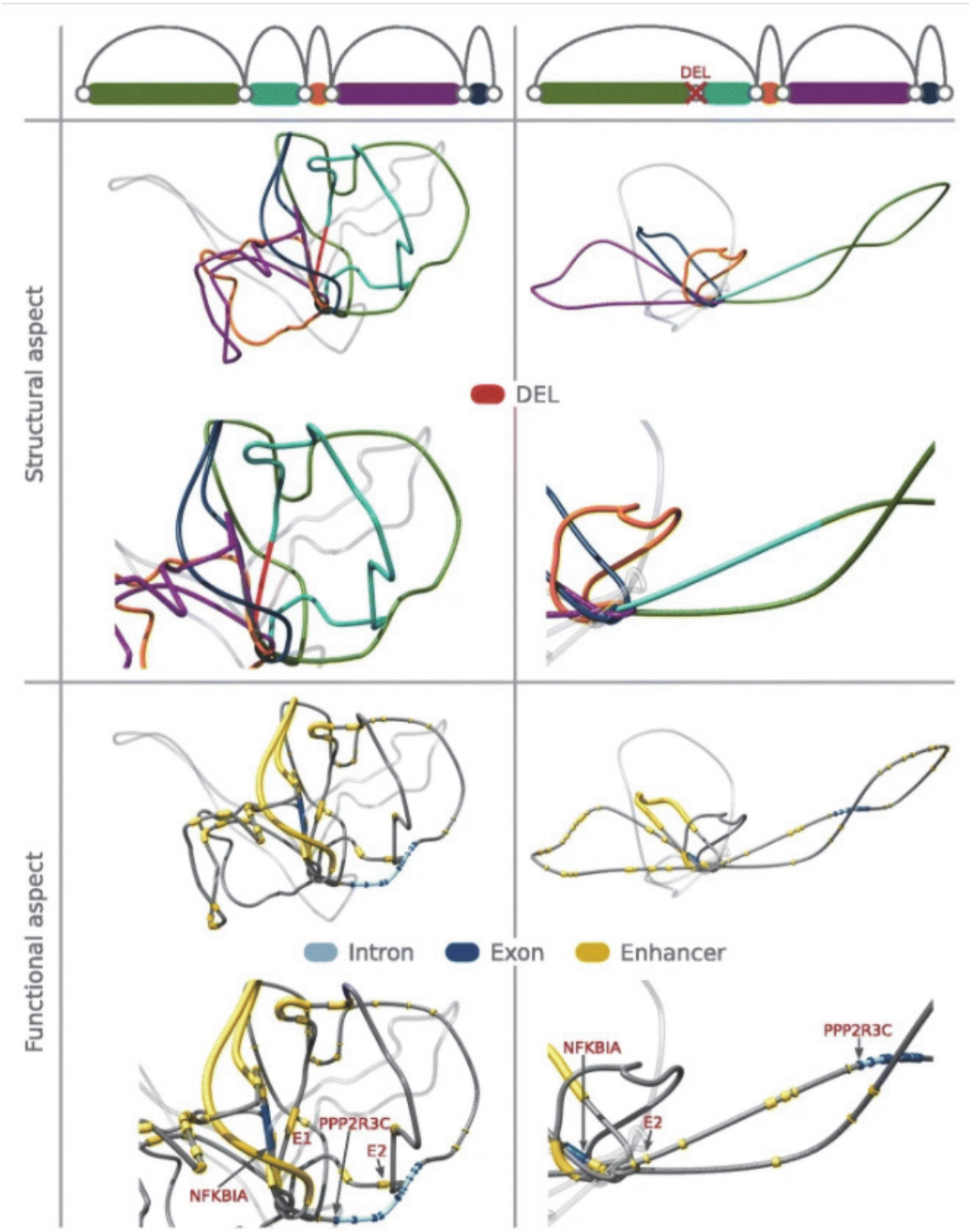


Figure 28: 3D models of the impact of a deletion (chr14:35605439-35615196) located in an intron of KIAA0391. While the deletion involves less than 10kB, it leads to a 3D structural change that completely obliterates a CTCF anchor. Left column: without deletion. Right column: with deletion. First row: linear schematic of chromosome with loops colored as on images in second row. Second row: 3D models of structural change. Third row: 3-D

model with genes and intronic, exonic and enhancer regions marked. Arrows indicate TSSs. Image from (Sadowski *et al.*, 2019).

Though the above example is for a sequence deletion, it demonstrates the vulnerability of chromatin to and the feasibility of small changes in non-coding regions to disrupt function by altering interactions between other chromosome sites and transcription regulation machinery. It also provides a potential explanation for a recurrent, over decade old observation in GWAS data: trait-associated variants are often in non-coding regions, sometimes far removed (in terms of linear genomic distance,) from the putative gene targets.

Because of work of groups like above, there is growing interest in mapping 450K data to chromatin anchor and loop structures. Wu and colleagues recently overlapped 450K data with various 3C-based datasets to generate a genome wide map of inferred promoter-anchored chromatin interactions (PAIs) (Wu, Y. *et al.*, 2020). Using this map, we were specifically interested in PAIs that overlapped with chromatin loops and topological associated domains (TADs) (data obtained directly from authors) given such dual associations may increase the likelihood of functional implications. However, given the small number of PAI with TAD overlap in their dataset, we looked at PAI and loop overlap only, (PAI and loop overlap totaled to 130 sites.)

2.6.3 Summary of methylation pattern mapping

We approach DNAm as an indicator of the structural adaptation of chromosomal regions to the net effect of intrinsic and extrinsic cellular forces. These adaptions lead to altered chromatin activity. To this end, we used the regioneR R package to estimate enrichment in chromatin activity or structure using permutation testing. Enrichment of overlap was defined as the number of overlaps between the component-representative sites and the features of interest relative to what would be the expected overlap with whole genome and the feature of interest. We employed the strategy where each feature set is circularly randomized within each chromosome. The overlap distribution in 10,000 random circular permutations was used to compute a *P*-value by comparing the observed value with the null distribution (using the human autosomal genome as provided in BSgenome.Hsapiens.UCSC.hg19.masked: Full masked genome sequences for Homo sapiens - UCSC version hg19. R package version 1.3.99.)

2.7 Replication

The ALSPAC and GenR cohort data were acquired in entirely independent studies. Thus, not all the covariates collected were homologous. Even with features in common between the discovery and replication cohorts, covariates can have substantially different distributions (Gao, C. *et al.*, 2018).

As such, we used covariates known to impact the child outcomes based on previous research in the GenR cohort.

Though there will be unavoidable differences in biological sample handling and DNAm processing, previous work performing epigenomic meta-analysis involving both ALSPAC and GenR have found results to be comparable (Joubert *et al.*, 2016; Sharp *et al.*, 2018). Moreover, the successful recovery of similar signals in GenR despite these technical differences would support the robustness of the findings. In fact, a recent study using 3 cohorts used machine learning to create a DNAm score (Rauschert *et al.*, 2020). This group actually used raw beta values only and found performance comparable to previous work that performed data harmonization. As such, there was no attempt to harmonize the pre-processing of DNAm of the cohorts except for low variance filtering and batch correction.

2.8 Disadvantages of pattern finding

In the end, pattern finding can suffer from statistical "mirages" just as candidate site/region studies do – it is a reality of crunching literally hundreds of thousands of numbers. Patterns are detected based on the statistical association between epigenetic changes and the subjects' exposure and/or phenotype matrix. Thus, the estimation of this association may be prone to false positives just as in candidate studies. In candidate studies, there is a widely accepted false positive correction technique. This is coupled with support of biological plausibility that is often tested through functional annotation testing, which is also subject to multiple hypothesis testing and thus also requires false positive correction. Associations estimated in this manner form the bulk of findings published in the DNAm literature surrounding maternal smoking and child outcomes (Knopik *et al.*, 2019).

In contrast, there is no such broadly accepted means of false positive correction in RF. Nor is there a specific technique for testing biological plausibility in pattern seeking. Thus, pattern seeking is better suited for hypothesis generation than hypothesis testing at this juncture. Despite these drawbacks, epigenetic studies of CCD are in dire need of exploration to make the knowledge translation leap. While the state of the art evolves and better defines what is a biologically and statistically robust pattern finding result, this work ventures to explore what contributions other methods besides traditional candidate studies can provide.

Chapter 3 Results

3.1 Mother-child characteristics

Among ALSPAC subjects, 15211 newborns have at least one item of collected data. Data from all these newborns and their mothers were included in composite construction.

DNAm data was processed on samples from a subset of 1018 children-mother pairs that form the ARIES project. In total, 914, 973 and 974 children had DNA methylation measured from cord blood (at birth), at around age 7 and age 17, respectively. Previous studies have compared the baseline characteristics of families in the overall ALSPAC versus the ARIES dataset (Appendix B). Compared with the ALSPAC mothers, the ARIES subset mothers were older, reported less MSP, attained a higher educational level and more likely employed in non-manual labour. The MSP-relevant features of 15211 infant-mother pairs used in this work is tabulated in Table 5 (see column labelled “Total”.)

Table 5: Descriptive statistics by maternal smoking in pregnancy classification

Sex	None (N=8754)	Periconception only (N=963)	2 or more trimesters (N=1898)	Throughout pregnancy (N=1815)	Unknown (N=1781)	Total (N=15211)	p - value
							0.044
N-Miss	22	1	7	0	487	517	
Male	4460 (51.1%)	476 (49.5%)	1019 (53.9%)	968 (53.3%)	650 (50.2%)	7573 (51.5%)	
Female	4272 (48.9%)	486 (50.5%)	872 (46.1%)	847 (46.7%)	644 (49.8%)	7121 (48.5%)	
Ethnicity							< 0.001
N-Miss	165	21	99	23	683	991	
Caucasian	8166 (95.1%)	891 (94.6%)	1664 (92.5%)	1705 (95.1%)	981 (89.3%)	13407 (94.3%)	
Other	423 (4.9%)	51 (5.4%)	135 (7.5%)	87 (4.9%)	117 (10.7%)	813 (5.7%)	
Gestational Age							< 0.001
N-Miss	2	0	1	0	589	592	
Mean (SD)	39.297 (2.401)	39.467 (2.122)	38.998 (3.031)	39.421 (1.805)	28.018 (13.761)	38.365 (5.504)	
Range	9.000 - 44.000	18.000 - 47.000	10.000 - 46.000	26.000 - 46.000	4.000 - 45.000	4.000 - 47.000	
Birth weight, z-score for gestional age							< 0.001
N-Miss	158	13	46	18	1085	1320	
Mean (SD)	0.495 (1.027)	0.589 (1.017)	0.235 (1.091)	0.032 (1.042)	0.189 (1.119)	0.392 (1.057)	
Range	-6.398 - 4.245	-3.585 - 3.729	-4.101 - 6.306	-3.278 - 3.248	-3.751 - 3.632	-6.398 - 6.306	
Birth weight, internal z-score							< 0.001
N-Miss	162	12	70	10	793	1047	
Mean (SD)	0.045 (0.504)	0.104 (0.502)	-0.081 (0.544)	-0.129 (0.476)	-0.061 (0.459)	0.003 (0.507)	
Range	-2.364 - 1.988	-2.051 - 1.772	-2.268 - 1.531	-2.165 - 1.380	-2.195 - 1.603	-2.364 - 1.988	
Birth length, internal z-score							< 0.001
N-Miss	245	26	109	23	813	1216	
Mean (SD)	0.109 (1.595)	0.385 (1.577)	-0.227 (1.677)	-0.393 (1.611)	-0.144 (1.378)	0.003 (1.606)	
Range	-9.251 - 5.474	-6.650 - 5.346	-7.624 - 5.133	-8.560 - 4.655	-5.793 - 4.091	-9.251 - 5.474	
Gestational weight gain							< 0.001
N-Miss	1937	235	738	410	1608	4928	

	None (N=8754)	Periconception only (N=963)	trimesters (N=1898)	pregnancy (N=1815)	Unknown (N=1781)	Total (N=15211)	*p*-value
Over	1771 (26.0%)	277 (38.0%)	380 (32.8%)	334 (23.8%)	48 (27.7%)	2810 (27.3%)	
Recommended	2749 (40.3%)	278 (38.2%)	421 (36.3%)	491 (34.9%)	66 (38.2%)	4005 (38.9%)	
Under	2297 (33.7%)	173 (23.8%)	359 (30.9%)	580 (41.3%)	59 (34.1%)	3468 (33.7%)	
Maternal grandmother - ever smoked							
N-Miss	455	29	180	52	1781	2497	
FALSE	4280 (51.6%)	562 (60.2%)	1139 (66.3%)	1234 (70.0%)	0	7215 (56.7%)	
TRUE	4019 (48.4%)	372 (39.8%)	579 (33.7%)	529 (30.0%)	0	5499 (43.3%)	
Maternal grandmother - smoked while pregnant							
N-Miss	496	32	190	55	1781	2554	
Don’t know	1178 (14.3%)	152 (16.3%)	260 (15.2%)	254 (14.4%)	0	1844 (14.6%)	
FALSE	5543 (67.1%)	610 (65.5%)	900 (52.7%)	809 (46.0%)	0	7862 (62.1%)	
TRUE	1537 (18.6%)	169 (18.2%)	548 (32.1%)	697 (39.6%)	0	2951 (23.3%)	
Maternal grandfather - ever smoked							
N-Miss	297	37	258	138	1781	2511	
FALSE	6383 (75.5%)	482 (52.1%)	581 (35.4%)	483 (28.8%)	0	7929 (62.4%)	
TRUE	2074 (24.5%)	444 (47.9%)	1059 (64.6%)	1194 (71.2%)	0	4771 (37.6%)	
Mother’s partner - smoked while pregnant							
N-Miss	290	12	161	34	1781	2278	
FALSE	8125 (96.0%)	863 (90.7%)	1477 (85.0%)	1555 (87.3%)	0	12020 (92.9%)	
TRUE	339 (4.0%)	88 (9.3%)	260 (15.0%)	226 (12.7%)	0	913 (7.1%)	
Others who smoke in household							
N-Miss	554	45	236	106	1781	2722	
FALSE	5992 (73.1%)	731 (79.6%)	1363 (82.0%)	1436 (84.0%)	0	9522 (76.2%)	
TRUE	2208 (26.9%)	187 (20.4%)	299 (18.0%)	273 (16.0%)	0	2967 (23.8%)	
Maternal education level a time of pregnancy							< 0.001
N-Miss	531	76	388	120	1534	2649	
Non-degree	7359 (89.5%)	763 (86.0%)	1186 (78.5%)	1278 (75.4%)	192 (77.7%)	10778 (85.8%)	
Degree	864 (10.5%)	124 (14.0%)	324 (21.5%)	417 (24.6%)	55 (22.3%)	1784 (14.2%)	

	None (N=8754)	Periconception only (N=963)	trimesters (N=1898)	pregnancy (N=1815)	Unknown (N=1781)	Total (N=15211)	*p*-value
Maternal financial concerns							< 0.001
N-Miss	755	105	432	186	1581	3059	
No strain	7445 (93.1%)	777 (90.6%)	1216 (82.9%)	1342 (82.4%)	163 (81.5%)	10943 (90.1%)	
Strain	554 (6.9%)	81 (9.4%)	250 (17.1%)	287 (17.6%)	37 (18.5%)	1209 (9.9%)	
Maternal psychopathology							< 0.001
N-Miss	246	41	107	78	1580	2052	
Denies	6749 (79.3%)	644 (69.8%)	1101 (61.5%)	1087 (62.6%)	141 (70.1%)	9722 (73.9%)	
Concerns	1759 (20.7%)	278 (30.2%)	690 (38.5%)	650 (37.4%)	60 (29.9%)	3437 (26.1%)	
Maternal substance use in pregnancy							< 0.001
N-Miss	34	0	1	0	1403	1438	
Denies	8365 (95.9%)	914 (94.9%)	1736 (91.5%)	1630 (89.8%)	377 (99.7%)	13022 (94.5%)	
Concerns	355 (4.1%)	49 (5.1%)	161 (8.5%)	185 (10.2%)	1 (0.3%)	751 (5.5%)	
Neighbourhood quality, ascending quality							< 0.001
N-Miss	624	57	280	86	1170	2217	
0	6 (0.1%)	0 (0.0%)	3 (0.2%)	11 (0.6%)	2 (0.3%)	22 (0.2%)	
1	20 (0.2%)	2 (0.2%)	9 (0.6%)	9 (0.5%)	5 (0.8%)	45 (0.3%)	
2	54 (0.7%)	7 (0.8%)	34 (2.1%)	37 (2.1%)	10 (1.6%)	142 (1.1%)	
3	113 (1.4%)	24 (2.6%)	57 (3.5%)	69 (4.0%)	19 (3.1%)	282 (2.2%)	
4	387 (4.8%)	53 (5.8%)	118 (7.3%)	132 (7.6%)	59 (9.7%)	749 (5.8%)	
5	331 (4.1%)	58 (6.4%)	131 (8.1%)	120 (6.9%)	41 (6.7%)	681 (5.2%)	
6	617 (7.6%)	85 (9.4%)	167 (10.3%)	192 (11.1%)	59 (9.7%)	1120 (8.6%)	
7	849 (10.4%)	83 (9.2%)	220 (13.6%)	218 (12.6%)	64 (10.5%)	1434 (11.0%)	
8	1102 (13.6%)	113 (12.5%)	231 (14.3%)	230 (13.3%)	92 (15.1%)	1768 (13.6%)	
9	1670 (20.5%)	200 (22.1%)	269 (16.6%)	320 (18.5%)	90 (14.7%)	2549 (19.6%)	
10	1993 (24.5%)	185 (20.4%)	238 (14.7%)	260 (15.0%)	104 (17.0%)	2780 (21.4%)	
11	872 (10.7%)	87 (9.6%)	125 (7.7%)	122 (7.1%)	55 (9.0%)	1261 (9.7%)	
12	116 (1.4%)	9 (1.0%)	16 (1.0%)	9 (0.5%)	11 (1.8%)	161 (1.2%)	
Maternal social status							< 0.001

	None (N=8754)	Periconception only (N=963)	trimesters (N=1898)	pregnancy (N=1815)	Unknown (N=1781)	Total (N=15211)	*p*-value
N-Miss	1822	243	800	580	1651	5096	
1	527 (7.6%)	28 (3.9%)	19 (1.7%)	17 (1.4%)	5 (3.8%)	596 (5.9%)	
2	2330 (33.6%)	203 (28.2%)	311 (28.3%)	300 (24.3%)	36 (27.7%)	3180 (31.4%)	
3	3009 (43.4%)	318 (44.2%)	450 (41.0%)	497 (40.2%)	52 (40.0%)	4326 (42.8%)	
4	443 (6.4%)	76 (10.6%)	112 (10.2%)	146 (11.8%)	14 (10.8%)	791 (7.8%)	
5	515 (7.4%)	85 (11.8%)	169 (15.4%)	210 (17.0%)	18 (13.8%)	997 (9.9%)	
6	106 (1.5%)	10 (1.4%)	37 (3.4%)	63 (5.1%)	5 (3.8%)	221 (2.2%)	
65	2 (0.0%)	0 (0.0%)	0 (0.0%)	2 (0.2%)	0 (0.0%)	4 (0.0%)	
Paternal social status							< 0.001
N-Miss	1218	184	689	447	1634	4172	
1	1053 (14.0%)	44 (5.6%)	57 (4.7%)	43 (3.1%)	8 (5.4%)	1205 (10.9%)	
2	2790 (37.0%)	253 (32.5%)	322 (26.6%)	340 (24.9%)	44 (29.9%)	3749 (34.0%)	
3	858 (11.4%)	101 (13.0%)	120 (9.9%)	110 (8.0%)	10 (6.8%)	1199 (10.9%)	
4	2047 (27.2%)	268 (34.4%)	499 (41.3%)	590 (43.1%)	60 (40.8%)	3464 (31.4%)	
5	620 (8.2%)	86 (11.0%)	154 (12.7%)	200 (14.6%)	18 (12.2%)	1078 (9.8%)	
6	156 (2.1%)	22 (2.8%)	53 (4.4%)	79 (5.8%)	6 (4.1%)	316 (2.9%)	
65	12 (0.2%)	5 (0.6%)	4 (0.3%)	6 (0.4%)	1 (0.7%)	28 (0.3%)	

Subject demographics. Columns 1-5 represent maternal smoking categories. Reported *p*-value: Continuous variables – ANOVA, categorical – Chi squared. "Miss" – missing data. Social status was derived from reported occupation according to the UK Registrar General's classification. From 1 to 3, the occupations refer to manual unskilled, semi-skilled manual and skilled occupations; 4 refers to skilled non-manual occupations; 5 refers to managerial and technical occupations; 6 refers to professional occupations and 65 refers to armed forces. Education – "Degree" refers to had a university degree at time of index pregnancy.

In Methods (Section 2.3: Mapping clinical data), we considered differences between categorical versus continuous descriptions of MSP. In the remainder of this section, we

explore the implications of using these two methods when describing the phenomenon of exposure to MSP.

3.1.1 MSP vulnerability using self-reported maternal smoking categories

We can see from Table 5 that there are significant differences in social status and maternal general health and pregnancy health between the MSP categorical groups. Moreover, our design is to construct MSP vulnerability patterns present in DNAm at the time of birth. As such, we are limited to the subsample of 914 subjects with cord blood samples. The breakdown of these infants within each of smoking categories is found in Table 6.

Table 6: Maternal smoking-birth weight categories - number of subjects with cord blood DNAm data

Smoking Category	Number of subjects
Non-smoker	707
Smoking in periconception	51
Smoking until T1	23
Smoking in 2 or more periods	100
Other (missing more than 1 response or irregular coding)	13

Three categories have less than 10% of the observations. This low portion is considered statistically important class imbalance that can impact model accuracy (Cerf, Gay, Selmaoui-Folcher, Crémilleux, & Boulicaut, 2013). The 13 subjects in the last category were interesting in that some report no smoking early but answered affirmatively later in pregnancy. While this could be true, this atypical pattern could also be an issue with database entry, reluctance of the mother to honestly report at the outset of the study or an honestly inaccurate response of the mother as found in other studies (for useful examinations of this issue, see Gorber, Schofield-

Hurwitz, Hardt, Levasseur, & Tremblay (2009) and Valeri *et al.* (2017). Also of note, 25 subjects were born under 37 completed weeks of gestation (i.e. pre-term) while all others were born at term or greater.

Next, we wanted to look at the relative impact of smoking in the cord sample subjects. Due to the distribution of birth weight based on gestational age in general populations, splines can be used to model this outcome (Villandré *et al.*, 2011). These results for the cord blood subjects are in Table 7. Looking at column 1, the smoking categories are related to birth weight. Looking at columns 3 to 7, it appears this hold true across gestational ages except at the latest gestational ages.

Table 7: Regression results from linear spline model of smoking category and sex on birth weight in cord sample subjects (n = 914)

	Estimate	Std. Error	t value	Pr(>|t|)
smokClass	-162.59	58.81	-2.765	**0.005821**
sex	-210.12	545.12	-0.385	0.699998
ns(GestAge, df = 5)1	2108.57	366.62	5.751	**1.25E-08**
ns(GestAge, df = 5)2	1909.99	388.78	4.913	**1.08E-06**
ns(GestAge, df = 5)3	2126.17	251.8	8.444	**< 2e-16**
ns(GestAge, df = 5)4	2857.56	832.12	3.434	**0.000624**
ns(GestAge, df = 5)5	569.57	499.62	1.14	0.254614
smokClass*sex	151.55	82.09	1.846	0.065231
sex:ns(GestAge, df = 5)1	-74.75	529.4	-0.141	0.887748
sex:ns(GestAge, df = 5)2	236.64	561.66	0.421	0.673631
sex:ns(GestAge, df = 5)3	-625.93	349.29	-1.792	0.073497
sex:ns(GestAge, df = 5)4	819.24	1194.76	0.686	0.4931
sex:ns(GestAge, df = 5)5	1313.69	612.77	2.144	**0.032336**

Values in bold have $p < 0.5$. Results using splines R package, (basis matrix for representing the family of piecewise-cubic splines performed with ns function using variable gestational age.)

Sex is a known major source of variability in methylation research. This variable did not have a significant impact on birth weight (our proxy for fetal development) in the DNAm convenience sample (Table 7). Due to the sampling procedure of ARIES from ALSPAC, the distribution of MSP exposure and sex is fairly equal (Table 8).

Table 8: ARIES cord blood DNAm samples - Sex distribution

	Non-smoker	Persistent
Male	357	62
Female	361	68

3.1.2 MSP vulnerability using typical-atypical categories

We next looked at subject characteristics by MSP-birth weight categories as described in Figure 12. In Table 9, we show the number of subjects in each category. Because this categorization seeks to optimize subject differences, only 359 subjects fall at the extremes of maternal smoking (i.e. non-smoking versus persistent smoking.) Using gestational age and sex corrected z-score for birth weight, we arbitrarily split birth weight such that children at or below -1 SD were categorized as lower birth weight than their peers.

Table 9: Typical-atypical categories based on MSP and birth weight in ARIES

Fetal tobacco exposure	Birth weight	Abbreviation	n
None	Low	NS - LBW	42
None	Appropriate	NS - AGA	260
Present	Low	S - LBW	16
Present	Appropriate	S - AGA	41

After further exploration of this categorization, we noted that other variables appeared to affect this categorization. For example, there was a disparate distribution of maternal pre-pregnancy weight across the categories (Figure 29). Specifically, there was an inverse association between maternal pre-pregnancy BMI and being in the NS-LBW group, (beta = -1.6 kg/m^2 average difference in mothers compared to mothers of infants in NS-AGA group, p = 0.000184.)

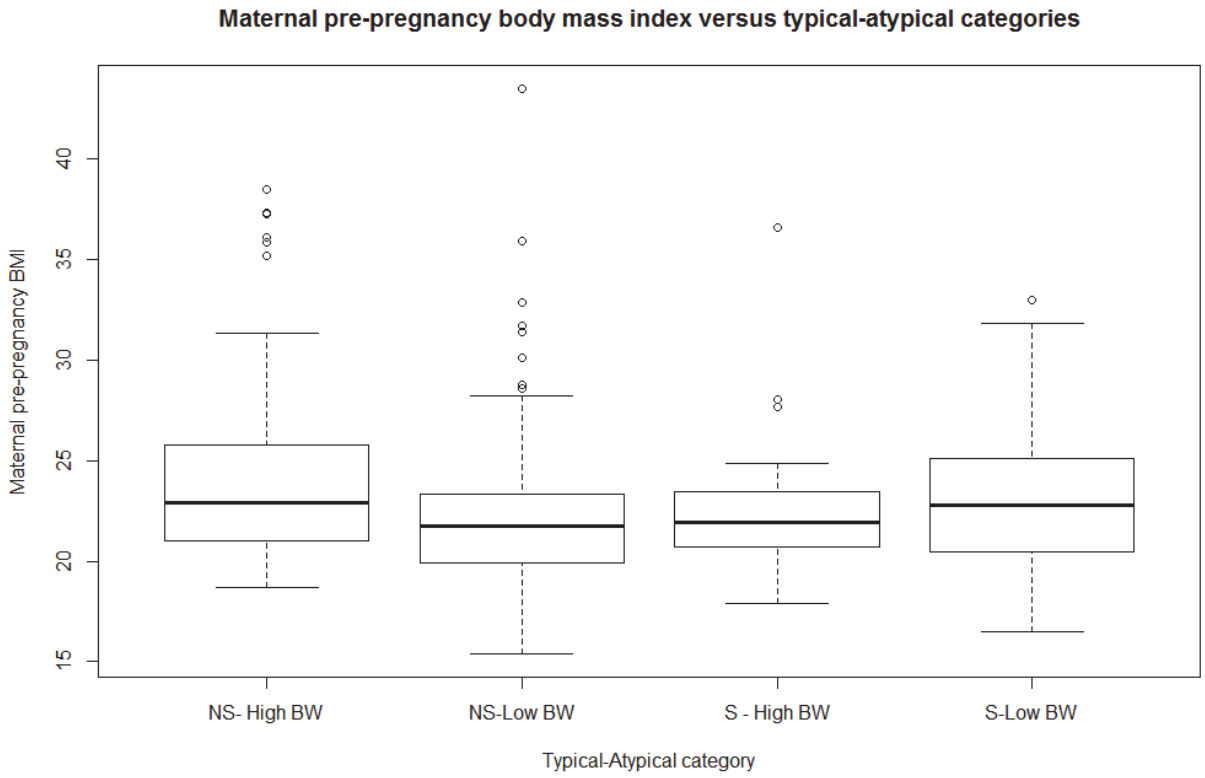


Figure 29: Maternal pre-pregnancy body mass index by typical-atypical category

To remove the influence of extremes of maternal pre-pregnancy BMI, obese and low BMI mothers were removed from the formative PLS-DA group.

Table 10: Typical-atypical MSP-birth weight categories - distribution based on maternal pre-gestational BMI.

Maternal weight before pregnancy	*NS- High BW*	*NS-Low BW*	*S - High*	*S-Low BW*
Low BMI	0	8	2	3
Normal	118	131	20	26
Overweight	35	18	2	9
Obese	14	6	1	2

Maternal pre-gestational weight classification using WHO reference guidelines (National Research Council, 2010).

3.1.3 MSP vulnerability using composite data

To create the vulnerability score, we used all available mother-infant data in the ALSPAC cohort (n = 15211). We used MIFAMD to impute missing data before conducting FAMD analysis. We used number of components = 2 (tuned using the estim_ncpFAMD function) and 50 imputed datasets. As discussed in the Methods section, we considered three sources of infant health relevant data related to MSP: pregnancy health related, MSP related and birth weight. We extracted 20 dimensions in total. For the sake of space, only data for five dimensions are shown.

Regarding the MSP variables, we recall the concerns relating to class imbalance and uncertainty in self-reported data discussed in Section 3.1.1. Another issue is that we cannot ascertain the accuracy of a more finely grained versus coarser breakdown of MSP categories in terms of the duration and which periods during periconception and gestation. The risk of bias caused by misclassification would be highly challenging to attenuate or even estimable in terms of direction or magnitude (Valeri *et al.*, 2017). Instead of bringing forward these issues, we decided to use consistency as a major basis of grouping this self-reported measure. Moving forward in multi-class data analysis, mothers who consistently reported smoking or not smoking throughout the survey period were classified as "Persistent" or "Non-smokers", respectively. Mothers who at one point or another reported smoking were classified as "Non-persistent".

In our first trial of using FAMD, we included all three types of data: maternal health related, MSP related and birth weight. We first look at the scree plot in Figure 30.

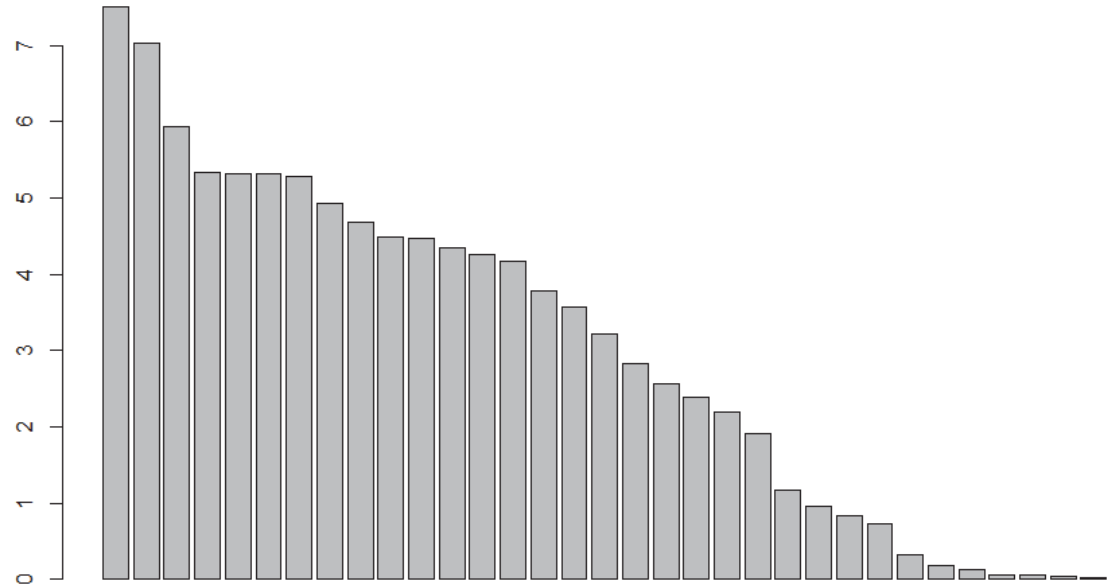


Figure 30: Boxplot of variance captured per FAMD dimensions using data from ALSPAC (n = 14694) using the following variables: maternal health, MSP related and birth weight factors. Y-axis is proportion (%) of variance captured.

We next look at these top two dimensions in Figure 31. The first dimension is mainly composed of MSP and social factors whereas the second dimension is dominated by maternal health and infant factors. Both dimensions together capture less than 20% of total variability. As well, there is only a small drop (less than 2%) between these top two dimensions and the third. Typically, successful factor analysis typically finds steeper drops in variance capture between top dimensions versus remaining dimensions (Pagès, 2004).

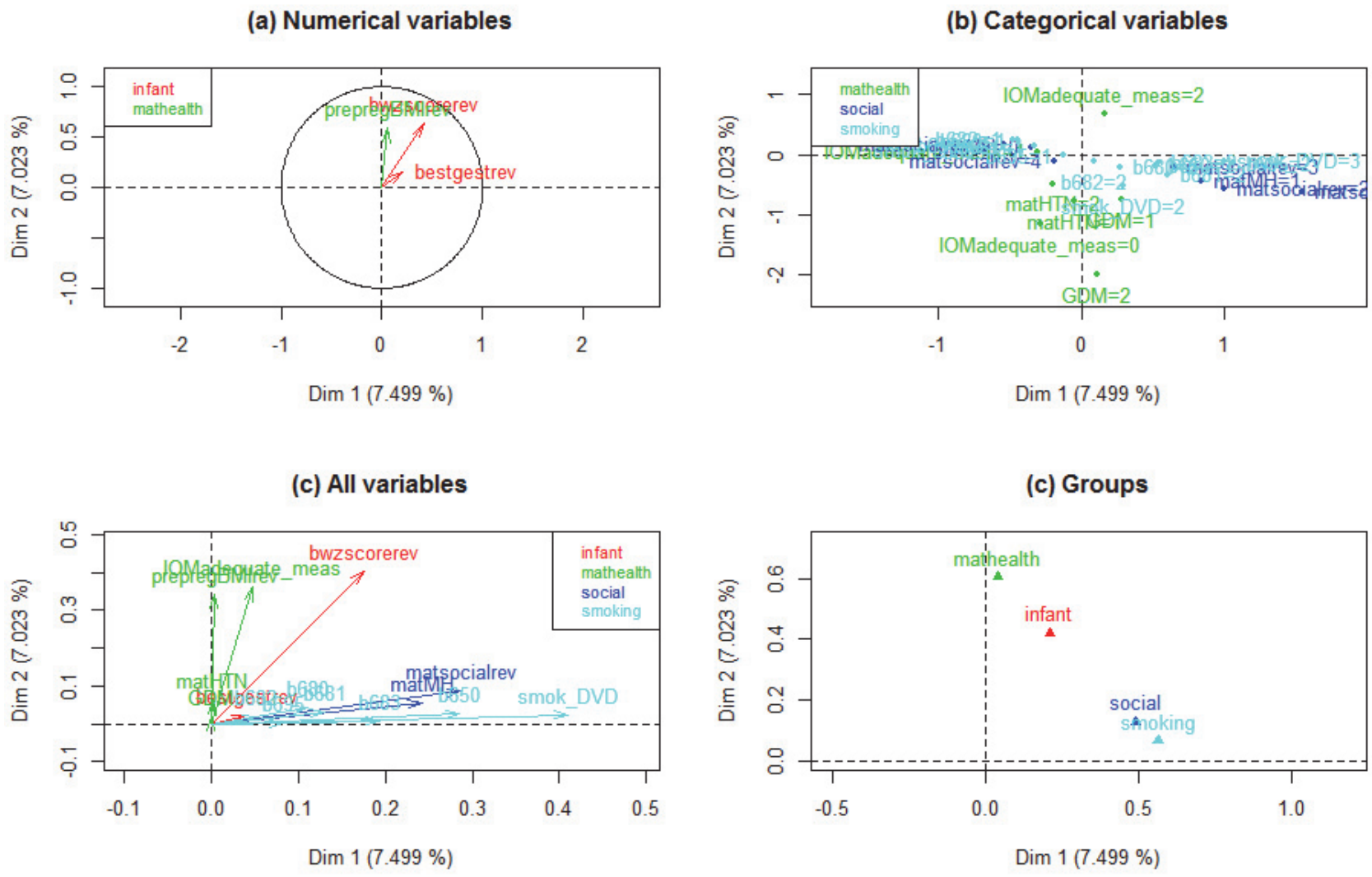


Figure 31: FAMD analysis using maternal health related, MSP-related and birth weight factors - Variable contribution of first two dimensions. Colours represent different groups of variables, (smoking related, pregnancy related, social factors and infant birth weight.)

We proceeded to systematically trial different combinations of features in composite creation. We used percent variability captured by the top dimensions and percent contribution of the feature to the dimensions to compare combinations. We found that the maternal health and most social features impaired performance. We also note that these features had the most missing data compared to other features like birth weight (see Table 5).

Proceeding with FAMD without maternal health and social features, we see the scree plot in Figure 32. Unlike the previous analysis with all the theoretically relevant variables included seen in Figure 30, the first two dimensions capture almost 40% of total data variability. As well, there is a larger gap in variance capture and eigenvalues between the top dimensions and the subsequent dimensions. We proceeded with this FAMD model that only included MSP-related features and infant birth weight. Please refer to Appendix F for measures of sampling adequacy.

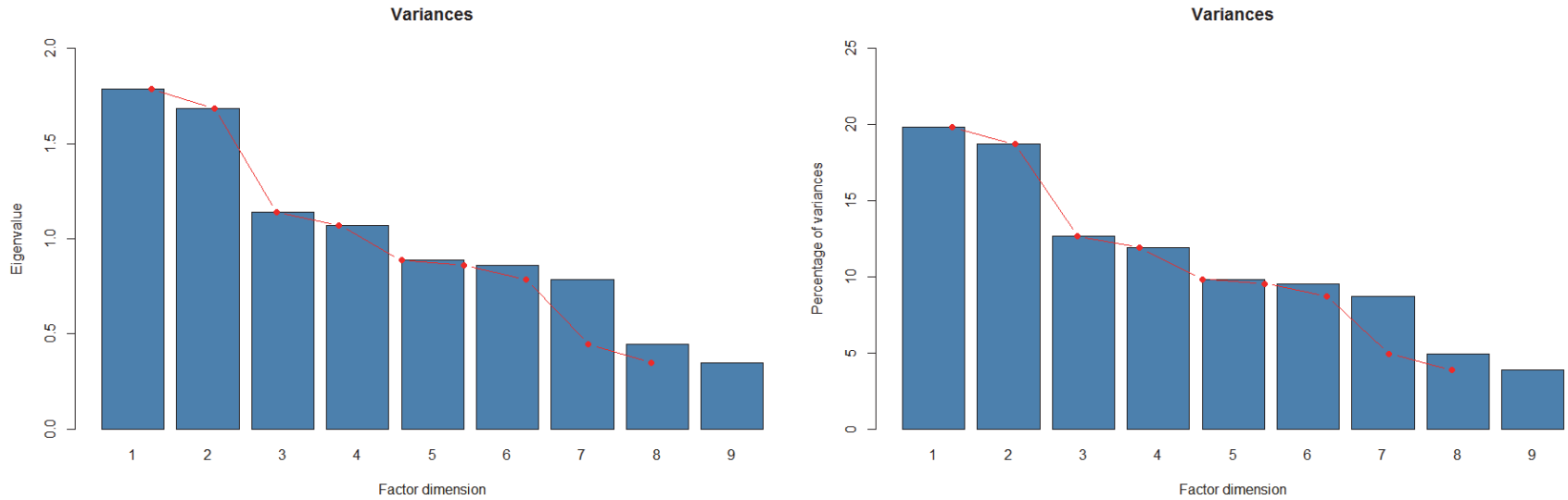


Figure 32: Scree plot of factor analysis of vulnerability data. Missing data imputed using 50 datasets using the MIFAMD function. Left: x-axis = Eigenvalues. Right: x-axis: Percentage of variance captured.

Next, we obtain a general overview of what each dimension represents by calculating the relative contribution of each variable to the dimension. The squared cosine shows the importance of a component for a given observation (see footnote 5). Higher values indicate a larger portion of importance. Figure 33 shows the contribution of each variable in determining the first two dimensions. The first dimension represents 19.8% of data variability and primarily captures grandmaternal data i.e. the grandmother's smoking history and whether the grandmother smoked while pregnant with the mother. In contrast, Dimension 2 captures similar information for the subject's own mother and represents 18.7% of data variability.

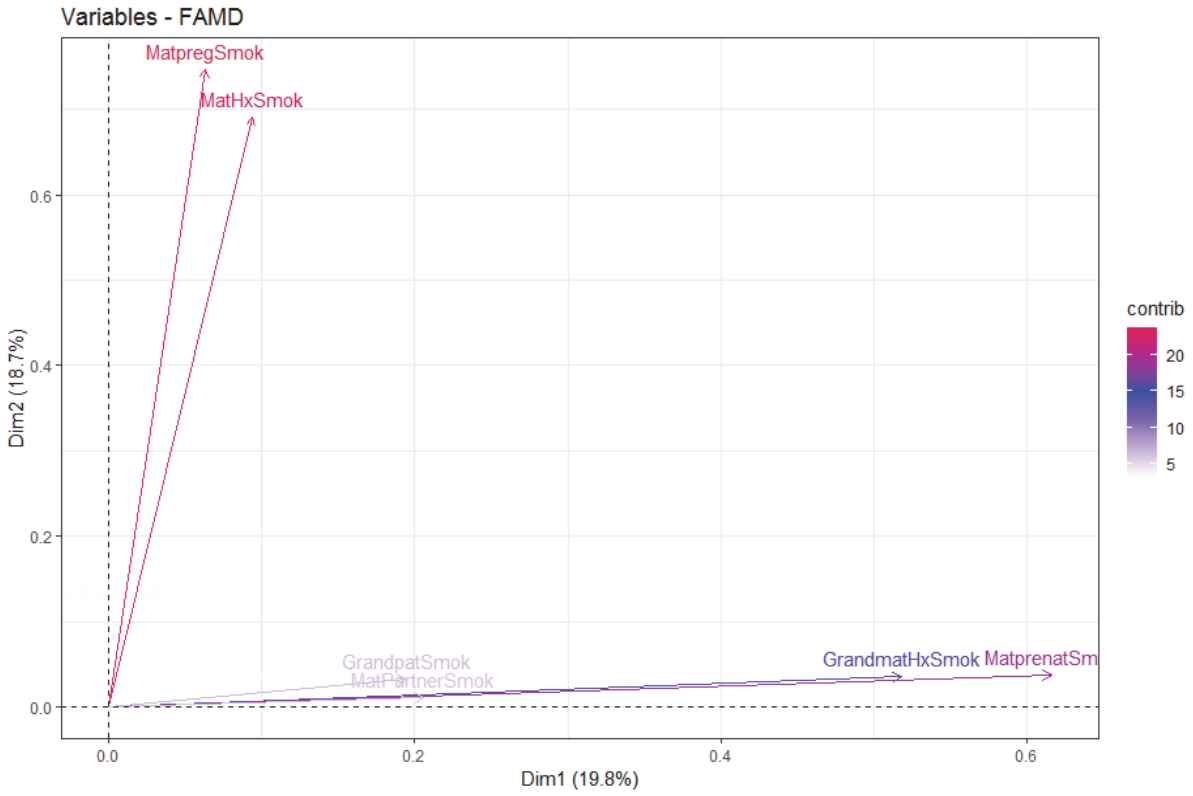


Figure 33: Variable contribution[5] to Dimension 1 and 2

We can also view each dimension separately to get an overview of the most prominent features of each dimension (Figure 34). If all variables contributed equally, then an equal split of 100% is calculated as the inverse of number of variables. With the eight MSP related variables, this is 12.5% (1 divided by 8). Some authors suggest that contribution over this value can be

[5] Contribution calculated as square cosine*100 / total square cosine of the dimension

considered important (Kassambara, 2017). For most dimensions, only about two variables are prominent using this cut-off (i.e. surpasses the red dotted line in Figure 34.)

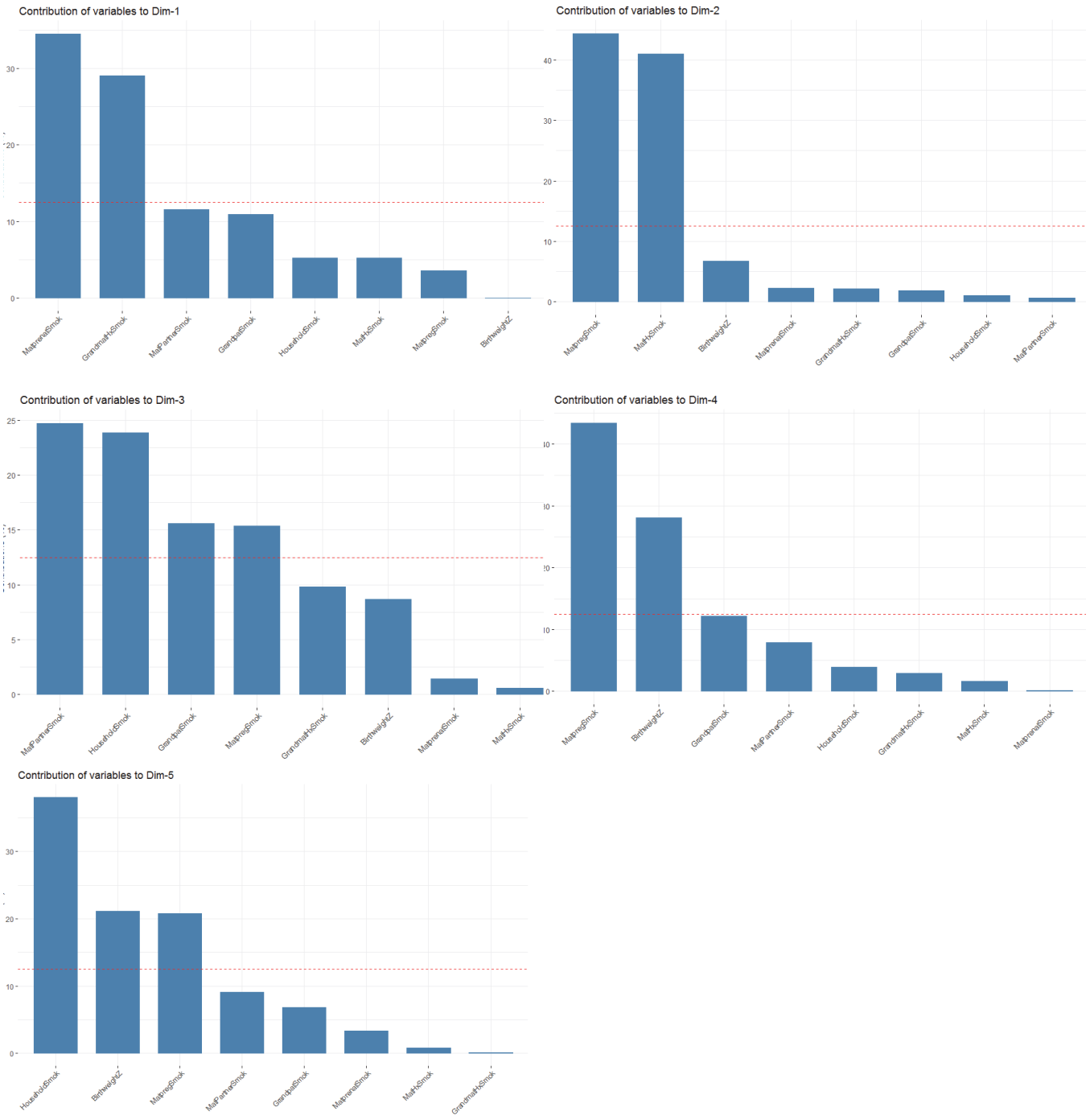


Figure 34: Barplot of relative contribution (y-axis: squared cosine*100/total squared cosine) of variables to dimension construction. The most relevant are typically considered over the inverse of the number of variables, (12% in this case - depicted by red dotted line.)

To better visualize gross directionality of factors relative to the dimension scores, we can view a correlation plot of the individual constituent variables and each of the five dimensions (Figure 35).

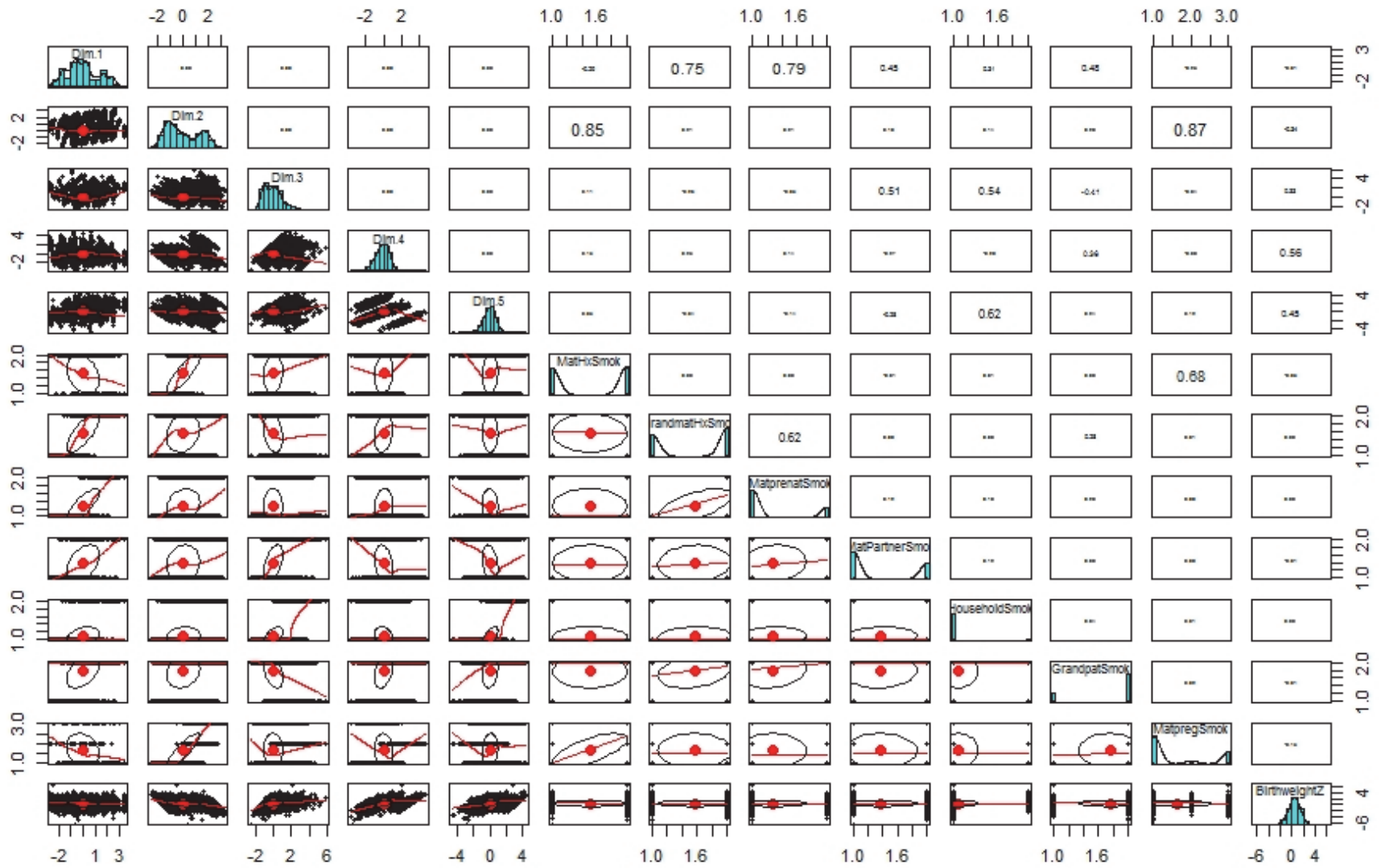


Figure 35: Correlation plot between FAMD dimensions and constituent variables. Pearson correlation values with font size reflecting the strength of correlation. X-axis – Dimension score. Y-axis – variable value (scaled and centered.)

However, take note that correlation is shown to only show a cursory overview of the data to show continuous and categorical variables together. Correlation was not used to perform factor analysis. Correlation may be misleading when visualizing categorical variables as there is no true ordinality to the data. For instance, Dimension 3 appears to have a low correlation value with maternal prenatal exposure to smoking – however, the blue bar graph shows that lower Dimension 3 scores are grouped with less maternal prenatal exposure. The clustering performed by FAMD is more appropriate for categorical data and better visualized in contribution plots like in Figure 31 and Figure 36.

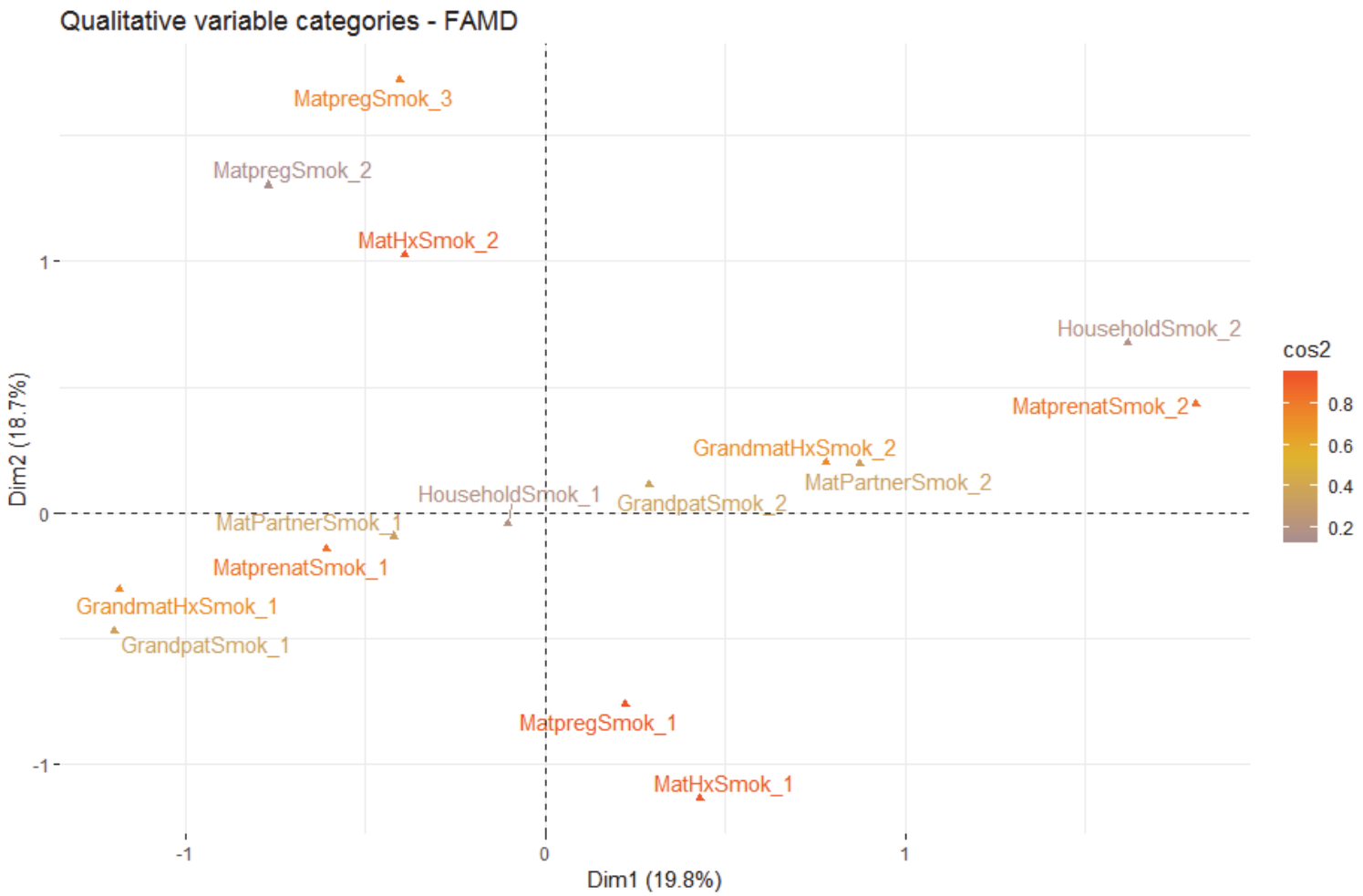


Figure 36: FAMD analysis using only MSP variables and birth weight. Plot of relative contribution of variables to dimension 1 and 2. Percent contribution indicated by colour legend. Variables that move from the centre along the same axis as the dimension are more influential. The numeric suffix represents the level of the variable, for instance GrandmatHxSmok_1 represents no grandmaternal history of smoking whereas GrandmatHxSmok_2 presents a postive history.

Using Dimension 1 as an example, it is positively related to grandmaternal smoking history and mother's own exposure to smoking from the grandmother (Figure 36.) Dimension 2 increases as maternal MSP and pre-conceptual history of smoking increases. Dimension 4 and 5 are both positively related to birth weight. However, Dimension 4 is higher when mothers report smoking during part but not the entire pregnancy. In contrast, Dimension 5 is lower when this type of smoking is reported. It is also higher when other household members smoke. Dimension 4 increases in relation to higher birth weight and what could be described as a "U-shaped" relation to maternal smoking in pregnancy. Dimensions 3 and 5 are the only dimensions to include relations in smoking in individuals besides the mother or grandmother. Table 11 summarizes the directionality of MSP exposure with each dimension.

Table 11: Summary of MSP exposure composite index - dimension characteristics

Dimension	Birth weight	Maternal Smoking in Pregnancy	Maternal Hx Smoking	Grandmaternal Hx of Smoking	Maternal prenatal exposure to smoking	Maternal partner smokes	Other household members smoke	Grandfather smokes
1	.	.	.	+	+	.	.	.
2	.	+	+	.	.	.	.	.
3	.	+/-	.	.	.	+	+	-
4	+	+/-	.	.	.	.	.	.
5	+	-/+	.	.	.	.	+	.

This table provides character of variable as dimension score increases. "." indicates no relation, "+" and "-" indicate a positive or negative correlation, respectively. In the Maternal smoking in pregnancy column (column 2), the "+/-" indicates a nonlinear relation i.e. low dimension scores are related to both no or high levels of reported MSP, but high values are related to medium levels of reported MSP. The converse is represented by "-/+".

3.2 DNA exploratory analysis

The distribution of beta values in our cord convenience samples is depicted in Figure 37. This beta distribution is consistent with other literature (Solomon *et al.*, 2018; Zhou, W. *et al.*, 2017).

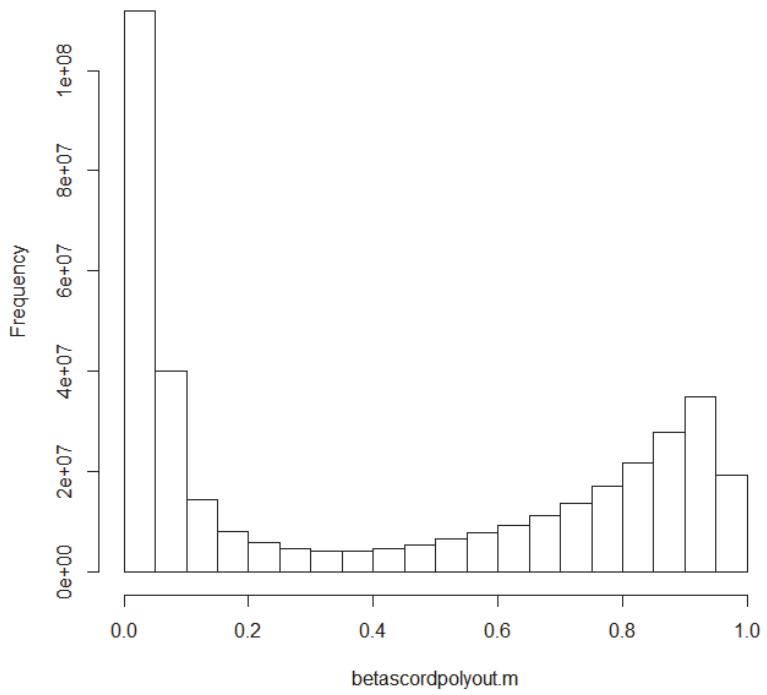


Figure 37: Histogram of beta values in cord blood samples (n = 914)

Next, we explored the DNAm data with PCA (Figure 38). The aim was to visualize if and what natural groupings existed in the cord DNAm based on the greatest variance in the dataset. Visual inspection of PCA components revealed no separation by MSP status of the samples. Looking at Figure 38, we can see that this analysis accounts for a small portion of DNAm variance, the first two principal components (PCs) together only summing to around 10%.

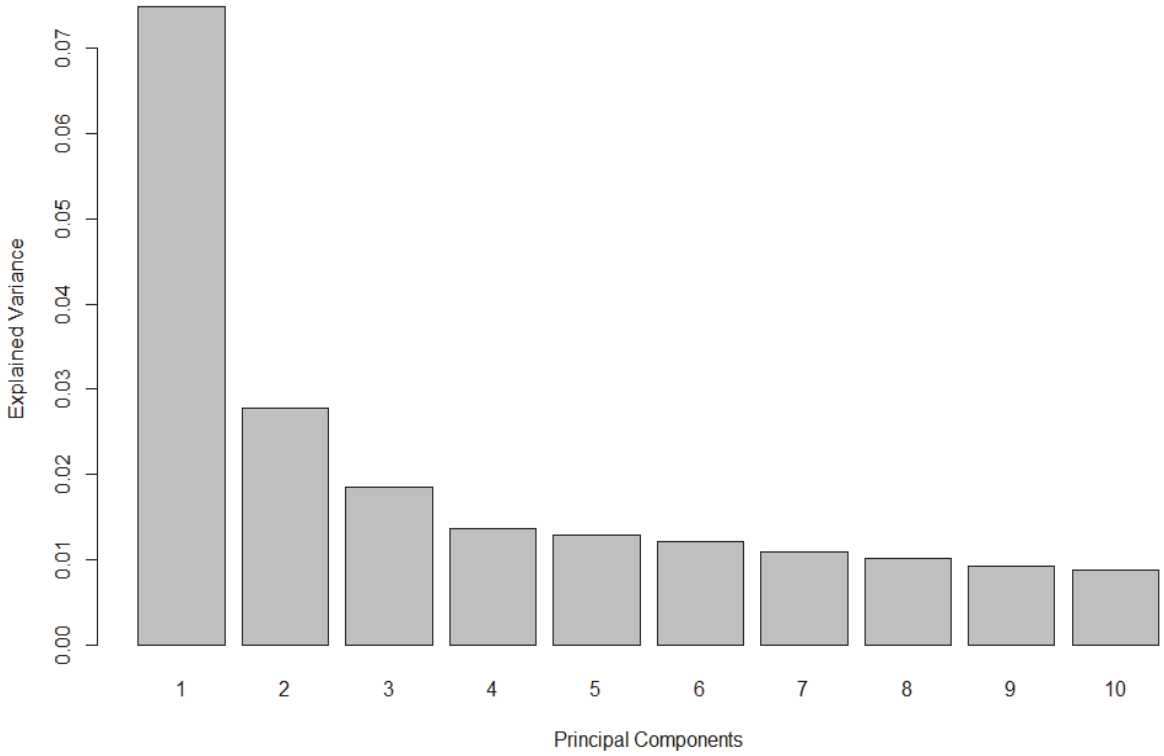


Figure 38: Scree plot. The first component captures the most methylation variability. However, the percentage captured is less than 10% which is considered low.

Instead of just viewing the variability captured by the components, we can visualize if the components separate subjects according to their MSP vulnerability. Using the top 2 components, we can plot each subject using the component scores. We can colour each subject data point according to their maternal reported MSP category. Little distinction can be seen in Figure 39.

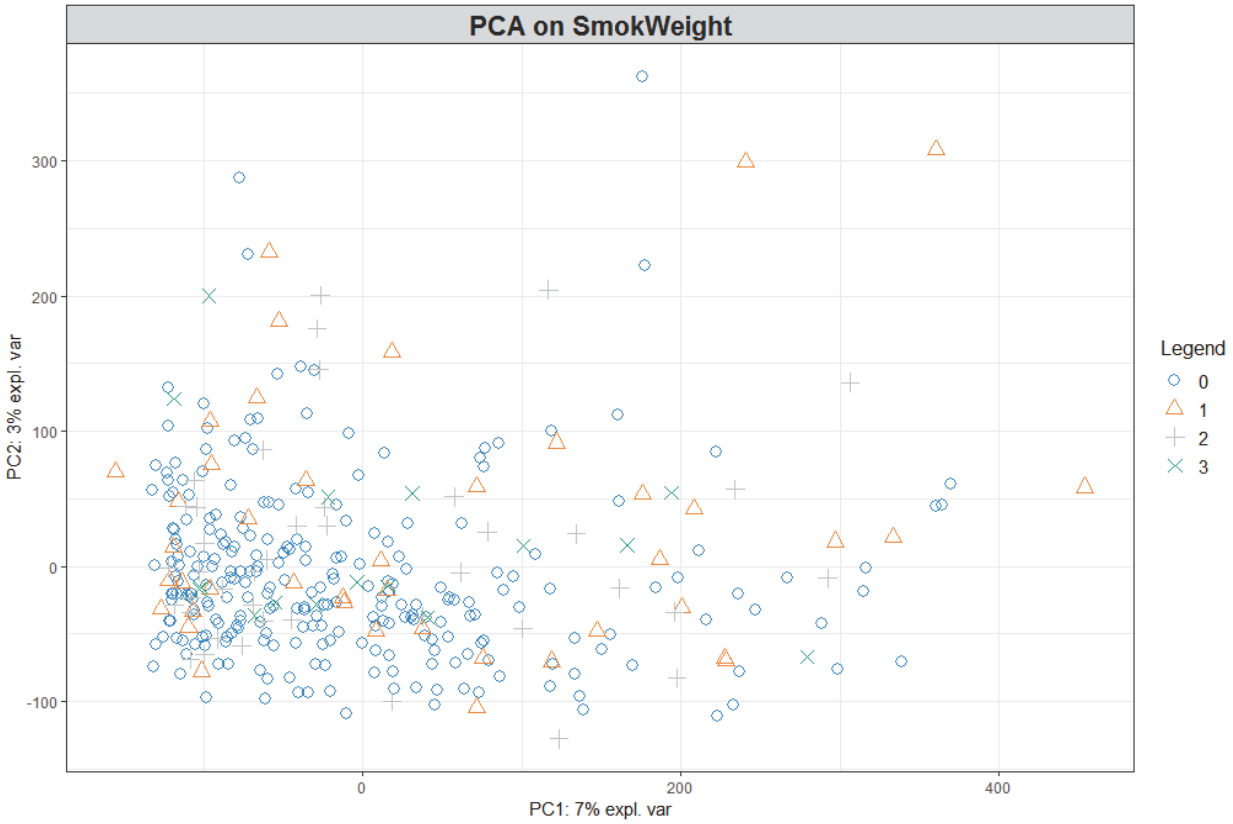


Figure 39: Score plot of principal components analysis of DNAm data, (first two principal components plotted.) Together, the first to components account for about 10% of total data variation. Legend: zero refers to non-smoking mother, 1 is periconception only, 2 is smoked in 1 or more trimesters, 3 is smoking consistently throughout pregnancy.

We also used multiple hypothesis testing to observe the linear relation between beta values and smoking category. The Manhattan plot in Figure 40 depicts the significance of association with beta values as the negative logarithm of the *p*-value (-log(*p*-value)) versus the chromosomal location for each of the tested CpG sites. Twelve CpGs passed the significance threshold after multiple testing correction (Bonferroni).

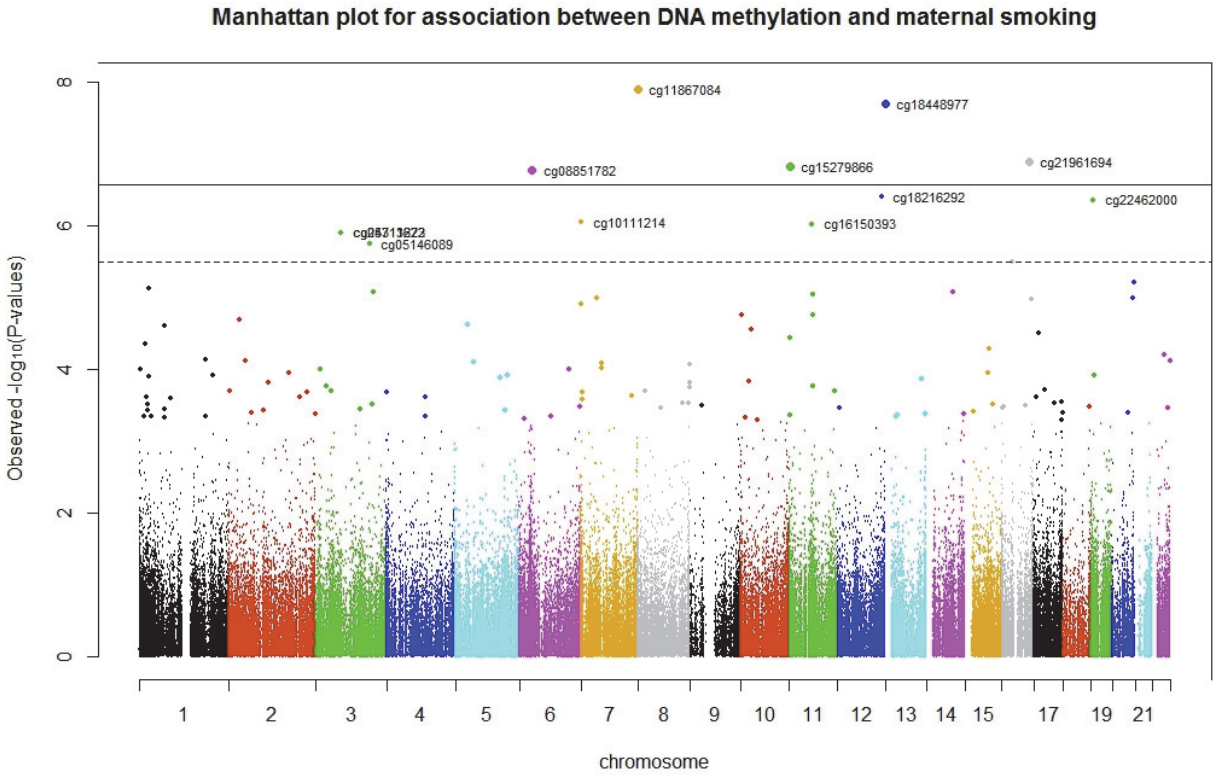


Figure 40: Manhattan plot of epigenome wide association analysis of smoking status in cord samples (n = 914). Solid horizontal line represents Bonferroni threshold; dotted horizontal line represents FDR correction (Benjamini & Hochberg method, p < 0.05) threshold. Generated using cpgAssoc R package.

This is in contrast to birth weight. When a similar analysis was performed relating to birth weight (Figure 41), 671 sites CpGs exceed significance the Bonferroni threshold.

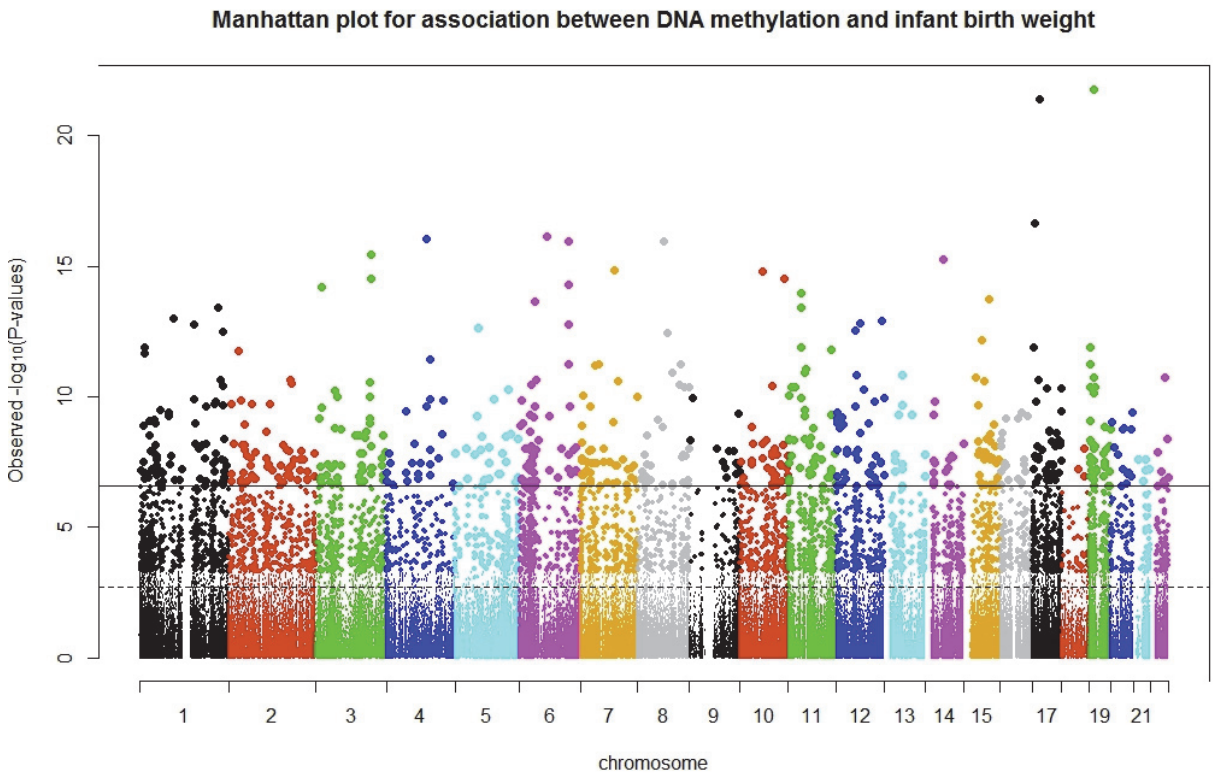


Figure 41: Manhattan plot of epigenome wide association analysis of infant birth weight (cohort based z-score) in cord samples (n = 914). Solid horizontal line represents Bonferroni threshold; dotted horizontal line represents FDR correction (Benjamini & Hochberg method, p < 0.05) threshold. Generated using cpgAssoc R package. DNAm vulnerability patterns using categories

Next, we will use PLS to directly observe co-variability between DNAm and MSP vulnerability. We start with maternal reported MSP and then proceed with typical-atypical MSP-birth weight categories. We will be using the categorical extension of PLS, PLS-DA. As with the PCA analysis above, we can similarly plot the PLS component scores of each subject and then colour data point according to the subject's MSP vulnerability category.

3.2.1 Analysis with PLS-DA and maternal reported MSP

We first use MSP based on maternal report as "bait" to guide the formation of the components using PLS-DA. The following figures, Figure 42 and Figure 43 show results from PLS-DA using cord DNAm data and maternal reported MSP for the ARIES and GenR cohorts, respectively.

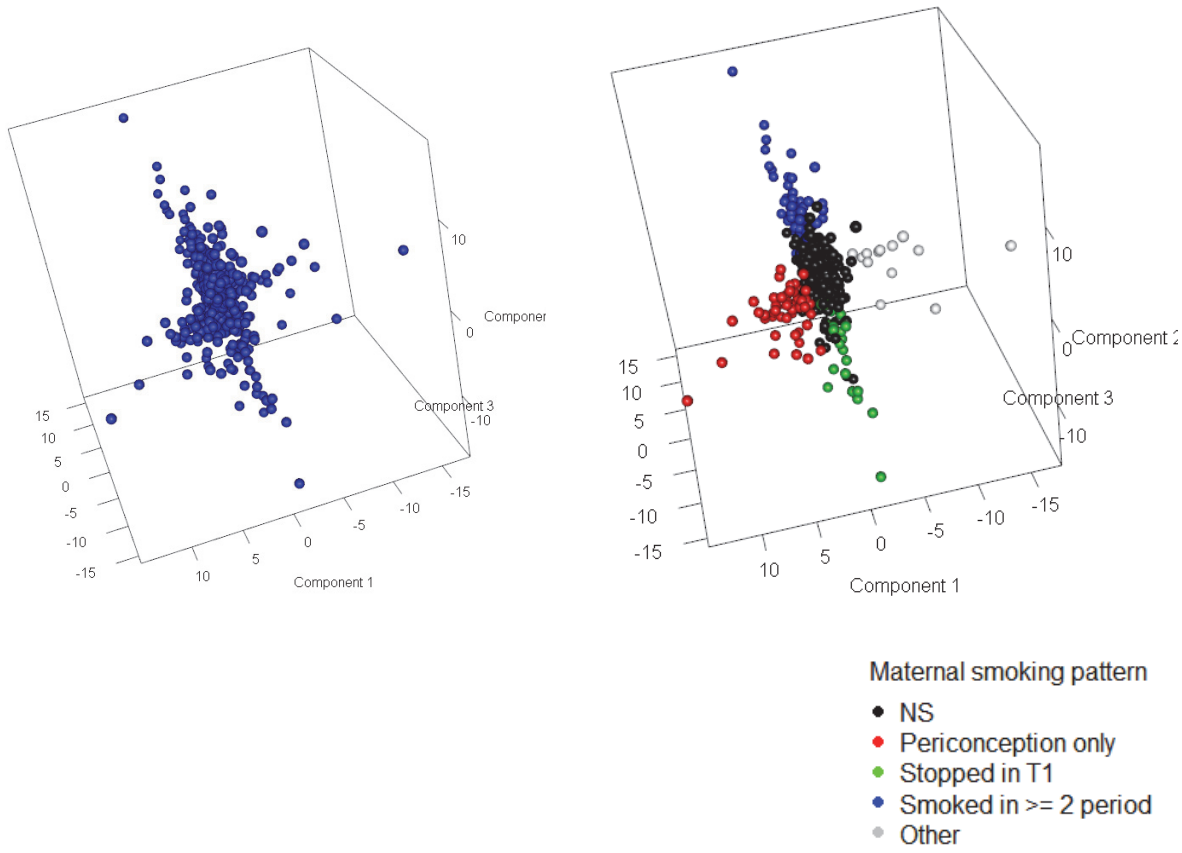


Figure 42: PLS-DA of DNAm and maternal reported MSP. Left: Plot of subject component scores. Right: same as left except each data point is coloured as per legend to demonstrate distinction between groups.

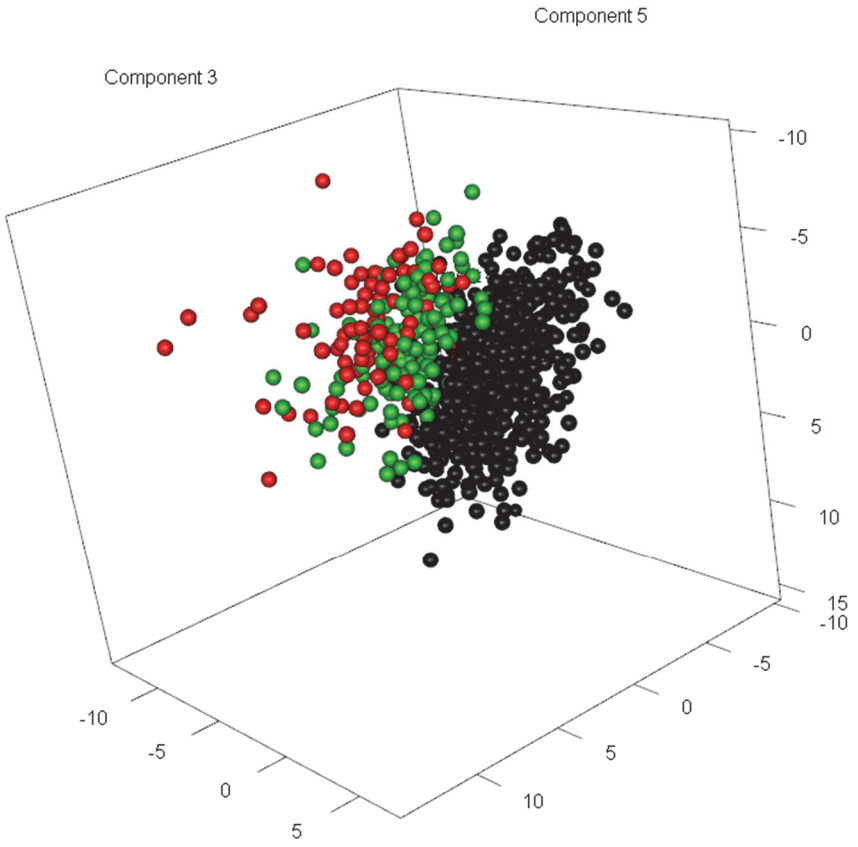


Figure 43: Generation R data - PLS-DA of cord DNAm data and maternal reported MSP. Black - non-smoker. Green - Quit in early pregnancy. Red - Smoker during pregnancy.

We can also colour the data points with other variables to see if the scores can separate subjects accordingly. As seen in Figure 44, the PLS-DA detected patterns did not appear related to social variables, such as social status or education level of either mother or father of the subject (for sake of space, only maternal education shown.) Neither paternal smoking before nor during the pregnancy showed any association. We also checked if BCD plate was related to any patterns in PLS-DA scores and found none. Subject sex did lead to some separation. We repeated PLS-DA after removing DNAm data from sex chromosomes X and Y. However, this difference persisted.

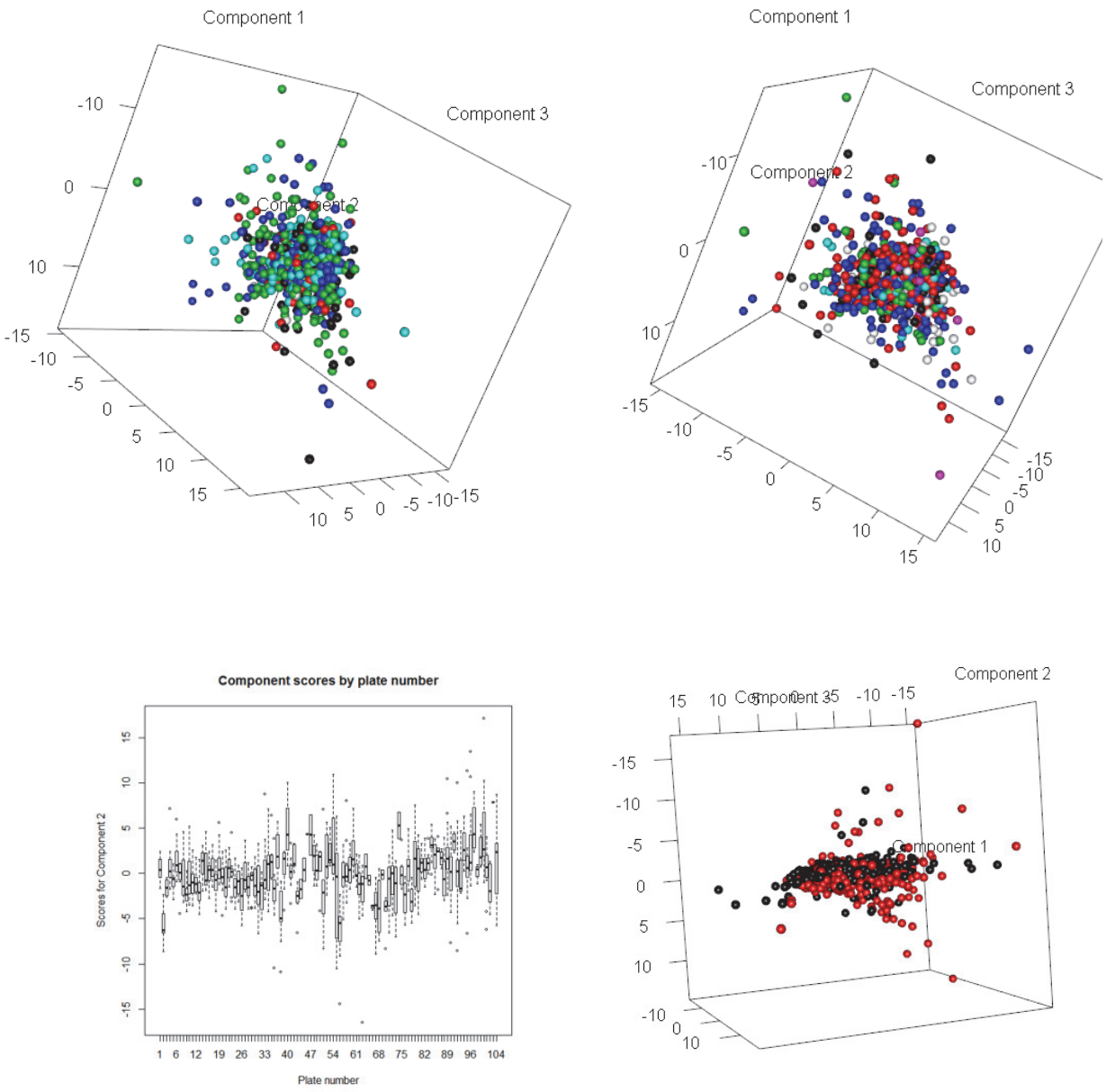


Figure 44: PLS-DA of cord DNAm related to maternal smoking categories. Colour are by the following variables: Top left - Maternal education level, top right - paternal smoking in before/during pregnancy, bottom left - microarray plate number (component 2 used as an example), bottom right – infant sex (after removal of DNAm data from X and Y chromosomes.)

We took a cursory look at whether the PLS-DA patterns had any relation to clinical outcomes. As an example, we look at regression models for prenatal growth (i.e. birth weight) in Figure 45 and a summary of results including early postnatal growth in Table 12. These associations were observed even after controlling for maternal BMI, maternal gestational weight gain, infant gestational age, and infant sex.

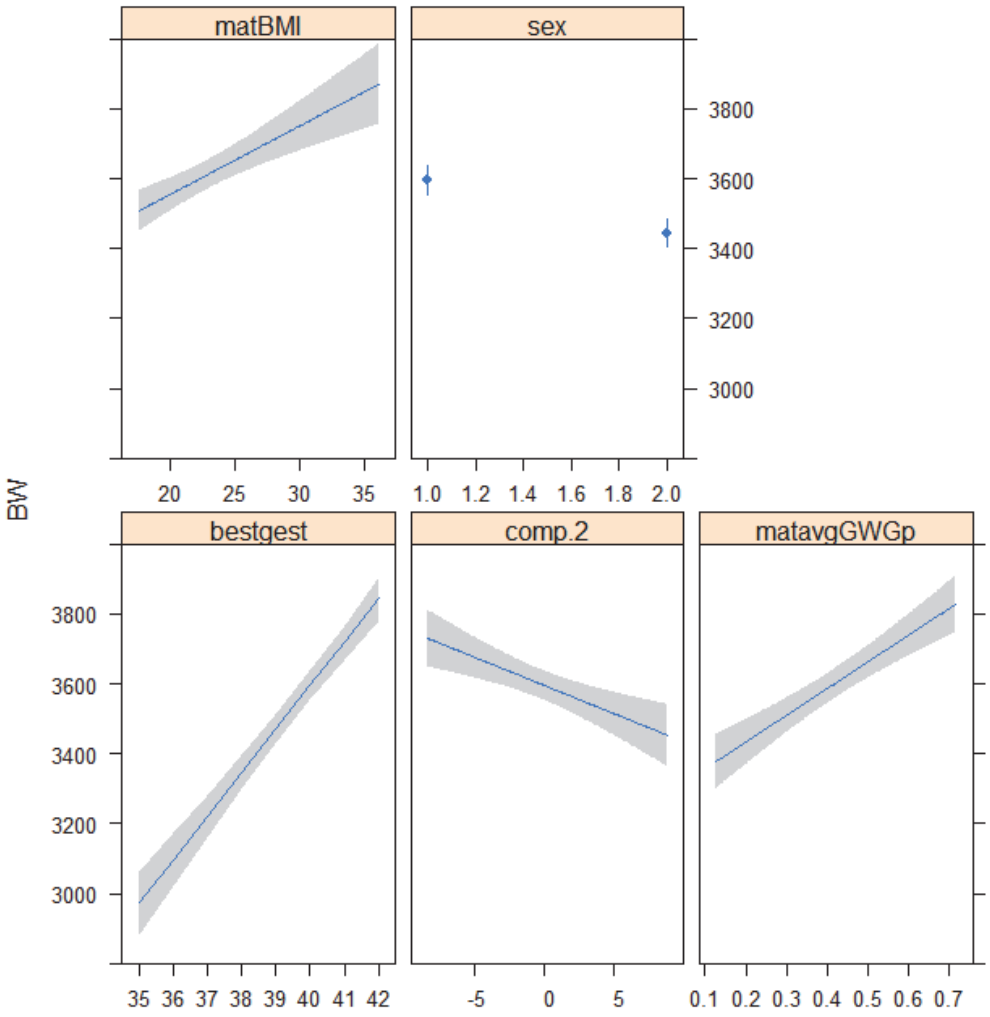


Figure 45: Linear plot - regression model to predict birth weight based on PLS-DA components (extracted using maternal reported MSP). Shading indicates residual error. Predictors (top to bottom, left to right): maternal BMI, infant sex, gestational age at birth, Component 2 scores, and maternal average weight gain.

Table 12: PLS-DA components (maternal reported MSP) linear relation to prenatal growth (birth weight) and postnatal growth in the first 3 months of life.

Outcome	B	SE	DNAm Pattern
Birth weight	2571.4*	519.3	2
Length growth	-1.6*	0.2	2
Weight growth	1.3*	0.2	3
*p = <.0001			

Covariates: maternal BMI, maternal average weight gain, infant sex, and gestational age at birth.

However, closer examination of the components revealed that the majority of variance was related to only two categories: smoking in periconception or "other". This latter group consists of subjects with no MSP data or inconsistent reporting (e.g. reporting no smoking in before or in early pregnancy but then reported smoking in later pregnancy.) Very little variance was related to subjects with no MSP exposure. Components 2 and 3, which were related to pre- and post-natal growth rates, do include relative increases in representation of variability related to first trimester smoking and smoking in greater than 1 period in pregnancy compared to Component 1. However, there remains little variability related to non-smokers in all three of these components.

Table 13: PLS-DA (cord blood and maternal smoking categories). Variance captures by components.

VARIABLE	NUMBER OF COMPONENTS (CUMULATIVE)							
	1	2	3	4	5	6	7	8
DNAm	0.3738	1.819	2.826	3.796	5.41	6.193	6.963	7.782
No smoking	2.8449	3.673	4.034	7.953	13.4	25.287	29.368	30.348
Smoking in Periconception	32.0895	33.635	41.228	41.268	41.27	41.876	42.911	45.342
Smoking to first trimester	0.5658	10.234	18.299	19.452	22.52	47.391	68.214	78.303
Smoking in 2 periods or more	0.2732	12.616	34.228	51.272	59.91	69.094	70.595	70.866
Other	31.6521	31.784	33.008	38.467	42.87	44.694	44.876	45.536

To further evaluate these results, we conducted model cross-validation using 10 random segments (Figure 46). Typically, a greater number of components should lead to better (i.e. lower) RMSEP-values. It is known in CV that a clear minimum of prediction errors may not be obtained, making the model selection difficult. The highly variable predictive error together with the poor representation of smoking versus non-smoking in the PLS-DA components suggests a poor model.

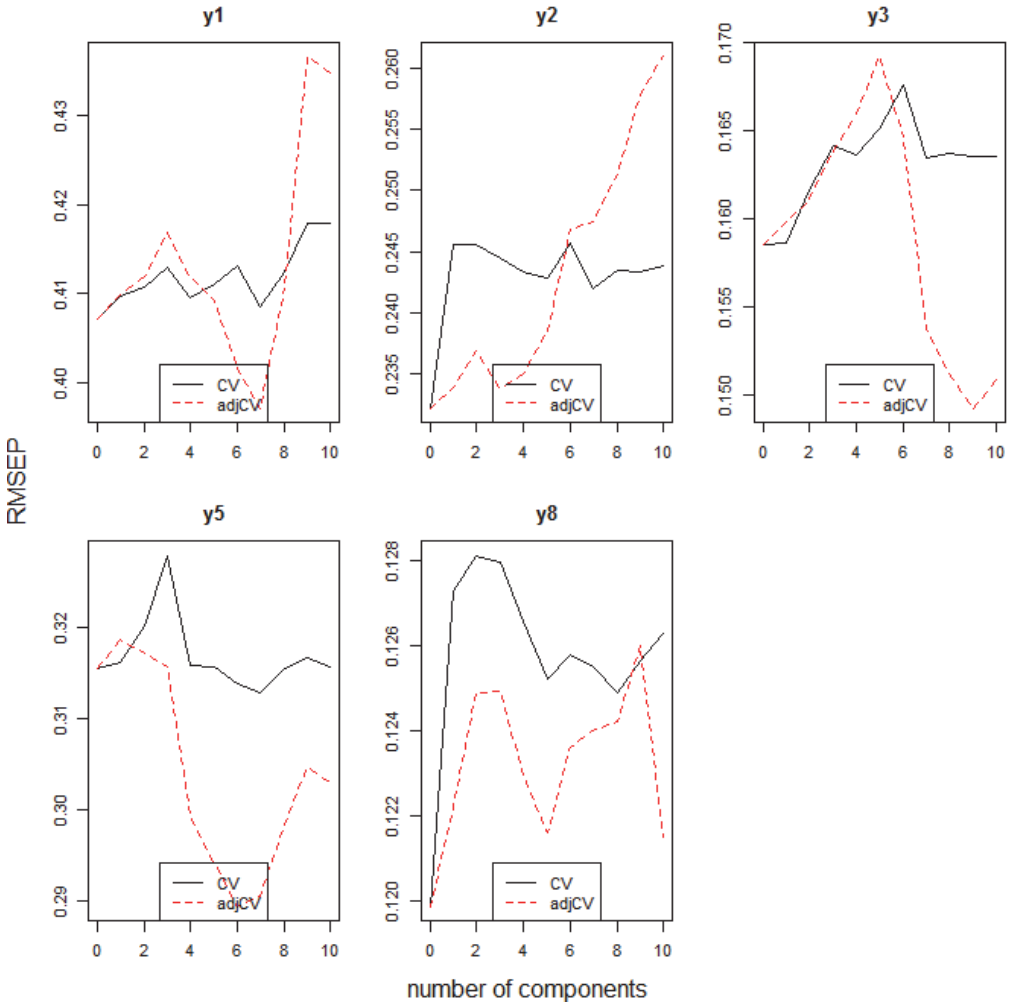


Figure 46: PLS-DA using maternal reported MSP categories - Performance using 10 fold cross validation. Lines show Root Mean Square Error of Prediction. X-axis: number of PLS-DA components tried.

3.2.2 Analysis with PLS-DA and typical-atypical MSP-birth weight categories

Next, we now change "bait" to the typical-atypical MSP-birth weight categories. We can see from Figure 47, there is some separation seen even with only two components. As well, about 9% of DNAm variability is captured with the first two components compared with less than 3% in the case of PLS-DA using maternal reported MSP categories.

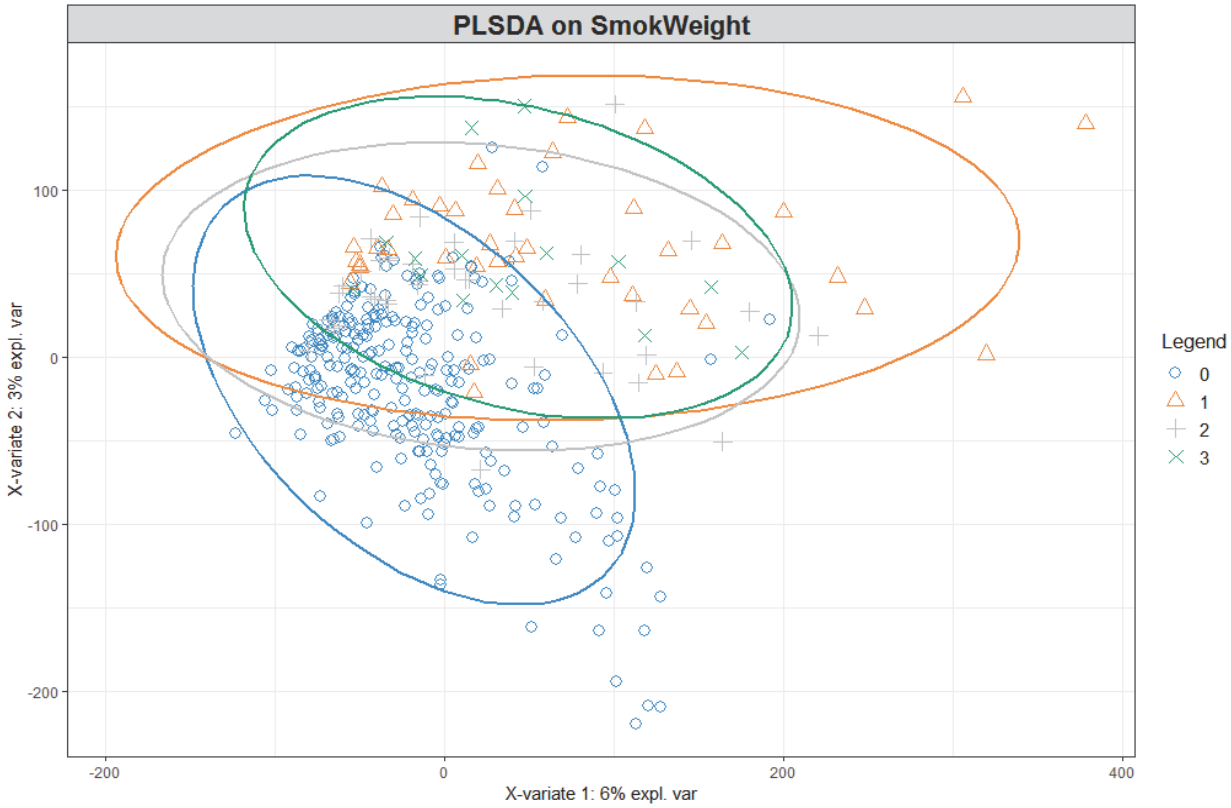


Figure 47: Scatterplot of PLS-DA scores using cord DNAm and typical-atypical MSP categories. Each axis also notes the percent explained variance of DNAm data. Legend: 0 – Non smoker, 1 – Smoking in periconception, 2 – smoking to first trimester, 3 – smoking in 2 or more periods.

By plotting three components as seen in Figure 48, we observe even clearer separation of subjects by typical-atypical category. This may be expected as this categorization uses the extremes of birth weight and MSP exposure. Due to this property, also note there are about only half as many data points as with maternal reported MSP analysis.

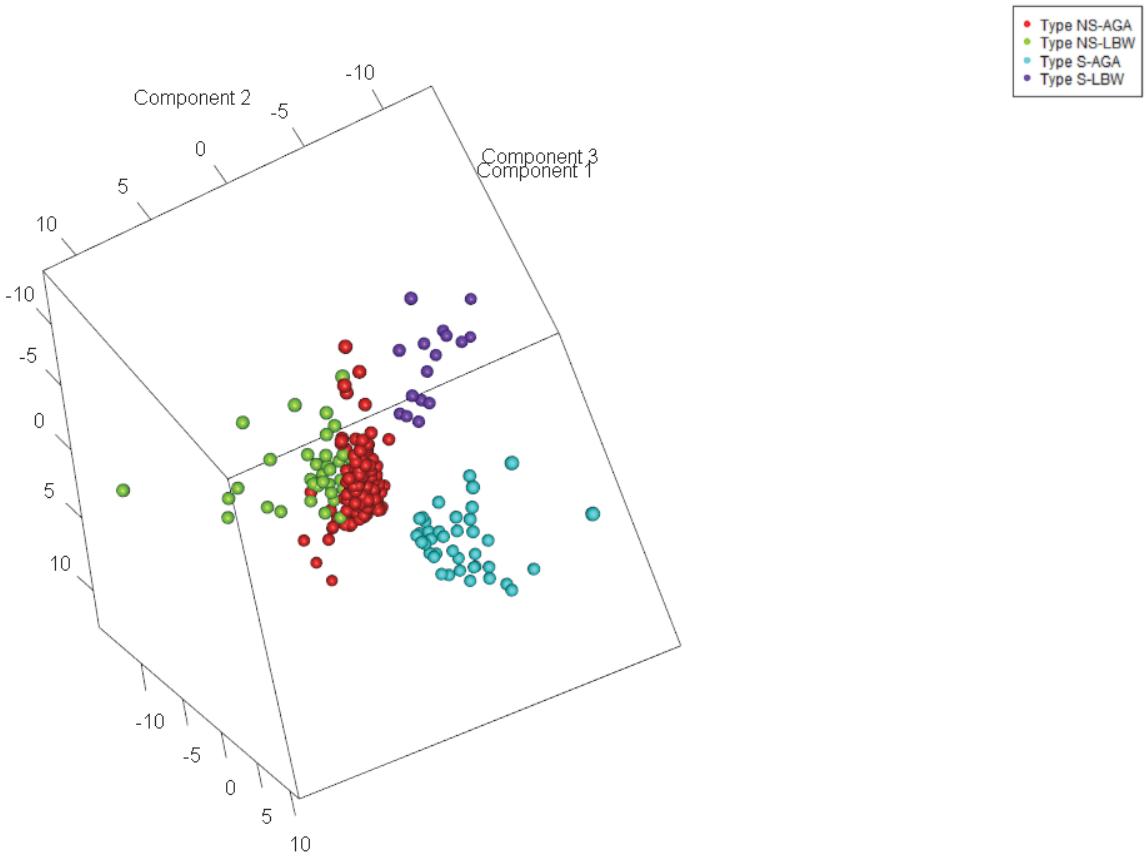


Figure 48: PLS-DA of cord DNA methylation data using typical-atypical categories.

As with the maternal reported MSP analysis, we again evaluated the model using cross validation. We again encountered the same problem as previously with no clear minimum of prediction errors. Suspecting we were encountering overfitting, we repeated the PLS analysis but this time using soft-thresholding penalization to promote sparsity (implemented using the sPLS function in the mixOmics R package.) This is shown in Figure 49. As we can see, introducing sparsity leads to a less distinct separation of points, especially of the smaller smoker categories.

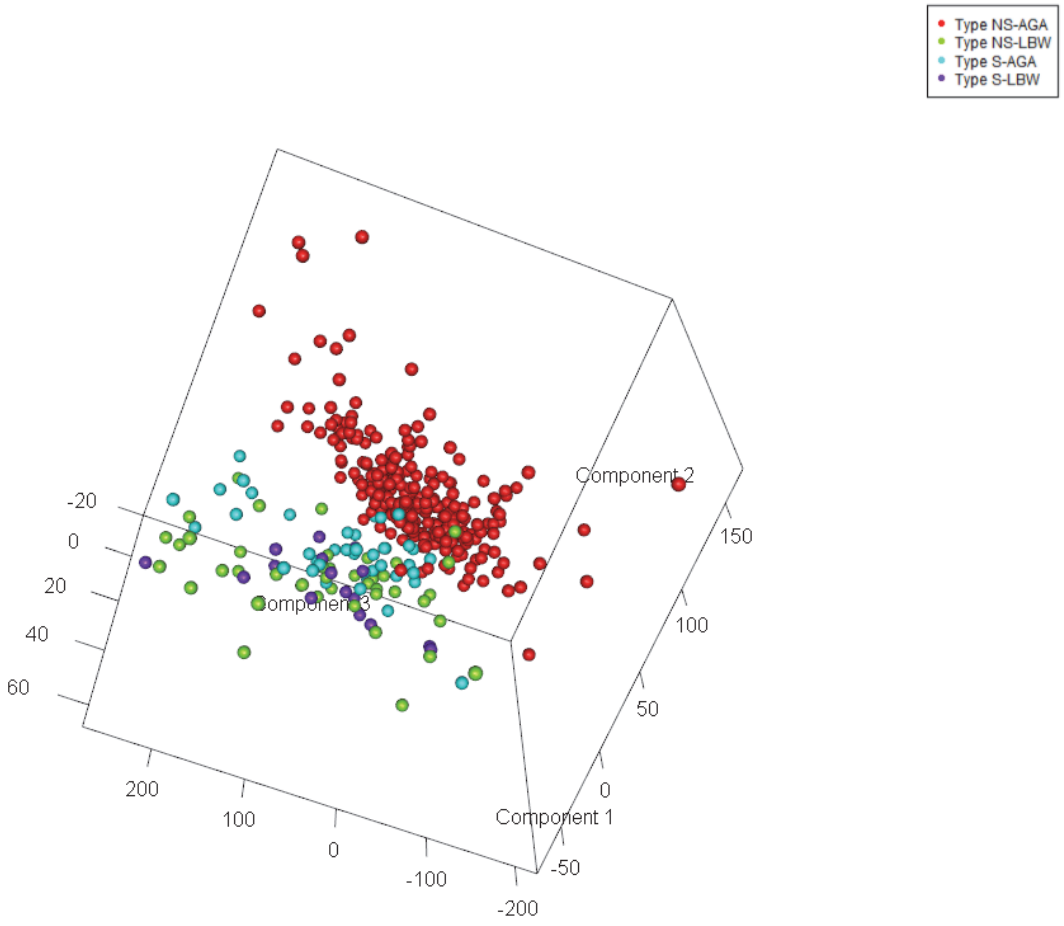


Figure 49: Sparse PLS-DA of cord DNA methylation data using typical-atypical categories.

We considered the matter of unbalanced samples (i.e. much smaller smoker-low birth weight group) and prediction error estimates. Rather than using MSEP, such situations may be better suited to estimating the Balanced Error Rate (BER). BER is appropriate in case of an unbalanced number of samples per class as it calculates the average proportion of wrongly classified samples in each class, weighted by the number of samples in each class (Cerf *et al.*, 2013). This is often compared to the overall error rate, which is simply the proportion of "correct" classifications across all the samples. As in Figure 46, we show results after conducting 10–fold CV but this time, using BER and overall error rate (Figure 50). While there is some improvement in overall error with adding more PLS-DA components, this is associated with a deterioration of BER. This observation in context of the results in Figure 49 likely represents worsening performance in the smaller classes (Cerf *et al.*, 2013).

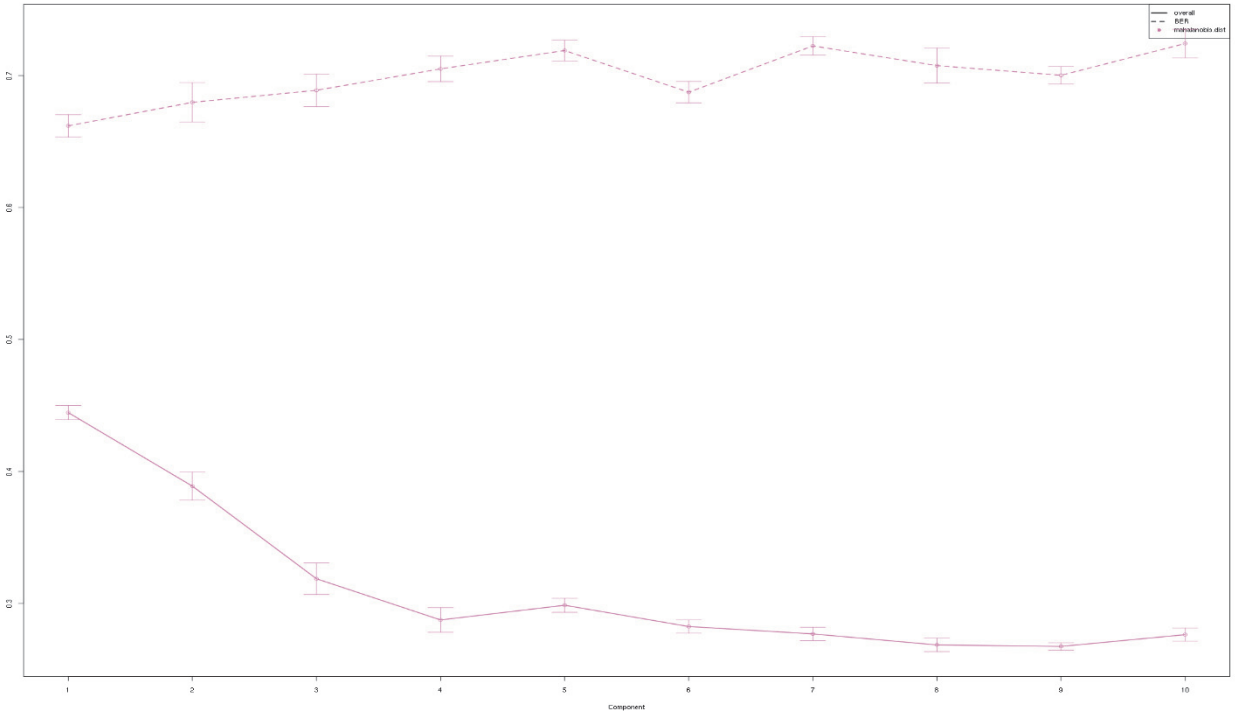


Figure 50: Sparse PLS-DA using typical-atypical MSP categories - Performance using 10 fold cross validation. Lines show global accuracy (continuous line) and balanced error rate (dashed line). X-axis: number of PLS-DA components tried.

3.2.3 Relation to covariates

Using these DNAm components derived from typical-atypical categories, we proceeded to evaluate their relation to covariates.

3.2.3.1 *Relation to cell type heterogeneity*

Cell type-attributed heterogeneity, along with genetic variation and age, are known to be major sources of variation in comparative blood-based DNAm studies (Jaffe & Irizarry, 2014; Tsai *et al.*, 2012). Looking at the correlation plot in Figure 51, we can see that Component 7 strongly relates to both CD4+ T cells and granulocytes. However, these cell counts are estimates based on an external cord blood reference. It appears the estimated values of these two cell types have a strong negative correlation.

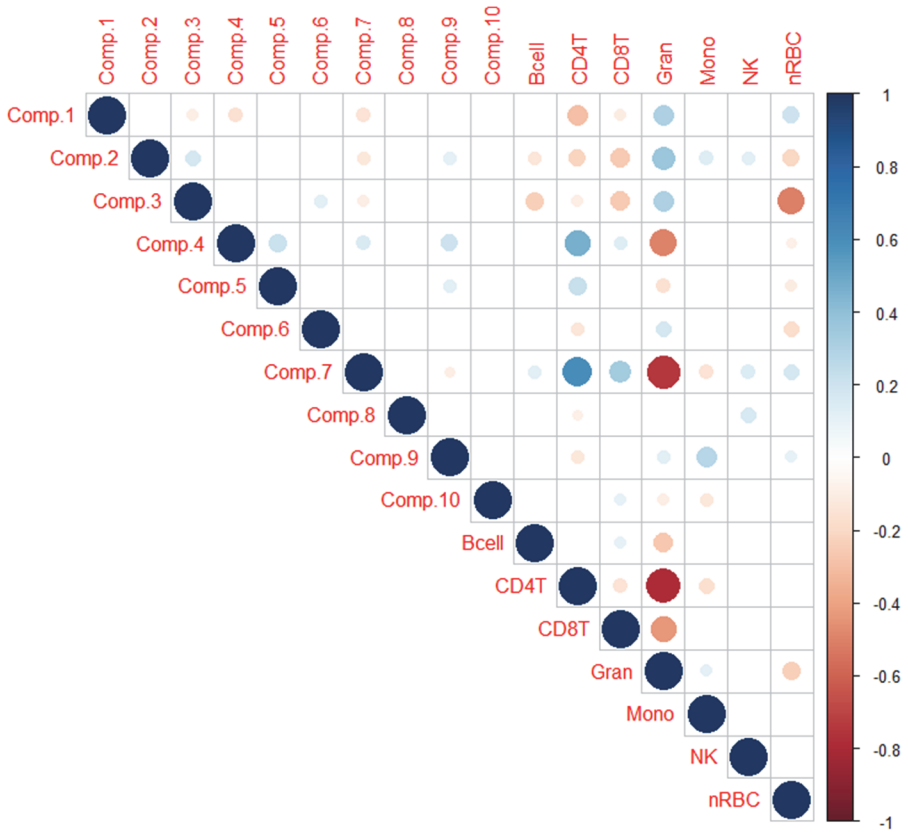


Figure 51: Correlation matrix. Cell count versus PLS-DA components from MSP-birth weight categories. Pearson correlations *p*-values are indicated by the circle diameter (i.e. larger diameter, lower *p*-values.) Colour scale indicates r value.

3.2.3.2 *Relation to infant sex*

As a reminder, we filtered probes on sex chromosomes from the DNAm data before PLS analysis. As well, the typical-atypical components used birth weight z-scores that were adjusted for gestational age and sex. Despite this, half of the components related to infant sex, Component 7 demonstrating the strongest relation. Recalling from above, it was also the component most correlated with cell type.

Table 14: ANOVA - Relation between infant sex and DNAm components (typical-atypical related).

	Comp1 (1)	Comp2 (2)	Comp3 (3)	Comp4 (4)	Comp5 (5)	Comp6 (6)	Comp7 (7)	Comp8 (8)	Comp9 (9)	Comp10 (10)
sex2	0.2*	-0.1	0.4*	0.4*	-0.1	0.2	-2.0***	0.4***	-0.1	0.2
	(0.1)	(0.1)	(0.2)	(0.2)	(0.1)	(0.1)	(0.2)	(0.1)	(0.2)	(0.1)
Constant	-0.1	0.1	0.1	-0.2	0.1	-0.04	0.8***	-0.2**	-0.1	-0.1
	(0.1)	(0.1)	(0.1)	(0.1)	(0.1)	(0.1)	(0.2)	(0.1)	(0.1)	(0.1)
Observations	916	916	916	916	916	916	916	916	916	916
R^2	0.01	0.001	0.01	0.004	0.000	0.004	0.1	0.01	0.000	0.003
Adjusted R^2	0.004	-0.000	0.01	0.003	-0.001	0.003	0.1	0.01	-0.001	0.002
Residual Std. Error	1.6	1.8	2.7	3.0	2.0	1.8	3.3	1.6	2.7	1.5
F Statistic	4.6*	0.6	5.9*	4.0*	0.3	3.7	84.9***	11.8***	0.3	2.7
Note:										*p<0.05; **p<0.01; ***p<0.001

3.2.3.3 *Relation to social factors*

Maternal education related to Component 2 (Table 15). In contrast, no component related to paternal social status (Table 16).

Table 15: ANOVA - Relation between maternal education and DNAm components (derived from typical-atypical mother-infant categories).

	Comp1 (1)	Comp2 (2)	Comp3 (3)	Comp4 (4)	Comp5 (5)	Comp6 (6)	Comp7 (7)	Comp8 (8)	Comp9 (9)	Comp10 (10)
matedu1	0.02	-0.5*	0.4	-0.2	-0.4	-0.2	-0.2	0.2	-0.2	0.1
	(0.2)	(0.2)	(0.3)	(0.4)	(0.3)	(0.2)	(0.4)	(0.2)	(0.4)	(0.2)
Constant	0.02	0.1*	0.3**	-0.04	0.1	0.1	-0.2	-0.01	-0.1	0.02
	(0.1)	(0.1)	(0.1)	(0.1)	(0.1)	(0.1)	(0.1)	(0.1)	(0.1)	(0.1)
Observations	897	897	897	897	897	897	897	897	897	897
R^2	0.000	0.01	0.001	0.000	0.003	0.001	0.000	0.001	0.001	0.000
Adjusted R^2	-0.001	0.005	0.000	-0.001	0.002	-0.000	-0.001	-0.000	-0.001	-0.001
Residual Std. Error	1.5	1.7	2.7	3.0	2.0	1.8	3.5	1.6	2.7	1.6
F Statistic	0.01	5.3*	1.1	0.2	2.5	0.8	0.1	0.6	0.4	0.3
Note:										*p<0.05; **p<0.01; ***p<0.001

Table 16: ANOVA - Relation between paternal social class (by occupation) and DNAm components (derived from typical-atypical mother-infant categories).

	Comp1 (1)	Comp2 (2)	Comp3 (3)	Comp4 (4)	Comp5 (5)	Comp6 (6)	Comp7 (7)	Comp8 (8)	Comp9 (9)	Comp10 (10)
patsocial2	0.1	-0.1	0.1	0.1	-0.3	-0.2	0.3	-0.1	-0.2	-0.1
	(0.2)	(0.2)	(0.3)	(0.3)	(0.2)	(0.2)	(0.4)	(0.2)	(0.3)	(0.2)
patsocial3	0.2	0.1	0.2	0.2	-0.7**	-0.3	0.3	-0.1	0.2	0.02
	(0.2)	(0.2)	(0.3)	(0.4)	(0.3)	(0.2)	(0.4)	(0.2)	(0.3)	(0.2)
patsocial4	-0.1	-0.3	0.1	0.1	-0.5*	-0.4	0.2	0.1	-0.1	0.1
	(0.2)	(0.2)	(0.3)	(0.3)	(0.2)	(0.2)	(0.4)	(0.2)	(0.3)	(0.2)
patsocial5	-0.5*	-0.4	-0.1	0.2	-0.3	-0.2	0.4	-0.02	0.3	-0.1
	(0.2)	(0.3)	(0.4)	(0.5)	(0.3)	(0.3)	(0.5)	(0.2)	(0.4)	(0.2)
patsocial6	0.2	-0.5	1.5*	1.7*	-0.5	-0.3	0.5	0.2	0.5	0.01
	(0.4)	(0.4)	(0.6)	(0.7)	(0.5)	(0.4)	(0.8)	(0.4)	(0.7)	(0.4)
Constant	0.02	0.3	0.2	-0.2	0.4*	0.3	-0.4	0.04	-0.1	0.02
	(0.1)	(0.2)	(0.2)	(0.3)	(0.2)	(0.2)	(0.3)	(0.1)	(0.2)	(0.1)
Observations	847	847	847	847	847	847	847	847	847	847
R^2	0.01	0.01	0.01	0.01	0.01	0.005	0.001	0.003	0.004	0.002
Adjusted R^2	0.01	0.005	0.001	0.001	0.004	-0.001	-0.005	-0.003	-0.002	-0.004
Residual Std. Error	1.5	1.8	2.7	3.0	2.1	1.8	3.5	1.6	2.8	1.6
F Statistic	2.0	1.8	1.2	1.1	1.6	0.8	0.2	0.5	0.7	0.4
Note:										*p<0.05; **p<0.01; ***p<0.001

3.2.4 Relation to clinical outcomes

We conducted the two-step approach to uncover DNAm components relevant to child outcomes as described in Methods (Section 2.5.5). We found Boruta frequently unable to select any important features. If any features were selected, the models performed poorly. As an example, the tables below provide the evaluation of models for school performance using the DNAm components related to typical-atypical categories. Below are the variables selected by Boruta and the RF performance metrics (Table 17 and Table 18, respectively.)

Table 17: Boruta selected variables (PLS-DA on cord DNAm using typical-atypical categories) for school outcomes.

Outcome	Boruta selected variables										
k1Math	patsocial										
k1Read	matedu	patsocial	Bcell	CD4T	CD8T	Gran	RBC	Comp1	Comp2	Comp3	Comp4
k1Sum	patsocial	CD8T									
k1Write	sex	patsocial	CD8T	NK	Comp8	Comp10					
k2Eng	Comp10										
k2Math	Comp7										
k2Sci	patsocial										
k3Eng	NULL										
k3Math	NULL										
k3Sci	NULL										

Table 18: Performance metrics for models of school performance. Components from DNA methylation at birth (PLS-DA using typical-atypical categories.)

Outcome	mtry	best.mtry	MSE	MSE.SD	R.squared	R.squared.SD
k1Math	2	4	0.4	0.02	0.01	0.02
k1Math	4	4	0.4	0.02	0.05	0.03
k1Math	6	4	0.4	0.02	0.04	0.03
k1Read	2	2	0.5	0.01	0.01	0.01
k1Read	9	2	0.5	0.02	0.02	0.03
k1Read	16	2	0.5	0.02	0.03	0.03
k1Sum	2	2	0.2	0.01	0.01	0.02
k1Sum	4	2	0.1	0.02	0.01	0.02
k1Sum	7	2	0.1	0.02	0.01	0.02
k1Write	2	2	0.3	0.02	0.05	0.03
k1Write	6	2	0.3	0.03	0.1	0.04
k1Write	11	2	0.3	0.03	0.1	0.04
k2Eng	2	2	0.02	0.00	-0.00	0.00
k2Eng		2				
k2Eng		2				
k2Math	2	2	0.04	0.01	0.02	0.01
k2Math		2				
k2Math		2				
k2Sci	2	6	0.05	0.01	-0.00	0.02
k2Sci	4	6	0.05	0.01	0.00	0.01
k2Sci	6	6	0.05	0.01	0.00	0.01

Metrics include MSE and R^2 (standard deviation in following column.) Results from different trials of *mtry* provided, with the most optimal *mtry* in the third column. When only one variable is selected, *mtry* can only be two. Eng – English, Sum – Summary score, Sci – Science.

A number of DNAm components passed Boruta selection, especially in the early school periods ("K1" stage in England.) However, among these, only Component 10 does not have a strong relation with cell type, sex or maternal education.

K2 English and Science models were very poor. K2 Math had only one selected variable, Component 7. Recall that this component related strongly to both granulocyte count and infant sex. Because no variables were selected for any K3 stage outcomes, we performed no further modeling and hence these are missing from the table.

3.3 DNAm vulnerability patterns using composite data

We now move to using our MSP composite as "bait". As the composite is a continuous measure, we used PLS regression (PLS-R) as opposed to discriminant analysis (PLS-DA). In addition, class imbalance is not a concern for continuous measures. Last, we now can use data on all 914 subjects with cord blood as we are not limited to only subjects whose mothers provided smoking data.

3.3.1 Analysis with PLS-R and MSP composite

Figure 52 provides the performance indices PRESS, R^2, MSEP and proportion of DNA methylation variability captured (Abdi, H. & Williams, 2013).

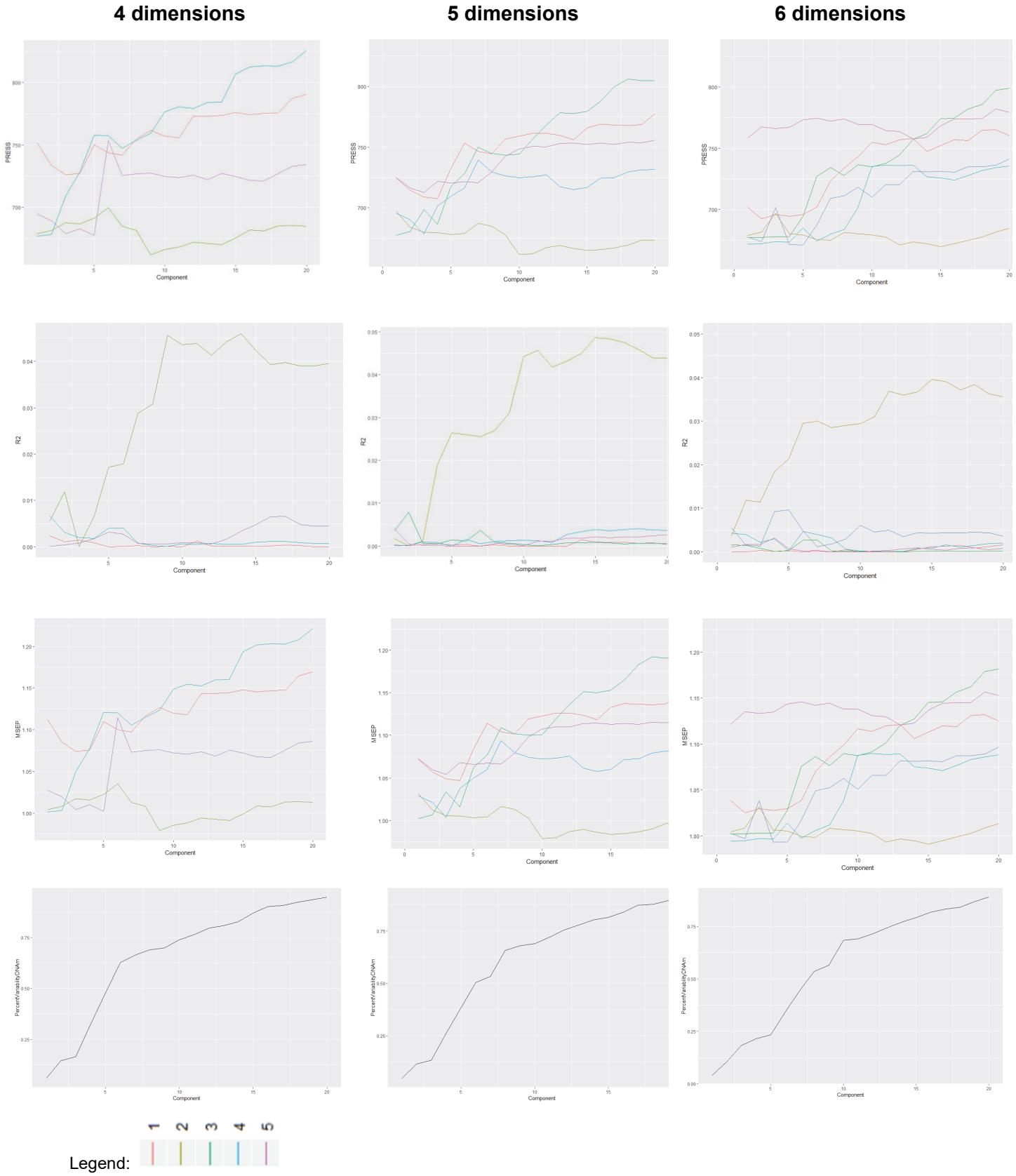


Figure 52: PLS regression performance metrics of relation between DNAm variability and MSP vulnerability composite (using 10-fold CV). Column name represents number of exposure dimensions used in PLS analysis. Each row graphs a different metric of fit versus number of DNAm components. Each line colour represents a different MSP composite. We trialed between 2 to 20 composite dimensions. However, for the sake of space, we show only results from trials using four, five or

six composite dimensions. (R function *perf* based on function of same name from package mixOmics implemented in package sgPLS.)

In terms of variability captured for the MSP composite and DNAm, we consider second and fourth rows of Figure 52. It appears that the composite-derived components provide a weaker relation to DNA methylation variability and composite dimensions than using the components derived from MSP alone or using the typical versus atypical categories. Looking at the predictive error (first and third rows), we do not see the expected decrease in error rate as number of components increase for most dimensions. However, unlike for the categorical MSP-derived components, the errors tend to stabilize for many of the dimensions instead of having an ongoing erratic pattern. This may suggest this model is less over-fit than the categorical MSP-derived models.

3.3.2 Relation to covariates

3.3.2.1 *Relation to cell type heterogeneity*

Looking only at correlations with $p < .05$ in Figure 53, the strongest association in cord blood was between component 3 and granulocyte count ($r = 0.3$.) There appears to be less correlation with cell type than compared with PLS-DA using typical-atypical categories (Figure 51).

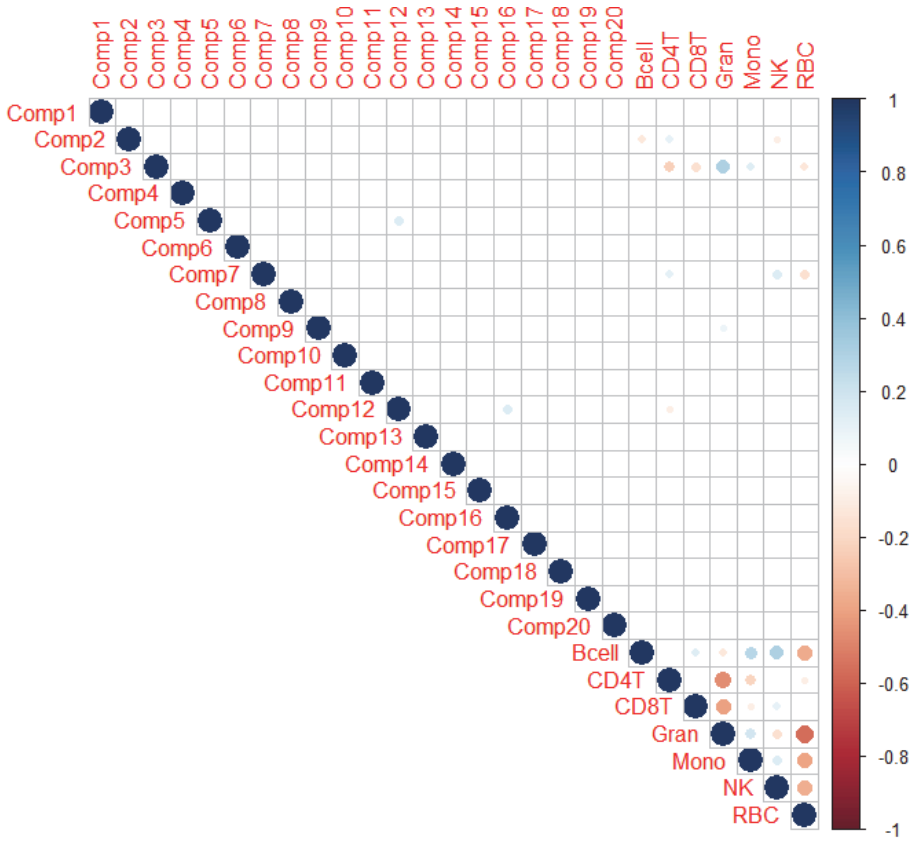


Figure 53: Correlation matrix between cord DNA methylation components (from MSP composite) and estimated cell type composition using meffil R package. Pearson correlations that have p<.05 are indicated with a circle. Colour scale indicates r-value.

3.3.2.2 *Relation to infant sex*

As seen in the PLS-DA results, subject sex had ongoing influence on DNAm components despite the removal of sex chromosomes from the DNAm data. This remains true for components from PLS analysis using the MSP composite. As seen in Table 19, Component 10 and to a lesser extent, Component 13, were associated with subject sex ($p < 0.05$).

Table 19: ANOVA between cord DNAm components and infant sex.

Relation between subject sex and latent variables -Cord

	Comp1 (1)	Comp2 (2)	Comp3 (3)	Comp4 (4)	Comp5 (5)	Comp6 (6)	Comp7 (7)	Comp8 (8)	Comp9 (9)	Comp10 (10)	Comp11 (11)	Comp12 (12)	Comp13 (13)	Comp14 (14)	Comp15 (15)	Comp16 (16)	Comp17 (17)	Comp18 (18)	Comp19 (19)	Comp20 (20)
sexlabel.xMale	-0.4	-0.6	-0.2	0.2	0.2	0.4	-0.2	0.2	-0.5	7.5***	-0.5	-0.1	0.9*	-0.5	0.3	-0.3	-0.2	-0.9	-0.3	-0.1
	(0.4)	(0.7)	(0.9)	(0.2)	(0.3)	(0.3)	(0.8)	(0.2)	(0.5)	(1.1)	(0.3)	(0.3)	(0.3)	(0.3)	(0.7)	(0.3)	(0.2)	(0.7)	(0.3)	(0.3)
Constant	0.3	0.1	-0.1	-0.1	-0.04	-0.3	0.2	-0.1	0.3	-3.1***	0.2	0.2	-0.5*	0.1	-0.2	0.2	0.2	0.5	-0.02	0.02
	(0.3)	(0.5)	(0.6)	(0.2)	(0.2)	(0.2)	(0.6)	(0.2)	(0.4)	(0.8)	(0.2)	(0.2)	(0.2)	(0.2)	(0.5)	(0.2)	(0.2)	(0.5)	(0.2)	(0.2)
Observations	914	914	914	914	914	914	914	914	914	914	914	914	914	914	914	914	914	914	914	914
R^2	0.001	0.001	0.000	0.001	0.001	0.003	0.000	0.001	0.001	0.05	0.002	0.000	0.01	0.003	0.000	0.001	0.001	0.002	0.001	0.000
Adjusted R^2	0.000	-0.000	-0.001	-0.000	-0.000	0.001	-0.001	-0.001	0.000	0.05	0.001	-0.001	0.01	0.002	-0.001	-0.000	0.000	0.001	0.000	-0.001
Residual Std. Error	6.3	11.0	13.5	3.4	3.8	3.8	12.2	3.5	7.7	16.5	4.9	5.2	5.3	4.7	10.1	4.4	3.5	10.8	4.7	4.2
F Statistic	1.1	0.7	0.1	1.0	0.6	2.3	0.1	0.5	1.1	47.0***	2.0	0.2	6.3*	2.7	0.3	0.9	1.1	1.5	1.1	0.2

Note: *p<0.05; **p<0.01; ***p<0.001

3.3.2.3 *Relation to social factors*

As we extracted the components from MSP-related factors that are correlated to social factors, we were interested to see what the relation was between variables such as maternal and paternal social factors as well as neighbourhood conditions, financial security and maternal psychopathology. There were varying degrees of missingness among these variables but social variables demonstrated no consistent pattern. For brevity, we only show the ANOVA tables for maternal education and paternal social status, (Table 20 and Table 21, respectively.)

Table 20: ANOVA - Maternal education by cord DNAm components.

Relation between maternal education and latent variables - Cord

	Comp1	Comp2	Comp3	Comp4	Comp5	Comp6	Comp7	Comp8	Comp9	Comp10	Comp11	Comp12	Comp13	Comp14	Comp15	Comp16	Comp17	Comp18	Comp19	Comp20
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)	(17)	(18)	(19)	(20)
matedu1	-1.2	-0.7	-1.7	-0.2	-0.3	0.1	-0.7	0.4	0.01	3.7	-0.03	-0.2	0.1	-0.8	0.3	0.6	0.5	0.5	-1.1	-0.6
	(0.8)	(1.4)	(1.7)	(0.4)	(0.5)	(0.5)	(1.6)	(0.5)	(1.0)	(2.2)	(0.6)	(0.7)	(0.7)	(0.6)	(1.3)	(0.6)	(0.5)	(1.4)	(0.6)	(0.5)
Constant	0.2	-0.1	0.03	0.05	0.1	-0.1	0.3	-0.1	-0.002	0.4	0.02	0.1	-0.1	-0.04	-0.2	0.003	-0.01	0.04	-0.1	0.005
	(0.2)	(0.4)	(0.5)	(0.1)	(0.1)	(0.1)	(0.4)	(0.1)	(0.3)	(0.6)	(0.2)	(0.2)	(0.2)	(0.2)	(0.4)	(0.2)	(0.1)	(0.4)	(0.2)	(0.1)
Observations	897	897	897	897	897	897	897	897	897	897	897	897	897	897	897	897	897	897	897	897
R^2	0.003	0.000	0.001	0.000	0.001	0.000	0.000	0.001	0.000	0.003	0.000	0.000	0.000	0.002	0.000	0.001	0.002	0.000	0.004	0.001
Adjusted R^2	0.001	-0.001	0.000	-0.001	-0.001	-0.001	-0.001	-0.000	-0.001	0.002	-0.001	-0.001	-0.001	0.001	-0.001	0.000	0.000	-0.001	0.002	0.000
Residual Std. Error	6.3	11.1	13.4	3.4	3.8	3.8	12.2	3.5	7.7	16.9	4.9	5.3	5.3	4.7	10.2	4.4	3.5	10.9	4.8	4.3
F Statistic	2.3	0.2	1.0	0.3	0.5	0.03	0.2	0.7	0.000	2.8	0.002	0.1	0.01	1.8	0.04	1.3	1.4	0.1	3.2	1.2

Note: $^{*}p<0.05$; $^{**}p<0.01$; $^{***}p<0.001$

Table 21: ANOVA - Paternal social status (as derived by occupation class) by cord DNAm components.

Relation between maternal education and latent variables - Cord

	Comp1 (1)	Comp2 (2)	Comp3 (3)	Comp4 (4)	Comp5 (5)	Comp6 (6)	Comp7 (7)	Comp8 (8)	Comp9 (9)	Comp10 (10)	Comp11 (11)	Comp12 (12)	Comp13 (13)	Comp14 (14)	Comp15 (15)	Comp16 (16)	Comp17 (17)	Comp18 (18)	Comp19 (19)	Comp20 (20)
patsocial2	0.3	-0.5	-0.2	0.4	-0.02	-0.6	0.2	0.2	-0.2	1.5	0.3	0.6	-0.2	0.1	0.9	-0.1	-0.3	0.1	-0.4	0.1
	(0.7)	(1.1)	(1.4)	(0.4)	(0.4)	(0.4)	(1.2)	(0.4)	(0.8)	(1.8)	(0.5)	(0.6)	(0.6)	(0.5)	(1.1)	(0.5)	(0.4)	(1.1)	(0.5)	(0.4)
patsocial3	-1.2	-1.4	-0.7	-0.1	0.5	-0.1	-0.03	-0.2	-1.9	1.8	-0.6	0.7	-0.5	0.7	2.1	0.1	-0.2	-0.2	-0.3	0.3
	(0.8)	(1.4)	(1.7)	(0.4)	(0.5)	(0.5)	(1.5)	(0.4)	(1.0)	(2.1)	(0.6)	(0.7)	(0.7)	(0.6)	(1.3)	(0.6)	(0.4)	(1.4)	(0.6)	(0.5)
patsocial4	0.3	-1.0	-0.2	0.3	0.04	-0.5	-0.8	-0.3	0.2	0.8	0.3	0.3	0.6	-0.2	2.2	-0.4	-0.2	1.5	-0.9	-0.2
	(0.7)	(1.2)	(1.5)	(0.4)	(0.4)	(0.4)	(1.3)	(0.4)	(0.8)	(1.8)	(0.5)	(0.6)	(0.6)	(0.5)	(1.1)	(0.5)	(0.4)	(1.2)	(0.5)	(0.5)
patsocial5	-1.2	-0.5	-0.9	-0.6	-0.9	-0.8	-0.9	0.2	0.5	-0.7	1.0	0.02	-0.5	0.8	1.5	0.8	-0.3	-0.4	-1.2	1.0
	(1.0)	(1.7)	(2.0)	(0.5)	(0.6)	(0.6)	(1.8)	(0.5)	(1.2)	(2.6)	(0.8)	(0.8)	(0.8)	(0.7)	(1.6)	(0.7)	(0.5)	(1.7)	(0.7)	(0.7)
patsocial6	-0.1	3.8	-1.5	-1.2	0.4	0.3	-1.7	-0.7	-0.3	-4.1	0.8	0.2	0.6	0.5	4.7	0.5	0.1	-0.4	-0.7	-1.4
	(1.5)	(2.7)	(3.2)	(0.8)	(0.9)	(0.9)	(2.9)	(0.8)	(1.9)	(4.0)	(1.2)	(1.3)	(1.3)	(1.1)	(2.5)	(1.1)	(0.8)	(2.6)	(1.2)	(1.0)
Constant	0.2	0.4	0.1	-0.1	0.1	0.3	0.5	-0.04	0.2	-0.1	-0.1	-0.3	-0.04	-0.2	-1.5	0.1	0.2	-0.5	0.4	-0.1
	(0.5)	(0.9)	(1.1)	(0.3)	(0.3)	(0.3)	(1.0)	(0.3)	(0.7)	(1.4)	(0.4)	(0.5)	(0.5)	(0.4)	(0.9)	(0.4)	(0.3)	(0.9)	(0.4)	(0.4)
Observations	847	847	847	847	847	847	847	847	847	847	847	847	847	847	847	847	847	847	847	847
R^2	0.01	0.01	0.001	0.01	0.01	0.01	0.002	0.004	0.01	0.004	0.01	0.003	0.01	0.005	0.01	0.01	0.001	0.004	0.01	0.01
Adjusted R^2	0.003	-0.001	-0.01	0.004	0.001	-0.000	-0.004	-0.002	0.002	-0.002	0.001	-0.003	-0.000	-0.001	0.002	-0.001	-0.005	-0.002	-0.000	0.002
Residual Std. Error	6.3	11.1	13.4	3.4	3.9	3.8	11.9	3.5	7.8	16.9	5.0	5.4	5.4	4.7	10.3	4.5	3.5	10.9	4.9	4.3
F Statistic	1.4	0.9	0.1	1.7	1.1	0.9	0.3	0.6	1.3	0.7	1.2	0.4	1.0	0.8	1.4	0.9	0.2	0.7	0.9	1.3

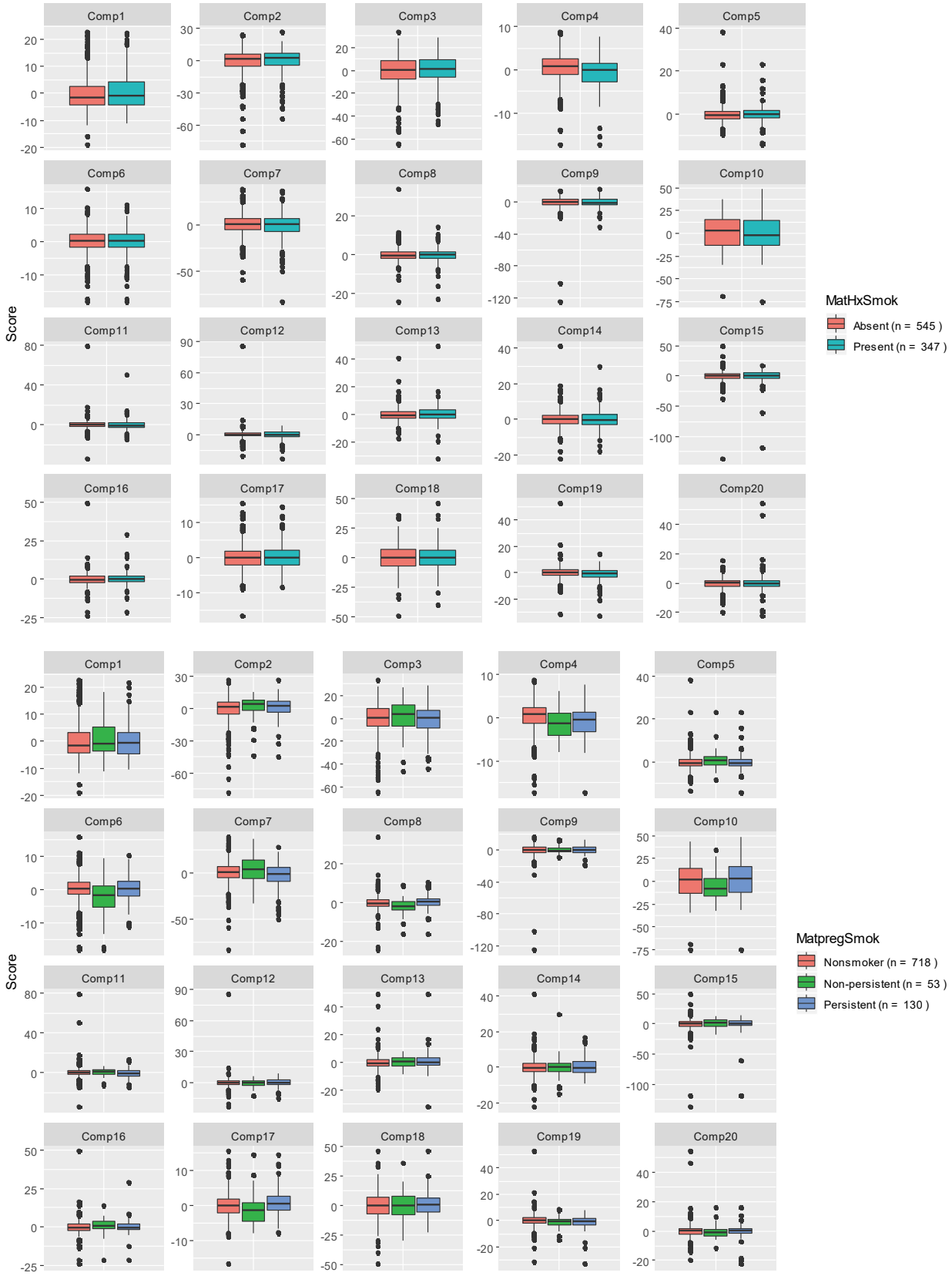
Note: *p<0.05; **p<0.01; ***p<0.001

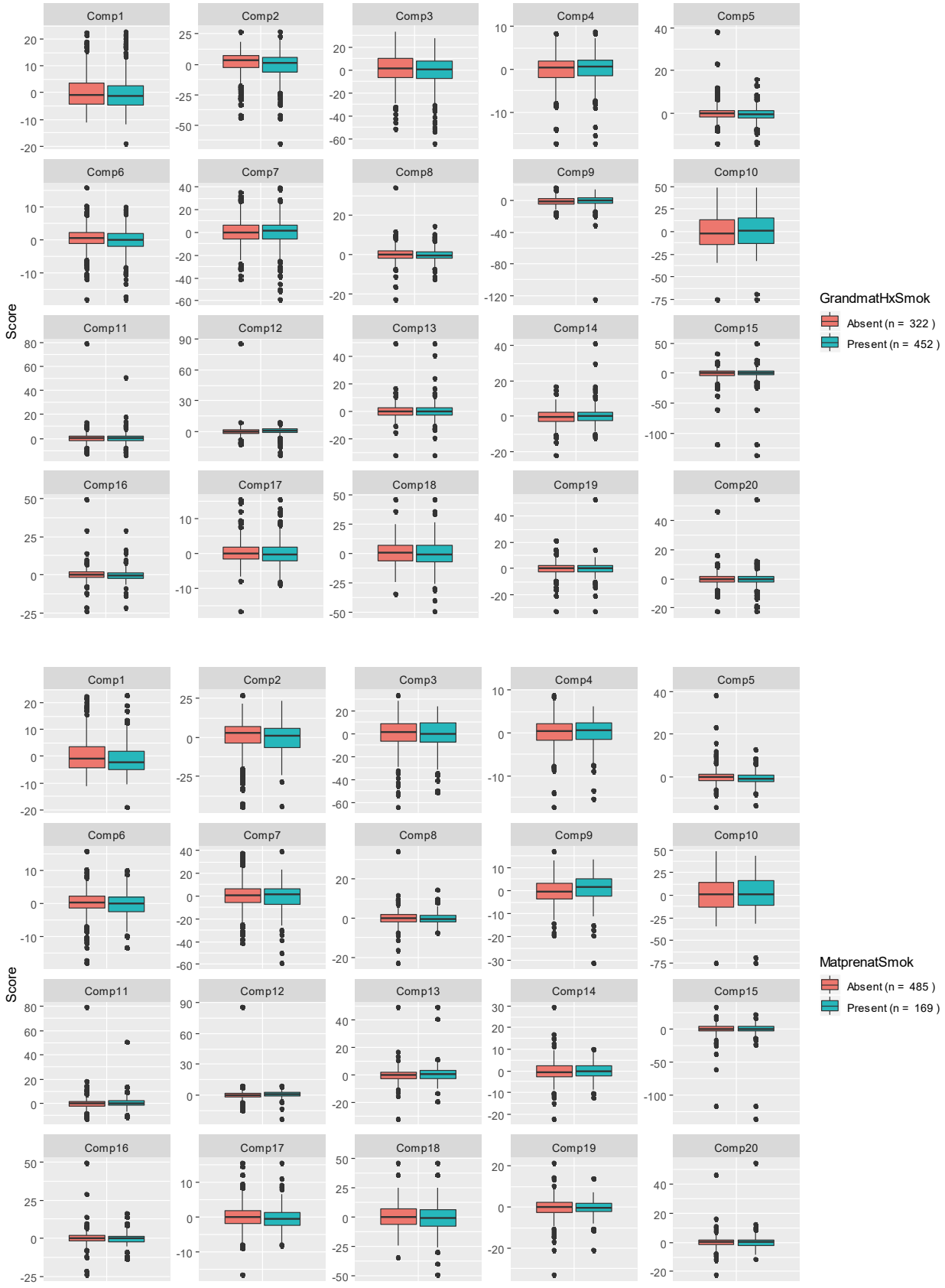
3.3.3 Relation to vulnerability composite

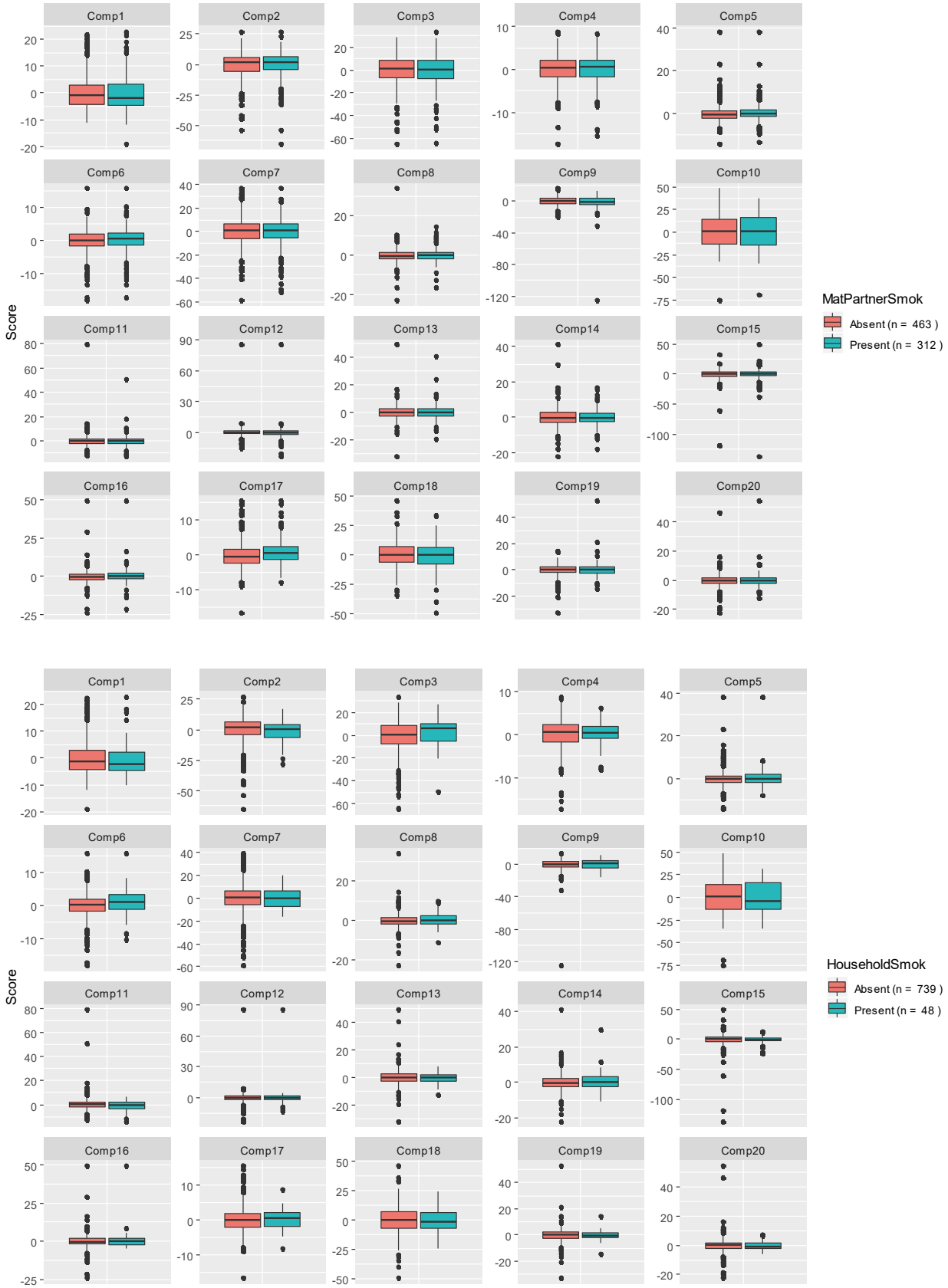
We further tuned the overlap between the MSP composite and DNAm to find the optimal number of dimensions and components to generate DNAm patterns. Again referring to Figure 52, we now look across the columns to find the number of dimensions that result in better PRESS, R^2, MSEP and proportion of variability captured. The 5-dimension solution appears to have slightly lower error rates with comparable variability capture. This would represent five clinical subtypes of infants' vulnerability to MSP-related factors – this would be a clinically reasonable figure. This is based on our *a priori* hypothesis that at least four possible extremes represent the spectrum between typical and atypical groups (see Figure 12). We also considered eigenvalues and variance explained (Figure 52) from the FAMD analysis. Thus, based on these results, as well as on composite features and theoretical basis, we proceeded with the 5-dimension composite solution.

We turn to the optimal number of components, testing number of components ranging from 2-50. Looking for the "hinge" points of each metric, we are seeking the point where increasing the number of components results in either a plateau or decrease in performance. Biologically, we also considered that each of the five dimensions of the exposure composite might arise from more than one biological process so we may expect the number of components to exceed that by multiples. Considering these points, we proceeded with the first 20 DNAm components.

We explored the relation of each component to the variables used to create the MSP vulnerability composite. The following figures plot the distribution of component scores based on these variables.


Comp1
Comp2
Comp3
Comp4
Comp5
Comp6
Comp7
Comp8
Comp9
Comp10
Comp11
Comp12
Comp13
Comp14
Comp15
Comp16
Comp17
Comp18
Comp19
Comp20
Score
MatHxSmok
Absent (n = 545)
Present (n = 347)
MatpregSmok
Nonsmoker (n = 718)
Non-persistent (n = 53)
Persistent (n = 130)


Comp1
Comp2
Comp3
Comp4
Comp5
Comp6
Comp7
Comp8
Comp9
Comp10
Comp11
Comp12
Comp13
Comp14
Comp15
Comp16
Comp17
Comp18
Comp19
Comp20
Score
GrandmatHxSmok
Absent (n = 322)
Present (n = 452)
MatprenatSmok
Absent (n = 485)
Present (n = 169)


Comp1
Comp2
Comp3
Comp4
Comp5
Comp6
Comp7
Comp8
Comp9
Comp10
Comp11
Comp12
Comp13
Comp14
Comp15
Comp16
Comp17
Comp18
Comp19
Comp20
Score
MatPartnerSmok
Absent (n = 463)
Present (n = 312)
HouseholdSmok
Absent (n = 739)
Present (n = 48)

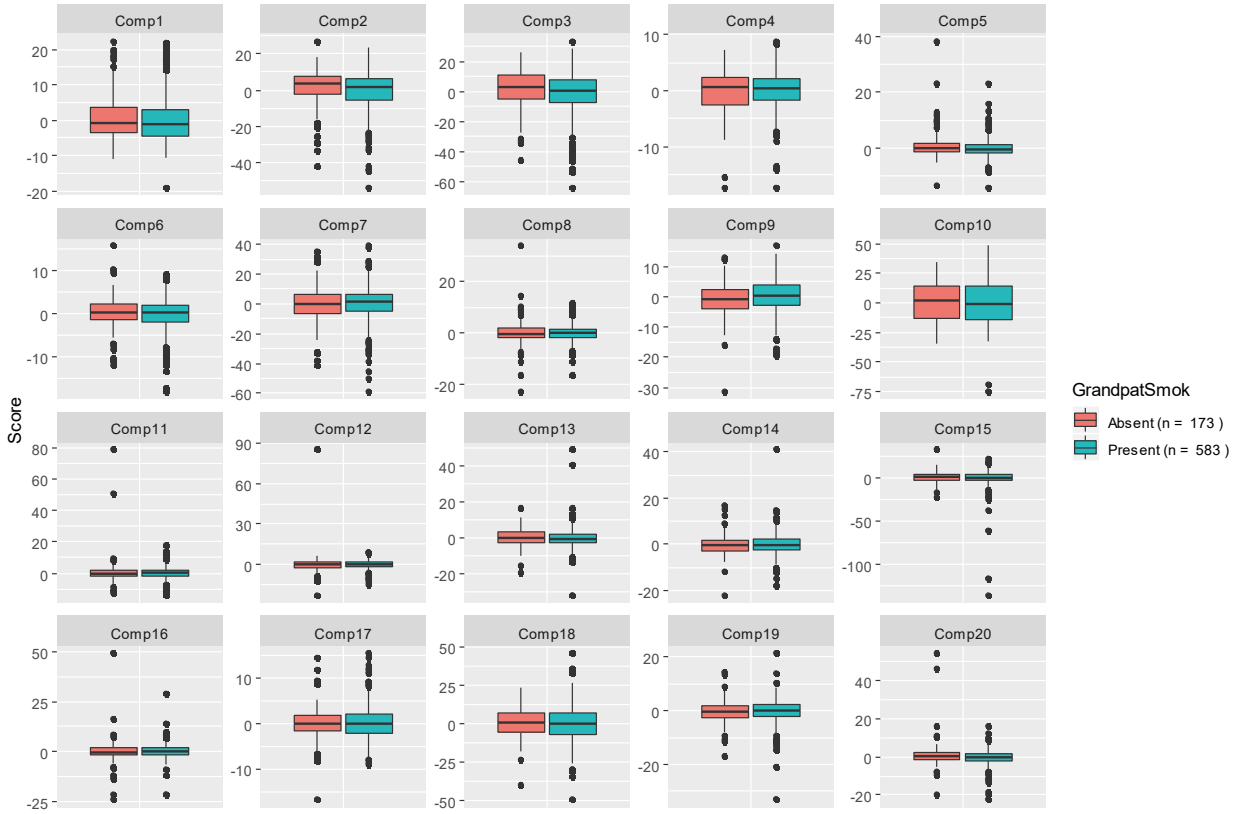


Figure 54: Boxplot of DNAm component scores versus variable (shown in legend) from MSP vulnerability composite. Legend: Report of this history was either Absent or Present for smoking of mother's partner (MatPartnerSmok), maternal history of smoking before pregnancy (MatHxSmok), maternal grandmother history of smoking (GrandmatHxSmok), maternal grandmother smoking while pregnant with mother (as reported by mother – MatprenatSmok), other household members besides parents who smoke (HouseholdSmok) and Maternal smoking during pregnancy. Sample size in brackets (after removing missing values for given variable.)

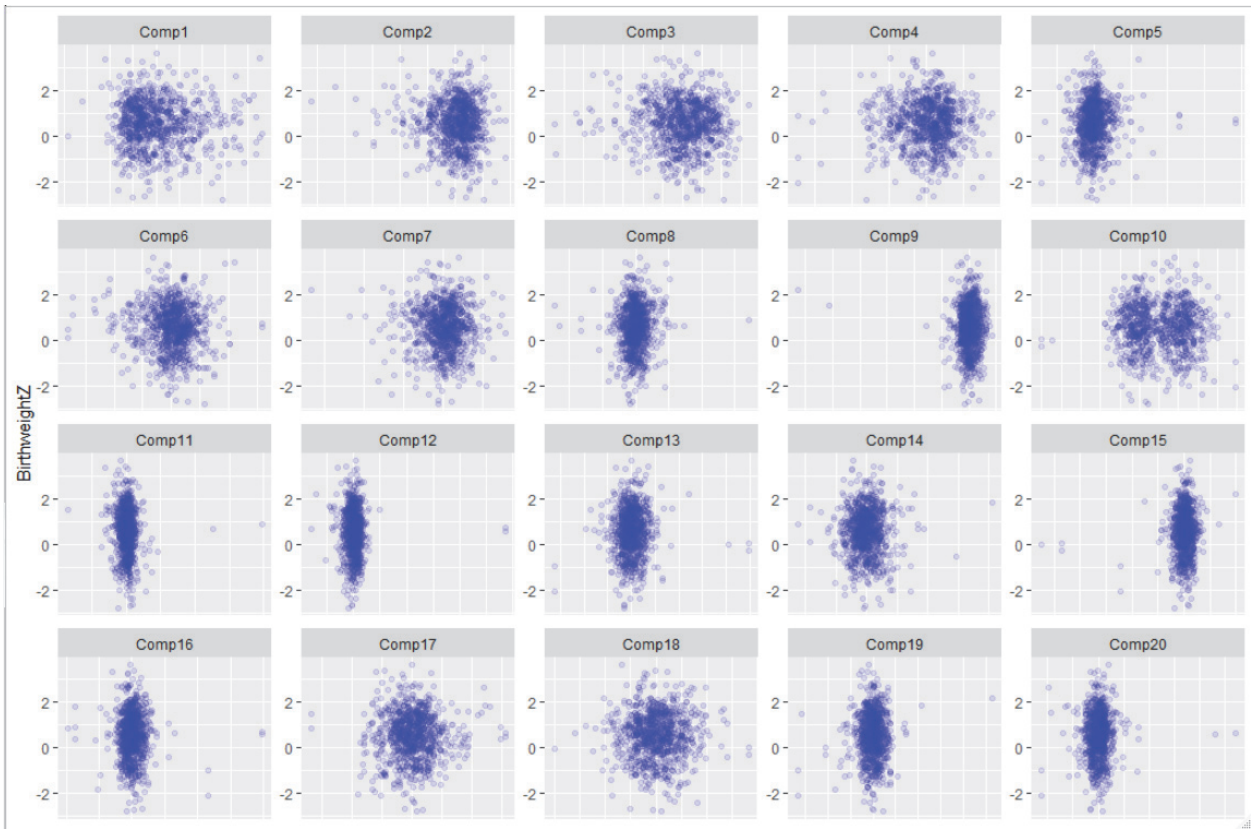


Figure 55: Scatterplot of DNAm component scores (x-axis) versus birth weight z-score (y-axis). N = 893.

We next sought to understand the MSP vulnerability subtype underlying each DNAm pattern. Looking at the Pearson correlation between each MSP dimension and the PLS components, Table 22 indicates the strongest relation between the two. The first MSP composite is most frequently represented (six components are most strongly related to this dimension.) In fact, the first 3 PLS components all represent this dimension the most (Figure 55). To remind the reader of the variables most representative of each dimension, Table 22 provides the contribution of constituent variables to each dimension (this is the tabular format of Figure 34)

Table 22: Relative variable contributions to each FAMD dimension representing MSP vulnerability

	Dim.1	Dim.2	Dim.3	Dim.4	Dim.5
BirthweightZ	0.0220139	6.691818	8.6639862	28.0441157	21.0907958
MatHxSmok	5.2459772	41.049266	0.5943018	1.6190297	0.7953177
GrandmatHxSmok	29.0232247	2.108929	9.8409139	2.9098393	0.0642217
MatprenatSmok	34.4921267	2.211654	1.4081536	0.0756129	3.3571985
MatPartnerSmok	11.5209179	0.631265	24.7175330	7.8570474	9.0555694
HouseholdSmok	5.2577192	1.038745	23.8554645	3.9020080	38.0508207
GrandpatSmok	10.9010463	1.898678	15.5767253	12.1614102	6.8105400
MatpregSmok	3.5369741	44.369646	15.3429217	43.4309368	20.7755362

This table displays the squared cosine*100/total squared cosine of the dimension to provide the percentage contribution.

Table 23: Top correlated FAMD dimension to DNA methylation component

Component	MSP vulnerability dimension
1	1
2	1
3	1
4	2
5	3
6	4
7	4
8	5
9	1
10	5
11	5
12	3
13	2
14	1
15	2
16	3
17	4
18	1
19	2
20	5

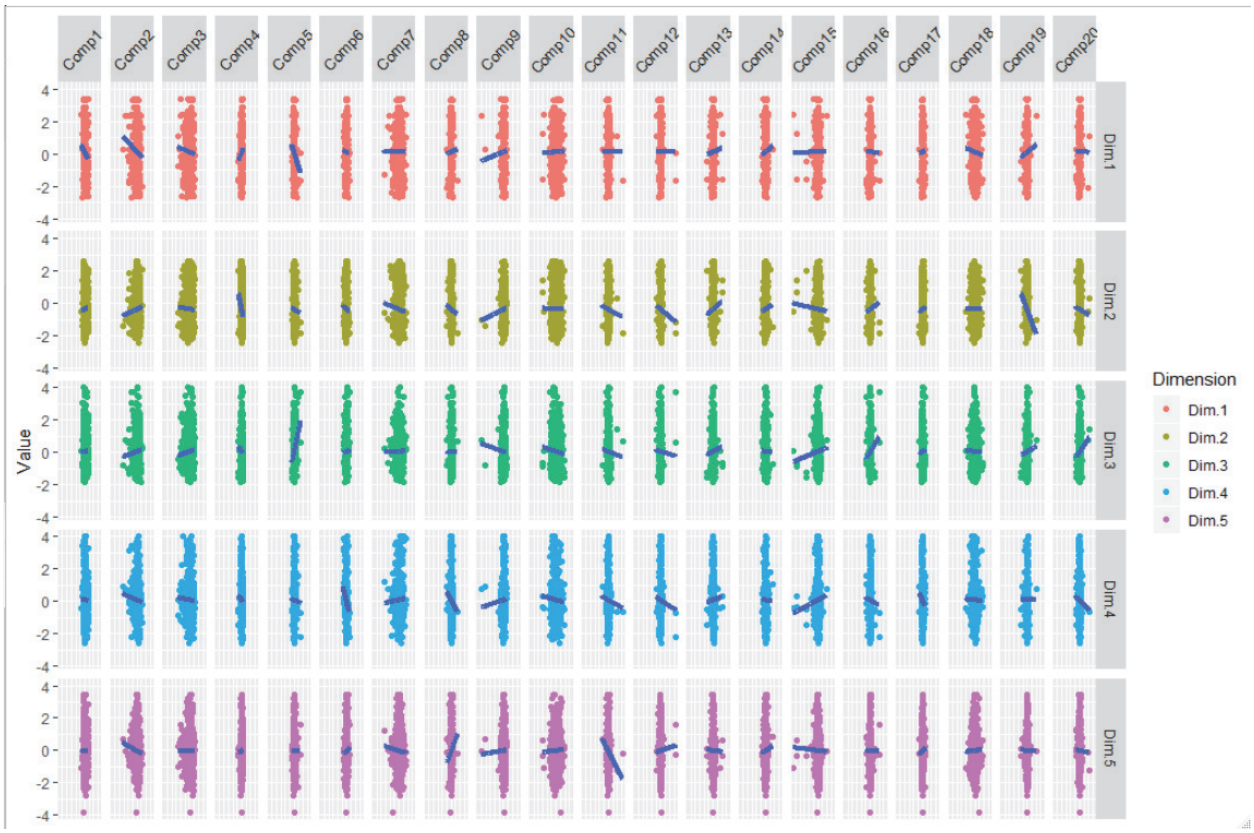


Figure 56: Scatterplot of DNAm scores (x-axis) versus MSP vulnerability dimension values (y-axis.) n = 914. Dark blue points indicate best fit line.

Recall that we used the *perf* function in the sgPLS package to perform 10-fold internal cross validation of our PLS model. This function provides insight into feature stability (from either the DNAm data or MSP dimensions) by indicating how often a feature is selected across all folds of validation. This offers another view of the relevance of dimensions to each component. Figure 57 shows the portion of folds that selected a given dimension for each component. With this view, we can see that while some components, (like Component 20,) "preferred" one dimension over others, other components selected 2 or more dimensions with no clear preference, (like Component 18.)

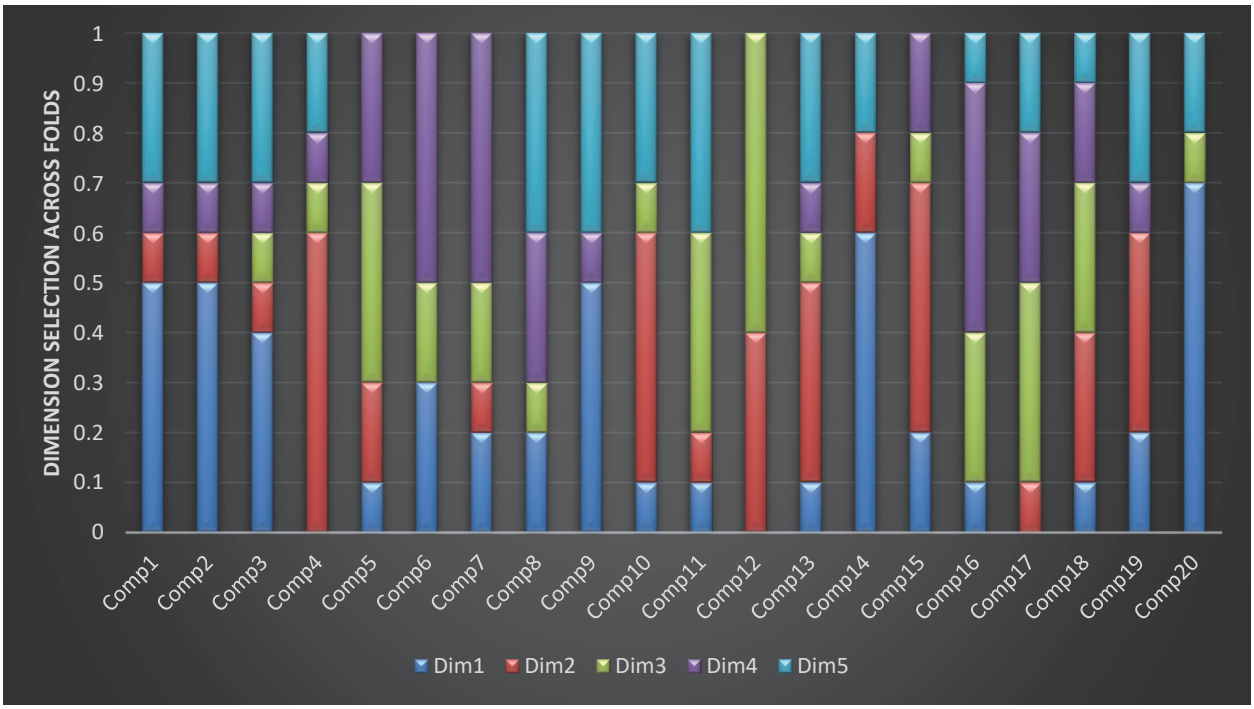


Figure 57: Barplot of feature stability measure of PLS model (from final tuning that used 5 dimensions and 20 components.) Y-axis is portion of times dimensions are selected across 10-fold internal cross validation by given component. Output from keepY result in perf function in sgPLS package.

The DNAm components extracted using the composite marker had more attenuated relations to covariates such as cell count, maternal education and sex compared the categorical MSP-derived components. In terms of performance, DNAm components captured less variability but appeared to less over-fit than the categorical-derived components. As such, we proceed with further analysis using composite-based DNAm patterns.

3.4 Clinical relevance of DNAm vulnerability patterns

In this chapter, we explore whether specific DNAm components predict child phenotype across diverse mental and physical outcomes. As described in Section 2.5.5, we used the two-step machine learning procedure to uncover if variables and which variables are relevant to predicting outcome.

The first step of this process is variable pre-selection with Boruta. Before we proceed, we wished to compare the effect of pre-selection versus no preselection. Metrics include the

number of trees needed to reduce error, R^2 and RMSE. A lower number of trees, higher R2 and lower MSE implies a better model. For the sake of space, we do not show model metrics for all the outcomes. As an example, we illustrate using waist circumference using cord DNAm components. Figure 58 shows the RF results with no preselection (i.e. all DNAm components and covariates are entered into the model.)

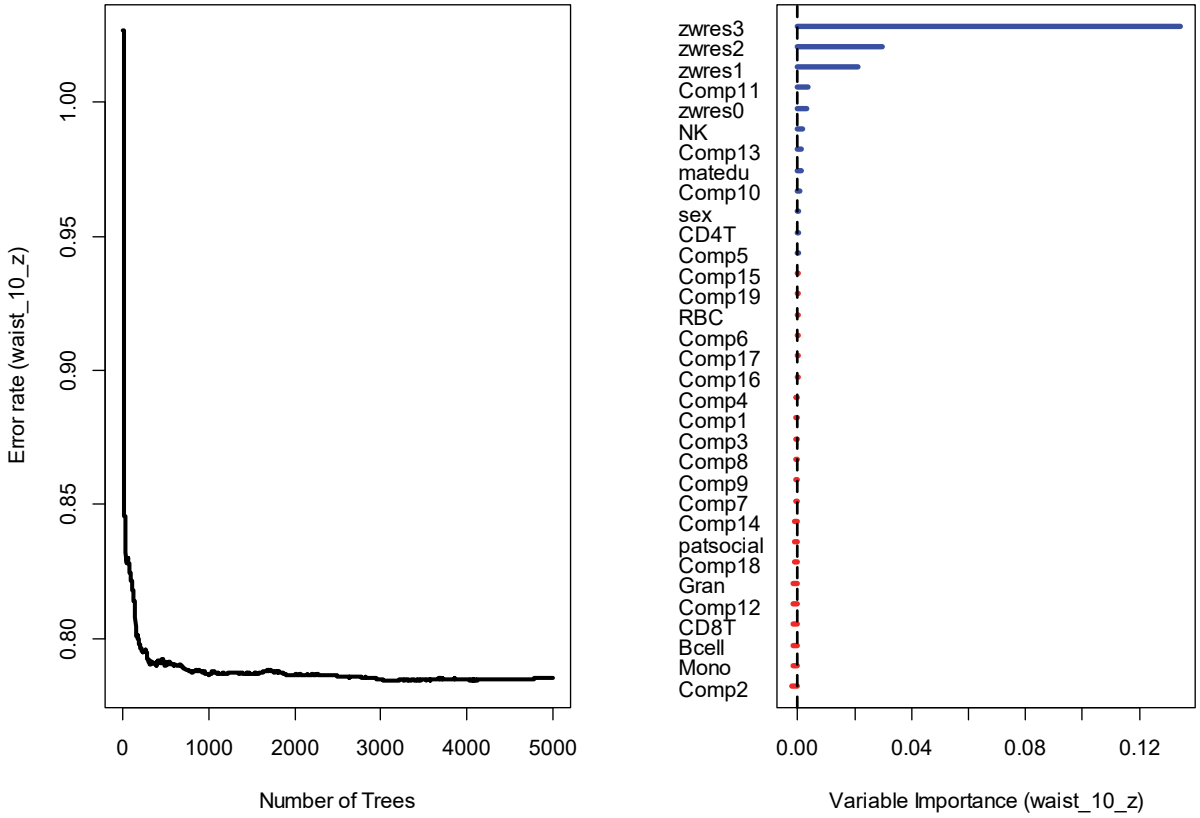


Figure 58: Random forest model of waist circumference (age 10). No preselection of variables. Variables: DNAm components from cord blood, sex, maternal education, paternal social status and early fetal and infant growth. Left: Error rate (calculated using OOB data) versus number of trees. Right: Variable importance (calculated using OOB data) versus variable importance (Breiman-Cutler method.) R package: rfsrc.

Second, we observe which variables are selected with Boruta (Figure 59). Variables to the right of the ShadowMax variable with green boxes consistently demonstrate relevance to the outcome that is likely unrelated to stochastic processes. Yellow boxes indicate "tentative" importance by the author of the package. As you may recall from Methods Section 2.5.4:

Contrast with biomarker discovery, the Boruta algorithm adds randomness by shuffling copies of all variables (which become the shadow variables) a specified number of runs. Variables labelled as tentative because RF could not decide whether the variable was more or less important than ShadowMax within the run number (in our case, 1000 runs.) We did not take these less certain variables forward in analysis. Therefore, in the case of waist circumference at age 10, the Boruta algorithm selected five variables as relevant: all early growth variables (including birth weight) and Component 11.

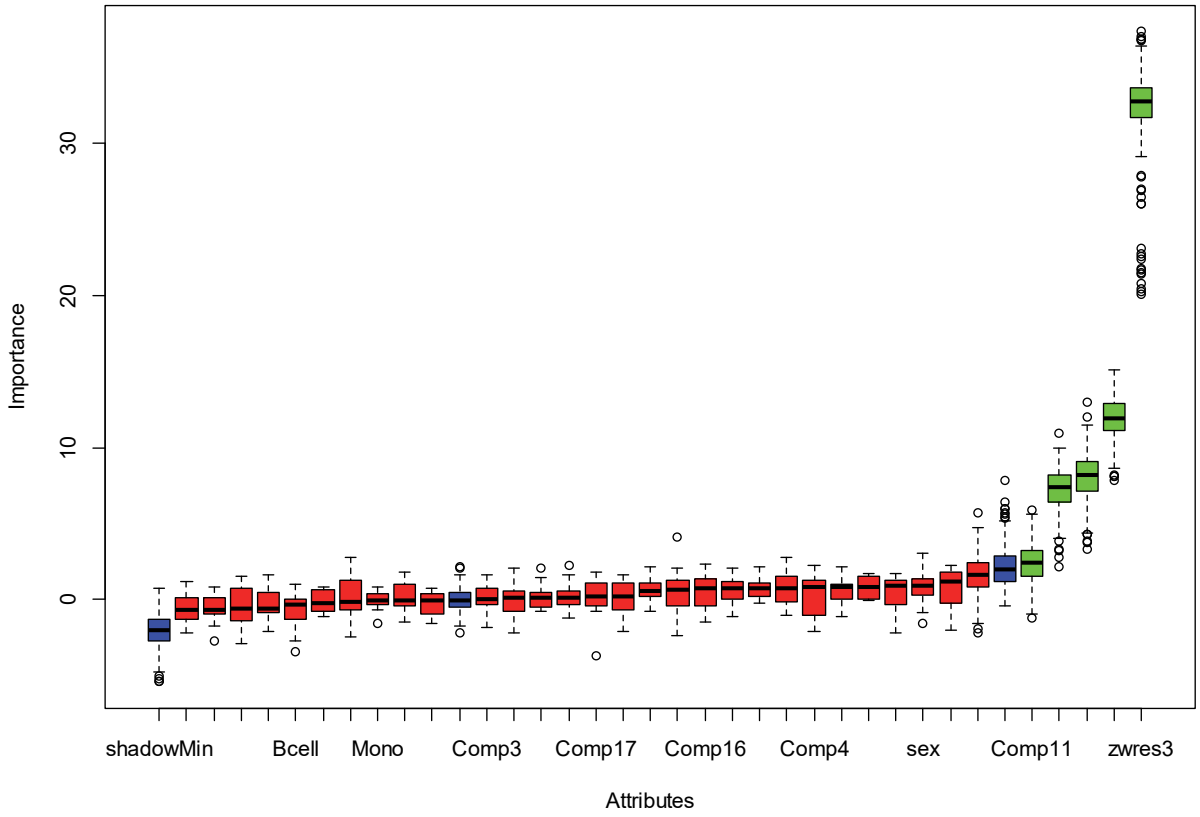


Figure 59: Boruta ranked variables for relevance to waist circumference at age 10 (n = 862.). Variables entered: DNAm components from cord blood, sex, maternal education, paternal social status and early fetal and infant growth. R package: Boruta.

Using only these Boruta selected variables, we conducted RF to obtain performance metrics (Figure 60). We note that compared to Figure 58, fewer trees are needed to drop the model error rate. Random forest models where there is more "certainty" in distinguishing between relevant and non-relevant variables require fewer decision trees to arrive at the final model.

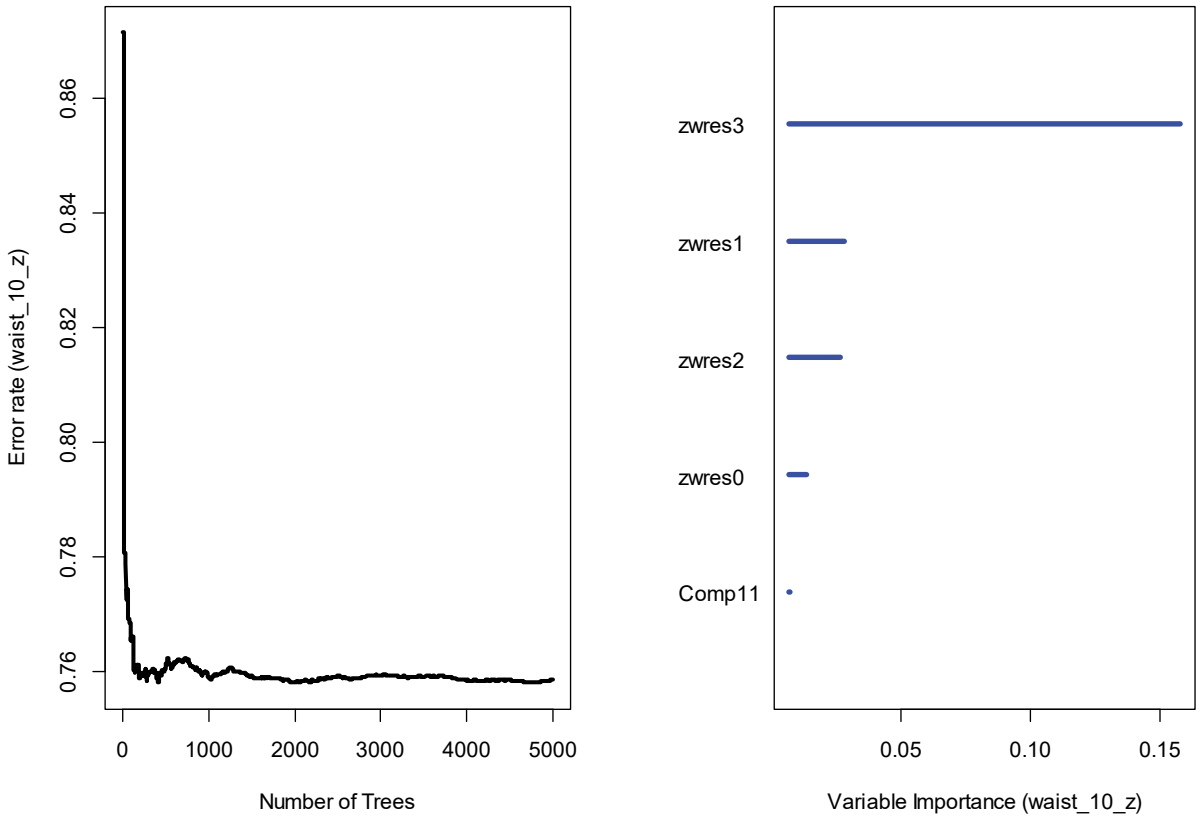


Figure 60: Random forest model of waist circumference (age 10). Boruta-selected variables only. Left: Error rate (calculated using OOB data) versus number of trees. Right: Variable importance (calculated using OOB data) versus variable importance (Breiman-Cutler method.) R package: rfsrc.

We compared RF models with and without Boruta feature selection (Table 25). Using the metric of variance explained and error rate, we can see that the error rate with and without Boruta is similar. However, the variance explained drops to 15.9% from 20.1% when modeling without Boruta. We observed overall better metrics with Boruta preselection across outcomes, regardless of the age of outcome measurement. As an example, Table 24 shows these metrics for waist circumference at various ages. In addition, computation times were much faster after Boruta selection. To maintain consistency, all results presented are from Boruta selected models.

Table 24: Comparison of performance between models with and without Boruta filter.

Age of measurement	Boruta filter?	*mtry*	RMSE	R^2

7	No	20	0.793	0.290
	Yes	2	0.786	0.314
9	No	20	0.880	0.171
	Yes	2	0.880	0.193
10	No	20	0.890	0.160
	Yes	2	0.879	0.194
11	No	20	0.867	0.165
	Yes	2	0.873	0.174

Outcome: waist circumference at various ages. Variables entered: cord DNAm components, estimated cell type composition, sex, weight trajectory (between birth and age 3), paternal social status and maternal education. R package: caret.

Like all RF algorithms, Boruta does not accept missing values. This necessary discarding of incomplete variables diminishes the sample size of models that are without Boruta preselection. Thus, we performed sensitivity analysis to check if these observations were related to the differences in sample size of Boruta-selected models versus those without pre-selection (e.g. n = 862 in the Boruta model compared with n = 805 without Boruta.) This is important for not only the power to detect relevant variables but also because excluded subjects change the impact of outlier and data noise effects on model specification. We repeated the analysis using the built-in impute function present in the randomForestSRC package. Table 25 shows a comparison between these three models, including with imputation (n = 914). The performance appears similar between the model without Boruta and the model with imputation. This makes the performance differences seen in Boruta-filtered models less likely due to change in sample size alone. Boruta filtered models perform better than either of the other two models in terms of percent variance explained and error rate.

However, excluded subjects change the impact of outlier and data noise effects on model specification. These results are at the right most column in Table 25. The metrics of RF with preselection and with data imputation demonstrate little difference (last two columns.)

Table 25: Random forest metrics - comparison of 3 models using cord blood DNAm patterns and waist circumference as outcome. Raw - Without Boruta preselection, Data imputation - With data imputation without Boruta preselection using built-in *impute* function. R package: rfsrc.

	Raw	Boruta filtered	Data imputation
Sample size	805	862	914
Number of trees	5000	5000	5000
Forest terminal node size	5	5	5
Average no. of terminal nodes	106.291	116.1642	120.7975
No. of variables tried at each split	11	2	11
Total no. of variables	33	5	33
Resample size used to grow trees	509	545	578
Number of random split points	10	10	10
% variance explained	15.95	20.15	15.32
Error rate	0.79	0.76	0.77

We also considered models to be likely over-fit if the metrics for the test set are dramatically worse than that of the training set. As an example, Table 26 shows the metrics for childhood anthropometric outcomes at various ages between 7 to 13 years (calculated using 5-fold cross validation with three repeats.) The test MSE and R2 values are slightly worse than the training values but remain close for most outcomes, making over-fitting less likely.

Table 26: Random forest performance with Boruta selected variables to predict child weight using DNAm components at Age 7 blood samples.

Outcome	Training.MSE	Training.R.squared	Test.MSE	Test.R.squared
weight_7_z	0.5	0.4	0.4	0.5
weight_8_z	0.6	0.4	0.5	0.3
weight_9_z	0.7	0.3	0.6	0.3
weight_10_z	0.7	0.3	0.6	0.3
weight_11_z	0.7	0.3	0.7	0.3
weight_13_z	0.7	0.2	0.8	0.2
height_7_z	0.2	0.8	0.2	0.8
height_8_z	0.2	0.7	0.3	0.8
height_9_z	0.3	0.7	0.3	0.7
height_10_z	0.3	0.6	0.3	0.7
height_11_z	0.4	0.6	0.4	0.5
height_13_z	0.4	0.5	0.5	0.5
bone_9_z	0.6	0.4	0.5	0.4
bone_11_z	0.7	0.3	0.6	0.3
bone_13_z	0.7	0.3	0.7	0.3
fat_9_z	0.7	0.2	0.8	0.2
fat_11_z	0.8	0.1	0.8	0.1
fat_13_z	0.8	0.1	0.9	0.02
lean_9_z	0.6	0.4	0.6	0.4
lean_11_z	0.7	0.3	0.7	0.3
lean_13_z	0.8	0.2	0.8	0.1
waist_7_z	0.7	0.3	0.6	0.3
waist_9_z	0.8	0.2	0.8	0.1
waist_10_z	0.8	0.2	0.8	0.1
waist_11_z	0.8	0.1	0.8	0.2

In the outcome column, numerical suffix indicates age at time of outcome measurement in years. Outcomes are all internally normalized by sex for the ALSPAC cohort. Outcomes are weight, height, bone mass (DEXA), fat mass (DEXA), lean mass (DEXA), and waist circumference.

As another safeguard against overfitting, we also checked the consistency of variable importance ranking after cross validation with rankings from Boruta and again using a different R package (randomForest). We excluded results from clearly poorly performing models (i.e. negative R2 and/or high MSE values), as well as models with highly divergent results in

importance ranking or test versus training metrics. We show in

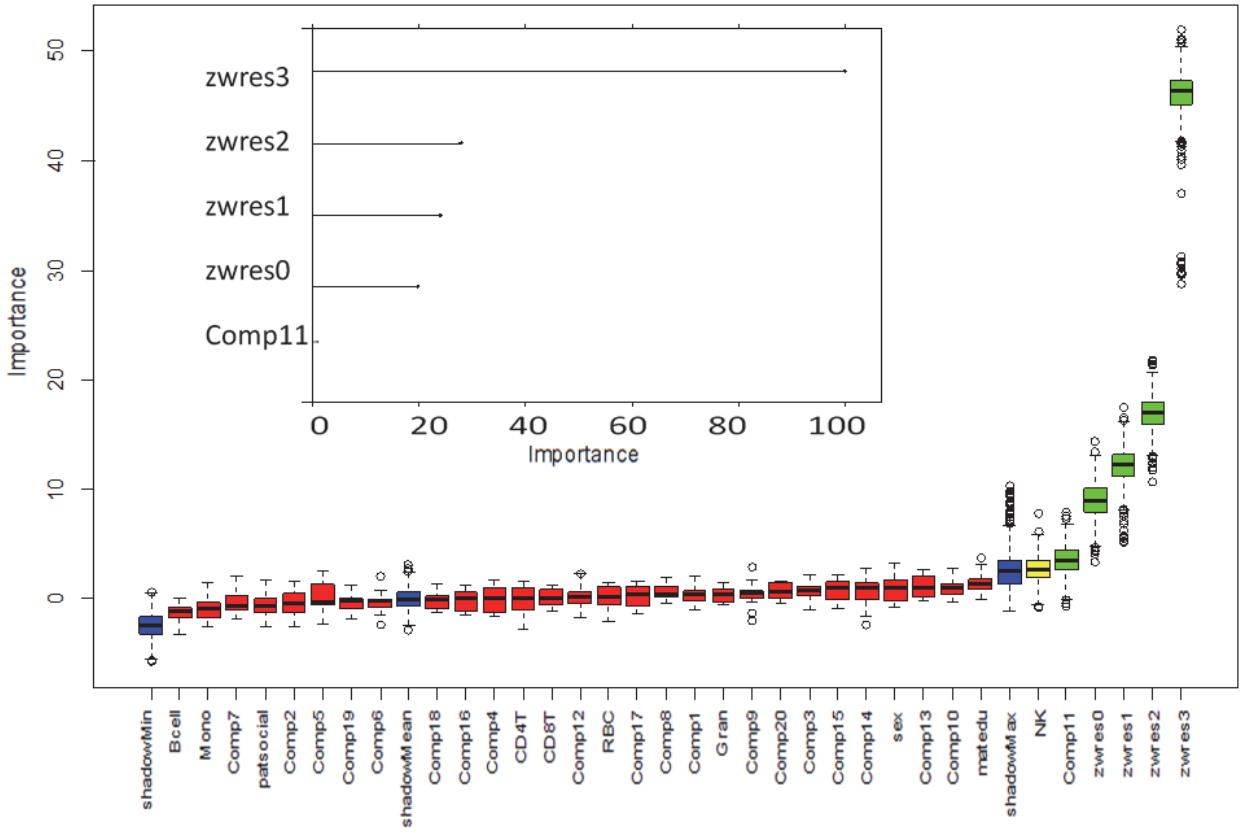


Figure 61 an example of these checks using the same outcome as in Figures 55 to 56 (waist circumference at age 10). We observe similar importance ranking after CV with repeats.

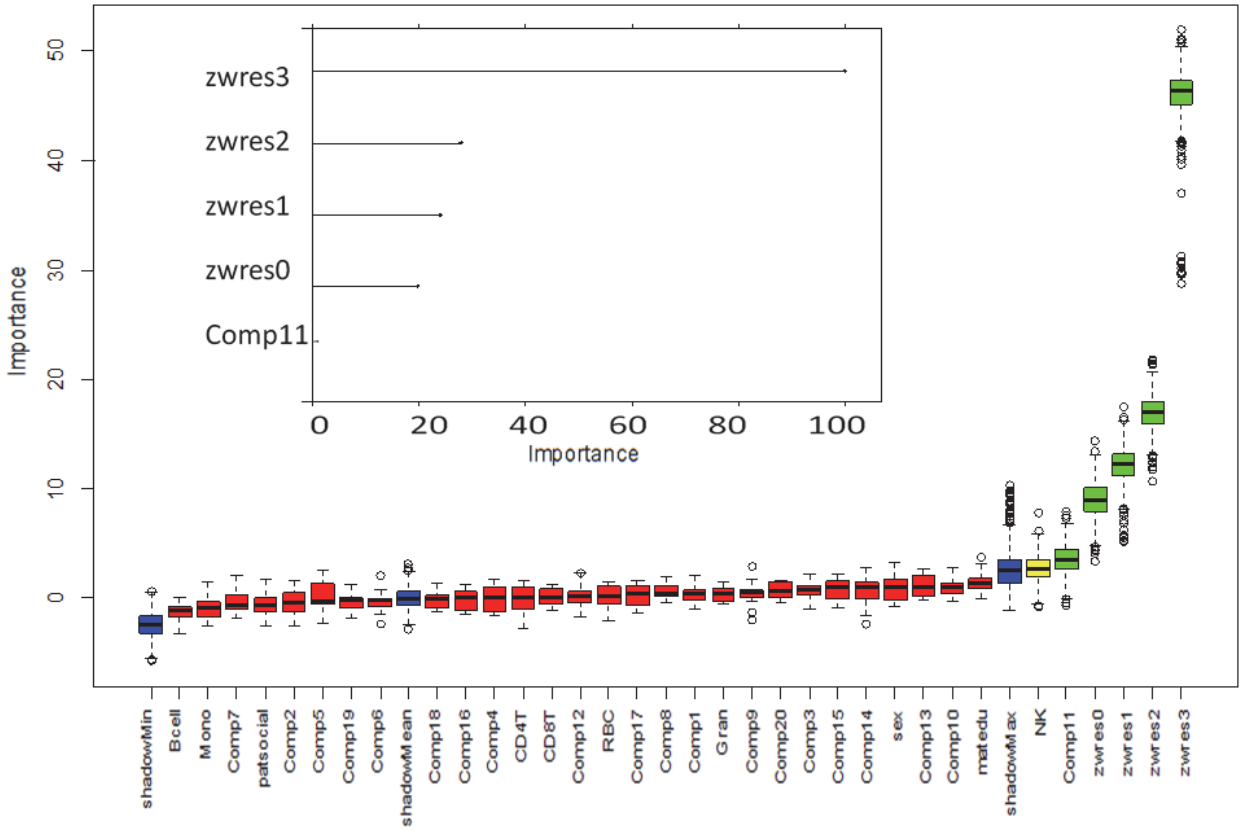


Figure 61: Plot of variable importance after 5-fold cross validation with three repeats. Outcome: waist circumference at age 10 (n = 862.) Variables: cord DNAm components, estimated cell type composition, sex, weight trajectory (between birth and age 3), paternal social status and maternal education. Ranking of importance of all variables (using package Boruta.) Inset: Ranking of importance of Boruta selected variables (using package randomforest.)

We now proceed with results for various physical and cognitive outcomes. We focus on results relating to cord DNAm components first but for the sake of space, tables and figures may include DNAm components derived from later ages that we will be refer to in subsequent sections.

3.4.1 Anthropometric outcomes

Table 27 summarizes the DNAm components selected by Boruta as relevant variables for weight, height and waist circumferences at various ages across childhood after passing the checks discussed at the introduction of this chapter. Only Component 11 and 14 passed this testing in cord blood DNAm, both of which relate to waist circumference.

Table 27: Boruta-selected DNAm components relevant to anthropometric measures at various ages.

	DNAm	Age 7	Age 8	Age 9
Weight	Age 15	*Comp11* *Comp7* Comp3	Comp2 Comp12 *Comp14* Comp3 Comp7	*Comp11* Comp2 Comp3 *Comp7*
	Age 7	Comp16 *Comp11* Comp12 Comp2 *Comp7* Comp3	*Comp9* Comp1 Comp2 *Comp7* Comp3	Comp16 *Comp9* *Comp7* Comp1 Comp2 Comp3
	Birth	Comp10		
Height	Age 15	*Comp5* Comp12 *Comp9* Comp1 Comp3 Comp7	Comp2 *Comp5* Comp12 *Comp9* Comp1 Comp3 *Comp7*	*Comp5* Comp1 *Comp9* Comp3 *Comp7*
	Age 7	*Comp11* Comp8 *Comp9* *Comp7* Comp3 Comp2 Comp1	*Comp11* Comp2 Comp3 *Comp7* *Comp9* Comp1	Comp8 Comp6 *Comp11* *Comp7* *Comp9* Comp3 Comp2 Comp1
	Birth			
Waist	Age 15	Comp2 *Comp14* *Comp9* *Comp11* Comp1 Comp3	Not measured	Comp1 *Comp9* Comp2 Comp3 *Comp7*
	Age 7	*Comp9* Comp1 Comp2 *Comp7* Comp3		*Comp9* *Comp7* Comp1 Comp3 Comp2
	Birth			

	DNAm	Age 10	Age 11	Age 13
Weight	Age 15	*Comp11* *Comp7*	*Comp5* *Comp14* *Comp11* *Comp7* Comp3	Comp2 *Comp14* Comp6 *Comp5* *Comp7* Comp3
	Age 7	Comp12 Comp16 *Comp9* Comp1 Comp2 *Comp7* Comp3	*Comp11* Comp8 Comp1 *Comp7* Comp2 Comp3	Comp16 Comp1 Comp3
	Birth			
Height	Age 15	Comp12 Comp1 *Comp5* *Comp9* *Comp7* Comp3	*Comp9* Comp1 *Comp7* Comp3	Comp19 Comp1 Comp2 Comp6 *Comp7* Comp3
	Age 7	*Comp11* Comp8 *Comp9* *Comp7* Comp3 Comp2 Comp1	*Comp11* *Comp9* Comp3 Comp2 *Comp7* Comp1	Comp19 Comp2 *Comp7* Comp1 Comp3
	Birth		Comp4	
Waist	Age 15	Comp1 *Comp11* Comp2 Comp3 *Comp7*	Comp1 *Comp14* *Comp7* Comp3	Not measured
	Age 7	*Comp9* Comp2 Comp1 *Comp7* Comp3	Comp1 *Comp7* Comp2 Comp3	
	Birth	*Comp11*	*Comp14*	

Outcome consists of weight, height and waist circumference (z-scores). This summary table only includes Boruta-selected DNAm components for the sake of clarity. Italics highlight components that persist at more than one DNAm collection age and/or more than one outcome. The columns indicate the child's age at the time of outcome measurement.

Consistent with previous literature (e.g. Bann *et al.*, 2014), the most important predictors of all anthropometric measures were birth size (length and weight) and rate of growth between birth and 3 years of life. These variables were consistently selected for all outcomes. There was also a relation between these growth outcomes and estimated leukocyte count, especially granulocytes and T cells. (For the sake of space, data not shown.)

We also looked at the results of DEXA scanning as a reliable means of estimating lean, fat and bone mass (Figure 62). Component 14 again reappears here, this time relating to lean mass at age 11. Like in Table 27, we summarize the results from the two-step analysis for body composition outcomes at various ages in Table 28.

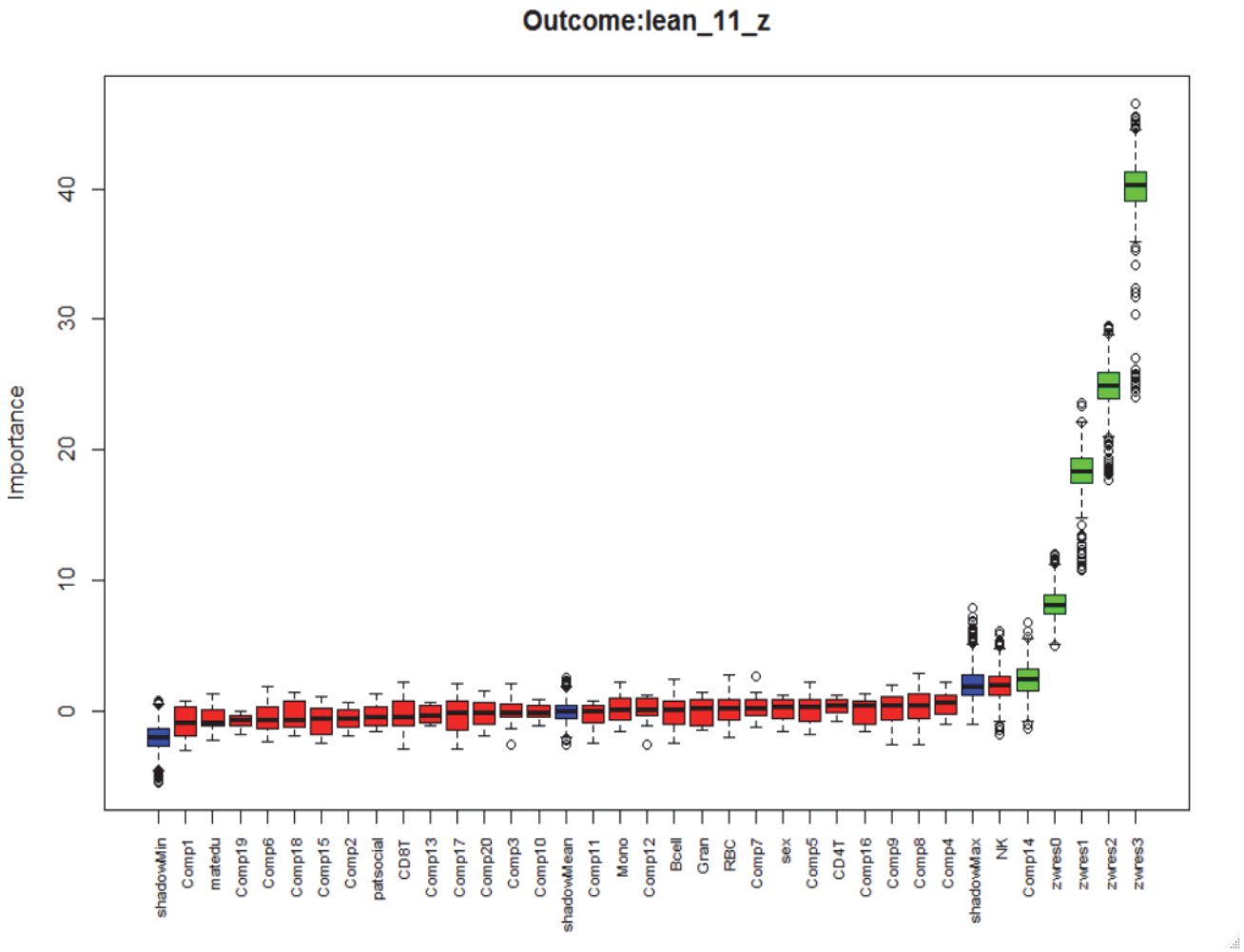


Figure 62: Boruta importance plot for lean mass (z-score) at age 11. Green boxes indicate selected variables.

Table 28: Boruta-selected DNAm components relevant to body composition at various ages.

	DNAm	Age 9	Age 11	Age 13
Lean	Age 15	Comp1 *Comp5* Comp6 *Comp9* Comp12	Comp1 *Comp5* Comp6 *Comp9* Comp12	Comp1 *Comp5* Comp6 Comp12 Comp19
	Age 7	*Comp11*	*Comp11* *Comp18*	Comp1 Comp2 Comp3 *Comp7* *Comp9*
	Cord		Comp14	
Fat	Age 15			Comp3 *Comp7*
	Age 7	Comp1 Comp2 Comp3 *Comp7*	Comp1 Comp2 Comp3 *Comp7*	Comp1 Comp2 Comp3 Comp6
	Cord			
Bone	Age 15	*Comp5*	Comp1 Comp3 *Comp5* *Comp7*	*Comp5* Comp6
	Age 7	Comp3 *Comp7*		Comp1 Comp3 *Comp7* Comp19
	Cord			

Outcome consists of lean, fat and bone mass as measured through DEXA scanning (z-scores). This summary table only includes Boruta-selected DNAm components for the sake of clarity. Italics highlight components that persist at more than one DNAm collection age and/or more than one outcome.

Knowing the importance of birth size to physical outcomes, we further explored these two components in context of birth weight. Partial dependence plots help visualize the interaction between DNAm components and variables in relation to outcomes. We show such a plot in that illustrates the predicted outcome values within quartiles of DNAm component scores while holding all other covariates at their average. Looking at Component 14 in the left column, we can see that around birth weight z-scores around or greater than negative one, lower Component 9 scores relate to higher lean mass. However, this separation by component score does not apply to lower birth weight infants. In the right column, Component 11 scores in the highest quartile relate to higher waist circumference but only within the average birth weight range. For lean mass and waist circumference, these plots demonstrate that the relation to DNAm patterns may not be uniform at all ranges of birth weight nor are they necessarily linear

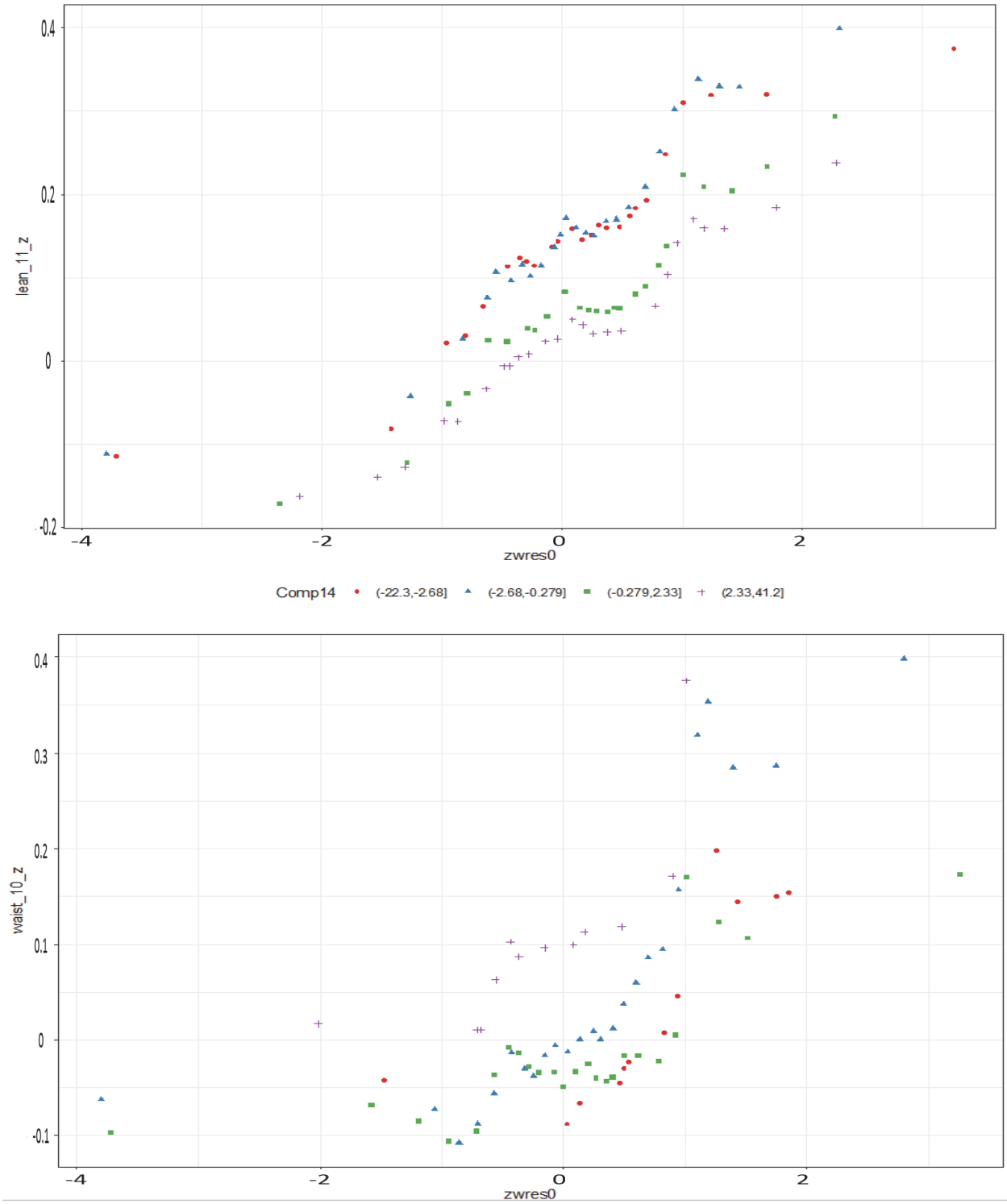


Figure 63: Partial dependence plot of outcome (y-axis) as a function of birth weight (z-score adjusted for sex and gestational age), conditional on quartile ranges of component scores from cord blood DNAm components. X-axis: Outcome as z-score (suffix indicates age in years.)

3.4.2 Neurodevelopment

Component 18 related to toddler development at 30 months. Components 8 and 11 were related to total IQ score (WISC) measured at age eight. Every model selected the paternal social status variable. All models for IQ at age four and development scores besides 30 months performed poorly, (results not shown.)

Table 29: Neurodevelopment outcomes - Random forest selected DNAm components

		DNAm Age								
Cognitive outcome - age 8	Total	Age 15	*Comp9*	*Comp18*	*Comp14*	Comp10	Comp3	Comp1	*Comp7*	Comp2
		Age 7	Comp6	Comp17	*Comp9*	Comp2	*Comp7*	Comp1	Comp3	
		Cord	Comp8	*Comp11*						
	Performance	Age 15	Comp10	*Comp9*	*Comp18*	*Comp14*	Comp1	Comp3	*Comp7*	Comp2
		Age 7								
		Cord	No components selected (except paternal social status)							
	Verbal	Age 15								
		Age 7	*Comp5*	*Comp14*	*Comp11*	*Comp9*	Comp2	*Comp7*	Comp1	Comp3
		Cord								
Development - 30 months		Age 15								
		Age 7	Comp15	Comp4	*Comp7*	Comp1	*Comp9*	Comp2	Comp13	Comp3	Comp10
		Cord	*Comp18*	Comp20						

This summary table only includes DNAm components. Grey boxes indicate possibly unreliable model.

3.4.3 Behaviour

Most models performed poorly (negative R^2 and high MSE). Better performing models demonstrated no consistent pattern of predictors, including DNAm components. Subject sex was the most common RF selected predictor. Results are omitted here for space considerations.

3.4.4 Academic performance

In cord blood, only DNAm Component 5 was selected more than once in relation to performance in Key Stage 2 and 3, (scores in English at both stages and Science at Key Stage 3.) Like for neurodevelopmental outcomes, paternal social status was selected in every model.

Table 30: Academic outcomes – Random forest selected DNAm components

		DNAm Age											
Key stage 1	Summary	Age 15	*Comp9*	Comp1	***Comp15***	*Comp18*	*Comp14*	Comp10					
		Age 7	Comp16	Comp13	Comp1	Comp10	Comp2	Comp3	***Comp7***				
		Cord	Comp20	*Comp14*	Comp2								
	Read	Age 15	Comp1	*Comp14*	*Comp9*	Comp10							
		Age 7	***Comp11***	Comp8	Comp16	Comp10	*Comp9*	Comp2					
		Cord	***Comp15***	*Comp14*	Comp19	Comp2							
	Write	Age 15	Comp4	***Comp15***	Comp1	*Comp18*	*Comp14*	Comp10					
		Age 7	Comp8	Comp16	Comp1	Comp13	***Comp7***	*Comp9*	Comp2	Comp3	Comp10		
		Cord	Comp17	Comp13									
	Math	Age 15	*Comp9*	Comp12	Comp3	*Comp18*	*Comp14*	Comp2	***Comp7***				
		Age 7	***Comp11***	Comp2	Comp1	***Comp7***	Comp3						
		Cord	Comp19	Comp2	***Comp15***								
Key stage 2	Math	Age 15	Comp4	*Comp14*	Comp1	Comp2	***Comp7***	Comp3					
		Age 7	Comp10	Comp13	*Comp9*	Comp2	Comp1	***Comp7***	Comp3				
		Cord	Only maternal education and paternal social status selected.										
	English	Age 15	Comp12	Comp6	Comp3	***Comp7***	*Comp18*	*Comp14*	Comp2	Comp1	Comp4	*Comp9*	Comp10
		Age 7	Comp4	Comp13	*Comp9*	Comp1	Comp2	***Comp7***	Comp3	Comp10			
		Cord	***Comp5***	Comp16	*Comp9*	Comp12	Comp2						
	Science	Age 15	***Comp11***	Comp4	*Comp9*	Comp1	Comp3	***Comp7***	Comp2				
		Age 7	Comp6	***Comp15***	Comp13	Comp10	*Comp9*	Comp1	Comp2	***Comp7***	Comp3		
		Cord	Comp20										
Key stage 3	Math	Age 15	***Comp11***	*Comp18*	*Comp9*	Comp3	***Comp7***	Comp1	Comp2				
		Age 7	***Comp15***	Comp6	Comp16	***Comp11***	Comp2	Comp1	***Comp7***	Comp3			
		Cord	***Comp11***										
	English	Age 15	Comp6	Comp3	*Comp18*	Comp4	*Comp9*	*Comp14*	Comp1	Comp10			
		Age 7	Comp16	*Comp9*	***Comp7***	Comp1	Comp3	Comp2	Comp13	Comp10			
		Cord	***Comp5***	Comp2	Comp12								
	Science	Age 15	Comp1	***Comp11***	*Comp14*	Comp12	*Comp9*	Comp2	Comp3	Comp8	Comp7		
		Age 7	*Comp9*	Comp12	Comp13	Comp2	***Comp7***	Comp1	Comp3				
		Cord	***Comp5***										

Academic outcome consists standardized testing performance at Key stage 1 (age 5-7 years), 2 (age 8-11 years) and 3 (age ~14 years). This summary table only includes Boruta-selected DNAm components for the sake of clarity. Italics highlight components that persist at more than one DNAm collection age and/or more than one outcome.

We note the presence of some DNAm patterns that were selected no matter the age of blood collection or clinical outcome, (e.g. Components 1, 2 and 3.) However, recall from the preceding covariate analysis that these components were the most related to cell type estimates. As well, Component 10 appears frequently but is also related to infant sex. This is in contrast to other components that appear less frequently but do not have strong relations to covariates (e.g. Component 9 and 11.) We will be focusing further analysis on this latter type of component.

3.4.5 Cell count composition revisited

We diverge at this time to recall that DNAm data from ages 7 and 15 years do not have a DNAm reference. As such, our analysis used a reference-free correction method (reFACTor) to mitigate the effects of cell type heterogeneity. To evaluate the difference between meffil and reFACTor, we repeated the correlation analysis of cord DNAm this time with reFACTor estimated cell type composition. Unlike the reference based cell type correction, several components stand out with strong correlation to estimated cell counts. We also performed two-step RF to observe the effect on relation of outcomes to cord blood components corrected with ReFACTor. Models were comparable to that with reference-based correction (Appendix C). We further compared component, dimension and estimated cell composition using reFACTor versus meffil (using adult reference "blood gse35069") and found only minor differences (Appendix C).

We were also wary of the relation between the MSP composite and cell count given the signal strength of this covariate in DNAm data. As the MSP composite forms the basis of construction of the DNAm components, we wanted to consider this possibility seriously. Fortunately, the two appear to have little correlation (Appendix C).

3.5 Comparison of performance: DNAm patterns versus MSP variables or composite in relation to child outcomes

We wanted to compare the ability of DNAm patterns to predict outcomes versus using the maternal report of smoking throughout pregnancy. Using the same two-step variable selection process, we entered the same control variables (i.e. sex and social factors) and instead of any DNAm variables, we used the maternal smoking variable as described in Methods and as used in the MSP vulnerability composite. In many cases, we found that the maternal smoking was not even selected by Boruta for further analysis. The sample size was often smaller as maternal smoking data was unavailable for all subjects whereas DNAm patterns can be

extracted for subjects with missing maternal data. We show an example using waist circumference in Figure 64. Like with DNAm component analysis (Figure 59), the top ranking variables are growth rate variables, with growth between age 1-3 years being the most important variable by a large margin.

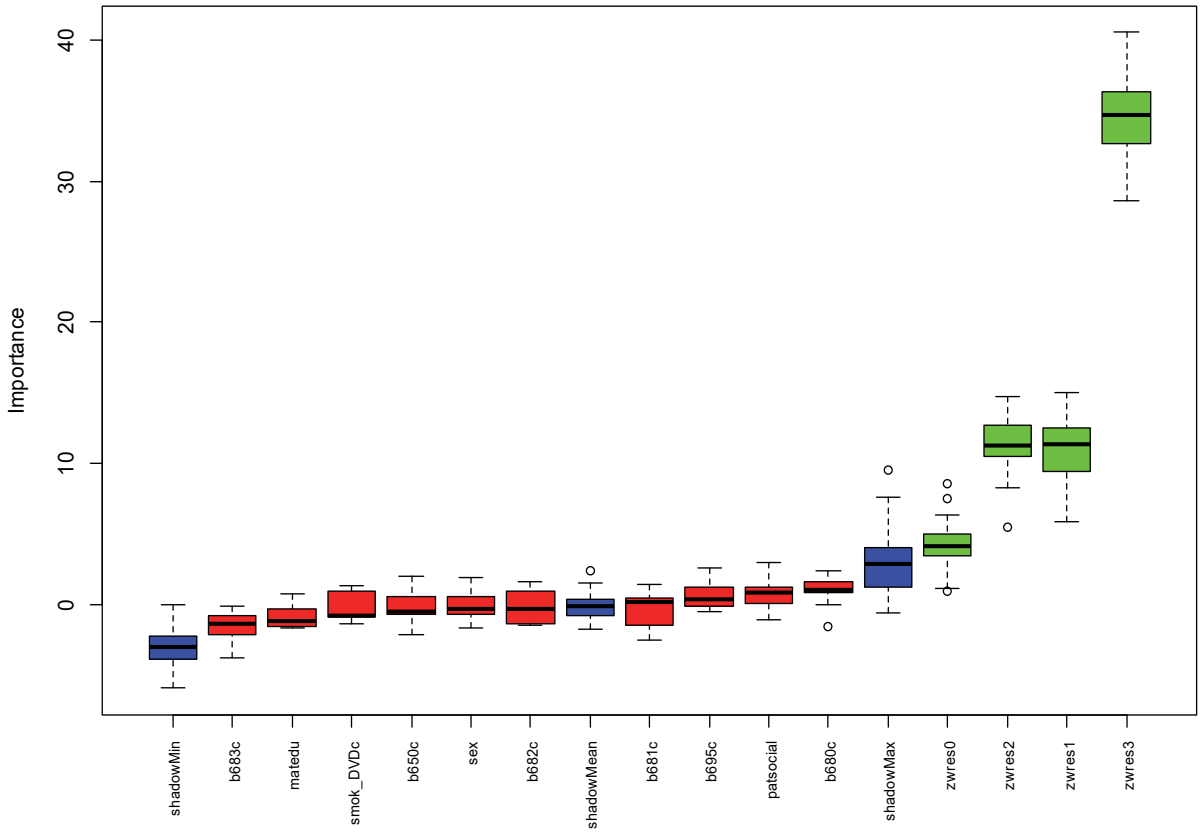


Figure 64: Boruta ranked variables for relevance to waist circumference at age 10. Variables entered: MSP-related variables (7 variables), sex, maternal education, paternal social status and early infant growth. R package: Boruta. (n = 533).

We also wanted to compare model performance using the MSP-related variables versus the MSP-composite derived DNAm components (Table 31). The results between the two are comparable. In addition, no variables were selected except for early growth measures (Table 32 and Figure 64). This was a consistent finding across most outcomes.

Table 31: Comparison of RF metrics between DNAm component versus clinical variables.

	DNAm components				MSP-related variables				MSP composite			
	MSE	MSE.SD	R2	R2 SD	MSE	MSE.SD	R2	R2 SD	MSE	MSE.SD	R2	R2 SD
Age 7	0.75	0.07	0.3	0.07	0.75	0.07	0.31	0.08	0.79	0.07	0.30	0.06
Age 9	0.85	0.07	0.18	0.04	0.85	0.05	0.19	0.08	0.87	0.08	0.19	0.06
Age 10	0.85	0.05	0.17	0.05	0.86	0.05	0.15	0.06	0.87	0.06	0.19	0.06
Age 11	0.86	0.04	0.15	0.05	0.87	0.05	0.14	0.05	0.87	0.04	0.16	0.05

Metrics from RF using outcome: waist circumference at all available data ages. On left, DNAm components. Middle: MSP variable-only. On right: Composite formed from MSP variables and birth weight.. Covariates for both analyses: sex, maternal education, paternal social status.

The sample size for DNAm component analysis is different for each outcome due to missing data. In the case of waist circumference at age 10, DNAm component analysis included 862 subjective while it was only 533 for variable-only analysis. To test whether the differences between DNAm component versus variable-only models were due to the change in sample size, we repeated the above analysis but again employing the built-in impute function present in the randomForestSRC package as in Section 3.4: Clinical relevance of DNAm vulnerability patterns. Results were consistent with and without imputation (data not shown.)

Table 32: Frequency of selection by Boruta as a relevant variable in models of waist circumference at ages 7, 9, 10 and 11.

DNAm component

Variable	Frequency
Comp11	1
Comp14	1
NK	2
zwres0	4
zwres1	4
zwres2	4
zwres3	4

MSP variable-only

Variable	Frequency
zwres0	3
zwres1	4
zwres2	4
zwres3	4

MSP composite

Variable	Frequency
Dim.1	4
Dim.2	2
Dim.3	4
Dim.4	4
Dim.5	3
zwres0	4
zwres1	4
zwres2	4
zwres3	4

We also performed the same analysis using the vulnerability composite. We found that model results were comparable to the other two models. Unlike with MSP variables alone however, the composite dimensions were often selected as relevant features by Boruta (Table 32).

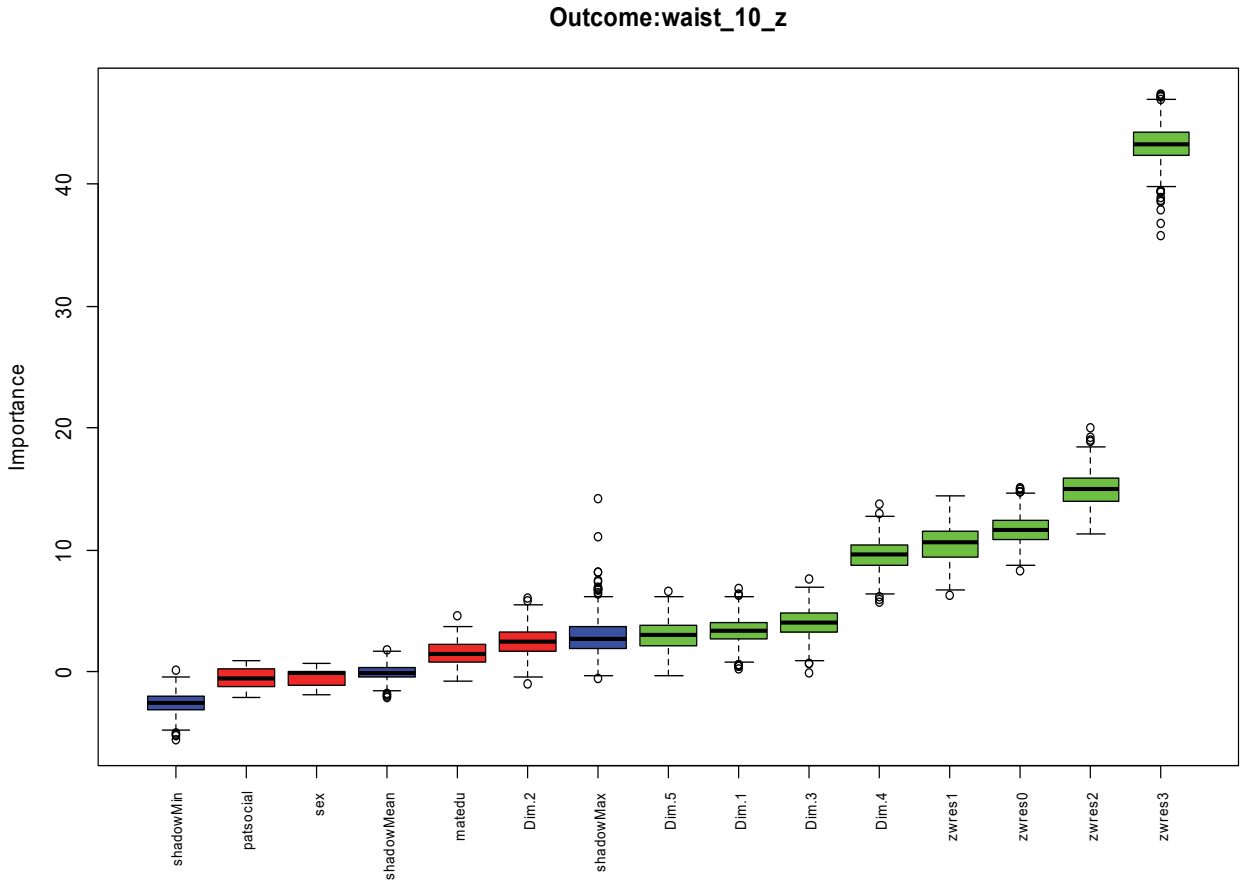


Figure 65: Boruta ranked variables for relevance to waist circumference at age 10. Variables entered: MSP dimensions, sex, maternal education, paternal social status and early infant growth. R package: Boruta. (n = 862).

3.6 DNAm patterns persist into mid and late childhood

Using the DNA patterns discovered in cord blood, we are able to use the PLS weightings as a "template" to extract the same patterns among CpG sites in DNAm data from other sources. In this way, we are able to calculate age-specific component scores for each individual's DNAm data collected from blood in mid and late childhood. In other words, these age-specific scores represent the DNAm patterns at age 7 and 15 years. We repeat the same clinical outcome testing for these later age DNAm patterns.

3.6.1 Late DNAm patterns and covariates

3.6.1.1 *Subject sex*

In Table 19, we showed the relation between sex and DNAm components from cord blood. We perform the same analysis for DNAm components from later ages (Table 33 and Table 34). We can see that Components 10 and 13 are stably related to sex at birth, age 7 and age 15. The number of components related to sex increased by 7 fold between birth to ages 7 and 15. While a number had $p < 0.001$, the beta coefficients were very small compared to Component 10.

Table 33: ANOVA between Age 7 DNAm components and infant sex.

Relation between subject sex and latent variables - Age 7

	Comp1 (1)	Comp2 (2)	Comp3 (3)	Comp4 (4)	Comp5 (5)	Comp6 (6)	Comp7 (7)	Comp8 (8)	Comp9 (9)	Comp10 (10)	Comp11 (11)	Comp12 (12)	Comp13 (13)	Comp14 (14)	Comp15 (15)	Comp16 (16)	Comp17 (17)	Comp18 (18)	Comp19 (19)	Comp20 (20)
sexlabel.xMale	0.1	0.7**	-0.7	1.1***	0.1	-0.3**	0.1	0.9***	-1.5***	29.5***	-1.0***	-0.02	2.1***	-0.7***	-2.7***	0.1	-0.4***	-4.4***	0.2*	0.3***
	(0.1)	(0.3)	(0.7)	(0.1)	(0.1)	(0.1)	(0.4)	(0.1)	(0.2)	(0.5)	(0.1)	(0.1)	(0.1)	(0.2)	(0.3)	(0.1)	(0.1)	(0.7)	(0.1)	(0.1)
Constant	12.5***	-48.8***	-13.7***	-1.3***	-2.1***	-0.1	5.0***	-1.4***	-4.9***	-6.6***	-6.0***	-8.0***	-14.8***	-3.1***	-2.5***	-2.4***	-1.8***	17.3***	-5.2***	1.1***
	(0.1)	(0.2)	(0.5)	(0.1)	(0.1)	(0.1)	(0.3)	(0.1)	(0.1)	(0.4)	(0.1)	(0.1)	(0.1)	(0.1)	(0.2)	(0.1)	(0.1)	(0.5)	(0.1)	(0.05)
Observations	973	973	973	973	973	973	973	973	973	973	973	973	973	973	973	973	973	973	973	973
R^2	0.002	0.01	0.001	0.2	0.002	0.01	0.000	0.1	0.1	0.8	0.1	0.000	0.2	0.02	0.1	0.002	0.02	0.04	0.004	0.02
Adjusted R^2	0.001	0.01	0.000	0.2	0.001	0.01	-0.001	0.1	0.1	0.8	0.1	-0.001	0.2	0.02	0.1	0.001	0.02	0.03	0.003	0.02
Residual Std. Error	1.6	3.9	10.4	1.2	1.2	1.4	6.2	1.1	2.5	7.8	1.5	1.2	1.9	2.5	4.3	1.2	1.3	11.3	1.4	1.1
F Statistic	1.8	7.3**	1.1	236.4***	2.4	8.8**	0.03	168.1***	90.0***	3,465.5***	117.9***	0.1	283.3***	22.1***	94.5***	1.5	19.3***	36.1***	3.9*	21.8***

Note: *p<0.05; **p<0.01; ***p<0.001

Table 34: ANOVA between Age 15 DNAm components and infant sex.

Relation between subject sex and latent variables -Age 15

	Comp1 (1)	Comp2 (2)	Comp3 (3)	Comp4 (4)	Comp5 (5)	Comp6 (6)	Comp7 (7)	Comp8 (8)	Comp9 (9)	Comp10 (10)	Comp11 (11)	Comp12 (12)	Comp13 (13)	Comp14 (14)	Comp15 (15)	Comp16 (16)	Comp17 (17)	Comp18 (18)	Comp19 (19)	Comp20 (20)
sexlabel.xMale	0.2	0.1	-2.4***	0.9***	-0.03	-0.2	0.5	0.9***	-1.3***	24.2***	-1.1***	0.004	2.0***	-0.9***	-2.1***	0.1	-0.4***	-3.6***	0.3*	0.3***
	(0.2)	(0.5)	(0.7)	(0.1)	(0.1)	(0.1)	(0.4)	(0.1)	(0.2)	(0.6)	(0.1)	(0.1)	(0.1)	(0.2)	(0.3)	(0.1)	(0.1)	(0.8)	(0.1)	(0.1)
Constant	-1.2***	0.5	1.8***	0.3***	0.4***	0.8***	8.3***	1.0***	-5.6***	-15.8***	-0.9***	-3.8***	-7.6***	6.8***	-0.5**	1.1***	-2.5***	31.1***	1.2***	2.7***
	(0.1)	(0.3)	(0.5)	(0.1)	(0.1)	(0.1)	(0.3)	(0.1)	(0.1)	(0.4)	(0.1)	(0.1)	(0.1)	(0.1)	(0.2)	(0.1)	(0.1)	(0.6)	(0.1)	(0.1)
Observations	974	974	974	974	974	974	974	974	974	974	974	974	974	974	974	974	974	974	974	974
R^2	0.001	0.000	0.01	0.1	0.000	0.004	0.001	0.1	0.04	0.6	0.1	0.000	0.2	0.02	0.1	0.001	0.02	0.02	0.005	0.01
Adjusted R^2	-0.000	-0.001	0.01	0.1	-0.001	0.003	0.000	0.1	0.04	0.6	0.1	-0.001	0.2	0.02	0.1	0.000	0.02	0.02	0.004	0.01
Residual Std. Error	3.0	7.0	10.7	1.4	1.2	1.3	7.0	1.2	3.2	8.9	1.8	1.4	2.3	3.2	4.0	1.3	1.5	12.3	1.8	1.4
F Statistic	0.7	0.1	12.5***	112.7***	0.1	3.5	1.4	139.9***	40.2***	1,799.4***	91.8***	0.002	191.2***	21.4***	68.0***	1.5	15.8***	21.0***	4.7*	14.5***

Note: *p<0.05; **p<0.01; ***p<0.001

3.6.1.2 *Social factors*

Like in cord blood, we used ANOVA to observe any relation between DNAm components at later ages and social factors. None were found (data not shown.)

3.6.1.3 *Batch effects*

While only weak batch effects were seen in cord blood after batch correction, such effects seemed far more prominent in blood at Age 7 and 15 as estimated through independent component analysis. As a reminder, Age 7 and 15 DNAm data was processed with the same normFact function (Renard & Absil, 2017) used for cord data to attenuate BCD plate batch effect specifically for those samples.

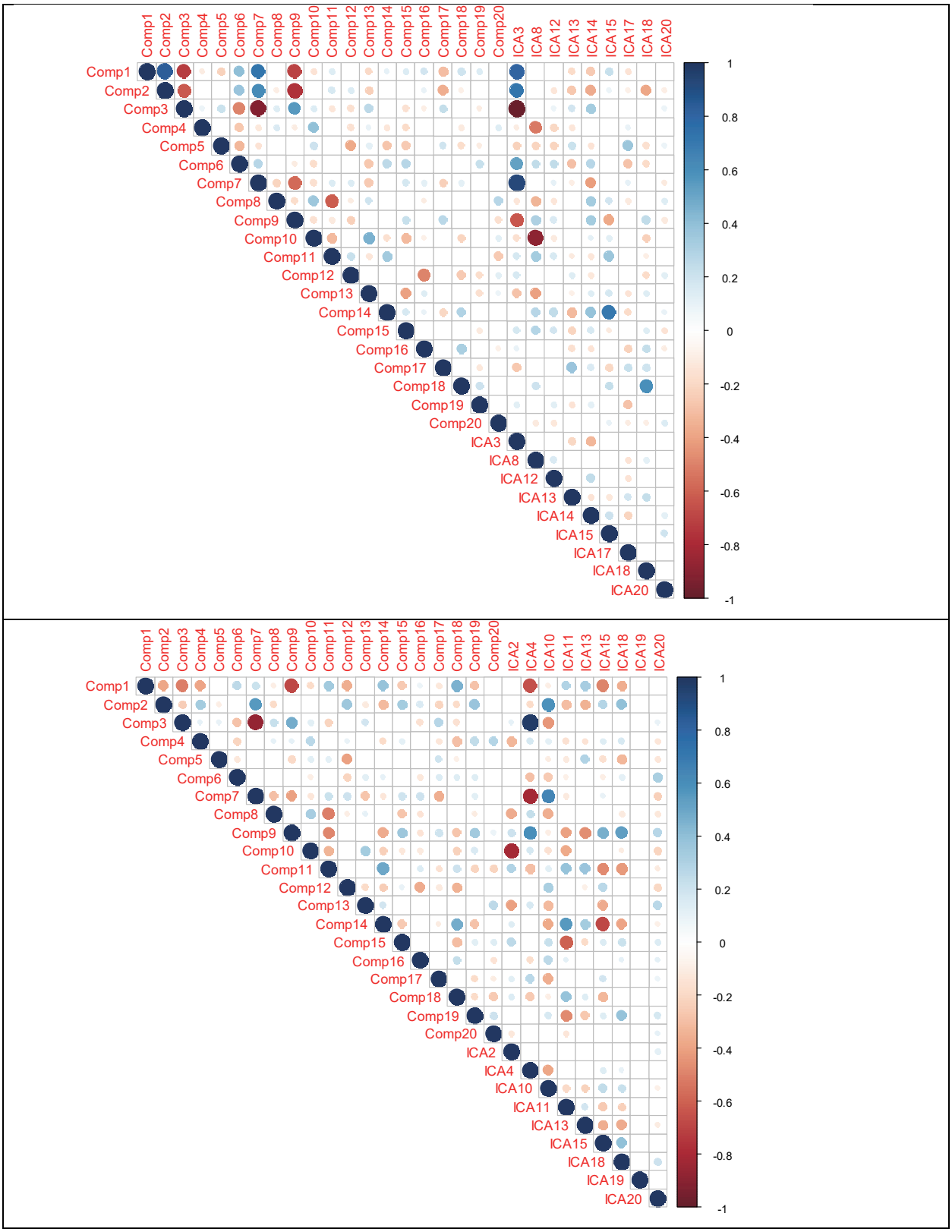


Figure 66: Correlation matrix between DNAm components and sample batch. Colour indicates Pearson correlation and circle size indicates *p*-value. DNAm data from age 7 (Top) and age 15 (Bottom.). R script normFact (Renard & Absil, 2017).

Components 1, 2, 3, 7, 10 and 14 were possibly related to batch effects in DNAm data from Age 7 and 15. Of note, Components 3, 10 and 14 are also related to sex at Age 7 and 15 (Table 33 and Table 34).

3.6.1.4 *Late DNAm patterns and cell count*

Figure 67 shows the correlation matrix between DNAm components and cell count, this time using a reference-free estimation method. Again, this was because there are no cell composition references available for age 7 and 15. We can see that Component 1, 2, 3 and 7 are strongly correlated with cell count using the reFACTor principal components.

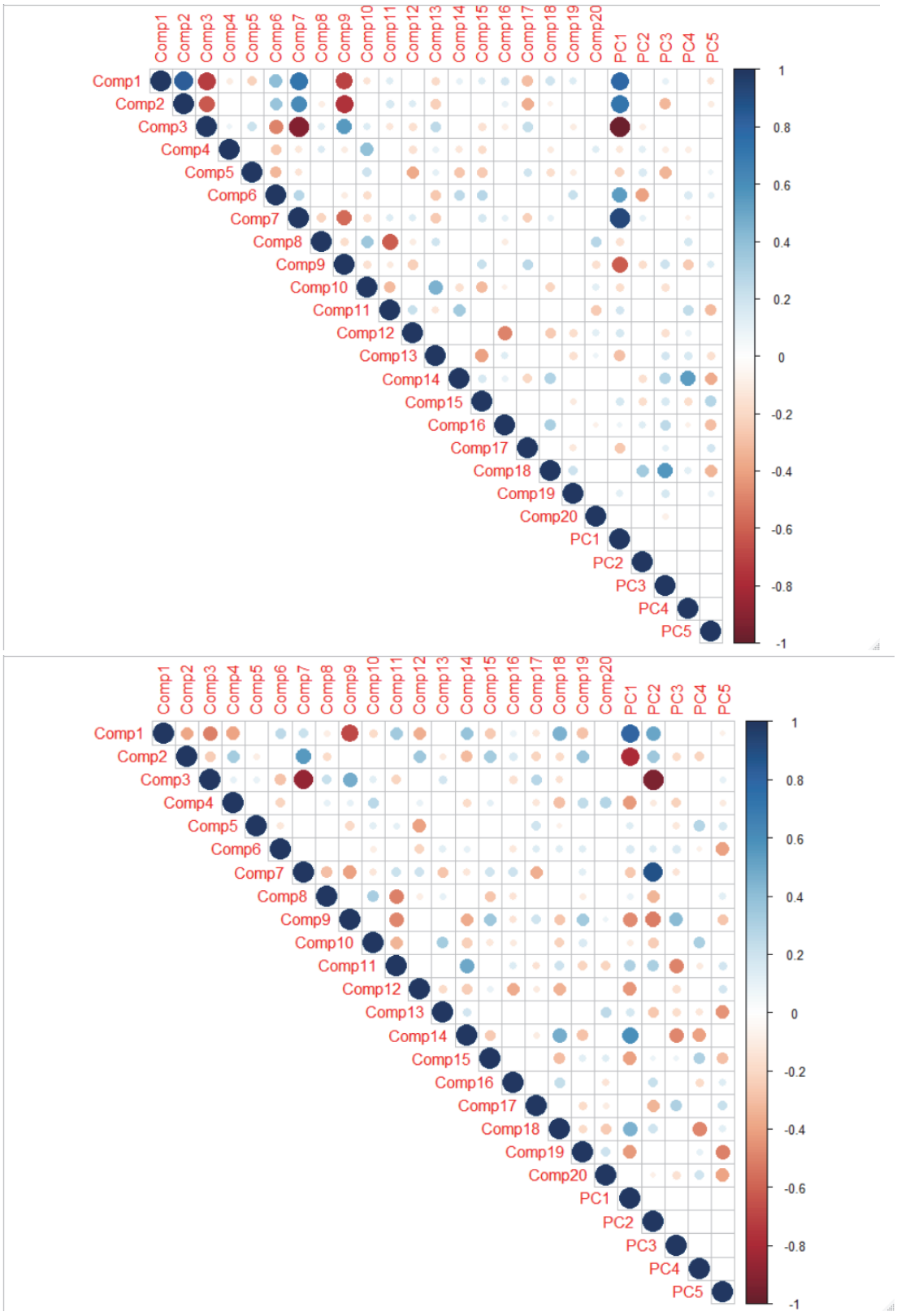


Figure 67: Correlation matrix between DNAm components and estimated cell count. Colour indicates Pearson correlation and circle size indicates *p*-value (R package reFACTor.) Top: Age 7 blood. Bottom: Age 15 blood.

3.6.2 Clinical relevance of mid and late childhood DNAm patterns

We performed the same two-step process as described for cord DNAm component analysis to model the relation to outcomes. Overall, we discovered a subset of components that persisted in their relation to concurrent or future outcomes at later ages. As well, the relation between outcomes in infancy and certain components at Age 7 and 15 persisted.

For example, DNAm Component 11 maintained its relation to anthropometric, body composition, cognitive outcomes and academic performance between cord, Age 7 and Age 15 DNAm samples, (Table 28 and Table 29.) As well, it does not seem to strongly relate to covariates.

While Component 11 was relatively steady, Component 5 almost seemed to "change track" with age. Component 5 in cord blood was related to a number of academic measures but no other outcomes. This component at Age 7 relates to age 8 Verbal skills. However, at Age 15, this component is unrelated to any neurocognitive outcomes. Instead, Component 5 at Age 15 is relevant to various anthropometric and body composition measures - at not just one but all 6 time points. This relation across physical outcomes over time was surprising as it did not exist in cord or Age 7 blood.

On the other hand, some later childhood outcomes were only related to Age 7 and 15 components and not cord components. This was true for DNAm components 7 and 9 which at Age 7 and 15 were related to cognitive scores at age 8, multiple anthropometric outcomes at various ages and academic performance from start of school to mid-childhood (multiple subjects at key stages 1, 2 and 3.)

Interestingly, some components would "skip" an age in terms of its relation to outcome. We take neurodevelopment as an example. Component 18 in cord is related to developmental scores at age 30 months. In Age 15 blood, component 18 was related to cognitive scores at age 8 years (Performance and Total scores). Component 14 displayed a similar pattern for anthropometric measures. In cord blood, it is related at lean mass at age 11. In Age 15 blood, it relates to weight and/or waist measures at ages 7, 8, 11 and 14. However, this relation should be viewed with caution as Component 14 was weakly associated with sex in DNAm data from samples at Age 7 and 15 (Table 33 and Table 34.) That said, there was no association between sex and Component 14 in cord blood. In addition, the association to sex appears

similar at Age 7 and 15 (for both, adjusted $R^2 = 0.02$, F statistic between 21-22, $p < 0.001$) yet at Age 7 there is no relation with Component 14 and any anthropometric measures. If sex was an important covariate at both Age 7 and 15, then we would expect the "false" relation with Component 14 with outcome to also appear in Age 7 blood.

3.6.3 DNAm patterns interact with child features

As seen in Results section 3.4.1: *Anthropometric outcomes*, we found that the DNAm components had differential relation to outcome depending on covariates. To further explore this phenomenon, we wanted to investigate whether this feature persisted in DNAm components extracted at later ages. To illustrate this, we employ dependence plot of the same outcomes as in Results section 3.4.1, (i.e. lean mass and waist circumference,) but this time using DNAm from blood at age 15. Figure 68 shows this data both for DNAm data from age 15 as well as from cord blood as shown in for better comparison. While different DNAm components relate to these outcomes at different ages, we can see that there exist differential relations between the component and outcome depending on the fetal growth (birth weight) at both DNAm ages.

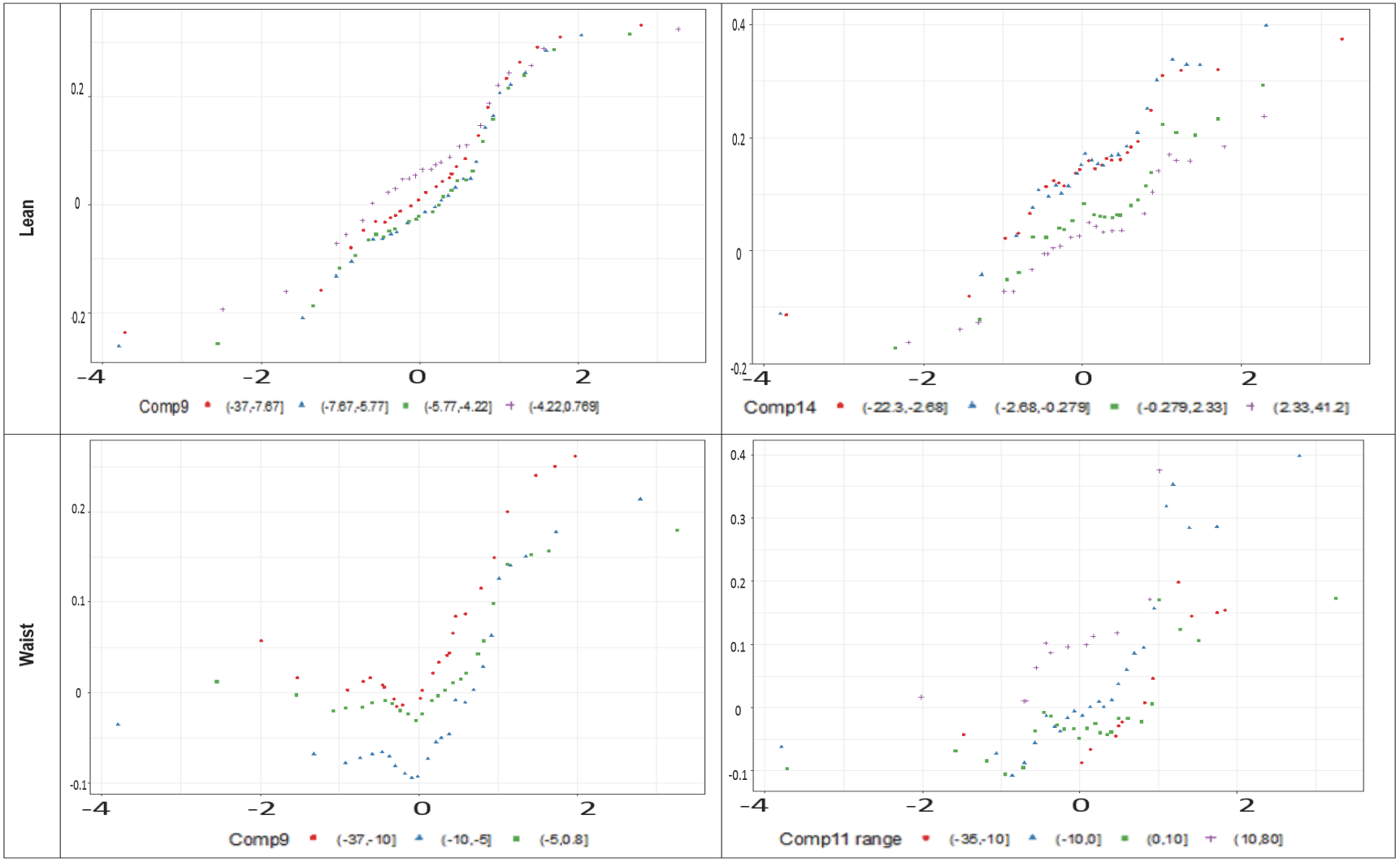


Figure 68: Partial dependence plot of outcome (y-axis) as a function of birth weight (z-score adjusted for sex and gestational age), conditional on ranges of component scores from blood at Age 15 (left column) or Cord (right column.) Z-axis: Outcome as z-scores for lean mass (top) and waist circumference (bottom.)

3.7 Molecular relevance

One of the biggest challenges in the study of complex diseases and epigenetics is how to link clinical and molecular relevance. Depending on the data set size and analytic method employed, dozens to hundreds of "significant" DNAm sites can be found related to infant exposure to MSP. This has been repeatedly shown in numerous cohorts around the world from various ethnic groups, (for reviews, see (Joubert *et al.*, 2016; Knopik *et al.*, 2019; Taal *et al.*, 2013).) What is the relation, if any, between the significant CpG "hits" and how do these interact with molecular mechanisms that can plausibly lead to altered phenotypes?

In order to better understand the possible molecular implications of the DNAm patterns suggested by the components, we mapped each components' representative CpG sites to their chromosomal location and its predicted chromatin states. Histone modifications and variants and regions of open chromatin are examples of epigenetic marks that help localize putative regulatory elements. When multiple marks are aggregated and placed in context with information such as TSSs or exon/intron boundaries, this generates a genome-wide map of what are called 'chromatin states' (Table 35). These states have been used to estimate chromatin activity and are known to coincide with critical genomic elements, such as promoters, enhancers, and transcribed, repressed, and repetitive regions (Ernst & Kellis, 2010). We used the Core 15 state model derived from five chromatin marks (H3K4me3, H3K4me1, H3K36me3, H3K27me3, H3K9me3) assayed in 127 epigenomes (Roadmap *et al.*, 2015).

STATE NO.	MNEMONIC	DESCRIPTION
1	TssA	Active TSS
2	TssAFlnk	Flanking Active TSS
3	TxFlnk	Transcr. at gene 5' and 3'
4	Tx	Strong transcription
5	TxWk	Weak transcription
6	EnhG	Genic enhancers
7	Enh	Enhancers
8	ZNF/Rpts	ZNF genes & repeats
9	Het	Heterochromatin
10	TssBiv	Bivalent/Poised TSS
11	BivFlnk	Flanking Bivalent TSS/Enh
12	EnhBiv	Bivalent Enhancer
13	ReprPC	Repressed PolyComb
14	ReprPCWk	Weak Repressed PolyComb
15	Quies	Quiescent/Low

Table 35: Chromatin state definitions and abbreviations.The active states (associated with expressed genes) consist of active TSS-proximal promoter states (TssA, TssAFlnk), a transcribed state at the 5' and 3' end of genes showing both promoter and enhancer signatures (TxFlnk), actively-transcribed states (Tx, TxWk), enhancer states (Enh, EnhG), and a state associated with zinc finger protein genes (ZNF/Rpts). The inactive states consist of constitutive heterochromatin (Het), bivalent regulatory states (TssBiv, BivFlnk, EnhBiv), repressed Polycomb states (ReprPC, ReprPCWk), and quiescent state (Quies). Image and description from (Roadmap *et al.*, 2015).

3.7.1 Genic and chromatin based context

For the sake of space, we will proceed with showing results for only a few illustrative components to demonstrate the general trend for most components. We take Components 7, 9, 11 and 14 as examples (Figure 69). We see that in terms of absolute counts, quiescent marks are the most common mark in patterns across chromosomal locations, followed by transcription marks (both weak and strong) and repressed Polycomb states. The next spike in frequency is in enhancer marks. This is the general trend seen across DNAm components.

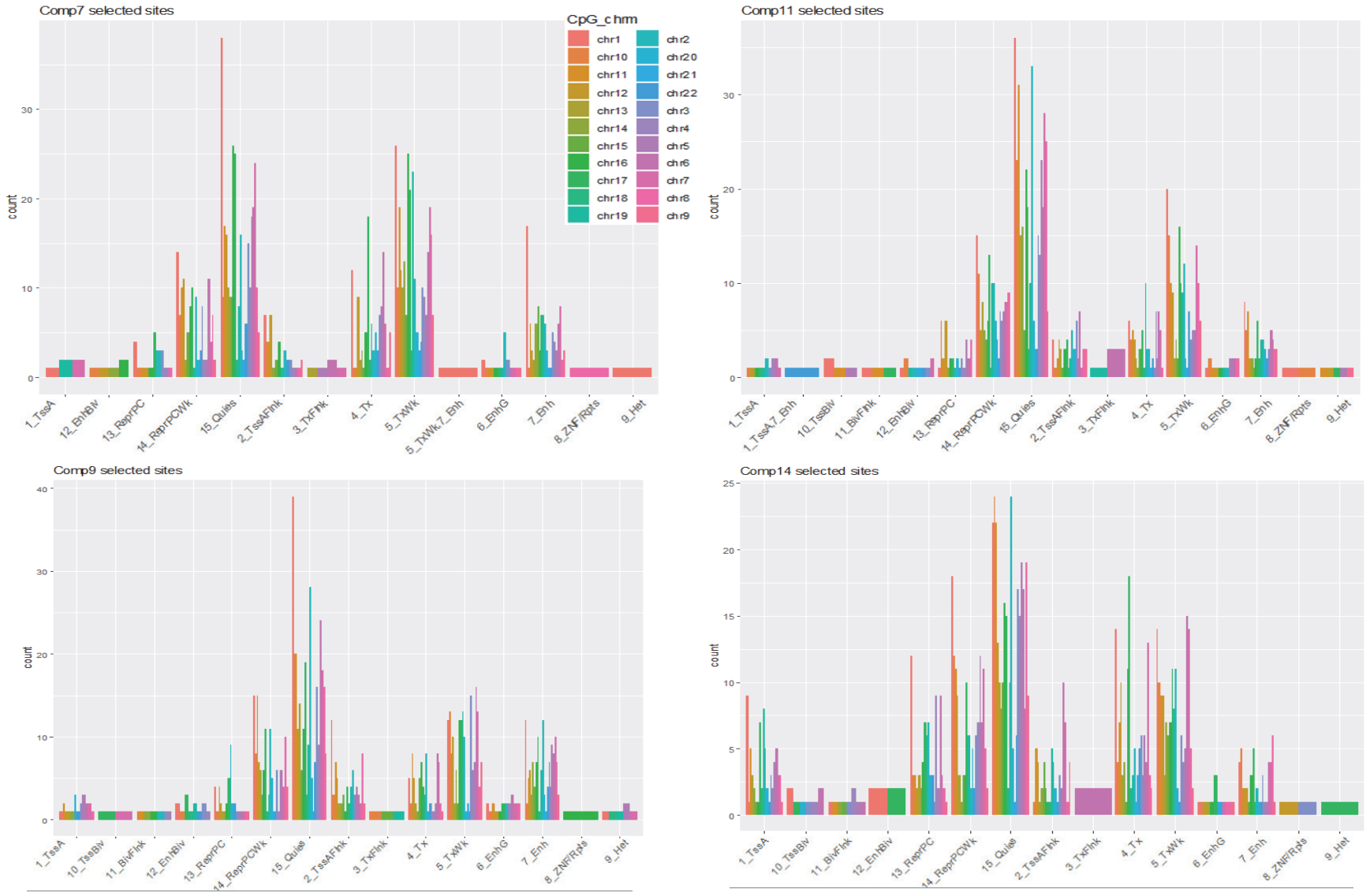


Figure 69: Histogram of cord DNAm components 7, 9, 11 and 14 categorized by chromatin state (using T cell reference genome E031 as example.)

The design of the 450K chip has a predilection for genic sites. Thus, it is important to frame the frequency of the chromatin marks of the components in context of that of the 450K chip itself. Figure 70 illustrates this comparison. It is clear the quiescent sites are by far the most frequent mark on the 450K chip, followed distantly by weak repressed polyComb marks. In contrast, quiescent and transcription marks are much closer in frequency in DNAm components. Enhancer and TxFlnk marks (the latter that demonstrates both promoter and enhancer signatures, see Table 35) are particularly notable for their appearance in the components as they are among the least frequent marks on the 450K chip.

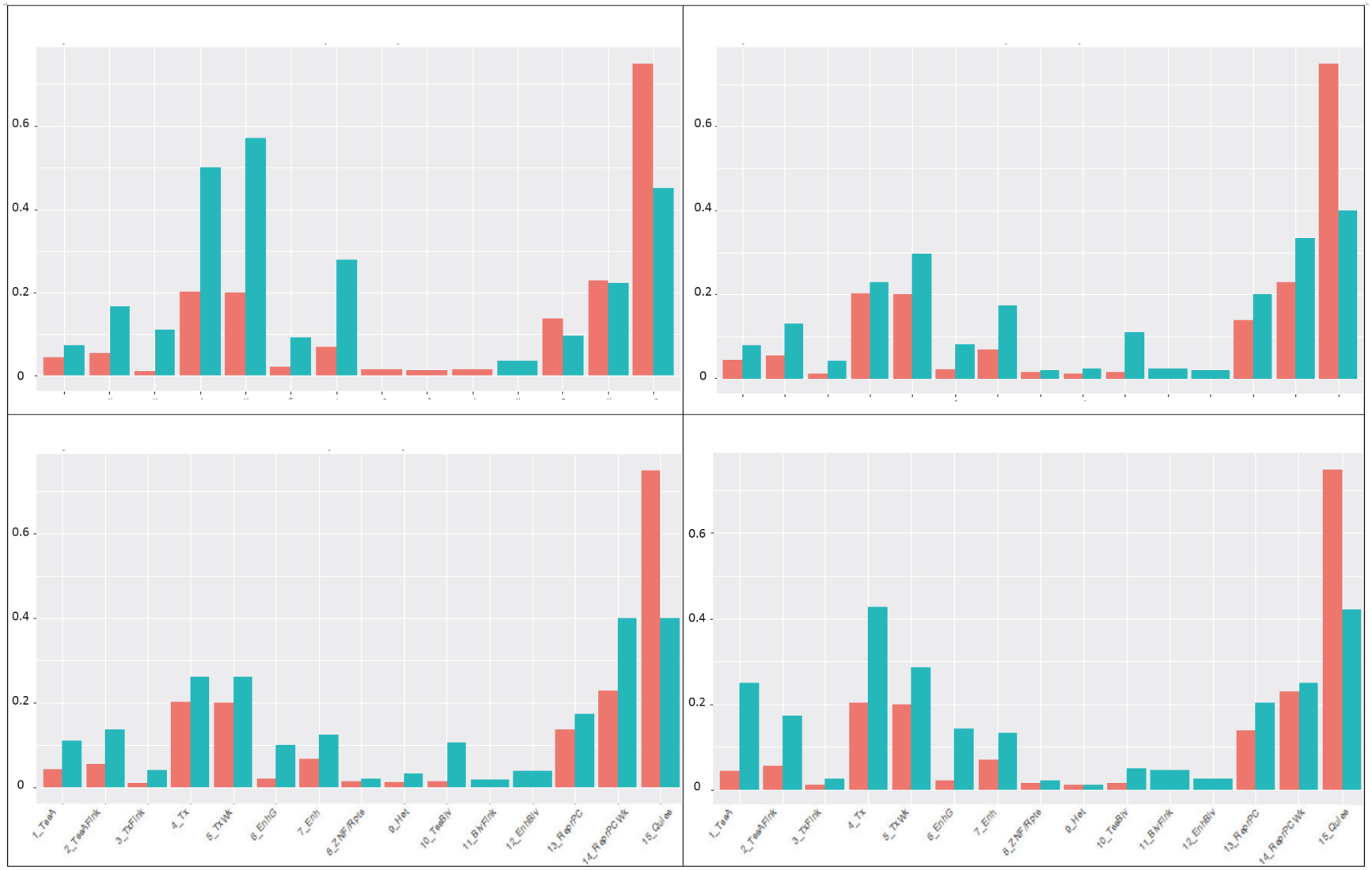
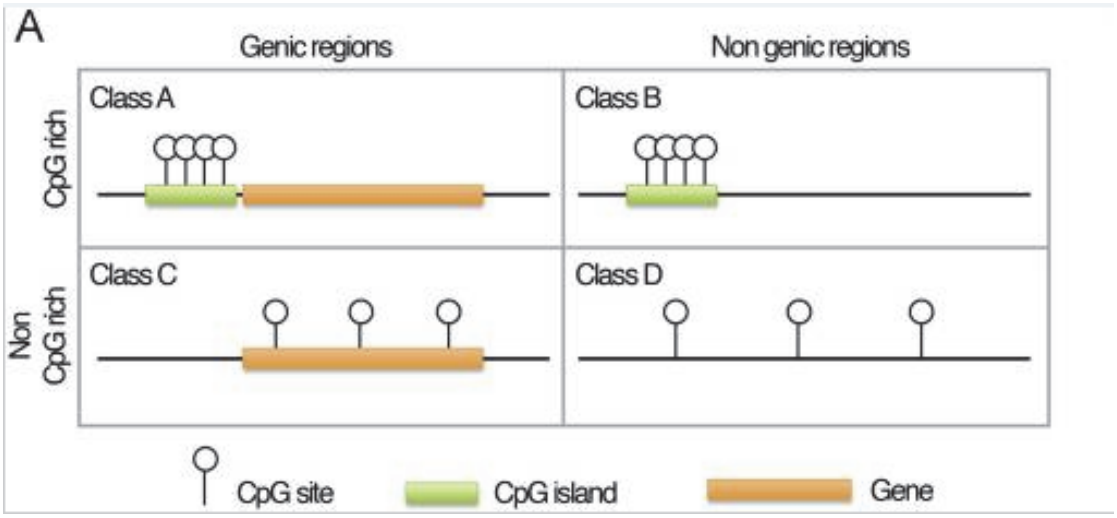


Figure 70: Comparison of chromatin mark frequency of Illumina 450K beadchip (post all probe filtering) and cord DNAm Component 7, 9, 14 and 11 (clockwise from top left.) X- and y-axis the same on all graphs. Legend: Orange – 450K, Green – DNAm component. Graph only shows marks with >1% frequency per chromosome for space considerations. Cell reference: cord T cells epigenome (E033).

To further investigate biological relevance in terms of epigenetic machinery, we were also curious about the CpG contexts of these components. We used the BOP categorization proposed by Bacalini *et al.* based on its location relative to other CpGs and genic areas (Bacalini *et al.*, 2015). This group found differential BOP analysis revealed more age-related sites in a meta-analysis of 3 studies using multivariate ANOVA compared to univariate analysis. Figure 71 shows schematically the definition of these BOPs as well as their distribution across the 450K chip. Note that the two classes in genic regions, Class A and C, are the most common class while the non-genic classes, Class B and D, are rarer.



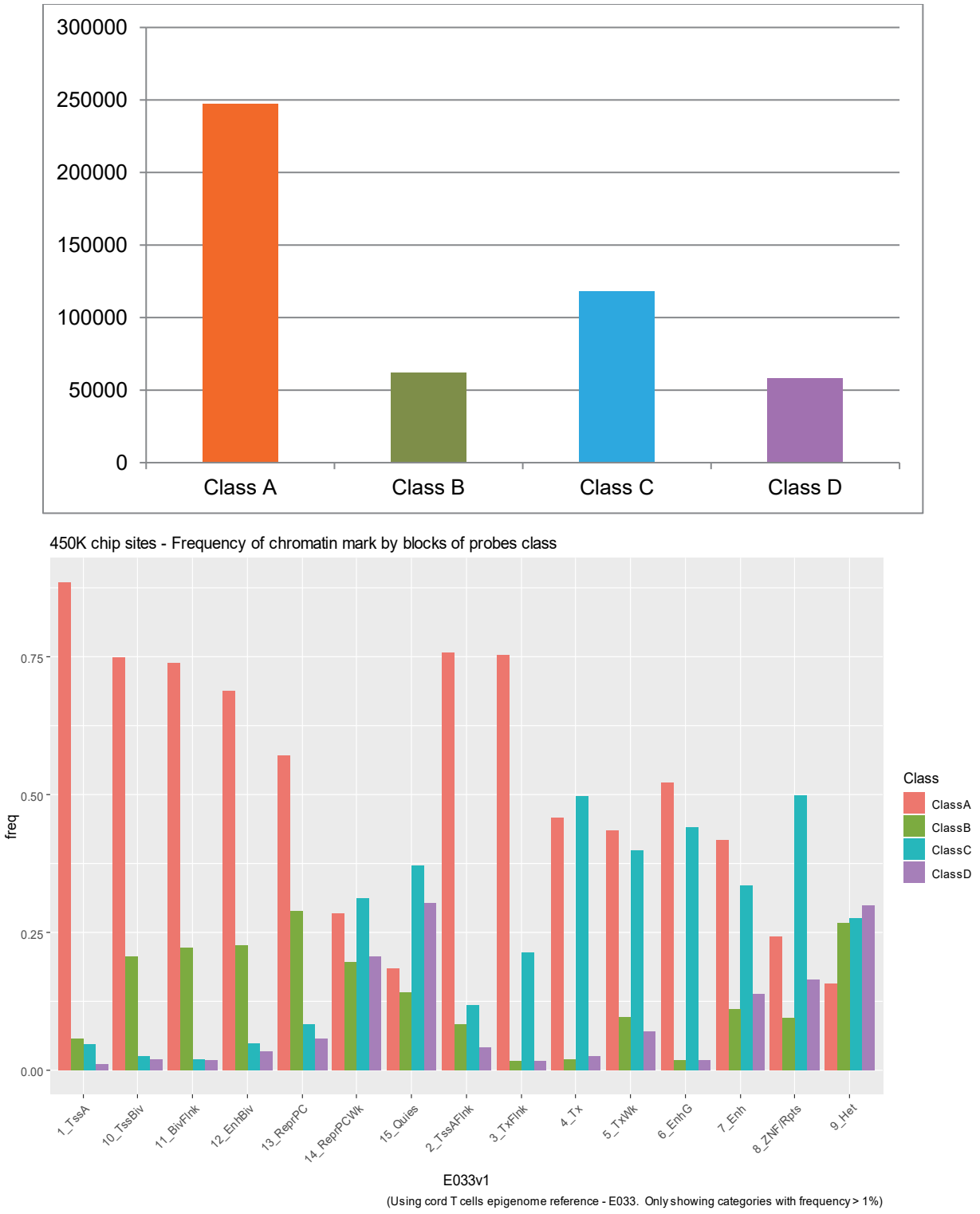
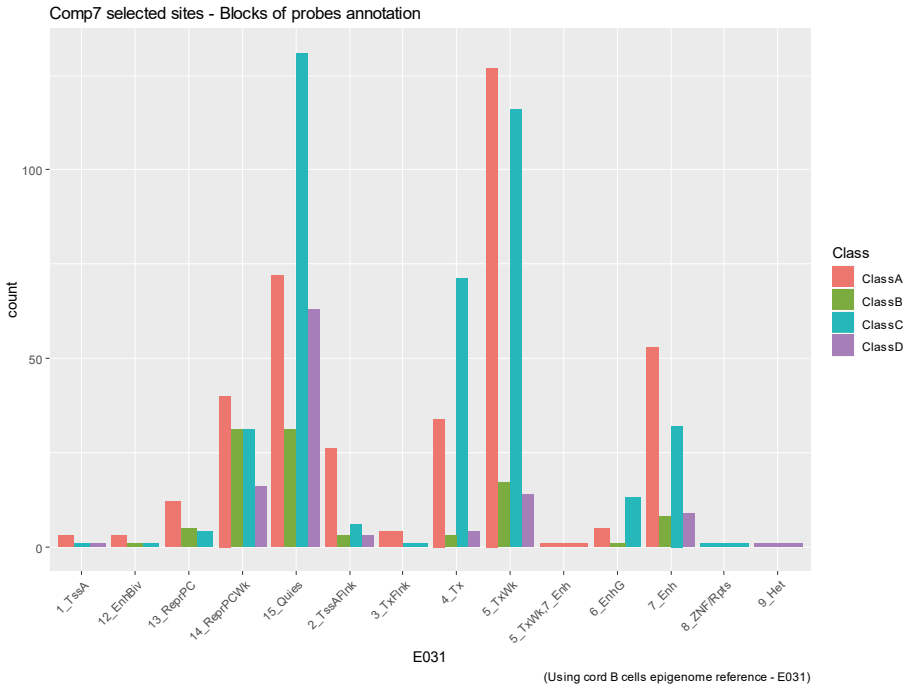
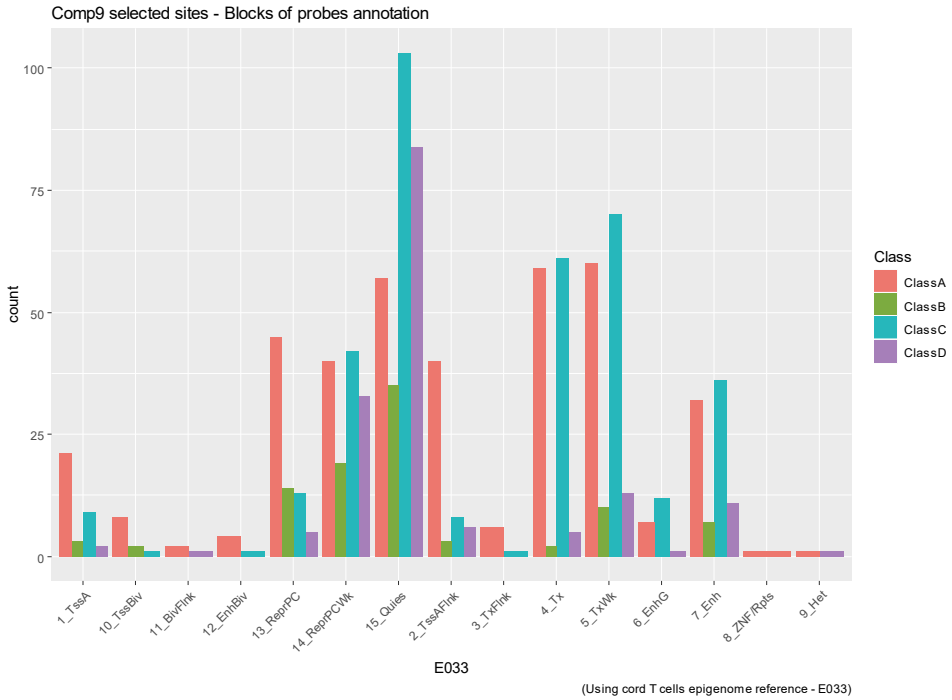
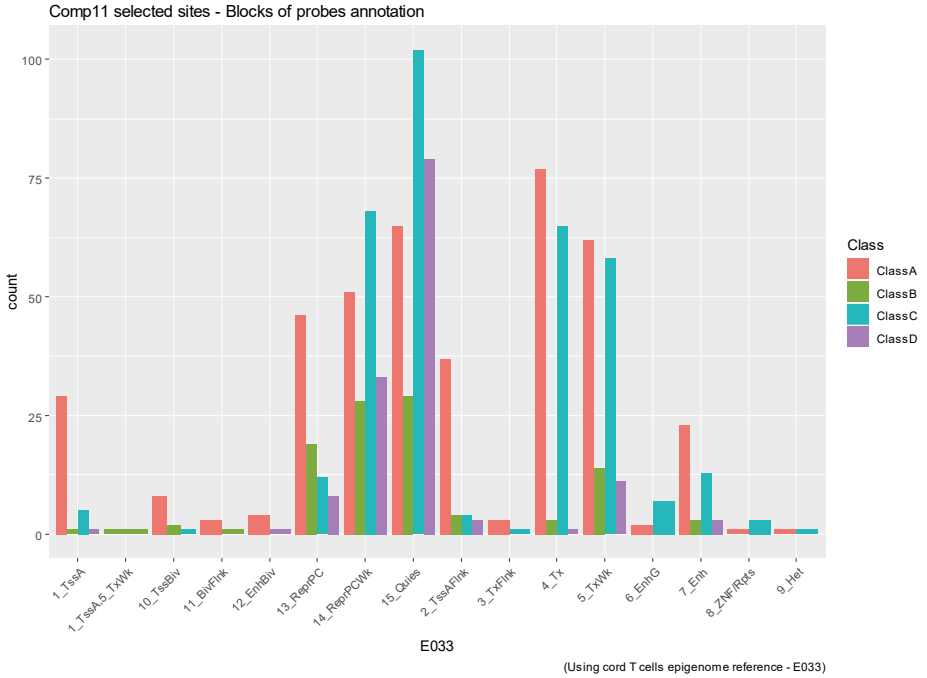


Figure 71: Top - Schematic of BOP classes based on genomic location. Figure from (Bacalini *et al.*, 2015). Middle – Histogram of all CpG probes on 450K bead chip in each BOP category. Bottom: Relative frequency of chromatin mark by BOP category (all CpG probes on 450K chip.) Relative frequency instead of count scale reasons. There were far more probes in Class A with the TssA mark (87635 probes) that made it difficult to see the other bars which were mostly in the hundreds to few thousands.

We attempted the analytic pipeline provided by (Bacalini *et al.*, 2015). No results passed the MANOVA or ANOVA statistical thresholds for relevance, (data not shown.) Taking a look at the distribution of the CpG sites representative of the DNAm patterns, however, we observed two unexpected findings (Figure 72). First, we found a notable shift in distribution of BOP categories comparing the 450K chip to the DNAm patterns. Most dramatically, we observe a move away from Class A sites, which are the CpG dense genic sites. We take special note that Class A is more than double the frequency on the 450K chip than any other class (Figure 71). In fact, Class A probes marked as active TSS account for about 18% of all sites alone. Second, there is a larger portion of CpG sites within the DNAm patterns in CpG poor regions, (Class C and D,) compared to the 450K chip. As well, despite Class A probes being the most common class in most chromatin mark categories, Class B and D still retained representation in the quiescent and repressed weak Polycomb categories. These non-genic classes are the least represented on the 450K chip (less than 15% of chip sites each.)




Comp9 selected sites - Blocks of probes annotation
count
100
75
50
25
0
1_TssA
10_TssBiv
11_BivFlnk
12_EnhBiv
13_ReprPC
14_ReprPCWk
15_Quies
2_TssAFlnk
3_TxFlnk
4_Tx
5_TxWk
6_EnhG
7_Enh
8_ZNF/Rpts
9_Het
E033
Class
ClassA
ClassB
ClassC
ClassD
(Using cord T cells epigenome reference - E033)


Comp11 selected sites - Blocks of probes annotation
count
100
75
50
25
0
1_TssA
1_TssA,5_TxWk
10_TssBiv
11_BivFlnk
12_EnhBiv
13_ReprPC
14_ReprPCWk
15_Quies
2_TssAFlnk
3_TxFlnk
4_Tx
5_TxWk
6_EnhG
7_Enh
8_ZNF/Rpts
9_Het
E033
Class
ClassA
ClassB
ClassC
ClassD
(Using cord T cells epigenome reference - E033)

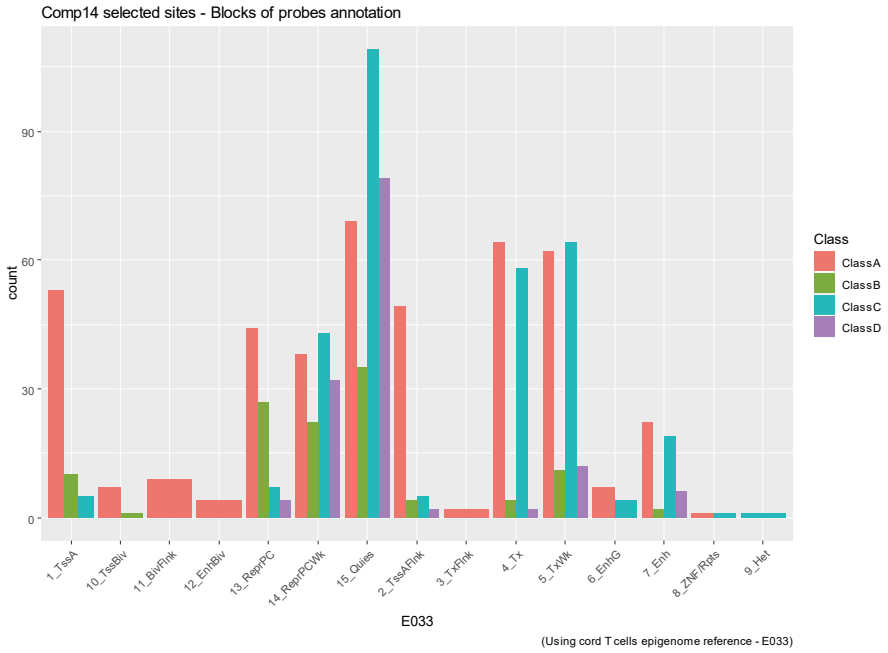
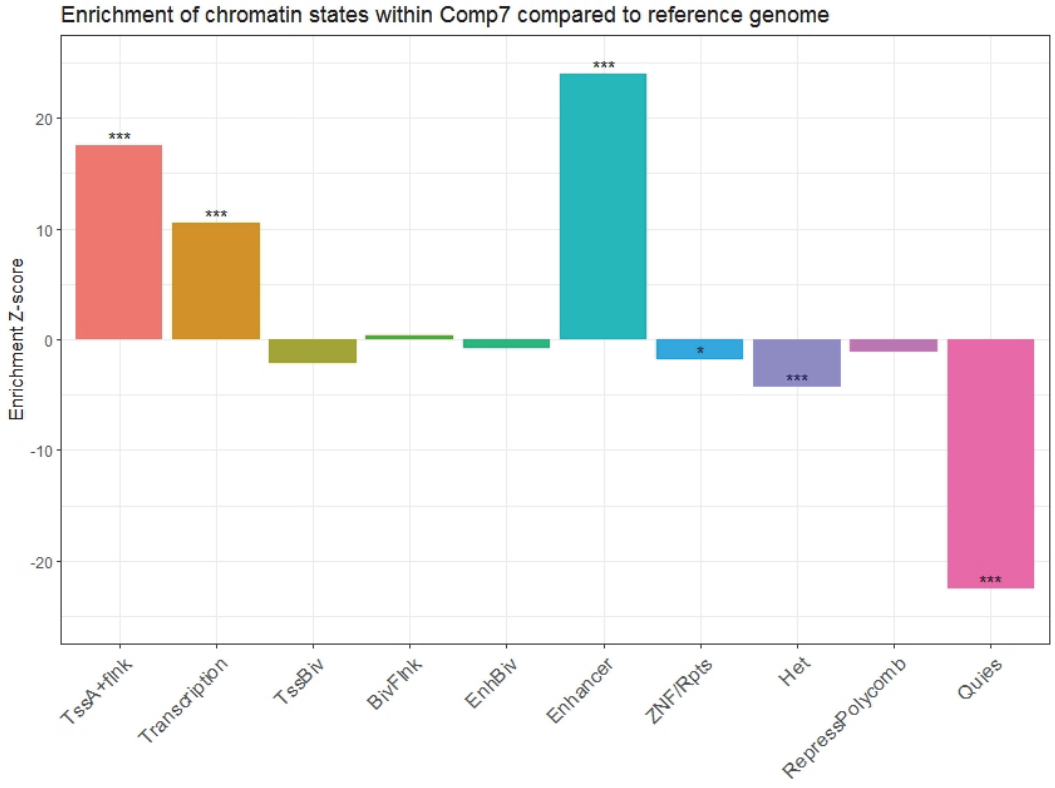
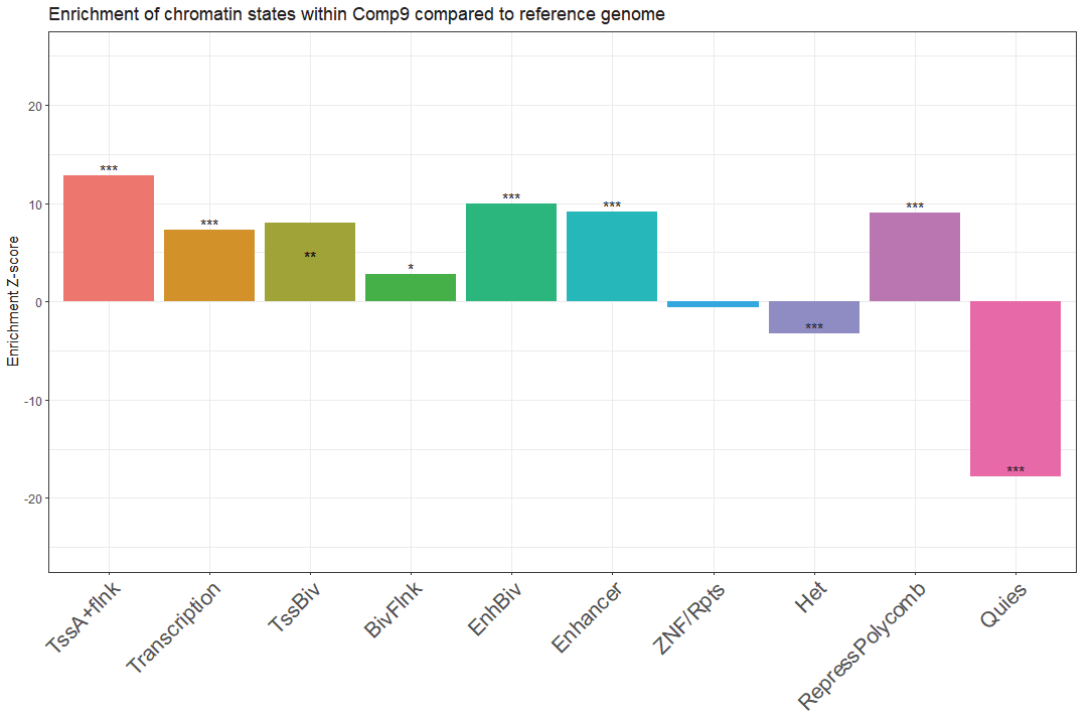
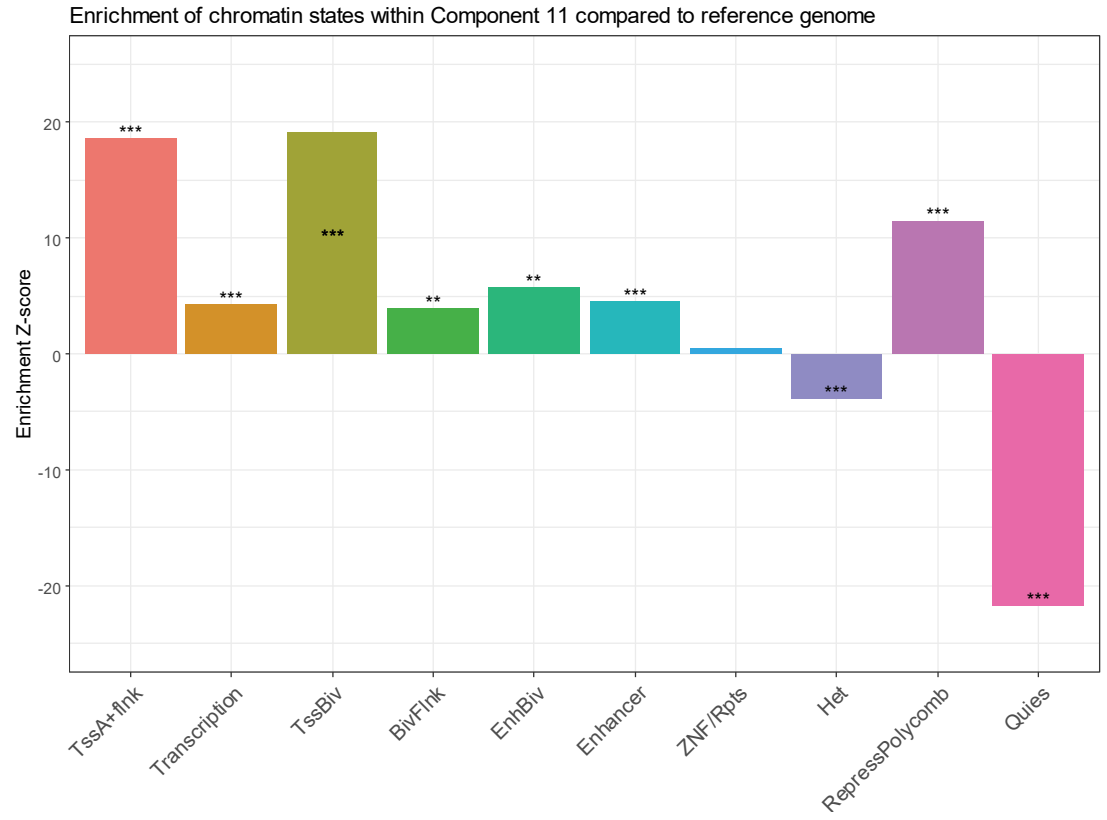


Figure 72: Histogram of representative CpG sites in cord DNAm components 7, 9, 11, and 14 categorized by BOP classes (using B cell reference genome E033 as example.)

We used permutation testing to test whether the DNAm patterns were enriched in particular chromatin states compared to the reference genomes (Figure 70). Results were similar regardless of reference used. We can see that patterns are enriched in both active marks (e.g. enhancer and active transcription features,) as well as inactive marks (e.g. bivalent and repressed Polycomb states,) compared to NIH Roadmap cord blood references. On the other hand, it had less than expected heterochromatin and quiescent state marks. To confirm whether these findings were specific to a certain cell type reference, we conducted this analysis using cord T or B cell as well as blood mononuclear cell reference genomes. Except for a few categories, most results were consistent between all references.


Enrichment of chromatin states within Comp7 compared to reference genome
Enrichment Z-score
20
10
0
-10
-20
***
***
***
*
***
***
TssA+flnk
Transcription
TssBiv
BivFlnk
EnhBiv
Enhancer
ZNF/Rpts
Het
RepressPolycomb
Quies


Enrichment of chromatin states within Comp9 compared to reference genome
Enrichment Z-score
20
10
0
-10
-20
***
***
**
*
***
***
***
***
***
TssA+flnk
Transcription
TssBiv
BivFlnk
EnhBiv
Enhancer
ZNF/Rpts
Het
RepressPolycomb
Quies


Enrichment of chromatin states within Component 11 compared to reference genome
Enrichment Z-score
20
10
0
-10
-20
***
***
***
**
**
***
***
***
***
TssA+flnk
Transcription
TssBiv
BivFlnk
EnhBiv
Enhancer
ZNF/Rpts
Het
RepressPolycomb
Quies

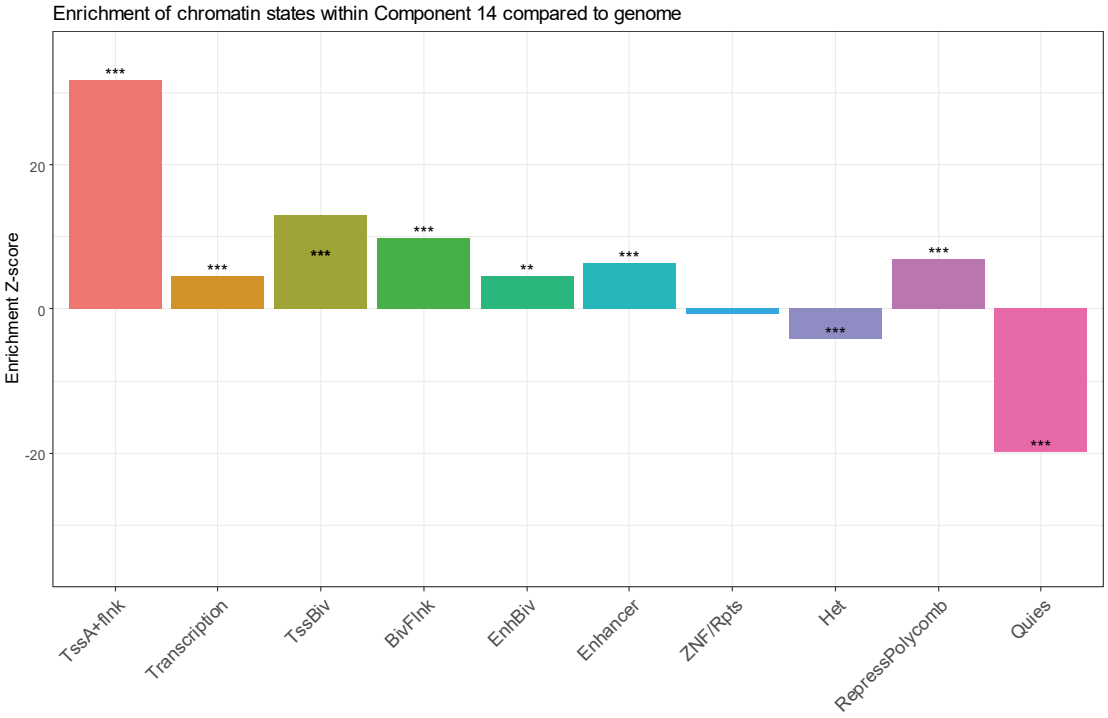

Enrichment of chromatin states within Component 14 compared to genome
Enrichment Z-score
20
0
-20
***
***
***
***
**
***
***
***
***
TssA+flnk
Transcription
TssBiv
BivFlnk
EnhBiv
Enhancer
ZNF/Rpts
Het
RepressPolycomb
Quies

Figure 73: Enrichment testing of component 7, 9, 11 and 14 loci as selected using sparse PLS using 1000 permutations. Asterisks indicate *p*-values: *(<.05), **(<.01),*(<.001). Results shown are for cord T cell reference (E033). Results using other references available upon request.**

3.7.2 Topology based context

We further explored the overlap of the component sites with functional chromatin structures. Recently, Yang *et al.* generated a genome-wide map of promoter-anchored chromatin interactions (PAIs) and among those, chromatin loops that overlap with 450K chip. Using the same permutation method, Figure 74 shows that there was indeed enrichment in some components for PAI/loop marks for some components. However, unlike for chromatin marks in the above analysis, enrichment for PAI using this mapping reference is not as consistent a feature among all components.

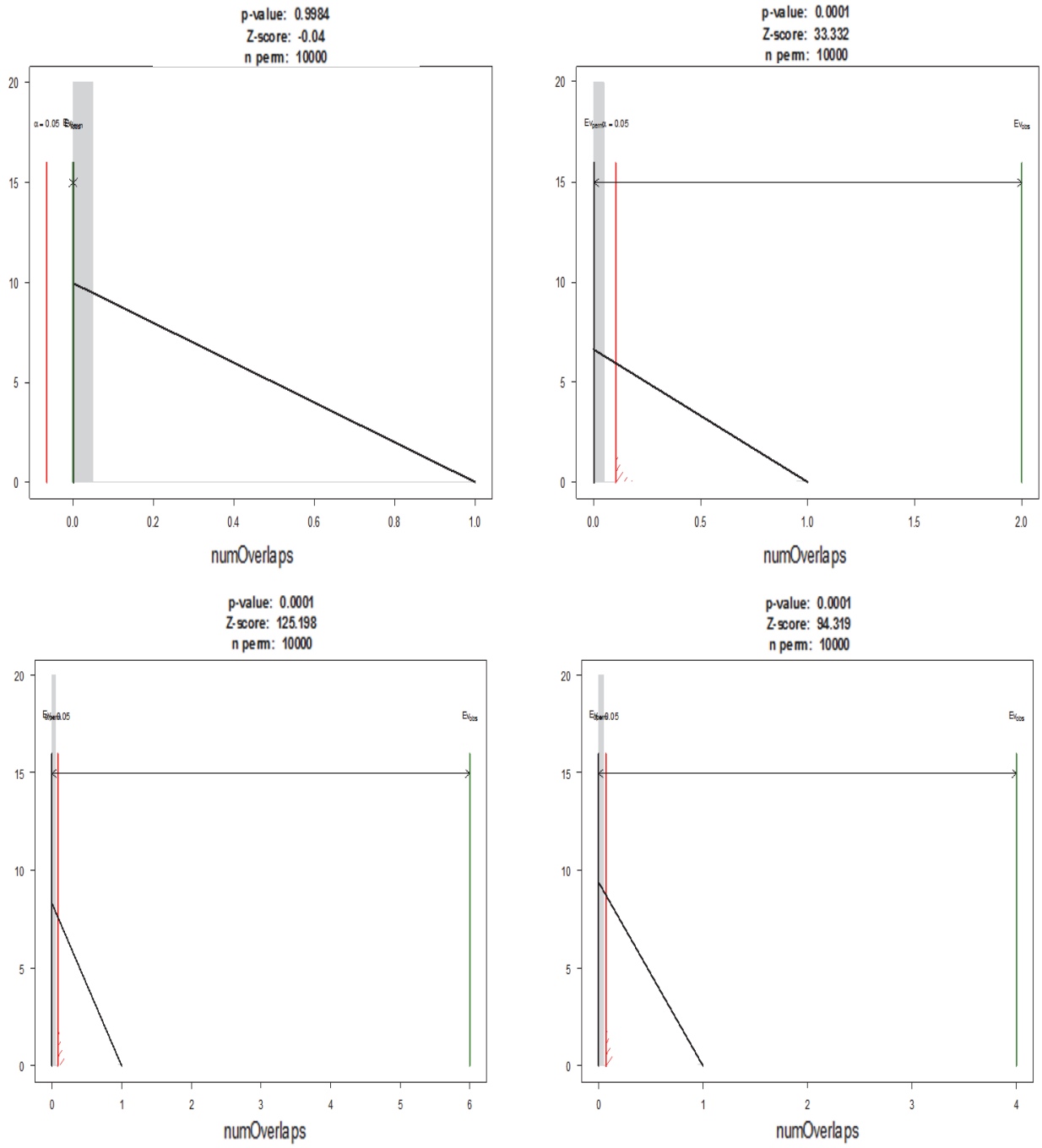
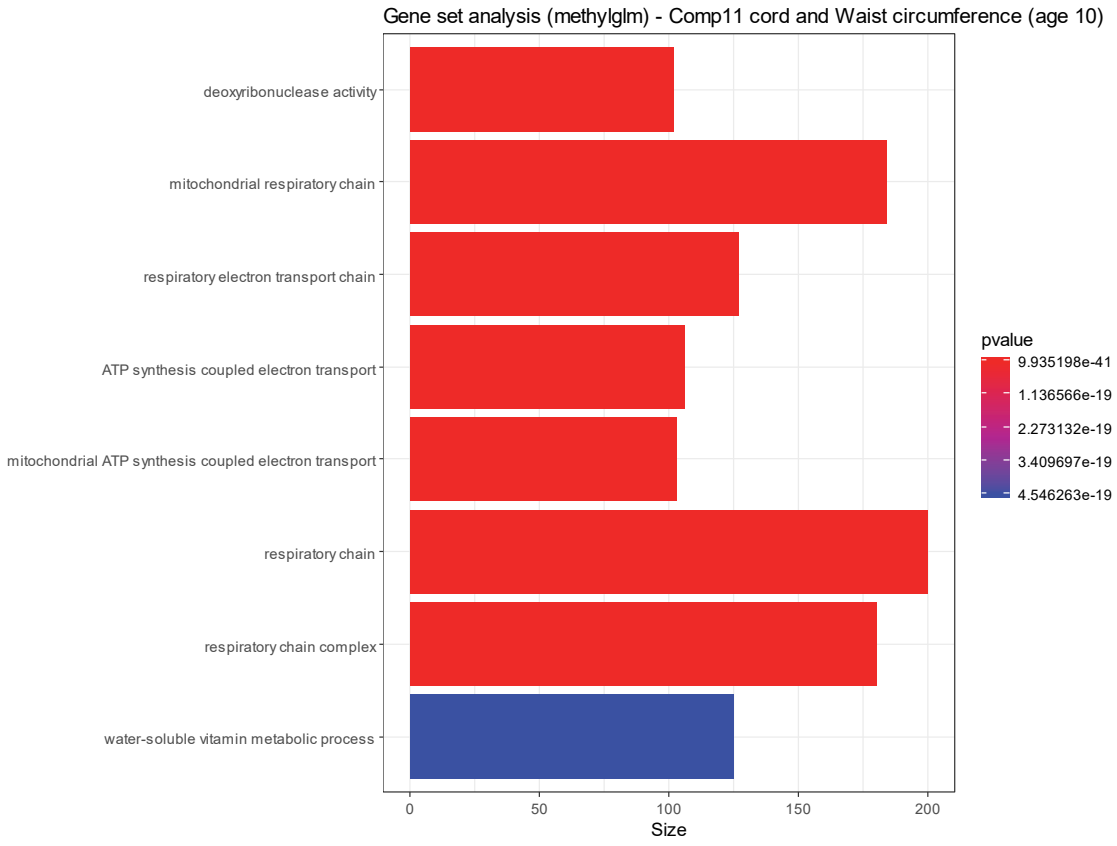


Figure 74: Permutation testing for enrichment at sites of PAI/loop chromatin locations relative to the human autosomal genome. P-value indicated on top of graph. Green vertical line: Observed number of overlaps. Red vertical line: Expected number of overlaps (alpha = 0.05). X-axis: Number of overlaps. Y-axis: Frequency. Cord DNAm components shown from clockwise from top left: Components 7, 9, 11 and 14. Among these, component 7 does not demonstrate more overlap than expected compared to the whole human genome.

3.7.3 Gene set enrichment

We also conducted similar gene set enrichment analysis as in previous epigenetic candidate studies. Figure 75 shows an example of this analysis using Component 11 and obesity related outcomes (waist circumference and lean mass). Like in previous DNA methylation literature for obesity, the most significant enrichment occurs in pathways related to DNA activity and regulation (for examples, see Keller *et al.*, 2017; Xu, X. *et al.*, 2013).



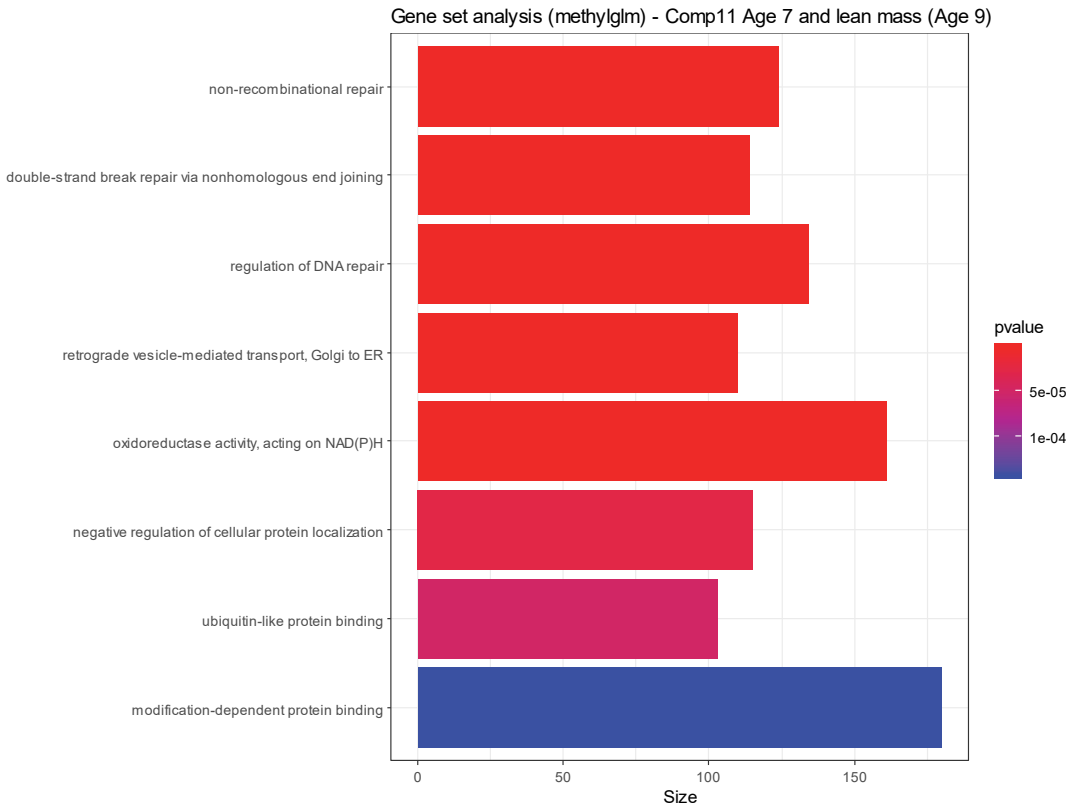


Figure 75: Ranking of gene importance by *p*-value from linear regression with a) waist circumference at age 10 and component 11 in Cord blood or b) lean mass at age 9 and component 11 in Age 7 blood as comparisons. Covariates included estimated cell types, sex, social status and bisulphite conversion batch. R package: methylGSA (default settings for minimum and maximum gene set size as well as gene set list, Gene Ontology).

3.8 Independent validation - Generation R cohort

Just as we extracted the DNAm patterns from blood at later ages in the ARIES, we performed the same procedure on the cord DNAm data available in the GenR cohort. We used the same two-step process with Boruta pre-selection to evaluate the relevance of variables for an anthropometric outcome, body mass index (BMI) in standard deviations (SDs) at age 5 (Figure 76). Covariates included the estimated cell counts, infant sex, SDs of both birth weight and BMI in infancy (measured between 13-17 months), as well as maternal education. Due to missing variable data, the model only included 686 of 969 subjects with available DNAm data. While the

outcomes are not exactly the same, we also juxtapose the performance metrics of this analysis with that of ARIES data for waist circumference at age 10 to get a better sense of the analysis in the two cohorts relative to one another.

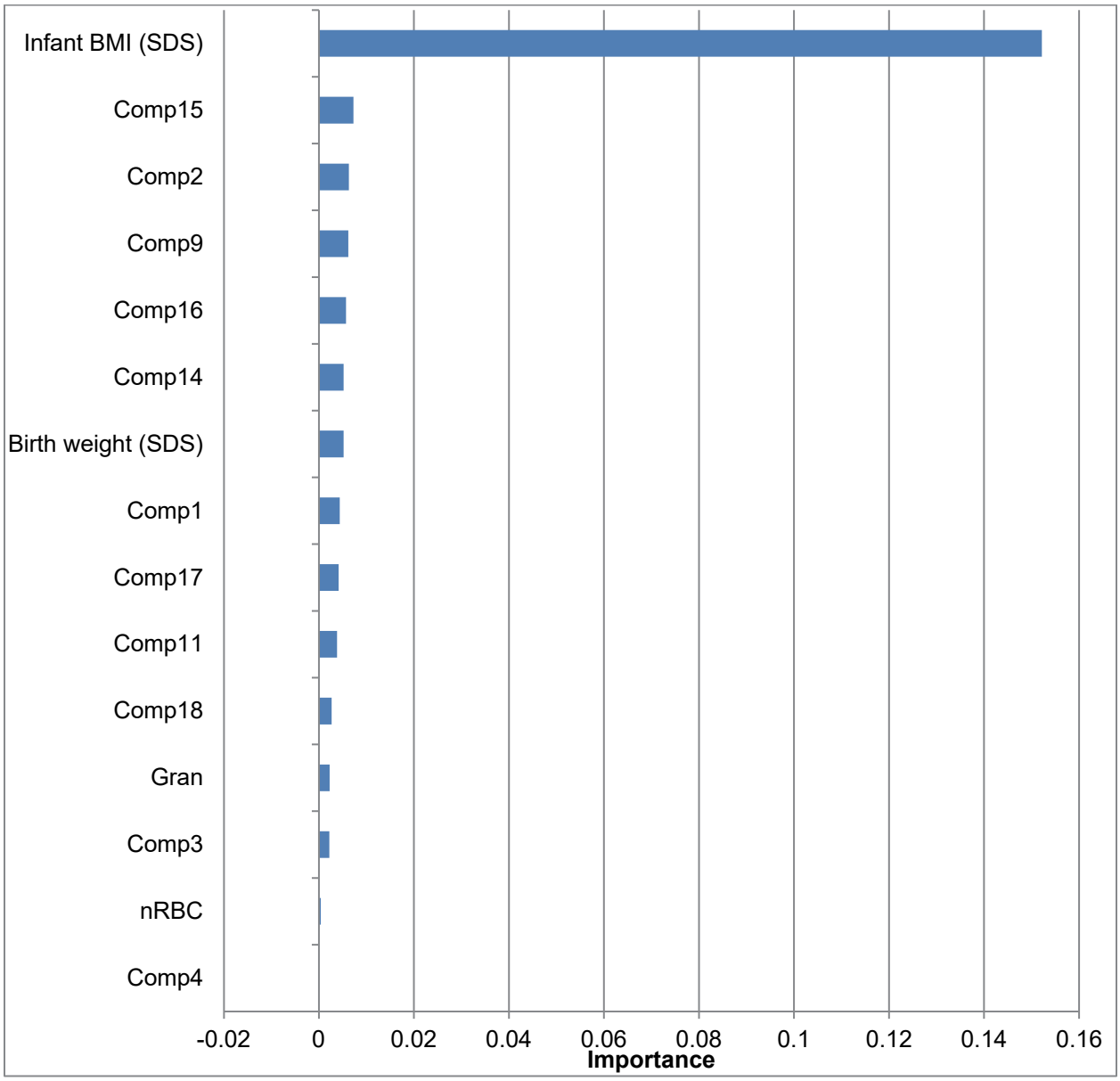


Figure 76: Generation R replication analysis - Age 5 BMI (SDS). Boruta selected variables - importance values (R package rfsrc.) Variables entered: GenR cord DNAm components, estimated cell count, sex, birth weight (SD), infant BMI (SD) and maternal education. (n = 686)

	GenR	ARIES
Sample size	686	862
Number of trees	5000	5000
Forest terminal node size	5	5
Average no. of terminal nodes	93.645	116.164
No. of variables tried at each split	5	2
Total no. of variables	15	5
Resample size used to grow trees	434	545
Number of random split points	10	10
% variance explained	22.05	20.15
Error rate	0.45	0.76

Figure 77: Random forest analysis summary - GenR outcome: Age 5 BMI (SDS). ARIES outcome: Waist circumference age 10.

Like in ARIES, early growth remains one of the most important variables selected. However, more DNAm components were selected by Boruta in the GenR (in total, 11 components were selected) data compared to ARIES for a single outcome. Interestingly, components 9, 11 and 14 were selected in GenR, all of which were also selected in ARIES for anthropometric outcomes such as weight, waist circumference as well as body composition measures of fat and lean mass. However, only component 11 was selected in cord blood in ARIES. Component 9 was selected for such outcomes in DNAm data at ages 7 and 15. Intriguingly, the error rate in GenR is much lower and the variance explained is slightly higher than that in ARIES.

Chapter 4 Discussion

4.1 Key findings

In the past half century, epigenetics has critically informed medicine in diseases like cancer and congenital syndromes. Despite fervid research worldwide and involving tens of thousands of subjects, there is no such translational success for CCDs. Traditional biomarker discovery approaches function well in clinical scenarios with clear distinctions between healthy and disease states. However, these approaches often apply assumptions that may be violated in conditions that have more complex and/or heterogeneous causes and manifestations such as in many CCDs.

We explore context-based views of vulnerability to disease and epigenetic patterns to better map the phenomenon of exposure to maternal smoking in pregnancy and its related clinical and biological topography of health consequences on children. In doing so, we discovered DNAm patterns with the following characteristics:

1. DNAm patterns found at birth traverse all of childhood.

As suggested by numerous studies, DNAm may serve a long-lasting, historical record of exposures. We used a novel multi-dimensional vulnerability composite using genetic and non-genetic variables related to MSP to "bait" the capture of DNAm patterns. These patterns persisted in blood in mid- and late-childhood. These results suggest that DNAm patterns may record not only exposures, but also the net effect of these exposures in the presence of protective and susceptibility factors.

2. DNAm patterns map to regions implicated in important changes in chromatin structure and function.

DNAm patterns covered a large portion of non-genic, CpG poor areas available on the 450K BeadChip. This is interesting because genic and CpG rich areas of the genome 1) are disproportionally over-represented by the 450K BeadChip and 2) characterize the vast majority of reported candidate sites identified by previous EWAS studies of complex traits. As well, DNAm patterns are enriched in specific forms of both active and repressive chromatin marks that may have direct biological relevance to stable changes in chromatin function. For example, the patterns were significantly enriched in enhancer marks that directly interact with gene promoters through elaborate chromatin looping structures.

Based on our current understanding of DNAm-mediated chromatin regulation, these patterns plausibly relate to dynamic processes implicated in stable yet environment sensitive changes in cellular phenotype.

3. There are common DNAm patterns shared among children with similar physical and mental health outcomes and shared among children in two independent cohorts.

Our findings in newborn cord blood support the common molecular origins of various CCD traits by demonstrating the pervasive effect of MSP across anthropometric and cognitive trajectories starting from fetal life. While research often separates physical and mental health domains, epidemiologic and biologic evidence suggests that CCDs may first manifest in one domain but later involve multiple domains. This is in keeping with the current paradigm that CCDs arise from a primary root that gives rise to a constellation of multi-system dysfunctions; this is in contrast to multiple primary dysfunctions. The longitudinal perspective is thus highly important to capture trait evolution.

An independent validation cohort replicated over half of DNAm components found in the discovery cohort. This replication succeeded despite little data harmonization of DNAm data. Future work will expand replication into other ethnic populations and health outcomes. Figure 78 depicts a summary schematic of our workflow.

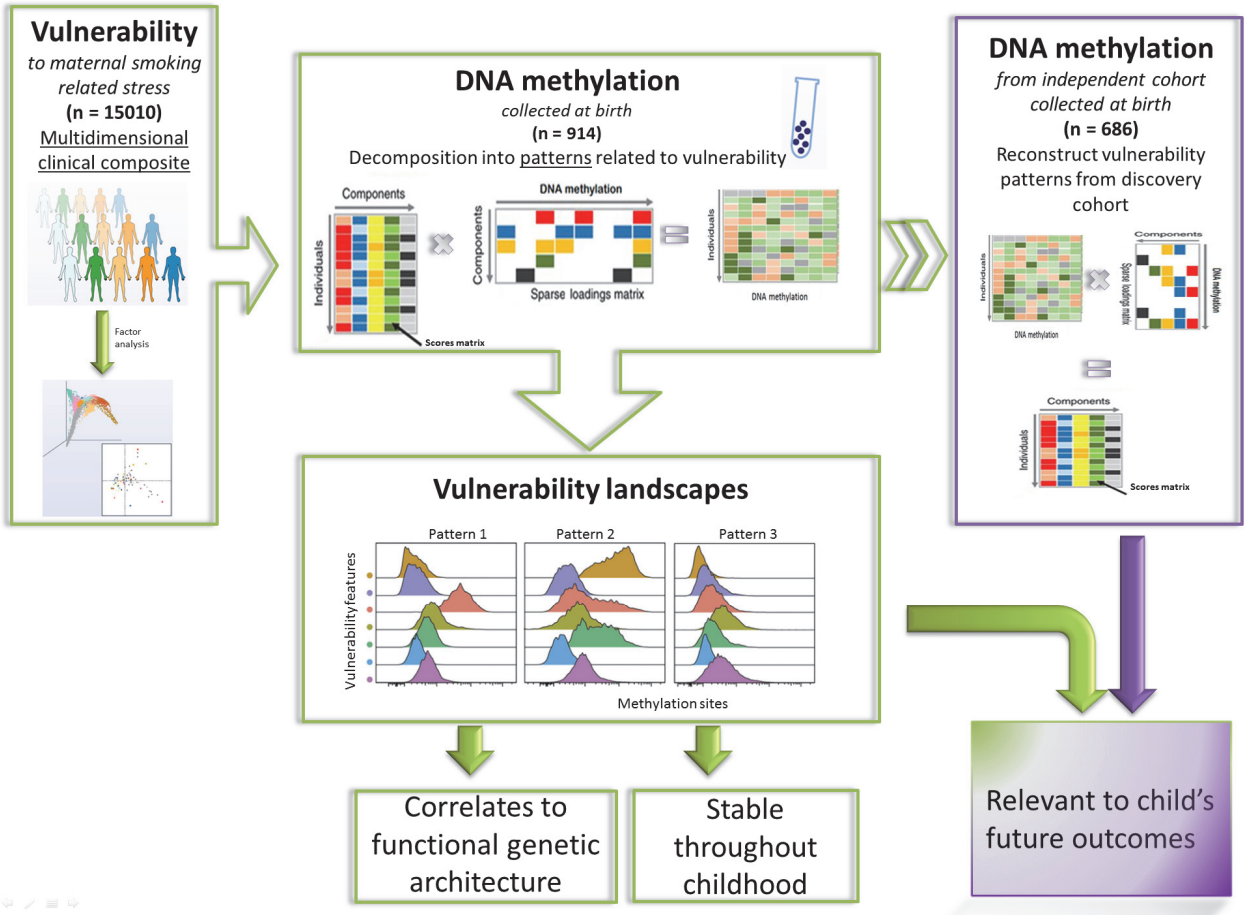


Figure 78: Schematic of thesis deliverables.

4.2 Vulnerability score using multi-class data

In this work, we explored different approaches to represent MSP-related risk to use as bait to capture "vulnerability informed" patterns of DNAm. We first explored using categorical measures of maternal reported MSP followed by PLS-DA, a technique commonly employed in cancer biomarker research. This led to DNAm patterns that provided sharp distinctions between infants with varying MSP exposure but the model was highly over-fit. Next, we attempted to map the extremes of MSP-related vulnerability and resilience by considering the impact on fetal health by including a proxy of intra-uterine growth: infant birth weight. We used typical-atypical MSP categories to group the outer lying mother-infant dyads in terms of MSP versus birth weight relations. While this led to some model improvements, the use of exposure-phenotype extremes in this data set lead to severe class imbalance (primarily due to low numbers of persistently smoking mothers.) As well, we found that covariates like cell type, infant sex or maternal education were associated with most of these typical-atypical related components. We next turned to a continuous measure of MSP vulnerability, a multi-class composite of MSP-related variables and birth weight. Using factor analysis, we arrived at a five dimension composite that appeared to capture five different and clinically plausible profiles that we will examine more closely here.

Referring to Table 11, we recall that Dimension 1 is positively related to grandmaternal smoking history and mother's exposure as a fetus to smoking from the grandmother. As such, it may relate to genetic or transgenerational MSP factors (i.e. grandmaternal smoking and maternal exposure to smoking as a fetus.) In contrast, Dimension 2 is related to maternal factors: the dimension score increases as report of MSP and pre-gestational history of smoking increases. There also appears to be an inverse relation to birth weight, albeit weak. Therefore, this dimension could be describing typical sensitivity to smoke exposure. Dimension 3 is mainly related to non-maternal line smoking. It is most related and positive correlated to smoking of the mother's partner and household members. Interestingly, it is inversely related to grandfather smoking but is not related to grandmother smoking. Last, it has an inverted U-shaped relation to MSP: low scores correspond to both no or high levels of reported MSP, but high values correspond to non-persistent MSP. This dimension may reflect direct environmental smoking background and may be an important contrast to the potentially genetic influences related to Dimension 1. Dimension 4 increases in relation to higher birth weight and demonstrates the same inverted U-shaped relation to MSP as Dimension 3. One may

speculate this reflects positive effects of maternal attempts to limit MSP. In contrast, Dimension 5 is positively associated with birth weight but also to household member smoking. Like Dimension 4, Dimension 5 may represent atypical resilience to risk, or seen another way, the presence of protective factors associated with these risks. Interestingly, unlike Dimension 3 and 4, higher scores in Dimension 5 correspond to non-smoking and persistent smoking status in pregnancy. In the case of a mother surrounded by other household smokers but abstains from smoking herself, this may offer a protective set up for her child. Mothers in the same situation who continue to smoke off and on in pregnancy may not similarly offer such protection. In the case of a fetus that thrives despite smoking in both the mother and household members, the fetus itself may have intrinsic resilience to this toxic exposure. There is a fascinating body of research specifically on patterns of maternal smoking in pregnancy. For example, Pickett *et al.* used a cohort of over 18,000 mother-infants pairs (Millennium Cohort Study) to investigate the relation between non-smokers, quitters and light or heavy smokers in pregnancy and infant temperament (Pickett, Wood, Adamson, D'Souza, & Wakschlag, 2008b). The authors described a "protective effect" of smokers who quit during pregnancy in relation to decreased risk of infant problem behaviours. Further, Massey *et al.* performed extensive laboratory and home based researcher observations of mothers and their responsiveness to their infants in a prospective pregnancy study (Massey *et al.*, 2018). This study used both maternal self-report, maternal urine and infant meconium at delivery to categorize smoking pattern groups, as well as to clarify the prenatal use of any other substances, such as alcohol, cannabis and opioids. They found spontaneous tobacco quitting associated with increased maternal responsiveness to her infant compared to all other groups of mothers. Interestingly, spontaneous quitters were also more responsive to their infants than even never-smoking mothers were. This was despite spontaneous quitters still having higher risk factors than never smokers do, such as more problems at school, with the law and at work. In this thoroughly profiled cohort, the authors speculate that this association was due to resilience conferred by successful quitting during pregnancy that is independent of nicotine addiction or social factors. This area requires further study but holds great potential for expanding the discovery of novel and clinically important targets beyond the “weakness” perspective to fostering strengths that promote thriving, healthy children.

Being a continuous variable, we had expected that birth weight would be one of the major contributors of the composite variables (Rhemtulla, Brosseau-Liard, & Savalei, 2012). We observed this directly when trialing the use of the R package VarSelLCM (Marbac & Sedki,

2019) to create the composite variable. This was unsurprising because many studies have found that non-continuous variables with less than four categories tend to be more weakly related to a latent factor than if it were continuous. VarSelLCM is a clustering method that attempts to separate individuals into mutually exclusive groups. However, factor analysis methods like FAMD attempt to group variables by their relation with each other. Individuals have scores for each group (which we call Dimension in this thesis) and the overall pattern of variables is described by the combination of Dimension scores. While both are pattern-finding methods and clustering is arguably easier to envision and interpret, it may not have fit our goal to build a multidimensional image of an infant's vulnerability to MSP-related phenomena.

We take special note of Dimensions 3, 4 and 5 that demonstrate an atypical relation to MSP i.e. there is no linear "dose-response" to direct MSP. Specifically, non-persistent smoking does not fall in between non-smoking and persistent smoking. This could be a spurious finding perhaps due to the bias and/or inaccuracy of self-reported MSP. Alternatively, the relation is real but seems atypical because one assumes that increasing MSP always has the same direction of effect in relation to other variables. This alludes back to our discussion in Section 2.3.2.1: Integrating variables - Methodologic assumptions regarding variable relations regarding researcher-imposed assumptions on the ordinality and weighting (i.e. importance) of a variable within a given context. In many cases, such arbitrary assumptions in a given population could introduce errors of unknown impacts. We argue that "non-linearity" is merely a matter of perspective. In fact, we posit this is one of the strengths of context-based analysis. We remind readers that these dimensions originate from data from eight variables analysed simultaneously. While humans can easily visualize positioning in 2-D and 3-D space, it is harder for us to conceptualize this in 8-D space. The "non-linearity" of the MSP relation may be difficult to reconcile in a dose-response paradigm, but mathematically represents the most optimal configuration to represent the variability in vulnerability profiles of these children.

4.3 DNA methylation patterns

We used PLS-R to overlap the vulnerability profiles with DNAm data to uncover what patterns emerge. As seen in Table 23 and Figure 57, Dimension 1 is the most represented dimension among the DNAm components. Dimension 2 and 5 are tied as the second most represented by number of DNAm components. We were surprised to find that a number of vulnerability dimensions related most strongly to various sources of smoke exposure instead of weight, like

Dimension 1 and 2. Nearly 75% of all the components are linked to either Dimension 1, 2 or 5. As well, each dimension takes almost an equal share of the components. Birth weight has repeatedly been shown to be related to cord DNAm, though the direction of this relation is unknown (Agha *et al.*, 2016; Haworth *et al.*, 2013; Kupers *et al.*, 2015; Michels, Harris, & Barault, 2011; Suter, M. *et al.*, 2010). This is in contrast to the weak or complete lack of findings of the relation between cord DNAm and smoke exposure from non-maternal sources, like grandparents (Joubert *et al.*, 2014), fathers (Bouwland-Both *et al.*, 2015; Joubert *et al.*, 2014; Lee, K. W. *et al.*, 2014; Rauschert *et al.*, 2019; Wu, C. C. *et al.*, 2019) and other household members.

The relation between Dimension 1 and DNAm was particularly surprising. Represented by relatively time distant MSP exposure (i.e. grandmaternal smoking characteristics,) we did not expect it to be such a prominent feature among DNAm patterns in relation to more proximal variables such as maternal smoking, (such as that represented by Dimension 2) or birth weight, (like in Dimension 4 and 5.) To be sure, the impact of grandmaternal exposures on grandchildren is supported by data from human cohorts from the past half century, with more specific findings for grandmaternal smoking in more recent cohorts and animal models (Babenko, Kovalchuk, & Metz, 2015; Drake & Walker, 2004; Hypponen, Smith, & Power, 2003; Rodgers & Bale, 2015; Tremblay, 2010)for examples and reviews.) Intergenerational effects specifically on human DNAm are also supported, for instance in a study of grandmaternal exposure to stress (Serpeloni *et al.*, 2017) and lead (Sen *et al.*, 2015). However, the results are very inconsistent in this arena. In their investigation of possible transgenerational effects of smoking, Joubert and colleagues found none of the 26 CpG sites examined showed any relation between grandmother smoking during pregnancy and grandchild cord DNAm methylation (Joubert *et al.*, 2014). The results of this work may be the first to contradict this earlier finding by Joubert *et al.*. Alternatively, it may offer evidence to encourage better matching of study design to hypothesis. This group started with data from the same 450K chip as in ARIES. However, they only pursued testing in 26 CpG sites that passed significance testing in relation to maternal smoking categories. We posit this pre-selection based on maternal data limits the scope of detection of grandmaternal effects and may prove a cautionary tale against discarding DNAm without better understanding of its biological and clinical contexts, (see Section 1.5: Mapping individual epigenetic data to genome wide patterns.) Nonetheless, there is a robust correlation between grandmaternal and maternal smoking found in both numerous epidemiologic and genetic studies (for review, see Buck *et al.*, 2019). So, if transgenerational

effects exist, it is not unreasonable that at least one of those 26 sites would have a relation to grandmaternal smoking. This would particularly be true in a well powered study like in this case, (Joubert *et al.* used data from 1042 newborns from the MoBa cohort.) In that case, we speculate whether the problem lies in the initial 26 selected CpGs. As discussed earlier, reliance on a single self-reported measure exposes the researcher to possibly biased estimates of true MSP-related impacts, both biologically and clinically. Inconsistencies between studies can be a good opportunity to re-evaluate the foundation of our hypotheses and study design. We suggest that a good place to start is improving the accuracy and relevance of measurements, such as considering the use of multivariate estimates to obtain the most comprehensive triangulation of a subject's individual susceptibility as explored in this work.

The relation between Dimension 3 and DNAm patterns is supported by previous work on the impact of second hand smoke exposure. For instance, Guerrero-Preston *et al.*. examined difference in global DNA methylation in cord blood between 3 maternal groups: non-smoker, second-hand smoker or active smoker (Guerrero-Preston *et al.*, 2010). They found a dose-like response in that progressively lower levels of methylation were found moving from the non-smoker, second hand smoke exposed and active smoker groups. To our knowledge, we are the first to identify DNAm patterns related to paternal or household member smoking.

We were fascinated by why there was unequal representation of certain dimensions by DNAm patterns. One possibility is that there is more than one overlapping pattern of variability in the newborn's DNAm that relate to the profiles represented by these dimensions. We speculate that if the dimensions represent an origin point for vulnerability to MSP-related risks, then it would reason that multiple altered biological processes would propagate from this common diathesis. Even if this diathesis resulted in a change in only one area of chromatin architecture, we have already seen in Section 3.7.2: Topology based context that this can lead to a cascade of changes in the physical interaction of chromatin and transcription machinery both locally and distal to the change. If this were the case, then DNAm differences could either mediate the effect of the diathesis or be the result of the diathesis related change in chromatin. Finer mapping of the overlap of DNA patterns with 3D chromatin data is needed to interrogate this possibility.

The fact that some dimensions are over-represented by PLS components may lead to a reasonable suspicion of confounding. It is true that the MSP variables are highly susceptible to confounding as they are strongly linked to socioeconomic factors. And because socioeconomic

factors are linked to child outcomes, one could argue that the PLS components could represent confounding rather than biological DNAm differences.

To dissect this possibility, we walk through the process of component construction bearing this argument specifically in mind. First, the creation of the MSP composite dimensions preceded and was entirely independent of the DNAm components. The clustering of certain variables could certainly be bound by confounding and not by biological processes. For instance, Dimension 3 is linked to 4 factors: grandpaternal smoking, household members smoking, maternal partner smoking as well as MSP. While researchers debate the possible biological reasons underlying the clustering of smoking behaviour, it very likely is strongly influenced by non-transmissible SES factors that lead to the coincidence of these risks, (for examples of studies countering these points, see (Cnattingius, 2004; Fertig, 2010; Freathy *et al.*, 2009; Hypponen *et al.*, 2003; Munafo, Freathy, Ring, St Pourcain, & Davey Smith, 2010; Shenassa, Papandonatos, Rogers, & Buka, 2015; Siahpush, 2006; Stroud *et al.*, 2014).) We have already discussed previous attempts to separate MSP-related effects, for instance to compare intra-uterine from environmental effects (e.g. see Brion *et al.*, 2010) or to control for selection bias (e.g. see Fertig, 2010). However, the reality is that exposure to MSP would unlikely occur in isolation. Furthermore, the effect of MSP would not homogeneously affect all individuals in a population. To this end, we ask to what clinically relevant end does attempting to study the isolated effect of MSP alone provide for translational medicine? At our current state of knowledge, we may better benefit by first asking: How does the constellation of MSP-related factors known to influence health outcomes affect the biological programming of the child? It is in this spirit that we attempt to represent MSP as a phenomenon in its most “natural” context instead of artificially dissecting maternal smoking from its associated factors. We argue that “confounders” actually provide important context in which to couch the biological effect of MSP. As such, we believe it is critical to include this context in order to truly unearth the reality of MSP effects on the fetus and child. We use the MSP composite to provide the coordinates of the unique vulnerability “space” occupied by each individual.

Second, we used the MSP composite to bait the capture of related DNAm patterns using PLS-R. As succinctly described by Hervé Abdi: “the goal of PLS regression is to predict Y from X and to describe their common structure.” (Abdi, Hervé, 2010). PLS-R does not seek to capture the most data variability. Instead, PLS-R is seeking a pattern of DNAm found among individuals that share similar coordinates in the “space” of MSP vulnerability. The PLS components are agnostic to what variables or class of variables the composite represents. PLS does not

consider how much variability from each constituent variable was captured by the dimension when decomposing the DNAm data. And there is no potential bias for PLS-R in our application to favour continuous or categorical MSP variables. Therefore, PLS components are not directly influenced by which variables most heavily contribute to a given dimension. In other words, PLS-R asks "what patterns in DNAm best describe the differences and similarities of children based on their position in the MSP vulnerability space?" rather than "how to use DNAm to account for the most variability in any single MSP variable like birth weight, maternal smoking, grandmaternal smoking, etc.?" We demonstrated the influence of MSP variables on outcome in Section 3.5: Comparison of performance: DNAm patterns versus MSP variables or composite in relation to child outcomes. When using the MSP related variables only (i.e. no DNAm components) in RF, few or none of the variables were selected above the covariates, (e.g. rate of infant growth, sex, cell count, etc.,) as relevant to outcome in the ARIES subset. Knowing this, we argue that whether the composite was formed from biological, social, nutritional and/or other factors, the DNAm pattern remains a molecular reflection of the composite profile. In this manner, we are able to attenuate direct exposure of our clinical outcome models to any single MSP-related variable and its inherent bias and measurement error. We argue this may improve the accuracy and robustness to clinically-irrelevant confounding in our detection of DNAm signals which translate to better outcome models.

Third, PLS is an iterative algorithm i.e. after the first component is extracted, it is "subtracted" from both the DNAm and composite projections before the algorithm seeks the second component and so on (Abdi, H. & Williams, 2013; Beaton, , , Saporta, & Abdi, 2019). The recurrent "selection" of certain dimensions is not due to it being "leftover" and then re-selected. It is due to the PLS model finding another projection that captures the next biggest overlap between DNAm and MSP composite data. This would support that there are different DNAm patterns underlying these recurring dimensions. In this way, a given vulnerability profile could be seen as causing more than one change in methylation-mediated chromatin structure. Another interpretation is that there is correlation in DNAm patterns of individuals who score similarly in Dimensions 1, 2 and 5. This group of individuals would drive up the chances of all three dimensions being targeted by the PLS model. Another way of thinking of this is that a given DNAm projection overlaps with more than one dimension. This could occur if a common DNAm pattern is related to more than one MSP profile, suggesting a common molecular disruption can be caused by different sources.

Put together, there are both biological and statistical concepts that could explain how and why certain MSP profiles are more often selected by the PLS model. We suggest that refining the MSP vulnerability mapping technique and re-testing this pattern finding in DNA methylation and other biomarker data (e.g. metabolite, RNA-seq, etc.) would be a robust means to test whether these vulnerability profiles do indeed point to common molecular intersections.

4.4 DNA methylation vulnerability relates to future child outcomes

Using RF to relate MSP-composite derived DNAm components to outcomes, we found a number of frequently appearing DNAm patterns in ARIES that appeared to be clinically relevant no matter the age of blood collection or clinical outcome, (e.g. Components 1, 2 and 3.) This was replicated in the GenR cohort. Like in other omic studies, there emerge biomarkers that appear almost ubiquitous (Cauchi *et al.*, 2016). They may emanate from biological pathways that intersect coincidentally or causally with the process under study. With more understanding of their interaction in time and intensity in context of more specific markers, it is possible they could also be useful diagnostically and/or therapeutically. However, the seemingly low specificity of these patterns within the scope of information and analysis available in this exploratory work remains a mystery that merits future investigation.

Leaving aside these ubiquitous components as well as those related to covariates, we found that Component 5 had pervasive relevance across outcomes but had an inconsistent pattern of timing, whether considering DNAm age and age of outcome assessment. Components 7 and 14 were also frequently called. However, these two components had weak but significant relations to covariates at Age 7 and 15 and therefore must be interpreted with extra caution. Components 9 and 11 were most frequently called as relevant to outcomes across DNAm ages and outcome measurement ages. Component 11 was the most consistent component in terms of relevance in cord, Age 5 and Age 17 blood. This was related to Dimension 5 (atypical resilience to MSP.) Components 9 and 14 were both related to Dimension 1 (genetic/transgenerational MSP factors.) All three of these components were selected in the GenR replication cohort, (further discussion below.) We speculate that the stability of these components across time within individuals and between individuals, even from an independent cohort, suggest they relate to an intrinsic property of the child and/or the early environment. This would also explain their persistence across outcome types.

Component 5 was most related to Dimension 3 (mainly non-maternal line smoking.) The variability in dose of environmental sources of smoke exposure may relate to the apparent pattern-less relevance of Component 5 to clinical outcomes and serves as an interesting contrast to components like Component 9 which were related to genetic/transgenerational MSP factors.

Component 5 also is a good point of discussion regarding reverse causality. For anthropometric measures, it was selected in blood at Age 7 and Age 15 and not at all in cord blood. Thus, it could be argued that only finding DNAm relations with greater temporal proximity to the outcome suggests that the outcome influenced DNAm rather than vice versa. Increasingly, this field has attempted to clarify the direction of causality between DNAm differences, exposures and outcomes. Enhanced pooling of epigenetic data and long standing longitudinal cohorts has allowed more power to conduct causal analyses such as Mendelian randomisation and cross-lagged models (Richardson *et al.*, 2019; Richmond, Al-Amin, Smith, & Relton, 2014; Wiklund *et al.*, 2018). However, these methods often require direct (i.e. non interactive) relations and adequate effect sizes in the setting of longitudinal human cohorts in order to avoid Type II errors. As such, these forms of interrogation may poorly fit the study of complex diseases. In this work, there are several aspects of the study design that argue against reverse causality.

First, the uncovered DNAm patterns were generated from the intersection of the MSP composite variable and cord blood DNAm data. Thus, the "weightings" or relative importance of a given DNAm site was determined by the overlap between the composite and values of methylation at the time of birth. Thus, the association between the patterns found in cord blood with later outcomes could not be due to reverse causation. However, for patterns found in blood at ages 7 and 15, there still exists the possibility that some DNAm variability in later life attributable to outcome related factors affected these older age DNAm patterns. We remind readers that the older age DNAm patters were derived from the "template" of weightings derived from *cord* blood. As such, this should attenuate the chance the relations between *cord-based* DNAm patterns seen in later childhood were caused by the health outcomes around this time of life.

Previous studies have shown that early exposures leave an enduring mark on DNAm. It could be the vulnerability composite is related to later outcomes and are unrelated to the DNAm patterns originating in cord blood. Therefore, there is low likelihood that outcome is importantly responsible for the Age 7 and Age 15 DNAm patterns.

Second, unlike previous studies, we performed no pre-selection of candidate sites using outcome to then create a score or other quantity to act as a "predictor". Our design was motivated purely by seeking exposure related patterns and then subsequently observing if these patterns are relevant to later development. Third, we had more DNAm samples with age. Thus, we are expanding our sample size as we test the association between later DNAm patterns and outcome. This increased power to detect associations may be responsible for the greater number of DNAm component "hits" at later ages.

Last, we are not the first to observe that DNAm at later ages is more strongly related to outcome. Many studies have observed this phenomenon where higher correlation of DNAm variability occurs the more temporally proximal the outcome measure (see (Agha *et al.*, 2016; Cao-Lei *et al.*, 2019; Reed *et al.*, 2020) for recent examples.) As well, it would be unreasonable to suspect that the physical context of the organism's current state has no relation to biological markers like DNAm, whether through causal or coincident pathways. The nucleus is a busy web of trafficking and connecting structures where molecules compete for space and equilibrium. Changes in one domain would likely have a ripple effect on other areas. Whether that leads to clinically relevant effects may depend on context but we do not argue the possibility of reverse effects. Instead, we are curious how reverse effects interact with MSP related effects and at what point do we expect the reverse effects to indicate that no causal MSP effects are present? This would be best clarified with functional testing but it is hard to imagine an answer equally applicable to all epigenetic candidates exists.

Relating specifically to Component 5, there are additional points that argue against reverse causality. First, very few components were selected in cord blood for anthropometric outcomes, so Component 5 is more like the rule than the exception in this case. Second, Component 5 was selected in cord blood for a number of academic performance time points. This makes it less likely that it's selection at age 7 and 15 was purely a consequence of the child's anthropometric outcomes. Third, another possibility is that one's epigenetic profile is more "set" at older ages. While epigenetic programming is ongoing throughout the lifespan, studies in monozygotic twins have clearly shown that DNAm divergence of these siblings widens with age (for example, see Talens *et al*, 2012.) This suggests that environment related methylation variation stably accumulates over time. One of our objectives was to identify genomic areas sensitive to MSP. According to our hypothesis, these areas should be susceptible to environmental changes, as opposed to more "conservative" areas. We speculate that the young newborn may not have yet "set" their trajectory in cord blood, but that as time passes

more and more children fall into the developmental path that their MSP-related origins were pushing them towards. Mathematically, the "clinical relevance" of a DNAm component relies on how many individuals in the sample demonstrate MSP-related (or is related to a covariate of MSP) outcomes. Hence, the apparent "increase" in relevant DNAm components relevant to outcome is actually just an increase in children more obviously following their MSP exposure trajectory. Therefore, the increasing concordance of differential DNAm with age could also be interpreted as a "lag" between cause and observable phenotype changes. In other words, the affected DNAm may need time before one can see it "set in". Is it because we fail to detect earlier phenotype changes or because more "hits" to the epigenome need to accumulate or do other effects "override" exposure-mediated earlier in life (e.g. early infant growth in the first 3-6 is strongly influenced by *in utero* environment/maternal factors and then becomes increasingly sensitive to environmental influences)? While this remains a mystery, the potential of DNAm to be the canary in the coal mine makes it an ideal pre-clinical disease biomarker.

The answer to these questions could potentially be better understood with methods like Mendelian randomisation (a specific instance of IV) analysis (Richmond *et al.*, 2014; Wehby *et al.*, 2011; Wiklund *et al.*, 2018). If these components could be statistically "tethered" to a suitable IV of the fetus and mother such as a genetic variant that is unrelated to clinical outcomes. As mentioned, it would require a very large sample size and would likely involve pooling multiple cohorts that all have collected the data on the same IV. Fortunately, with the DNAm pattern template, only DNAm data are required after that requirement. The cohorts would not require any MSP related data.

That we found little to no relation with behaviour was unsurprising. There is a relatively large mass of EWAS literature regarding child mental health outcomes ranging from ADHD and autism to infant temperament traits. While the initial findings that were reported were often positive and typically using only a few methylation sites, more recent findings have provided more genome-wide coverage and have been less consistent, (for examples, see Bale *et al.*, 2010; Devlin, Brain, Austin, & Oberlander, 2010; Petronis, 2010 versus Knopik *et al.*, 2019; Hamza *et al.*, 2019; Meijer *et al.*, 2020; Taylor, R. M. *et al.*, 2020 and evidence of causality in human studies is weak (Cecil *et al.*, 2018; Knopik *et al.*, 2019). While it is unclear if this is due to changes in methodology, technique or publication bias, we do know that pediatric mental health is an important but highly heterogeneous outcome to study. Among child outcomes, it may have the most complex and poorly understood relation to social determinants of health, including MSP (Bradley & Corwyn, 2008; Cents *et al.*, 2013; Crockenberg & Leerkes, 2004;

Hanington, Ramchandani, & Stein, 2010; Jansen *et al.*, 2009; Lee, L. C., Halpern, Hertz-Picciotto, Martin, & Suchindran, 2006; Palmer *et al.*, 2016; Rueda & Rothbart, 2009). We speculate whether this outcome may also benefit from re-evaluation as to the precision and accuracy of measurement, with consideration of integrating genetic influences (e.g. Bale *et al.*, 2010; Oberlander *et al.*, 2010; Rodgers & Bale, 2015; Stergiakouli & Thapar, 2010) and the evolution of natural versus pathologic changes in cognition and behaviour that occur in childhood over time (De Pauw, Mervielde, & Van Leeuwen, 2009; Guerin, Gottfried, & Thomas, 1997; Lemery, Goldsmith, Klinnert, & Mrazek, 1999).

In this first pass effort to explore integrated MSP-methylation-outcome analysis, we did not include outcomes like asthma and allergy. While there has replicated findings of relations between MSP - child DNAm (Joubert, Bonnie R. *et al.*, 2016; Reese, S. E. *et al.*, 2017) and child DNAm – asthma (Nicodemus-Johnson *et al.*, 2016; Reese, Sarah E. *et al.*, 2019), the link suggesting DNAm mediates MSP and outcome remains questionable. To our knowledge, only Neophytou and colleagues (2019) have successfully demonstrated a mediation effect. However, this was only at a single locus and was only related to parental report of physician diagnosed asthma status (located in the *AHRR* locus, cg05575921 indirect OR was 1.18 (95% CI = 1.07, 1.68).) In addition, this finding failed to correspond to associations to clinical measures of asthma control and lung function tests (i.e. confidence intervals all crossed zero.) This area continues to undergo active research; however, there is so far no evidence of a connection. However, asthma and allergy are primary inflammatory diseases. With numerous smoking related DNAm hits related to inflammatory pathways and our current state of ignorance regarding the mechanisms of action of DNAm in complex trait processes, asthma and allergy may require specific study design considerations to best tease apart pathogenic versus bystander effects.

Last, we noted with surprise that several DNAm components were related to height, (e.g. Components 5, 7 and 11.) Unlike adiposity, numerous epidemiologic studies suggest that genetic factors account for up to 90% of adult height variation (Lanktree *et al.*, 2011; Shah *et al.*, 2015). In contrast, adiposity has demonstrated heritability of only between 40-70% (Herrera & Lindgren, 2010) as well as a growing body of research supporting epigenetic (including differential DNAm) influences (for example, see Sharp *et al.*, 2015). However, it is plausible that epigenetics also modifies height. Simeone *et al.* observed that over 80% of height associated genes contain CpG islands, suggesting they are susceptible to DNAm-mediated regulation (Simeone & Alberti, 2014). Other epigenetic differences previously implicated in influencing

adult height include micro-RNA and open chromatin regions (Muthuirulan & Capellini, 2019). Across human populations, height has an intricate relation with social status, infectious disease and nutrition. While height may or may not have a causal role in this web, the possibility that stature has modifiable aspects is intriguing and highly worthwhile for human health and social well-being. Our work further supports exploration of this possibility.

4.5 Effect size

Our study was designed for hypothesis generation with the goal of exploring various mapping methods for clinical and DNAm data. We employed methods that do not lend themselves to standard calculations of effect size on childhood outcomes. We did so with the belief that effect size is not necessarily a major contributor of importance, as evidenced by the publication of impactful EWAS literature with effect sizes as low in magnitude as 1% in both adult (Esposito *et al.,* 2016; Stringhini *et al.*, 2015; Rakyan *et al.* 2011) and pediatric (see excellent recent review by Shanthikumar et al. 2020) outcome studies. In fact, small effect size motivates the filtering of methylation probes displaying low variance.

Nonetheless, we venture to discuss the lack of effect size information provided by our work. The best estimates of effect size assume linear and direct relations between the exposure, biomarker and/or outcome. As discussed in the literature by groups such as Rakyan *et al.* (2011), the estimate size is highly dependent on the sample size. To expect a large and direct relation between our DNAm patterns and outcome is unreasonable for several reasons.

First, as discussed in Section 1.5.2, the estimated effect size in context of high noise and with our sample size is likely very small. To rely on effect size to judge importance would expose the work to an anticipated high chance of false negative results.

Second, blood is a mixed cell organ. We use it as a proxy for epigenetic changes in tissues related to growth and mental development, which involve experimentally inaccessible tissues such as endocrine and nervous tissues. Therefore, our measurement tool aims to capture an indirect and incomplete “shadow” of the change in epigenetic poise of the target tissue. Moreover, our capture of epigenetic signals employs a non-categorical measure of vulnerability. Our view of vulnerability is a multidimensional and interactive composite of genetic and non-genetic factors. In essence, we seek to capture surrogate DNAm patterns that reflect the predilection for specific cell fates in a distinct target organ that likely is mixed cell type itself.

This scenario is poorly suited to the use of a method that assumes direct and non-interactive relations.

Third, as discussed in Section 1.2.1, current biomarkers largely function by detecting the by-products of cellular damage or dysfunction, e.g. serum enzyme levels due to cell lysis or increased production of inflammatory signals due to injury. We challenge the paradigm that clinical tools should depend on the aftermath of injury. We posit that it is possible to detect vulnerable molecular poise that can alert clinicians of a maladaptive health trajectory before disease onset. For instance, Rakyan and colleagues (2011) identified 132 differentially methylated positions that distinguished type 1 diabetes status in discordant monozygotic twins in blood monocytes. They found these sites displayed differential methylation before diabetes diagnosis in a separate data set (Rakyan *et al.* 2011). However, this was using a clear binary outcome in adults. Given that we are linking epigenetic poise and childhood antecedents of potential disease, we employ entities which are mathematically "fuzzy" i.e. are difficult to characterize using dose-response type models between risk and outcome.

However, we were curious about the impact of candidate sites identified in other studies of childhood outcomes. We surveyed the literature for "effect sizes" of DNAm in this regard. Using mediation analysis, Kupers *et al.* found that differential methylation at 3 CpGs at the GFI1 locus explained 12-19% of the lower birthweight in smoking mothers (Kupers *et al.*, 2015). This was observed in a meta-analysis of three large pregnancy cohorts: GECKO, ARIES & GenR. Later, Valeri and colleagues attempted to replicate these findings in the MoBa cohort (Valeri *et al.*, 2017). After adding statistical correction for misclassification of maternal smoking, this group found a much weaker evidence of methylation mediated differences in birth weight than reported by Kupers *et al.* As well, Valeri *et al.* went on to perform further analysis using maternal cotinine levels available to the MoBa study to "correct" the self-report of 18 mothers who denied smoking in pregnancy. This subsequent analysis showed no significant mediation effect by methylation.

More recently, Reed *et al.* used DNAm scores to investigate the direction of influence between DNAm and birth weight and BMI in later childhood using the same ARIES data. The scores were calculated by "multiplying the CpG site methylation levels in that profile with the corresponding published effects estimates and then summing the products."(Reed *et al.*, 2020) They found the DNAm scores explained the most concurrent BMI variance as the child ages (i.e. cord DNAm scores explained 1% of birth weight variance while adolescent DNAm scores

explained roughly 3% of adolescent BMI variance.) The authors also compared the mothers' DNAm scores and mothers' BMI and found the value to be much higher at 10%. As well, the authors used cross-lagged models and Mendelian randomisation and found data to suggest that BMI is predictive of later DNAm. The group concluded BMI was more likely influencing the DNAm scores rather than vice versa. While our methodology was not designed to evaluate direct effects and thus cannot provide comparable estimates of DNAm "effect size" per se, the low effect size and low predictive ability of DNAm seen in the study by Reed and colleagues demonstrates the importance of considering reverse causality. It also underlines how study design and asking what questions greatly influences what answers you obtain. This work is motivated by the idea that biological and clinical perspectives of CCD need better accommodation in the analysis of epigenetic data. The results of this work is just an initial foray into how we can design and analyse DNAm data to better answer questions regarding causality and direct or interactive effects on health.

4.6 Persistence

Other studies have demonstrated sustained MSP-related DNAm changes over time (Bauer *et al.*, 2016; de Vocht *et al.*, 2015; Joubert, Bonnie R. *et al.*, 2016). However, this is not a consistent finding among environmental epigenetic studies. Several studies have found cord DNAm at specific sites related to specific exposures, but found no evidence of these relations in DNAm levels measured in later life. For example, Alfano and colleagues used the same ARIES data and found no overlap between time points between their identified methylation marker for socioeconomic position (maternal education) and DNAm either using level of association (FDR-corrected $p < 0.05$) or through longitudinal analysis (i.e. average difference in methylation to yearly change in lowest/highest education level.)

The observed persistence of DNAm patterns over time is particularly striking considering that this means not only "surviving" physiologic shifts in DNAm during rapid phases of childhood development, but also be detectable through technical variability from the use of heterogeneous tissues. There is known low reproducibility between certain tissues. For instance, even within the same gene, (intron 3 *AHRR* gene,) Novakovic and colleagues observed tissue specificity for sites of differential DNAm related to MSP between infant cord, infant buccal epithelium and placenta (Novakovic *et al.*, 2014). In addition, previous literature has also shown tissue specificity for unreliable probes. By analysing tissue from three time points, we sought the

same pattern of methylation in DNA in two if not three different sources depending on the subject: cord blood, buffy coat and whole blood. Besides their tissue subtype specificity, these tissues also exhibit variability secondary to mixed cell heterogeneity.

4.7 Replication

Using the GenR cohort, we found evidence of replication for several patterns in predicting mid-childhood BMI. There are several critical aspects to this replication. First, the replication analysis simply used the genome-wide "template" of MSP related patterns from ARIES in that very little data harmonization was attempted. Besides employing the same batch removal technique and filtering of low variance probes, the entire preprocessing of DNAm data from quality control to normalization was the current standard protocol for that cohort. This preprocessing agnostic feature of the MSP-related patterns supports their biological role. This is in contrast to finding patterns only fitted for ARIES data or due to a artefact encountered by chance.

Second, the association models had a number of differences in control variables between GenR and ALSPAC. Even within the same study population, researchers often find altering control variables can importantly change model estimates. In fact, this many times leads to attenuation of previously "significant" relations. That the patterns persisted to relate to child growth despite these differences, even in an entirely independent cohort, supports a robust relation to biological mechanisms.

Third, despite the above-mentioned differences in handling data and modeling between the two cohorts, the relative ranking of importance of the patterns was roughly similar. If the patterns reflected a technical artefact or other data bias that happened to be present in both studies, one could potentially see some patterns emerge as relevant in both data sets. But it would be less likely for such bias to so similarly affect both cohorts to create the same relative pattern rankings. The two cohorts are from different countries, were recruited about two decades apart and under different research stewardship (Kooijman *et al.*, 2016b). Recent international epigenetic and genetic consortium have promoted more uniform raw data collection and extraction, processing and analysis. For instance, the Canadian led International Human Epigenome Consortium has a dedicated team, the Assay Standards Working Group, to standardize assay protocols and quality control. However, ARIES is among the first pioneers of epigenetic cohorts and pre-dates many of the new standards to which more modern cohorts

now conform. It is for such disparities between cohorts that many studies attribute the failure of replication in independent cohorts. For example, Alfano *et al.* were unable to replicate the parental socioeconomic position related DMP found in ARIES in the ENVIRONAGE birth cohort (Alfano *et al.*, 2019).

While replication of MSP-related differential methylation sites/regions is reported quite frequently (Joubert, Bonnie R. *et al.*, 2016; Knopik *et al.*, 2019; Lee, K. W. *et al.*, 2014), this is not the case for outcome related hits (for reviews, see Dall'Aglio *et al.*, 2018 and compare contradictory findings of Kupers *et al.*, 2015 and Valeri *et al.*, 2017). This may be due in part to the low reliability of certain CpGs leading to false negatives (Sugden *et al.*, 2019). This makes replication an even more challenging task even in the presence of "true" differences.

Fourth, the improved prediction error rate seen in GenR relative to ARIES may simply be due to the close temporal juxtaposition of the infant growth variable and BMI at age 5. However, we did not note relative improvement in performance with younger childhood outcomes in ARIES models, arguing against temporal proximity of outcome to explanatory variables (like infant growth rate) inflating predictive ability. Still looking at the prediction rate, let's assume infant growth most strongly drives this model performance metric. In that case, we may expect that the DNAm components created in one cohort and then validated in another cohort would worsen the prediction error, just like how a training model performs better than a testing model. The fact that DNAm components "survived" variable selection without model degradation in two cohorts despite the presence of a very strong variable such as infant growth may support its true relevance to the outcome.

4.8 Relevance to molecular function

In seeking molecular references for these DNAm components, we first localized the patterns in context of the 450K BeadChip and then in their predicted chromatin state. Then, we further mapped the patterns to one of the basic structural units of chromatin-based gene regulation, TADs and chromatin loops.

Despite the high prevalence of probes found in CpG rich, genic regions on the 450K chip, there was a consistent and dramatic shift away from these probes in the DNAm patterns. On the contrary, the DNAm patterns had a stronger predilection for non CpG rich sites compared to the 450K chip (Figure 72). These findings were consistent regardless of which cell references used

and across most DNAm patterns. We found these results interesting given the known statistical bias in EWAS for selecting probes in CpG rich regions (Bacalini *et al.*, 2015; Geeleher *et al.*, 2013; Silva-Martínez *et al.*, 2017). Therefore, we believe that the design of the 450K chip was not a strong influence on the emergence of the observed DNAm patterns, as also evidenced by the very different distribution of sites seen using both BOP and chromatin mark annotation. This is an important sanity check that supports the notion that these DNAm patterns are purposefully created individual elements rather than a statistical "trimming" of sites that would follow the microarray distribution.

Moreover, we note that CpG poor areas of the genome, especially those in non-genic regions, are believed to regulate chromatin structure (Xue *et al.*, 2020). Thus, the localization of our DNAm patterns may lend support to our hypothesis that clinically-informed pattern finding can detect changes in chromatin architecture. This is in contrast to traditional EWAS where candidate sites are largely focused on sites/regions in CpG dense, genic areas. We speculate that alterations in chromatin architecture may have pervasive and perhaps persistent effects on cellular phenotype. In comparison, the context dependent nature of epigenetic regulation renders a significant degree of uncertainty regarding the effect of methylation differences at a single site or region on a given nearby gene. Re-looking at previously observed associations between methylation and nearby transcription levels, researchers increasingly question whether these are causal versus coincidental relations given the high inconsistency seen between studies of the same genic areas (see (Xu, H. *et al.*, 2020) for a recent review.)

The enrichment in enhancer and bivalent domains is particularly interesting as the 450K chip has a low representation of such marks (Figure 70). As well, enhancer states cover on average approximately 3% of each reference epigenome; bivalent states even less than that. In comparison, about 68% of reference epigenomes are covered by a quiescent state (Roadmap *et al.*, 2015). In a multiomic study of smoking versus non-smoking mothers and their children, Bauer *et al.* found that MSP-related differential methylation in enhancer regions were about twice more frequent than in promotors (Bauer *et al.*, 2016) using WGBS data, (which unlike 450K chip data, is genome comprehensive and therefore not biased to manufacturer curated CpG sites.) Interestingly, most of the enhancers identified were intragenic and targeted distal genes. Moreover, the same study found that enhancer methylation is more often functionally translated in terms of RNA expression than that in promoters or non-regulatory elements (Bauer *et al.*, 2016). Bivalent loci are so-called as they are "poised" between active transcription and stable repression. Typically, they consist of large regions of H3 lysine 27 methylation harboring

smaller regions of H3 lysine 4 methylation. Bivalent marks are the subject of intense research due to their known involvement in the dramatic change from pluripotent to lineage specific cells (Harikumar & Meshorer, 2015) and post-differentiation cellular plasticity (Tritschler, Theis, Lickert, & Bottcher, 2017). Data also suggest that these regions have clinical implications, such as in cancer (Avgustinova & Benitah, 2016) and neurodegenerative (Yang, X. W., 2016) disease. Chromosomal domains that may be poised to switch to and from "active" and "inactive" states would be ideal candidates for exposure-sensitive genomic targets (Prickaerts *et al.*, 2016). Referring again to the study above by Bauer and colleagues, the authors described chromatin state transitions from birth to around age four by using ChIP-seq to map genome wide histone modifications. They found children exposed to MSP had significantly more transitions into bivalent states than those without this exposure. This is in line with our finding of enrichment of bivalent domains as a clinically relevant DNAm pattern related to MSP.

Similarly, the low representation of quiescent marks is notable given it is one of the most frequent marks on the chip and in reference genomes. As well, looking at the maternal methylome, smoking leads to increased chromatin state transitions to the quiescent state from the time of delivery to 4 years later compared to non-smoking mothers, which is the opposite of what is observed in their offspring (Bauer *et al.*, 2016). A chromatin state change over time relative to exposures suggests an environment sensitive epigenetic mechanism at play. Thus, this finding in mothers may suggest that the significantly low representation of quiescent marks seen in the DNAm patterns is not a general response to smoke exposure but may be unique to MSP exposure in the fetus. Among the EWAS investigating MSP exposure, many of the most reproducible candidate sites target general pathways which could be suspicious for a non-specific response to tobacco exposure i.e. may not be directly related to disease susceptibility. For instance, *CYP1A1* and *AHRR* are both involved in the xenobiotic metabolism pathways of numerous toxins found in cigarette smoke. The altered epigenetic regulation of genes involved in such pathways is unsurprising with MSP exposure – however, it does not necessarily signify a causal influence on pathophysiologic processes for the fetus. As well, the altered methylation of this gene is found in the placenta (Suter, M. *et al.*, 2010) and in mothers (Bauer *et al.*, 2016). Findings that demonstrate a distinction between the effects of MSP on mothers compared to their infants may shed light on more suitable candidates for the study of DOHaD in the MSP model.

Advances in 3D mapping of the genome herald a new frontier in defining the mechanics of omic data and their function in disease pathology (Hu & Tee, 2017; Tak & Farnham, 2015). We look

again at the enrichment in enhancer marks, now from the point of view of potential functional relevance. Enhancers are non-coding epigenetic regulatory elements (~ 50–1500 base pairs) that form long-range DNA loop structures that spatially and temporally regulate gene expression by engaging in numerous physical interactions with gene promoters, both proximal and distal to its sequence location. Such chromatin loop structures are believed to be among the basic regulatory units that form the foundation upon which more intricate chromatin architecture is built. Given its expansive regulatory role, it is therefore unsurprising that numerous and diverse diseases associated with genetic variants are enriched in enhancer regions. These include autoimmune disorders, diabetes, cancers and neurodegenerative diseases (Fu, Tessneer, Li, & Gaffney, 2018; Hu & Tee, 2017). Moreover, enhancers are also sensitive to diverse environmental signals (Hah *et al.*, 2015; Klengel *et al.*, 2013; Lu, McComish, Burdon, Taylor, & Korner, 2019). The exciting and rising tide of research investigating 3D genomic features such as enhancers may hold the key to unlocking the functional relevance of disease-associated genetic variants and phenotype (Fu *et al.*, 2018; Tak & Farnham, 2015; Weaver *et al.*, 2017).

Given the enrichment in chromatin marks like enhancers that physically interact with promoters to direct chromatin conformation, we were curious how these patterns mapped to 3-D chromatin structure. As discussed in Section 2.6.2: Mapping methylation to chromatin topology, Wu and colleagues recently created the first 450K BeadChip data set annotating areas of promoter anchored chromatin interactions (PAIs). We found that most identified DNAm patterns were enriched in PAIs, but also chromatin loop structures (Figure 74).

Recently, Czamara and colleagues used data from 4 large pregnancy cohorts to map 450K chip sites whose cord methylation levels correlate with the largest differences in gene expression, specific prenatal environmental factors (i.e. pregnancy characteristics) and the closest SNP genotype (Czamara *et al.*, 2019). In this way, the authors offer a map of gene expression correlated with variably methylated regions according to gene and environment influences. They further filtered this list to those CpG sites that overlap with SNPs with putatively functional consequences on regulatory marks (using DeepSEA variants that consider likelihood of the presence of histone marks, DNase hypersensitive regions or TF binding for a given 1 kb sequence.) Using the same 15-core state chromatin annotation as in our analysis, they found their identified CpG sites were enriched in heterochromatin, repressed Polycomb, TSS and enhancer marks. Besides their finding regarding heterochromatin, our results overlap.

Another recent study used over 2000 database and publication sources to examine the correlation between gene expression and methylation across different tissues and ages (Wang, K., 2019). Thousands of significantly correlated CpG-gene expression pairs were found in each tissue using FDR correction, however, only 37 sites were consistently ranked highly in at least three tissues. This small number may reflect the tissue specific nature of CpG methylation and impact on gene regulation. We speculate this also alludes to the limitation of sequence-based annotation of methylation in providing indication of local chromatin architecture, another caution again single site-based analysis. Interestingly, the significant CpG-gene expression pairs demonstrated a commonality regardless of tissue in their enrichment for 15-core state chromatin annotation. This study found that pairs with negative methylation-expression correlation were enriched in active regions containing active transcription and enhancer marks while positively correlated ones were enriched in bivalent enhancer and repressed polycomb marks. This enrichment overlaps with the enrichment found in our DNAm patterns and supports the molecular viability of these patterns in modifying cellular phenotype.

Because our goal was to uncover patterns of DNAm differences related to the exposure composite that delineate changes to chromatin structure leading to stable MSP related changes in cell fate, the proximal relation of the DNAm component loci to local gene annotations is less relevant. For example, Figure 75 shows results from gene set analysis of Component 11 sites. Not surprisingly, few categories were significantly enriched and most were related to generic cell functions.

We wonder at our finding of significant depletion in the DNAm components of quiescent and heterochromatin marks. This is in keeping with the study by (Bauer *et al.*, 2016). These inactive states, especially heterochromatin, are typically associated with constitutive gene silencing such that these areas are laden with repressive marks and are highly condensed. They also tend to be physically separated from transcriptional factories. So far, inactive chromatin is less frequently reported to be involved in dynamic, environment sensitive changes compared to active chromatin. However, it is a thriving area of interest in cancer and toxicology, where dramatic changes in the distribution of such areas of relative safety from DNA damage can pre-dispose a cell to apoptosis or aberrant replication (Liu, W. & Irudayaraj, 2020; Williamson, Zhu, & Yuan, 2018). It is reasonable that MSP is an overall chromatin activating event given that DNA damage is one of the hallmarks of tobacco toxicity.

We did attempt to compare mapping of our components with the candidate sites identified in previous studies. For instance, using the 28 MSP differentially methylated CpGs found in cord blood by (Richmond *et al.*, 2015) in ARIES, we found 5 sites overlapped with PAI sites (permutation testing p = 0.001.) None overlapped with loop structures. However, is analyzing all 28 sites at once correct? We struggled with how to group the candidate sites: by region or chromosome? By CpG density or proximity to genes? Do they represent a single process or reflect multiple processes? Because of our *a priori* hypothesis that we could "bait" the extraction of clinically-relevant DNAm patterns with specific vulnerability profiles, the relation of CpG sites identified in the DNAm patterns can be subjected to logical interrogation of function. Without this *a priori* context, an unguided mining of statistical associations may render the discovery of the clinical and functional relevance of isolated sites or regions elusive.

At this point, there is a paucity of techniques to combine multi-omic data to estimate functional relevance. Giambartolomei *et al.* developed the method multiple-trait-colocalization (moloc) to colocalize omic data and genetic variants within a user defined genomic region to draw statistical estimates of causality (Giambartolomei *et al.*, 2018). The moloc method uses summary level data of the association of a given trait with SNPs from public databases, (i.e. effect size estimates and standard errors from quantitative trait locus (QTL) mapping.) These traits could be DNAm, gene expression, or phenotype. The moloc algorithm estimates the probability of a relation between the posited causal trait and one or more traits within the same region. In this way, the authors used brain tissue data to find 52 novel candidate genes that were related to three traits: schizophrenia diagnosis and DNA methylation and RNA levels (posterior probability > 0.8). This could be interpreted as finding genetic variants that influence schizophrenia phenotype through methylation. However, this relation could also be due to pleiotropy where the variant influenced each trait independently and there is no functional relation between the traits. The moloc method has been applied in various contexts, including making causal inferences regarding DNAm-mediated effects of MSP on complex traits in the ARIES cohort. A recent study employed Mendelian randomization to link MSP-related differential methylation at 412 CpG sites (FDR<0.01) to 643 complex traits using GWAS data (Richardson *et al.*, 2019). Of the 22 CpGs-trait associations that passed multiple testing correction ($p < 1.89 \times 10^{-7}$), they discovered two gene regions where the same variant linked CpG methylation, gene expression and lung function measures using moloc. The authors further showed using Mendelian randomization that the direction of association was likely MSP →methylation→lung function. As well, they observed a relation between DNAm and lung

function at age 8 and 15, however, the relation was stronger at age 15. The authors interpreted this to mean reverse causation i.e. that possible adolescent initiation of smoking or SHS was responsible for this observation.

Another method using QTL data to improve functional mapping is EpiXcan (Zhang, W. *et al.*, 2019). This method seeks to improve transcriptomic imputation methods that model the combined effect of multiple SNPs in proximity to the TSS on local transcription. EpiXcan leverages chromatin state annotation data by converting the association between state and SNPs into penalty factors to inform the gene expression imputation algorithm. This improvement leads to better annotation of genetic risk variants that are in noncoding regions. As well, the authors found improved tissue- and disease-specific relevance of identified candidate genes compared to previous transcriptomic imputation methods.

The moloc method assumes the causal genetic variant lies within the genomic region with the other traits and that at most one variant is responsible for each trait in the region i.e. only pairwise relations are considered. As well, the use of QTL data as the basis of association and using distance-based analysis units introduces two issues: 1) gene-centric bias and 2) the assumption that for a trait to be relevant, it should affect the traits nearby, (e.g. a relevant CpG site should impact the transcription of proximal genes.) This may limit the ability to detect multi-way interactions and interactions that may fall outside the user-defined genomic window, (such as due to a chromatin looping structure.) While QTL mapping based methods may limit context-based relevance of methylation on chromatin architecture, these techniques all leverage publically available data. This is a major advantage given the limitations of cost and feasibility in amassing enough data in prospective human cohorts. Our pattern finding technique fits this need to optimize use of global efforts to characterize cross-tissue and cross-population omic profiles if we are to find functional and thus medically useful links.

Pre-existing databases are also being leveraged to elucidate the relation between traits and toxins or pharmaceuticals. For example, this can facilitate high volume scanning of thousands of environmental pollutants for potential links to a trait of interest. The Comparative Toxicogenomics Database (CTD) has collated over a million relations between chemicals and 33 biologic substrates, (e.g. mRNA, proteins, etc.) As a recent example, Smith and colleagues paired chemical and mRNA data from CTD to overlap a candidate set of differentially expressed genes during sensitive periods of neuroplasticity using enrichment analysis (Smith, M. R. *et al.*, 2020). They identified 50 chemicals that consisted of both known neurotoxins as well as novel

candidates. Given the high prevalence of childhood neurodevelopmental disorders, the authors argue this type of analysis could systematically prevent far-reaching harms on a societal level by drawing attention to potential toxins, especially commonly used chemicals like antimicrobials, (e.g. applied directly or through agricultural applications.) Similar to CTD, there are databases generated by computational pharmacology that integrate information from electronic health records, clinical trial data, and drug interaction and adverse effect reporting with biologic data such as gene expression, protein networks and genes. One such repository is the Connectivity Map (Lamb, J. *et al.*, 2006), a library of 1.5M gene expression profiles linked to about 8000 pharmacologic compounds. Initially, such data was used to better predict drug-related adverse effects. However, these databases are increasingly mined for drug repurposing, a field aimed at finding new uses for existing drugs. This relatively new area has garnered much excitement given its potential to provide rapid benefit to society, financial benefit to the pharmaceutical industry and benefit to the development of treatments for rare diseases that are by nature hampered by low numbers of patients (Hodos, Kidd, Shameer, Readhead, & Dudley, 2016) . For instance, the authors of the EpiXcan study discovered 43 gene-trait links (Zhang, W. *et al.*, 2019) in their database scan. They curated 1309 compounds from the Connectivity Map library which they considered “capable of perturbing the expression” of those 43 candidates. Using enrichment analysis, they found a number of compounds with plausible links to the trait in question. For instance, the “Coronary Artery Disease” trait was enriched for drug targets that have genetic association with heart disease, hypercholesterolemia, abdominal obesity, and myocardial infarction. Zhang and colleagues went on to overlap the identified trait-compound candidates with compounds likely to have a disease modifying effect on the trait. They found several enrichments, for example they found an overlap between compounds for coronary artery disease that are considered disease modifying for childhood obesity. They authors believe this supports the use of their pipeline for predicting drug repurposing and overall the utility of integrating gene-based data to understand the biologic networks underpinning disease processes to systematically guide therapeutic discovery.

In the end, genes are the archetypal tether that links multiple omic layers. We are wary of gene-centric views of data because we simply “don’t know what we don’t know” and could unknowingly neglect a whole space of biologically important information. We look to the exciting field of mapping 1D epigenomic data to surmise 3D chromatin structure (Xu, H. *et al.*, 2020). As well, there is also the ever growing are of code biology to shed more light on the functional implications of molecular mapping in CCD. Code biology is dedicated to delineate the relation

between two “worlds of organic molecules” through a set of rules (for a review, see (Barbieri, 2003). Though arguably the most famous success of code biology is the genetic code, this field has enthusiastically tackled the neural code, the sugar code, the cytoskeleton codes, and the splicing codes, as examples. We take the work of Prakash *et al.* in the histone code to illustrate the 3 steps in this methodology (Prakash & Fournier, 2018): “Here code is identified, where an input system (histone modifications) is translated into an output system (chromatin states) via adaptors (epigenetic regulators or TFs). Such a code has distinct importance in gene regulation and consequently for the cellular phenotype.” It is hoped that this formal methodology will soon crack the “epigenetic code” and advance the efforts to assess and modify patient-specific disease risk in CCDs by judiciously mining the expanding universe of omic information.

4.9 Impact

Patient-specific recognition of disease susceptibility is the key to resource-efficient prevention. By using individual biologic susceptibility obtained before overt cellular damage in childhood, one can deliver targeted intervention during the most sensitive and formative periods of human development. This “front loaded” investment narrows the chance of organ dysfunction and spread to other organs, thus offering optimal and compounded health gains population wide.

This work links phenotypic mapping between clinical and epigenetic features of children from birth to mid and then late childhood. This mapping localized DNAm patterns that may be relevant to both physical and mental health outcomes. We show that the clinical composite relates to outcomes in later childhood in a similar manner to the DNAm maps. As well, we demonstrate that these DNAm maps can be transferred to an independent cohort. Last, these DNAm patterns map to areas of chromatin structure previously implicated in environment-sensitive regulation. We believe this supports proof of concept that DNAm maps may replace the use of smoking-related risk data and localize clinically-relevant molecular pathways underlying the origins of complex traits. We suggest that the approach to estimate the importance of DNAm patterns taken in this study should be seen as complementary to the conventional statistical techniques used in EWAS in CCD.

Machine based learning is becoming the norm rather than the exception in high dimensional omic data analysis (see Rauschert *et al.* (2020) as a very recent example.) This work also

employs machine learning but we do not at all suggest this is or will be the panacea to the translation gap between epigenetics and CCD research. Our study is novel in that we avoid exposing the selection of DNAm signals to any single MSP-related variable and/or outcome variable. In doing so, we aim to attenuate exposure to the inherent assumptions, biases and errors of any single measurement tool. Instead, our goal with the data available is to best describe the clinical and molecular constellation of MSP-related events that relate to complex traits. Among conventional and machine learning techniques, we selected the method that best matched this goal with the data currently available in the discovery cohort. We are hopeful that as cohort databases and analytic methods evolve, so will approaches to finding clinically relevant molecular targets. Most importantly, we call for new modes of evaluating biomarker utility beyond relative risk or dose-response effects. Such metrics require individual-independent categorization that sacrifices the clinical information embedded in individual-context measures. In the study of CCDs, we fear this risks putting the cart before the horse.

The ethical dilemmas of the post-genomic era pose a fascinating and at times perplexing realm of research (see Dupras, Joly, & Rial-Sebbag (2020) for a recent review.) Cautionary perspectives evoke sombre images of epigenetic markers being wielded to stigmatize certain populations. For instance, this work could fuel accusations that smoking mothers are harming their children as if adversity is due to free and voluntary choices. In another example, the possibility that molecular vulnerability can pass from one generation may lead to insinuations of hopelessness for the future generation of that group. While our hope is that epigenetics can reduce health disparities, the just use of scientific research requires constant vigilance and care.

4.10 Limitations

This work suggests a potential MSP associated epigenetic network rooted early in life that relates to later physical and cognitive differences in childhood. However, it cannot differentiate between whether MSP-related DNAm patterns are causally related to later outcomes, the consequence of outcome or merely coincidental.

The thrust of our work is to improve delineation of a given individual's multifactorial vulnerability to adverse contexts. This vulnerability concept begs a discussion of the contrast between quantitative or biochemical measures of smoking versus self-report. The qualitative aspect of the situation of a mother who would report smoking and her consistency in reporting this at different times in her pregnancy is clearly distinct from attempts to quantify fetal exposure. The

correlation between maternal self-report and biochemical markers of smoking and their consequent net tobacco-related chemical exposure of the fetus is a field unto itself. Actual clinically-relevant fetal exposure requires consideration of a plethora of factors, including tobacco product and delivery choice, gestational developmental stage, duration, concurrent exposure to prescribed or illicit agents as well as xenobiotic interactions including pharmacogenetics involving the mother, placenta and fetus (for examples, see (Agrawal *et al.*, 2010; Knopik *et al.*, 2019; Marceau *et al.*, 2016; Werler, Pober, & Holmes, 1985). The landscape of this research is changing rapidly. For example, the increased use of vaping and the legalisation of marijuana in Canada have had a major impact on women of child-bearing age in recent years and it is known smoking is associated with other substance use (Rodriguez & Smith, 2019). One limitation of our study is that we have used self-reported data as the single source of information regarding smoking during pregnancy. Various studies comparing self-reported smoking with biochemical measures provides reassurance for using this measure in terms of its clinical correlation with offspring outcomes (Gorber *et al.*, 2009; Guerrero-Preston *et al.*, 2010; Keskitalo *et al.*, 2009). At the same time, other studies estimate that around 20% of mothers under-report smoking in pregnancy (Dietz *et al.*, 2011; Shipton *et al.*, 2009). We are interested by the misclassification correction performed by (Valeri *et al.*, 2017) in the MoBa cohort which resulted in a contradiction of the mediation effect of MSP on birth weight by DNAm found by (Kupers *et al.*, 2015). MoBa had the benefit of both self-report and cotinine levels to estimate MSP. However, that is not to say that biochemical markers are “better”. For example, markers like cotinine measured at only a few intervals of pregnancy do not necessarily translate 100% to fetal MSP effects. We speculate that a more nuanced composite of MSP would integrate both these markers as well as various clinical and genetic data. For instance, one could attempt to characterize the “dose”, the “mediator” (i.e. the placenta) and the “recipient” to better estimate clinical relevance. First, the “dose” of exposure on the fetus may include SNPs of xenobiotic enzymes of mothers and infant, as well various maternal SNPs previously used in IV analysis as a proxy for smoking quantity and ease of quitting (for example, maternal rs1051730 was used in (Brand *et al.*, 2019).) This genetic data was available in ARIES and it is becoming common for epigenetic cohorts to also collect genotype data simultaneously given the relative low costs and high-throughput of SNP microarray chips. Second, indication of placental effects (e.g. placental thickness and umbilical blood flow on Doppler based on sonography or treatment of maternal hypertension) may be accessible through health databases. However, such data may only be available in resource-rich urban settings where such prenatal management is part of routine standard of care. Last, the impact on the fetus can be extracted

from routine prenatal visits in most developed countries including gestational weight gain in different trimesters and basic fetal ultrasound biometry like estimated fetal weight, femur length, and head and abdominal circumference. This work's vulnerability composite suggests that transgenerational, historical and environmental smoking variables and birth weight variables are relevant. Thus, including such variables along with the aforementioned would be an important exploration of a richer, multiclass mapping of MSP vulnerability. With such a composite, self-reported and/or biologically estimated MSP would be but one of many factors forming the child's multidimensional profile and thus MSP associated error would be attenuated.

Future work should also examine the interaction of concurrent smoke exposure, for example through SHS, third hand smoke (exposure to residue left on surfaces exposed to smoke) and adolescent smoking. A previous ARIES study found that differential DNAm by MSP in cord blood at 5 CpGs persisted in adolescence whether the data included or excluded adolescents exposed to personal smoking or SHS (Richmond *et al.*, 2015). However, the relevance of this finding on physical or cognitive outcomes is unknown and warrants investigation. SHS is less often directly studied but more often a control variable. In our work, the vulnerability composites Dimensions 3 and 5 were the only dimensions to include relations to smoking in non-maternal line relationships. These dimensions made interesting contributions to our results, particularly in considering potential resilience factors at play. As well, Dimension 5 is among the most represented profiles in cord DNAm. Our work strongly supports further direct study of all sources of smoking to clarify this phenomenon. The entity of third-hand smoke (THS) is particularly relevant to childhood exposures and has received very little research attention in epigenetics thus far. Children are particularly susceptible to THS compared to adults given their oromotor developmental stage that increases their proximity to high exposures surfaces like carpet and walls and the transfer of residue from surfaces, toys and hands to their mouths and eyes.

The reader may recall we struggled to include several variables that have direct relation to placental function and infant outcomes. These included factors like gestational diabetes, hypertension and weight gain. There are several possible reasons. First, these variables suffered the most missing data. This would very likely affect composite construction. Second, it may be that these variables together truly do not form reliable latent constructs. Third, this may have been a specific limitation of our factor analysis method. It could possibly affect the model algorithm and estimate convergence when categorical variables, particularly those with merely 2 or 3 levels, outnumber continuous ones by such a large margin. The methodologies for multi-

class data analysis is evolving with better accommodation of data types especially nominal versus ordinal non-continuous variables (see the PLS-correspondence analysis-regression method proposed by (Beaton *et al.*, 2019) as an example.) We also look forward to more sophisticated developments in this area with regards to missing data given the strong influences of non-random missingness and selection bias in MSP research (Fang *et al.*, 2010; Fertig, 2010; Knopik, Valerie S., 2009; Valeri *et al.*, 2017). It is clear that deeper consideration of the properties of such vulnerability-related variables and multiclass multivariate data analysis is needed in future work.

Another limitation is that we have no confirmatory or functional assessment of the DNAm patterns, such as targeted pyrosequencing, profiling of TF binding, 3C of chromatin structure or gene expression. These data are unavailable in ARIES. We acknowledge that the DNAm patterns identified in this work are purely due to statistical estimations. One possible reason for the observed relations is that these dimensions coincide with another MSP-unrelated variable, (e.g. a covariate such as batch,) that also has a molecular signal. This “covariate” signal becomes the easiest manner for DNAm data to explain the variance of the MSP composite. We used state-of-the-art techniques to remove variability due to known batch effects and ARIES employs a purpose built laboratory information management system as well as a semi-random approach in plating the 450K BeadChip to attenuate batch effects. However, it remains that other known and unknown technical artifacts may overlap with the MSP composite.

There are a number of animal models demonstrating clear methylation-mediated gene expression differences between nicotine exposure groups, (for recent examples, see (Buck *et al.*, 2019; Zeng *et al.*, 2020).) Human studies are far more limited. We found only two studies that attempted to complete the circle from discovery to functional relevance. Bauer and colleagues employed RNA-seq, WGBS and ChIP-seq to determine genome wide gene expression, DNAm and histone modifications in cord blood in children of smoking versus non-smoking mothers (Bauer *et al.*, 2016). Comparing smoking versus non-smoking in pregnancy groups, they discovered 8409 DMRs in cord blood using a moderated t-statistic ($p > 0.1$) and permutation analysis. This group also found some DMRs were genotype associated. After removing these, 1404 DMRs remained. These DMRs were associated with significant chromatin state changes from birth to age four when comparing non-smoking to smoking mothers.

The authors also followed up the observation that the *JNK* gene appeared to have both genetic and methylation-mediated differential expression. They noted a relation between *JNK2* enhancer demethylation in both cord and age 4 blood with increased risk of wheeze after age 4. They used an *in vitro* model with peripheral blood mononuclear cells to confirm enhancer demethylation upon exposure to cigarette extract. They also used a *JNK2*-/- mice to demonstrate less airway inflammation compared to wild type.

They found very few differentially expressed genes in children by smoking category, (<10 genes identified using multiple testing threshold of 10% FDR, BH correction.) The paucity of differentially expressed genes led the authors to instead target the downstream pathways targeted by DMRs. The authors did not directly state how many of these pathways demonstrated differential expression among gene members between the smoking and non-smoking group. However, they pointed out that the Wingless-Type MMTV Integration Site Family (*WNT*) pathway gene members did distinguish the two groups of children.

This study found very low correlation (ranging between 5-10%) between DMR methylation and gene expression. This study also found that the correlation between RNA expression of target genes and differential DNAm sites increased over time, suggesting that DNAm changes precede functional change such as transcription. However, only 16 children had all three of the above omic data extracted.

More recently, Vives-Usano and colleagues used data from the large Human Early-life Exposome project that represents the collaboration of six population based birth cohorts across Europe (Vives-Usano *et al.*, 2020). Using DNA methylation and gene and miRNA transcription, plasma proteins, and sera and urinary metabolites data from up to 1203 children, this group found 41 differentially methylated CpGs that localized to 18 unique loci based on a 5% FDR comparing 2 groups: any or sustained maternal smoking in pregnancy versus non-smoking. All loci have been previously reported but one (*Formin 1* gene). Of these, only differential DNAm at five loci were related to proximal gene expression. Interestingly, this association was not with the "closest" gene by base pair distance except in one case. Despite finding a link between MSP to DNAm, and then DNAm to gene expression, this group found no link between MSP and child serum metabolites or blood gene/miRNA expression. The only further association with MSP found was with urinary alanine and lactate levels in childhood. The authors did not report any testing for association between methylation and metabolites. Despite having by far the

largest pool of children of any MSP-related multi-omic study, the authors suspect that the lack of overall novel findings was due to inadequate power.

This brings us to the point that cost is a major barrier to obtaining biological data. As well, the study by Bauer *et al.* (2016) illustrates the importance of follow-up over time in order to unearth the molecular seeds that underlie pathogenesis. This cost of adequate sample size in longitudinal over cross sectional data is multiplicative in the discovery stage, rendering confirmation of biologic functional relevance of possible candidates out of reach of most research budgets. Another major barrier to collection of biological data is feasibility. For this reason, numerous studies use methylation patterns in blood as the surrogate for the target tissue like brain, adipose, liver, pancreatic and muscle tissues (for examples, see Bansal *et al*. (2017), Shanthikumar *et al*. (2020), Forest *et al*. (2018), Keller *et al*. (2017), Yuen *et al*. (2011) and Horvath *et al*. (2012)). As in these studies and discussed in Section 2.1.2.2(a), researchers note that the greatest source of methylation variation in human EWAS is cell type (Horvath *et al*., 2012; Roadmap *et al*., 2015). Given the tissue specificity of DNAm, it begs what, if any, relation our findings have to cellular poise of the target organ? Studies employing animal models and post-mortem human samples demonstrate some concordance between blood and target tissues. This is unsurprising in disorders involving the immune system like oncologic, allergic, and autoimmune (the latter which include certain forms of diabetes, rheumatoid arthritis and inflammatory bowel disease,) diseases (for examples, see Reinius *et al*., 2012). As well, studies show blood-brain correlation of differential DNAm for psychiatric outcomes (Fuchikami *et al*., 2011; Melas *et al*., 2012; Unternaehrer *et al*., 2012). Together, these disease groups account for a prominent majority of CCDs globally. Thus, whether as a direct actor, mediator or bystander of various disease processes involving diverse organ systems, blood is a major organ that taps into nearly every part of the human body. This is unlike buccal or sperm cells, for which DNAm data has also been non-invasively collected but with much more limited theoretic or experimental connections to CCDs. There is also the growing potential of DNAm data obtained from cell-free DNA for various diseases (Lehmann-Werman *et al.,* 2016). Studies have isolated cell-free DNA from non-invasive sampling of saliva, sputum, urine, stool and seminal fluid, for example. However, at this time, blood and specifically circulating immune cells, remains one of the most promising human tissues to study CCD-related shifts in epigenetic poise. That said, tissue specificity remains an important consideration when evaluating the clinical relevance of blood-based DNAm biomarkers.

Fortunately, unlike biological data with its inherent cost and feasibility restraints, clinical data availability is surging. We are in the age of electronic medical records and high consumer interest in self-monitoring of health status. As well, the sophistication of behaviour analysis, especially through the internet and hand held devices, has become both increasingly effortless and invasive in terms of privacy. While the ethics of this is beyond our scope, it is a fact that there is no end of novel and rich clinical data that extends far beyond standardized questionnaires and clinical checkpoints. The ability to mine more comprehensive and continuous data may rapidly overcome the errors and assumptions that plague data that is sparse in detail and time points.

A recurring issue throughout this work is the limited coverage the 450K chip. Even the newer chip created by the same company, the EPIC BeadChip, covers under 3% of approximately 28 million CpG sites in the human genome (Jiao *et al.*, 2018; Solomon *et al.*, 2018; Wang, X. M. *et al.*, 2019; Zhou, W. *et al.*, 2017). Moreover, non-genic CpG remain relatively under-represented though this coverage is improved relative to the 450K chip. Using WGBS, Bauer and colleagues found that less than 5% of MSP-related DMRs were covered by the 450K chip. Given the current track record, it is fair to question whether pursuing CCD research with this technology will ever be clinically fruitful. The study of CCDs requires population based data and enough statistical power. At this time, microarrays are the only viable option. Gene expression related to tobacco exposure fades with time, while DNAm differences studies with microarrays are still notable up to 40 years later. Chromatin conformation is simply too costly and low throughput to be useful for population-based exploratory studies. DNA methylation is stable yet environment sensitive. It is unique among biomarkers. This study in many ways is looking at how we can most efficiently and effectively extract useful information from limited resources. However, it is unlikely at this state of knowledge that DNAm arrays alone will provide enough information to guide clinical action.

During the writing of this thesis, the next generation of Illumina DNAm BeadChips was released, the EPIC BeadChip (Jiao *et al.*, 2018; Solomon *et al.*, 2018; Wang, X. M. *et al.*, 2019; Zhou, W. *et al.*, 2017). This chip covers >90% of sites covered by the 450K chip, but has several technical differences including a highly different ratio of Type I:Type II probes. Illumina no longer supports the 450K chip so opportunities for replication in future cohorts using the same tool will increasingly diminish. The 450K BeadChip covered about 98% of genes but only 1.7% of CpGs present in the human genome. This means the DNAm patterns represent an incomplete shadow of methylation topography across the genome. The EPIC array assays

nearly double the number of CpGs (>850,000). A wider appraisal of methylation, especially in non-genic regions, using the EPIC BeadChip may reveal more functionally and clinically relevant relations.

Methods are constantly being developed and refined to include various species of RNA and various sequence, expression, or functional data, (e.g. proteomics, microbiome, lipomics, etc.,) to better characterize the human molecular profile. We look forward to techniques that can accommodate these layers of high dimensional data along a lower dimensional individual axis and then a sparse time axis, such as in non-negative tensor factorization. Such multi-axis and multi-layer analysis is common in fields like functional neurophysiology. We highly anticipate the adaption of such techniques in omic research to provide richer and more dynamic individual mapping.

We also must be wary of making generalizations from potentially non-representative samples. The ARIES cohort contained about half as many mothers who reported smoking than in the whole ALSPAC, (14.3% versus 30.2%.) An important next step for this work is the "re-discovery" of DNAm patterns in a population representative dataset rather than convenience sample.

Studies have shown poor reproducibility between platforms of certain individual CpG methylation of the same sample. This has been found in both solid and blood tissues (Solomon *et al.*, 2018) and regardless of normalisation procedures (Cheung, Burgers, Young, Cockell, & Reynard, 2019). For instance, Cheung and colleagues found that about half of CpG sites showed low correlation ($r < 0.2$) between the 450K and EPIC arrays within the same cartilage samples. This phenomena is believed to be a technical artefact. It most severely affects low variance and extreme value (i.e. 0% or 100%) methylation sites, as well as those measured by Type 1 probes (the 450K uses two types of probes.) Low reproducibility will make future comparisons and meta-analyses with different assays more challenging. CpG sites with poor reproducibility vary between tissue types and therefore the specification of a common list of CpGs to exclude is unavailable. Researchers must exercise caution when investigating hits and perform confirmatory testing with methods such as targeted pyrosequencing and in independent samples (Solomon *et al.*, 2018).

RF regression is a powerful exploratory tool in data mining given its three-pronged ability to handle high-dimensional data, compute measures of importance and account for interactions between predictors. This matches well with the environment responsive epigenetic changes

thought to underlie CCD etiology. Our goal was not to optimize predictive power per se – but to explore contributors to the continuum of disease manifestation generated by genetic-epigenetic-environmental influences present in children. However, the interpretability of RF results has been criticized as more challenging than conventional regression analyses. For instance, while the latter can be relatively easily probed by plotting the association between variables on the basis of the regression coefficients, the former provides the importance of a predictor that contains its complex interaction structure with all other predictors included. Simulation studies have further shown that small interaction effects contribute to the overall predictive accuracy, but that current measures are unable to specifically isolate these effects (Wright, Ziegler, & König, 2016).

Therefore, the clinical outcome results of this study must be couched in context of the other variables entered into the model. Thus, exact replication is unexpected in different cohorts with different variables measured with varying error. However, we suggest that exact replication is neither necessary nor even proof of validation. Exact replication could still be the result of chance. We tender that reliable validation is finding compatible and consistent results even with the use of different methods and data.

We wish to clarify that our use of context-based detection is emphatically not to dodge identifying one or more specific factors associated with high levels of harm. It would be highly useful to find a clear “X leads to Y via Z” solution. For instance, individuals with a variant of the D-aminolevulinic acid dehydrogenase gene are highly susceptible to environmental lead exposure, a neurotoxin with potent effects on fetal and child development. In the US starting in the 1960’s, studies found that the prevalence of this variant was as high as 20% in some populations (Lustberg & Silbergeld, 2002). This led to a federal level drop in the lead safety threshold from 60 to 10 µg/dl. This move has been subsequently mirrored by numerous countries worldwide, such that severe pediatric lead poisoning in developed countries is now rare. This is one of the most dramatic examples of how molecular identification of a specific diathesis lead to a major health impact worldwide. Exposure to MSP and its related factors is a long-standing and complex global health problem – to remove its root causes is an unfortunate impossibility. However, elucidating why and how some individuals suffer more than others is feasible. Understanding the molecular basis of vulnerability would allow the development of targeted counter measures that could viably lead to change in public health and/or industry standards. While it is unlikely tobacco manufacturers would cease to employ key neuroactive substances such as nicotine in their products, pragmatic and enforceable standards can only

come from understanding if and how the origins of harm overlap with the business bottom line. We argue that at our current state of knowledge, characterization of molecular vulnerability to MSP is the key to unlocking resilience on a population level.

Chapter 5 Summary

As a science, medicine adopts many of its principal tenets from the scientific method originating from the 1600s. For example, there is a strong distinction in the use of language describing the patient history and physical exam compared to the diagnostic assessment. In its best form, the former should collate information with no attribution of cause or relevance to the patient complaint. On the other hand, the latter involves judgement and its associated assumptions. It also accounts for potential errors by including a differential diagnosis. The poor translation of epigenetics to CCD may relate to a neglect of basic scientific method and believing we can jump to the diagnosis without a careful and complete assessment of information. The main thrust of this work stems from acknowledging ignorance of the clinical and molecular implications of the origins of health risk. We attempt to minimize assumptions about what we do not know while optimizing the description of what we do know.

We ventured to sharpen the overlap of clinical data with genome-wide shifts in methylation to visualize the molecular architecture underlying various profiles of vulnerability to adversity. We mapped vulnerability using a composite measure that combined multiple views of clinical exposure while also offering the statistical advantages of a continuous measure. We used the MSP exposure composite to express not only *in utero* tobacco exposure, but also familial and individual susceptibility to MSP related effects. We purposefully selected mapping methods to minimize researcher assumptions. We avoided imposing if and what "levels" of MSP-related risk are considered clinically contributory, as well as what and where are relevant sites of DNAm. While these posed challenges to data integration and computational cost during the exploratory phase, in the end our pipeline is very feasible and would be easily up-scaled to higher throughput studies. This work could be translated to further develop understanding of pathologic mechanisms associated with other risks and hopefully to re-train research focus on the biological roots of disease rather than the after-effects of accumulating system damage.

It is implausible that epigenetics can offer a single, parsimonious explanation of all the interconnections between risks and disease. This work suggests epigenetic mechanisms are a

promising crossroads to probe the early origins of disease vulnerability– but there are boundless opportunities to improve the precision and comprehensiveness of clinico-biological mapping. In the ever-expanding universe of omic study and human disease, the new challenge is to converge and interpret these data as accurately as possible, humbly wary of bias and error. Our endeavour advances new practical and theoretical considerations to address the formidable problem of connecting multi-omic signals with complex traits – a critical obstacle to forging diagnostic and therapeutic translation.

References

Abdi, H., & Williams, L. J. (2013). Partial least squares methods: Partial least squares correlation and partial least square regression. *Methods in Molecular Biology (Clifton, N.J.), 930*, 549-579. doi:10.1007/978-1-62703-059-5_23 [doi]

Abdi, H. (2010). Partial least squares regression and projection on latent structure regression (PLS regression). *Wiley Interdisciplinary Reviews: Computational Statistics, 2*(1), 97-106.

Abrevaya, J. (2006). Estimating the effect of smoking on birth outcomes using a matched panel data approach. *Journal of Applied Econometrics, 21*(4), 489-519.

Adams, D., Altucci, L., Antonarakis, S. E., Ballesteros, J., Beck, S., Bird, A., . . . Caricasole, A. (2012). BLUEPRINT to decode the epigenetic signature written in blood. *Nature Biotechnology, 30*(3), 224.

Agha, G., Hajj, H., Rifas-Shiman, S. L., Just, A. C., Hivert, M., Burris, H. H., . . . DeMeo, D. L. (2016). Birth weight-for-gestational age is associated with DNA methylation at birth and in childhood. *Clinical Epigenetics, 8*(1), 118.

Agrawal, A., Scherrer, J. F., Grant, J. D., Sartor, C. E., Pergadia, M. L., Duncan, A. E., . . . Xian, H. (2010). The effects of maternal smoking during pregnancy on offspring outcomes. *Preventive Medicine, 50*(1-2), 13-18. doi:10.1016/j.ypmed.2009.12.009

Akhmetov, I., & Bubnov, R. V. (2015). Assessing value of innovative molecular diagnostic tests in the concept of predictive, preventive, and personalized medicine. *EPMA Journal, 6*(1), 19.

Alfano, R., Guida, F., Galobardes, B., Chadeau-Hyam, M., Delpierre, C., Ghantous, A., . . . Nawrot, T. S. (2019). Socioeconomic position during pregnancy and DNA methylation signatures at three stages across early-life: Epigenome-wide association studies in the ALSPAC birth cohort. *International Journal of Epidemiology, 48*(1), 30-44.

Anderson, D. R., Burnham, K. P., & Thompson, W. L. (2000). Null hypothesis testing: Problems, prevalence, and an alternative. *The Journal of Wildlife Management,* 912-923.

Annunziato, A. (2008). DNA packaging: Nucleosomes and chromatin. *Nature Education, 1*(26)

Aristizabal, M. J., Anreiter, I., Halldorsdottir, T., Odgers, C. L., McDade, T. W., Goldenberg, A., . . . Sokolowski, M. B. (2019). Biological embedding of experience: A primer on epigenetics. *Proceedings of the National Academy of Sciences,* 201820838.

Aryee, M. J., Jaffe, A. E., Corrada-Bravo, H., Ladd-Acosta, C., Feinberg, A. P., Hansen, K. D., & Irizarry, R. A. (2014). Minfi: A flexible and comprehensive bioconductor package for the analysis of infinium DNA methylation microarrays. *Bioinformatics, 30*(10), 1363-1369.

Aten, J. E., Fuller, T. F., Lusis, A. J., & Horvath, S. (2008). Using genetic markers to orient the edges in quantitative trait networks: The NEO software. *BMC Systems Biology, 2*, 34. doi:10.1186/1752-0509-2-34

Avgustinova, A., & Benitah, S. A. (2016). The epigenetics of tumour initiation: Cancer stem cells and their chromatin. *Current Opinion in Genetics & Development, 36*, 8-15. doi:10.1016/j.gde.2016.01.003 [doi]

Babenko, O., Kovalchuk, I., & Metz, G. A. (2015). Stress-induced perinatal and transgenerational epigenetic programming of brain development and mental health. *Neuroscience & Biobehavioral Reviews, 48*, 70-91.

Bacalini, M. G., Boattini, A., Gentilini, D., Giampieri, E., Pirazzini, C., Giuliani, C., . . . Garagnani, P. (2015). A meta-analysis on age-associated changes in blood DNA methylation: Results from an original analysis pipeline for infinium 450k data. *Aging, 7*(2), 97-109. doi:100718 [pii]

Bakulski, K. M., Feinberg, J. I., Andrews, S. V., Yang, J., Brown, S., L McKenney, S., . . . Fallin, M. D. (2016). DNA methylation of cord blood cell types: Applications for mixed cell birth studies. *Epigenetics, 11*(5), 354-362. doi:10.1080/15592294.2016.1161875

Bale, T. L., Baram, T. Z., Brown, A. S., Goldstein, J. M., Insel, T. R., McCarthy, M. M., . . . Nestler, E. J. (2010). Early-life programming and neurodevelopmental disorders. *Biological Psychiatry, 68*(4), 314-319. doi:10.1016/j.biopsych.2010.05.028

Bann, D., Wills, A., Cooper, R., Hardy, R., Aihie Sayer, A., Adams, J., . . . NSHD Scientific and Data Collection Team. (2014). Birth weight and growth from infancy to late adolescence in relation to fat and lean mass in early old age: Findings from the MRC national survey of health and development. *International Journal of Obesity (2005), 38*(1), 69-75. doi:10.1038/ijo.2013.115 [doi]

Barbieri, M. (2003). The organic codes-an introduction to semantic biology. *Genetics and Molecular Biology, 26*, 105-106.

Barker, D., & Osmond, C. (1995). The maternal and infant origins of cardiovascular disease. *Marmot M,*

Barouki, R., Gluckman, P. D., Grandjean, P., Hanson, M., & Heindel, J. J. (2012). Developmental origins of non-communicable disease: Implications for research and public health. *Environmental Health, 11*(1), 42.

Bauer, T., Trump, S., Ishaque, N., Thürmann, L., Gu, L., Bauer, M., . . . Mallm, J. (2016). Environment-induced epigenetic reprogramming in genomic regulatory elements in smoking mothers and their children. *Molecular Systems Biology, 12*(3)

Beaton, D., Saporta, G., & Abdi, H. (2019). A generalization of partial least squares regression and correspondence analysis for categorical and mixed data: An application with the ADNI data. *bioRxiv* 598888. doi:10.1101/598888

Belsky, J., & Pluess, M. (2009). Beyond diathesis stress: Differential susceptibility to environmental influences. *Psychol Bull, 135*(6), 885-908. doi:10.1037/a0017376

Bennett, G. G., Herring, S. J., Puleo, E., Stein, E. K., Emmons, K. M., & Gillman, M. W. (2010). Web-based weight loss in primary care: A randomized controlled trial. *Obesity, 18*(2), 308-313.

Biederman, J., Monuteaux, M. C., Faraone, S. V., & Mick, E. (2009). Parsing the associations between prenatal exposure to nicotine and offspring psychopathology in a nonreferred sample. *The Journal of Adolescent Health : Official Publication of the Society for Adolescent Medicine, 45*(2), 142-148. doi:10.1016/j.jadohealth.2008.12.003

Bird, A. (1992). The essentials of DNA methylation. *Cell, 70*(1), 5-8.

Bohr, N. (1913). I. on the constitution of atoms and molecules. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science, 26*(151), 1-25.

Bollen, K. A., Noble, M. D., & Adair, L. S. (2013). Are gestational age, birth weight, and birth length indicators of favorable fetal growth conditions? A structural equation analysis of filipino infants. *Statistics in Medicine, 32*(17), 2950-2961.

Bookman, E. B., McAllister, K., Gillanders, E., Wanke, K., Balshaw, D., Rutter, J., . . . for the NIH G x E Interplay Workshop participants. (2011). Gene-environment interplay in common complex diseases: Forging an integrative model-recommendations from an NIH workshop. *Genetic Epidemiology,* doi:10.1002/gepi.20571; 10.1002/gepi.20571

Booth, J. N., Tomporowski, P. D., Boyle, J., Ness, A. R., Joinson, C., Leary, S. D., & Reilly, J. J. (2014). Obesity impairs academic attainment in adolescence: Findings from ALSPAC, a UK cohort. *International Journal of Obesity, 38*(10), 1335-1342.

Bosch, N. M., Riese, H., Reijneveld, S. A., Bakker, M. P., Verhulst, F. C., Ormel, J., & Oldehinkel, A. J. (2012). Timing matters: Long term effects of adversities from prenatal period up to adolescence on adolescents' cortisol stress response. the TRAILS study. *Psychoneuroendocrinology, 37*(9), 1439-1447.

Boulesteix, A. L., & Strimmer, K. (2007). Partial least squares: A versatile tool for the analysis of high-dimensional genomic data. *Briefings in Bioinformatics, 8*(1), 32-44. doi:10.1093/bib/bbl016

Bouwland-Both, M. I., van Mil, N. H., Tolhoek, C. P., Stolk, L., Eilers, P. H., Verbiest, M. M., . . . Steegers-Theunissen, R. P. (2015). Prenatal parental tobacco smoking, gene specific DNA methylation, and newborns size: The generation R study. *Clinical Epigenetics, 7*(1), 83-z. eCollection 2015. doi:10.1186/s13148-015-0115-z [doi]

Boyce, W. T., & Ellis, B. J. (2005). Biological sensitivity to context: I. an evolutionary-developmental theory of the origins and functions of stress reactivity. *Development and Psychopathology, 17*(2), 271-301.

Bradley, R. H., & Corwyn, R. F. (2008). Infant temperament, parenting, and externalizing behavior in first grade: A test of the differential susceptibility hypothesis. *J Child Psychol Psychiatry, 49*(2), 124-31. doi:10.1111/j.1469-7610.2007.01829.x

Brand, J. S., Gaillard, R., West, J., McEachan, R. R., Wright, J., Voerman, E., . . . Lawlor, D. A. (2019). Associations of maternal quitting, reducing, and continuing smoking during pregnancy with longitudinal fetal growth: Findings from mendelian randomization and parental negative control studies. *PLoS Medicine, 16*(11), e1002972.

Breiman, L. (2001). Random forests. *Machine Learning, 45*(1), 5-32. doi:1010933404324

Brion, M. J., Victora, C., Matijasevich, A., Horta, B., Anselmi, L., Steer, C., . . . Davey Smith, G. (2010). Maternal smoking and child psychological problems: Disentangling causal and noncausal effects. *Pediatrics, 126*(1), 57. doi:10.1542/peds.2009-2754

Buchanan, A. V., Weiss, K. M., & Fullerton, S. M. (2006). Dissecting complex disease: The quest for the philosopher's stone? *International Journal of Epidemiology, 35*(3), 562-571. doi:dyl001 [pii]

Buck, J. M., Sanders, K. N., Wageman, C. R., Knopik, V.S., Stitzel, J. A., & O'Neill, H. C. (2019). Developmental nicotine exposure precipitates multigenerational maternal transmission of nicotine preference and ADHD-like behavioral, rhythmometric, neuropharmacological, and epigenetic anomalies in adolescent mice. *Neuropharmacology, 149*, 66-82.

Burdge, G. C., & Lillycrop, K. A. (2010). Bridging the gap between epigenetics research and nutritional public health interventions. *Genome Medicine, 2*(11), 80. doi:10.1186/gm201

Burns, J. (Ed.). (1970). *Towards a theoretical biology* Edinburgh Univ. Press.

Butcher, L. M., & Beck, S. (2015). Probe lasso: A novel method to rope in differentially methylated regions with 450K DNA methylation data. *Methods, 72*, 21-28.

Camerota, M., & Bollen, K. A. (2016). Birth weight, birth length, and gestational age as indicators of favorable fetal growth conditions in a US sample. *PloS One, 11*(4), e0153800.

Camerota, M., & Willoughby, M. T. (2019). Prenatal risk predicts preschooler executive function: A cascade model. *Child Development,*

Cao-Lei, L., Elgbeili, G., Szyf, M., Laplante, D. P., & King, S. (2019). Differential genome-wide DNA methylation patterns in childhood obesity. *BMC Research Notes, 12*(1), 174.

Cauchi, M., Weber, C. M., Bolt, B. J., Spratt, P. B., Bessant, C., Turner, D. C., . . . Morgan, G. (2016). Evaluation of gas chromatography mass spectrometry and pattern recognition for the identification of bladder cancer from urine headspace. *Analytical Methods, 8*(20), 4037-4046.

Cecil, C. A., Walton, E., Pingault, J., Provençal, N., Pappa, I., Vitaro, F., . . . Tiemeier, H. (2018). DRD4 methylation as a potential biomarker for physical aggression: An epigenome-wide, cross-tissue investigation. *American Journal of Medical Genetics Part B: Neuropsychiatric Genetics, 177*(8), 746-764.

Cents, R. A., Diamantopoulou, S., Hudziak, J. J., Jaddoe, V. W., Hofman, A., Verhulst, F. C., . . . Tiemeier, H. (2013). Trajectories of maternal depressive symptoms predict child problem behaviour: The generation R study. *Psychological Medicine, 43*(1), 13-25. doi:10.1017/S0033291712000657; 10.1017/S0033291712000657

Cerf, L., Gay, D., Selmaoui-Folcher, N., Crémilleux, B., & Boulicaut, J. (2013). Parameter-free classification in multi-class imbalanced data sets. *Data & Knowledge Engineering, 87*, 109-129.

Chambeyron, S., & Bickmore, W. A. (2004). Chromatin decondensation and nuclear reorganization of the HoxB locus upon induction of transcription. *Genes & Development, 18*(10), 1119-1130. doi:10.1101/gad.292104 [doi]

Chatterton, Z., Hartley, B. J., Seok, M., Mendelev, N., Chen, S., Milekic, M., . . . Brennand, K. (2017). In utero exposure to maternal smoking is associated with DNA methylation alterations and reduced neuronal content in the developing fetal brain. *Epigenetics & Chromatin, 10*(1), 4.

Chen, T., & Dent, S. Y. (2014). Chromatin modifiers and remodellers: Regulators of cellular differentiation. *Nature Reviews.Genetics, 15*(2), 93-106. doi:10.1038/nrg3607 [doi]

Chen, Y. A., Lemire, M., Choufani, S., Butcher, D. T., Grafodatskaya, D., Zanke, B. W., . . . Weksberg, R. (2013). Discovery of cross-reactive probes and polymorphic CpGs in the illumina infinium HumanMethylation450 microarray. *Epigenetics : Official Journal of the DNA Methylation Society, 8*(2), 203-209. doi:10.4161/epi.23470; 10.4161/epi.23470

Cheung, K., Burgers, M. J., Young, D. A., Cockell, S., & Reynard, L. N. (2019). Correlation of infinium HumanMethylation450K and MethylationEPIC BeadChip arrays in cartilage. *Epigenetics*, 1-10.

Chhabra, D., Sharma, S., Kho, A. T., Gaedigk, R., Vyhlidal, C. A., Leeder, J. S., . . . Tantisira, K. G. (2014). Fetal lung and placental methylation is associated with in utero nicotine exposure. *Epigenetics, 9*(11), 1473-1484.

Choukrallah, M., Sewer, A., Talikka, M., Sierro, N., Peitsch, M. C., Hoeng, J., & Ivanov, N. V. (2018). *Epigenomics in tobacco risk assessment: Opportunities for integrated new approaches* doi://doi.org/10.1016/j.cotox.2019.01.001

Civelek, M., & Lusis, A. J. (2014). Systems genetics approaches to understand complex traits. *Nature Reviews Genetics, 15*(1), 34.

Class, Q. A., Lichtenstein, P., Långström, N., & D'onofrio, B. M. (2011). Timing of prenatal maternal exposure to severe life events and adverse pregnancy outcomes: A population study of 2.6 million pregnancies. *Psychosomatic Medicine, 73*(3), 234.

Cnattingius, S. (2004). The epidemiology of smoking during pregnancy: Smoking prevalence, maternal characteristics, and pregnancy outcomes. *Nicotine & Tobacco Research, 6*(Suppl_2), S125-S140.

Collings, C. K., & Anderson, J. N. (2017). Links between DNA methylation and nucleosome occupancy in the human genome. *Epigenetics & Chromatin, 10*(1), 18.

Cook, J. L., Green, C. R., de la Ronde, S., Dell, C. A., Graves, L., Ordean, A., . . . Wong, S. (2017). Epidemiology and effects of substance use in pregnancy. *Journal of Obstetrics and Gynaecology Canada : JOGC = Journal D'Obstetrique Et Gynecologie Du Canada : JOGC, 39*(10), 906-915. doi:S1701-2163(17)30508-X [pii]

Cook, P. R., & Marenduzzo, D. (2018). Transcription-driven genome organization: A model for chromosome structure and the regulation of gene expression tested through simulations. *Nucleic Acids Research, 46*(19), 9895-9906.

Crockenberg, S. C., & Leerkes, E. M. (2004). Infant and maternal behaviors regulate infant reactivity to novelty at 6 months. *Developmental Psychology, 40*(6), 1123-1132.

Czamara, D., Eraslan, G., Page, C. M., Lahti, J., Lahti-Pulkkinen, M., Hämäläinen, E., . . . Reynolds, R. M. (2019). Integrated analysis of environmental and genetic influences on cord blood DNA methylation in new-borns. *Nature Communications, 10*(1), 1-18.

Dall'Aglio, L., Muka, T., Cecil, C. A., Bramer, W. M., Verbiest, M. M., Nano, J., . . . Tiemeier, H. (2018). The role of epigenetic modifications in neurodevelopmental disorders: A systematic review. *Neuroscience & Biobehavioral Reviews, 94*, 17-30.

de Goede, O. M., Razzaghian, H. R., Price, E. M., Jones, M. J., Kobor, M. S., Robinson, W. P., & Lavoie, P. M. (2015). Nucleated red blood cells impact DNA methylation and expression analyses of cord blood hematopoietic cells. *Clinical Epigenetics, 7*, 95-6. eCollection 2015. doi:10.1186/s13148-015-0129-6 [doi]

De Pauw, S. S., Mervielde, I., & Van Leeuwen, K. G. (2009). How are traits related to problem behavior in preschoolers? similarities and contrasts between temperament and personality. *J Abnorm Child Psychol, 37*(3), 309-25. doi:10.1007/s10802-008-9290-0

de Vocht, F., Simpkin, A. J., Richmond, R. C., Relton, C., & Tilling, K. (2015). Assessment of offspring DNA methylation across the lifecourse associated with prenatal maternal smoking using bayesian mixture modelling. *International Journal of Environmental Research and Public Health, 12*(11), 14461-14476. doi:10.3390/ijerph121114461 [doi]

Debruyne, M., Höppner, S., Serneels, S., & Verdonck, T. (2017). Outlyingness: Why do outliers lie out? *arXiv Preprint arXiv:1708.03761,*

Degenhardt, F., Seifert, S., & Szymczak, S. (2019). Evaluation of variable selection methods for random forests and omics data sets. *Briefings in Bioinformatics, 20*(2), 492-503. doi:10.1093/bib/bbx124 [doi]

Dennis, R., & Forzani, L. (2018). *Partial least squares prediction in high-dimensional regression.*

Devlin, A. M., Brain, U., Austin, J., & Oberlander, T. F. (2010). Prenatal exposure to maternal depressed mood and the MTHFR C677T variant affect SLC6A4 methylation in infants at birth. *PloS One, 5*(8), e12201. doi:10.1371/journal.pone.0012201

Dhanasekaran, K., Kumari, S., & Kanduri, C. (2013). Noncoding RNAs in chromatin organization and transcription regulation: An epigenetic view. *Sub-Cellular Biochemistry, 61*, 343-372. doi:10.1007/978-94-007-4525-4_15 [doi]

Di Pierro, M., Cheng, R. R., Aiden, E. L., Wolynes, P. G., & Onuchic, J. N. (2018). De novo prediction of human chromosome structures: Epigenetic marking patterns encode genome architecture. *Biophysical Journal, 114*(3), 597a.

Diaz-Uriarte, R., & Alvarez de Andres, S. (2006). Gene selection and classification of microarray data using random forest. *BMC Bioinformatics, 7*, 3-3. doi:1471-2105-7-3 [pii]

Dietz, P. M., Homa, D., England, L. J., Burley, K., Tong, V. T., Dube, S. R., & Bernert, J. T. (2011). Estimates of nondisclosure of cigarette smoking among pregnant and nonpregnant women of reproductive age in the United States. *American Journal of Epidemiology, 173*(3), 355-359.

Dover, G. J. (2009). The barker hypothesis: How pediatricans will diagnose and prevent common adult-onset diseases. *Transactions of the American Clinical and Climatological Association, 120*, 199-207.

Drake, A. J., & Walker (2004). The intergenerational effects of fetal programming: Non-genomic mechanisms for the inheritance of low birth weight and cardiovascular risk. *Journal of Endocrinology, 180*(1), 1-16. doi:10.1677/joe.0.1800001

Duda, R. O., Hart, P. E., & Stork, D. G. (2001). Pattern classification wiley. *New York, 680*

Dupras, C., Joly, Y., & Rial-Sebbag, E. (2020). Human rights in the postgenomic era: Challenges and opportunities arising with epigenetics. *Social Science Information, 59*(1), 12-34.

Durif, G., Modolo, L., Michaelsson, J., Mold, J. E., Lambert-Lacroix, S., & Picard, F. (2018). High dimensional classification with combined adaptive sparse PLS and logistic regression. *Bioinformatics (Oxford, England), 34*(3), 485-493. doi:10.1093/bioinformatics/btx571 [doi]

Dutta, B., Wallqvist, A., & Reifman, J. (2012). PathNet: A tool for pathway analysis using topological information. *Source Code for Biology and Medicine, 7*(1), 10.

Edgar, R. D., Jones, M. J., Meaney, M. J., Turecki, G., & Kobor, M. S. (2017). BECon: A tool for interpreting DNA methylation findings from blood in the context of brain. *Translational Psychiatry, 7*(8), e1187.

Ekpu, V. U., & Brown, A. K. (2015). The economic impact of smoking and of reducing smoking prevalence: Review of evidence. *Tobacco use Insights, 8*, 1-35. doi:10.4137/TUI.S15628 [doi]

Ellis, B. J., Essex, M. J., & Boyce, W. T. (2005). Biological sensitivity to context: II. empirical explorations of an evolutionary-developmental theory. *Development and Psychopathology, 17*(2), 303-328.

Ernst, J., & Kellis, M. (2010). Discovery and characterization of chromatin states for systematic annotation of the human genome. *Nature Biotechnology, 28*(8), 817-825.

Esposito, E. A., Jones, M. J., Doom, J. R., MacIsaac, J. L., Gunnar, M. R., & Kobor, M. S. (2016). Differential DNA methylation in peripheral blood mononuclear cells in adolescents

exposed to significant early but not later childhood adversity. *Development and Psychopathology, 28*(4 Pt 2), 1385.

Everson, T. M., Vives-Usano, M., Seyve, E., Cardenas, A., Lacasaña, M., Craig, J. M., . . . Heude, B. (2019). Placental DNA methylation signatures of maternal smoking during pregnancy and potential impacts on fetal growth. *BioRxiv,* 663567.

Fang, H., Johnson, C., Chevalier, N., Stopp, C., Wiebe, S., Wakschlag, L. S., & Espy, K. A. (2010). Using propensity score modeling to minimize the influence of confounding risks related to prenatal tobacco exposure. *Nicotine & Tobacco Research : Official Journal of the Society for Research on Nicotine and Tobacco, 12*(12), 1211-1219. doi:10.1093/ntr/ntq170

Feng, Y. Q., Desprat, R., Fu, H., Olivier, E., Lin, C. M., Lobell, A., . . . Bouhassira, E. E. (2006). DNA methylation supports intrinsic epigenetic memory in mammalian cells. *PLoS Genetics, 2*(4), e65. doi:10.1371/journal.pgen.0020065

Fertig, A. R. (2010). Selection and the effect of prenatal smoking. *Health Economics, 19*(2), 209-226.

Fischle, W., Wang, Y., & Allis, C. D. (2003). Histone and chromatin cross-talk. *Current Opinion in Cell Biology, 15*(2), 172-183.

Fop, M., & Murphy, T. B. (2018). Variable selection methods for model-based clustering. *Statistics Surveys, 12*, 18-65.

Forest, M., O'Donnell, K. J., Voisin, G., Gaudreau, H., MacIsaac, J. L., McEwen, L. M., . . . Greenwood, C. M. T. (2018). Agreement in DNA methylation levels from the illumina 450K

array across batches, tissues, and time. *Epigenetics, 13*(1), 19-32. doi:10.1080/15592294.2017.1411443 [doi]

Forray, A., & Foster, D. (2015). Substance use in the perinatal period. *Current Psychiatry Reports, 17*(11), 91-5. doi:10.1007/s11920-015-0626-5 [doi]

Fortin, J. P., & Hansen, K. D. (2015). Reconstructing A/B compartments as revealed by hi-C using long-range correlations in epigenetic data. *Genome Biology, 16*, 180-y. doi:10.1186/s13059-015-0741-y [doi]

Freathy, R. M., Ring, S. M., Shields, B., Galobardes, B., Knight, B., Weedon, M. N., . . . Hattersley, A. T. (2009). A common genetic variant in the 15q24 nicotinic acetylcholine receptor gene cluster (CHRNA5-CHRNA3-CHRNB4) is associated with a reduced ability of women to quit smoking in pregnancy. *Human Molecular Genetics, 18*(15), 2922-2927. doi:10.1093/hmg/ddp216

Frieser, M. J., Wilson, S., & Vrieze, S. (2018). Behavioral impact of return of genetic test results for complex disease: Systematic review and meta-analysis. *Health Psychology, 37*(12), 1134.

Fu, Y., Tessneer, K. L., Li, C., & Gaffney, P. M. (2018). From association to mechanism in complex disease genetics: The role of the 3D genome. *Arthritis Research & Therapy, 20*(1), 216-x. doi:10.1186/s13075-018-1721-x [doi]

Fuchikami, M., Morinobu, S., Segawa, M., Okamoto, Y., Yamawaki, S., Ozaki, N., . . . Tsuchiyama, K. (2011). DNA methylation profiles of the brain-derived neurotrophic factor (BDNF) gene as a potent diagnostic biomarker in major depression. *PloS One, 6*(8), e23881.

Gao, C., Sun, H., Wang, T., Tang, M., Bohnen, N. I., Müller, M. L., . . . Spino, C. (2018). Model-based and model-free machine learning techniques for diagnostic prediction and classification of clinical outcomes in parkinson's disease. *Scientific Reports, 8*(1), 7129.

Gao, L., Liu, X., Millstein, J., Siegmund, K. D., Dubeau, L., Maguire, R. L., . . . Hoyo, C. (2018). Self-reported prenatal tobacco smoke exposure, AXL gene-body methylation, and childhood asthma phenotypes. *Clinical Epigenetics, 10*(1), 98.

Geeleher, P., Hartnett, L., Egan, L. J., Golden, A., Raja Ali, R. A., & Seoighe, C. (2013). Gene-set analysis is severely biased when applied to genome-wide methylation data. *Bioinformatics (Oxford, England), 29*(15), 1851-1857. doi:10.1093/bioinformatics/btt311 [doi]

Geiman, T. M., & Robertson, K. D. (2002). Chromatin remodeling, histone modifications, and DNA methylation—how does it all fit together? *Journal of Cellular Biochemistry, 87*(2), 117-125.

Genuer, R., Poggi, J., & Tuleau-Malot, C. (2015). VSURF: An R package for variable selection using random forests. *The R Journal, 7*(2), 19-33. Retrieved from https://hal.archives-ouvertes.fr/hal-01251924

Giambartolomei, C., Zhenli Liu, J., Zhang, W., Hauberg, M., Shi, H., Boocock, J., . . . Pasaniuc, B. (2018). A bayesian framework for multiple trait colocalization from summary association statistics. *Bioinformatics, 34*(15), 2538-2545.

Gilliland, F. D., Li, Y., & Peters, J. M. (2001). Effects of maternal smoking during pregnancy and environmental tobacco smoke on asthma and wheezing in children. *American Journal of Respiratory and Critical Care Medicine, 163*(2), 429-436.

Golding, J. (1990). Children of the nineties. A longitudinal study of pregnancy and childhood based on the population of avon (ALSPAC). *West of England Medical Journal, 105*(3), 80-82.

Gorber, S. C., Schofield-Hurwitz, S., Hardt, J., Levasseur, G., & Tremblay, M. (2009). The accuracy of self-reported smoking: A systematic review of the relationship between self-reported and cotinine-assessed smoking status. *Nicotine & Tobacco Research : Official Journal of the Society for Research on Nicotine and Tobacco, 11*(1), 12-24. doi:10.1093/ntr/ntn010

Greene, C. S., Penrod, N. M., Kiralis, J., & Moore, J. H. (2009). Spatially uniform relieff (SURF) for computationally-efficient filtering of gene-gene interactions. *BioData Mining, 2*(1), 5.

Gromski, P. S., Muhamadali, H., Ellis, D. I., Xu, Y., Correa, E., Turner, M. L., & Goodacre, R. (2015). A tutorial review: Metabolomics and partial least squares-discriminant analysis--a marriage of convenience or a shotgun wedding. *Analytica Chimica Acta, 879*, 10-23. doi:10.1016/j.aca.2015.02.012 [doi]

Guerin, D. W., Gottfried, A. W., & Thomas, C. W. (1997). Difficult temperament and behaviour problems: A longitudinal study from 1.5 to 12 years. *International Journal of Behavioral Development, 21*(1), 71-90. doi:10.1080/016502597384992

Guerrero-Preston, R., Goldman, L. R., Brebi-Mieville, P., Ili-Gangas, C., Lebron, C., Witter, F. R., . . . Sidransky, D. (2010). Global DNA hypomethylation is associated with in utero exposure to cotinine and perfluorinated alkyl compounds. *Epigenetics : Official Journal of the DNA Methylation Society, 5*(6), 539-546.

Gutierrez-Arcelus, M., Lappalainen, T., Montgomery, S. B., Buil, A., Ongen, H., Yurovsky, A., . . . Planchon, A. (2013). Passive and active DNA methylation and the interplay with genetic variation in gene regulation. *Elife, 2*, e00523.

Hah, N., Benner, C., Chong, L. W., Yu, R. T., Downes, M., & Evans, R. M. (2015). Inflammation-sensitive super enhancers form domains of coordinately regulated enhancer RNAs. *Proceedings of the National Academy of Sciences of the United States of America, 112*(3), 297. doi:10.1073/pnas.1424028112 [doi]

Hales, C. N., & Barker, D. J. (1992). Type 2 (non-insulin-dependent) diabetes mellitus: The thrifty phenotype hypothesis. *Diabetologia, 35*(7), 595-601.

Hamza, M., Halayem, S., Bourgou, S., Daoud, M., Charfi, F., & Belhadj, A. (2019). Epigenetics and ADHD: Toward an integrative approach of the disorder pathogenesis. *Journal of Attention Disorders, 23*(7), 655-664.

Hanington, L., Ramchandani, P., & Stein, A. (2010). Parental depression and child temperament: Assessing child to parent effects in a longitudinal population study. *Infant Behavior & Development, 33*(1), 88-95. doi:10.1016/j.infbeh.2009.11.004

Hannon, E., Lunnon, K., Schalkwyk, L., & Mill, J. (2015). Interindividual methylomic variation across blood, cortex, and cerebellum: Implications for epigenetic studies of neurological and neuropsychiatric phenotypes. *Epigenetics, 10*(11), 1024-1032.

Harikumar, A., & Meshorer, E. (2015). Chromatin remodeling and bivalent histone modifications in embryonic stem cells. *EMBO Reports, 16*(12), 1609-1619. doi:10.15252/embr.201541011 [doi]

Harper, K. N., Peters, B. A., & Gamble, M. V. (2013). Batch effects and pathway analysis: Two potential perils in cancer studies involving DNA methylation array analysis. *Cancer Epidemiology, Biomarkers & Prevention : A Publication of the American Association for Cancer Research, Cosponsored by the American Society of Preventive Oncology,* doi:10.1158/1055-9965.EPI-13-0114

Haworth, K. E., Farrell, W. E., Emes, R. D., Ismail, K. M., Carroll, W. D., Borthwick, H. A., . . . Fryer, A. A. (2013). Combined influence of gene-specific cord blood methylation and maternal smoking habit on birth weight. *Epigenomics, 5*(1), 37-49. doi:10.2217/epi.12.72; 10.2217/epi.12.72

Heinonen, K., Räikkönen, K., Pesonen, A., Andersson, S., Kajantie, E., Eriksson, J. G., . . . Lano, A. (2011). Longitudinal study of smoking cessation before pregnancy and children's cognitive abilities at 56 months of age. *Early Human Development, 87*(5), 353-359.

Herrera, B. M., & Lindgren, C. M. (2010). The genetics of obesity. *Current Diabetes Reports, 10*(6), 498-505.

Holbrook, J. D., Huang, R. C., Barton, S. J., Saffery, R., & Lillycrop, K. A. (2017). Is cellular heterogeneity merely a confounder to be removed from epigenome-wide association studies? *Epigenomics, 9*(8), 1143-1150. doi:10.2217/epi-2017-0032 [doi]

Hore, V., Viñuela, A., Buil, A., Knight, J., McCarthy, M. I., Small, K., & Marchini, J. (2016). Tensor decomposition for multiple-tissue gene expression experiments. *Nature Genetics, 48*, 1094. Retrieved from https://doi.org/10.1038/ng.3624

Horsthemke, B. (2006). Epimutations in human disease. *DNA methylation: Development, genetic disease and cancer* (pp. 45-59) Springer.

Horsthemke, B. (2018). A critical view on transgenerational epigenetic inheritance in humans. *Nature Communications, 9*(1), 1-4.

Houle, D., Govindaraju, D. R., & Omholt, S. (2010). Phenomics: The next challenge. *Nature Reviews Genetics, 11*(12), 855.

Howells, L., Musaddaq, B., McKay, A. J., & Majeed, A. (2016). Clinical impact of lifestyle interventions for the prevention of diabetes: An overview of systematic reviews. *BMJ Open, 6*(12), e013806-013806. doi:10.1136/bmjopen-2016-013806 [doi]

Hu, Z., & Tee, W. W. (2017). Enhancers and chromatin structures: Regulatory hubs in gene expression and diseases. *Bioscience Reports, 37*(2), 10.1042/BSR20160183. Print 2017 Apr 30. doi:BSR20160183 [pii]

Huang da, W., Sherman, B. T., & Lempicki, R. A. (2009). Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nature Protocols, 4*(1), 44-57. doi:10.1038/nprot.2008.211

Huang, J., Marco, E., Pinello, L., & Yuan, G. C. (2015). Predicting chromatin organization using histone marks. *Genome Biology, 16*, 162-z. doi:10.1186/s13059-015-0740-z [doi]

Hubé, F., & Francastel, C. (2018). Coding and non-coding RNAs, the frontier has never been so blurred. *Frontiers in Genetics, 9*, 140.

Huncharek, M., Haddock, K. S., Reid, R., & Kupelnick, B. (2010). Smoking as a risk factor for prostate cancer: A meta-analysis of 24 prospective cohort studies. *American Journal of Public Health, 100*(4), 693-701. doi:10.2105/AJPH.2008.150508 [doi]

Husson, F., Josse, J., Le, S., Mazet, J., & Husson, M. F. (2016). Package 'FactoMineR'. *An R Package, 96*, 698.

Hypponen, E., Smith, G. D., & Power, C. (2003). Effects of grandmothers' smoking in pregnancy on birth weight: Intergenerational cohort study. *BMJ (Clinical Research Ed.), 327*(7420), 898. doi:10.1136/bmj.327.7420.898

Iles-Caven, Y., Golding, J., Gregory, S., Emond, A., & Taylor, C. M. (2016). Data relating to early child development in the avon longitudinal study of parents and children (ALSPAC), their relationship with prenatal blood mercury and stratification by fish consumption. *Data in Brief, 9*, 112-122.

Iovleff, S. (2019). MixAll: Clustering mixed data with missing values [computer software]

Issa, J. J. (2007). DNA methylation as a therapeutic target in cancer. *Clinical Cancer Research, 13*(6), 1634-1637.

Jaffe, A. E., & Irizarry, R. A. (2014). Accounting for cellular heterogeneity is critical in epigenome-wide association studies. *Genome Biology, 15*(2), R31-r31. doi:10.1186/gb-2014-15-2-r31 [doi]

Jansen, P. W., Raat, H., Mackenbach, J. P., Jaddoe, V. W., Hofman, A., Verhulst, F. C., & Tiemeier, H. (2009). Socioeconomic inequalities in infant temperament: The generation R study. *Social Psychiatry and Psychiatric Epidemiology, 44*(2), 87-95. doi:10.1007/s00127-008-0416-z

Jauniaux, E., & Burton, G. J. (2007). Morphological and biological effects of maternal exposure to tobacco smoke on the feto-placental unit. *Early Human Development, 83*(11), 699-706.

Jenkins, T. G., James, E. R., Alonso, D. F., Hoidal, J. R., Murphy, P. J., Hotaling, J. M., . . . Aston, K. I. (2017). Cigarette smoking significantly alters sperm DNA methylation patterns. *Andrology, 5*(6), 1089-1099.

Jiao, C., Zhang, C., Dai, R., Xia, Y., Wang, K., Giase, G., . . . Liu, C. (2018). Positional effects revealed in illumina methylation array and the impact on analysis. *Epigenomics, 10*(5), 643-659. doi:10.2217/epi-2017-0105 [doi]

Joehanes, R., Just, A. C., Marioni, R. E., Pilling, L. C., Reynolds, L. M., Mandaviya, P. R., . . . London, S. J. (2016). Epigenetic signatures of cigarette smoking. *Circulation.Cardiovascular Genetics, 9*(5), 436-447. doi:CIRCGENETICS.116.001506 [pii]

Joo, J. E., Wong, E. M., Baglietto, L., Jung, C. H., Tsimiklis, H., Park, D. J., . . . Southey, M. C. (2013). The use of DNA from archival dried blood spots with the infinium HumanMethylation450 array. *BMC Biotechnology, 13*, 23-23. doi:10.1186/1472-6750-13-23 [doi]

Joubert, B. R., Felix, J. F., Yousefi, P., Bakulski, K. M., Just, A. C., Breton, C., . . . London, S. J. (2016). DNA methylation in newborns and maternal smoking in pregnancy: Genome-wide consortium meta-analysis. *American Journal of Human Genetics, 98*(4), 680-696. doi:10.1016/j.ajhg.2016.02.019 [doi]

Joubert, B. R., Haberg, S. E., Bell, D. A., Nilsen, R. M., Vollset, S. E., Midttun, O., . . . London, S. J. (2014). Maternal smoking and DNA methylation in newborns: In utero effect or epigenetic inheritance? *Cancer Epidemiology, Biomarkers & Prevention : A Publication of the American Association for Cancer Research, Cosponsored by the American Society of Preventive Oncology, 23*(6), 1007-1017. doi:10.1158/1055-9965.EPI-13-1256; 10.1158/1055-9965.EPI-13-1256

Joubert, B. R., Haberg, S. E., Nilsen, R. M., Wang, X., Vollset, S. E., Murphy, S. K., . . . London, S. J. (2012). 450K epigenome-wide scan identifies differential DNA methylation in newborns related to maternal smoking during pregnancy. *Environmental Health Perspectives, 120*(10), 1425-1431. doi:10.1289/ehp.1205412; 10.1289/ehp.1205412

Joubert, B. R., Felix, J. F., Yousefi, P., Bakulski, K. M., Just, A. C., Breton, C., . . . Xu, C. (2016). DNA methylation in newborns and maternal smoking in pregnancy: Genome-wide consortium meta-analysis. *The American Journal of Human Genetics, 98*(4), 680-696.

Jung, I., Schmitt, A., Diao, Y., Lee, A. J., Liu, T., Yang, D., . . . Chee, S. (2019). A compendium of promoter-centered long-range chromatin interactions in the human genome. *Nature Genetics,* 1-8.

Kapranov, P., Cheng, J., Dike, S., Nix, D. A., Duttagupta, R., Willingham, A. T., . . . Hofacker, I. L. (2007). RNA maps reveal new RNA classes and a possible function for pervasive transcription. *Science, 316*(5830), 1484-1488.

Karimpour-Fard, A., Epperson, L. E., & Hunter, L. E. (2015). A survey of computational tools for downstream analysis of proteomic and other omic datasets. *Human Genomics, 9*(1), 28.

Kassambara, A. (2017). *Practical guide to principal component methods in R: PCA, M (CA), FAMD, MFA, HCPC, factoextra* STHDA.

Kazakevych, J., Sayols, S., Messner, B., Krienke, C., & Soshnikova, N. (2017). Dynamic changes in chromatin states during specification and differentiation of adult intestinal stem cells. *Nucleic Acids Research, 45*(10), 5770-5784. doi:10.1093/nar/gkx167 [doi]

Keller, M., Hopp, L., Liu, X., Wohland, T., Rohde, K., Cancello, R., . . . Eichelmann, F. (2017). Genome-wide DNA promoter methylation and transcriptome analysis in human adipose tissue unravels novel candidate genes for obesity. *Molecular Metabolism, 6*(1), 86-100.

Kelly, A. D., & Issa, J. J. (2017). The promise of epigenetic therapy: Reprogramming the cancer epigenome. *Current Opinion in Genetics & Development, 42*, 68-77. doi:S0959-437X(17)30036-9 [pii]

Keskitalo, K., Broms, U., Heliovaara, M., Ripatti, S., Surakka, I., Perola, M., . . . Kaprio, J. (2009). Association of serum cotinine level with a cluster of three nicotinic acetylcholine receptor genes (CHRNA3/CHRNA5/CHRNB4) on chromosome 15. *Human Molecular Genetics, 18*(20), 4007-4012. doi:10.1093/hmg/ddp322

Klein, H., & Hebestreit, K. (2016). An evaluation of methods to test predefined genomic regions for differential methylation in bisulfite sequencing data. *Briefings in Bioinformatics, 17*(5), 796-807.

Klengel, T., Mehta, D., Anacker, C., Rex-Haffner, M., Pruessner, J. C., Pariante, C. M., . . . Binder, E. B. (2013). Allele-specific FKBP5 DNA demethylation mediates gene-childhood trauma interactions. *Nature Neuroscience, 16*(1), 33-41. doi:10.1038/nn.3275 [doi]

Kline, J., Stein, Z., & Hutzler, M. (1987). Cigarettes, alcohol and marijuana: Varying associations with birthweight. *International Journal of Epidemiology, 16*(1), 44-51.

Knopik, V.S., Marceau, K., Bidwell, L. C., & Rolan, E. (2019). Prenatal substance exposure and offspring development: Does DNA methylation play a role? *Neurotoxicology and Teratology, 71*, 50-63. doi:S0892-0362(17)30153-8 [pii]

Knopik, V.S. (2009). Maternal smoking during pregnancy and child outcomes: Real or spurious effect? *Developmental Neuropsychology, 34*(1), 1-36. doi:10.1080/87565640802564366

Kooijman, M. N., Kruithof, C. J., van Duijn, C. M., Duijts, L., Franco, O. H., van IJzendoorn, M. H., . . . Mackenbach, J. P. (2016a). The generation R study: Design and cohort update 2017. *European Journal of Epidemiology, 31*(12), 1243-1264.

Kooijman, M. N., Kruithof, C. J., van Duijn, C. M., Duijts, L., Franco, O. H., van IJzendoorn, M. H., . . . Mackenbach, J. P. (2016b). The generation R study: Design and cohort update 2017. *European Journal of Epidemiology, 31*(12), 1243-1264.

Kramer, M. S. (1987). Intrauterine growth and gestational duration determinants. *Pediatrics, 80*(4), 502-511.

Kupers, L. K., Xu, X., Jankipersadsing, S. A., Vaez, A., la Bastide-van Gemert, S., Scholtens, S., . . . Snieder, H. (2015). DNA methylation mediates the effect of maternal smoking during pregnancy on birthweight of the offspring. *International Journal of Epidemiology, 44*(4), 1224-1237. doi:10.1093/ije/dyv048 [doi]

Kursa, M., Rudnicki,W. (2010). Feature selection with the boruta package. *Journal of Statistical Software, 36*(11) doi:10.18637/jss.v036.i11

Kursa, M. B. (2014). Robustness of random forest-based gene selection methods. *BMC Bioinformatics, 15*(1), 8.

Kyle, U. G., & Pichard, C. (2006). The dutch famine of 1944-1945: A pathophysiological model of long-term consequences of wasting disease. *Current Opinion in Clinical Nutrition and Metabolic Care, 9*(4), 388-394. doi:10.1097/01.mco.0000232898.74415.42 [doi]

Lamb, J., Crawford, E. D., Peck, D., Modell, J. W., Blat, I. C., Wrobel, M. J., . . . Ross, K. N. (2006). The connectivity map: Using gene-expression signatures to connect small molecules, genes, and disease. *Science, 313*(5795), 1929-1935.

Lamb, N. (2007). What is epigenetics? Retrieved from https://ghr.nlm.nih.gov/primer/howgeneswork/epigenome

Lanktree, M. B., Guo, Y., Murtaza, M., Glessner, J. T., Bailey, S. D., Onland-Moret, N. C., . . . Keating, B. J. (2011). Meta-analysis of dense genecentric association studies reveals common and uncommon variants associated with height. *American Journal of Human Genetics, 88*(1), 6-18. doi:10.1016/j.ajhg.2010.11.007

Lappalainen, T., & Greally, J. M. (2017). Associating cellular epigenetic models with human phenotypes. *Nature Reviews Genetics, 18*(7), 441.

Laucht, M., Esser, G., & Schmidt, M. H. (1997). Developmental outcome of infants born with biological and psychosocial risks. *Journal of Child Psychology and Psychiatry, 38*(7), 843-853.

Lay, F. D., Liu, Y., Kelly, T. K., Witt, H., Farnham, P. J., Jones, P. A., & Berman, B. P. (2015). The role of DNA methylation in directing the functional organization of the cancer epigenome. *Genome Research, 25*(4), 467-477. doi:10.1101/gr.183368.114 [doi]

Lazarovici, A., Zhou, T., Shafer, A., Machado, A. C. D., Riley, T. R., Sandstrom, R., . . . Stamatoyannopoulos, J. A. (2013). Probing DNA shape and methylation state on a genomic scale with DNase I. *Proceedings of the National Academy of Sciences, 110*(16), 6376-6381.

Le Clec'h, S., Oszwald, J., Decaens, T., Desjardins, T., Dufour, S., Grimaldi, M., . . . Lavelle, P. (2016). Mapping multiple ecosystem services indicators: Toward an objective-oriented approach. *Ecological Indicators, 69*, 508-521.

Lee, K. W., & Pausova, Z. (2013). Cigarette smoking and DNA methylation. *Frontiers in Genetics, 4*, 132.

Lee, K. W., Richmond, R., Hu, P., French, L., Shin, J., Bourdon, C., . . . Gaunt, T. (2014). Prenatal exposure to maternal cigarette smoking and DNA methylation: Epigenome-wide association in a discovery sample of adolescents and replication in an independent cohort at birth through 17 years of age. *Environmental Health Perspectives, 123*(2), 193-199.

Lee, L. C., Halpern, C. T., Hertz-Picciotto, I., Martin, S. L., & Suchindran, C. M. (2006). Child care and social support modify the association between maternal depressive symptoms and early childhood behaviour problems: A US national study. *Journal of Epidemiology and Community Health, 60*(4), 305-310. doi:10.1136/jech.2005.040956

Lehmann-Werman, R., Neiman, D., Zemmour, H., Moss, J., Magenheim, J., Vaknin-Dembinsky, A., . . . Zetterberg, H. (2016). Identification of tissue-specific cell death using methylation patterns of circulating DNA. *Proceedings of the National Academy of Sciences, 113*(13), E1826-E1834.

Lemery, K. S., Goldsmith, H. H., Klinnert, M. D., & Mrazek, D. A. (1999). Developmental models of infant and childhood temperament. *Developmental Psychology, 35*(1), 189-204.

Lewis, C. M., Whitwell, S. C., Forbes, A., Sanderson, J., Mathew, C. G., & Marteau, T. M. (2007). Estimating risks of common complex diseases across genetic and environmental factors: The example of crohn disease. *Journal of Medical Genetics, 44*(11), 689-694.

Leybovitz-Haleluya, N., Wainstock, T., Landau, D., & Sheiner, E. (2018). Maternal smoking during pregnancy and the risk of pediatric cardiovascular diseases of the offspring: A population-based cohort study with up to 18-years of follow up. *Reproductive Toxicology (Elmsford, N.Y.), 78*, 69-74. doi:S0890-6238(17)30727-X [pii]

Li, J., Tran, M., & Siwabessy, J. (2016). Selecting optimal random forest predictive models: A case study on predicting the spatial distribution of seabed hardness. *PloS One, 11*(2), e0149089. doi:10.1371/journal.pone.0149089 [doi]

Li, L., Peters, H., Gama, A., Carvalhal, M. I., Nogueira, H. G., Rosado-Marques, V., & Padez, C. (2016). Maternal smoking in pregnancy association with childhood adiposity and blood pressure. *Pediatric Obesity, 11*(3), 202-209. doi:10.1111/ijpo.12046 [doi]

Linnekamp, J. F., Butter, R., Spijker, R., Medema, J. P., & van Laarhoven, H. W. M. (2017). Clinical and biological effects of demethylating agents on solid tumours - A systematic review. *Cancer Treatment Reviews, 54*, 10-23. doi:S0305-7372(17)30004-X [pii]

Liquet, B., de Micheaux, P. L., Hejblum, B. P., & Thiebaut, R. (2016). Group and sparse group partial least square approaches applied in genomics context. *Bioinformatics (Oxford, England), 32*(1), 35-42. doi:10.1093/bioinformatics/btv535 [doi]

Liu, S., Zhang, L., Quan, H., Tian, H., Meng, L., Yang, L., . . . Gao, Y. Q. (2018). From 1D sequence to 3D chromatin dynamics and cellular functions: A phase separation perspective. *Nucleic Acids Research, 46*(18), 9367-9383. doi:10.1093/nar/gky633 [doi]

Liu, W., & Irudayaraj, J. (2020). Perfluorooctanoic acid (PFOA) exposure inhibits DNA methyltransferase activities and alters constitutive heterochromatin organization. *Food and Chemical Toxicology,* 111358.

Loucoubar, C., Grant, A. V., Bureau, J., Casademont, I., Bar, N. A., Bar-Hen, A., . . . Badiane, A. (2017). Detecting multi-way epistasis in family-based association studies. *Briefings in Bioinformatics, 18*(3), 394-402.

Lu, M., McComish, B. J., Burdon, K. P., Taylor, B. V., & Korner, H. (2019). The association between vitamin D and multiple sclerosis risk: 1,25(OH)2D3 induces super-enhancers bound by VDR. *Frontiers in Immunology, 10*, 488. doi:10.3389/fimmu.2019.00488 [doi]

Lunetta, K. L., Hayward, L. B., Segal, J., & Van Eerdewegh, P. (2004). Screening large-scale association study data: Exploiting interactions using random forests. *BMC Genetics, 5*(1), 32.

Lustberg, M., & Silbergeld, E. (2002). Blood lead levels and mortality. *Archives of Internal Medicine, 162*(21), 2443-2449.

Maccani, J. Z., Koestler, D. C., Houseman, E. A., Marsit, C. J., & Kelsey, K. T. (2013). Placental DNA methylation alterations associated with maternal tobacco smoking at the RUNX3 gene are also associated with gestational age. *Epigenomics, 5*(6), 619-630.

Manolio, T. A., Collins, F. S., Cox, N. J., Goldstein, D. B., Hindorff, L. A., Hunter, D. J., . . . Chakravarti, A. (2009). Finding the missing heritability of complex diseases. *Nature, 461*(7265), 747.

Manrai, A. K., Cui, Y., Bushel, P. R., Hall, M., Karakitsios, S., Mattingly, C. J., . . . Thomas, D. C. (2017). Informatics and data analytics to support exposome-based discovery for public health. *Annual Review of Public Health, 38*, 279-294.

Marbac, M., & Sedki, M. (2019). VarSelLCM: An R/C package for variable selection in model-based clustering of mixed-data with missing values. *Bioinformatics, 35*(7), 1255-1257.

Marceau, K., Palmer, R. H., Neiderhiser, J. M., Smith, T. F., McGeary, J. E., & Knopik, V.S. (2016). Passive rGE or developmental gene-environment cascade? an investigation of the role of xenobiotic metabolism genes in the association between smoke exposure during pregnancy and child birth weight. *Behavior Genetics, 46*(3), 365-377.

Margueron, R., & Reinberg, D. (2010). Chromatin structure and the inheritance of epigenetic information. *Nature Reviews.Genetics, 11*(4), 285-296. doi:10.1038/nrg2752; 10.1038/nrg2752

Markunas, C. A., Xu, Z., Harlid, S., Wade, P. A., Lie, R. T., Taylor, J. A., & Wilcox, A. J. (2014). Identification of DNA methylation changes in newborns related to maternal smoking during pregnancy. *Environmental Health Perspectives, 122*(10), 1147-1153. doi:10.1289/ehp.1307892 [doi]

Martino, D., Ben-Othman, R., Harbeson, D., & Bosco, A. (2019). Multiomics and systems biology are needed to unravel the complex origins of chronic disease. *Challenges, 10*(1), 23.

Massey, S. H., Mroczek, D. K., Burns, J. L., Clark, C. A., Espy, K. A., & Wakschlag, L. S. (2018). Positive parenting behaviors in women who spontaneously quit smoking during pregnancy: Clues to putative targets for preventive interventions. *Neurotoxicology and Teratology, 67*, 18-24.

McCann, S. E., Liu, S., Wang, D., Shen, J., Hu, Q., Hong, C., . . . Zhao, H. (2013). Reduction of dietary glycaemic load modifies the expression of microRNA potentially associated with energy balance and cancer pathways in pre-menopausal women. *British Journal of Nutrition, 109*(4), 585-592.

McMillen, I. C., & Robinson, J. S. (2005). Developmental origins of the metabolic syndrome: Prediction, plasticity, and programming. *Physiological Reviews, 85*(2), 571-633. doi:10.1152/physrev.00053.2003

Mehmood, T., Bohlin, J., Kristoffersen, A. B., Sæbø, S., Warringer, J., & Snipen, L. (2012). Exploration of multivariate analysis in microbial coding sequence modeling. *BMC Bioinformatics, 13*(1), 97.

Melas, P. A., Rogdaki, M., Ösby, U., Schalling, M., Lavebratt, C., & Ekström, T. J. (2012). Epigenetic aberrations in leukocytes of patients with schizophrenia: Association of global DNA methylation with antipsychotic drug treatment and disease onset. *The FASEB Journal, 26*(6), 2712-2718.

Meijer, M., Klein, M., Hannon, E., van der Meer, D., Hartman, C., Oosterlaan, J., . . . Mill, J. (2020). Genome-wide DNA methylation patterns in persistent attention-deficit/hyperactivity disorder and in association with impulsive and callous traits. *Frontiers in Genetics, 11*, 16.

Mevik, B., & Cederkvist, H. R. (2004). Mean squared error of prediction (MSEP) estimates for principal component regression (PCR) and partial least squares regression (PLSR). *Journal of Chemometrics, 18*(9), 422-429.

Meyer, D., & Wien, F. T. (2015). Support vector machines. *The Interface to Libsvm in Package e1071, 28*

Michels, K. B., Harris, H. R., & Barault, L. (2011). Birthweight, maternal weight trajectories and global DNA methylation of LINE-1 repetitive elements. *PloS One, 6*(9), e25254. doi:10.1371/journal.pone.0025254

Min, J. L., Hemani, G., Davey Smith, G., Relton, C., & Suderman, M. (2018). Meffil: Efficient normalization and analysis of very large DNA methylation datasets. *Bioinformatics, 34*(23), 3983-3989.

Mirowsky, J., & Ross, C. E. (2015). Education, health, and the default american lifestyle. *Journal of Health and Social Behavior, 56*(3), 297-306. doi:10.1177/0022146515594814 [doi]

Mishra, A., & Hawkins, R. D. (2017). Three-dimensional genome architecture and emerging technologies: Looping in disease. *Genome Medicine, 9*(1), 1-14.

Miyake, K., Kawaguchi, A., Miura, R., Kobayashi, S., Tran, N. Q. V., Kobayashi, S., . . . Yamagata, Z. (2018). Association between DNA methylation in cord blood and maternal smoking: The hokkaido study on environment and children's health. *Scientific Reports, 8*(1), 5654.

Moore, J. H., Asselbergs, F. W., & Williams, S. M. (2010). Bioinformatics challenges for genome-wide association studies. *Bioinformatics, 26*(4), 445-455.

Müller, F., Scherer, M., Assenov, Y., Lutsik, P., Walter, J., Lengauer, T., & Bock, C. (2019). RnBeads 2.0: Comprehensive analysis of DNA methylation data. *Genome Biology, 20*(1), 1-12.

Munafo, M. R., Freathy, R. M., Ring, S. M., St Pourcain, B., & Davey Smith, G. (2010). Association of COMT Val108/158Met genotype and cigarette smoking in pregnant women. *Nicotine & Tobacco Research : Official Journal of the Society for Research on Nicotine and Tobacco,* doi:10.1093/ntr/ntq209

Murgatroyd, C., & Spengler, D. (2011). Epigenetics of early child development. *Frontiers in Psychiatry / Frontiers Research Foundation, 2*, 16. doi:10.3389/fpsyt.2011.00016

Murphy, S. K., Adigun, A., Huang, Z., Overcash, F., Wang, F., Jirtle, R. L., . . . Hoyo, C. (2012). Gender-specific methylation differences in relation to prenatal exposure to cigarette smoke. *Gene, 494*(1), 36-43.

Murray, C. J., & Lopez, A. D. (1997). Global mortality, disability, and the contribution of risk factors: Global burden of disease study. *Lancet, 349*(9063), 1436-1442. doi:10.1016/S0140-6736(96)07495-8

Muthuirulan, P., & Capellini, T. D. (2019). Complex phenotypes: Mechanisms underlying variation in human stature. *Current Osteoporosis Reports, 17*(5), 301-323.

National Research Council. (2010). *Weight gain during pregnancy: Reexamining the guidelines* National Academies Press.

Neophytou, A. M., Oh, S. S., Hu, D., Huntsman, S., Eng, C., Rodríguez-Santana, J. R., . . . Burchard, E. G. (2019). In utero tobacco smoke exposure, DNA methylation, and asthma in latino children. *Environmental Epidemiology (Philadelphia, Pa.), 3*(3)

Nguyen, T., Shafi, A., Nguyen, T., & Draghici, S. (2019). Identifying significantly impacted pathways: A comprehensive review and assessment. *Genome Biology, 20*(1), 1-15.

Nicodemus-Johnson, J., Myers, R. A., Sakabe, N. J., Sobreira, D. R., Hogarth, D. K., Naureckas, E. T., . . . Ober, C. (2016). DNA methylation in lung cells is associated with asthma endotypes and genetic risk. *JCI Insight, 1*(20), e90151. doi:10.1172/jci.insight.90151 [doi]

Northstone, K., Guggenheim, J. A., Howe, L. D., Tilling, K., Paternoster, L., Kemp, J. P., . . . Williams, C. (2013). Body stature growth trajectories during childhood and the development

of myopia. *Ophthalmology, 120*(5), 1064-73.e1. doi:10.1016/j.ophtha.2012.11.004; 10.1016/j.ophtha.2012.11.004

Novakovic, B., Ryan, J., Pereira, N., Boughton, B., Craig, J. M., & Saffery, R. (2014). Postnatal stability, tissue, and time specific effects of AHRR methylation change in response to maternal smoking in pregnancy. *Epigenetics, 9*(3), 377-386.

Ober, C., & Vercelli, D. (2011). Gene-environment interactions in human disease: Nuisance or opportunity? *Trends in Genetics : TIG, 27*(3), 107-115. doi:10.1016/j.tig.2010.12.004

Oberlander, T. F., Papsdorf, M., Brain, U. M., Misri, S., Ross, C., & Grunau, R. E. (2010). Prenatal effects of selective serotonin reuptake inhibitor antidepressants, serotonin transporter promoter genotype (SLC6A4), and maternal mood on child behavior at 3 years of age. *Archives of Pediatrics & Adolescent Medicine, 164*(5), 444-451.

O'Reilly, E., Tuzova, A. V., Walsh, A. L., Russell, N. M., O'Brien, O., Kelly, S., . . . Perry, A. S. (2019). epiCaPture: A urine DNA methylation test for early detection of aggressive prostate cancer. *JCO Precision Oncology, 2019*, 10.1200/PO.18.00134. Epub 2019 Jan 14. doi:10.1200/PO.18.00134 [doi]

Pagès, J. (2004). Analyse factorielle de données mixtes. *Revue De Statistique Appliquée, 52*(4), 93-111.

Pagès, J. (2014). *Multiple factor analysis by example using R* Chapman and Hall/CRC.

Palmer, R. H., Bidwell, L. C., Heath, A. C., Brick, L. A., Madden, P. A., & Knopik, V.S. (2016). Effects of maternal smoking during pregnancy on offspring externalizing problems: Contextual effects in a sample of female twins. *Behavior Genetics, 46*(3), 403-415. doi:10.1007/s10519-016-9779-1 [doi]

Peters, T. J., Buckley, M. J., Statham, A. L., Pidsley, R., Samaras, K., Lord, R. V., . . . Molloy, P. L. (2015). De novo identification of differentially methylated regions in the human genome. *Epigenetics & Chromatin, 8*(1), 6.

Petronis, A. (2010). Epigenetics as a unifying principle in the aetiology of complex traits and diseases. *Nature, 465*(7299), 721.

Pickett, K. E., Wood, C., Adamson, J., D'Souza, L., & Wakschlag, L. S. (2008a). Meaningful differences in maternal smoking behaviour during pregnancy: Implications for infant behavioural vulnerability. *J Epidemiol Community Health, 62*(4), 318-24. doi:10.1136/jech.2006.058768

Pickett, K. E., Wood, C., Adamson, J., D'Souza, L., & Wakschlag, L. S. (2008b). Meaningful differences in maternal smoking behaviour during pregnancy: Implications for infant behavioural vulnerability. *J Epidemiol Community Health, 62*(4), 318-24. doi:10.1136/jech.2006.058768

Pidsley, R., Y Wong, C. C., Volta, M., Lunnon, K., Mill, J., & Schalkwyk, L. C. (2013). A data-driven approach to preprocessing illumina 450K methylation array data. *BMC Genomics, 14*, 293-293. doi:10.1186/1471-2164-14-293; 10.1186/1471-2164-14-293

Pina, C., Pinto, F., Feijo, J. A., & Becker, J. D. (2005). Gene family analysis of the arabidopsis pollen transcriptome reveals biological implications for cell growth, division control, and gene expression regulation. *Plant Physiology, 138*(2), 744-756. doi:10.1104/pp.104.057935

Pluess, M., & Belsky, J. (2011). Prenatal programming of postnatal plasticity? *Dev Psychopathol, 23*(1), 29-38. doi:10.1017/s0954579410000623

Polderman, T. J., Benyamin, B., de Leeuw, C. A., Sullivan, P. F., van Bochoven, A., Visscher, P. M., & Posthuma, D. (2015). Fifty years of twin studies: A meta-analysis of the heritability of human traits. *Nature Genetics, 47*, 702-709.

Portela, A., & Esteller, M. (2010). Epigenetic modifications and human disease. *Nature Biotechnology, 28*(10), 1057-1068. doi:10.1038/nbt.1685 [doi]

Prakash, K., & Fournier, D. (2018). Evidence for the implication of the histone code in building the genome structure. *Biosystems, 164*, 49-59.

Presson, A. P., Sobel, E. M., Papp, J. C., Suarez, C. J., Whistler, T., Rajeevan, M. S., . . . Horvath, S. (2008). Integrated weighted gene co-expression network analysis with an application to chronic fatigue syndrome. *BMC Systems Biology, 2*, 95. doi:10.1186/1752-0509-2-95

Price, M. E., Cotton, A. M., Lam, L. L., Farre, P., Emberly, E., Brown, C. J., . . . Kobor, M. S. (2013). Additional annotation enhances potential for biologically-relevant analysis of the illumina infinium HumanMethylation450 BeadChip array. *Epigenetics & Chromatin, 6*(1), 4-4. doi:10.1186/1756-8935-6-4; 10.1186/1756-8935-6-4

Price, N. D., Magis, A. T., Earls, J. C., Glusman, G., Levy, R., Lausted, C., . . . Zhou, Y. (2017). A wellness study of 108 individuals using personal, dense, dynamic data clouds. *Nature Biotechnology, 35*(8), 747.

Prickaerts, P., Adriaens, M. E., Beucken, T. V. D., Koch, E., Dubois, L., Dahlmans, V. E. H., . . . Voncken, J. W. (2016). Hypoxia increases genome-wide bivalent epigenetic marking by specific gain of H3K27me3. *Epigenetics & Chromatin, 9*, 46-0. eCollection 2016. doi:10.1186/s13072-016-0086-0 [doi]

Qi, Y. (2012). Random forest for bioinformatics. *Ensemble machine learning* (pp. 307-323) Springer.

Ragland, D. R. (1992). Dichotomizing continuous outcome variables: Dependence of the magnitude of association and statistical power on the cutpoint. *Epidemiology (Cambridge, Mass.), 3*(5), 434-440.

Rahmani, E., Zaitlen, N., Baran, Y., Eng, C., Hu, D., Galanter, J., . . . Halperin, E. (2016). Sparse PCA corrects for cell type heterogeneity in epigenome-wide association studies. *Nature Methods, 13*(5), 443-445. doi:10.1038/nmeth.3809 [doi]

Rai, B. (2017). Feature selection and predictive modeling of housing data using random forest. *World Academy of Science, Engineering and Technology, International Journal of Social, Behavioral, Educational, Economic, Business and Industrial Engineering, 11*(4), 919-923.

Rakyan, V. K., Down, T. A., Balding, D. J., & Beck, S. (2011). Epigenome-wide association studies for common human diseases. *Nature Reviews.Genetics, 12*(8), 529-541. doi:10.1038/nrg3000; 10.1038/nrg3000

Reinius, L. E., Acevedo, N., Joerink, M., Pershagen, G., Dahlén, S., Greco, D., . . . Kere, J. (2012). Differential DNA methylation in purified human blood cells: Implications for cell lineage and studies on disease susceptibility. *PloS One, 7*(7), e41361.

Rao, S., Chiu, T., Kribelbauer, J. F., Mann, R. S., Bussemaker, H. J., & Rohs, R. (2018). Systematic prediction of DNA shape changes due to CpG methylation explains epigenetic effects on protein–DNA binding. *Epigenetics & Chromatin, 11*(1), 6.

Rauschert, S., Melton, P. E., Burdge, G. C., Craig, J. M., Godfrey, K. M., Holbrook, J. D., . . . Pennell, C. (2019). Maternal smoking during pregnancy induces persistent epigenetic changes into adolescence, independent of postnatal smoke exposure and is associated with cardiometabolic risk. *Frontiers in Genetics, 10*, 770.

Rauschert, S., Melton, P. E., Heiskala, A., Karhunen, V., Burdge, G., Craig, J. M., . . . Beilin, L. J. (2020). Machine learning-based DNA methylation score for fetal exposure to maternal smoking: Development and validation in samples collected from adolescents and adults. *Environmental Health Perspectives, 128*(9), 097003.

Raviram, R., Rocha, P. P., Muller, C. L., Miraldi, E. R., Badri, S., Fu, Y., . . . Skok, J. A. (2016). 4C-ker: A method to reproducibly identify genome-wide interactions captured by 4C-seq experiments. *PLoS Computational Biology, 12*(3), e1004780. doi:10.1371/journal.pcbi.1004780 [doi]

Reed, Z. E., Suderman, M. J., Relton, C. L., Davis, O. S., & Hemani, G. (2020). The association of DNA methylation with body mass index: Distinguishing between predictors and biomarkers. *Clinical Epigenetics, 12*, 1-13.

Reese, S. E., Zhao, S., Wu, M. C., Joubert, B. R., Parr, C. L., Haberg, S. E., . . . London, S. J. (2017). DNA methylation score as a biomarker in newborns for sustained maternal smoking during pregnancy. *Environmental Health Perspectives, 125*(4), 760-766. doi:10.1289/EHP333 [doi]

Reese, S. E., Xu, C., Herman, T., Lee, M. K., Sikdar, S., Ruiz-Arenas, C., . . . Ullemar, V. (2019). Epigenome-wide meta-analysis of DNA methylation and childhood asthma. *Journal of Allergy and Clinical Immunology, 143*(6), 2062-2074.

Relton, C. L., & Davey Smith, G. (2010). Epigenetic epidemiology of common complex disease: Prospects for prediction, prevention, and treatment. *PLoS Medicine, 7*(10), e1000356. doi:10.1371/journal.pmed.1000356

Relton, C. L., Gaunt, T., McArdle, W., Ho, K., Duggirala, A., Shihab, H., . . . Reik, W. (2015). Data resource profile: Accessible resource for integrated epigenomic studies (ARIES). *International Journal of Epidemiology, 44*(4), 1181-1190.

Renard, E., & Absil, P. (2017). (2017). Comparison of location-scale and matrix factorization batch effect removal methods on gene expression datasets. Paper presented at the *2017 IEEE International Conference on Bioinformatics and Biomedicine (BIBM),* 1530-1537.

Rhemtulla, M., Brosseau-Liard, P. É, & Savalei, V. (2012). When can categorical variables be treated as continuous? A comparison of robust continuous and categorical SEM estimation methods under suboptimal conditions. *Psychological Methods, 17*(3), 354.

Richardson, T. G., Richmond, R. C., North, T. L., Hemani, G., Davey Smith, G., Sharp, G. C., & Relton, C. L. (2019). An integrative approach to detect epigenetic mechanisms that putatively mediate the influence of lifestyle exposures on disease susceptibility. *International Journal of Epidemiology, 48*(3), 887-898. doi:10.1093/ije/dyz119 [doi]

Richmond, R. C., Al-Amin, A., Smith, G. D., & Relton, C. L. (2014). Approaches for drawing causal inferences from epidemiological birth cohorts: A review. *Early Human Development, 90*(11), 769-780. doi:10.1016/j.earlhumdev.2014.08.023 [doi]

Richmond, R. C., Simpkin, A. J., Woodward, G., Gaunt, T. R., Lyttleton, O., McArdle, W. L., . . . Relton, C. L. (2015). Prenatal exposure to maternal smoking and offspring DNA methylation across the lifecourse: Findings from the avon longitudinal study of parents and children (ALSPAC). *Human Molecular Genetics, 24*(8), 2201-2217. doi:10.1093/hmg/ddu739 [doi]

Roadmap, E. C., Kundaje, A., Meuleman, W., Ernst, J., Bilenky, M., Yen, A., . . . Kellis, M. (2015). Integrative analysis of 111 reference human epigenomes. *Nature, 518*, 317. Retrieved from https://doi.org/10.1038/nature14248

Rodgers, A. B., & Bale, T. L. (2015). Germ cell origins of posttraumatic stress disorder risk: The transgenerational impact of parental stress experience. *Biological Psychiatry, 78*(5), 307-314.

Rodriguez, J. J., & Smith, V. C. (2019). Epidemiology of perinatal substance use: Exploring trends in maternal substance use. *Seminars in Fetal & Neonatal Medicine,* doi:S1744-165X(19)30015-0 [pii]

Rodriguez-Bernal, C. L., Rebagliato, M., Iniguez, C., Vioque, J., Navarrete-Munoz, E. M., Murcia, M., . . . Ballester, F. (2010). Diet quality in early pregnancy and its effects on fetal growth outcomes: The infancia y medio ambiente (childhood and environment) mother and child cohort study in spain. *The American Journal of Clinical Nutrition, 91*(6), 1659-1666. doi:10.3945/ajcn.2009.28866; 10.3945/ajcn.2009.28866

Rogers, J. M. (2019). Smoking and pregnancy: Epigenetics and developmental origins of the metabolic syndrome. *Birth Defects Research,*

Roseboom, T., de Rooij, S., & Painter, R. (2006). The dutch famine and its long-term consequences for adult health. *Early Human Development, 82*(8), 485-491. doi:S0378-3782(06)00184-8 [pii]

Roseboom, T. J., Van Der Meulen, Jan HP, Ravelli, A. C., Osmond, C., Barker, D. J., & Bleker, O. P. (2001). Effects of prenatal exposure to the dutch famine on adult disease in later life: An overview. *Twin Research and Human Genetics, 4*(5), 293-298.

Royston, P., Altman, D. G., & Sauerbrei, W. (2006). Dichotomizing continuous predictors in multiple regression: A bad idea. *Statistics in Medicine, 25*(1), 127-141.

Rueda, M. R., & Rothbart, M. K. (2009). The influence of temperament on the development of coping: The role of maturation and experience. *New Dir Child Adolesc Dev, 2009*(124), 19-31. doi:10.1002/cd.240

Rye, M., Sandve, G. K., Daub, C. O., Kawaji, H., Carninci, P., Forrest, A. R., . . . FANTOM consortium. (2014). Chromatin states reveal functional associations for globally defined transcription start sites in four human cell lines. *BMC Genomics, 15*, 120-120. doi:10.1186/1471-2164-15-120 [doi]

Sadowski, M., Kraft, A., Szalaj, P., Wlasnowolski, M., Tang, Z., Ruan, Y., & Plewczynski, D. (2019). Spatial chromatin architecture alteration by structural variations in human genomes at the population scale. *Genome Biology, 20*(1), 148.

Sanavia, T., Aiolli, F., Da San Martino, G., Bisognin, A., & Di Camillo, B. (2012). Improving biomarker list stability by integration of biological knowledge in the learning process. *BMC Bioinformatics, 13 Suppl 4*, S22-S22. doi:10.1186/1471-2105-13-S4-S22; 10.1186/1471-2105-13-S4-S22

Sandoval, J., Heyn, H., Moran, S., Serra-Musach, J., Pujana, M. A., Bibikova, M., & Esteller, M. (2011). Validation of a DNA methylation microarray for 450,000 CpG sites in the human genome. *Epigenetics : Official Journal of the DNA Methylation Society, 6*(6), 692-702.

Schoft, V. K., Chumak, N., Mosiolek, M., Slusarz, L., Komnenovic, V., Brownfield, L., . . . Tamaru, H. (2009). Induction of RNA-directed DNA methylation upon decondensation of constitutive heterochromatin. *EMBO Reports, 10*(9), 1015-1021. doi:10.1038/embor.2009.152; 10.1038/embor.2009.152

Schvartzman, J. M., Thompson, C. B., & Finley, L. W. S. (2018). Metabolic regulation of chromatin modifications and gene expression. *The Journal of Cell Biology, 217*(7), 2247-2259. doi:10.1083/jcb.201803061 [doi]

Scornet, E., Biau, G., & Vert, J. (2015). Consistency of random forests. *The Annals of Statistics, 43*(4), 1716-1741.

Selvin, S. (1996). Statistical power and sample-size calculations. *Statistical Analyses of Epidemiologic Data, 2*, 95-100.

Sen, A., Heredia, N., Senut, M., Land, S., Hollocher, K., Lu, X., . . . Ruden, D. M. (2015). Multigenerational epigenetic inheritance in humans: DNA methylation changes associated with maternal exposure to lead can be transmitted to the grandchildren. *Scientific Reports, 5*, 14466.

Serpeloni, F., Radtke, K., de Assis, S. G., Henning, F., Nätt, D., & Elbert, T. (2017). Grandmaternal stress during pregnancy and DNA methylation of the third generation: An epigenome-wide association study. *Translational Psychiatry, 7*(8), e1202.

Shah, S., Bonder, M. J., Marioni, R. E., Zhu, Z., McRae, A. F., Zhernakova, A., . . . Visscher, P. M. (2015). Improving phenotypic prediction by combining genetic and epigenetic associations. *American Journal of Human Genetics, 97*(1), 75-85. doi:10.1016/j.ajhg.2015.05.014 [doi]

Shanthikumar, S., Neeland, M. R., Maksimovic, J., Ranganathan, S. C., & Saffery, R. (2020). DNA methylation biomarkers of future health outcomes in children. *Molecular and Cellular Pediatrics, 7*(1), 1-11.

Sharp, G. C., Arathimos, R., Reese, S. E., Page, C. M., Felix, J., Kupers, L. K., . . . Zuccolo, L. (2018). Maternal alcohol consumption and offspring DNA methylation: Findings from six general population-based birth cohorts. *Epigenomics, 10*(1), 27-42. doi:10.2217/epi-2017-0095 [doi]

Sharp, G. C., Lawlor, D. A., Richmond, R. C., Fraser, A., Simpkin, A., Suderman, M., . . . Relton, C. L. (2015). Maternal pre-pregnancy BMI and gestational weight gain, offspring DNA methylation and later offspring adiposity: Findings from the avon longitudinal study of parents and children. *International Journal of Epidemiology, 44*(4), 1288-1304. doi:10.1093/ije/dyv042 [doi]

Shenassa, E. D., Papandonatos, G. D., Rogers, M. L., & Buka, S. L. (2015). Elevated risk of nicotine dependence among sib-pairs discordant for maternal smoking during pregnancy: Evidence from a 40-year longitudinal study. *Epidemiology (Cambridge, Mass.), 26*(3), 441.

Shipton, D., Tappin, D. M., Vadiveloo, T., Crossley, J. A., Aitken, D. A., & Chalmers, J. (2009). Reliability of self reported smoking status by pregnant women for estimating smoking prevalence: A retrospective, cross sectional study. *Bmj, 339*, b4347.

Siahpush. (2006). The association of smoking with perception of income inequality, relative material well-being, and social capital. *Social Science Medicine, 63*(11)

Silva, T. C., Coetzee, S. G., Gull, N., Yao, L., Hazelett, D. J., Noushmehr, H., . . . Berman, B. P. (2019). ELMER v.2: An R/bioconductor package to reconstruct gene regulatory networks from DNA methylation and transcriptome profiles. *Bioinformatics (Oxford, England), 35*(11), 1974-1977. doi:10.1093/bioinformatics/bty902 [doi]

Silva-Martínez, G. A., Zaina, S., & Lund, G. (2017). Array probe density and pathobiological relevant CpG calling bias in human disease and physiological DNA methylation profiling. *Briefings in Functional Genomics, 17*(1), 42-48.

Silveira, P. P., Pokhvisneva, I., Parent, C., Cai, S., Rema, A. S. S., Broekman, B. F., . . . Meaney, M. J. (2017). Cumulative prenatal exposure to adversity reveals associations with a broad range of neurodevelopmental outcomes that are moderated by a novel, biologically informed polygenetic score based on the serotonin transporter solute carrier family C6, member 4 (SLC6A4) gene expression. *Development and Psychopathology, 29*(5), 1601-1617.

Simeone, P., & Alberti, S. (2014). Epigenetic heredity of human height. *Physiological Reports, 2*(6), e12047.

Smith, M. R., Yevoo, P., Sadahiro, M., Readhead, B., Kidd, B., Dudley, J. T., & Morishita, H. (2020). Systematic analysis of environmental chemicals that dysregulate critical period

plasticity-related gene expression reveals common pathways that mimic immune response to pathogen. *Neural Plasticity, 2020*, 1673897. doi:10.1155/2020/1673897 [doi]

Smith, M. T., McHale, C. M., & de la Rosa, R. (2019). Using exposomics to assess cumulative risks from multiple environmental stressors. *Unraveling the exposome* (pp. 3-22) Springer.

Smith, T. F., Maccani, M. A., & Knopik, V.S. (2012). Maternal smoking during pregnancy and offspring health outcomes: The role of epigenetic research in informing legal policy and practice. *Hastings LJ, 64*, 1619.

Smith, Z. D., Chan, M. M., Mikkelsen, T. S., Gu, H., Gnirke, A., Regev, A., & Meissner, A. (2012). A unique regulatory phase of DNA methylation in the early mammalian embryo. *Nature, 484*(7394), 339.

Sofer, T., Maity, A., Coull, B., Baccarelli, A., Schwartz, J., & Lin, X. (2012). Multivariate gene selection and testing in studying the exposure effects on a gene set. *Statistics in Biosciences, 4*(2), 319-338. doi:10.1007/s12561-012-9072-7

Solomon, O., MacIsaac, J., Quach, H., Tindula, G., Kobor, M. S., Huen, K., . . . Holland, N. (2018). Comparison of DNA methylation measured by illumina 450K and EPIC BeadChips in blood of newborns and 14-year-old children. *Epigenetics, 13*(6), 655-664.

Stergiakouli, E., & Thapar, A. (2010). Fitting the pieces together: Current research on the genetic basis of attention-deficit/hyperactivity disorder (ADHD). *Neuropsychiatric Disease and Treatment, 6*, 551-560. doi:10.2147/NDT.S11322

Strecher, V., Wang, C., Derry, H., Wildenhaus, K., & Johnson, C. (2002). Tailored interventions for multiple risk behaviors. *Health Education Research, 17*(5), 619-626.

Stringhini, S., Polidoro, S., Sacerdote, C., Kelly, R. S., Van Veldhoven, K., Agnoli, C., . . . Panico, S. (2015). Life-course socioeconomic status and DNA methylation of genes regulating inflammation. *International Journal of Epidemiology, 44*(4), 1320-1330.

Stroud, L. R., Papandonatos, G. D., Shenassa, E., Rodriguez, D., Niaura, R., LeWinn, K. Z., . . . Buka, S. L. (2014). Prenatal glucocorticoids and maternal smoking during pregnancy independently program adult nicotine dependence in daughters: A 40-year prospective study. *Biological Psychiatry, 75*(1), 47-55.

Stunnenberg, H. G., Abrignani, S., Adams, D., de Almeida, M., Altucci, L., Amin, V., . . . Arima, T. (2016). The international human epigenome consortium: A blueprint for scientific collaboration and discovery. *Cell, 167*(5), 1145-1149.

Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., . . . Mesirov, J. P. (2005). Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences of the United States of America, 102*(43), 15545-15550. doi:0506580102 [pii]

Sugden, K., Hannon, E. J., Arseneault, L., Belsky, D. W., Broadbent, J. M., Corcoran, D. L., . . . Caspi, A. (2019). Establishing a generalized polyepigenetic biomarker for tobacco smoking. *Translational Psychiatry, 9*(1), 92-9. doi:10.1038/s41398-019-0430-9 [doi]

Susser, E. S., & Lin, S. P. (1992). Schizophrenia after prenatal exposure to the dutch hunger winter of 1944-1945. *Archives of General Psychiatry, 49*(12), 983-988.

Suter, M. A., Anders, A. M., & Aagaard, K. M. (2013). Maternal smoking as a model for environmental epigenetic changes affecting birthweight and fetal programming. *Molecular Human Reproduction, 19*(1), 1-6. doi:10.1093/molehr/gas050 [doi]

Suter, M., Abramovici, A., & Aagaard-Tillery, K. (2010). Genetic and epigenetic influences associated with intrauterine growth restriction due to in utero tobacco exposure. *Pediatric Endocrinology Reviews : PER, 8*(2), 94-102.

Suter, M., Ma, J., Harris, A., Patterson, L., Brown, K. A., Shope, C., . . . Aagaard-Tillery, K. M. (2011). Maternal tobacco use modestly alters correlated epigenome-wide placental DNA methylation and gene expression. *Epigenetics : Official Journal of the DNA Methylation Society, 6*(11), 1284-1294. doi:10.4161/epi.6.11.17819; 10.4161/epi.6.11.17819

Suzuki, K., Kondo, N., Sato, M., Tanaka, T., Ando, D., & Yamagata, Z. (2011). Gender differences in the association between maternal smoking during pregnancy and childhood growth trajectories: Multilevel analysis. *International Journal of Obesity, 35*(1), 53.

Suzuki, M., Liao, W., Wos, F., Johnston, A. D., DeGrazia, J., Ishii, J., . . . Greally, J. M. (2018). Whole-genome bisulfite sequencing with improved accuracy and cost. *Genome Research, 28*(9), 1364-1371.

Taal, H. R., de Jonge, L. L., van Osch-Gevers, L., Steegers, E. A., Hofman, A., Helbing, W. A., . . . Jaddoe, V. W. (2013). Parental smoking during pregnancy and cardiovascular structures and function in childhood: The generation R study. *International Journal of Epidemiology, 42*(5), 1371-1380. doi:10.1093/ije/dyt178; 10.1093/ije/dyt178

Tajbakhsh, J. (2011). DNA methylation topology: Potential of a chromatin landmark for epigenetic drug toxicology. *Epigenomics, 3*(6), 761-770. doi:10.2217/epi.11.101 [doi]

Tak, Y. G., & Farnham, P. J. (2015). Making sense of GWAS: Using epigenomics and genome engineering to understand the functional relevance of SNPs in non-coding regions of the human genome. *Epigenetics & Chromatin, 8*, 57-4. eCollection 2015. doi:10.1186/s13072-015-0050-4 [doi]

Talens, R. P., Christensen, K., Putter, H., Willemsen, G., Christiansen, L., Kremer, D., . . . Heijmans, B. T. (2012). Epigenetic variation during the adult lifespan: Cross-sectional and longitudinal data on monozygotic twin pairs. *Aging Cell, 11*(4), 694-703.

Taylor, C. M., Kordas, K., Golding, J., & Emond, A. M. (2017). Data relating to prenatal lead exposure and child IQ at 4 and 8 years old in the avon longitudinal study of parents and children. *Neurotoxicology, 62*, 224-230. doi:S0161-813X(17)30155-9 [pii]

Taylor, J. M., & Yu, M. (2002). Bias and efficiency loss due to categorizing an explanatory variable. *Journal of Multivariate Analysis, 83*(1), 248-263.

Taylor, R. M., Smith, R., Collins, C. E., Mossman, D., Wong-Brown, M. W., Chan, E., . . . Drysdale, K. (2020). Global DNA methylation and cognitive and behavioral outcomes at 4 years of age: A cross-sectional study. *Brain and Behavior, 10*(4), e01579.

Tchasovnikarova, I. A., & Kingston, R. E. (2018). Beyond the histone code: A physical map of chromatin states. *Molecular Cell, 69*(1), 5-7.

Tehranifar, P., Wu, H., McDonald, J. A., Jasmine, F., Santella, R. M., Gurvich, I., . . . Terry, M. B. (2018). Maternal cigarette smoking during pregnancy and offspring DNA methylation in midlife. *Epigenetics, 13*(2), 129-134.

Teschendorff, A. E., & Relton, C. L. (2018). Statistical and integrative system-level analysis of DNA methylation data. *Nature Reviews Genetics, 19*(3), 129.

Teschendorff, A., Renard, E., & Absil, P.,A. (2014). Supervised normalization of large-scale omic datasets using blind source separation. (pp. 465-497) doi:10.1007/978-3-642-55016-4_17

Timms, J. A., Relton, C. L., Sharp, G. C., Rankin, J., Strathdee, G., & McKay, J. A. (2019). Exploring a potential mechanistic role of DNA methylation in the relationship between in utero and post-natal environmental exposures and risk of childhood acute lymphoblastic leukaemia. *International Journal of Cancer, 145*(11), 2933-2943.

Tiong, K. L., & Yeang, C. H. (2019). MGSEA - a multivariate gene set enrichment analysis. *BMC Bioinformatics, 20*(1), 145-6. doi:10.1186/s12859-019-2716-6 [doi]

Tobi, E. W., Slieker, R. C., Stein, A. D., Suchiman, H. E. D., Slagboom, P. E., Van Zwet, E. W., . . . Lumey, L. H. (2015). Early gestation as the critical time-window for changes in the prenatal environment to affect the adult human blood methylome. *International Journal of Epidemiology, 44*(4), 1211-1223.

Toledo-Rodriguez, M., Lotfipour, S., Leonard, G., Perron, M., Richer, L., Veillette, S., . . . Paus, T. (2010). Maternal smoking during pregnancy is associated with epigenetic modifications of the brain-derived neurotrophic factor-6 exon in adolescent offspring. *American Journal of Medical Genetics.Part B, Neuropsychiatric Genetics : The Official Publication of the International Society of Psychiatric Genetics, 153B*(7), 1350-1354. doi:10.1002/ajmg.b.31109

Touleimat, N., & Tost, J. (2012). Complete pipeline for infinium((R)) human methylation 450K BeadChip data processing using subset quantile normalization for accurate DNA methylation estimation. *Epigenomics, 4*(3), 325-341. doi:10.2217/epi.12.21 [doi]

Trajkovski, I., Lavrac, N., & Tolar, J. (2008). SEGS: Search for enriched gene sets in microarray data. *Journal of Biomedical Informatics, 41*(4), 588-601. doi:10.1016/j.jbi.2007.12.001

Tremblay, R. E. (2010). Developmental origins of disruptive behaviour problems: The 'original sin' hypothesis, epigenetics and their consequences for prevention. *Journal of Child Psychology and Psychiatry, and Allied Disciplines, 51*(4), 341-367. doi:10.1111/j.1469-7610.2010.02211.x

Tritschler, S., Theis, F. J., Lickert, H., & Bottcher, A. (2017). Systematic single-cell analysis provides new insights into heterogeneity and plasticity of the pancreas. *Molecular Metabolism, 6*(9), 974-990. doi:10.1016/j.molmet.2017.06.021 [doi]

Trivedi, R., & Asch, S. M. (2019). Can we improve patient adherence by harnessing social forces? *Journal of General Internal Medicine, 34*(6), 785-786.

Tsai, P. C., Spector, T. D., & Bell, J. T. (2012). Using epigenome-wide association scans of DNA methylation in age-related complex human traits. *Epigenomics, 4*(5), 511-526. doi:10.2217/epi.12.45 [doi]

Turner, M. C., Nieuwenhuijsen, M., Anderson, K., Balshaw, D., Cui, Y., Dunton, G., . . . Jerrett, M. (2017). Assessing the exposome with external measures: Commentary on the state of the science and research recommendations. *Annual Review of Public Health, 38*, 215-239.

Uhler, C., & Shivashankar, G. V. (2017). Regulation of genome organization and gene expression by nuclear mechanotransduction. *Nature Reviews Molecular Cell Biology, 18*(12), 717-727.

Unternaehrer, E., Luers, P., Mill, J., Dempster, E., Meyer, A. H., Staehli, S., . . . Meinlschmidt, G. (2012). Dynamic changes in DNA methylation of stress-associated genes (OXTR, BDNF) after acute psychosocial stress. *Translational Psychiatry, 2*(8), e150.

Vacha-Haase, T., & Thompson, B. (2004). How to estimate and interpret various effect sizes. *Journal of Counseling Psychology, 51*(4), 473-481. doi:10.1037/0022-0167.51.4.473

Valeri, L., Reese, S. L., Zhao, S., Page, C. M., Nystad, W., Coull, B. A., & London, S. J. (2017). Misclassified exposure in epigenetic mediation analyses. does DNA methylation mediate effects of smoking on birthweight? *Epigenomics, 9*(3), 253-265.

Van Belle, G., Fisher, L. D., Heagerty, P. J., & Lumley, T. (2004). *Biostatistics: A methodology for the health sciences* John Wiley & Sons.

van der Meer, D., Hoekstra, P. J., Van Donkelaar, M., Bralten, J., Oosterlaan, J., Heslenfeld, D., . . . Hartman, C. A. (2017). Predicting attention-deficit/hyperactivity disorder severity from psychosocial stress and stress-response genes: A random forest regression approach. *Translational Psychiatry, 7*(6), e1145.

Vattikuti, S., Guo, J., & Chow, C. C. (2012). Heritability and genetic correlations explained by common SNPs for metabolic syndrome traits. *PLoS Genetics, 8*(3), e1002637.

Vazsonyi, A. T., & Belliston, L. M. (2006). The cultural and developmental significance of parenting processes in adolescent anxiety and depression symptoms. *Journal of Youth and Adolescence, 35*(4), 491.

Villandré, L., Hutcheon, J. A., Trejo, M. E. P., Abenhaim, H., Jacobsen, G., & Platt, R. W. (2011). Modeling fetal weight for gestational age: A comparison of a flexible multi-level spline-based model with other approaches. *The International Journal of Biostatistics, 7*(1)

Villar, J., Cheikh Ismail, L., Victora, C. G., Ohuma, E. O., Bertino, E., Altman, D. G., . . . International Fetal and Newborn Growth Consortium for the 21st Century (INTERGROWTH-21st). (2014). International standards for newborn weight, length, and head circumference by gestational age and sex: The newborn cross-sectional study of the INTERGROWTH-21st project. *Lancet (London, England), 384*(9946), 857-868. doi:10.1016/S0140-6736(14)60932-6 [doi]

Vives-Usano, M., Hernandez-Ferrer, C., Maitre, L., Ruiz-Arenas, C., Andrusaityte, S., Borràs, E., . . . Coen, M. (2020). In utero and childhood exposure to tobacco smoke and multi-layer molecular signatures in children. *BMC Medicine, 18*(1), 1-19.

Waddington, C. H. (1957). The strategy of the genes: A discussion of some aspects of theoretical biology.

Wade, M., Madigan, S., Akbari, E., & Jenkins, J. M. (2015). Cumulative biomedical risk and social cognition in the second year of life: Prediction and moderation by responsive parenting. *Frontiers in Psychology, 6*, 354.

Wakschlag, L. S., Pickett, K. E., Cook Jr, E., Benowitz, N. L., & Leventhal, B. L. (2002). Maternal smoking during pregnancy and severe antisocial behavior in offspring: A review. *American Journal of Public Health, 92*(6), 966-974.

Wang, Y., Miller, D. J., & Clarke, R. (2008). Approaches to working in high-dimensional data spaces: Gene expression microarrays. *British Journal of Cancer, 98*(6), 1023-1028. doi:10.1038/sj.bjc.6604207

Wang, K. (2019). Various relationships between dna methylation and gene expression in different tissues and ages. *European Neuropsychopharmacology, 29*, S821.

Wang, X. M., Tian, F. Y., Fan, L. J., Xie, C. B., Niu, Z. Z., & Chen, W. Q. (2019). Comparison of DNA methylation profiles associated with spontaneous preterm birth in placenta and cord blood. *BMC Medical Genomics, 12*(1), 1-3. doi:10.1186/s12920-018-0466-3 [doi]

Wang, Y., Zhou, P., Wang, L., Li, Z., Zhang, Y., & Zhang, Y. (2012). Correlation between DNase I hypersensitive site distribution and gene expression in HeLa S3 cells. *PloS One, 7*(8)

Weaver, I. C., Korgan, A. C., Lee, K., Wheeler, R. V., Hundert, A. S., & Goguen, D. (2017). Stress and the emerging roles of chromatin remodeling in signal integration and stable transmission of reversible phenotypes. *Frontiers in Behavioral Neuroscience, 11*, 41. doi:10.3389/fnbeh.2017.00041 [doi]

Wehby, G. L., Fletcher, J. M., Lehrer, S. F., Moreno, L. M., Murray, J. C., Wilcox, A., & Lie, R. T. (2011). A genetic instrumental variables analysis of the effects of prenatal smoking on birth weight: Evidence from two samples. *Biodemography and Social Biology, 57*(1), 3-32.

Welter, D., MacArthur, J., Morales, J., Burdett, T., Hall, P., Junkins, H., . . . Hindorff, L. (2013). The NHGRI GWAS catalog, a curated resource of SNP-trait associations. *Nucleic Acids Research, 42*(D1), D1001-D1006.

Weniger, M., Engelmann, J. C., & Schultz, J. (2007). Genome expression pathway analysis tool--analysis and visualization of microarray gene expression data under genomic, proteomic and metabolic context. *BMC Bioinformatics, 8*, 179. doi:10.1186/1471-2105-8-179

Werler, M. M., Pober, & Holmes, L. B. (1985). Smoking and pregnancy. *Teratology, 32*(3), 473-481.

West-Eberhard, M. J. (2003). Developmental plasticity and evolution. (pp. 20). New York: Oxford University Press.

West-Eberhard, M. J. (2003). *Developmental plasticity and evolution* Oxford University Press.

Wiklund, P., Karhunen, V., Richmond, R. C., Rodriguez, A., De Silva, M., Wielscher, M., . . . Järvelin, M. (2018). DNA methylation links prenatal smoking exposure to later life health outcomes in offspring. *bioRxiv,* 428896. doi:10.1101/428896

Wild, C. P. (2005). No title. *Complementing the Genome with an "exposome": The Outstanding Challenge of Environmental Exposure Measurement in Molecular Epidemiology,*

Williams, G. M., O'Callaghan, M., Najman, J. M., Bor, W., Andersen, M. J., Richards, D., & Chinlyn, U. (1998). Maternal cigarette smoking and child psychiatric morbidity: A longitudinal study. *Pediatrics, 102*(1), e11.

Williamson, A. K., Zhu, Z., & Yuan, Z. (2018). Epigenetic mechanisms behind cellular sensitivity to DNA damage. *Cell Stress, 2*(7), 176.

Winchester, S. B., Sullivan, M. C., Roberts, M. B., Bryce, C. I., & Granger, D. A. (2018). Long-term effects of prematurity, cumulative medical risk, and proximal and distal social forces on individual differences in diurnal cortisol at young adulthood. *Biological Research for Nursing, 20*(1), 5-15.

Winham, S. J., Colby, C. L., Freimuth, R. R., Wang, X., De Andrade, M., Huebner, M., & Biernacka, J. M. (2012). SNP interaction detection with random forests in high-dimensional genetic data. *BMC Bioinformatics, 13*(1), 164.

Wold, S., Martens, H., & Wold, H. (1983). The multivariate calibration problem in chemistry solved by the PLS method. *Matrix pencils* (pp. 286-293) Springer.

Wolpert, D. H., & Macready, W. G. (1997). No free lunch theorems for optimization. *IEEE Transactions on Evolutionary Computation, 1*(1), 67-82.

Wright, M. N., Ziegler, A., & König, I. R. (2016). Do little interactions get lost in dark random forests? *BMC Bioinformatics, 17*(1), 145.

Wu, C. C., Hsu, T. Y., Chang, J. C., Ou, C. Y., Kuo, H. C., Liu, C. A., . . . Yang, K. D. (2019). Paternal tobacco smoke correlated to offspring asthma and prenatal epigenetic programming. *Frontiers in Genetics, 10*, 471. doi:10.3389/fgene.2019.00471 [doi]

Wu, Y., Qi, T., Wang, H., Zhang, F., Zheng, Z., Phillips-Cremins, J. E., . . . Yang, J. (2019). Promoter-anchored chromatin interactions predicted from genetic analysis of epigenomic data. *bioRxiv,* 580993. doi:10.1101/580993

Wu, Y., Qi, T., Wang, H., Zhang, F., Zheng, Z., Phillips-Cremins, J. E., . . . Zeng, J. (2020). Promoter-anchored chromatin interactions predicted from genetic analysis of epigenomic data. *Nature Communications, 11*(1), 1-12.

Xu, C., & Corces, V. G. (2018). Nascent DNA methylome mapping reveals inheritance of hemimethylation at CTCF/cohesin sites. *Science (New York, N.Y.), 359*(6380), 1166-1170. doi:10.1126/science.aan5480 [doi]

Xu, H., Zhang, S., Yi, X., Plewczynski, D., & Li, M. J. (2020). Exploring 3D chromatin contacts in gene regulation: The evolution of approaches for the identification of functional enhancer-promoter interaction. *Computational and Structural Biotechnology Journal,*

Xu, X., Su, S., Barnes, V. A., De Miguel, C., Pollock, J., Ownby, D., . . . Wang, X. (2013). A genome-wide methylation study on obesity: Differential variability and differential methylation. *Epigenetics, 8*(5), 522-533.

Xue, Y., Yang, Y., Tian, H., Quan, H., Liu, S., Zhang, L., & Gao, Y. Q. (2020). Domain segregated 3D chromatin structure and segmented DNA methylation in carcinogenesis. *bioRxiv,*

Yang, C., Li, C., Wang, Q., Chung, D., & Zhao, H. (2015). Implications of pleiotropy: Challenges and opportunities for mining big data in biomedicine. *Frontiers in Genetics, 6*, 229.

Yang, X. W. (2016). Life and death rest on a bivalent chromatin state. *Nature Neuroscience, 19*(10), 1271-1273. doi:10.1038/nn.4396 [doi]

Yu, C., Woo, H. J., Yu, X., Oyama, T., Wallqvist, A., & Reifman, J. (2017). A strategy for evaluating pathway analysis methods. *BMC Bioinformatics, 18*(1), 1-11.

Zaren, B., Lindmark, G., & Bakketeig, L. (2000). Maternal smoking affects fetal growth more in the male fetus. *Paediatric and Perinatal Epidemiology, 14*(2), 118-126.

Zeng, Z., Meyer, K. F., Lkhagvadorj, K., Kooistra, W., Reinders-Luinge, M., Xu, X., . . . Hylkema, M. N. (2020). Prenatal smoke effect on mouse offspring Igf1 promoter methylation from fetal stage to adulthood is organ and sex specific. *American Journal of Physiology-Lung Cellular and Molecular Physiology, 318*(3), L549-L561.

Zhang, B., Hong, X., Ji, H., Tang, W., Kimmel, M., Ji, Y., . . . Wang, X. (2018). Maternal smoking during pregnancy and cord blood DNA methylation: New insight on sex differences and effect modification by maternal folate levels. *Epigenetics, 13*(5), 505-518.

Zhang, L., Xie, W. J., Liu, S., Meng, L., Gu, C., & Gao, Y. Q. (2017). DNA methylation landscape reflects the spatial organization of chromatin in different cells. *Biophysical Journal, 113*(7), 1395-1404. doi:S0006-3495(17)30911-6 [pii]

Zhang, W., Voloudakis, G., Rajagopal, V. M., Readhead, B., Dudley, J. T., Schadt, E. E., . . . Roussos, P. (2019). Integrative transcriptome imputation reveals tissue-specific and shared biological mechanisms mediating susceptibility to complex traits. *Nature Communications, 10*(1), 3834-7. doi:10.1038/s41467-019-11874-7 [doi]

Zhang, Y., Qi, G., Park, J., & Chatterjee, N. (2018). Estimation of complex effect-size distributions using summary-level statistics from genome-wide association studies across 32 complex traits. *Nature Genetics, 50*(9), 1318.

Zhi, D., Aslibekyan, S., Irvin, M. R., Claas, S. A., Borecki, I. B., Ordovas, J. M., . . . Arnett, D. K. (2013). SNPs located at CpG sites modulate genome-epigenome interaction. *Epigenetics, 8*(8), 802-806. doi:10.4161/epi.25501 [doi]

Zhong, H., Kim, S., Zhi, D., & Cui, X. (2019). Predicting gene expression using DNA methylation in three human populations. *PeerJ, 7*, e6757.

Zhou, D., Li, Z., Yu, D., Wan, L., Zhu, Y., Lai, M., & Zhang, D. (2015). Polymorphisms involving gain or loss of CpG sites are significantly enriched in trait-associated SNPs. *Oncotarget, 6*(37), 39995-40004. doi:10.18632/oncotarget.5650 [doi]

Zhou, F., Wang, W., Shen, C., Li, H., Zuo, X., Zheng, X., . . . Chen, M. (2016). Epigenome-wide association analysis identified nine skin DNA methylation loci for psoriasis. *Journal of Investigative Dermatology, 136*(4), 779-787.

Zhou, W., Laird, P. W., & Shen, H. (2017). Comprehensive characterization, annotation and innovative use of infinium DNA methylation BeadChip probes. *Nucleic Acids Research, 45*(4), e22. doi:10.1093/nar/gkw967 [doi]

Zhu, Y., Chen, Z., Zhang, K., Wang, M., Medovoy, D., Whitaker, J. W., . . . Wang, W. (2016). Constructing 3D interaction maps from 1D epigenomes. *Nature Communications, 7*, 10812. doi:10.1038/ncomms10812 [doi]

Appendix A Random Forest tuning

Table 36: Performance of random forest models for different tuning parameters. (Example using outcome Height (z-score) at age 10.)

mtry	*mtry* (numeric)	Split train:test	*ntree*	Train data % var explained	Test Data RMSE	Test Data % var explained
p	10	70:30	500	64.78	0.3	63.32
p/3 (default)	**3**	**70:30**	**500**	**65.13**	**0.29**	**64.59**
p/3 (default)	3	70:30	200	64.48	0.29	63.98
p	10	50:50	500	59.4	0.3	68.4
p/3 (default)	3	50:50	500	60.2	0.31	67.02
p/3 (default)	3	50:50	200	58.8	0.3	67.45
p	10	80:20	500	64.4	0.33	65.1
p/3 (default)	3	80:20	500	64.99	0.32	65.8
p/3 (default)	3	80:20	200	64.3	0.33	64.7

Appendix B Comparison of maternal baseline characteristics in ALSPAC mothers included and excluded from ARIES

Characteristic	ALSPAC mothers included in ARIES	ALSPAC mothers not included in ARIES
	Mean (SD) , Median (IQR) or %*	Mean (SD) , Median (IQR) or %*
Reported pre-pregnancy BMI	(n=944)	(n=10633)
	22.8 (3.7)	22.9 (3.8)
Pregnancy stage-specific GWG	(n=971)	(n=11512)
0 to 18 weeks (kg/wk)	0.3 (0.2)	0.3 (0.2)
18 to 28 weeks (kg/wk)	0.5 (0.2)	0.5 (0.2)
28 weeks to delivery (kg/wk)	0.5 (0.2)	0.5 (0.2)
Total GWG (kg)	(n=939)	(n=11486)
	12.6 (4.4)	12.5 (4.8)
IoM categories of GWG	(n=881)	(n=9401)
Below recommended GWG	322	3146
Recommended GWG	334	3671
Over recommended GWG	225	2584
Offspring sex	(n=1018)	(n=13041)
Male	48.8%	51.9%
Female	51.2%	48.1%
Gestational age at delivery (weeks)	(n=1018)	(n=13614)
	40 (39 – 41)	40 (39 – 41)
Parity	(n=989)	(n=12002)
Nulliparous	46.4%	44.6%
Multiparous	53.6%	55.5%
Age at delivery (years)	(n=986)	(n=10838)
	29.6 (4.4)	28.2 (4.8)
Occupation	(n=901)	(n=9188)
Manual occupation	14.0%	20.5%
Non-manual occupation	86.0%	79.5%
Smoking status	(n=1006)	(n=12166)
Never before or during pregnancy	86.7%	73.3%
Before pregnancy or during 1st trimester only	3.6%	7.3%
Throughout pregnancy	9.7%	19.4%

*SD = standard deviation; IQR = interquartile range

Table excerpt from (Sharp *et al.*, 2015). IoM categories of GWG: recommendations for gestational weight gain: Mothers were categorized as having gained the recommended, less than recommended and in excess of the recommended weight during gestation depending on their pre-pregnancy BMI. IoM: Institute of Medicine.

Appendix C Additional information: Cord blood and cell type composition

Comparison of models using either reference (using R package meffil) versus reference-free (using R package reFACTor) estimates of cell composition. Shown are examples using models for height outcomes at various ages. Other outcomes available upon request.

Cord results

Correlation table between DNAm components (cord blood) and reFACTor estimated cell count. We trialed both n=6 and n=7, the latter because newborns have an additional DNA-containing blood cell, nucleated red blood cells compared to the typical blood cell population of non-infants. There are only minor differences between the two. For Age 7 and Age 15 DNAm, we used n=6.

N = 6

row	column	cor	p
Comp8	Comp11	-0.07	0.026
Comp5	Comp12	0.14	0.000
Comp5	Comp16	0.07	0.047
Comp12	Comp16	0.15	0.000
Comp15	Comp19	0.07	0.032
Comp1	PC1	0.12	0.000
Comp2	PC1	-0.19	0.000
Comp3	PC1	0.22	0.000
Comp10	PC1	0.09	0.006
Comp12	PC1	0.08	0.017
Comp1	PC2	-0.20	0.000
Comp2	PC2	0.09	0.008
Comp3	PC2	0.26	0.000
Comp9	PC2	0.07	0.037
Comp16	PC2	-0.08	0.023
Comp17	PC2	-0.07	0.034
Comp6	PC3	0.10	0.004

N = 7

row	column	cor	p
Comp8	Comp11	-0.07	0.026
Comp5	Comp12	0.14	0.000
Comp5	Comp16	0.07	0.047
Comp12	Comp16	0.15	0.000
Comp15	Comp19	0.07	0.032
Comp1	PC1	0.11	0.001
Comp2	PC1	-0.19	0.000
Comp3	PC1	0.23	0.000
Comp10	PC1	0.09	0.006
Comp12	PC1	0.08	0.015
Comp1	PC2	-0.21	0.000
Comp2	PC2	0.09	0.005
Comp3	PC2	0.25	0.000
Comp9	PC2	0.07	0.035
Comp16	PC2	-0.08	0.022
Comp17	PC2	-0.07	0.032
Comp6	PC3	0.09	0.004

Comp7	PC3	0.21	0.000		Comp7	PC3	0.21	0.000
Comp10	PC4	0.09	0.010		Comp10	PC4	-0.08	0.016
Comp13	PC4	-0.22	0.000		Comp13	PC4	0.21	0.000
Comp15	PC4	0.17	0.000		Comp15	PC4	-0.17	0.000
Comp17	PC4	-0.08	0.019		Comp17	PC4	0.08	0.015
Comp18	PC4	-0.08	0.011		Comp18	PC4	0.09	0.008
Comp19	PC4	0.08	0.020		Comp19	PC4	-0.08	0.012
Comp4	PC5	-0.09	0.005		Comp4	PC5	0.09	0.005
Comp7	PC5	0.08	0.013		Comp7	PC5	-0.08	0.013
Comp9	PC5	-0.11	0.001		Comp9	PC5	0.10	0.003
Comp10	PC5	0.08	0.016		Comp10	PC5	-0.08	0.015
Comp12	PC5	0.07	0.025		Comp12	PC5	-0.07	0.034
Comp13	PC5	-0.12	0.000		Comp13	PC5	0.13	0.000
Comp15	PC5	-0.07	0.023		Comp15	PC5	0.07	0.028
Comp17	PC5	0.08	0.017		Comp17	PC5	-0.08	0.017
Comp19	PC5	-0.16	0.000		Comp19	PC5	0.16	0.000
Comp7	PC6	0.07	0.034		Comp7	PC6	-0.07	0.042
Comp14	PC6	0.07	0.031		Comp16	PC6	-0.10	0.003
Comp15	PC6	0.07	0.044		Comp3	PC7	0.09	0.008
Comp16	PC6	0.08	0.010		Comp14	PC7	0.12	0.000

Table below: random forest model performance metrics for anthropometric outcomes at various ages - Cord DNAm components with ReFACTor estimated cell counts (k = 6). Compared to values for same model using the meffil estimated cell counts (Table 26), values are similar.

Outcome	Training.MSE	Training.R.squared	Test.MSE	Test.R.squared
weight_7_z	0.5	0.5	0.5	0.5
weight_8_z	0.6	0.4	0.5	0.4
weight_9_z	0.7	0.3	0.6	0.3
weight_10_z	0.7	0.3	0.6	0.3
weight_11_z	0.7	0.3	0.7	0.3
weight_13_z	0.7	0.2	0.7	0.3
height_7_z	0.2	0.8	0.2	0.8
height_8_z	0.2	0.8	0.2	0.7
height_9_z	0.3	0.6	0.3	0.7
height_10_z	0.3	0.7	0.4	0.6
height_11_z	0.4	0.6	0.4	0.6
height_13_z	0.4	0.6	0.5	0.5
bone_9_z	0.5	0.4	0.7	0.3
bone_11_z	0.7	0.3	0.6	0.3
bone_13_z	0.8	0.2	0.7	0.3
fat_9_z	0.7	0.2	0.8	0.1
fat_11_z	0.8	0.1	0.9	0.1
fat_13_z	0.8	0.1	0.8	0.1
lean_9_z	0.6	0.4	0.6	0.3
lean_11_z	0.7	0.3	0.7	0.3
lean_13_z	0.8	0.2	0.8	0.2
waist_7_z	0.6	0.3	0.7	0.3
waist_9_z	0.8	0.2	0.7	0.1
waist_10_z	0.8	0.2	0.7	0.1
waist_11_z	0.7	0.2	0.9	0.1

For Age 7 and Age 15 DNAm data, we trialed two adult blood references, blood gse35069 and blood gse35069 complete, the latter of which differs by replacing granulocytes with eosinophils and neutrophils. We found that the former generated fewer negative cell composition estimates. We report results using the former reference in all analysis with meffil below.

Age 7 results

Table below: random forest model performance metrics for height outcomes at various ages – Age 7 DNAm components with meffil estimated cell counts.

Outcome	MSE	MSE.SD	R.squared	R.squared.SD
height_7_z	0.5	0.03	0.8	0.03
height_8_z	0.5	0.04	0.7	0.04
height_9_z	0.6	0.03	0.7	0.04
height_10_z	0.6	0.03	0.6	0.04
height_11_z	0.7	0.03	0.6	0.03
height_13_z	0.7	0.04	0.5	0.04

Variables selected by Boruta for height measured at various ages (model using meffil estimated cell counts):

Variables selected by outcome and age: height

$height_7_z "zhres0" "zhres1" "zhres2" "zhres3" "Comp1" "Comp2" "Comp3" "Comp7" "Comp8" "Comp9" "Comp11" "CD4T" "CD8T" "Gran"

$height_8_z "zhres0" "zhres1" "zhres2" "zhres3" "Comp1" "Comp2" "Comp3" "Comp7" "Comp9" "Comp11" "CD4T" "Gran"

$height_9_z "zhres0" "zhres1" "zhres2" "zhres3" "Comp1" "Comp2" "Comp3" "Comp7" "Comp8" "Comp9" "CD4T" "CD8T" "Gran" "Mono"

$height_10_z "zhres0" "zhres1" "zhres2" "zhres3" "Comp1" "Comp2" "Comp3" "Comp7" "Comp8" "Comp9" "Comp13" "CD4T" "CD8T" "Gran" "Mono"

$height_11_z "zhres0" "zhres1" "zhres2" "zhres3" "Comp1" "Comp2" "Comp3" "Comp7" "Comp9" "Comp11" "Comp13" "Comp18" "CD4T" "CD8T" "Gran"

$height_13_z "zhres0" "zhres1" "zhres2" "zhres3" "Comp1" "Comp2" "Comp3" "Comp7" "Comp9" "Comp10" "CD4T" "CD8T" "Gran"

Table below: random forest model performance metrics for height outcomes at various ages – Age 7 DNAm components with reFACTor estimated cell counts (k = 5).

Outcome	mtry	MSE	MSE.SD	R.squared	R.squared.SD
height_7_z	4	0.4	0.02	0.8	0.020103486402191
height_7_z	4	0.4	0.02	0.8	0.0191611638509662
height_8_z	4	0.5	0.04	0.8	0.0312217950148276
height_8_z	4	0.5	0.04	0.8	0.0311926089125626
height_9_z	4	0.5	0.03	0.7	0.0236873546587589
height_9_z	4	0.5	0.03	0.7	0.0223383089217606
height_10_z	4	0.6	0.03	0.7	0.0342588878890713
height_10_z	4	0.6	0.03	0.7	0.0340476859245589
height_11_z	4	0.6	0.04	0.6	0.0430164396355809
height_11_z	3	0.6	0.04	0.6	0.0427943895206408
height_13_z	4	0.7	0.03	0.5	0.0526437929630131
height_13_z	3	0.7	0.03	0.5	0.0522359733732524

Variables selected by Boruta for height measured at various ages (model using reFACTor estimated cell counts) from Age 7 DNAm data:

Variables selected by outcome and age: height

$height_7_z "zhres0" "zhres1" "zhres2" "zhres3" "PC1" "Comp1" "Comp2" "Comp3" "Comp7" "Comp8" "Comp9" "Comp11"

$height_8_z "zhres0" "zhres1" "zhres2" "zhres3" "PC1" "PC4" "Comp1" "Comp2" "Comp3" "Comp7" "Comp9" "Comp11"

$height_9_z "zhres0" "zhres1" "zhres2" "zhres3" "PC1" "PC4" "Comp1" "Comp2" "Comp3" "Comp7" "Comp9" "Comp11"

$height_10_z "zhres0" "zhres1" "zhres2" "zhres3" "PC1" "PC4" "Comp1" "Comp2" "Comp3" "Comp7" "Comp8" "Comp9"

$height_11_z "zhres0" "zhres1" "zhres2" "zhres3" "PC1" "PC4" "Comp1" "Comp2" "Comp3" "Comp7" "Comp11"

$height_13_z "zhres0" "zhres1" "zhres2" "zhres3" "PC1" "Comp1" "Comp2" "Comp3" "Comp7" "Comp9" "Comp10"

Age 15 results

Table below: random forest model performance metrics for height outcomes at various ages – Age 15 DNAm components with meffil estimated cell counts.

Outcome	MSE	MSE.SD	R.squared	R.squared.SD
height_7_z	0.5	0.03	0.8	0.03
height_8_z	0.5	0.03	0.7	0.03
height_9_z	0.6	0.03	0.7	0.03
height_10_z	0.6	0.03	0.6	0.03
height_11_z	0.6	0.04	0.6	0.04
height_13_z	0.7	0.03	0.6	0.04

Variables selected by Boruta for height measured at various ages (model using meffil estimated cell counts) from Age 15 DNAm data:

Variables selected by outcome and age: height

$height_7_z "zhres0" "zhres1" "zhres2" "zhres3" "Comp1" "Comp3" "Comp5" "Comp7" "Comp9" "Comp12" "CD8T" "Gran"

$height_8_z "zhres0" "zhres1" "zhres2" "zhres3" "Comp1" "Comp3" "Comp5" "Comp7" "Comp9" "Comp11" "Comp12" "Gran"

$height_9_z "zhres0" "zhres1" "zhres2" "zhres3" "Comp1" "Comp3" "Comp5" "Comp7" "Comp9" "Comp12" "Bcell" "Gran"

$height_10_z "zhres0" "zhres1" "zhres2" "zhres3" "Comp1" "Comp3" "Comp5" "Comp7" "Comp9" "Bcell" "CD8T" "Gran"

$height_11_z "zhres0" "zhres1" "zhres2" "zhres3" "Comp1" "Comp3" "Comp7" "Comp9" "Bcell" "CD8T" "Gran"

$height_13_z "zhres0" "zhres1" "zhres2" "zhres3" "Comp1" "Comp2" "Comp3" "Comp6" "Comp7" "Comp9" "Comp19" "Bcell" "Gran"

Table below: random forest model performance metrics for height outcomes at various ages – Age 15 DNAm components with reFACTor estimated cell counts (k = 5).

Outcome	mtry	MSE	MSE.SD	R.squared	R.squared.SD
height_7_z	4	0.4	0.01	0.8	0.0127964431829585
height_7_z	3	0.4	0.01	0.8	0.0135363711011232
height_8_z	4	0.5	0.03	0.8	0.0297921841803997
height_8_z	4	0.5	0.03	0.8	0.0308536933034139
height_9_z	4	0.5	0.03	0.7	0.0319020946614102
height_9_z	3	0.5	0.03	0.7	0.0317500898428896
height_10_z	4	0.6	0.04	0.7	0.0448607984768266
height_10_z	3	0.6	0.04	0.7	0.0442672136897325
height_11_z	3	0.6	0.03	0.6	0.0305048620131254
height_11_z	3	0.6	0.03	0.6	0.0304578310695841
height_13_z	4	0.7	0.03	0.6	0.0421369398469073
height_13_z	4	0.7	0.03	0.6	0.0406628598877004

Variables selected by Boruta for height measured at various ages (model using reFACTor estimated cell counts) from Age 15 DNAm data:

```
Variables selected by outcome and age: height
$height_7_z "zhres0" "zhres1" "zhres2" "zhres3" "PC2" "Comp1" "Comp3" "Comp5" "Comp7" "Comp9" "Comp12"
$height_8_z "zhres0" "zhres1" "zhres2" "zhres3" "PC1" "PC2" "Comp1" "Comp3" "Comp5" "Comp7" "Comp9" "Comp12"
$height_9_z "zhres0" "zhres1" "zhres2" "zhres3" "PC2" "Comp1" "Comp3" "Comp5" "Comp7" "Comp9" "Comp12"
$height_10_z "zhres0" "zhres1" "zhres2" "zhres3" "PC2" "Comp1" "Comp3" "Comp5" "Comp7" "Comp9"
$height_11_z "zhres0" "zhres1" "zhres2" "zhres3" "PC2" "Comp1" "Comp3" "Comp7" "Comp9"
$height_13_z "zhres0" "zhres1" "zhres2" "zhres3" "PC1" "PC2" "Comp1" "Comp2" "Comp3" "Comp6" "Comp7" "Comp9"
```

Model metric values are are similar between models using meffil (with reference) as with reFACTor (reference-free) estimated cell counts. There may be a trend to slightly smaller standard deviations in the reFACTor models for the age 7 height outcomes.

There are slight differences between Boruta selected variables. For example in Age 7 DNAm data, Component 11 is selected in models of height at age 7, 8, 9 and 11 using reFACTor estimated cell counts. Models using meffil estimated counts are similar except Component 11 is not selected for height outcome at age 9. Whether these observed shifts are due to the cell count estimation method or a degree of variability in the random forest process is unclear.

Outcome	MSE	MSE.SD	R.squared	R.squared.SD
height_7_z	0.5	0.03	0.8	0.03
height_8_z	0.5	0.03	0.7	0.03
height_9_z	0.6	0.03	0.7	0.03
height_10_z	0.6	0.03	0.6	0.03
height_11_z	0.6	0.04	0.6	0.04
height_13_z	0.7	0.03	0.6	0.04

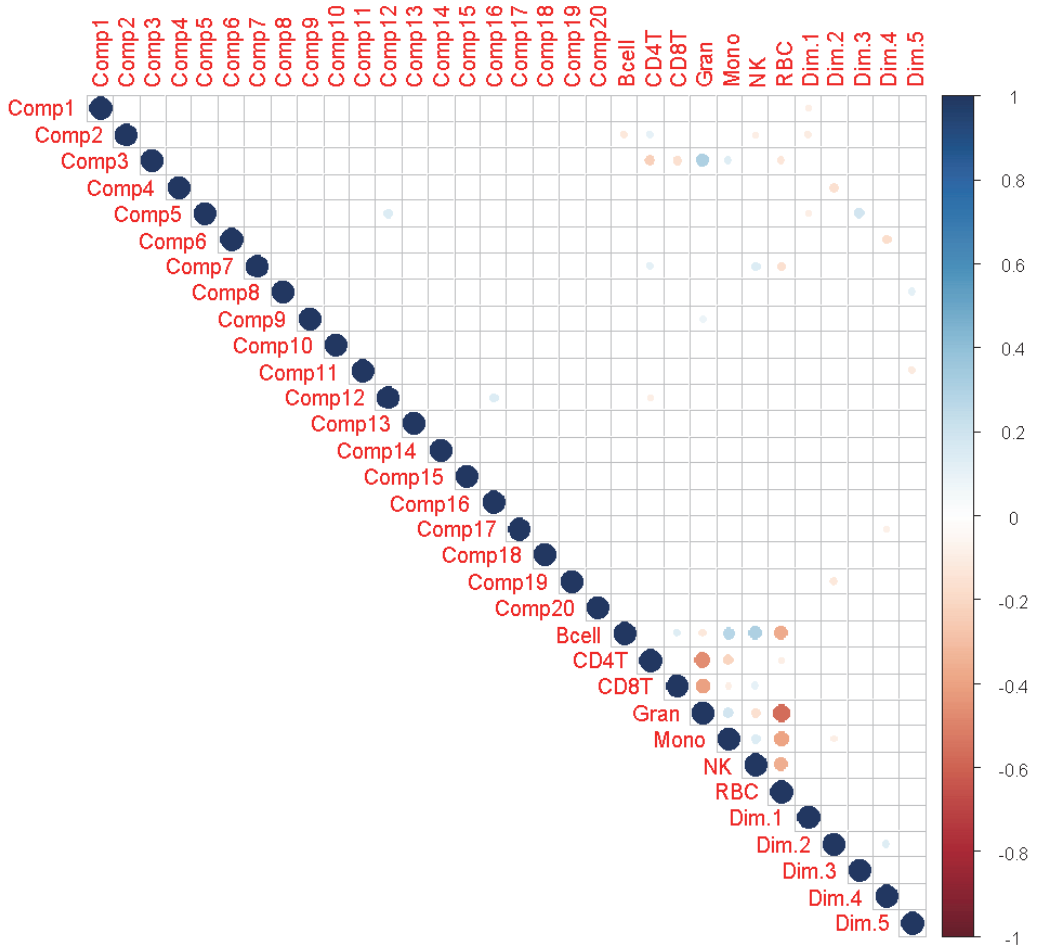


Figure 79: Correlation matrix between cord DNA methylation components (from MSP composite), MSP composite dimensions and estimated cell type composition using meffil R package and MSP composite dimensions. Pearson correlations that have p<.05 are indicated with a circle. Colour scale indicates r-value.

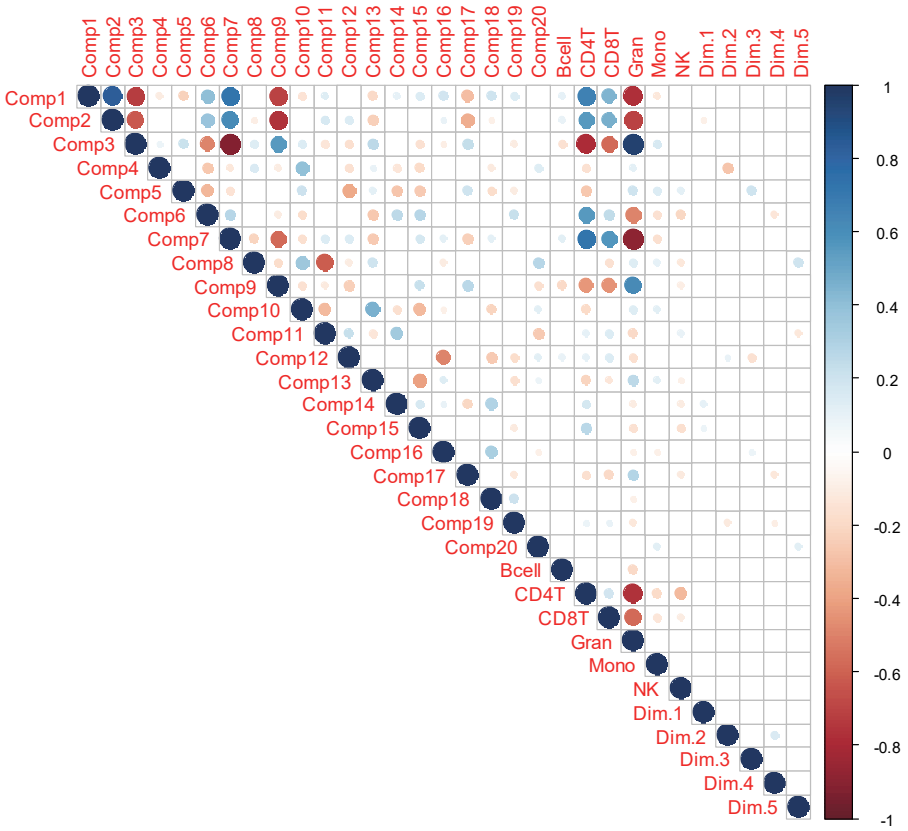


Figure 80: Correlation matrix between DNA methylation components at Age 7 (from MSP composite), MSP composite dimensions and estimated cell type composition using meffil R package (Reference: blood gse35069 - adult.) Pearson correlations that have p<.05 are indicated with a circle. Colour scale indicates r-value.

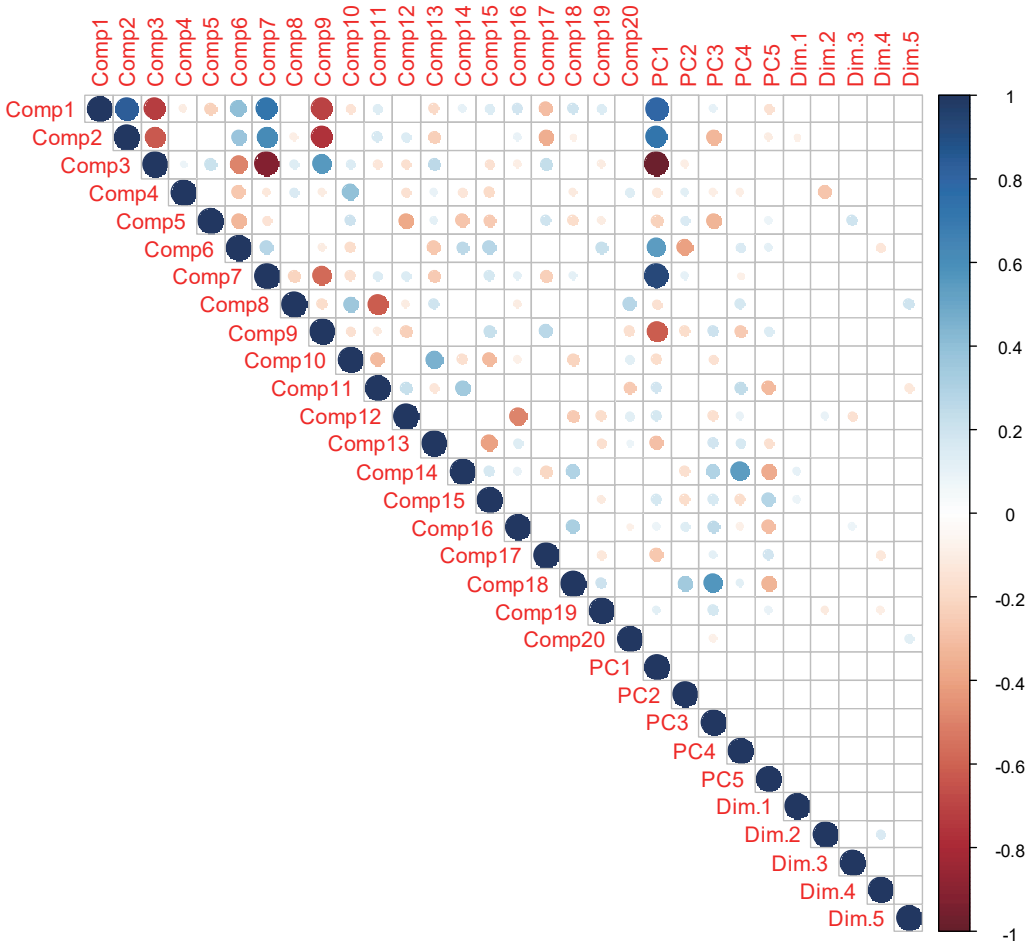


Figure 81: Correlation matrix between DNA methylation components at Age 7 (from MSP composite), MSP composite dimensions and estimated cell type composition using reFACTor R package (k = 5.) Pearson correlations that have p<.05 are indicated with a circle. Colour scale indicates r-value.

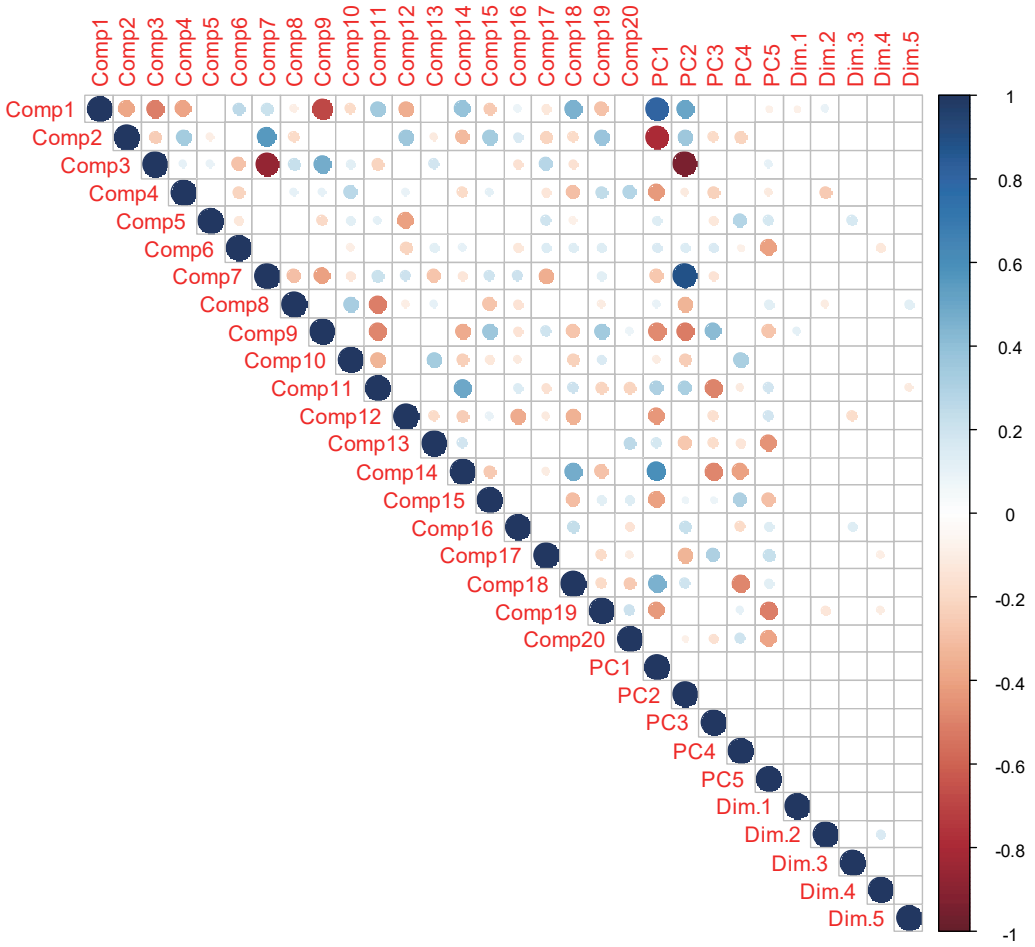


Figure 82: Correlation matrix between DNA methylation components at Age 15 (from MSP composite), MSP composite dimensions and estimated cell type composition using reFACTor R package (k = 5.) Pearson correlations that have p<.05 are indicated with a circle. Colour scale indicates r-value.

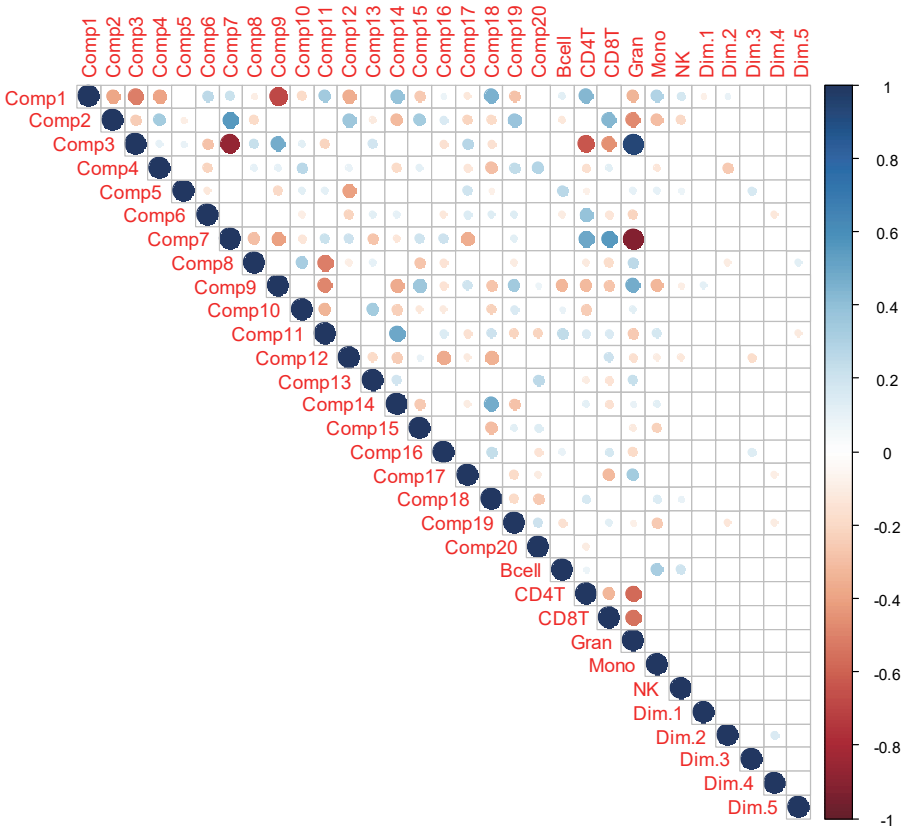
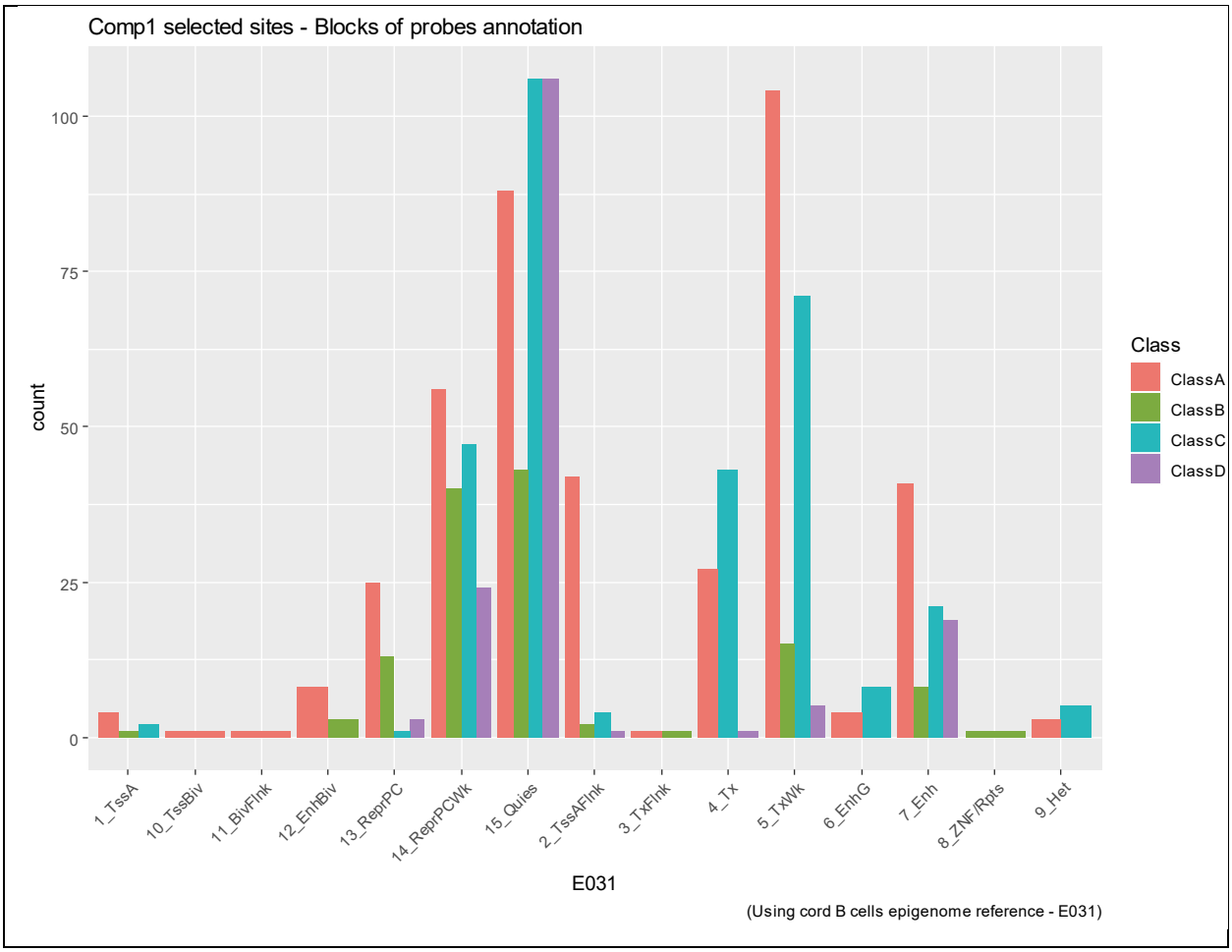
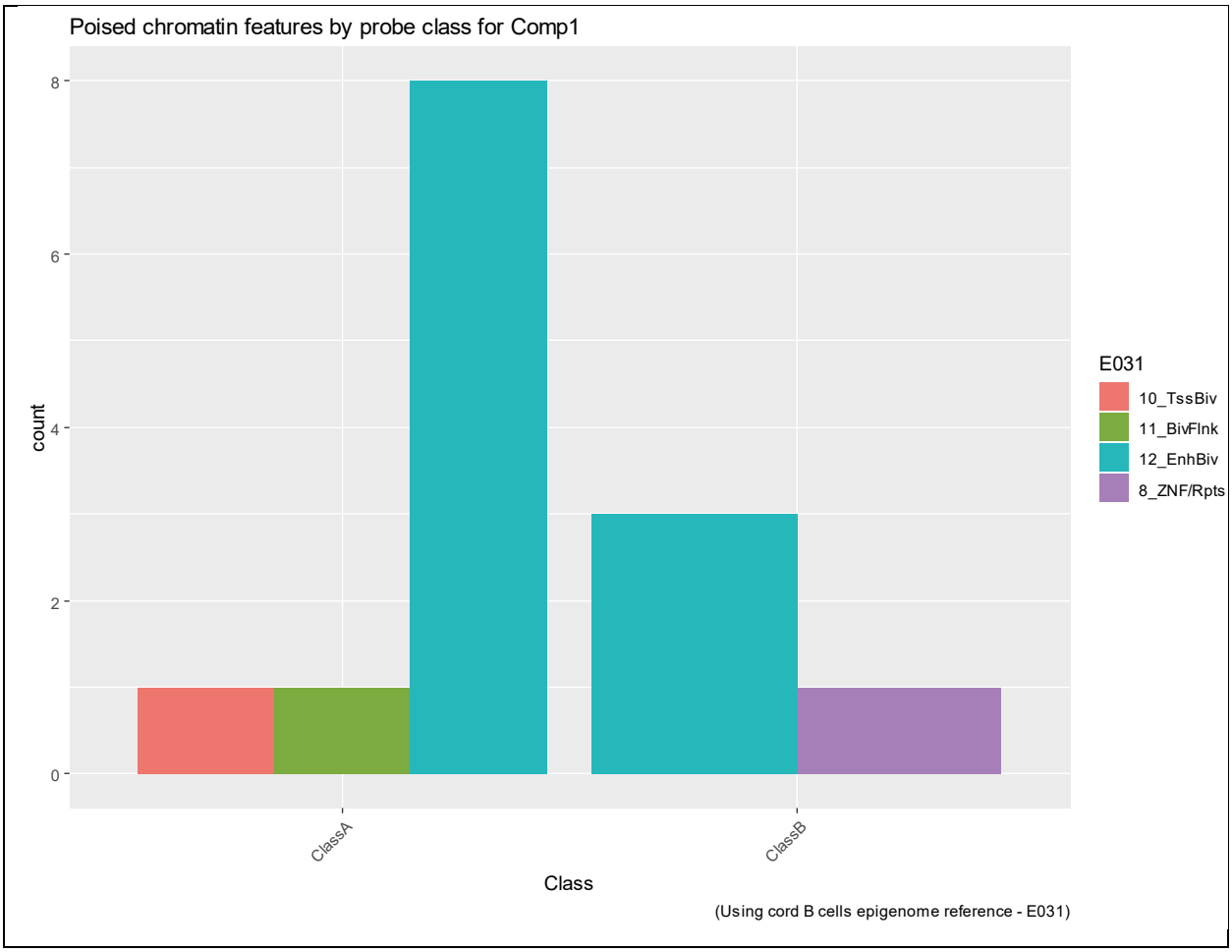


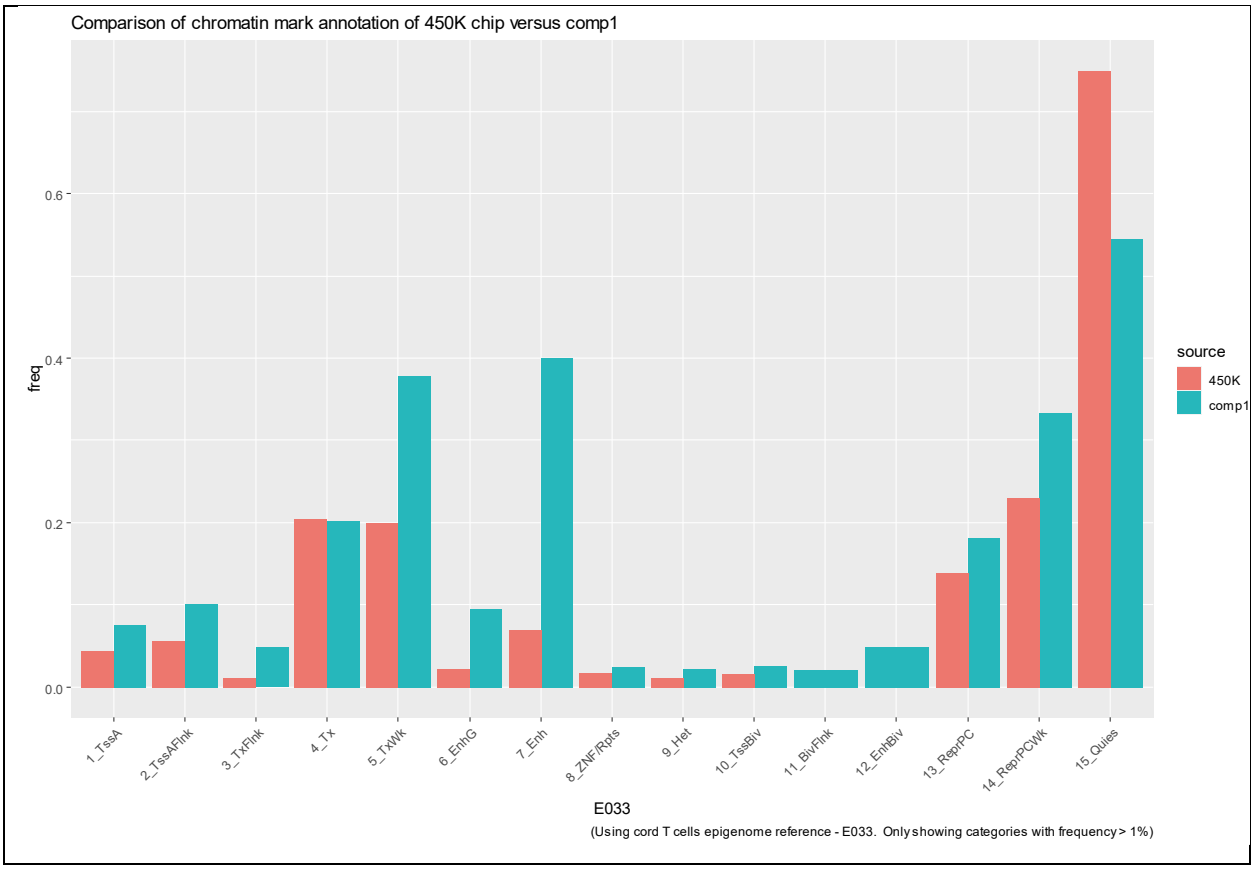
Figure 83: Correlation matrix between DNA methylation components at Age 15 (derived from MSP composite), MSP composite dimensions and estimated cell type composition using meffil R package (Reference: blood gse35069 - adult.) Pearson correlations that have p<.05 are indicated with a circle. Colour scale indicates r-value.

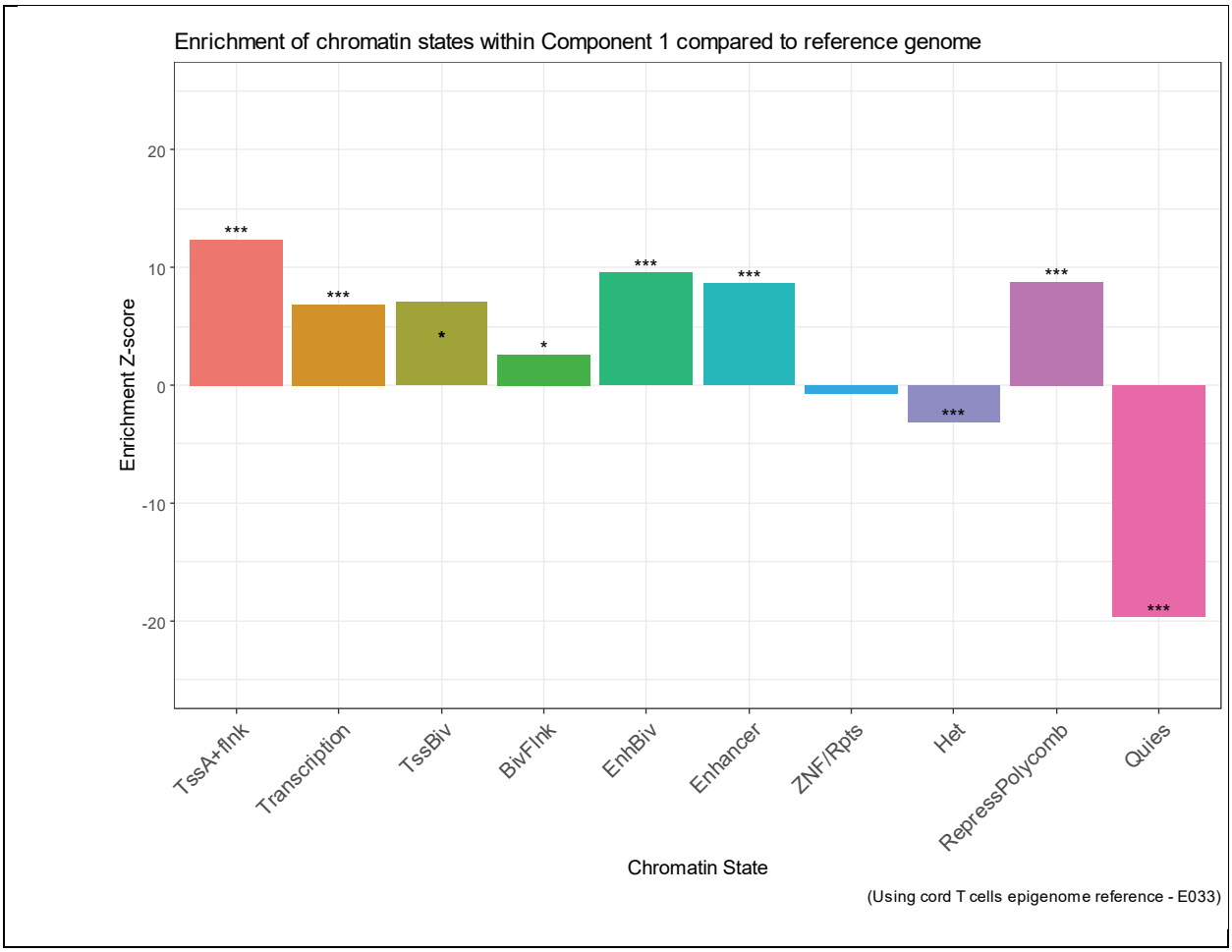
Appendix D Chromatin characteristics of component 1.

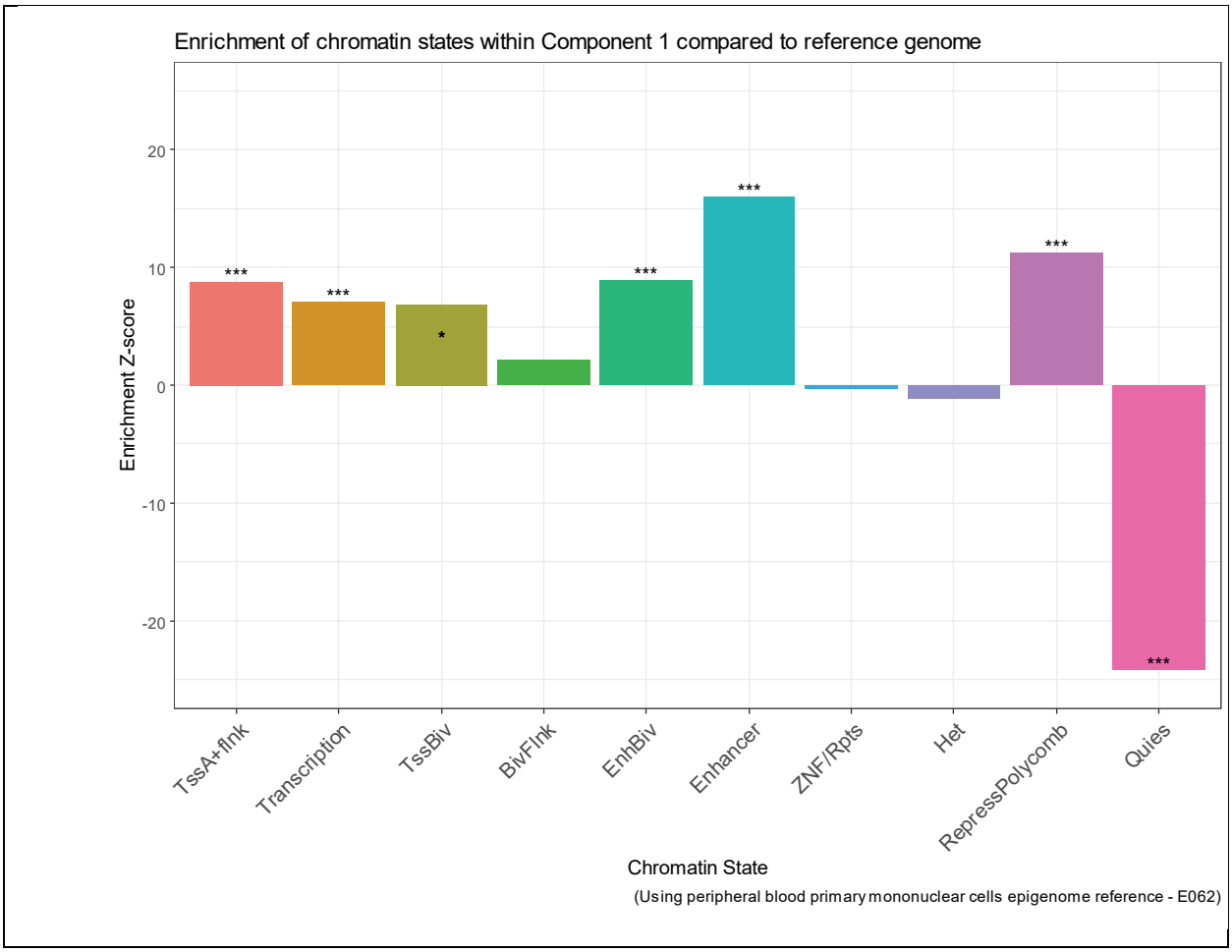
Chromatin characteristics of Component 1


Comp1 selected sites - Blocks of probes annotation
count
100
75
50
25
0
1_TssA
10_TssBiv
11_BivFlnk
12_EnhBiv
13_ReprPC
14_ReprPCWk
15_Quies
2_TssAFlnk
3_TxFlnk
4_Tx
5_TxWk
6_EnhG
7_Enh
8_ZNF/Rpts
9_Het
E031
(Using cord B cells epigenome reference - E031)
Class
ClassA
ClassB
ClassC
ClassD


Poised chromatin features by probe class for Comp1
count
8
6
4
2
0
ClassA
ClassB
Class
(Using cord B cells epigenome reference - E031)
E031
10_TssBiv
11_BivFlnk
12_EnhBiv
8_ZNF/Rpts


Comparison of chromatin mark annotation of 450K chip versus comp1
freq
0.0
0.2
0.4
0.6
1_TssA
2_TssAFlnk
3_TxFlnk
4_Tx
5_TxWk
6_EnhG
7_Enh
8_ZNF/Rpts
9_Het
10_TssBiv
11_BivFlnk
12_EnhBiv
13_ReprPC
14_ReprPCWk
15_Quies
E033
(Using cord T cells epigenome reference - E033. Only showing categories with frequency > 1%)
source
450K
comp1


Enrichment of chromatin states within Component 1 compared to reference genome
Enrichment Z-score
20
10
0
-10
-20
***
***
*
*
***
***
***
***
***
TssA+flnk
Transcription
TssBiv
BivFlnk
EnhBiv
Enhancer
ZNF/Rpts
Het
RepressPolycomb
Quies
Chromatin State
(Using cord T cells epigenome reference - E033)


Enrichment of chromatin states within Component 1 compared to reference genome
Enrichment Z-score
20
10
0
-10
-20
***
***
*
***
***
***
***
TssA+flnk
Transcription
TssBiv
BivFlnk
EnhBiv
Enhancer
ZNF/Rpts
Het
RepressPolycomb
Quies
Chromatin State
(Using peripheral blood primary mononuclear cells epigenome reference - E062)

Appendix E Comparison of results with reFACTor

Tables show correlation with DNAm components with reFACTor components. Only correlation with p < 0.05 displayed. Left: k = 6, Right: k = 7

k=6, ncomp = 6

row	column	cor	p
Comp8	Comp11	-0.07	0.026
Comp5	Comp12	0.14	0.000
Comp5	Comp16	0.07	0.047
Comp12	Comp16	0.15	0.000
Comp15	Comp19	0.07	0.032
Comp1	PC1	0.12	0.000
Comp2	PC1	-0.19	0.000
Comp3	PC1	0.22	0.000
Comp10	PC1	0.09	0.006
Comp12	PC1	0.08	0.017
Comp1	PC2	-0.20	0.000
Comp2	PC2	0.09	0.008
Comp3	PC2	0.26	0.000
Comp9	PC2	0.07	0.037
Comp16	PC2	-0.08	0.023
Comp17	PC2	-0.07	0.034
Comp6	PC3	0.10	0.004
Comp7	PC3	0.21	0.000
Comp10	PC4	0.09	0.010
Comp13	PC4	-0.22	0.000
Comp15	PC4	0.17	0.000
Comp17	PC4	-0.08	0.019
Comp18	PC4	-0.08	0.011
Comp19	PC4	0.08	0.020

k = 7, ncomp = 7

row	column	cor	p
Comp8	Comp11	-0.07	0.026
Comp5	Comp12	0.14	0.000
Comp5	Comp16	0.07	0.047
Comp12	Comp16	0.15	0.000
Comp15	Comp19	0.07	0.032
Comp1	PC1	0.11	0.001
Comp2	PC1	-0.19	0.000
Comp3	PC1	0.23	0.000
Comp10	PC1	0.09	0.006
Comp12	PC1	0.08	0.015
Comp1	PC2	-0.21	0.000
Comp2	PC2	0.09	0.005
Comp3	PC2	0.25	0.000
Comp9	PC2	0.07	0.035
Comp16	PC2	-0.08	0.022
Comp17	PC2	-0.07	0.032
Comp6	PC3	0.09	0.004
Comp7	PC3	0.21	0.000
Comp10	PC4	-0.08	0.016
Comp13	PC4	0.21	0.000
Comp15	PC4	-0.17	0.000
Comp17	PC4	0.08	0.015
Comp18	PC4	0.09	0.008
Comp19	PC4	-0.08	0.012

Comp4	PC5	-0.09	0.005
Comp7	PC5	0.08	0.013
Comp9	PC5	-0.11	0.001
Comp10	PC5	0.08	0.016
Comp12	PC5	0.07	0.025
Comp13	PC5	-0.12	0.000
Comp15	PC5	-0.07	0.023
Comp17	PC5	0.08	0.017
Comp19	PC5	-0.16	0.000
Comp7	PC6	0.07	0.034
Comp14	PC6	0.07	0.031
Comp15	PC6	0.07	0.044
Comp16	PC6	0.08	0.010

Comp4	PC5	0.09	0.005
Comp7	PC5	-0.08	0.013
Comp9	PC5	0.10	0.003
Comp10	PC5	-0.08	0.015
Comp12	PC5	-0.07	0.034
Comp13	PC5	0.13	0.000
Comp15	PC5	0.07	0.028
Comp17	PC5	-0.08	0.017
Comp19	PC5	0.16	0.000
Comp7	PC6	-0.07	0.042
Comp16	PC6	-0.10	0.003
Comp3	PC7	0.09	0.008
Comp14	PC7	0.12	0.000

Appendix F Measures of sampling adequacy for factorization

We examined the factorability of the eight composite variables using standard criteria. First, four out of eight items correlated at least .3 with at least one other item (see correlation matrix below). Second, the Bartlett's test of sphericity was significant ($\chi 2$ (28) = 98.2, p = 9.98 x 10^{-10}). Third, the overall Kaiser-Meyer-Olkin factor adequacy was 0.53. Given these overall metrics, we proceeded with factor analysis with these eight items.