TISSUE-SPECIFIC INVESTIGATIONS OF DNA METHYLATION VARIATION IN HUMAN NEUROBIOLOGICAL DISEASES

by

Sumaiya Islam

M.Sc., Georgetown University, 2010

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF

THE REQUIREMENTS FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

in

THE FACULTY OF GRADUATE AND POSTDOCTORAL STUDIES

(Medical Genetics)

THE UNIVERSITY OF BRITISH COLUMBIA

(Vancouver)

October 2018

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

Tissue-specific investigations of DNA methylation variation in human neurobiological diseases

submitted by Sumaiya Islam in partial fulfillment of the requirements for

the degree of Doctor of Philosophy

in Medical Genetics

Examining Committee:

Dr. Michael S. Kobor, Department of Medical Genetics

Supervisor

Dr. Wendy P. Robinson, Department of Medical Genetics

Supervisory Committee Member

Dr. Anthony G. Phillips, Department of Psychiatry

University Examiner

Dr. Clare L. Beasley, Department of Psychiatry

University Examiner

Dr. Carmen Marsit, Departments of Environmental Health & Epidemiology

External Examiner

Additional Supervisory Committee Members:

Dr. Martin Hirst, Department of Microbiology and Immunology

Supervisory Committee Member

Dr. David Huntsman, Department of Pathology and Laboratory Medicine

Supervisory Committee Member

Abstract

Epigenomic variation represents an emerging focus in human health research, particularly in regards to neurobiological disease susceptibility and pathogenesis. DNA methylation (DNAm), which involves the covalent attachment of a methyl group to a cytosine primarily at CpG dinucleotides, has been widely assessed in the context of epigenome-wide association studies (EWASs), with DNAm associations identified across a broad range of disease states, environmental exposures and genetic backgrounds. However, DNAm profiling in neurobiological diseases is challenged by the fact that DNAm variation is highly tissue-specific and target brain tissues may be difficult or impossible to collect from postmortem samples, in living individuals undergoing treatment interventions or in pediatric populations. As such, the use of cell-culture models or accessible peripheral tissues such as blood or buccal swabs represent alternative approaches used in human neurobiological DNAm studies to identify potential biomarkers of disease or treatment response.

The overarching aim of my dissertation was to apply and evaluate various tissue-specific approaches to investigate DNAm variation across different neurobiological diseases. To this end, I performed four separate studies to assess disease-associated DNAm from a) post-mortem brain samples, b) primary brain-derived cell culture models and c) accessible peripheral tissues. Specifically, I examined DNAm patterns related to Huntington's disease pathogenesis and tissue-specific *Huntingtin* gene expression in postmortem human cortex samples. I subsequently compared DNAm profiles from glioblastoma multiforme tumours and matched primary cell cultures enriched for brain-tumour initiating cell populations, identifying a homeobox-enriched signature of differential DNAm between the paired samples. Beyond brain-specific DNAm patterns, I also explored the use of a disease-relevant blood cell type, $CD3^+$ T-lymphocytes, to detect DNAm alterations associated with alcohol dependence in patients undergoing a clinical intervention. Finally, I assessed DNAm variability and the influence of genetic variation on DNAm in peripheral blood and buccal epithelial cells from two pediatric cohorts, highlighting a number of potential considerations and practical implications for the appropriate design and interpretation of early-life EWAS analyses in these tissues. Overall, these findings provide evidence to implicate DNAm variation in neurological function and pathology as well as present potential opportunities for the identification of novel biomarkers in accessible tissues.

Lay Summary

An emerging focus in human health research is epigenetics, the study of modifications to DNA structure and regulation. Epigenetic patterns can vary greatly between tissues, making tissue source an important consideration in epigenetic studies. This can be challenging, particularly for brain disease research, as human brain tissues can often only be obtained from deceased individuals and are difficult or impossible to collect in certain contexts, such as from children or living individuals undergoing a treatment intervention. This thesis explores the use of different sample sources, including brain tissues, brain-derived cell culture models and accessible tissues like blood or cheek swabs, to study disease-associated patterns of DNA methylation, an epigenetic mechanism involving the addition of chemical marks to DNA. Specifically, this work provides evidence to associate DNA methylation to various disease contexts and highlights the potential of DNA methylation patterns to serve as markers of disease risk or progression.

Preface

All data chapters in this thesis (Chapters 2-5) are presented in manuscript format, as they are currently published (Chapters 2 & 4) or under submission (Chapters 3 & 5).

Portions of **Chapter 1** (introduction) have been adapted from previously published work:

- Islam SA, Lussier AA, Kobor MS. (2018). Epigenetic analysis of human postmortem brain. In Webster MJ & Huitinga I (Eds). *Handbook of Clinical Neurology: Brain Banking Neurological and Pyschiatric Disorders.* (pp 237-261). Elsevier Inc. Reprinted with permission from Elsevier Inc (License Number: 4403920771953).
- Lussier AA*, Islam SA*, Kobor MS. (2017). Genetics and epigenetics of development. In Gibbs R & Kolb B (Eds). *The Neurobiology of Brain and Behavioural Development*. (pp 153-210). Elsevier Inc. *Authors contributed equally. Reprinted with permission of Creative Commons Attribution 4.0 International License (http://creativecommons.org /licenses/by/4.0/).

A version of **Chapter 2** has been published as:

- De Souza RAG*, Islam SA*, McEwen LM, Mathelier A, Hill A, Mah SM, Wasserman WW, Kobor MS, Leavitt BR. (2016). DNA methylation profiling in human Huntington's disease brain. *Human Molecular Genetics*. *Authors contributed equally. Reprinted with permission of Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/).

Collection of animal tissues used in this work were approved by the University of British Columbia's Animal Care Committee (Certificate: A14-0031). Collection of all human samples used in this study were approved by the University of British Columbia, Children and Women's Hospital Ethics board (Certificates: H06-70467 and H05-70532) and the Vancouver Coastal Health Authority Research study board (Certificate: V09-0129). The experimental and analytical design for this study was developed by myself and R. De Souza. The collection of human and animal tissues, cell culture experiments, qRT-PCR and ChIP-qPCR were performed by R. De Souza, with assistance from S. Franciosi, G. Lu and A. Hill. DNA methylation arrays were run by S. Mah and L. Lam while L. McEwen designed and performed the pyrosequencing assays. A. Mathelier analyzed the publically-available ChIP-seq data, under the supervision of W.

Wasserman. I performed all statistical and bioinformatics analyses of the DNA methylation array and pyrosequencing data. R. De Souza and I jointly wrote the manuscript and prepared all publication figures. B. Leavitt and M. Kobor supervised all steps of the process and provided critical feedback during manuscript preparation.

Chapter 3 is original and unpublished. Collection of all human samples used in this study were approved by the joint University of British Columbia and Children and Women's Hospital Ethics board (Certificate: H14-02694). I developed the design and research questions for this study alongside A. Wang, S. Weiss and M. Kobor. Sample collection was performed in Calgary, AB by A. Wang. DNA methylation arrays were performed by L. McEwen and J. MacIsaac. Generation and analysis of *MGMT* RNA-Seq data was performed by Y. Shen at Canada's Michael Smith Genome Sciences Centre, under the direction of S. Jones and M. Marra. I was responsible for all bioinformatics analyses and manuscript preparation, with critical feedback from S. Weiss and M. Kobor.

A version of **Chapter 4** has been published as:

- Brückmann C*, Islam SA*, MacIsaac JL, Morin AM, Karle KN, Santo AD, Wüst R, Lang I, Batra A, Kobor MS[§], Nieratschker V[§]. (2017). DNA methylation signatures of chronic alcohol dependence in purified $CD3^+$ T-cells of patients undergoing alcohol treatment. *Scientific Reports*. *Authors contributed equally; [§]Authors jointly supervised work. Reprinted with permission of Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/).

  Collection of all human samples used in this study were approved by the joint University of British Columbia and Children and Women's Hospital Ethics board (Certificate: H16-02333). This study was conceived and designed by V. Nieratschker and M. Kobor with input from myself and C. Brückmann. Recruitment of study participants and sample preparation was led by C. Brückmann, with assistance by K. Karle, A. Di Santo, R. Wüst, I. Lang and A. Batra. DNA methylation arrays were run by J. MacIsaac and pyrosequencing was performed by A. Morin. I was responsible for statistical analysis of all data. I generated all publication figures and wrote the manuscript with C. Brückmann. V. Nieratschker and M. Kobor provided critical feedback at all stages of manuscript preparation.

Chapter 5 is original and unpublished. Collection of all human samples used in this study were approved by the joint University of British Columbia and Children and Women's Hospital Ethics board (Certificates: H07-01317 and H07-02773). I developed the design and research questions for this study alongside S. Goodman and M. Kobor. Participant recruitment for both cohorts and sample collection were led by R. Barr, T. Boyce and M. Kobor. Genotyping and DNA methylation arrays were performed by J. MacIsaac. S. Goodman and I jointly analyzed all data and co-wrote the manuscript, under the supervision of M. Kobor.

Chapter 6 (discussion) contains adapted excerpts from the following publications:

- Islam SA, Lussier AA, Kobor MS. (2018). Epigenetic analysis of human postmortem brain. In Webster MJ & Huitinga I (Eds). *Handbook of Clinical Neurology: Brain Banking Neurological and Pyschiatric Disorders.* (pp 237-261). Elsevier Inc. Reprinted with permission from Elsevier Inc (License Number: 4403920771953).
- Lussier AA*, Islam SA*, Kobor MS. (2017). Genetics and epigenetics of development. In Gibbs R & Kolb B (Eds). *The Neurobiology of Brain and Behavioural Development*. (pp 153-210). Elsevier Inc. *Authors contributed equally. Reprinted with permission of Creative Commons Attribution 4.0 International License (http://creativecommons.org /licenses/by/4.0/).

Given that Chapters 2-5 remain largely unchanged from their published version or submitted manuscript copies, I have retained the use of plural first person pronouns in these sections. In the remainder of the dissertation, singular first person pronouns are employed.

Table of Contents

List of Tables

List of Figures

List of Abbreviations

450K	Illumina Infinium HumanMethylation450K Beadchip Array
5hmC	5-hydroxymethylcytosine
5mC	5-methylcytosine
AD	Alcohol dependence
ADORA2A	Adenosine A (2A) receptor
ASD	Autism spectrum disorders
AUDIT	Alcohol use disorder identification test
BDNF	Brain-derived neurotrophic factor
BEC	Buccal epithelial cell
BH	Benjamini-Hochberg
BMIQ	Beta-mixture quantile normalization
BTIC	Brain tumour initiating cell
C3ARE	Cleaning, Carrying, Changing, Attending, Reading and Expressing
caC	carboxycytosine
CETS	Cell epigenotype-specific
CGIs	CpG islands
ChIP	Chromatin immunoprecipitation
ChIP-qPCR	Chromatin immunoprecipitation qPCR
ChIP-seq	ChIP-sequencing
CLS	Calgary laboratory services
CpG	Cytosine-phosphate-guanine dinucleotide
CPT	Cell preparation tubes
Cvexp	Experimental coefficient of variation
Cvtotal	Total coefficient of variation
DAVID	Database for Annotation, Visualization and Integrated Discovery
DMRs	Differentially methylated regions
DNAhm	DNA hydroxymethylation

DNAm	DNA methylation
DNMTs	DNA methyltransferases
DSM-IV	Diagnostic and statistical manual of mental disorders (4th Edition)
EDTA	Ethylenediaminetetracetic
ENCODE	Encyclopedia of DNA Elements
ESCs	Embryonic stem cells
EWAS	Epigenome-wide association study
FASD	Fetal alcohol spectrum disorder
fC	formylcytosine
FDR	False discovery rate
FGF	Fibroblast growth factor
FKBP5	FK605 binding protein 5
G-CIMP	Glioma CpG island methylator phenotype
GBM	Glioblastoma multiforme
GECKO	Gene Expression Collaborative Kids Only
GO	Gene ontology
GR	Glucocorticoid receptor
GSI	Global distress level
GWAS	Genome-wide association study
GxE	Gene by environment
HD	Huntington's disease
Hdh	murine huntingtin gene
HEK	Human embryonic kidney
HTT	Huntingtin gene
ICF	Immunodeficiency, Centromere Instability, Facial anomalies
IHEC	International Human Epigenome Consortium
MBD	methyl-CpG-binding domain
MC-Seq	Methyl-capture sequencing
mCH	Methylated CH (where C = Cytosine; H = Adenine/Cytosine/Thymine)

MDD	Major depressive disorder
meDIP	methylated DNA immunoprecipitation
MGMT	O6-methylguanine-DNA-methyltransferase
MMR	Mismatch repair
mQTL	Methylation quantitative trait locus
MS	Methylation-specific
MS-Cseq	Methylation-specific clone sequenciing
MSP	Methylation-specific PCR
NF	Normalization factor
NF-kB	Nuclear factor kappa-light-chain-enchancer of activated B cells
NIH	National Institutes of Health
NPCs	Neural progenitor cells
NSCs	Neural stem cells
OCDS	Obsessive compulsive drinking scale
OPCs	Oligodendrocyte precursor cells
ox-BS	oxidative bisulfite
PBMC	Peripheral blood mononuclear cells
PC	Principal component
PCA	Principal component analysis
PD	Parkinson's disease
PMI	Post-mortem interval
PRC2	Polycomb repressor complex 2
PsychChip	Illumina Infinium PsychChip Beadchip Array
PTSD	Post-traumatic stress disorder
PWM	Position weight matrix
RNA-Seq	RNA-Sequencing
RRBS	Reduced representation bisulfite sequencing
RT-qPCR	Quantitative real time PCR
SCL-90-R	Symptom checklist-90-R

SFM	Serum-free medium
SNP	Single nucleotide polymorphism
SWAN	Subset-quantile-within-array-normalization
T1	Timepoint 1
T2	Timepoint 2
TAB	TET-assisted bisulfite
TCGA	The Cancer Genome Atlas
tDMRs	Tissue-specific differentially methylated regions
TET	Ten-Eleven-Translocation
TF	Transcription factor
TFBSs	Transcription factor binding sites
TMZ	Temozolomide
UTR	Untranslated region
WGBS	Whole genome bisulfite sequencing

Acknowledgements

The completion of this PhD dissertation could not have been possible without the support of a number of people. Firstly, my heartfelt gratitude to my supervisor, Dr. Michael S. Kobor for graciously welcoming me into his lab during a difficult transition period, continuously pushing me to challenge myself and always encouraging me to have fun while doing good science. I have learned so much from you and I thank you for all the lifelong lessons. My sincere thanks to my wonderful committee members, Dr Wendy Robinson, Dr. Martin Hirst and Dr. David Huntsman for your continuous support, guidance and kindness.

I would like to acknowledge all of my wonderful collaborators: Dr. Sam Weiss, Dr. Alice Wang, Dr. Artee Luchman, Dr. Yaoqing Shen, Dr. Steve Jones, Dr. Vanessa Nieratschker, Dr. Christof Brückmann, Dr. Blair Leavitt, Dr. Rebecca De Souza, Dr. Tom Boyce and Dr. Ron Barr. I would also like to thank the members of Medical Genetic graduate program, including the many professors, fellow graduate students, and in particular, Cheryl Bishop, whose kindness and encouragement has meant a lot to me.

A profound thank you to former and current members of the Kobor lab who have enriched my PhD experience with meaningful friendships, wonderful support and touching memories. Thank you Koborites – Alex L., Alex M., Alice, Alyssa, Anthony, Asha, Cath, Chloe, David, Evan, Grace, Helena, Hilary, Joe, Josh, Julie, Katia, Kristy, Lisa, Maria, Maggie, Meaghan, Mina, Nicole, Olivia, Phoebe, Rachel C., Rachel E., Ruiwei, Sachini, Samantha, Sarah G., Sarah M., Tanya and Yasmin. A special thank you to Rachel Edgar, Meaghan Jones, Lisa McEwen and Mina Park who not only provided critical feedback and support throughout my PhD but are also so inspiring as my colleagues and friends. I would also like to specifically acknowledge Nicole Gladish and Daniel Radiloff who stood with me during a particularly difficult time in my PhD with integrity and courage – your solidarity and enduring friendship means the world to me.

Throughout my PhD, I have been so fortunate to have an incredible support system of friends and peers. A heartfelt thank you to Grace Tharmarajah, Jack Hickmott, Magda Price,

Pooja Mohan, Chaini Konwar, Dominik Sommerfeld, Mandi Schmidt, Jessica Pillsworth, Kailin Webb and Iram Jabbar. Your love, encouragement and care have enlightened, strengthened and inspired me throughout my graduate school experience.

Finally, the greatest blessing in my PhD and in my life has been my family. To my older sister Shezana, thank you for being my inspiration, for always looking out for me and for bringing Givo, Sha and Sabine into our beautiful growing family. To my little sister Nishi, thank you for being a quiet but strong voice of reason and calm in my life. To my husband Atif, thank you for being my most cherished source of comfort and strength and for being my very best friend. Finally, to my parents, my Abbu and Ammu: everything that I have been able to achieve or do in my life is because of you. Thank you for your immense love and sacrifice.

Dedication

To my parents for all their love and sacrifice.

Chapter 1: Introduction

1.1 Dissertation context and aims

Epigenetics represents an increasingly attractive focus for the study of brain dysfunction and complex neurobiological diseases, due, in part, to its potential to serve as a mechanistic interface between genomic function and environmental context. One of the most studied epigenetic marks in humans is DNA methylation (DNAm), which has been widely assessed in the form of epigenome-wide association studies (EWASs). Playing a key role in gene expression regulation, DNAm patterns exhibit high tissue- and cell-type specificity and are associated with the preservation of the cellular memory required for developmental stability (1–3). Conversely, DNAm is also subject to dynamic variation in that methylation at specific sites can change in response to environmental influences (4, 5). As such, DNAm serves as one of the most promising epigenetic candidates for the biological mediation of gene-environment (G x E) interactions (6). Taken together, it is this tissue-specific and environmentally-responsive control of gene regulation which makes DNAm a prime research focus for the study of neurobiological disease phenotypes in human population cohorts.

Despite its increasing popularity, DNAm profiling in human neurobiological diseases is challenged by the fact that DNAm variation is highly tissue-specific and target brain tissues may be difficult to collect from postmortem samples, particularly in pediatric populations (7). Although the use of brain tissues may afford the opportunity to profile potential disease-related mechanisms, brain samples can only be assessed retrospectively, thereby precluding longitudinal analyses in living individuals throughout disease progression or across treatment interventions. As such, the use of cell-culture models or accessible peripheral tissues such as blood or buccal swabs represent alternative approaches used in human neurobiological DNAm studies to identify potential biomarkers of disease or treatment response (8).

The overarching aim of my dissertation was to apply and evaluate the use of various tissue sources to investigate DNAm variation across different types of neurobiological disease states including neurodegenerative, neuropsychiatric and neurological pathologies. To this end, I performed four separate studies to assess disease-associated DNAm from a) postmortem brain samples, b) primary brain-derived cell culture models and c) accessible peripheral tissues. Specifically, I examined DNAm patterns in postmortem human cortex samples in order to

understand the potential role of DNAm in Huntington's disease pathogenesis and tissue-specific transcriptional regulation of the *Huntingtin* gene locus (Chapter 2). I subsequently compared DNAm profiles from glioblastoma multiforme tumours and matched primary cell cultures enriched for brain-tumour initiating cell (BTIC) populations, with the hopes of identifying DNAm variation implicated in the stem-like functions of BTICs (Chapter 3). Beyond brain-derived DNAm patterns, I also explored the use of a disease-relevant blood cell type, $CD3^+$ T-lymphocytes, to detect DNAm alterations associated with alcohol dependence in patients undergoing a clinical intervention (Chapter 4). Finally, I assessed DNAm variability and the influence of genetic variation on DNAm in peripheral blood and buccal epithelial cells from two pediatric cohorts, with the goal of characterizing the tissue-specific nature of genetic influences on DNAm variability in these pediatric tissues (Chapter 5).

Taken together, this body of work seeks to highlight the unique advantages and challenges of various tissue-specific approaches to DNAm studies in neurobiological diseases. Moreover, this suite of analyses aims to assess DNAm variation in neurological function and pathology as well as present potential opportunities for the identification of novel biomarkers in accessible tissues.

1.2 Epigenetics: an emerging focus in the study of human neurobiological disease

1.2.1 Epigenetic mechanisms link cellular function and environmental influences

Over the last decade, the emerging field of neuroepigenetics has garnered considerable interest for its potential to provide insights into the molecular mechanisms underpinning neurological function and brain-related pathologies (9). While genetic variation comprises the inherited basis of cellular function and activity, epigenetics is considered the regulatory overlay of the genome that fine-tunes gene activity in response to external signals (10). The concept of epigenetic regulation was first proposed by Conrad Waddington in the early 1940s to describe how the developmental patterning of multicellular organisms are shaped by 'epigenetic landscapes' that drive cellular differentiation along a programmed trajectory towards specific cell-type lineages (11). Since its inception, the field of epigenetics has flourished into an active area of study aimed at characterizing gene regulation and biological variation, particularly as it relates to human neurobiological diseases (12).

Today, epigenetics is operationally defined as mitotically heritable modifications of DNA and its regulatory components, including chromatin and non-coding RNA, that potentially modulate cellular states or fate through gene expression changes, without changing the DNA sequence itself (2, 6, 13, 14). Notably, Waddington's original hypothesis still holds true: the identity and the functional specification of the ~200 different cell types in the human body are largely dictated by the unique epigenomic profiles and corresponding transcriptional programs of each cellular subtype (15). In this manner, epigenetic mechanisms dually serve to allow for dynamic tissue- and cell type-specific variation, as well as the preservation of the cellular memory required for developmental stability. In addition, epigenetic regulation is now becoming increasingly recognized as a potential biological mediator of environmental influences, which can contribute to sculpting the epigenome across the life course (4, 5). Indeed, epigenetic marks may be particularly malleable during certain developmental periods in early life, resulting in epigenetic changes that may be linked to health risks and disease in later life (6, 10). Taken together, epigenetic mechanisms exist in a seeming paradox between the stability of cellular identity and plasticity of environmental responses, modulating cellular functions through both short- and long-term responses to stimuli (10). It is this cell-specific and environmentally-responsive control of gene regulation which makes epigenetics an emerging focus for molecular studies in human health research, particularly in the context of complex, environmentally-conditioned neurobiological and psychiatric phenotypes (Figure 1.1).

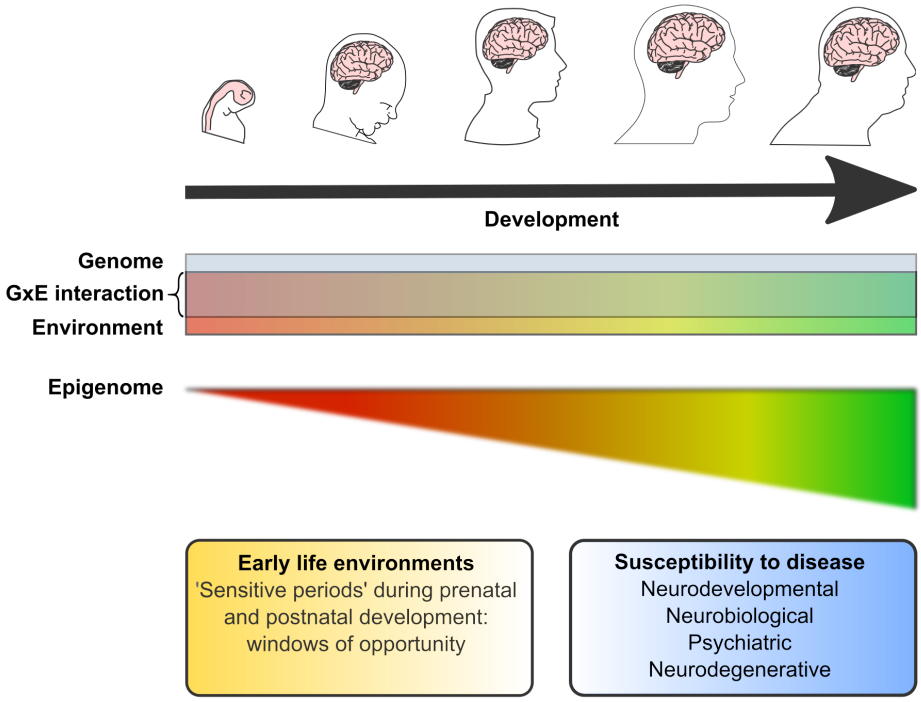


Figure 1.1 Biological variation and brain-related phenotypes are influenced by multiple sources throughout brain development.

Genetic variation (blue bar) is inherited at birth and remains mostly stable throughout the lifetime. By contrast, the environment presents shifting conditions that can influence long term health and behavior (rainbow bar). The intersection of these two influences represents gene by environment (GxE) interactions (overlap). Together, these factors are reflected in the epigenome, which is highly malleable in response to environmental conditions and strongly influenced by genetic variation. Epigenetic variability increases across the lifecourse, with different developmental windows conferring differential sensitivity to GxE and environmental influences (widening triangle gradient). These windows of opportunity for developmental programming of epigenetic patterns and subsequent health are more vulnerable during early life, which ranges from preconception and prenatal life to postnatal environments. In turn, these effects can also influence vulnerability to disease, including neurodevelopmental disorders, neurobiological dysregulation, psychiatric disease, and neurodegeneration later in life. Taken together, the influences of both the genome and shifting environments are reflected in the epigenome, which can shape development and vulnerability to disease.

1.2.2 Epigenome-wide association studies

Deriving from the burgeoning field of epigenetic epidemiology, epigenome-wide association studies (EWASs) systematically apply genome-wide approaches to identify epigenetic loci that are associated with a particular complex disease, phenotype or environmental exposure (16). While EWAS analyses have become increasingly popular and have generated a number of successfully replicated findings, including epigenetic differences associated with tobacco smoking exposure, there has also been growing recognition of various methodological

and analytical issues to consider when performing an EWAS (16–23). Similar to genome-wide association studies (GWASs), EWASs may be challenged by technical and statistical issues including batch effects, harmonization of data collected on multiple platforms, low power to detect due to insufficient sample size, multiple testing barriers and identification of functionally-relevant associations (16). However, unlike GWASs, which interrogate relatively stable genetic variation, EWAS analyses are further complicated by the fact that epigenetic patterns can vary by biological factors including tissue type, cell composition within a tissue, age, sex, environmental exposure, and genetic background (20, 24, 25). Due to this added complexity, care and consideration of these factors are warranted in the design of EWASs and in the interpretation of EWAS findings (16, 20, 25). While EWAS analyses may be performed on any type of epigenetic mark, the vast majority of EWASs to date have interrogated DNA methylation due to its relative stability and ease of measurement using genome-scale, quantitative technologies (26, 27).

1.3 DNA methylation (DNAm)

1.3.1 Genomic distribution and establishment of DNAm

DNA methylation (DNAm) is arguably the most studied epigenetic mark in humans and involves the covalent attachment of a methyl group to the 5' position of cytosine, typically at CpG dinucleotide sites (28). These CpG dinucleotides occur relatively infrequently in the genome as a consequence of natural selection against DNAm-induced sequence mutability as methylated cytosines can undergo spontaneous deamination to thymine (29–31). Areas with comparatively high CpG content in the genome have been termed "CpG islands" (CGIs) and these CGIs are thought to exist as regions that were either never methylated or only transiently methylated in the germline while the rest of the genome experienced a progressive loss of CpGs at methylated sequences across evolution (29–31). Importantly, the DNAm status of the ~ 28 million CpG sites in the human genome is often dependent on genomic context (26, 32). For example, CGIs, which are associated with approximately 50-70% of known promoters, tend to contain low levels of methylation in somatic cells, while non-island CpGs exhibit generally higher methylation levels (29, 30, 33).

The establishment and maintenance of DNAm patterns are carried out by a highly conserved family of enzymes known as DNA methyltransferases (DNMTs). In mammals, 3

major DNMTs have been identified, DNMT1, DNMT3A and DNMT3B, which are characterized by a conserved stretch of amino acids in the C-terminal catalytic domain that target the 5' carbon of cytosine (34). As the most abundant form in adult cells, DNMT1 maintains DNAm patterns during cell division by binding hemi-methylated CpG sites following DNA replication and methylating the cytosine on the newly-synthesized daughter strand (35). In contrast to DNMT1's role in DNAm maintenance, DNMT3A and DNMT3B establish *de novo* genome-wide DNAm patterns following embryo implantation (36). These enzymes show equal affinity for hemi-methylated or non-methylated DNA, often in conjunction with other factors, and are essential for early development, as deleting their encoding genes causes embryonic lethality in mice (36, 37).

1.3.2 Associated functions of DNAm

DNAm is characterized as an important regulatory mechanism of genome function with key roles in various developmental processes including tissue differentiation, imprinting and X-chromosome inactivation (1). In general, DNAm is associated with the regulation of gene expression, although its effects on transcription are highly dependent on genomic context (38, 39). For example, DNAm at gene promoters is typically associated with gene expression silencing, although its role may be more variable within gene bodies (15, 26). Conversely, in regions of lower CpG density which flank CGIs, known as "island shores", high DNAm levels are generally associated with highly expressed genes, especially if the associated CGI is lowly methylated (39–41). While the exact mechanisms remain mostly unknown, transcriptional silencing by DNAm may potentially occur through the direct blocking of transcription factor binding or the recruitment of transcriptional repressors to promoter, enhancers, or insulator regions (15, 42). Emerging evidence shows that when comparing a single gene across a population, the association between DNAm and gene expression can be negative, positive, or non-existent, highlighting the complex relationship between DNAm and transcription (38, 43, 44). Moreover, DNAm can be both active, by being a likely cause of gene expression variation, or passive, by being a consequence or an independent mark of gene expression levels (43, 44). In addition to its role in transcriptional control, DNAm at exon-intron boundaries has been associated with altered mRNA splicing, and its presence within certain exons potentially regulates alternative transcriptional start sites (45–48). Finally, DNAm in repetitive elements, which comprise more than half of the human genome including intergenic sequences, tends to

occur at relatively high levels and is associated with maintenance of chromosome structure and genomic integrity (49, 50).

1.3.3 Tissue specificity of DNAm

Tissue specificity represents a salient feature of DNAm, as different tissues and cell types acquire distinct DNAm landscapes during differentiation (51). At present, considerable research efforts are being made to elucidate the tissue specificity of DNAm patterns with respect to individual CpGs as well as inter-individual variation within a tissue (52–54). Specifically, it has been shown that genome-wide DNAm differences between tissues within an individual greatly exceed differences within a tissue across a population, as substantiated by the clustering of samples according to tissue type over individual in multi-tissue comparisons (55–57). Site-specific and regional analyses have identified tissue-specific differentially methylated regions (tDMRs) across the genome, which primarily occur in regions of intermediate CpG density, with the majority of tDMRs present in CpG island shores (58). Together, these results are in line with the finding that tissue identity is the largest driver of variance in genome-wide DNAm profiles (59).

Following tissue type, cell composition within a tissue is the second largest driver of DNAm variation (59). As stem cells divide and differentiate into specific terminal cell types, DNAm patterns become increasingly cell-type specific, with cells from related lineages showing more similar DNAm profiles (51, 52). Indeed, DNAm patterns may serve as a form of cellular memory, as demonstrated by the recapitulation of hematopoietic lineage hierarchy using DNAm profiles of blood cell types (60). These findings suggest a role for DNAm in lineage specification, but they also highlight important implications for DNAm study design and analysis as unaccounted differences in cell composition within a tissue may result in the identification of spurious DNAm associations or mask true associations (20, 23–25). As a result, a number of bioinformatic approaches have been developed to either directly estimate or adjust for cell-type heterogeneity from different tissue sources in EWAS analyses (61–70). Cell-type correction for DNAm studies in brain samples are particularly pertinent as brain tissues are highly heterogeneous, comprising of many different cell types including neurons, astrocytes, oligodendrocytes, oligodendrocyte precursor cells (OPCs), microglia and vascular cells, and can often exhibit disease-specific alterations in cell proportions such as loss of specific neurons in

distinct brain regions, in conjunction with increased proliferation of glial cells (gliosis), as observed in various neurodegenerative pathologies (71). Taken together, DNAm studies of tissue specificity and cell composition are rapidly evolving to enhance our understanding of epigenetic mechanisms in tissue-specific biology and lineage specification.

1.3.4 Interplay between genetic variation and DNAm

There is growing interest in DNAm as a potential mediator of gene-environment (G x E) interactions, defined as genetic or environmental effects on phenotype or outcome that are dependent on each other. More specifically, certain genes can moderate an environment's influence on a particular individual, or environmental influences can only be revealed among individuals of a particular genotype (72). For example, a link between childhood maltreatment and an allelic risk variant for post-traumatic stress disorder (PTSD) was established in the *FKBP5* gene, which encodes a chaperone of the glucocorticoid receptor (GR), a key mediator of the stress response (73). This association is potentially mediated through a decrease in methylation of a CpG located in the intron of the *FKBP5* risk allele, leading to the suppression of GR function, dysregulation of stress responsivity, and increased risk for PTSD (73). This work not only provided one of the first demonstrations of epigenetics as a molecular mediator in G x E interplay, but also pointed to the functional effects of a methylation quantitative trait locus (mQTL), defined as an allelic variant that correlates with CpG methylation levels. A number of studies have explored the occurrence of mQTLs across different populations, developmental stages and tissues (43, 74–77). In the context of the brain, mQTLs tend to occur as *cis* associations in different brain regions and may underlie risk loci of various neuropsychiatric diseases, such as schizophrenia and bipolar disorder (78–83). In this regard, genetic variation represents an additional contributor to DNAm patterns, with genetic influences accounting for an estimated 20-80% of DNAm variance within a tissue (84–88) . When modeling variability in human neonatal methylomes, the inclusion G x E interaction terms account for up to 75% of the variably methylated regions between individuals over models containing only G or E terms, suggesting that G x E interactions play an important role in mediating the genomic response to external stimuli and potentially shaping developmental trajectories in early life (89). Finally, more recent demonstration of epigenetic mediation of G x E interaction in the context of substance use intervention programs in youth has highlighted the potential positive impact of

considering G x E effects in the design of intervention schemes and prevention strategies (90). Taken together, these observations provide a compelling framework for further investigating the biological implications of genetic and epigenetic interplay.

1.3.5 Non-CpG DNAm

While DNAm primarily occurs in the context of CpG dinucleotides, it can also occur at CpH (where H = A/C/T) sites. Indeed, both the maintenance DNMT1 and de novo DNMT3A/B enzymes have been shown to methylate non-CpG cytosines *in vitro* (91, 92). Previous studies have shown that methylated CH dinucleotides (mCH) occur in cultured embryonic stem cells (ESCs) and induced pluripotent stem cells (93–97). Moreover, analysis of adult human and mouse CNS neurons found that mCH is specifically enriched in neurons compared to other cell types, as non-CpG methylation is nearly absent in nonneuronal adult somatic cells, but can reach up to ~ 25% of all cytosines in neurons of the adult mouse dentate gyrus (92, 96, 98). Levels of mCH increase rapidly during early postnatal brain development (mouse, ~2-4 weeks; human 0-2 years), suggesting that mCH potentially plays an important role in the regulation of postnatal brain development. These changes are associated with a transient rise in DNMT3A levels, as knockdown of this enzyme results in significant loss of mCH, but not methylated CpG levels (92). Genome-wide profiling also showed that in neurons, mCH is present throughout the 5' upstream, gene-body, and 3' downstream regions of genes, where it is negatively correlated with gene expression (92, 98). Furthermore, *in vitro* plasmid reporter gene analyses have shown that CpH methylation is associated with transcriptional repression in mouse neurons (92). However, mCH is not associated with gene silencing in all cell types, as non-CpG methylation in ESCs positively correlates with gene expression (94). It is thought that the distinct distribution and role in gene expression of mCH in different cell types relates to differences in the relative abundance and activity of specific "readers" and "writers" of non-CpG methylation (99). Furthermore, in addition to CpH methylation, very recent research has detected the presence of methylated adenosine nucleotides in vertebrates, suggesting that that DNA modification variants may be more diverse than previously thought (100–103).

1.3.6 Additional cytosine modifications

Apart from methyl groups, additional cytosine modifications have been identified and are potentially implicated in the process of DNA demethylation. Although the mechanisms underlying the establishment and maintenance of DNAm by DNMTs have been well characterized, the process of DNA demethylation remains unclear. Thought to involve both active and passive pathways, this phenomenon is vital for typical development and genetic regulation, particularly in the brain (104–106). For example, neuronal activity-induced DNA demethylation of specific promoters and expression of corresponding genes such as brain-derived neurotrophic factor (*BDNF*) and fibroblast growth factor (*FGF*) occurs through the action of Gadd45b and represents an activity-dependent form of modulating neurogenesis in the adult brain (107). Passive DNA demethylation can occur due to a lack of DNMT1 activity, resulting in a gradual loss of DNAm over several rounds of replication (108). In addition, DNA demethylation may potentially occur through the oxidation of 5mC, catalyzed by the Ten-Eleven-Translocation (TET) family of enzymes (109, 110). This process generates a series of oxidized cytosine base variants, including 5-hydroxymethylcytosine (5hmC), formylcytosine (fC), and carboxycytosine (caC) (109, 111). The oxidized site can then be removed by thymine DNA glycosylase to create an abasic site, which undergoes base excision repair to yield an unmodified cytosine (112). Alternatively, 5hmC can be converted to hydroxymethyluracil by activation-induced deaminase prior to base excision repair (113). Although the exact details of active DNA demethylation remain unclear, the emerging evidence points to a process involving the coordinated activity of a number of key enzymatic players and intermediate modified cytosine species. In addition to their potential role in DNA demethylation, these cytosine variants may also play a role in modulating chromatin structure or recruiting various factors to key regions of the genome (114). For instance, various members of the methyl-CpG-binding domain (MBD) protein family display different affinities for 5hmC, and given their role in recruiting different chromatin modifying complexes, hmC could potentially alter chromatin landscapes throughout the genome (115). Interestingly, DNA hydroxymethylation (DNAhm) is present at high levels in pluripotent cells and the brain, where it has been implicated in neural stem cell functions, although its exact functional role remains to be uncovered (110, 111, 116). Genome-wide mapping of DNAhm in various brain regions, including the frontal cortex, hippocampus, and cerebellum, identified an enrichment of 5hmC in gene bodies, which was positively

associated with gene transcription, particularly at developmentally activated genes (98, 117). Finally, active DNA demethylation and TET activity is associated with memory formation and addiction in mice, further supporting its functional role in neural activity (118).

1.3.7 Methods for assessing DNA modifications

Different approaches have been used to study DNA modification patterns, ranging from bulk levels to targeted and genome-wide techniques. Methods to investigate bulk levels of DNA modifications involve immunoblotting and fluorescence methods as well as the analysis of repetitive elements to obtain a snapshot of levels across the entire genome (119, 120). However, these are now being superseded by high-throughput, genome-wide approaches, which can profile DNA modification levels, often at single base resolution, across the genome.

For the most part, genome-wide methods rely on the chemical treatment of cytosine residues to obtain DNA modification information. Bisulfite conversion is the primary method used to investigate DNAm, converting unmodified cytosines to uracil, while leaving cytosines with DNA modifications unaffected. However, this technique also results in the protection of hydroxymethylated residues, and the resulting uracil/cytosine ratio reflects both types of modifications. As such, additional biochemical treatments have been developed to deconstruct this confound, including TET-assisted bisulfite (TAB) and oxidative bisulfite (ox-BS) conversion, which only protect 5hmC from conversion to uracil (121, 122). These are being used in conjunction to identify the ratio of DNA modifications at a given cytosine residue.

Several techniques are currently being used in the genome-wide analysis of DNA modifications using the above methods. The current gold standard, whole genome bisulfite sequencing (WGBS), provides cytosine modification information for all sites within the genome, but requires high sequencing coverage and may often be too costly to perform in large cohort studies (94). By contrast, other techniques use enrichment approaches to limit analyses to certain genomic regions. Reduced-representation bisulfite sequencing (RRBS) uses methylation-insensitive enzymes to digest DNA and size selection to enrich regions of high CpG density for sequencing (123, 124). This results in lower overall coverage of the genome, covering about 2 million CpGs, with an enrichment of CGIs and promoters, which may be less sensitive to environmental influences, but are more closely linked to gene expression patterns (123, 125). Enrichment can also be performed through capture methods, such as Agilent's SureSelect

Human Methyl-Seq or NimbleGen's SeqCap Epi Enrichment System, which use custom designed oligonucleotides to select relevant regions of the genome (126). In addition to these sequencing-based approaches, commercial arrays from Illumina (GoldenGate [~1,500 sites], 27K [~27,000 sites], 450K [~485,000 sites], and most recently, the EPIC [~860,000 sites]) provide quantitative DNAm or DNAhm data across the genome, and are the most commonly used method for EWASs (127–130). Furthermore, more targeted methods, such as pyrosequencing, also rely on these chemical conversions to obtain quantitative epigenetic data at specific loci (126). In contrast to these methods, methylated DNA immunoprecipitation (meDIP) uses antibodies raised against DNAm or DNAhm to select regions of the genome enriched for these cytosine modifications (46, 131–133). While this method is unbiased towards CpG-rich regions and can reduce the complexity of the dataset by omitting unmethylated regions, it is unable to provide quantification at single-CpG resolution, thereby representing a largely qualitative approach to assess cytosine modifications across the genome (126) .

1.4 DNAm studies of human neurobiological diseases

1.4.1 DNAm associations with neurodevelopmental disorders

The etiological basis for a number of neurodevelopmental disorders involves the disruption of genes encoding epigenetic factors (134). For example, mutations in the X-linked gene *MECP2*, which encodes a methyl-CpG-binding protein, causes Rett syndrome, a childhood disorder associated with a broad range of developmental, cognitive and neurologic deficits (135). Alternatively, loss-of-function mutations in the DNA methyltransferase gene *DNMT3B* are responsible for Immunodeficiency, Centromere Instability, Facial anomalies (ICF) syndrome, a disorder involving facial dysmorphism, immunoglobin deficiency and defective brain development (136, 137). In addition, perturbations affecting DNAm at imprinted loci, which exhibit parent-of-origin expression regulation, can result in imprinting disorders with developmental impairments such as Prader-Willi and Angelman syndrome (138, 139). Collectively, these examples support the growing body of evidence which implicates epigenetic dysfunction in neurodevelopmental pathologies.

Another common group of neurodevelopmental disorders is autism spectrum disorders (ASD), which are characterized by severe deficits in social interaction, communication and behavioral patterns that are restrictive and stereotypical (140). Increased interest in epigenetic

profiling in ASD samples has been prompted by the observation that there is notable overlap in clinical features and symptoms between ASD and other neurodevelopmental disorders, such as Angelman, Fragile X, and Rett syndromes, which are caused by epigenetic perturbations or mutations in genes encoding epigenetic factors (141). Candidate gene DNAm studies in ASD using superior temporal gyrus and cerebral cortex samples previously reported increased promoter DNAm at the *RELN* and *MECP2* genes, respectively (142, 143). On a genome-wide level, interrogation of DNAm in ASD dorsolateral prefrontal cortex, temporal cortex and cerebellum samples using the Illumina 450K array revealed numerous differentially methylated sites enriched in areas of low CpG density as well as 4 differentially methylated regions (DMRs), which reached genome-wide significance, with 3 out of 4 DMRs replicating in an independent cohort (144, 145).

Beyond brain-derived DNAm measures, there is growing interest in exploring DNAm associations to neurodevelopmental disorders in other tissues. For example, the analysis of DNAm profiles in buccal epithelial cells (BECs) of children with fetal alcohol spectrum disorder (FASD), which is characterized by growth retardation, facial dysmorphologies, neurological abnormalities and intellectual impairments, revealed widespread FASD-associated alterations in the BEC methylome (146, 147). In the context of ASD, differential DNAm was observed in BECs at genes that were expressed in the brain and encoded protein products that were previously implicated in ASD pathology (148). Moreover, DNAm measures in other peripheral tissues such as blood and saliva have identified links between DNAm variation and childhood neurodevelopmental outcomes, demonstrating the potential for DNAm biomarkers of neurodevelopmental impairment in these accessible tissues (149–152). Finally, the use of cord blood and placental DNAm profiles have allowed for the exploration of early life biomarkers of pre- and early postnatal stress exposures on fetal or infant brain development (152, 153).

1.4.2 DNAm signatures of neuropsychiatric disorders

Increasing interest in epigenomic contributions to complex, environmentally-conditioned neuropsychiatric diseases has been potentiated by the growing recognition that epigenetics may serve as a mediator between genomic variation and environmental influences (154–158). Specifically, epigenetic associations have been investigated for a number of different psychiatric

conditions, including schizophrenia, bipolar disorder, major depressive disorder and various addiction disorders (154–156).

Candidate gene analyses in postmortem brain have reported significant DNAm alterations associated with schizophrenia and bipolar disorder at specific loci including *RELN*, *COMT*, *SOX10, HTR2A, HTR1A, BDNF, HCG9, KCNQ3* and *DAT1* (159–168). However, these candidate DNAm associations have not been consistently replicated, likely owing to differences in methodology, the specific CpG dinucleotides interrograted, brain regions examined and clinical populations from which the postmortem samples were collected (169–171). In order to achieve more comprehensive and unbiased assessments of DNAm variation in major psychosis, one of the earliest genome-wide analyses of psychosis used CpG-island micorarrays to analyze postmortem frontal cortex tissue from schizophrenic, bipolar disorder and control subjects, identifying psychosis-associated DNAm alterations in genes related to glutamatergic and GABAergic neurotransmission, neuronal development and metabolism (172). These results were further corroborated by a subsequent finding that *DNMT1* and *TET1*, two genes encoding enzymes that respectively methylate and hydroxymethylate CpGs, were overexpressed in the brain of schizophrenic and bipolar disorder patients and that DNMT1 showed increased binding to a subset of GABAergic (i.e. *GAD1*) and glutamatergic (i.e. *BDNF*) gene promoters in the cerebral cortex but not the cerebellum (173). More recent genome-wide DNAm profiling using commercial microarrays, namely Illumina 27K and 450K arrays, in prefrontal cortex, dorsolateral prefrontal cortex, frontal cortex, cerebellum and hippocampus tissues have reported widespread DNAm aberrations associated with schizophrenia at numerous genes involved in neurodevelopmental processes and GABAergic neurotransmission, including a large number of *cis*-mQTLs which overlapped with risk SNPs implicated in schizophrenia (81, 82, 174–179). Taken together, these findings implicate DNAm alterations related to early neurodevelopmental programs and neurotransmission regulation, along with genetic risk variants, as potential contributors to schizophrenia and bipolar disorder pathophysiology.

In addition to schizophrenia and bipolar disorder, DNA modifications have also been assessed in the context of major depressive disorder (MDD) and addiction disorders. Specifically, high-throughput microarray profiling of ~3.5 million CpGs in postmortem frontal cortex MDD samples identified 244 MDD-associated DMRs which were highly enriched for neuronal growth and development genes (180). Hippocampal DNAm differences in male

offspring were also observed in relation to maternal depression during pregnancy as well as increased DNAm at specific loci (i.e. *BDNF*) in Wernicke area in relation to suicidal behavior, signifying epigenetic associations with a broad range of depression-related behavioral phenotypes (181, 182). More recently, DNAhm variation was assessed in postmortem prefrontal cortex samples of depressed individuals, although no individual 5hmC site reached genome-wide significance (183). In regards to addiction disorders, co-methylated modules enriched for genes involved in neural development and transcriptional regulation were identified in prefrontal cortex tissues of male alcohol dependent patients (184). In addition, *cis*-mQTLs underlying the prodynorphin gene (*PDYN*) were associated with alcohol dependence status in dorsolateral prefrontal cortex tissues (185). Overall, these studies provide compelling evidence for the potential role of brain-specific DNA modifications in mediating the combined contribution of genetic and environmental factors in neuropsychiatric disorders.

The identification of epigenetic biomarkers associated with psychiatric conditions represents a prominent focus in neuropsychiatric disease research and has largely been explored through methylomic profiling of various surrogate tissues, including BECs and peripheral blood. In blood leukocytes, candidate gene analyses have examined DNAm associations to MDD in various genes including *BDNF*, immune-related markers and *GLUT1/4*, while genome-wide analyses in buccal samples of discordant monozygotic (MZ) twins revealed differential DNAm associated with adolescent depression at a novel CpG underlying the *STK23C* gene, which encodes a serine/threonine kinase of unknown function (186–188). In the context of schizophrenia, genome-wide scans of blood DNAm identified 172 novel site-specific associations between DNAm and schizophrenia, after adjustment for potential confounding factors such as age, sex, ethnicity, smoking, batch and cell type heterogeneity; importantly, these associations were subsequently validated in an independent replication cohort (189). Furthermore, systematic integration of genetic and epigenetic variation in whole blood of schizophrenic patients demonstrated the co-localization of allelic associations for schizophrenia and differential DNAm, which was primarily enriched for immune-related loci (190). Finally, blood-based EWAS analyses for chronic alcoholism reported decreased global DNAm levels in peripheral lymphocytes, with gene-specific hypermethylation at *vasopressin, DAT, HERP* and *SNCA*, in alcohol dependent patients over controls, although the majority of these analyses did not account for potential blood cell-type proportion differences, thereby limiting the

interpretability of the findings (191–195). Taken together, these results highlight the potential for DNAm measures to serve as biomarkers of neuropsychiatric disease risk or progression.

1.4.3 DNAm variation in neurological malignancies

The characterization of epigenetic modifications, particularly DNAm alterations, in cancer has helped advance our understanding of tumour biology as well as prompted the development of prognostic biomarkers for various cancers. In general, tumours exhibit global hypomethylation of their DNAm profiles, which has been associated with chromosomal instability and increased tumour frequency, along with gene-specific hypermethylation, which has been associated with transcriptional silencing of key tumour suppressor genes such as *RB1*, *PTEN* and *MLH1,* amongst others (196–201). While there are observed exceptions to this pattern, these findings of widespread epigenetic changes have contributed to the formulation of a model which suggests that some genes are epigenetically disrupted at the earliest stages of tumourigenesis, even before mutations, causing altered differentiation throughout tumour evolution (202). More recently, this model has been refined to describe three types of genes implicated in epigenetic dysregulation in cancer: 1) 'epigenetic mediators', which are disrupted in function or expression early in tumour development, 2)'epigenetic modifiers', which are responsive to changes in the cellular environment and often linked to nuclear architecture and 3)'epigenetic modulators', which are directly involved in the dysregulation of epigenetic machinery in cancer (203). To date, considerable work has been made to examine DNAm variation in various tumour types, including, pediatric and adult neurological cancers (204).

In the context of pediatric brain tumours, highly prevalent blocks of hypomethylation are linked to increased transcription of various genes (ie *RUNX2*, *OTX2*) that are commonly mis-expressed in certain subtypes of medulloblastoma, a malignant tumour type of the cerebellum that primarily affects children (205). In another form of childhood brain cancer, neuroblastoma, genome-wide DNAm analyses revealed the presence of distinct CpG methylation patterns that associated with survival outcomes and other molecular features such as oncogenic *MYCN* amplification as well as aberrant hypermethylation of known or candidate tumour suppressor genes (ie *TERT*, *CAMTA1*, *CHD5* and *KIF1B*) (206–208). The DNA methylome within a subset of infant hindbrain tumours, known as ependymomas, displays aberrant hypermethylation at CpG islands, specifically at loci encoding ESC targets regulated by the Polycomb repressor

complex 2 (PRC2), suggesting that epigenomic changes could underlie the disruption of cell state and differentiation processes associated with ependymoma development (209). Similar to pediatric brain tumours, adult brain malignancies, particularly gliomas, also exhibit patterns of global hypomethylation with site-specific hypermethylation, which largely occurs at promoter CGIs (210). For example, low-grade gliomas often possess a hypermethylation signature at 1503 promoter CpGs, called the glioma-CpG island methylator phenotype (G-CIMP), which disproportionately occurs in conjunction with mutations in the epigenetic modulator genes *IDH1/2* and has been correlated with improved clinical prognosis over non-G-CIMP tumours (211, 212). In a high-grade form of adult glioma, known as glioblastoma multiforme (GBM), widespread but variable loss of DNAm is observed, affecting up to 50% of CpG sites across the genome, along with aberrant hypermethylation in various genes such as *HOXA11, CD81, PRKCDBP, TES, MEST, TNFRSF10A* and *FZD9* (213, 214). DNAm profiling in glioma cell lines have largely corroborated findings of aberrant gene-specific hypermethylation, although in certain instances, data from primary gliomas and subsequent xenograft models were not consistent with *in vitro* culture results, thereby signifying potential differences in DNAm measures from different tumour-derived sources (215–221). Finally, promoter DNAm-mediated silencing of the DNA repair gene *MGMT* in gliomas, particularly GBM tumours, is linked to favourable response to alkylating chemotherapy treatment and is widely accepted as an effective predictive biomarker for glioma management (222–225). Beyond its potential as a prognostic marker, DNAm-based signatures have also been applied to define specific glioma subtypes as well as more broadly classify tumours of the central nervous system (226–228). Taken together, these results underscore the potential for DNAm measures to contribute to the molecular characterization of neurological malignancies as well as serve as predictive biomarkers for neuro-oncological care.

1.4.4 DNAm profiles in neurodegenerative diseases

Studies of DNA modifications in human neurodegenerative disorders have become increasingly prevalent due to their potential to provide insights in potential molecular underpinnings of these degenerative brain disorders (229). Specifically, DNA modifications have been assessed in the context of various neurodegenerative pathologies including Huntington's Disease (HD), Alzheimer's disease and Parkinson's Disease (PD).

In the context of HD pathogenesis, aberrant DNAm has been implicated in HD-specific transcriptional dysregulation as well as age-related HD degeneration (230). Specifically, early work using MeDIP in postmortem putamen specimens reported increased DNAm and reduced DNAhm levels in the 5'UTR of a candidate gene, *ADORA2A,* a gene which exhibits reduced expression in HD patients over age-matched controls (231). DNAm profiling in mature sperm cells revealed substantial intra- and inter-individual DNAm variation at tested sites in the promoter of the causative *huntingtin* (*HTT*) gene locus (232). Another study performed in peripheral blood samples reported a lack of association between *HTT* DNAm and age of disease onset, although more recent analyses performed in various brain cortical regions demonstrated an association between HD status and increased epigenetic age acceleration, a hypothesized measure of biological aging (233, 234).

Apart from HD, DNAm alterations have also been implicated in Alzheimer's disease brain pathology with early immunoblotting analyses showing reductions in global DNAm and DNAhm from entorhinal cortex, temporal neocortex and hippocampus of Alzheimer's disease patients over controls, although attempted replication of these findings have produced conflicting results (235–240). Other candidate gene studies in Alzheimer's disease postmortem brain tissues reported significant disease-associated differential DNAm at various genes implicated in Alzheimer's pathogenesis including *APP, GSK3B, MAPT, PP2AC, APOE, DNMT1, MTHFR, PGC-1α* and *TREM2,* amongst others (241–247). Genome-wide DNAm profiling in superior temporal gyrus tissues using the Illumina 450K array found numerous differentially methylated regions, the majority of which showed Alzheimer's-related increases in DNAm over controls, with notable overlap of hits to previous studies (248–250). Beyond brain-specific DNAm analyses, work in peripheral blood leukocytes has reported Alzheimer's-associated differential DNAm at specific gene promoters, such as *BDNF* and *OPRK1*, although these findings have had limited reproducibility and remain largely inconclusive (251–253).

In regards to PD, early candidate gene studies of DNAm in postmortem substantia nigra, putamen and cortex samples revealed disease-related DNAm decreases at intronic CpGs of *SNCA* and *TNFα*, key genes associated with PD risk, although these associations has not been fully replicated or may be specific to distinct brain regions (254–257). Interestingly, sequestration of nuclear DNMT1 by α-synuclein, encoded by *SNCA*, in brain samples from PD and Lewy body dementia patients has been attributed to global loss of DNAm at numerous genes

including *SNCA* itself, *SEPW1* and *PRKAR2A* (258). On a genome-wide scale, DNAm profiling in postmortem brain samples have identified DNAm changes associated with PD risk variants at *PARK16*, *GPNMB* and *STX1B* genes, signifying possible combined contribution of genetic and epigenetic variation on PD pathophysiology (259). Another DNAm study performed in putamen and cortex showed decreased DNAm and concomitant increase in expression of *CYP2E1* in PD brain over controls (260). In addition, comparative analysis of DNAm profiles from matched brain and blood samples of PD patients and healthy controls revealed widespread differential DNAm at genes previously associated with PD pathology, including sites that exhibited high blood-brain DNAm concordance, suggesting that these associations may serve as potential blood-based PD biomarkers (261).

Finally, DNA methylomes of prefrontal cortex samples from different neurogenerative disorders (i.e. Alzheimer's disease, PD, and Dementia with Lewy bodies) showed that the similar aberrant CpG methylation patterns across different disease entities targeted a defined gene set, signifying that common epigenomic alterations may contribute to distinct neurodegenerative states (262). Overall, these results provide compelling evidence for the role of DNA modification variation in neurodegenerative diseases, particularly as it pertains to neuronal-specific pathology.

1.5 Various tissue sources to study DNAm variation in neurobiological diseases

Increasing interest in epigenomic profiling of neurobiological disease has dually prompted the rapid rise of EWAS analyses in neuropathological disease cohorts as well as the exploration of different tissue sources to examine disease-associated DNAm variation. In regards to the latter trend, various tissue sources have been used in neurobiological EWASs including target brain tissues, brain-derived cell culture models and peripheral surrogate tissues. Importantly, each of these tissue sources offer unique opportunities and challenges in regards to feasibility and other methodological constraints as well as the interpretation of identified associations.

1.5.1 Target brain tissues

As neurobiological diseases primarily manifest in the brain, the use of human brain samples is logically the most appropriate tissue source for DNAm studies of neuropathological disorders. Importantly, the use of target brain tissues provides the unique opportunity to profile

DNAm alterations which are more likely to be mechanistically implicated in disease risk or pathogenesis over those measured in surrogate tissues. As a result, brain tissue is often regarded as the "gold standard" for DNAm analyses in neurobiological diseases (8). Beyond imaging-based applications, it is currently not possible to perform in-depth molecular characterizations of the brain epigenome in living subjects, therefore only postmortem brain samples have been used to date for human DNAm studies (263). Despite recent advances in brain banking efforts, the limited availability and applicability of postmortem brain tissues present a number of challenges and key considerations for DNAm studies performed in human brain samples (264).

One of the major problems with epigenetic analyses in brain tissues is the relative scarcity of suitable material and the lack of replication potential. The limited availability of well-characterized brain tissue collections restricts cohort sample sizes such that studies have low power to detect or replicate efforts, especially if effect sizes of disease-related DNAm associations are subtle (8). Specifically, it is challenging to obtain appropriate and adequate control tissue samples that are matched to case samples for various demographic factors such as age, sex, genetic ancestry and prior exposures including smoking, which represent potential confounding variables in EWAS analyses (265). In particular, controls are often skewed towards older individuals, which may lead to age-related confounds in identified DNAm associations if not appropriately controlled for in the study design and analysis (8). Moreover, case classifications are frequently poorly defined and the absence of disease symptoms or relevant co-morbities are not always clinically confirmed (266). These methodological constraints highlight the need for increased coordination, collaboration and consensus in the collection and appropriate characterization of brain tissue samples for epigenetic studies.

Beyond sample size limitations, the nature of postmortem sampling has implications for the interpretation of DNAm measures. As brain tissues are sampled after death, often after the disease has occurred, DNAm in these tissues can only be measured retrospectively, with no opportunity for longitudinal tracking of DNAm variation in pre-symptomatic stages or across disease progression within an individual. Such retrospective postmortem sampling precludes the possibility of making causal inferences from identified DNAm associations as the establishment of causal connections generally require prospective, often longitudinal, analyses (8). Moreover, DNAm patterns detected in postmortem tissues may represent variation that arises as a consequence of the disease and may be confounded with other biological signals such as cause of

death, tissue pH, agonal state or postmortem interval (267–270). Careful consideration of these pre- and postmortem factors during sampling, storage and analysis may help to discern disease-related DNAm patterns from unrelated technical variation.

1.5.2 Brain-derived cell culture models

Brain-derived cell culture systems have been extensively used in neuroscience research as tools to study brain disease pathogenesis, often through the direct manipulation of molecular targets and pathways, as well as a platform for drug screening (271). These models can comprise of 2-dimensional primary cultures with mixed cell lineages isolated from healthy or diseased brain resections at different developmental stages or immortalized cell lines of neuronal, microglial, astrocytic and oligodendrocytic cell types (272, 273). Alternatively, 3-dimensional neural cultures can be established by proliferating neural stem cells (NSCs) and neural progenitor cells (NPCs) in non-adherent, serum-free growth medium, stimulating the formation of clonal spheres, known as neurospheres (274). Neurosphere formation have been widely used as an *in vitro* culture system to study neurogenesis and neural development (275–277). The neurosphere assay has also been used to isolate and expand stem-like cell populations from resected brain tumours, particularly gliomas, known as brain tumour-initiating cells (BTICs) (278, 279). As BTICs are hypothesized to drive tumour formation and treatment resistance, while preserving the molecular and phenotypic properties of the parental tumour, the BTIC culture system represents a valuable model for *in vitro* examination of tumour biology (280–282). An advanced application of 3D culture techniques has generated cerebral organoids, or ‘mini-brains’ which recapitulate the organization and cell-specific epigenomic profiles of the developing human brain and are rapidly becoming an essential tool for human neurobiological disease modeling (283–285).

Although cell culture models are useful tools for in-depth molecular characterization and manipulation of disease-relevant pathways, these *in vitro* systems also present a number of obstacles, particularly in regards to DNAm analyses. Firstly, cell culture-induced artifacts can contribute to technical noise in DNAm measures. A previous study which examined RRBS data from a diverse collection of 82 human cell lines and tissues, including matched primary cell lines and tissues, reported cell culture-induced DNAm differences occurring predominately in intragenic CpGs of genes involved in cell proliferation (286). A secondary challenge of *in vitro*

models is that they are generally unable to recapitulate the complex interactions between various cell types and the local microenvironment, which may, directly or indirectly, contribute to disease pathogenesis (287). For example, glial cells play a key role in neuronal development and function, relying on cellular interactions that likely cannot be mimicked *in vitro* (288, 289). These results underscore the need to verify *in vitro* findings in *in vivo* systems as well as standardize cell culture practices to minimize technical variation.

1.5.3 Accessible peripheral tissues

In addition to understanding human DNAm variation in target brain tissues, there is increasing interest in studying neuroepigenetic alterations in peripheral tissues as they relate to brain-derived measures or phenotypes. This has been largely motivated by the fact that postmortem human brain tissues are relatively scarce, particularly in pediatric cohorts, in comparison to more readily accessible tissues that may be used as surrogates, such as blood, saliva, or buccal swabs. The increased availability of these surrogate tissues often allows for large sample sizes and greater replication potential. Furthermore, peripheral tissues can be used for longitudinal tracking of DNAm changes associated with phenotypes of interest across time or treatment interventions. As such, readily accessible peripheral tissues have become increasingly attractive for the exploration of epigenetic biomarkers, which may aid in early disease detection and monitoring (290).

A key pre-requisite in the use of surrogate tissues for epigenetic biomarker discovery of brain-related phenotypes is that epigenetic patterns from peripheral tissues should reflect those in the brain. Recent studies have sought to address this by exploring DNAm concordance between matched human blood and brain tissues, reporting mixed patterns in which some DNAm sites are highly concordant between tissues while others are discordant (59, 81, 291, 292). These results have shown that variable CpGs are more likely positively correlated in DNAm signal between blood and brain, although such concordance between tissues is likely enriched for genetically-mediated DNAm (59, 81, 291, 292). Importantly, various publically available resources have been created to provide DNAm profiles of matched blood and brain tissues at individual CpGs and allow researchers to discern if DNAm patterns observed in blood are similar in brain (81, 292). This approach has been demonstrated in previous studies of psychosis in discordant monozygotic twins to show that blood-based DNAm findings could be replicated in postmortem

brain tissues (293, 294). However, other studies comparing DNAm signal between matched blood and other surrogate tissues, such as saliva or buccal epithelial cells (BECs), have argued that alternate peripheral tissues which share common ontogenetic properties with brain may serve as more appropriate surrogates than blood; for example, it is proposed that DNAm profiles from BECs may be more suitable proxies for brain DNAm as they both arise from the ectodermal germ layer (68, 148, 295). Additional cross-tissue comparisons of DNAm variation may further contribute to the evaluation of surrogate tissues and promote epigenetic biomarker discovery in the context of human neurobiological phenotypes.

1.6 Dissertation overview

The overarching goal of this thesis was to implement and assess various tissue-specific approaches to investigate DNAm variation across different types of neurobiological disease states including neurodegenerative, neuropsychiatric and neurological pathologies. The experimental data will be presented through four separate chapters, as outlined in Section 1.1. Specifically, I performed four separate studies to characterize disease-associated DNAm from a) postmortem brain samples, b) primary brain-derived cell culture models and c) accessible peripheral tissues. In **Chapter 2**, I examine DNAm patterns related to Huntington's disease pathogenesis and tissue-specific *Huntingtin* gene expression in postmortem human cortex samples. In **Chapter 3**, I compare DNAm profiles from GBM tumours and matched primary cell cultures enriched for BTIC populations, identifying a homeobox-enriched signature of differential DNAm between the paired samples. Beyond brain-specific DNAm patterns, in **Chapter 4**, I also explore the use of a disease-relevant blood cell type, $CD3^+$ T-lymphocytes, to identify DNAm changes related to alcohol dependence in patients undergoing a clinical intervention. For **Chapter 5**, I investigate DNAm variability and the influence of genetic variation on DNAm in peripheral blood and buccal epithelial cells from two pediatric cohorts, highlighting a number of potential considerations and practical implications for the appropriate design and interpretation of early-life EWAS analyses in these tissues. Finally, in **Chapter 6**, I draw on the main findings from each data chapter to provide an integrated discussion of limitations and future considerations for DNAm studies of neurobiological disease states.

Chapter 2: DNA methylation profiling in human Huntington's disease brain

2.1 Background and rationale

Huntington's disease (HD) is a progressive adult-onset neurological disorder with an estimated worldwide prevalence of at least 2.71 per 100 000 individuals (296). Clinically, HD is defined by characteristic motor disturbances, most notably chorea, which typically begins to present when the individual is of 30 to early 40 years of age (297). Motor and cognitive deficits generally progress in a predictable manner until the decline of the individual ~20 years after the age of onset (298). Genetically, HD is linked to a single causative mutation, a polyglutamine-coding CAG repeat expansion in the first exon of the huntingtin (*HTT*) gene (299). Wild-type alleles of *HTT* carry a CAG repeat length of ~17–20 repeats, whereas repeat lengths of >35 are considered pathogenic (300). However, repeat lengths between 35 and 39 are considered intermediate alleles, as they do not display almost complete penetrance seen with lengths >40 (301). Importantly, the CAG repeat length has been inversely correlated to the age of onset, in which individuals with longer CAG repeats typically have younger ages of onset than those with shorter repeat lengths (302, 303). Although this inverse correlation with the CAG repeat length accounts for 60– 70% of the variability in the age of onset seen in HD patients, the remaining ~30% of the age of onset variability is thought to be attributed to genetic, epigenetic and/or environmental variation (304, 305). For example, a single-nucleotide polymorphism (SNP) in a nuclear factor kappa-light-chain-enhancer of activated B cells (NF-κB)-binding site within the *HTT* promoter has been recently shown to alter the transcriptional activity of the *HTT* promoter and to affect the age of onset in HD, increasing or decreasing the age of onset depending on its presence on the wild-type or mutant *HTT* allele, respectively (306). Such findings illustrate the role of gene dosage, specifically the ratio of wild-type to mutant *HTT*, in HD pathogenesis as well as highlight the importance of elucidating additional HD disease modifiers that may affect *HTT* gene expression.

Gene expression patterns of *HTT*, particularly in relation to HD pathogenesis, have been widely studied, although its transcriptional regulation remains poorly understood (307–309). Despite being associated with a primarily neurological disorder, *HTT* itself is a ubiquitously expressed gene, with highest expression found in brain and testes tissues (307–309). In addition,

it has been suggested that *HTT* gene expression is subject to inter-individual variation, as a study of wild-type, inbred, male mice found considerable inter-individual variability of mouse *HTT* (Hdh) expression in forebrain (309). Aside from its wild-type expression patterns, mutant *HTT* has been associated with transcriptional dysregulation, particularly in brain tissues, and has been found to aberrantly interact with various transcription factors (310–314). Collectively, these results highlight the importance of understanding the molecular mechanisms that underlie *HTT*-associated transcriptional gene regulation. Given that *HTT*-related expression patterns exhibit both tissue-specific and inter-individual variation, it has been suggested that epigenetic modifications may play an important role, particularly in the context of HD neuropathology (230).

Epigenetics refers to mitotically heritable modifications to DNA and DNA packaging that alters the accessibility of DNA and potentially regulates gene transcription without changing the underlying DNA sequence (13). DNA methylation is arguably the most studied epigenetic mark and involves the covalent attachment of a methyl group onto the 5′ carbon of cytosine, typically at CpG dinucleotides (26, 29, 315). CpG sites are nonrandomly distributed across the human genome and tend to be enriched in regions known as CpG islands (26, 29). The relationship between DNA methylation and gene expression, which is now recognized to be more complex than originally thought, is often dependent on genomic context (26, 38). For example, DNA methylation at gene promoters is generally associated with gene expression silencing; however, its role within gene bodies may be more variable, including splicing-related regulation at introns (26). Importantly, tissue-specific expression patterns are largely associated with tissue-specific DNA methylation (43, 316). Indeed, it has been shown that the main drivers of DNA methylation variance are tissue identity followed by cellular heterogeneity within a tissue (59). Moreover, DNA methylation patterns are subject to inter-individual variability in a manner that is tissue-specific (317). Taken together, we hypothesize that DNA methylation serves as an integral feature of *HTT* transcriptional regulatory circuitry and may play a critical role in establishing individual-specific and tissue-specific patterns of *HTT* gene expression.

In the context of HD pathogenesis, aberrant DNA methylation has been proposed to potentially contribute to transcriptional dysregulation observed in HD brain tissues (230). Indeed, a genome-wide study using cell lines derived from transgenic HD mouse striatal neurons found several transcriptionally dysregulated genes to have aberrant DNA methylation (318). An

additional study conducted in human brain samples, specifically the striatum, identified altered DNA methylation patterns in the adenosine A (2A) receptor (*ADORA2A*), a gene that shows reduced expression in HD patients (231). Although these studies draw a potential link between DNA methylation and HD-related transcriptional dysregulation, they did not account for cell-type heterogeneity in the DNA methylation profiles across tissue specimens nor did they address how the HD mutation may impact local DNA methylation patterns at the *HTT* locus itself. Related to the latter point, a previous report did describe intra- and inter-individual DNA methylation variation in the *HTT* locus in sperm cells, although the functional consequences and maintenance of this differential DNA methylation after fertilization have yet to be explored (232). Another study that focused on the *HTT* locus reported a lack of association between *HTT* DNA methylation and age of disease onset, although their analyses were performed in human peripheral blood samples as opposed to HD brain tissues (233). Overall, studies of disease-related alterations in genome-wide DNA methylation profiles of human HD brain tissues as well as local DNA methylation variation at the *HTT* gene locus across somatic tissues represent unaddressed areas of inquiry that may further advance our understanding of *HTT* transcriptional regulation in HD pathology.

In this study, we sought to better understand the role of DNA methylation both in human HD pathogenesis and in the transcriptional regulation of the *HTT* locus. Specifically, we aimed to address two main research questions: (i) Are there genome- wide DNA methylation changes that arise due to HD-related pathogenesis in human brain tissues? (ii) Are there DNA methylation differences at the *HTT* gene locus which contribute to tissue-specific *HTT* gene expression patterns? Here we present a systematic assessment of HD-associated DNA methylation alterations in human cortex samples. While the primary brain structures affected by the HD mutation are the caudate and putamen (collectively known as the striatum), secondary pathogenesis occurs in additional structures, including the cortex, as the disease progresses (319, 320). Owing to the advanced pathogenic stage of the majority of the HD individuals in our cohort, it is likely that the remaining striatal tissues would be significantly altered in cell-type heterogeneity and the amount of discernable striatum tissue remaining compared with control individuals, prompting us to select cortex as our primary brain tissue for this study. While we found limited evidence of HD-associated DNA methylation changes in human cortex tissues after adjustment for cell-type heterogeneity, we observed a general association between DNA

methylation and age of disease onset in HD cortex samples. We also sought to characterize how DNA methylation variation may contribute to tissue-specific transcriptional regulation at the *HTT* gene locus. Notably, we identified site-specific differential DNA methylation between matched cortex and liver samples that spanned the *HTT* gene region, particularly underlying a newly discovered CTCF-binding site within the *HTT* proximal promoter. Collectively, our results suggest that DNA methylation may be associated with the age of disease onset in cortex samples, although we were unable to detect HD-associated DNA methylation at probed sites. Moreover, our data suggest that DNA methylation may, in part, contribute to tissue-specific *HTT* transcription through differential CTCF binding.

2.2 Materials and Methods

2.2.1 Human samples

Human samples were taken from the UBC HD Biobank in accordance with institutional ethics policies. HD individuals in cohort were selected based on the availability of cortex and liver tissues, availability of a control individual matched for age and no known concomitant conditions. These samples were frozen at the time of their collection and stored at −80°C for long-term storage. The human samples used, age, sex, CAG size on both *HTT* alleles (calculated using methods in (321)) and PMI are given in Table 2.1. DNA methylation age for these samples was calculated using a well-established online epigenetic age prediction software (322).

2.2.2 Mouse samples

Mouse cortex samples were isolated from FVB, littermate, male mice at 3 months of age in accordance with institutional ethics policies. Tissues were then flash frozen following dissection and stored at -80°C until processed for RNA.

2.2.3 DNA isolation and DNA methylation arrays

Genomic DNA was isolated using the DNeasy Blood & Tissue kit (Qiagen) as per manufacturer's instructions. DNA was then purified and concentrated using the DNA Clean & Concentrator kit (Zymo Research). Sample yield and purity was measured after each step using a Nanodrop ND-1000 (Thermo Scientific). 750 ng of DNA was used for bisulfite conversion using

the Zymo Research EZ DNA Methylation Kit (Zymo Research). Following bisulfite conversion, samples were randomized and 160 ng of bisulfite-converted DNA was applied to the Illumina Infinium HumanMethylation450K Beadchip array, as per manufacturer's protocols (Illumina) (129).

2.2.4 450K data quality control and normalization

Raw intensity values from the arrays were imported into Illumina GenomeStudio V2011.1 software and subjected to initial quality control checks for array staining, extension and bisulfite conversion followed by color correction and background adjustment using control probes contained on the 450K array. Subsequent processing and analysis were performed in R (Version 3.0.1)(323). Probe filtering was performed in which 11 648 probes on the sex chromosomes and 65 probes targeting SNPs were removed from the dataset. Additionally, probes with detection P-values >0.01, probes with missing beta values and probes for which less than three beads contributed to the signal in any sample were eliminated (a total of 6144). Recent re-annotation of the Illumina 450K array was used to filter 30 685 probes that are known to be polymorphic at the target CpG (20 150), or probes which have nonspecific *in silico* binding to the sex chromosomes (10 535)(324). Together, these measures eliminated 48 542 probes, leaving a total of 437 035 probes for further analysis. Following data subsetting to produce a cortex-only dataset and a matched tissue (cortex and liver) dataset, quantile normalization was conducted using the lumi R package, after which subset-within-array normalization (SWAN) was used to correct for probe-type differences (325, 326). ComBat was then used to remove chip-to-chip effects that were detected in the cortex-only dataset, which was run across two chips (327). The 450K data has been made publicly available (GEO GSE79064).

2.2.5 Principal component analysis and neuron/glial cell-type correction

PCA is a data-reduction method that can be used to decompose the measured methylation signals into a set of linearly independent principal component (PC) patterns that are ranked according to how much variance in the data they represent. The top-ranked PCs in methylation data can often be correlated with known traits in the cohort, such as tissue type, cellular composition, or disease state (328). To investigate the effect of cell type, particularly in our cortex-only dataset, we utilized the CETS algorithm which estimates neuron versus glia

proportions in brain tissue samples using methylation profiles of 10 000 brain cell-type-specific Illumina 450K probes (note: these probes were subsequently filtered out of the datasets resulting in a final count of 427 242 probes for differential methylation analysis) (69). Subsequent PCA of our cortex-only dataset revealed that the top-ranking PC (accounting for 35.6% of variance in the methylation profiles) was significantly associated with neuron/glial cell-type proportions. To account for changes in DNA methylation due to differences in inter-individual brain cell-type composition in the cortex samples, the methylation values for each of the probes were regressed on the estimated cell-type (i.e. neuron) proportions for each sample, as previously described (329). The residuals of each regression model were applied to the mean value of each data series to obtain the 'corrected' methylation data. PCA was subsequently used to check that the correlation of the cell-type proportions were minimal (i.e. correlated to PC12 accounting for a negligible 0.00043% of the variance) in the cortex-only dataset. PCA was additionally used to check for correlation of other known meta-variables (i.e. sex, age of death, age of HD onset, disease status, PMI, chip) with the underlying methylation patterns of the cortex-only and matched datasets, respectively. Note that for all PCA analyses, the top-ranking PC (denoted as PC0) was negated as it is not informative of inter-individual variance in the DNA methylation data (38).

2.2.6 Differential methylation analyses of 450K data

The 'Bumphunting' method was implemented for the detection of DMRs in HD cases versus controls (330). Differentially methylated probes were identified using the R limma package's moderated t-statistics with empirical Bayesian variance estimation in the genome-wide analyses (331). In the candidate gene region (*HTT*; chr4: 2973107-3258169) analyses, individual linear models were fit for each of the 87 probes underlying the region. Specifically, in the cortex-only dataset, a linear model was fit for each probe's methylation measures with HD status as the main effect. To reduce the burden of multiple test correction for analyses in our limited sample number of HD cortex tissues, a pre-filtering step was applied to remove invariable probes with an inter-quantile range (from 10th to 90th percentile) < 0.05. This left 83 939 variable probes in which individual linear models were fit to test the association between the age of onset (adjusted for sex) and DNA methylation in HD cortex tissues. In the matched dataset, differentially methylated probes between cortex versus liver samples were identified using paired analysis for

individual linear models that were fit for each of the 87 probes underlying the region. For all tests, the resulting P-values were adjusted using the Benjamini–Hochberg false discovery rate (FDR) method (332). A power calculation was conducted using an established method designed to perform sample size estimations for microarray data while controlling for false discovery rate (333). All statistical analyses were performed on transformed M-values (334).

2.2.7 Pyrosequencing validation

PyroMark Assay Design 2.0 (Qiagen) software was used to design the bisulfite pyrosequencing assays (primers are listed in Supplementary Table 2.2). HotstarTaq DNA polymerase kit (Qiagen) was used to amplify the target region using the biotinylated primer set with the following PCR conditions: 15 min at 95°C, 45 cycles of 95°C for 30s, 58°C for 30s, and 72°C for 30s and a 5 min 72°C extension step. The amplicon was then electrophoresed on an agarose gel for confirmation of both the presence and quality of the product. Streptavidin-coated Sepharose beads were bound to the biotinylated-strand of the PCR product and then washed and denatured to yield single-stranded DNA. Sequencing primers were introduced to allow for pyrosequencing (PyromarkTM Q96 MD pyrosequencer, Qiagen). Pyro Q-CpG software (Qiagen) was used to generate quantitative methylation levels of the targeted CpG dinucleotides of interest. Primers used for all pyrosequencing assays are listed in Supplementary Table 2.2.

2.2.8 Quantitative real time PCR (RT-qPCR)

Mouse, human and HEK293 cell line samples were processed for RNA extraction using the protocol detailed in the PureLink® RNA mini kit (Invitrogen) with the following modifications: (i) tissue homogenization was achieved using a Fastprep Homogenizer (Thermo Scientific), (ii) homogenization of HEK293 cells was achieved using a 22 gauge needle and syringe, (3) to increase RNA yield and purity, DNase was used to degrade any residual genomic DNA in the prep column using the PureLink® DNase set (Invitrogen). For human and mouse individuals, three samples of each tissue type were processed and analyzed to ensure the RT-qPCR results reflected gene expression across the tissue. The concentration and purity of RNA was assessed using a Nanodrop spectrometer (Thermo Scientific). Reverse transcription was performed using the SuperScript® VILOTM cDNA synthesis kit (Invitrogen). Quantitative analysis of mRNA expression was performed using Fast SYBR®Green master mix according to

the manufacturer's instructions (Applied Biosystems). Amplification of cDNA was performed using the StepOne Plus real-time PCR system (Applied Biosystems). Primers used are provided in Supplementary Table 2.1. Quantification of mRNA levels was calculated using the standard curve method with 10-fold serial dilutions comprising a portion of each sample used in the study. Normalization of the quantified mRNA levels was accomplished using a normalization factor generated by the GeNorm program provided in the qBase software package (Biogazelle). The normalization factor was generated on a per sample basis, using amplification of a series of control genes in separate wells.

2.2.9 Identification of CTCF binding site using available ChIP-seq datasets

Using available ChIP-seq datasets for CTCF from ENCODE and PAZAR and the transcription factor (TF) binding profile for CTCF from JASPAR, the proximal promoter region of *HTT*, chr4: 3074800–3078250, was screened for CTCF-binding sites (335–337). The CTCF TFBSs were retrieved from the MANTA database (338). Precisely, the JASPAR CTCF TF-binding profile (retrieved as a position frequency matrix) was transformed into a position weight matrix (PWM) using the jaspar Bio-Python module and the DNA sequence was scanned on both strands with the PWM; sites harboring a PWM score above a 85% relative threshold were predicted as transcription factor binding sites (TFBSs) (337, 339).

2.2.10 ChIP-qPCR of CTCF binding site

ChIP assays were carried out as previously described (340). Briefly, ChIP assays were performed on cortex and liver tissues from two individuals from the matched dataset or from 4 male mice containing the *HTT* transgene, YAC18 mice, employing a specific mouse monoclonal antibody recognizing CTCF (Abcam Cat# ab70303) and using the EpiQuik Tissue Chromatin Immunoprecipitation Kit (Epigentek), including mouse IgG as a nonspecific control (341). The precipitated DNA was analyzed by quantitative real-time PCR using oligonucleotides recognizing a CTCF binding site located within the *HTT* promoter region: 5′-CTGAGGACCC CAAGTGTGAC-3′ and 5′-CAGGTCGGGACTCATTCCT-3′. ChIP quantitative real-time PCR data were analyzed by the Percent Input Method. Mouse samples were further normalized by IgG to control for differences across tissues.

2.2.11 HEK293 stable cell line generation

The *HTT* promoter-luciferase construct was generated and described in detail in a previous study (306). Briefly, the construct carries a portion of the *HTT* promoter 3.7 kb upstream of the translational start site. The sequence for this *HTT* promoter was isolated from an HD patient and represents haplotype A, the most common on the HD allele (342). To prepare the *HTT* promoter-luciferase construct for use with the FLP-In System (Invitrogen), the construct was digested using the restrictions enzymes SalI and EcoRV-HF (a blunt restriction enzyme) (NEB). SalI and another blunt restriction enzyme, AfeI, were used to digest the pcDNA5/FRT construct from the FLP-In System in order to excise the FLP recombinase docking site (FRT site) and the ATG-less Hygromycin resistance gene. This portion of the pcDNA5/FRT construct was then ligated into the *HTT* promoter-luciferase construct using T4 DNA ligase (NEB) to generate the *HTT* promoter-luciferase FRT construct. HEK293 cells with a single genomically integrated FRT docking site were purchased from Invitrogen, Flp-InTM-293 cell line. Following the FLP-In System protocol, the *HTT* promoter luciferase FRT construct was dual transfected with the pOG44 construct from the FLP-In System, which expresses a modified FLP recombinase gene, into the Flp- InTM-293 cell line. Transfection was conducted using the TransIT® LT1 reagent (Mirus). The clonal selection of successful intergrants was performed using hygromycin antibiotic (Invitrogen) selection. The lack of β-galatosidase expression was used to further confirm successful integration.

2.2.12 siRNA-mediated knockdown

Three siRNA constructs for CTCF, a siRNA against HPRT and a scramble siRNA were obtained from Origene (Catalogue numbers: SR307273, SR30003, SR30004). The TransIT-TKO® transfection re- agent (Mirus) was used to transfect each siRNA into the HEK293 stably expressing cells detailed above. In the case of CTCF, a cocktail of equal amounts all three siRNA variants was used. 48 h after transfection, cells were collected and processed for RT-qPCR as detailed above. Successful transfection of siRNA was assessed using HPRT knockdown and the HRPT siRNA (data not shown).

2.3 Results

2.3.1 Selection of human tissue samples

To assess the association of DNA methylation on HD pathogenesis and its potential role in *HTT* transcriptional regulation, we utilized a cohort of age-matched HD and control individuals from the HD BioBank at the University of British Columbia. Details of this cohort are provided in Table 2.1. The analysis of DNA derived from this cohort generated an initial 450K DNA methylation sample set, which was subsequently parsed into two distinct datasets for downstream analyses (Supplementary Figure 2.1). The first dataset was used to investigate differential DNA methylation patterns in the forebrain cortex between HD cases versus controls in the form of an epigenome-wide association study (EWAS) (herein referred to as the cortex-only dataset; detailed in Table 2.1). The cortex-only dataset consisted of six control and seven HD samples, the latter of which had an average mutant *HTT* CAG repeat length of 46 and an average age of disease onset of ~45 years. The second dataset, composed of matched cortex and liver samples from a subset of individuals in our cohort, was used to identify potential tissue-specific differentially methylated sites within the *HTT* gene locus (herein referred to as the matched dataset; detailed in Table 2.1). The matched dataset comprises matched cortex and liver tissues from five individuals, four of which carried the HD mutant expansion and one healthy individual with a normal *HTT* CAG repeat length. All samples were analyzed on the Illumina HumanMethylation450K Beadchip array and resulting readouts were processed and normalized to yield two distinct datasets comprising ~430 000 probes genome-wide.

Sample ID	CAG size (HD first if applicable)	Sex	Age	Age of onset	Age of onset relative to CAG size[a]	Symptomatic HD years	PMI (hr)	Used in 450K methylation array	Used in *HTT* expression RT-PCR	Used in pyro-sequencing
Individuals utilized on Illumina 450K array										
COB 05	19/20	M	75	NA	NA	NA	10.8	Y		
COB 20/30	17/20	M	74	NA	NA	NA	6.25	Y		
COB 22/25	18/23	F	77	NA	NA	NA	12	Y		
COB 51	17/23	M	54	NA	NA	NA	12.5	Y		
COB 59	17/19	M	21	NA	NA	NA	8.5	Y		
COB 125	15/17	M	74	NA	NA	NA	2	Y		
HDB 119	41/17	M	74	62	Late	12	3	Y	Y	Y
HDB 159	42/21	M	69	54	Mean	15	7	Y		
HDB 162	42/15	F	69	46	Early	23	7	Y		
HDB 165	42/31	F	71	37	Early	34	15	Y		
HDB 166	43/17	F	72	61	Late	11	8	Y	Y	Y
HDB 167	50/23	M	52	35	Late	17	15	Y	Y	Y
HDB 176	62/19	M	29	23	Mean	6	3.5	Y	Y	Y
Mean	Control = 6 HD = 7	M = 9 F = 4	**62.4**	**45.4**		**16.9**	**8.5**			
Validation individuals										
HDB 156	51	M	26	Pre-symptomatic	Pre-symptomatic	Pre-symptomatic	48			Y
HDB 175	53	M	39	24	Mean	15	3.5			Y
HDB 178	44	M	68	47	Mean	21	14			Y
Mean	49.3	M = 3	**44.3**	**35.5**		**18.0**	**21.8**			

Table 2.1 Human cohort sample characteristics

[a]Calculated as described in (306). In short, an age of onset ratios [observed age of onset/expected age of onset based on CAG size as described in (303)] were calculated for each individual within the UBC HD biobank. All of the individuals were then categorized into percentiles based on their age of onset ratios; mean (40-60th percentile), early (15-40th percentile), late (60-85th percentile).

2.3.2 Normalization of neuronal cell-type proportion differences between HD and control cortex methylation profiles

To investigate HD-associated changes in the DNA methylation profiles, we focused exclusively on the cortex samples. Given that cell-type heterogeneity is the second largest contributor to DNA methylation variance after tissue type, we corrected for the inter-individual differences in neuronal versus non-neuronal cell proportions in the cortex methylation profiles (59). While this approach is generally recommended when assessing DNA methylation in brain samples, it might be particularly relevant for HD epigenetic research since neuronal degradation and gliosis are well-described neuropathological features of the disease that increase in severity and spread to brain structures outside of the striatum as the disease progresses (343). Given that the average length of time between the age of onset and the age of death of the HD individuals in our cohort is 16.9 years, and that HD progression from initial diagnosis of symptoms to patient death is ~20 years, it is likely that neuronal degeneration and gliosis are present in the majority of the cortex samples of our HD individuals (298). To ascertain such proportional differences in neuronal versus glial cell types in the cortex samples, we implemented the cell epigenotype-specific (CETS) algorithm, which estimates the ratio of neuronal versus glial cell-type proportions based on underlying DNA methylation profiles of reference probes (69). As suspected, we observed a trend towards a decrease in neuronal proportions (and by converse, an increase in glial proportions) in the HD cortex samples compared with controls, although this did not reach statistical significance (Figure 2.1A). To determine whether cell-type heterogeneity influenced DNA methylation patterns in the cortex samples, principal component analysis (PCA) was performed on our cortex-only 450K dataset (Figure 2.1B). PC1 was significantly correlated to neuronal cell proportion ($P < 0.001$), and accounted for a substantial 35.6% of variance in the cortex-only DNA methylation dataset. In order to normalize neuronal proportion differences in the cortex DNA methylation profiles, a published strategy was used to regress these cell-type differences out of the DNA methylation data (Figure 2.1C) (329). Confirming the efficiency of the CETS-based adjustment, PCA following cell-type correction indicated that neuronal proportions correlated only to PC12 and accounted for a negligible 0.00043% of the variance in the corrected cortex-only dataset (Figure 2.1D). This adjusted cortex-only dataset thus represented cortex DNA methylation profiles whose inter-individual cell-type differences have

been normalized to the best of our abilities, allowing the use of this corrected dataset for all further analyses.

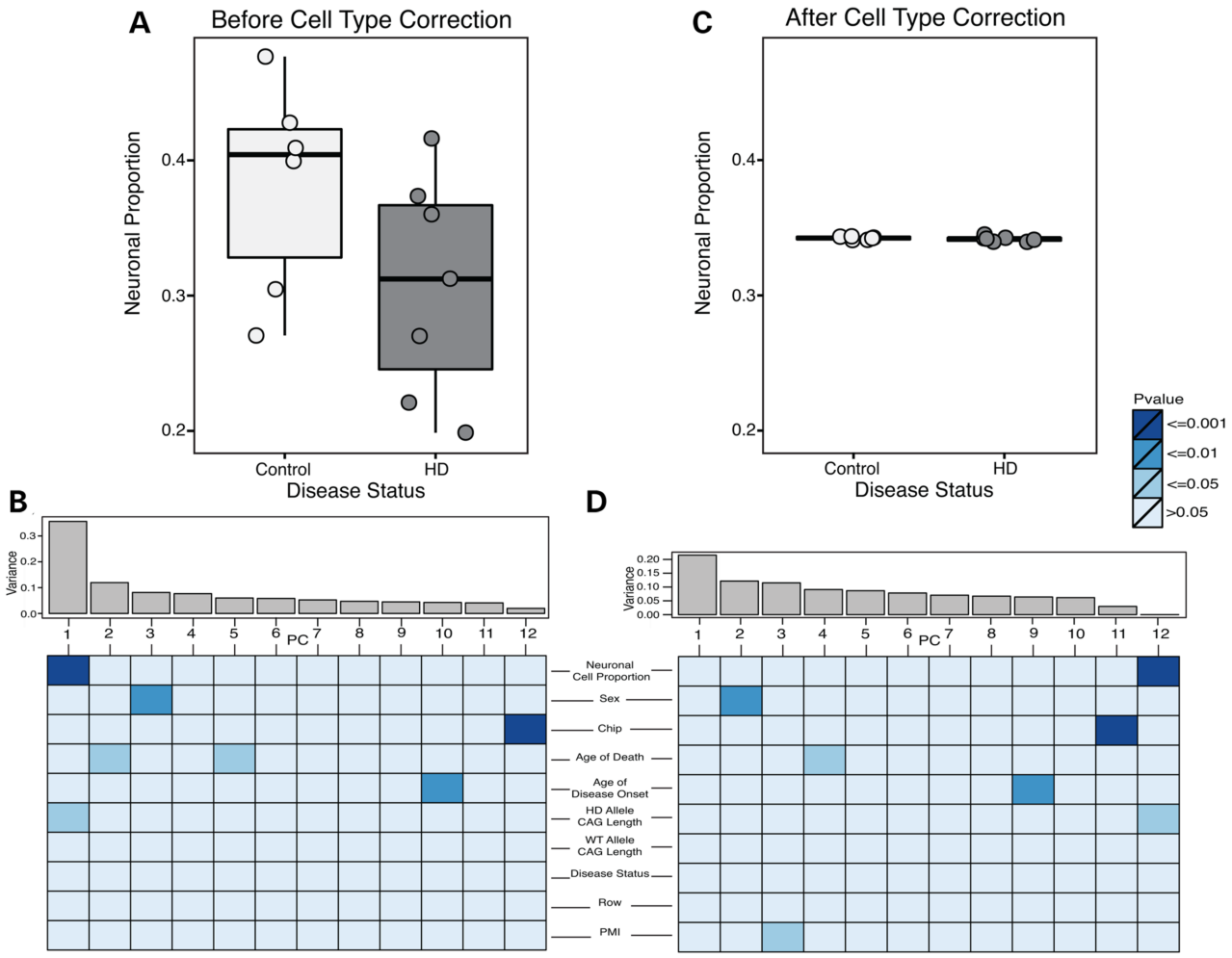


Figure 2.1 Correction for neuronal proportion differences between HD and control cortex methylation profiles.

(A) Estimations of neuronal cell proportions based on underlying reference methylation profiles of the uncorrected cortex-only 450K dataset using the CETS algorithm. A non-significant trend towards decreased neuronal cell proportions in HD cases compared with controls is observed (Mann–Whitney U test, ns). (B) PCA of the uncorrected cortex-only 450K dataset shows the correlation of known phenotypic and technical variables to PCs (bottom heatmap), each representing an incremental proportion of the variance in the methylation data (top scree plot). Neuronal cell proportions significantly correlate with PC1 representing 35.6% of the methylation variance in the uncorrected cortex-only dataset. (C) Estimations of neuronal cell proportions in the corrected cortex-only methylation dataset show that differences in brain cell composition between HD cases and controls have been normalized (Mann–Whitney U test, ns). (D) PCA of the corrected cortex-only 450K dataset shows the correlation of known phenotypic and technical variables to PCs (bottom heatmap), each representing an incremental proportion of the variance in the methylation data (top scree plot). Neuronal cell proportions significantly correlated with PC12 representing a negligible 0.00043% of the methylation variance in the corrected cortex-only dataset.

2.3.3 Assessment of HD-associated differential DNAm in cortex tissue

Using the cell-type corrected cortex methylation profiles, we next assessed HD-associated genome-wide DNA methylation changes in the cortex-only dataset. Specifically, we implemented a local regression smoothing method called 'Bumphunting' to detect differentially methylated regions (DMRs) between HD and control cortex samples (330). This method did not identify any significant HD-associated DMRs in the cortex samples. Consistent with these results, linear modeling of all 427 242 probes in the cortex-only dataset did not detect any significant differentially methylated sites after multiple test correction (FDR < 0.05). This was supported by the uniform P-value distribution of the linear regression analyses (Supplementary Figure 2.2A). Moreover, unsupervised hierarchical clustering of global DNA methylation profiles failed to differentiate groupings between the HD and cortex samples (Supplementary Figure 2.2B). This was further corroborated by the lack of correlation between disease status and any of the PCs underlying the DNA methylation patterns in our cortex-only dataset (Figure 2.1D).

In addition to assessing the relationship between HD and global DNA methylation patterns in the cortex, we also explored the association of DNA methylation variation on the age of disease onset. Interestingly, when we examined the cortex DNA methylation profiles of the HD individuals only, the age of disease onset correlated with PC4, representing 14.5% of the DNA methylation variation in HD cortex samples, although the correlation was only nominally significant (Spearman's rank correlation, P = 0.048) (Supplementary Figure 2.3A). Moreover, we noted a leftward skewed P-value distribution for the effect of age of onset on DNA methylation in the HD cortex samples, signifying an enrichment of low P-values beyond the expectation by chance (Supplementary Figure 2.3B). Perhaps not surprisingly given the low sample numbers, probe-wise linear modeling on pre-filtered variable probes failed to identify differentially methylated sites associated with the age of onset (adjusted for sex) after multiple test correction (FDR < 0.05). This suggested that although DNA methylation in HD cortex samples was likely associated with the age of disease onset, we were underpowered to identify site-specific changes due to our low sample size of HD brain tissues. Related to the age of onset variation, we also considered the impact of DNA methylation age, a biological age prediction derived from a subset of DNA methylation sites, in our cortex samples, as this age estimate has been recently associated with age-related diseases and mortality (322, 344, 345). We used a well-described

DNA methylation age predictor on our cortex methylation profiles and found that there were no significant differences in DNA methylation age between the HD and control samples (Supplementary Figure 2.3C) (322).

Finally, we specifically queried the *HTT* gene region for HD-associated differential methylation. Given that many transcriptional regulatory regions, such as enhancers, are known to reside outside of the proximal promoter where DNA methylation levels tend to be more dynamic and subject to variation, we expanded our search beyond the proximal promoter of *HTT* in order to assess potential epigenetic variation in the local chromosomal environment of the *HTT* gene (52). We focused on the genomic region encompassing the first exon of the gene preceding *HTT* (*GRK4*) through to the gene following *HTT* (*MANSTD1*) as being representative of the local genomic context of the *HTT* gene locus, chr4: 2973107–3258169. This region comprises 87 sites covered by 450K probes in the cortex-only dataset. When tested individually, none of the 87 probed sites in this candidate genomic region were differentially methylated between HD and control cortex samples. Overall, our analyses using the cortex DNA methylation profiles showed evidence of an association between DNA methylation and age of disease onset but did not identify any detectable HD-associated DNA methylation changes at probed sites.

2.3.4 Disease-associated and inter-individual transcriptional variation in *HTT*

Although beyond the association with the age of onset, no HD-associated DNA methylation alterations in our cortex samples were found at the *HTT* locus, we investigated whether there is a difference in *HTT* expression levels between HD cases and controls in human cortex tissue. Using our cortex-only sample set, we found that there was no significant difference in *HTT* mRNA levels between control and HD individuals (Figure 2.2A). We also considered the possibility that the lack of *HTT* expression differences between HD cases and controls may be due to a lack of inter-individual variability in *HTT* mRNA expression in our cohort, contrary to a previous report of high *HTT* expression variation in whole-brain samples of inbred mice (309). Using individuals from our cortex-only dataset, we implemented a previously described testing method for measuring inter-individual expression variation whereby the variability contributed by experimental variation, referred to as experimental coefficient of variation (CV_{exp}), could be subtracted from the total variation between individuals, referred to as total coefficient of variation (CV_{total}) (309). We found a CV_{total} of 26.4% and a CV_{exp} of 7.5% for *HTT* mRNA

expression in the human cortex samples. This indicated that of the observed *HTT* mRNA expression variation, 71.5% ((26.4–7.5)/(26.4)) of the variation is likely attributed to inter-individual expression variation. This suggested that considerable *HTT* mRNA expression variation existed across individuals in our cortex-only sample set and thus, likely did not underlie the lack of HD-associated transcriptional variation observed.

Interestingly, the CV_{total} we observed in our human cortex-only cohort (CV_{total} of 26.4%) was lower than the 30% previously observed in whole-brain samples of male inbred mice (309). This was a surprising result given the inherent increase in genetic and environmental heterogeneity in a human cohort compared with inbred, environmentally controlled, laboratory mice. Upon reflection of the previous reported *Hdh* results, we noted the use of a single reference gene, *18S*, when compared with our generation of a normalization factor based on several reference genes (309, 346). We revisited their analysis to compare both normalization methods on 8 wild-type FVB male mice using whole-brain samples and found that normalization with *18S* alone yielded CV_{total} results comparable with the published results while the use of a normalization factor yielded a substantially lower CV_{total} measure (27.8% when normalized with 18S and 6.3% when a normalization factor is used) (309). Moreover, normalizing *18S* mRNA expression itself with a normalization factor generated a CV_{total} of 18.2%, suggesting that the use of *18S* alone is a poor choice for *HTT* RT-qPCR normalization. This difference in normalization approach likely explains the discrepancy between our results and the previously published results in inbred mice (309).

Finally, as these samples comprise post-mortem human tissues, we also considered other potential factors contributing to *HTT* expression variability such as the impact of post-mortem interval (PMI) and neuronal proportion differences on our *HTT* expression readouts. We found no significant association between these measures and *HTT* mRNA levels, respectively (Figure 2.2B-D). Taken together, our analyses in cortex tissues found minimal evidence of HD-related DNA methylation differences at probed sites other than a general association to age of disease onset and no *HTT* mRNA differences between HD cases and controls.

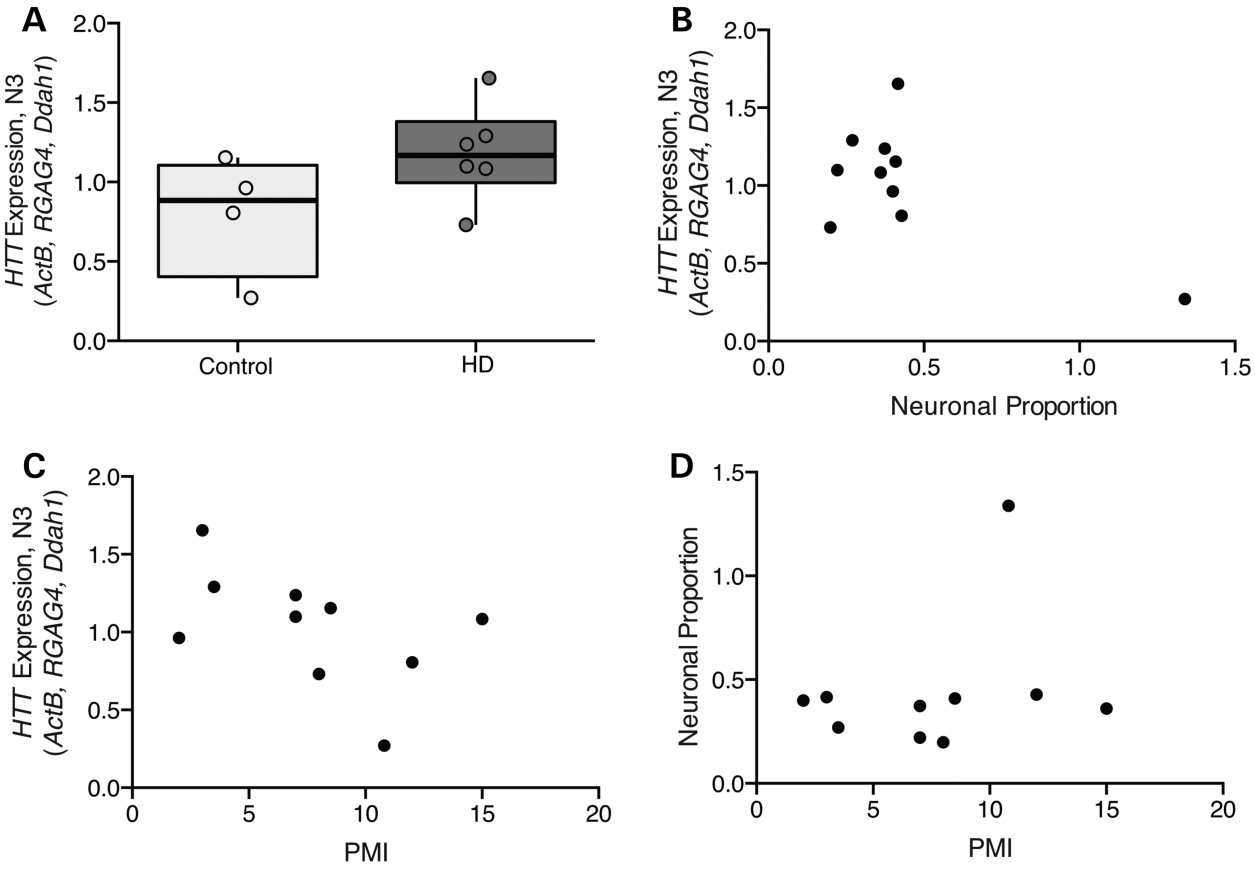


Figure 2.2 Comparison of *HTT* expression between HD and control individuals and associations between *HTT* expression, PMI and neuronal proportion.

(A) RT-qPCR analysis of *HTT* transcript levels in individuals in the cortex only dataset, normalized by a normalization factor generated by 3 (NF3) normalization genes [indicated on y axis (Wilcoxon paired signed-rank test, ns)]. (B) Association of *HTT* expression to predicted neuronal proportions of cortex samples showed a lack of correlation (Spearman correlation, ns) (C) Association of *HTT* expression to PMI showed a lack of correlation (Spearman correlation, ns) (D) Association of predicted neuronal proportions in cortex samples to PMI showed a lack of correlation (Spearman correlation, ns).

2.3.5 Identification of tissue-specific transcriptional differences and tissue-specific differential DNAm at *HTT* locus

We next examined tissue-specific differences at the *HTT* gene locus using paired cortex and liver samples in our matched dataset. To confirm that the samples in our matched dataset exhibit tissue-specific *HTT* expression patterns, we assessed *HTT* mRNA levels in the matched tissues and found that the average expression of *HTT* was higher in cortex versus liver, as consistent with previous reports (paired t-test, $P < 0.05$) (Figure 2.3A).

To investigate tissue-specific DNA methylation differences at the *HTT* locus using matched cortex and liver samples in our matched dataset, paired testing by linear regression

analysis was performed for the 87 probes underlying the candidate *HTT* genomic region (chr4: 2973107–3258169). We identified 38 differentially methylated probes (at FDR < 0.05) with DNA methylation differences ranging from 10 to 40% between cortex and liver tissues (Figure 2.3B and 2.4A). Consistent with these results, unsupervised hierarchical clustering grouped samples based on tissue type for this region, indicating that this region was differentially methylated between cortex and liver tissues (Figure 2.3C). While a few of the identified differentially methylated probes were found in the *HTT* promoter region, the majority of the identified probes occurred within the gene body of *HTT*, primarily towards the 3′ end of the gene (Figure 2.3C). Of note, our matched tissue dataset comprises both HD cases and a healthy control. Given that we were unable to detect DNA methylation differences between HD cases versus controls in our cortex only dataset, we hypothesized that any variation identified in the matched dataset would be largely due to tissue-specific variation. This was supported by our observation that the tissue type was significantly correlated to PC1, accounting for 85.9% of the DNA methylation variance in the matched dataset (Figure 2.4B).

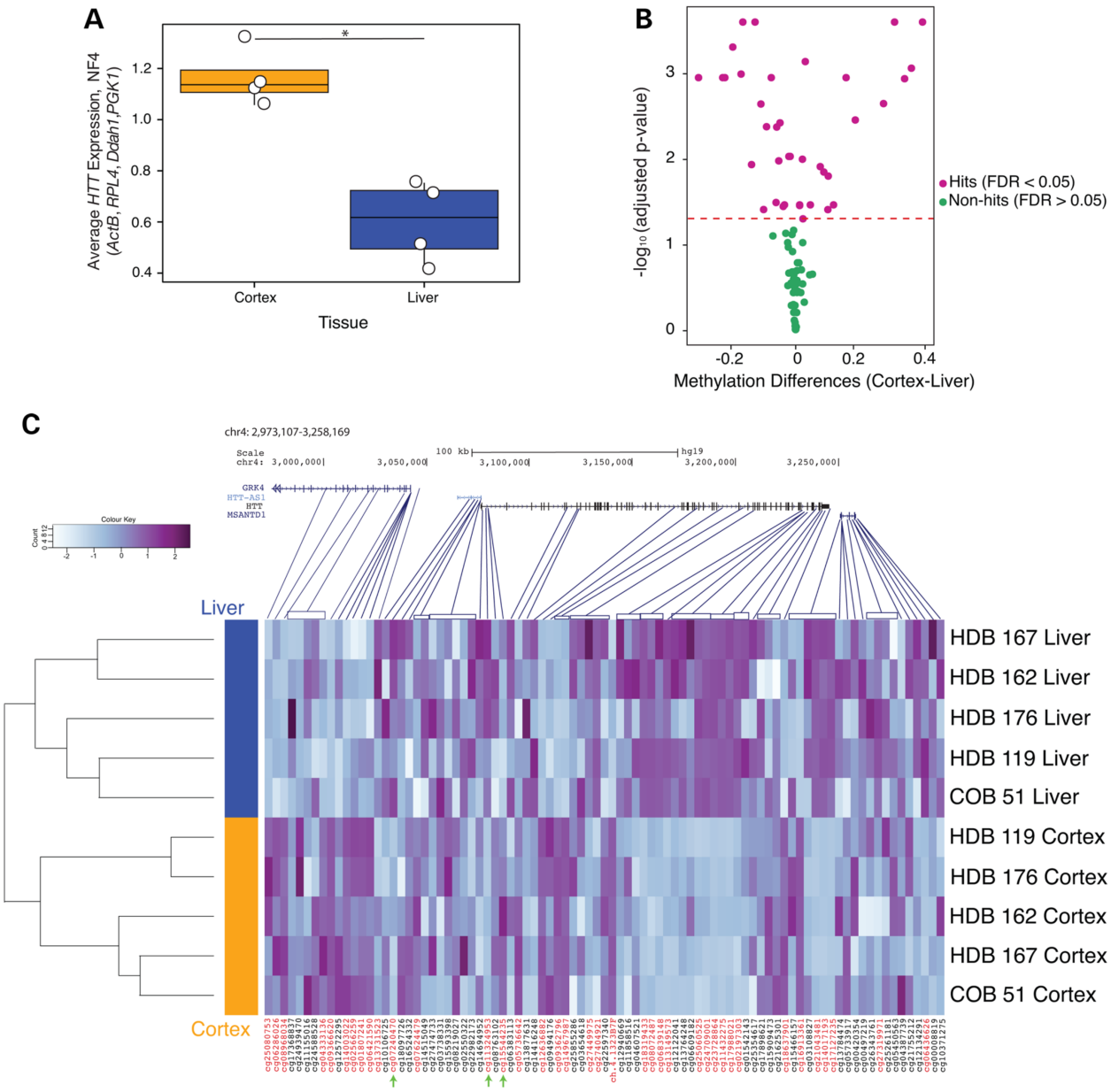


Figure 2.3 Identification of tissue-specific *HTT* expression differences and differentially methylated probes underlying the *HTT* gene region in cortex and liver tissues.

(A) RT-qPCR analysis of *HTT* transcript levels in human cortex and liver tissues from individuals (n = 4) utilized in the matched 450K dataset (averaged from triplicate samples per individual), normalized by a normalization factor generated by 4 (NF4) normalization genes (indicated on y axis) (Paired t-test, *P < 0.05). (B) Volcano plot depicting differences in methylation levels (cortex–liver) for each probe in the *HTT* gene region (indicated on x axis) against adjusted P-value (indicated on y axis, on –log10 scale). Dashed red line denotes FDR threshold of 0.05. Magenta points represent statistically significant hits (FDR < 0.05) while green points represent non-statistically significant sites (FDR > 0.05). (C) Heatmap representing the 87 probes represented on the 450K methylation array in *HTT* gene region (chr4: 2973107–3258169). Gene features of the genomic region are represented in the upper portion of the figure. A total of 38 of the probes (shown in red) were found to be differentially methylated between the matched cortex (yellow) and liver (blue) samples (FDR < 0.05). Unsupervised hierarchical clustering successfully segregated samples based on the tissue type. Probes selected for pyrosequencing verification indicated by green arrows.

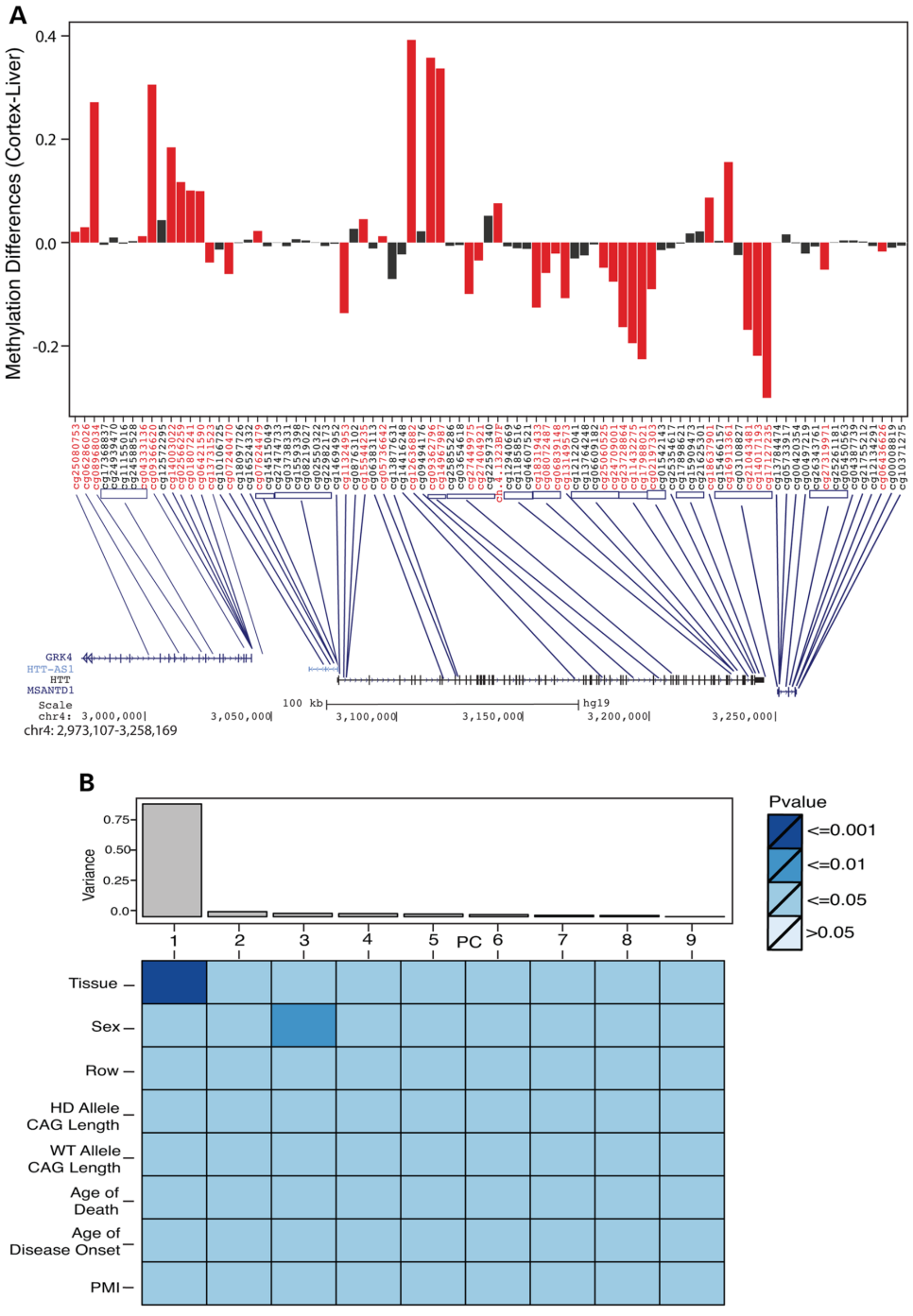


Figure 2.4 DNA methylation differences (cortex–liver) of probes in *HTT* gene region and PCA of matched 450K dataset.

(A) Differences in methylation levels (cortex–liver) for each probe in the *HTT* gene region (chr4: 2973107–3258169). Gene features of the genomic region are represented in the lower portion of the figure. Probes that are significantly differentially methylated between cortex and liver are shown in red. (B) PCA of the matched 450K dataset after brain-cell-type correction shows the correlation of known phenotypic and technical variables to PCs (bottom heatmap), each representing an incremental proportion of the variance in the methylation data (top scree plot). Tissue type (cortex versus liver) significantly correlate with PC1 representing 85.9% of the methylation variance.

2.3.6 Pyrosequencing validation of tissue-specific DNAm

To confirm the tissue-specific differential DNA methylation loci from the matched dataset using a different methodology, we performed pyrosequencing in the original cortex and liver tissue samples at three sites (cg07240470, cg11324953 and cg15544235; these probes are indicated by green arrows in Figure 2.3C). Pyrosequencing data were highly correlated to the data from the 450K array at all three of the selected CpG sites, further substantiated by Bland–Altman plots that showed an unbiased agreement (Supplementary Figure 2.4). Consistent with adjacent CpG sites often showing related DNA methylation patterns, CpGs in close proximity to our target validation CpG were significantly positively correlated (Spearman correlation, $P < 0.0001$) (Supplementary Figure 2.5) (347). We calculated the mean DNA methylation values across the correlated CpGs in each pyrosequencing assay and found these to be significantly different between cortex and liver tissues of our original individuals as well as in three additional HD individuals used for validation (Wilcoxon paired signed-rank test, $P < 0.05$) (Figure 2.5). Taken together, pyrosequencing thus verified the original tissue-specific differential DNA methylation hits from our matched dataset and extended it to neighboring CpGs not present on the 450K array.

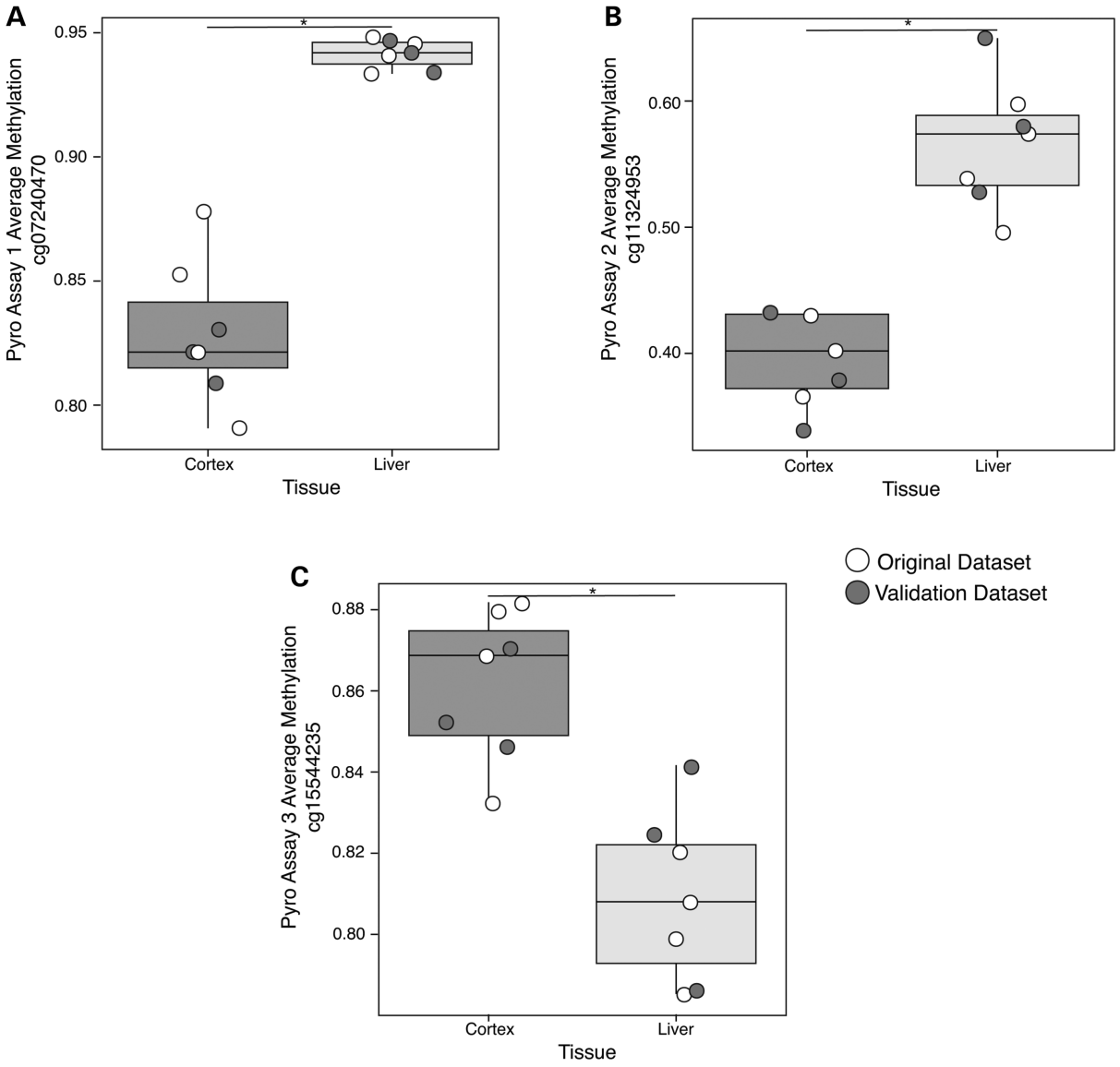


Figure 2.5 Confirmation of tissue-specific DNA methylation hits by pyrosequencing.

Boxplots showing difference in averaged DNA methylation levels in correlated CpGs containing the following Illumina 450K array probed sites (A) cg07240470 (B) cg11324953 and (C) cg15544235. Original samples used in 450K matched dataset are colored in white and validation samples are shaded in grey (Wilcoxon paired signed-rank test, *adjusted P-value < 0.05).

2.3.7 Identification of a differentially methylated CTCF binding site in *HTT* promoter

Following the identification of tissue-specific *HTT* DNA methylation patterns, we sought to identify potential transcriptional regulatory elements that could functionally link DNA methylation variation at the *HTT* gene locus to tissue expression variation. Given that CTCF is a transcription factor whose function is known to be partly regulated by cell-type-specific differential DNA methylation, we screened the proximal *HTT* promoter using a candidate

approach for putative CTCF transcription factor binding sites (TFBSs) (348, 349). We utilized available CTCF ChIP-sequencing (ChIP-seq) datasets collected from both ENCODE and PAZAR to identify candidate CTCF-binding locations and then screened these candidate binding regions for putative CTCF TFBSs through the use of the CTCF TF-binding profile from JASPAR (335–337, 349). By combining the available ChIP-seq data and the TF binding profile, we were able to identify putative CTCF TFBSs based on both identified interaction of CTCF with the genomic location and motif similarity. Importantly, we identified a putative CTCF-binding site ~1.2 kb from the *HTT* translational start site (Figure 2.6A). We noted that this CTCF TFBS is conserved in Rhesus monkeys and partially conserved in mouse, underscoring the potential functional importance of this site (Supplementary Figure 2.6A). As the specific CpG site within this CTCF-binding site and two other nearby CpGs were not covered in the Illumina 450K array, we used bisulfite pyrosequencing to assess the DNA methylation levels at these sites (Figure 2.6A). We found that these CpG sites showed higher DNA methylation levels in the liver (~30–60%) over cortex (~10–25%) both in the original individuals in our matched dataset and in three additional individuals (Figure 2.6B-D) (Wilcoxon paired signed-rank test, $P < 0.05$). In addition, unsupervised hierarchical clustering separated the samples by tissue type at these three specific CpG sites (Supplementary Figure 2.6B-C). Of note, we observed that the DNA methylation levels at these CpG sites were not significantly different between HD cases and control in cortex samples, signifying minimal evidence of possible HD-related effects at these sites (Supplementary Figure 2.6D) (Mann–Whitney U test, ns). We also noted a lack of association between the age of onset and CpG methylation underlying the CTCF TFBS in HD cortex tissue (Spearman correlation, ns). These results indicated that a CTCF-binding site not previously identified at the *HTT* proximal promoter was differentially methylated between cortex and liver tissues in manner that is likely not dependent on HD status.

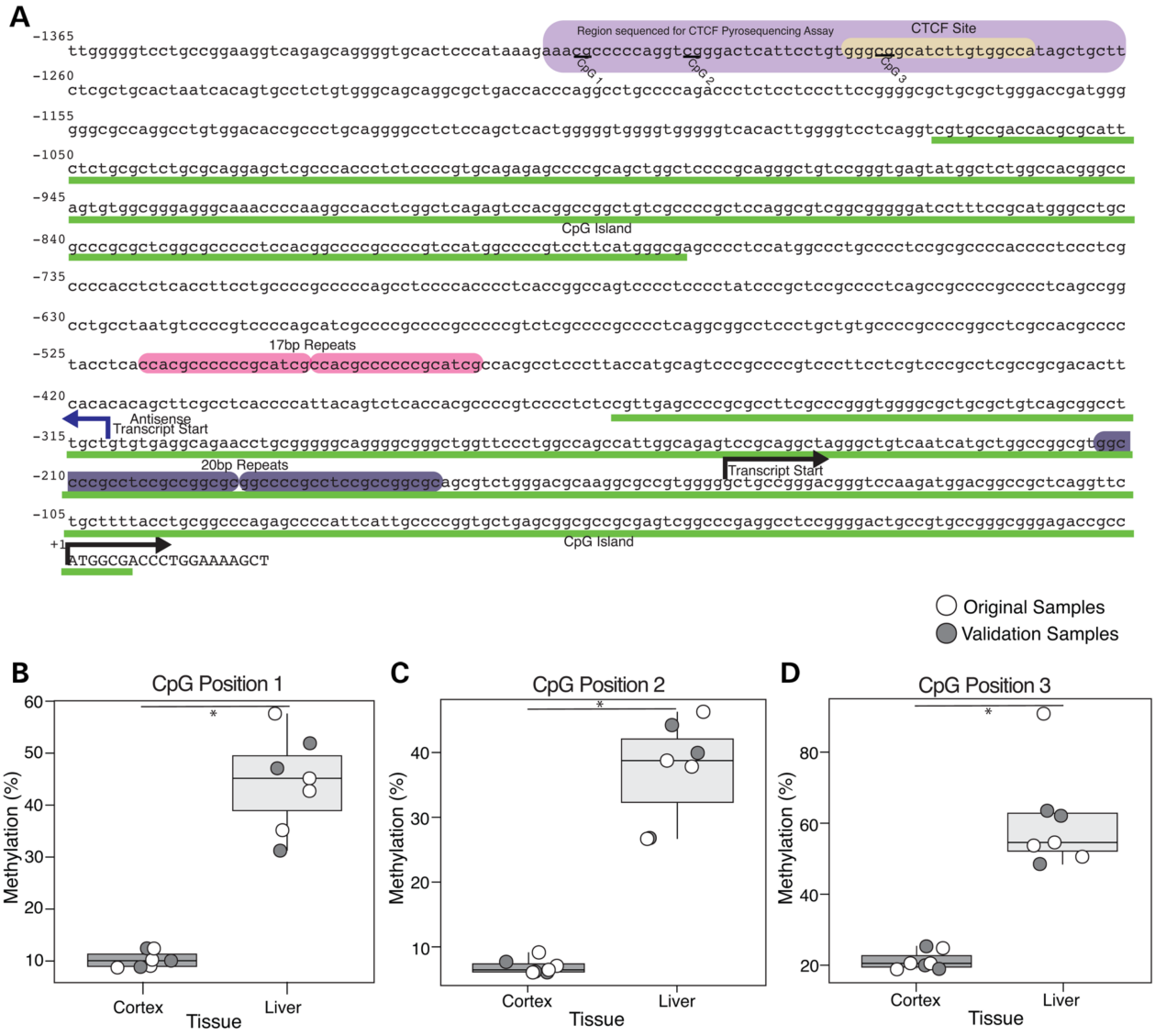


Figure 2.6 Identification of a differentially methylated CTCF-binding site within the *HTT* promoter.

(A) Schematic of *HTT* proximal promoter, orientated by the transcriptional start site. Underlined in green are two CpG islands in the region. Also shown are two repeat sections in the region, a 20 and 17 bp repeat, the transcript start site and an antisense transcript start site. The identified CTCF site is shown within the sequence used for pyrosequencing, the three CpG sites within the pyrosequencing assay are underlined. (B–D) The average methylation for each pyrosequenced CpG, indicated in (A), for the original individuals used in the 450K analysis (colored in white) and validation individuals (shaded in grey) (Wilcoxon paired signed-rank test, *P < 0.05).

2.3.8 Identification of CTCF TFBS occupancy *in vivo* and assessment of functional impact of CTCF on *HTT* promoter function *in vitro*

To assess the occupancy of the CTCF protein at this CTCF TFBS between liver and cortex tissues, we conducted a chromatin immunoprecipitation qPCR (ChIP-qPCR) assay in cortex and liver tissues from two representative HD individuals in the matched dataset. Using

primers designed to specifically amplify the CTCF-binding site, we found a percent enrichment of the identified CTCF-binding site of 16.4% (±11.6%) in cortex tissue and 4.2% (±2.2%) in the liver, suggesting a higher occupancy of CTCF in cortex where DNA methylation of the site is decreased compared with liver (Figure 2.7B). We similarly performed ChIP-qPCR in cortex and liver tissues of YAC18 mice, which express a full-length human *HTT* transgene, including the *HTT* proximal promoter, on a yeast artificial chromosome (341). As with the human samples, we found a higher percent enrichment of the identified CTCF-binding site in mouse cortex (51.8% ± 32.8%) when compared with liver (3.2% ± 0.7%) (Figure 2.7A). Together, these results suggest that the CTCF site within the *HTT* promoter is more highly occupied in cortex over liver tissue.

Next, to determine whether CTCF could functionally regulate *HTT* promoter activity, we assessed the effect of siRNA-mediated knockdown of CTCF in human embryonic kidney cells (HEK293 cells) using a previously described *HTT* promoter-luciferase construct that contains the identified CTCF site (306). HEK293 cells were modified to carry a single, genomic integration of the *HTT* promoter-luciferase construct, thereby allowing for the assessment of *HTT* promoter function in a fully integrated, genomic context. Treatment with CTCF-targeting siRNAs resulted in a significant knockdown of CTCF mRNA compared with both untreated and scramble siRNA-treated cells (Figure 2.7C) (one-way ANOVA, $P < 0.0001$). Silencing of CTCF resulted in a decreased function of the *HTT* promoter-luciferase construct as assessed by decreased luciferase reporter transcript expression compared with both untreated and scramble siRNA treated cells (Figure 2.7D) (one-way ANOVA, $P < 0.0001$). Combined, our data suggest that CTCF may, in part, regulate *HTT* promoter function and may contribute to tissue-specific *HTT* expression in manner that is dependent on differential DNA methylation.

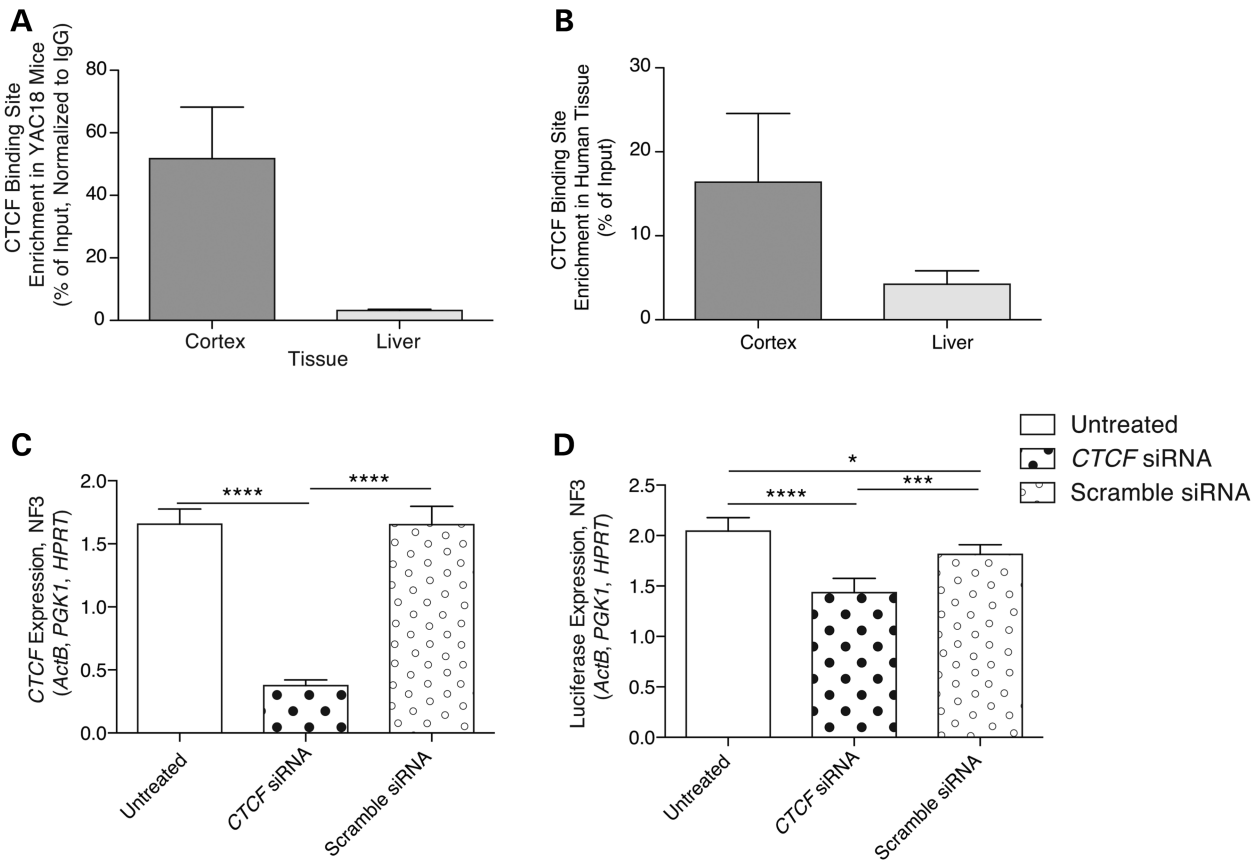


Figure 2.7 CTCF binding site is differentially enriched in cortex versus liver in human and mouse tissues and knockdown of CTCF decreases *HTT* promoter function.

(A and B) ChIP-qPCR of CTCF occupancy of the identified CTCF-binding site of 4 YAC18 mice (A) or two individuals from the matched dataset (B) in cortex and liver tissues, calculated using percent of input genomic DNA. The percent enrichment mouse: cortex—51.8% ± 32.9%; liver—3.2% ± 0.7%. Percent enrichment human: cortex—16.4% ± 11.8%; liver—4.2% ± 2.2%. (C and D) RT-qPCR of CTCF expression (C) and Luciferase reporter expression (D) in HEK 293 cells stably expressing a *HTT* promoter and transfected for three siRNAs targeting CTCF. Data are normalized to a normalization factor of 3 (NF3), normalization genes (indicated on the y-axis). For each treatment n = 3. (One-way ANOVA with Tukey post-hoc test, *P < 0.05, ***P < 0.001, ****P < 0.0001.)

2.4 Discussion

In this study, we utilized genome-wide DNA methylation profiles of human cortex tissue, with a subset of matched liver tissues, from a cohort of HD and control individuals in order to identify potential HD-related DNA methylation aberration in the brain as well as tissue-specific DNA methylation variation at the *HTT* gene locus. We were unable to identify any HD-associated DNA methylation alterations at probed sites after correction for neuronal proportion differences of the cortex tissues, but did find evidence of a general association between the age of disease onset and DNA methylation in HD cortex samples. When comparing DNA

methylation between matched cortex and liver tissues, we discovered site-specific differential DNA methylation spanning the promoter and intragenic regions of the *HTT* gene, including a new differentially methylated CTCF-binding site in the *HTT* proximal promoter. This CTCF site displayed increased occupancy in cortex tissue, where *HTT* expression is higher, when compared with the liver and silencing of CTCF resulted in reduced function of an *HTT* promoter–reporter construct. Together, our data suggest that DNA methylation, as ascertained by the technology platform utilized here, may have minimal association with the HD status but may be associated with the age of disease onset in cortex tissue. In addition, DNA methylation may, in part, contribute to tissue-specific HTT gene expression patterns through differential CTCF occupancy.

Our EWAS analyses in cortex tissues showed a lack of association between HD status and DNA methylation variation which is seemingly in contradiction to two recent studies that have assessed DNA methylation changes in the context of the HD mutation (231, 318). While both of these studies have laid the groundwork upon which further investigation into DNA methylation changes in HD can be conducted, it is possible that alternative interpretations to the ones being put forward in the publications may, in part, explain these results. The first study utilized both human HD striatum samples and striatal samples from the R6/2 mouse model of HD to assess the effect of the HD mutation on the DNA methylation of a single gene, *ADORA2A* (231). Given that the average Vonsattel stage for the human HD samples used in the study was ~3 (maximum Vonsattel stage rating is 4) and that Vonsattel staging is based, in part, on striatal atrophy, it is likely that the neuronal versus glial proportions in their HD samples differed significantly from their control striatum (350). In our study, we found a trend towards a decrease in neuronal cell-type proportion in HD versus control cortex, a brain structure affected in the later stages of disease pathogenesis, unlike the striatum that displays the primary neuronal degeneration seen at the earliest stages of HD (319, 320). Consequently, it is possible that the reported differential methylation of the *ADORA2A* gene may be largely driven by differences in brain cell-type proportions and not disease status. Indeed, our analyses showed that any HD-related effects on CpG methylation at *ADORA2A* were eliminated after correction for brain-cell-type heterogeneity. The second recent HD DNA methylation study utilized a genome-wide approach to assess DNA methylation changes in an immortalized mouse striatal cell line, expressing either the mutant or wild-type human huntingtin protein (318). While mouse striatal cell lines do recapitulate some aspects HD pathogenesis, it is unclear to what extent DNA

methylation changes identified in mouse cell line models can be extrapolated to the human genome. Furthermore, *in vitro* cell culturing has been shown to alter DNA methylation patterns, further confounding interpretation of these results from immortalized cell lines (286, 351). It is worth noting that an another potential source of discrepancy between our findings in cortex tissue and these earlier studies in striatal samples is that DNA methylation alterations may factor into HD-related changes in striatum more so than in cortex (231, 318). To address this, a future study using striatum of HD patients at early stages of pathogenesis, before significant neuronal loss occurs within the striatum, would be beneficial. Finally, previous findings reported a null association between the age of disease onset and DNA methylation at the *HTT* locus in human HD samples while our results did show evidence of a general association between the age of disease onset and global DNA methylation variation in HD cortex tissue (233). Although this earlier study also utilized human samples, their analyses were limited to the *HTT* gene only and were performed in peripheral blood, while our analyses assessed DNA methylation variation across ~430 000 sites genome-wide using disease-relevant human brain cortex tissue.

It is worth noting that our EWAS analyses in cortex tissue had a few inherent limitations. Firstly, due to the limited availability of human cortex tissue, it is likely that our study is underpowered to detect genome-wide HD-associated DNA methylation differences, even of moderate to high effect size, as suggested by recent work detailing power calculations and sample size considerations for EWAS analyses (352). In our EWAS results, we noted that the top hit had an adjusted P-value of ~0.9, signifying that even at a more relaxed FDR threshold, it is unlikely that we would be able to detect any HD-related DNA methylation differences at queried sites in our dataset. Assuming that HD-related DNA methylation changes are small, occurring in only 0.1% of methylation sites in the cortex genome with subtle effect sizes of ~1%, we calculated that at least 70 samples per group would likely be required in order to detect such site-specific changes at an FDR of 5% and power of 80%. Moreover, as the 450K array covers only 1.7% of total CpGs in the genome with limited representation of enhancers and intergenic regions, it is also possible that HD-relevant differential DNA methylation exists at sites beyond those interrogated in our analyses (129). Our second study limitation arose from the fact that we were unable to directly assess DNA methylation levels within the CAG tract of mutant *HTT* loci. Although DNA methylation most commonly occurs in the context of CpG dinucleotides, non-CpG methylation occurs at CpHpG sites (in which H = A, C or T) and exists at appreciable

levels in the human brain, particularly within neuronal genomes (92, 98, 99). Given that the HD mutation is a result of a CAG repeat expansion, it is plausible that the expanded HD CAG tract in human cortex tissue may be affected by aberrant non-CpG methylation. Since none of the probes present on the 450K array targeted the *HTT* CAG tract and pyrosequencing of the *HTT* CAG tract would likely produce irregular results due to the large number of repeats, we were unable to ascertain whether differential DNA methylation of the *HTT* CAG tract occurs in HD cases versus control cortex tissue. In addition, as most HD individuals are heterozygous for the HD mutation, any future work aimed at directly assessing *HTT* CAG tract DNA methylation will necessitate phasing the results to either the mutant or wild-type allele. Finally, a third technical limitation in our study is that bisulfite conversion of DNA for 450K and pyrosequencing analyses generates composite readouts of both the canonical 5-methylcytosine (5mC) mark of DNA methylation as well as its oxidized derivative, 5-hydroxymethylcytosine (5hmC) (32). As such, we were unable to distinguish between the relative contribution of 5mC or 5hmC variation to HD pathogenesis. Given that 5hmC tends to be most abundant in brain than in other adult tissues, it is possible that we were unable to resolve true 5mC differences between HD versus control cortex samples (109, 117, 353). Overall, future adequately powered studies that can directly assess non-canonical forms of DNA methylation, including non-CpG methylation at the *HTT* CAG tract and genome-wide 5hmC measures, may help clarify the possible association of these cytosine variants with epigenomic variation in HD brain tissues. Despite these technical limitations, our EWAS analyses in cortex tissue represents the first comprehensive assessment of DNA methylation variation in human HD brain tissues and, accordingly, provides a founding framework for investigating additional sources of epigenetic variation in HD.

Our examination of *HTT* transcriptional differences in human cortex tissues revealed no *HTT* mRNA alterations between HD cases and controls, yet notable *HTT* mRNA variation across individuals. The latter result is consistent with previous findings showing considerable levels of inter-individual expression variation in *Hdh* mRNA from genetically identical mice; however, the total amount of variation, CV_{total}, was lower than previously reported in mice (309). We noted a difference in normalization method utilized between our study and the previous results, namely the use of a single reference gene, *18S*, when compared with our generation of a normalization factor based on several reference genes (309, 346). After revisiting the previous analysis and comparing both methods of normalization, we found the previously reported results to be

replicable only when *18S* alone was used for normalization (309). Furthermore, normalizing *18S* mRNA expression itself with a normalization factor yielded considerable expression variation, suggesting that the use of *18S* alone is a poor choice for the normalization of RT-qPCR data from mouse brain tissues.

It is important to note that our quantification of *HTT* mRNA represented combined mutant and wild-type *HTT* allele transcript levels. As we did not investigate the relative wild-type and mutant *HTT* ratios in our samples, there is a possibility that allele-specific *HTT* transcriptional variation existed in our samples. Alterations in the relative ratios of mutant and wild-type *HTT* have recently been shown to affect the age of onset in HD patients, providing the first evidence of an expression modifier contributing to the variation in the age of onset seen in HD (306). Given that we did not find evidence of any disease-associated DNA methylation variation at probed sites within the *HTT* gene locus or genome-wide, it is unlikely that the presence of altered ratios of *HTT* mRNA in our samples were attributed to DNA methylation differences. This is consistent with previous work showing that allele-specific transcriptional differences are largely driven by genetic regulatory variation as opposed to DNA methylation variation (43).

Our analyses using matched human tissues show that the *HTT* gene locus experiences tissue-specific DNA methylation patterning. Specifically, we identified 38 sites within the *HTT* gene locus that were differentially methylated between matched cortex and liver samples. One potential caveat in these analyses is that we primarily used matched cortex and liver tissues from HD individuals. As we were unable to detect any site-specific association between HD status and DNA methylation at probes within the *HTT* gene locus in cortex tissue, we postulated that this would also hold true in other tissues. While it is certainly possible that the HD mutation has a differential effect on local *HTT* DNA methylation in the liver, the matched control tissues in our cohort did not differ significantly from their HD tissue counterparts. Although not conclusive, this suggests that it would be unlikely that the HD mutation had a differential effect on local *HTT* gene DNA methylation in the liver when compared with cortex.

Using available ChIP-seq datasets, we identified a putative CTCF TFBS within the *HTT* proximal promoter and found that this TFBS was differentially methylated in a tissue-specific manner. An estimated 30–60% of identified CTCF sites have been found to display cell-type-specific occupancy of CTCF that may, in part, be due to tissue-specific DNA methylation of the

CpGs underlying the CTCF-binding sequence (348, 354). Using ChIP- qPCR in both humans and transgenic mice, we found that the CTCF site in the *HTT* promoter displayed higher enrichment in cortex than in liver tissues. Combining this result with our pyrosequencing measures, where we found low levels of CpG methylation of the CTCF site in cortex when compared with the liver, suggests that CTCF occupancy at the *HTT* promoter is regulated by tissue-specific differential DNA methylation. Furthermore, *in vitro* CTCF knockdown revealed a decrease in promoter function, indicating that CTCF may be a transcriptional regulator of the *HTT* gene. In particular, this result suggests that reduced occupancy of the CTCF TFBS in the *HTT* promoter would be accompanied by lower *HTT* expression, a theory supported by our ChIP-qPCR and mRNA expression results in the liver where both CTCF enrichment and *HTT* expression are decreased over cortex. Future studies aimed at addressing the consequences of DNA methylation variation at this CTCF-binding site in relation to CTCF occupancy and *HTT* expression both *in vitro* and *in vivo* would help further clarify the regulation of differential expression of *HTT* across tissue types. Furthermore, as CTCF is also known to be involved in the mediation of long-range DNA interactions, bringing distant transcriptional regulatory regions into contact with the promoter, it is likely that regulatory regions outside the *HTT* proximal promoter may play an important role in expression of *HTT* across human tissues (349).

Chapter 3: Comparative DNA methylation profiling in glioblastoma multiforme tumours and matched brain tumour initiating cells

3.1 Background and rationale

Glioblastoma multiforme (GBM), also known as Grade IV astrocytoma, is the most common and aggressive brain malignancy in humans. Even with advances in standard of care treatment, the median survival for newly diagnosed GBM cases is less than two years (355). These dismal survival outcomes are predominantly attributed to the poor treatment response and acquired resistance to existing therapies, which constitute the main challenges in effective clinical management of the disease (356). Accumulating evidence suggests that tumoural heterogeneity may underlie treatment failure by conferring redundant signaling mechanisms (357–359). Large-scale efforts to characterize the molecular and genomic heterogeneity of GBM have led to the identification of common genetic alterations and frequently dysregulated signaling pathways in these tumours (360–362). Despite these efforts, the reasons for poor treatment outcomes and development of therapy resistance in GBM are still largely unknown, and therefore, represent a considerable barrier in understanding disease development and progression.

One of the most promising discoveries in investigating the potential source of treatment failure and therapy resistance in GBM has been the identification of brain tumour initiating cells (BTICs) (278, 279). Within the bulk GBM tumour, BTICs are thought to comprise a subpopulation of cells with "stem-like" properties of self-renewal and multipotency. They also display tumorigenic potential, as they are able to form secondary tumours that recapitulate the molecular and histological characteristics of the parental tumour when injected at limiting dilutions into immunodeficient mice (281, 363, 364). It is hypothesized that BTICs constitute a "disease reservoir" that drives treatment resistance and spawns disease recurrence during the clinical course of GBM (365, 366). Moreover, as BTICs are directly derived from parental tumour resections and can be cultured long-term, they have been proposed as an attractive *in vitro* model system to study GBM biology (367, 368). To date, substantial focus has been placed on elucidating the diverse genetic alterations and transcriptomic changes in these tumorigenic stem-like populations in comparison to their tumour counterparts (368, 369). However, there has

been increasing appreciation for the potential role of epigenetic factors as an additional source of the poorly understood biological variation within these tumour initiating cell populations.

DNA methylation (DNAm) is arguably the most studied epigenetic mark in humans and involves the covalent attachment of a methyl group to the 5'position of cytosine, typically at cytosine-guanine dinucleotide (CpG) sites (28). These CpG dinucleotides occur relatively infrequently in the genome, thus areas with comparatively high CpG content are termed "CpG islands" (CGIs) (29–31). DNAm levels of specific CpGs is often dependent on their localization within the genome (26). For example, CGIs tend to be less methylated compared to non-island CpGs, and are associated with approximately 70% of known promoters (29, 30). Importantly, DNAm is associated with gene expression, although its role in transcriptional regulation can depend on genomic context (38, 39, 370). For instance, DNAm in promoter regions tends to be associated with gene expression silencing, while its role is more variable within gene bodies (15, 370). Finally, as DNAm is involved in cell fate specification and differentiation, DNAm profiles often exhibit high discordance between tissue and cell types, with tissue identity being the largest driver of DNAm variance followed by cell composition within a tissue (52, 59). Additional common contributors to DNAm variation include environmental exposures, age, ethnicity, technical confounds or a combination of these factors (20, 22, 23, 25, 371). In addition, DNAm has been shown to be under strong genetic regulation, with DNAm levels at individual CpGs associating with genetic polymorphisms known as methylation quantitative trait loci (mQTL) (43, 81, 82, 84, 372). Finally, apart from the canonical 5-methylcytosine (5mC) mark of DNAm, oxidized cytosine variants such as 5-hydroxymethylcytosine (5hmC) have been detected at relatively high levels in pluripotent cells and the brain, where it has been in implicated in active DNA demethylation and neural stem cell activity, although its exact functional role is undefined (109–111, 116). Taken together, DNAm may comprise an important component of cell-specific transcriptional regulatory circuitry, which may be exploited by tumour cells to potentiate and sustain a malignant state (198, 370, 373).

In the context of GBM tumours, pronounced DNAm aberrations have been observed including a number of markers of prognostic significance. For example, when measured in GBM tumours, promoter DNAm of the O^6-methylguanine-DNA methyltransferase (*MGMT*) gene, which encodes a dealkylating DNA repair enzyme, has been associated with loss of *MGMT* expression and favourable outcomes in response to alkylating chemotherapeutics such as

temozolomide (222–224). In addition, low-grade gliomas often exhibit a concerted hypermethylation signature at 1503 promoter located CpG sites, called the glioma-CpG island methylator phenotype (G-CIMP), which has been correlated with improved clinical prognosis over non-G-CIMP tumours (211). Beyond absolute differences in DNAm levels, stochastic increases in DNAm variability have been reported for GBM tumours, and more broadly across other cancer types, compared to their normal tissue counterparts, suggesting a general disruption in the tumour epigenome (374, 375). Finally, dysregulated DNAm at a homeobox gene, *HOXA10*, has been shown to moderate the expression of a HOX-dominant, stem-cell signature in GBM (376). Notably, this HOX-dominant expression signature has been linked to increased treatment resistance and sustained proliferation of BTIC populations from GBM tumours (377, 378). Collectively, these findings underscore the potential importance of DNAm in GBM pathogenesis, particularly in regards to its prognostic utility.

Despite the considerable progress that has been made in profiling DNAm patterns in GBM tumours, very little is currently known about DNAm variation in BTICs. Preliminary DNAm analyses comparing BTICs to normal neural stem cells or unmatched bulk tumours have reported significantly increased BTIC DNAm levels at genes associated with developmental processes, transcriptional regulation, nervous system development and regulation of cellular metabolic pathways (221, 379). However, to date, genome-wide DNAm profiles from BTICs have not been directly compared to their matched GBM tumour counterparts. Such comparisons may help identify epigenetic variation implicated in stem-like functions of BTICs as well as inform the evaluation of BTICs as an appropriate experimental model system to study GBM biology.

In this study, we aimed to characterize how DNAm profiles of BTICs and their parental GBM tumours differ and whether these differences, in part, explain the stem-like and tumorigenic properties of BTICs. Conversely, we also sought to determine the extent to which CpG methylation is concordant between matched samples and how DNAm at key prognostic DNAm markers such as the *MGMT* promoter may vary between BTICs and GBM tumours. To address these questions, we leveraged the strength of profiles from matched BTIC and bulk GBM tumour samples measured on the Illumina Infinium HumanMethylation450K (450K) array platform. We found substantial DNAm variation between paired BTICs and tumours in terms of their DNAm levels and variability at individual CpGs. Notably, we observed a HOX-enriched

signature of differential methylation as well as greater DNAm variability in BTICs over tumours. Despite these differences, the association between DNAm and gene expression at the prognostic *MGMT* promoter region was consistent between matched samples, signifying that BTICs may, to some extent, recapitulate the transcriptional regulatory circuitry in certain regions.

3.2 Materials and Methods

3.2.1 Patient tumour samples and matched brain tumour initiating cell culture

Frozen tumour samples from patients with newly diagnosed and recurrent glioblastoma (Supplementary Table 3.1) were obtained from the Tumor Tissue Bank of the Arnie Charbonneau Cancer Institute (Calgary, Alberta, Canada) and transported to the BTIC Core Facility (Calgary, Alberta, Canada) to culture BTICs as described previously (367, 380, 381). In brief, BTIC lines were initiated in defined culture serum-free medium (SFM). All cultures grew as non-adherent spheres and were cryopreserved in 10% DMSO (Sigma-Aldrich, St. Louis, MO, USA) in SFM for use in later experiments. All established BTIC lines used within this study were validated for identity by short tandem repeat analysis performed by Calgary Laboratory Services (CLS) after each thaw, as per American Association for Cancer Research recommendations. This study has Institutional review board approval under the "Brain Tumor and Related Tissue Bank protocol-V2" and approved by Foothills Hospital and the Conjoint Health Research Ethics Board.

3.2.2 DNA isolation and DNAm arrays

DNA was isolated using DNeasy (Qiagen, Hilden, Germany) as per manufacturer's instructions. DNA was then purified and concentrated using the DNA Clean & Concentrator kit (Zymo Research, Irvine, CA, USA). Sample yield and purity was measured after each step using a Nanodrop ND-1000 (Thermo Scientific, Irvine, CA). Approximately 750 ng of DNA was used for bisulfite conversion using the Zymo Research EZ DNA Methylation Kit (Zymo Research, Irvine, CA, USA). Following bisulfite conversion, samples were randomized and 160 ng of bisulfite-converted DNA was applied to the Illumina Infinium HumanMethylation450 (450K) Beadchip array and ran in two separate batches, as per manufacturer's protocols (Illumina, San Diego, CA, USA) (129).

3.2.3 Data quality control and normalization

Raw intensity values from the arrays were imported into Illumina GenomeStudio V2011.1 software and subjected to initial quality control checks for array staining, extension and bisulfite conversion followed by colour correction and background adjustment using control probes contained on the 450K array. Data were extracted in the form of an average beta value matrix and imported into R Statistical software (Version 3.1.1) for subsequent processing and analysis (http://www.r-project.org). Probe filtering was performed in which 11,648 probes on the sex chromosomes and 65 probes examining single nucleotide polymorphisms (SNPs) were removed from the dataset. Additionally, probes with detection p-values greater than 0.01, probes with missing beta values, and probes for which less than three beads contributed to the signal in any sample were eliminated (a total of 25,068). Recent re-annotation of the Illumina HumanMethylation450k array (324) was used to filter 18,393 probes that are known to be polymorphic at the target CpG and 10,208 probes which have nonspecific *in silico* binding to the sex chromosomes. Together, these measures eliminated 65,382 probes, leaving a total of 420,195 probes for further analysis. The data was then subset into a separate dataset comprising of BTIC lines with matched tumours (35 pairs and 3 technical replicates, n = 73) for downstream processing and analysis. Quantile normalization was not deemed appropriate for this dataset based on the global differences in distribution between the BTIC and tumour samples, as assessed using the quantro package (382). Beta-mixture quantile normalization (BMIQ) was used to normalize differences between Type I and Type II probes on the 450K array (383).

3.2.4 Principal component analysis and correction for technical variation

Principal component analysis (PCA) was performed in which technical variation between batches were detected. To correct for such batch effects, ComBat was then used, protecting for sample status (BTIC or tumour)(327). Subsequent PCA confirmed the successful removal of these technical artifacts from the 450K data. As a check of the processing steps, the correlation between 3 replicate pairs were calculated and found to improve after each stage of normalization and processing, indicating normalization and batch correction successfully removed noise from the data.

3.2.5 Statistical and bioinformatic analyses of DNAm data

Unsupervised hierarchical clustering was performed using Euclidean distance measures with complete linkage. We performed site-specific identification of differentially methylated probes using Wilcoxon signed-rank tests across all CpGs in processed 450K dataset. To test for differentially methylated regions (DMRs), a well-established method for cancer DMR detection, Bumphunting, was used with default settings including specification of a smoothing window 300-900 bp in size with 1000 per mutations (330). All differential methylation analyses were performed on M-values, which are logit-transformed beta values and represent less heteroscedastic measures over beta values (334). Beta values were used for visualization and reporting purposes as they represent percent methylation (0 = no methylation, 1 = fully methylated). We tested for significance of overlap between individual differentially methylated sites and DMRs using 10,000 Monte Carlo simulations to build an expected overlap to compare against observed overlap.

Functional categories enriched in genes mapping to significant DMRs were identified using the functional annotation and clustering tool of the Database for Annotation, Visualization, and Integrated Discovery (DAVID) Version 6.8 (https://david.ncifcrf.gov/) (384, 385). The probability that a gene ontology (GO) biological process term was overrepresented was determined using a modified Fisher's exact test (386).

Inter-individual variability of each CpG was calculated as the range between the 10th and 90th percentile beta values for each CpG, referred to as "reference range" (372). This method captures variability across the bulk of samples while being largely robust to outlier samples. A paired Wilcoxon signed-rank test was performed to generate a significance value for difference between groups. Fligner-Killeen tests were used to compare probewise DNAm variability differences between matched tumours and BTICs. Significant differentially variable CpGs were annotated to underlying genomic features using a previously described method and tested for genomic context enrichment using 10,000 Monte Carlo simulations to build an expected overlap for comparison against observed overlap (39, 292).

Probewise Spearman's correlations were calculated on beta values between the matched GBM tumours and BTICs. To test for enrichment of mQTL associations in significantly correlated CpGs, we used mQTL previously identified in adult human brain tissues (cerebellum, frontal cortex, pons and temporal cortex), comprising of 3916 unique CpGs under observed

genetic influence (79). Using 10,000 Monte Carlo simulations, we built an expected overlap of the mQTL-associated CpGs and significantly correlated CpGs, to compare with the observed overlap.

To assess DNAm at the *MGMT* promoter region, we examined DNAm at 17 CpGs underlying a region shown to have prognostic significance in GBM (chr10:131,264,786-131,265,769; Human Genome GRCh37/hg19 Assembly) (387–389). *MGMT* gene expression levels were obtained from a previous RNA-Sequencing (RNA-Seq) dataset from these samples (submitted), using methods as previously described (390). Spearman correlations were performed to test association between site-specific DNAm at the *MGMT* promoter and *MGMT* gene expression levels.

For all tests, the resulting p-values were adjusted using the Benjamini-Hochberg (BH) false discovery rate (FDR) method (332).

3.3 Results

3.3.1 Cohort sample characteristics and DNAm array data

To compare DNAm profiles between GBM tumours and matched BTICs, we examined tumour specimens and BTIC cultures from a cohort of 35 adult patients (n = 24 males; 74%) diagnosed with primary (n = 24, 69%) or recurrent (n = 10, 28%) GBM; one patient had unknown recurrence status (Table 3.1). Age at diagnosis ranged from 35 to 84 years (median age = 59 years). We generated DNAm profiles from GBM tumours and BTIC samples using the Infinium HumanMethylation450 (450K) array on bisulfite-converted DNA, which quantifies a composite measure of 5mC and 5hmC at over 485,000 sites genome-wide. We applied data quality control processing to filter out poor performing probes, probes that bind directly or non-specifically to sex chromosomes and probes that hybridize to SNPs at the target CpG (324). BMIQ normalization was used to remove probe-type differences and further processing by ComBat was performed to minimize detected batch effects in the data (327, 383). This final 450K dataset, consisting of 420,195 probes, was used for all subsequent analyses.

Characteristic	Number (%)	
All patients	35 (100)	
Age at diagnosis (y)		
35 – 50	6 (17)	(min = 35 y)
51 – 65	20 (57)	(median = 59 y)
66 – 85	9 (26)	(max = 84 y)
Sex		
Male	26 (74)	
Female	9 (26)	
Diagnosis		
Primary	24 (69)	
Recurrent	10 (28)	
Unknown	1 (3)	

Table 3.1 Patient cohort descriptives

3.3.2 BTICs and GBM tumours exhibited substantially differential DNAm profiles

Using the processed 450K dataset, we first sought to identify differences in DNAm measures between BTIC samples and GBM tumours at the genome-level. Based on unsupervised hierarchical clustering of DNAm profiles, matched samples largely clustered by tissue source (BTIC versus tumour), indicating that intra-individual tissue differences in DNAm exceeded inter-individual DNAm differences in the same tissue (Figure 3.1A). These results were corroborated by PCA, which showed that tissue source was significantly associated with PC1 (Wilcoxon signed-rank test, $p = 5.8 \times 10^{-11}$), comprising 23.7% of the variance in DNAm (Supplementary Figure 3.1). Collectively, these results indicated that GBM tumours and matched BTIC cultures exhibited genome-wide differences in their DNAm profiles.

In order to identify site-specific differences in DNAm between matched tumours and BTICs at various FDR thresholds and absolute DNAm difference (delta beta) cutoffs, we performed Wilcoxon signed-rank tests across all 420,195 CpGs in processed 450K dataset (Table 3.2). Using stringent thresholds of FDR ≤ 0.001 and delta beta $\geq 20\%$ established in previous cancer 450K analyses, we detected 60,826 differentially methylated sites between tumours and BTICs (129).Of these 60,826 significant differentially methylated sites, 43,535 sites (71.5%) showed lower DNAm, while 17,291 sites (28.5%) had higher DNAm in tumours over matched BTICs, with absolute DNAm differences ranging from 20% to 82% (Figure 3.1B).

Although site-specific DNAm analysis can be informative in identifying individual CpGs associated with a disease phenotype, the detection of differentially methylated regions (DMRs) represent an additional means of assessing differential DNAm which has been routinely used for

cancer DNAm studies (58, 391, 392). To test for DMRs, we used Bumphunting, a well-established method for cancer DMR detection, which defined over 1086 DMRs, of which 872 were significant at FDR ≤ 0.05 (330). It is worth noting that a less conservative FDR threshold was used for the DMR analysis in contrast to the stringent FDR cutoff applied for the site-specific analysis due to the fact that controlling the FDR at the site-level does not directly relate to region-level control, especially when the region itself has to be defined, as it is for the Bumphunting method (393, 394). As we observed in the site-specific differential DNAm, the majority (856 out of 872, 98%) of these significant DMRs exhibited lower average delta beta across the region in tumours over BTICs, with absolute DNAm differences in DMRs ranging from 23.5% to 69% (Figure 3.1C). We mapped these 872 DMRs to 747 unique genes and found that DMRs occurred across all genomic contexts (upstream, promoter, 5'UTR, gene body, 3'UTR) (Figure 3.1C). The top-ranking DMR at chr12:4381435-4382425, with averaged delta beta of 60%, was located in the promoter region of the *CCDN2* gene, a loci previously described as genetically aberrant in GBM and implicated in other cancer types (Figure 3.1D) (362, 395). Taken together, these site-specific and regional analyses revealed substantially differential DNAm between GBM tumours and matched BTICs.

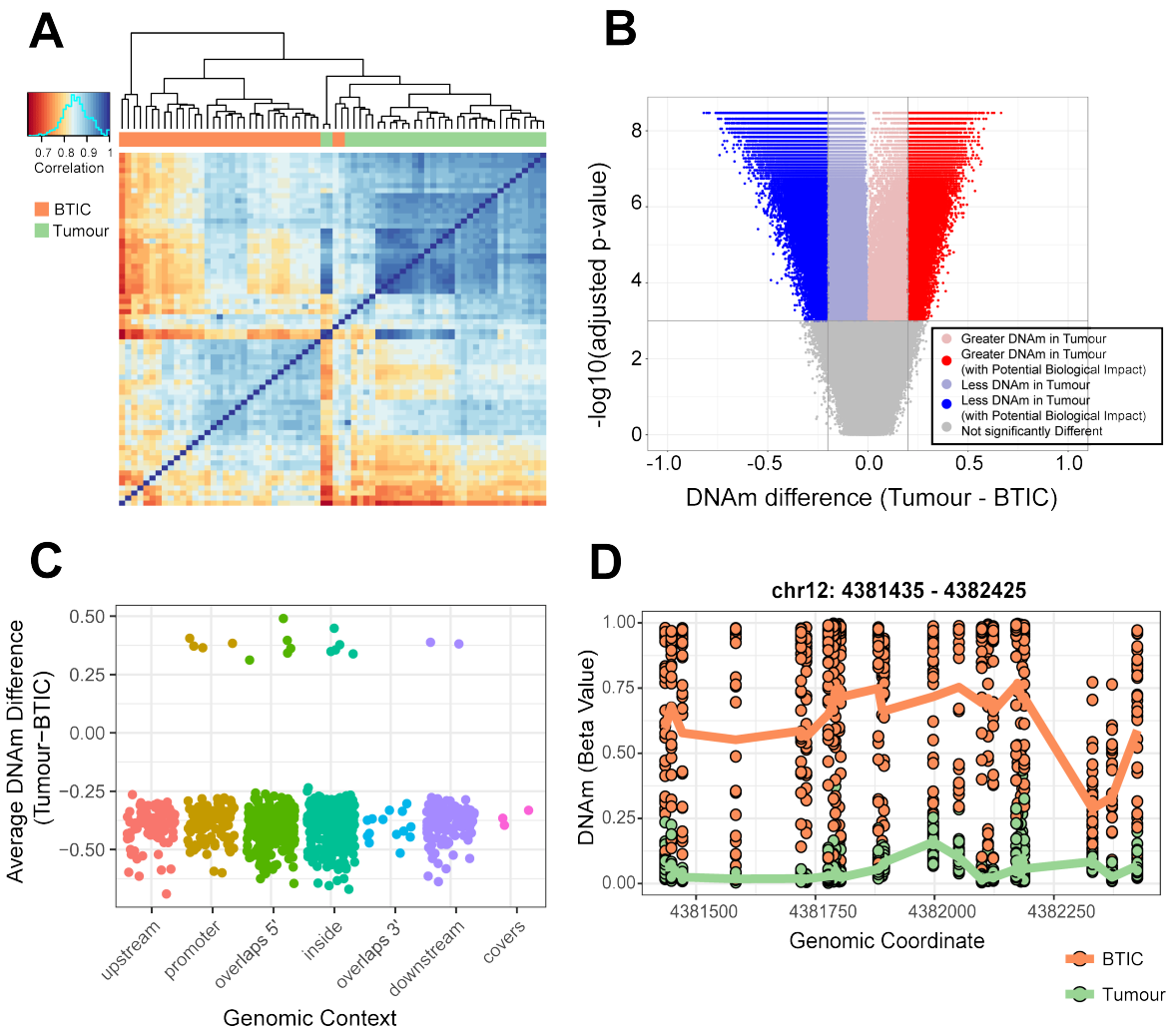


Figure 3.1 BTICs and matched GBM tumours exhibit substantial differences in their genome-wide DNAm profiles.

A) Unsupervised hierarchical clustering of global DNAm profiles largely clustered BTICs (orange) apart from their parental tumours (green). B) Volcano plot depicting differences in DNAm levels (Tumour - BTIC) for all interrogated CpGs in processed 450K dataset (indicated on x axis) against adjusted P-value (indicated on y axis, on $-\log_{10}$ scale). Gray line denotes FDR threshold of 0.001. Coloured points represent statistically significant hits (FDR < 0.001), with darker hues corresponding to hits with absolute DNAm differences > 0.2, while gray points represent non-statistically significant sites (FDR > 0.001). C) All 872 significant DMRs (FDR < 0.05) plotted against their genomic context of their location (x-axis) and average differences in DNAm levels (Tumour - BTIC) across the DMR. D) The top-ranking DMR at chr12:4381435-4382425, located in the promoter region of the *CCDN2* gene, with averaged 60% DNAm levels in BTICs (orange) over matched tumours (green).

	FDR ≤ 0.1	FDR ≤ 0.05	FDR ≤ 0.001
No threshold	230542	208311	129887
Delta beta ≥ 0.05	174409	165687	118114
Delta beta ≥ 0.1	132317	129443	101930
Delta beta ≥ 0.2	62874	62867	60826

Table 3.2 Number of site-specific DNAm hits at various FDR and delta beta thresholds

3.3.3 Differential DNAm was enriched for HOX genes

To delineate biological processes and pathways that are enriched in genes underlying our identified DMRs, we performed functional annotation and GO enrichment analyses using the DAVID online tool (384, 385). Using a modest cutoff of FDR ≤ 0.05, we uncovered 8 significant functional annotation clusters, which comprised of terms involving homeobox, DNA-binding, positive regulation of transcription from RNA polymerase II promoter, fork-head transcription factors, homeobox protein, T-box transcription factors, helix-loop-helix DNA-binding motif and steroid hormone receptor activity (Table 3.3). Notably, the term "homeobox" was represented in two enrichment clusters, including the top-ranking cluster with a high enrichment score of 28.18 (BH-adjusted p-value = 9.3×10^{-31}) and the fifth cluster with an enrichment score of 3.57 (BH-adjusted p-value = 7.3×10^{-3}), signifying that *HOX* genes were disproportionately over-represented in DNAm differences between BTICs and GBM tumours.

To further examine HOX-associated differential DNAm between BTICs and tumours, we focused on significant DMRs which mapped to homeobox genes. Of the 872 significant DMRs identified in our differential methylation analyses, we found 26 DMRs (3%) mapping to HOX genes (Table 3.4). Specifically, there were DMRs underlying multiple genes from the HOXA cluster on chromosome 7p14, HOXC cluster on chromosome 12q13 and HOXD cluster on chromosome 2q31, as well as two other members of the homeobox family (Figure 3.2A and Table 3.4). All of the HOX-associated DMRs showed increased average DNAm in BTICs over matched tumour across the region (Figure 3.2A and Table 3.4). The four top-ranking HOX DMRs were *HOXA6* (chr7:27187269-27188020), *HOXA11* (chr7: 27224568-27225143), *HOXD9* (chr2:176986956-176987605) and *HOXA11-AS* (chr7:27231819-27233141) which had an average DNAm increase of 42.5%, 33.3%, 38.5% and 39.6% in BTICs over GBM tumours, respectively (Figure 3.2B). Collectively, these results supported a potential role for *HOX* genes as an important source of DNAm variation between bulk GBM tumours and BTICs.

Annotation Cluster	Feature/Description	Enrichment Score	Count	P-value	BH-adjusted p-value
1	**Homeobox**	28.18	61	2.4E-33	9.3E-31
2	**DNA-binding**	15.32	168	2.9E-29	5.6E-27
3	**Positive regulation of transcription from RNA polymerase II promoter**	95	95	7.1E-19	2.0E-15
4	**Transcription factor, Fork-head**	10	10	3.7E-5	1.5E-2
5	**Homeobox protein, antennapedia type, conserved type**	3.57	7	4.2E-5	7.3E-3
6	**Transcription factor, T-box**	3.54	6	2.1E-4	2.4E-2
7	**DNA-binding region: Helix-loop-helix motif**	2.91	17	1.2E-4	3.3E-2
8	**Steroid hormone receptor activity**	2.13	9	6.9E-4	3.3E-2

Table 3.3 Functional annotation clusters from DAVID GO analysis.
Top clusters with Benjamini-Hochberg (BH) corrected p-value < 0.05 are listed.

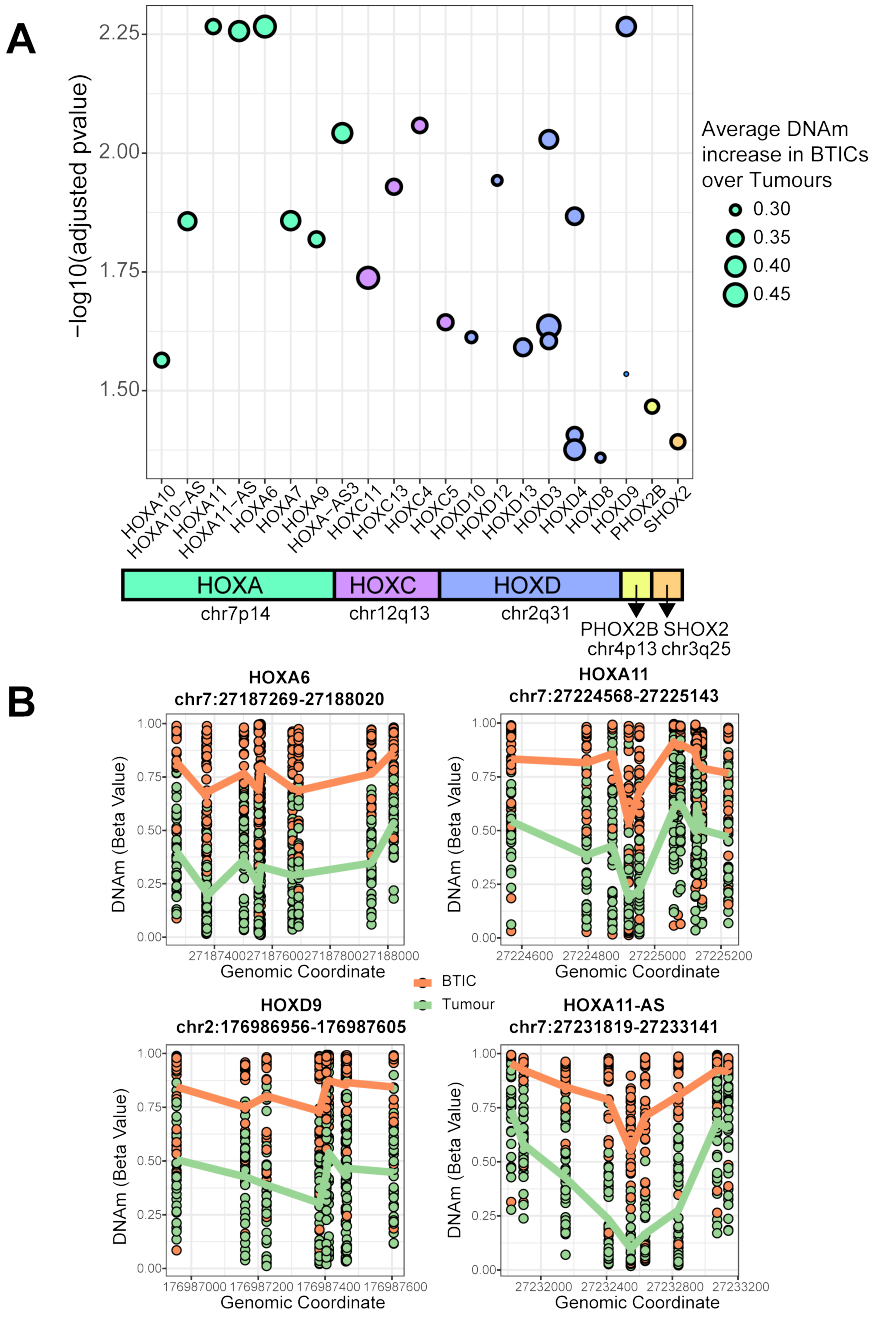


Figure 3.2 Hox-enriched differential DNAm between BTICs and matched GBM tumours

A) Manhattan plot of identified Homeobox-associated DMRs plotted in order of their genomic location on each of the HOX chromosome clusters (x-axis) against their adjusted P-value (indicated on y axis, on $-\log_{10}$ scale). The size of the points corresponds to average increase in DNAm levels in BTICs over matched tumour across the DMR. B) Smoothened scatterplot of DNAm levels (beta values, y-axis) by genomic coordinate (x-axis) for the four top-ranking HOX DMRs, *HOXA6* (chr7:27187269-27188020), *HOXA11* (chr7: 27224568-27225143), *HOXD9* (chr2:176986956-176987605) and *HOXA11-AS* (chr7:27231819-27233141), which had an average DNAm increase of 42.5%, 33.3%, 38.5% and 39.6% in BTICs (orange) over GBM tumours (green).

Genomic Sequence[a]	Adjusted p-value	# CpGs	Average delta beta	Associated gene	Genomic region
chr12:54338743-54339167	0.01	4	-0.342	HOXC13	overlaps exon upstream
chr12:54369102-54369638	0.02	4	-0.427	HOXC11	inside exon
chr12:54424902-54425418	0.02	4	-0.342	HOXC5	inside intron
chr12:54446019-54446308	0.01	7	-0.336	HOXC4	inside intron
chr2:176956396-176957055	0.03	6	-0.365	HOXD13	promoter
chr2:176963948-176964720	0.01	10	-0.298	HOXD12	overlaps 5'
chr2:176981064-176981654	0.02	6	-0.305	HOXD10	overlaps 5'
chr2:176986956-176987605	0.01	9	-0.385	HOXD9	overlaps 5'
chr2:176987918-176988505	0.03	5	-0.285	HOXD9	overlaps exon downstream
chr2:176994142-176994764	0.04	8	-0.295	HOXD8	overlaps 5'
chr2:177015992-177016837	0.01	6	-0.36	HOXD4	overlaps 5'
chr2:177017331-177017449	0.04	4	-0.342	HOXD4	overlaps exon upstream
chr2:177021702-177022056	0.04	4	-0.406	HOXD4	downstream
chr2:177024117-177024502	0.02	4	-0.454	HOXD3	upstream
chr2:177027440-177027931	0.02	5	-0.347	HOXD3	promoter
chr2:177029459-177030228	0.01	8	-0.377	HOXD3	inside intron
chr3:157821919-157822072	0.04	4	-0.332	SHOX2	covers exon(s)
chr4:41746614-41747092	0.03	4	-0.323	PHOX2B	inside exon
chr7:27187269-27188020	0.01	11	-0.425	HOXA6	overlaps 5'
chr7:27191914-27192549	0.01	6	-0.394	HOXA-AS3	inside exon
chr7:27194614-27195036	0.01	5	-0.383	HOXA7	overlaps exon upstream
chr7:27205114-27205262	0.01	7	-0.341	HOXA9	overlaps 5'
chr7:27209338-27209828	0.01	6	-0.366	HOXA10-AS	overlaps exon downstream
chr7:27219224-27219704	0.03	4	-0.33	HOXA10	overlaps exon downstream
chr7:27224568-27225143	0.01	11	-0.333	HOXA11	overlaps 5'
chr7:27231819-27233141	0.01	9	-0.396	HOXA11-AS	downstream

Table 3.4 Identified HOX-associated differentially methylated regions between BTICs and matched GBM tumours

[a]Genome coordinates from Human Genome GRCh37/hg19 Assembly; 5' denotes 5-prime untranslated region

3.3.4 BTIC DNAm was more variable than matched GBM tumour DNAm

Apart from examining differences in absolute DNAm levels, we also sought to assess differences in inter-individual DNAm variability between matched BTICs and GBM tumours as previous work have demonstrated that DNAm variability is highly tissue-specific (57, 59, 68, 292, 295, 317, 396). To measure inter-individual DNAm variability, we calculated a reference range value for each CpG, which is defined as the range between the 10th and 90th percentile beta values for each site (372). We chose to use this reference range measure as it reflects inter-individual variability across the bulk of samples and is largely robust to outlier samples. At the genome-wide level, we found that BTIC DNAm had significantly higher reference range than GBM tumour DNAm, with a 6% greater average reference range in BTICs over tumours Figure 3.3A; Wilcoxon signed-rank test, $p = 2.2 \times 10^{-16}$). To assess reference range differences at individual CpGs, we performed a Fligner-Killeen test at each site and observed that there were 131,307 CpGs that had significantly differential variability between BTICs and matched tumours at FDR ≤ 0.05. These significant differentially variable sites were enriched at intergenic regions and at 2kb regions flanking CGIs (known as 'shores'); conversely, they were depleted at promoters, 3'UTRs, CGIs and regions located 2-4 kb from CGIs (known as 'shelves') (FDR $\leq$ 0.05; Supplementary Figure 3.2). The majority of these variable CpGs (85.2%) had increased reference range measures in BTICs over matched GBM tumours. The six top sites, ranked by highest reference range value in BTICs, exhibited 23-84% higher reference range measures in BTICs over matched tumours (Figure 3.3B). Overall, these findings suggested that BTIC DNAm was more variable than tumour DNAm at both the genome-wide level and across a higher proportion of individual CpGs.

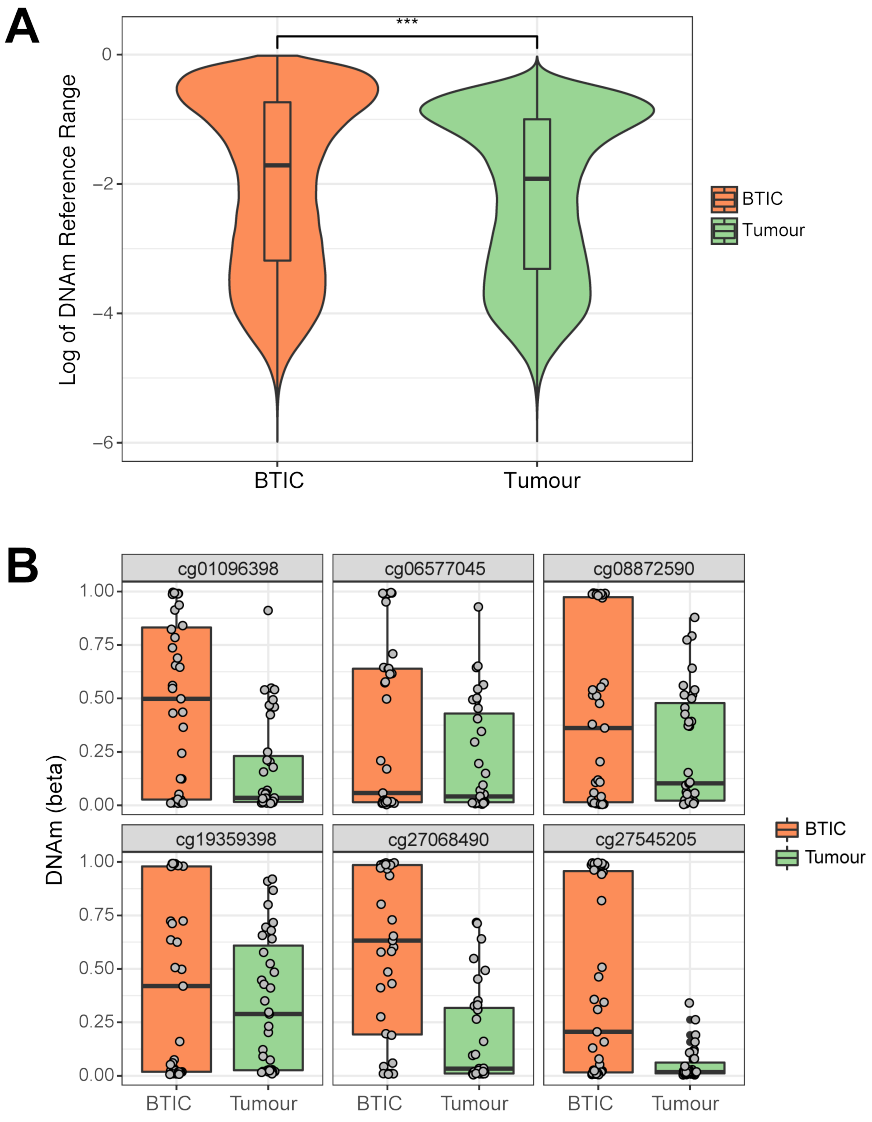


Figure 3.3 BTIC DNAm was more variable than tumour DNAm.

A) Distribution of log of DNAm reference range measures of all interrogated CpGs (y-axis) for BTICs (orange) and tumours (green), showing that BTICs had significantly higher average DNAm reference range of 27.7% (-1.28 on the $\log_{10}$ scale) over tumours which had 21.3% (-1.55 on the $\log_{10}$ scale (***p = 2.2 x 10^{-16}, Wilcoxon signed-rank test,). B) The six top differentially variable sites, ranked by highest reference range value in BTICs (cg01096398, cg06577045, cg27545205, cg19359398, cg27068490 and cg08872590), which had 46%, 43%, 84%, 23%, 46% and 38% higher reference range measures in BTICs (orange) over matched tumours (green).

3.3.5 Identification of concordant CpGs between matched BTICs and GBM tumours

As BTICs have been proposed as an experimental *in vitro* model system to study GBM tumour biology, we rationalized that in addition to examining differences in their DNAm profiles, it may be informative to leverage the strength of our matched tissue design in our study

to explore concordance in DNAm signal between BTICs and GBM tumours. To this end, performed probewise Spearman's correlations between matched samples and assessed how DNAm concordance relates to DNAm variability between these samples. At increasing reference range thresholds, we observed progressively greater enrichment of positively correlated CpGs (Figure 3.4A). Of the 420,195 sites evaluated in our processed 450K dataset, 30,004 CpGs (7.1%) were significantly correlated after multiple test correction (FDR ≤ 0.05). The majority (29,875 out of 30,004; 99.6%) of these CpGs were highly positively correlated with correlation coefficients ranging from 0.48 to 0.98. We enumerated the significantly correlated sites at various thresholds of correlation coefficients and reference range, noting that there were 15 CpGs which exhibited high inter-individual DNAm variability (reference range ≥ 0.5) and high positive correlation (Spearman rho ≥ 0.9) between matched BTICs and tumours (Figure 3.4B and Supplementary Table 3.1). These CpGs may represent sites that are predictive of GBM tumour DNAm when measured in BTICs. Contrary to the high proportion of positively correlated sites, there were only 129 (0.4% of 30,004) sites which were negatively correlated with rho values ranging from -0.48 to -0.64 (Supplementary Table 3.1). None of these negatively correlated sites exhibited high inter-individual DNAm variability (reference range ≥ 0.5). Collectively, these findings suggest that CpGs with greater variability were more likely to be positively correlated between BTICs and tumours.

Given that prior work have suggested that strong concordance in DNAm signal between tissues, particularly at variable CpGs, may be attributed to genetic variation, we speculated that a proportion of our significantly correlated sites between BTICs and tumours may be under genetic influence (75, 77, 79, 89, 292). To explore potential genetic contributions to DNAm concordance between BTICs and GBM tumours, we assessed the degree of overlap between our 30,004 significantly correlated sites and 3619 CpGs that have been previously associated with mQTLs across various human brain regions including cerebellum, frontal cortex, pons and temporal cortex (79). We found a modest, yet significantly greater than expected by chance overlap (288 out of 30,004 sites, ~1%; $p = 0.033$) between our correlated sites and mQTL-associated CpGs in brain tissues. Of note, the 288 overlapping sites which were associated with brain mQTLs exhibited a subtle shift towards increased reference range values over non-mQTL-associated CpGs, although the shift was not statistically significant (Mann-Whitney U test, $p = 0.21$) (Supplementary Figure 3.3). Taken together, these results indicated that shared genetic variation

potentially contributes to DNAm concordance between BTICs and tumours, likely at variable CpGs.

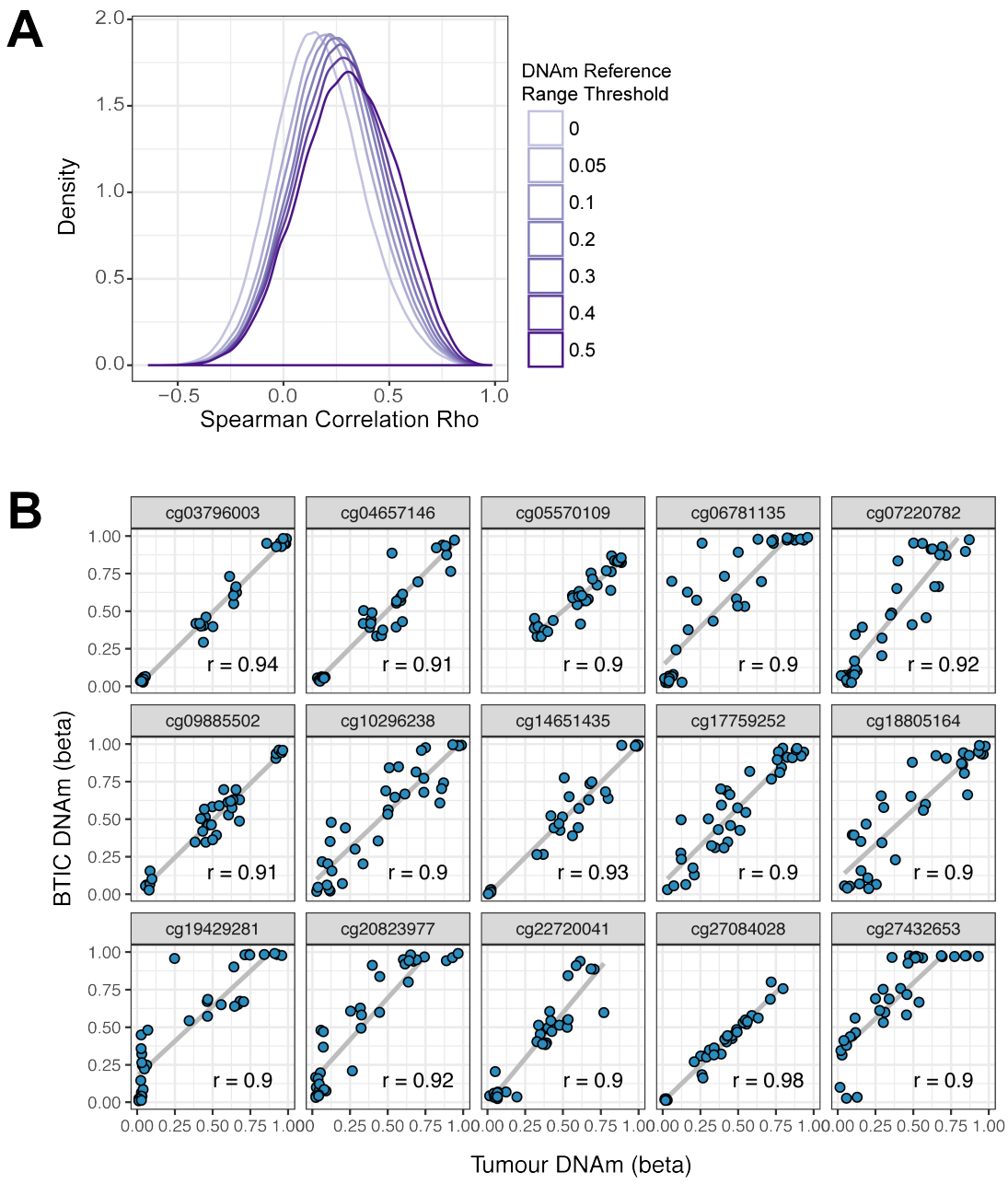


Figure 3.4 Identification of correlated CpGs between matched BTICs and GBM tumours.

A) Density distribution of Spearman's correlation rho between matched BTICs and tumours showing progressively higher enrichment of positively correlated CpGs at increasing reference range cut-offs. Reference range thresholds were set at 0.05, 0.1, 0.2, 0.3, 0.4 and 0.5 (as indicated by the darkening purple hues). B) Scatterplot of DNAm for tumour (x-axis) versus BTIC (y-axis) for 15 CpGs, which exhibited high inter-individual DNAm variability (reference range ≥ 0.5) and high positive correlation (Spearman rho values indicated for each site) between matched samples.

3.3.6 The relationship between DNAm and gene expression at the *MGMT* promoter was consistent between BTICs and GBM tumours

In addition to our unbiased comparison of genome-wide BTIC DNAm to matched tumour DNAm, we were particularly interested in assessing biological variation at the *MGMT* gene promoter, a well-established prognostic DNAm biomarker for GBM when measured in tumours (222–224). As promoter methylation-mediated gene silencing of *MGMT* in GBM tumours is associated with favourable outcomes, we sought to specifically interrogate both promoter DNAm levels and gene expression levels of *MGMT* in BTICs and matched tumours (223, 224, 387, 397, 398). For DNAm measures, we focused on the *MGMT* promoter region at chr10:131,264,786-131,265,769, a region that has been previously shown to have prognostic significance, which was covered by 17 probes in our 450K dataset (387–389). Although this region was not identified in our DMR analyses, we observed modest differential DNAm between BTICs and GBM tumours across this promoter region, with BTICs possessing, on average, 15% higher DNAm levels over tumours, though this difference was not statistically significant ($p > 0.05$) (Figure 3.5A). Using RNA-Seq readouts of the *MGMT* locus, we found that BTICs had significantly decreased *MGMT* gene expression levels (Wilcoxon signed-rank test, $p = 2.83 \times 10^{-7}$) (Figure 3.5B). We performed correlation testing of site-specific DNAm measures in the promoter region and *MGMT* gene expression readouts, finding that 10 out of the 17 tested CpGs (53%) exhibited significant associations between DNAm levels and gene expression (Spearman rank correlation, FDR ≤ 0.05) (Figure 3.5C). Of these, 5 sites were significantly negatively correlated in both BTICs and GBM tumours. Most notably, the trend in relationship between site-specific DNAm and gene expression measures was consistent between BTICs and matched GBM tumours (Figure 3.5C). Together, these findings suggest that despite having differences in gene-specific DNAm and mRNA levels, BTICs may be able to recapitulate the relationship between site-specific DNAm and gene expression as matched GBM tumours, as demonstrated at the prognostic *MGMT* promoter region.

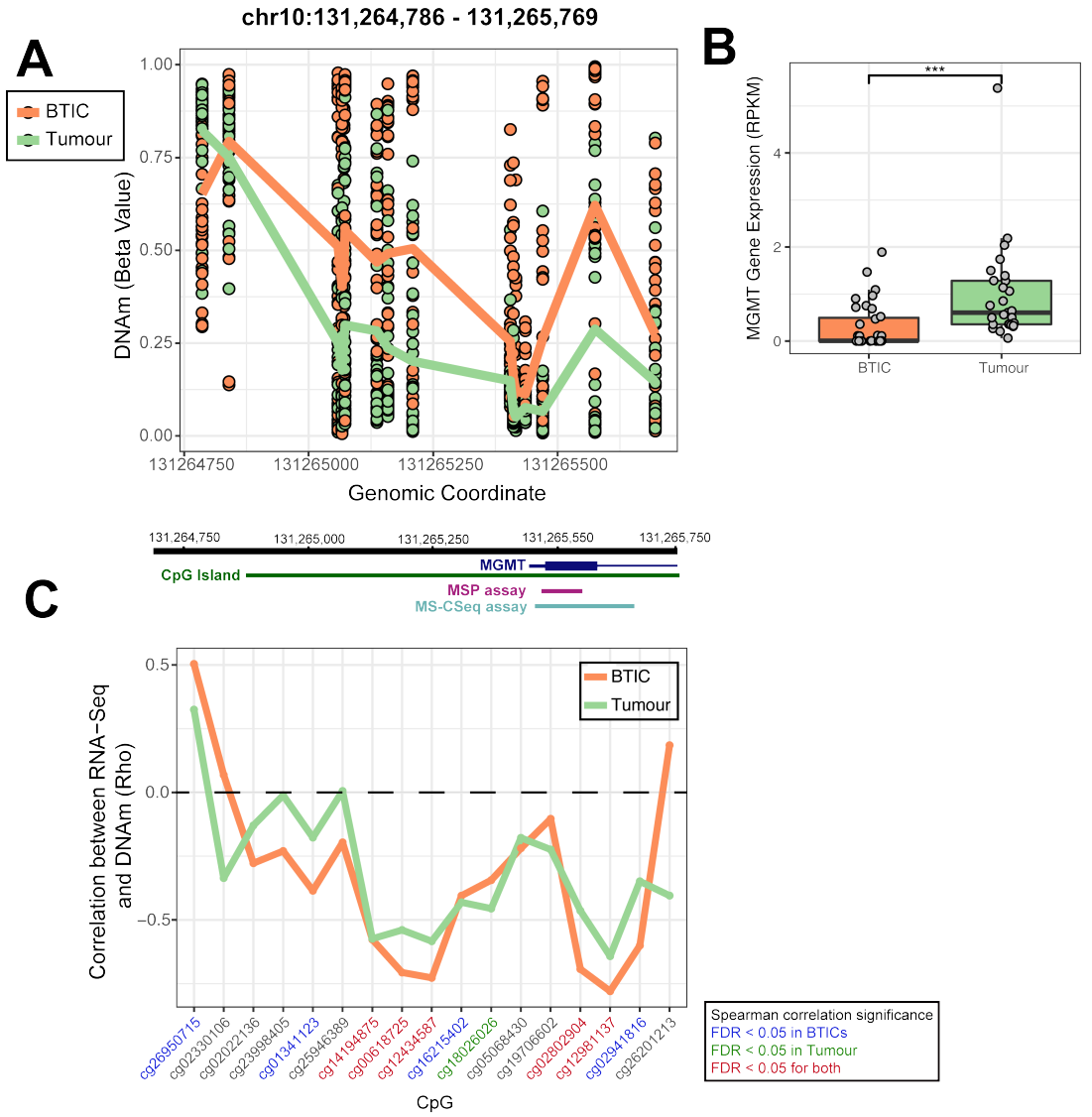


Figure 3.5 The relationship between DNAm and gene expression at the MGMT promoter was consistent between BTICs and GBM tumours.

A) Smoothened scatterplot of DNAm levels (beta values, y-axis) by genomic coordinate (x-axis) for the *MGMT* promoter region at chr10:131,264,786-131,265,769 (Human Genome GRCh37/hg19 Assembly), which had an average DNAm increase of 15% in BTICs (orange) over GBM tumours (green) across the region. *MGMT* gene features of the genomic region are depicted in the schematic below (blue), along with underlying CpG island (green) and regions assayed by methylation-specific (MS) PCR (MSP, purple) and MS-clone sequencing (MS-CSeq, turquoise) (387, 388). B) Boxplot of *MGMT* RNA-Seq readouts (RPKM) (y-axis) for BTICs and tumours, showing significantly lower expression levels in BTICs (orange) over matched tumour (green)(***$p = 2.83 \times 10^{-7}$, Wilcoxon signed-rank test). C) Smoothened dot plot of Spearman correlation rho values between *MGMT* gene expression (RPKM) and beta values of the 17 CpG probes from 450K dataset at *MGMT* promoter region (chr10:131,264,786-131,265,769; Human Genome GRCh37/hg19 Assembly) (y-axis) ordered by genomic coordinate location (x-axis) (relating to gene schematic from A). Sites indicated in red were significantly correlated between gene expression and DNAm at FDR < 0.05 for both BTICs and tumours, sites indicated in blue denote significantly correlated sites (FDR < 0.05) in BTICs exclusively and sites shown in green denote significantly associated sites (FDR < 0.05) in tumours only.

3.4 Discussion

The generation of patient-derived primary BTIC cultures offers a potential model system to study GBM biology, particularly in terms of their molecular features. However, at present, it is largely unclear how the epigenomes of these tumour-initiating cell subpopulations differ from their parental bulk tumour tissue. In this study, we compared genome-wide DNAm profiles from matched BTICs and bulk GBM tumours in terms of their site-specific levels, variability and concordance. We observed substantial differences between paired BTICs and tumours in their global DNAm profiles, identifying a differential HOX-enriched DNAm signature with multiple homeobox genes from the HOXA, HOXC and HOXD clusters as well as other HOX gene family members. Beyond DNAm levels, we also examined differences in DNAm variability, finding that BTIC DNAm was more variable than tumours with increasingly variable CpGs more likely to be positively correlated between matched samples and potentially under genetic influence. Finally, we found that the association between DNAm and gene expression at the prognostic *MGMT* promoter region was consistent between BTICs and tumours, suggesting that processes implicated in transcriptional control may, to some extent, be conserved at certain regions between BTICs and parental tumours. Overall, our data highlight DNAm as an additional constituent in the biological variation between BTICs and bulk tumour tissue.

Our finding of a HOX-enriched differential methylation signature between BTICs and matched tumour tissue expands on a body of work which implicates HOX gene dysregulation in GBM pathogenesis and BTIC function (279, 376–379, 399–401). HOX genes are a highly conserved family of genes encoding homeodomain transcription factors that are involved in the regulation of body patterning during embryogenesis and throughout development (402). Previous studies have linked aberrant overexpression of HOX genes with gliomagenesis and treatment resistance (279, 399, 400). Notably, a stem-cell-related, gene expression signature dominated by HOX genes is predictive of poor survival in GBM patients, irrespective of their age or tumour *MGMT* DNAm status (377). Activation of this HOX expression signature in GBM occurs through dysregulated action of epignenetic regulators such as Trithorax proteins at *HOXA10* as well as inappropriate PI3-kinase signaling through *HOXA9* (378, 379, 401). We noted a strong gene overlap between the HOX expression signature and our HOX-enriched DNAm alterations (ie *HOXD10, HOXD4, HOXD8, HOXA9, HOXA7* and *HOXA10*; 6 out of 26 loci; 23%). Moreover, our observation of HOX-enriched hypermethylation in BTICs is consistent with

previous reports of increased DNAm at HOX genes in BTICs over normal neural stem cells (379). However, unlike previous comparisons to normal tissues, our work leverages the strength of matched samples to highlight DNAm as a potential source of epigenetic variation between bulk tumour tissue and BTIC cultures. As other forms of epigenetic regulation, including histone modifications and microRNA activity, have been associated with the HOX expression signature, it is tempting to speculate that differential DNAm may represent part of a larger, coordinated epigenomic program that aberrantly alters HOX gene function to confer tumorigenic, stem-like properties in BTICs over non-tumorigenic bulk tumour cells (376).

As intratumoural cellular heterogeneity is a key feature of GBM malignancy, we anticipate that our differential DNAm findings may be attributed, in part, to cellular heterogeneity. GBM tumours are highly heterogeneous in their cellular composition, consisting of an admixture of neoplastic astrocytes, infiltrating immune cells and stromal components, which can dilute the 'tumoural purity' of each tumour sample (403, 404). Within their malignant fraction, GBM tumours contain cellular niches that are enriched for distinct functional properties including resistance to DNA-damaging radiotherapy, adaptation to hypoxia, transient quiescence and self-renewal (278, 405–409). In regards to their latter stem-like attributes of quiescence and self-renewal, single-cell transcriptomic analysis of GBM samples has demonstrated that cellular states within an *in vivo* tumour represent a continuous gradient of stemness features while *in vitro* BTIC cultures emulate extremes of stem-like profiles (410). As our DNAm analyses were performed on whole tumour resections and BTIC cultures, our findings of differential DNAm levels and variability likely reflect both cell-type differences and stem-like states between matched BTICs and bulk tumour cells. Although methods to deconvolute the 'tumoural purity' of solid tumours from DNAm profiles has been developed, it is unclear whether these approaches would be applicable for primary cell cultures such as BTICs (404, 411).

Apart from differences in DNAm levels, our results also identified differences in DNAm variability between matched BTICs and tumours. Specifically, we observed that nearly one-third of all queried probes (131,307 out of 420,195; 31%) had differential DNAm variability between BTICs and tumours, with the majority of sites having greater variability in BTICs over tumours. The high proportion of variable CpGs is in line with previous studies which showed that unlike their normal tissue counterparts, cancer samples exhibit highly variable CpG methylation levels in a large proportion of the genome, indicating substantial stochastic DNAm variation in the

tumour epigenome (374, 412). Our study also sought to delineate DNAm concordance between matched BTICs and tumours, finding that strong positive correlation in DNAm signal were more likely to occur at variable CpGs. Notably, these correlated sites tended to be enriched for CpGs under genetic influence, as demonstrated by the overlap of correlated sites and previously described mQTL-associated CpGs in the brain. Although we were underpowered to perform mQTL analyses in our sample set, these findings suggest that 1) BTICs retain genomic information from their parental parents, to the level of single nucleotide polymorphisms, as shown in previous reports and 2) common genetic variation between BTICs and tumours may have similar effects on DNAm levels between matched samples (368, 379). However, it is worth noting that the previously described list of brain mQTL used in the enrichment analysis was generated on the 27K Infinium array platform, which has reduced genomic coverage compared to the 450K array (79). As our samples were analyzed on the larger 450K platform, our analyses have likely underestimated the degree of overlap between the correlated sites and mQTL-associated CpGs in the brain.

Concordance in biological variation at the *MGMT* promoter is of particular interest in the context of GBM, as promoter DNAm and subsequent silencing of *MGMT* expression in tumours are well-accepted clinical indicators of response to alkylating chemotherapies and favourable prognosis (222–224). Our finding that the association between site-specific DNAm and gene expression measures was consistent between BTICs and matched GBM tumour suggests that processes implicated in transcriptional control may, to some extent, be conserved at certain regions between BTICs and parental tumours. Additionally, our data indicate that the prognostic utility of the *MGMT* promoter region may be deployed in BTIC models, although we were unable to directly test this hypothesis in survival analyses as patient outcomes were not available. Of note, previous 450K profiling of the *MGMT* promoter in GBM tumours, also reported prognostic significance for two of the same CpGs in our analysis (cg12434587 and cg12981137), which underlie the region that was formerly interrogated by methylation-specific (MS) PCR and MS-clone sequencing assays for clinical assessment (387–389).

It is worth noting that our study had a number of technical challenges and limitations. Firstly, we were unable to resolve DNAm signal from other forms of DNA modifications, such as DNA hydroxymethylation (DNAhm), because the bisulfite conversion process of DNA generates composite measures of both canonical 5mC and 5hmC. Variation in DNAhm, which

has been shown to exist at appreciable levels in human brain tissues and pluripotent stem cells, may be particularly relevant in GBM pathology as the Ten-Eleven Translocation (TET) enzyme that catalyzes 5hmC can be indirectly inhibited by the action of a commonly mutated gene in gliomas, *IDH1/2* (98, 117, 413–416). Thus, future work aimed at delineating 5hmC signal from 5mC levels may help clarify the relative contribution of these cytosine modifications to the epigenetic variation in matched BTICs and GBM tumours. Our second limitation arises from the fact that we are unable to rule out potential cell culture-induced artifacts in our DNAm data. It is worth noting that these culture-induced effects have largely been attributed to passage number and immortalization of cell lines, as opposed to primary cultures (286, 417, 418). Nevertheless, we attempted to reduce *in vitro* culture-induced variation by ensuring a limited number of passages (< 10) and uniform neurosphere assay conditions were used, as supported by the strong average correlation (Spearman rho = 0.99) in genome-wide DNAm profiles between BTIC replicates. Finally, as we were underpowered to perform genome-wide association analyses between DNA sequence variation to DNAm in our samples, we were unable to determine potential genetic regulation of DNAm, which has been demonstrated in other cancer types (419).

Despite these limitations, our study provides the starting groundwork for the exploration of epigenetic variation between BTICs and parental GBM tumours. In our analyses, we observed widespread, HOX-enriched DNAm alterations between BTICs and matched tumours as well as potential conservation of DNAm-mediated gene silencing of *MGMT* prognostic marker between matched samples. In this manner, BTICs may provide a unique opportunity to dually identify novel therapeutic targets as well as implement existing prognostication strategies for GBM management.

Chapter 4: DNA methylation signatures of chronic alcohol dependence in purified $CD3^+$ T-cells of patients undergoing alcohol treatment

4.1 Background and Rationale

Alcohol dependence (AD) is a severe disorder that has long-lasting detrimental consequences, resulting in considerable health, economic and societal burden. According to the World Health Organization, alcohol related diseases account for approximately 3.3 million deaths per year (WHO, 2014). Although this number is alarmingly high, studies indicate that problematic drinking behaviour still is underestimated (420). To date, treatment options are limited and the effectiveness of existing alcohol treatment programs is often less than optimal or difficult to assess, warranting a need for improvement.

The pathogenesis of AD is complex and includes genetic as well as non-genetic factors. Evidence is emerging that the interaction between underlying genetic factors and environmental stimuli (gene x environment, GxE) in particular plays a major role in addiction-related disease states(421–423). Such findings have prompted considerable inquiry into the biological basis of GxE influences, with epigenetic regulation providing one of the most compelling candidate mechanisms for the mediation of GxE effects (6, 10).

One of the most frequently studied epigenetic mechanisms is DNA methylation (DNAm), which involves the covalent addition of a methyl group to the 5′ position of a cytosine, primarily in the context of a cytosine-phosphate-guanine (CpG) dinucleotide. CpG dinucleotides are especially prevalent in CpG islands, genomic regions of approximately 1000 base pairs (bp) with a CG content greater than 50% (29). CpG islands are associated with 50–70% of human gene promoters and increased DNAm in these regions is generally correlated with a decreased transcription of the respective gene (26, 424). Furthermore, methylated regions adjacent to CpG islands, called CpG island shores (up to 2 kb in either direction) or shelves (from 2 to 4 kb in either direction), may contribute to and potentiate epigenetic effects on gene expression (39, 425, 426). In recent years, there has been increasing appreciation for the complexity of the relationship between DNAm and gene expression regulation, which tends to be highly dependent on genomic context (26, 38). DNAm profiles of genetic regions can vary substantially between different cell types (427). It has been shown that after tissue origin, cellular heterogeneity within

a tissue is a major driver of DNAm variance, highlighting the need to account for cellular composition in DNAm analyses (59, 428).

Several biological factors including age, sex and ethnicity also have a profound impact on DNAm patterns (429–431). In addition, a number of lifestyle-based environmental exposures, including smoking and alcohol consumption, are associated with variation in DNAm (17–19, 184, 432–444). In particular, DNAm alterations in AD patients have been documented in a number of epigenetic studies in human populations. For example, candidate gene analyses reported differential DNAm of the dopamine and serotonin transporters, the nerve growth factor *NGF*, leptin and most recently *GDAP1* in AD patients compared to healthy controls (433, 435, 441, 443, 444). In the context of epigenome-wide association studies (EWAS), previous studies found widespread AD-associated DNAm differences at single sites, differentially methylated regions (DMRs) and in "bulk" DNAm, representing mean global total levels of DNAm (184, 434, 436–438, 442). One study assessed DNAm alterations in peripheral blood mononuclear cells (PBMCs) of AD patients participating in a short-term alcohol treatment program compared to healthy controls, and reported differential methylation at 56 CpG sites in patients prior to treatment compared to controls. Although no statistically significant DNAm differences were observed in patients before and after the alcohol treatment program, 49 of the 56 differential sites reverted back in patients post-treatment to levels similar to controls (432). Together, these previous studies identified a multitude of AD-associated differentially methylated sites, however, they did not account for cell type heterogeneity in their analyses, thereby potentially resulting in associations that are confounded by inter- and intra-individual differences in cellular composition. Most recently, a study involving 13,317 participants from 13 distinct cohorts analysed DNAm profiles in monocytes and whole blood. This analysis, which was adjusted for cell composition, revealed hundreds of AD-associated differentially methylated CpG sites (438).

Although all these previous studies support a potential link between DNAm variation and AD, a number of questions have yet to be explored: I) Are there signatures of AD in a disease-relevant blood cell type? II) Does treatment result in reversion of differential DNAm back to the levels found in controls? III) Importantly, can such AD-associated differential DNAm be replicated in independent cohorts, signifying the robustness of the identified genome-wide hits, and IV) Can the differential DNAm from a purified blood cell type also be detected in whole blood samples, indicating the potential relevance of these associations in other blood cell types?

To address these questions, we assessed genome-wide DNAm profiles of purified $CD3^+$ T-cells of a well-characterized cohort of long-term chronic AD patients participating in a clinical 3-week alcohol treatment program, along with the profiles of healthy controls closely matched for sex, age, ethnicity and smoking behaviour. We restricted our analyses to T-cells due to the known effects of chronic alcohol abuse in modulating the number, activity and relative subtype abundance levels of these immune cells (445). For example, short-term binge drinkers as well as chronic AD patients exhibit a reduced number of peripheral T-cells (446). In addition, a shift from $CD4^+$ and $CD8^+$ naïve T-cells towards memory T-cells is observed in AD patients (447). Furthermore, alcohol consumption influences T-cell activation, leading to elevated numbers of activated $CD8^+$ T-cells, which may contribute to chronic inflammation (445, 447). For these reasons, heightened susceptibility to infections, including tuberculosis, pneumonia and HIV is observed in those patients (445, 448). T-cells have also been used previously in similar epigenetic studies due to their regulatory function in neuroimmune mechanisms (181, 449). Furthermore, by comparing the patients before and after 3 weeks of participating in a clinical alcohol treatment program, we sought to identify differentially methylated sites that may play a potential role in alcohol withdrawal and early recovery. In order to test whether our findings were robust, we validated four of our top-ranked hits by pyrosequencing, replicated the top-ranking hits in an independent second cohort of AD patients and matched controls and additionally confirmed the top-ranking hits in whole blood DNA of our cohort samples.

4.2 Materials and Methods

4.2.1 Study cohorts

The discovery study cohort was comprised of 24 male AD patients (mean age 47.5 ± 10.1 years) participating in a 3-week in-patient alcohol treatment program at the Clinic for Psychiatry and Psychotherapy in Tuebingen, Germany. AD was diagnosed according to the fourth edition of the Diagnostic and Statistical Manual of Mental Disorders (DSM-IV). Twenty-three population based, sex- and age-matched healthy controls (mean age 46.9 ± 10.3 years) were recruited from Tuebingen and the surrounding area. The replication study cohort was comprised of 13 male AD patients (mean age 50.9 ± 9.1 years) and 12 matched healthy controls (mean age 45.3 ± 16.2 years). In addition, the smoking behaviour (measured as cigarettes per day) of both groups was matched. Subjects with a dependence other than nicotine and patients with any psychiatric

disorder necessitating psychotropic medication were excluded from the study. All subjects were of Caucasian origin and gave written informed consent after recovering from alcohol intoxication (patients) or prior to participation in the study (controls), which was approved by the ethics committee of the University of Tuebingen and was conducted in accordance with the Declaration of Helsinki.

After recovery from alcohol intoxication and at the time of study inclusion, respectively (time point 1, T1), patients and controls answered a self-administered phenotypic and demographic questionnaire, the Alcohol Use Disorder Identification Test (AUDIT), assessing alcohol consumption, and the Symptom Checklist-90-R (SCL- 90-R) questionnaire, assessing the global distress level (GSI) (450, 451). Patients also answered the obsessive compulsive drinking scale (OCDS-G) questionnaire, reflecting obsession and compulsivity related to craving and drinking behavior (452). OCDS-G and SCL-90-R were reassessed after three weeks (± 2 days) of participation in the alcohol treatment program (time point 2, T2). Controls with AUDIT scores >15 were excluded, as a higher value is suggestive for problematic alcohol intake.

At T1 and T2, peripheral venous blood was drawn from patients in Ethylenediaminetetraacetic (EDTA) and Mononuclear Cell Preparation tubes (CPT, both BD, Franklin Lakes, NJ, USA). EDTA and CPT blood samples from the controls were drawn at study inclusion. Samples for whole blood DNA extraction were kept at −80 °C until further usage.

4.2.2 $CD3^+$ T-cell purification and DNA isolation

Immediately after blood draw, PBMCs were first separated via centrifugation of the CPT tubes for 20 min at 1650 × g. $CD3^+$ T-cells were then purified from PBMCs following the positive isolation protocol using Dynabeads CD3 (Invitrogen, Carlsbad, CA, USA). The cells were subsequently lysed and DNA was prepared using the QIAamp DNA Mini Kit (Qiagen, Hilden, Germany) according to standard protocol.

4.2.3 Bisulfite conversion and Illumina 450K DNAm arrays

T-cell DNA (750 ng) was bisulfite converted using the Zymo Research EZ DNA Methylation Kit (Zymo Research, Irvine, CA, USA). DNA yield and purity was assessed using a Nanodrop ND-1000 (Thermo Fisher Scientific, Waltham, MA, USA). Samples were

subsequently randomized and 160 ng of bisulfite-converted DNA was applied to the Illumina Infinium HumanMethylation450K (450 K) Beadchip array, as per manufacturer's protocols (Illumina, San Diego, CA, USA)(129).

4.2.4 DNAm array data quality control and normalization

Quality control, processing and differential DNAm analysis of 450K array data was performed as previously described (453, 454). Briefly, raw intensity values from the arrays were imported into Illumina GenomeStudio V2011.1 software and subjected to initial quality control checks for array staining, extension and bisulfite conversion followed by color correction and background adjustment using control probes contained on the 450K array. Subsequent processing and analysis were performed in R Version 3.2.1 (http://www.r-project.org). Profiles from 65 probes targeting single nucleotide polymorphisms (SNPs) were used to ensure T1 and T2 samples were indeed matched from the same individual. The 65 SNP probes were also filtered out of the dataset. Additional probe filtering was performed in which poor performing probes including those with detection *P*-values greater than 0.01, probes with missing beta values, and probes for which less than three beads contributed to the signal in any sample were eliminated (a total of 13 903). Recent re-annotation of the Illumina 450K array was used to filter 19 343 probes that are known to be polymorphic at the target CpG (324). Probes which have nonspecific in silico binding to the sex chromosomes were assessed in a post-hoc analysis following differential DNAm analysis to ensure they did not overlap with identified hits (324). Together, quality control checks eliminated 33 311 probes, leaving a total of 452 266 probes for further analysis. Following quality control processing, quantile normalization was conducted using the lumi R package after assessment using the quantro package indicated that quantile normalization was appropriate for this dataset (325, 382). Differences between Type I and Type II probes on the 450K array were normalized using Subset-quantile Within Array Normalization (SWAN)(326). ComBat was then used to remove chip and row effects, while protecting sample group. Removal of technical variation was assessed by principal component analysis (PCA) (327).

4.2.5 Blood cell type deconvolution

To test for potential contamination of bead-purified $CD3^+$ T-cell samples by other blood cell types, a well-established algorithm was used to bioinformatically estimate cell type composition based on underlying reference DNAm profiles (63, 455). In addition, the 450K data was subjected to advanced DNA methylation age analysis in blood using a publicly available DNA methylation age predictor tool in order to obtain predicted abundance measures of additional blood cell types including plasma blasts, $CD8^+CD28^-CD45RA^-$ (memory and effector) T-cells, naïve $CD8^+$ T-cells and naïve $CD4^+$ T-cells (322). Upon detection of potential non-T-cell contamination in a fraction of samples, we removed this cell-type composition variation by regressing probewise DNAm on estimated cell type proportions, as previously described (456). The residuals of each regression model were applied to the mean value of each data series to obtain the 'corrected' DNAm data. PCA was subsequently used to check that the presence of the cell type proportions in DNAm variation was minimal in the corrected dataset. PCA was additionally used to check for correlation of other known meta-variables (i.e. sample group, age, daily smoking) with the underlying DNAm patterns of the uncorrected and corrected 450K datasets, respectively. Note that for all PCA analyses, the top-ranking PC (denoted as PC0) was negated as it is not informative of inter-individual variance in the DNAm data (38).

4.2.6 Differential DNAm analyses of 450K dataset

The cell-type corrected 450K dataset was subsetted into controls versus patients (T1), patients (T1) versus patients (T2) and controls versus patients (T2) sample sets, respectively, prior to differential DNAm analysis. In the genome-wide analyses, differentially methylated probes were identified using the R limma package's moderated t-statistics with empirical Bayesian variance estimation (331). Specifically, in the comparison of controls versus patients (T1), a linear model was fit for each probe's DNAm measures with sample group as the main effect, adjusted for age and smoking levels. In the comparison of patients (T1) and patients (T2) samples, differentially methylated probes were identified using paired testing in linear regression analysis. For both of these comparisons, differentially methylated regions (DMRs) were detected using DMRcate package which uses the moderated t-statistics generated in their respective limma analyses (457). In the comparison of controls versus patients (T2), we sought to assess which differentially methylated sites between controls and patients (T1) exhibited reversion in

the patient (T2) samples such that their DNAm levels were comparable to controls. To address this, we specifically tested the 59 hits identified between controls versus patients (T1) (FDR < 0.1 and DNAm difference > 5%) by fitting individual linear models for each of the 59 probes. For all tests, the resulting *P*-values were adjusted using the Benjamini-Hochberg False Discovery Rate (FDR) method (332). All statistical analyses were performed on transformed M-values (334). The 450 K data has been made publicly available on the Gene Expression Omnibus database (GSE98876).

4.2.7 Pyrosequencing-based validation and replication in T-cells

500 ng T-cell DNA was bisulfite-converted using the Epitect Fast Bisulfite Conversion Kit (Qiagen) as described earlier (441). For amplification of the region of interest, PCR was conducted using the PyroMark PCR Kit (Qiagen) with the following primers: forward (fwd): 5′-GTTATGGTTGGGTTTTTGGG-3′, reverse (rev): 5′-Bio-CCTATCTCCTCAAACAAAAACTAAAAA-3′, sequencing (seq): 5′-AGTTAGGGATTATAGTGTAGTTG-3′ (cg07280807); fwd: 5′-GTGTTTGTGGGAATGTTTTTTATA-3′, rev: 5′-Bio-CACACTACACTTTCATTTTCTATCAA-3′, seq: 5′-TTTTTAGATATATAAATTTTTTTTTT-3′ (cg18752527) and fwd/seq: 5′-GTTATTTATAAAGGAGGGTGAGATTA-3′, rev: 5′-Bio-AACCACTACTCCTATAAAACCCCAC-3′ (cg16529483/cg24496423). A detailed list of PCR primers and programs is provided in Supplementary Table 4.4. Specificity of the PCR was verified by agarose gel electrophoresis including a negative control. Pyrosequencing was conducted on a PyroMark Q24 according to standard protocol using PyroMark Gold Reagents (both Qiagen). Each sample was measured in triplicates; an intra-sample deviation of ≥3% led to the exclusion of the deviating measurement. For each site, measurements of DNA with known methylation levels of 0%, 25%, 50%, 75% and 100% were obtained (Epitect Control DNA, Qiagen). Correlations between the 450 K dataset and pyrosequencing were tested using the Spearman's correlation test.

4.2.8 Pyrosequencing-based validation and replication in whole blood

DNA was prepared from EDTA tubes using QIAamp DNA Blood Maxi Kit (Qiagen) according to manufacturer's protocol. Afterwards, bisulfite conversion and pyrosequencing was carried out as described above.

4.2.9 Questionnaire evaluation

The AUDIT score is the sum of all 10 items of the questionnaire. The GSI score represents the sum of all the subscales of the SCL-90-R divided by the number of answered items (usually 90). For the OCDS score, the higher value of four item pairs (Items 1 and 2, 7 and 8, 9 and 10, and 12 and 13) was added up with the remaining items, leading to a potential range of 0 to 40. Up to one missing item was allowed and replaced by adding the mean of all other items.

4.3 Results

4.3.1 Study cohorts and DNAm array normalization

To identify AD-associated DNAm variation, we utilized a discovery and replication cohort of AD patients and healthy controls, who were closely matched for age, sex and smoking behaviour. Demographic and AD-relevant characteristics as well as AUDIT and GSI scores of both cohorts are provided in Table 4.1A-B. To measure the effectiveness of the 3-week alcohol treatment program, we compared both GSI and OCDS scores in the discovery cohort at the beginning and after treatment. We found that both values decreased significantly, indicating a reduced alcohol craving and a better overall psychological well-being post-treatment (Table 4.1C).

In order to assess the association of AD with genome-wide DNAm in our discovery cohort, we measured site-specific DNAm at over 450,000 CpGs using the Illumina 450 K array. To test for potential cellular heterogeneity in the bead-purified $CD3^+$ T-cell samples, we used the Houseman blood deconvolution algorithm to estimate cell-type proportions, observing up to 32% of contaminating non-T-cell DNA in a fraction of our samples, although these proportions were not correlated with group status (Supplementary Figure 4.1). Regression-based adjustment of 450 K data resulted in the removal of these cell type associations as assessed by PCA (Supplementary Figure 4.2). The adjusted dataset thereby represented DNAm profiles from T-

cells whose inter-individual cell type differences had been normalized to the best of our abilities for subsequent analyses.

	a) Discovery study cohort			b) Replication study cohort		
	Controls (N = 23)	Patients (N = 24)	*P*-value	Controls (N = 12)	Patients (N = 13)	*P*-value
age	46.9 ± 10.3	47.5 ± 10.1	0.8	45.3 ± 16.2	50.9 ± 9.1	0.4
active smokers	18 (78%)	19 (79%)	0.9	8 (67%)	9 (69%)	0.9
cigarettes per day	13.8 ± 12.6	15.2 ± 10.7	0.7	8.9 ± 8.0	10.5 ± 9.4	0.7
Years of alcohol dependence		10.6 ± 9.4			14.6 ± 11.7	
Days since last drink before hospital admission		1.2 ± 0.6			0.3 ± 0.4	
Standard drinks consumed each day in the week before hospital admission		13.7 ± 8.3			19 ± 11.4	
AUDIT	5.9 ± 3.8	24 ± 6.5	4E-15	2.8 ± 2.3	28.0 ± 4.9	3E-14
GSI	0.15 ± 0.14	0.72 ± 0.45	6E-07	0.10 ± 0.09	0.11 ± 0.10	0.9
	c) Results after 3-week alcohol treatment in the discovery cohort					
	Patients (T1)		Patients (T2)		*P*-value (paired testing)	
GSI	0.72 ± 0.45		0.41 ± 0.52		0.036	
OCDS	19.3 ± 6.6		12.0 ± 4.9		3E-05	

Table 4.1 Description of a) the discovery study cohort and b) the replication study cohort. c) Results after 3-week alcohol treatment program in the discovery cohort.
Errors are given as standard deviation. Abbreviations: AUDIT, alcohol use disorder identification test; GSI, global severity index; OCDS, obsessive compulsive drinking scale.

4.3.2 Identification of AD-associated differential DNAm

Based on site-specific analyses of the T-cell DNAm array profiles, we identified 59 differentially methylated CpG sites between patients (T1) and controls with DNAm differences (Δ-beta) of at least 5% to increase the likelihood of biological relevance (FDR < 0.1). Of these 59 hits, 28 sites showed higher methylation, while 31 sites had lower methylation in patients compared to controls. Differences in DNAm ranged from 5 to 14% (Figure 4.1A). The top 10 hits, ranked by Benjamini-Hochberg (BH)-adjusted *P*-value significance, are listed in Table 4.2A. A complete list of all 59 significant hits (FDR < 0.1) is provided in Supplementary Table 4.1. The top-ranked hit (cg18752527) exhibited a DNAm difference of 6.6% and was located within the intragenic region of the *HECW2* gene.

In addition to single CpG sites, we identified 29 significant DMRs (FDR < 0.01, Δ-beta > 5%) using DMRCate. These DMRs contained 153 CpG sites, of which 8 were also identified as differentially methylated in the site-specific analysis between controls and patients (T1) (Supplementary Table 4.2). Interestingly, 4 of these overlapping 8 hits were in the *SRPK3* gene region (Figure 1B).

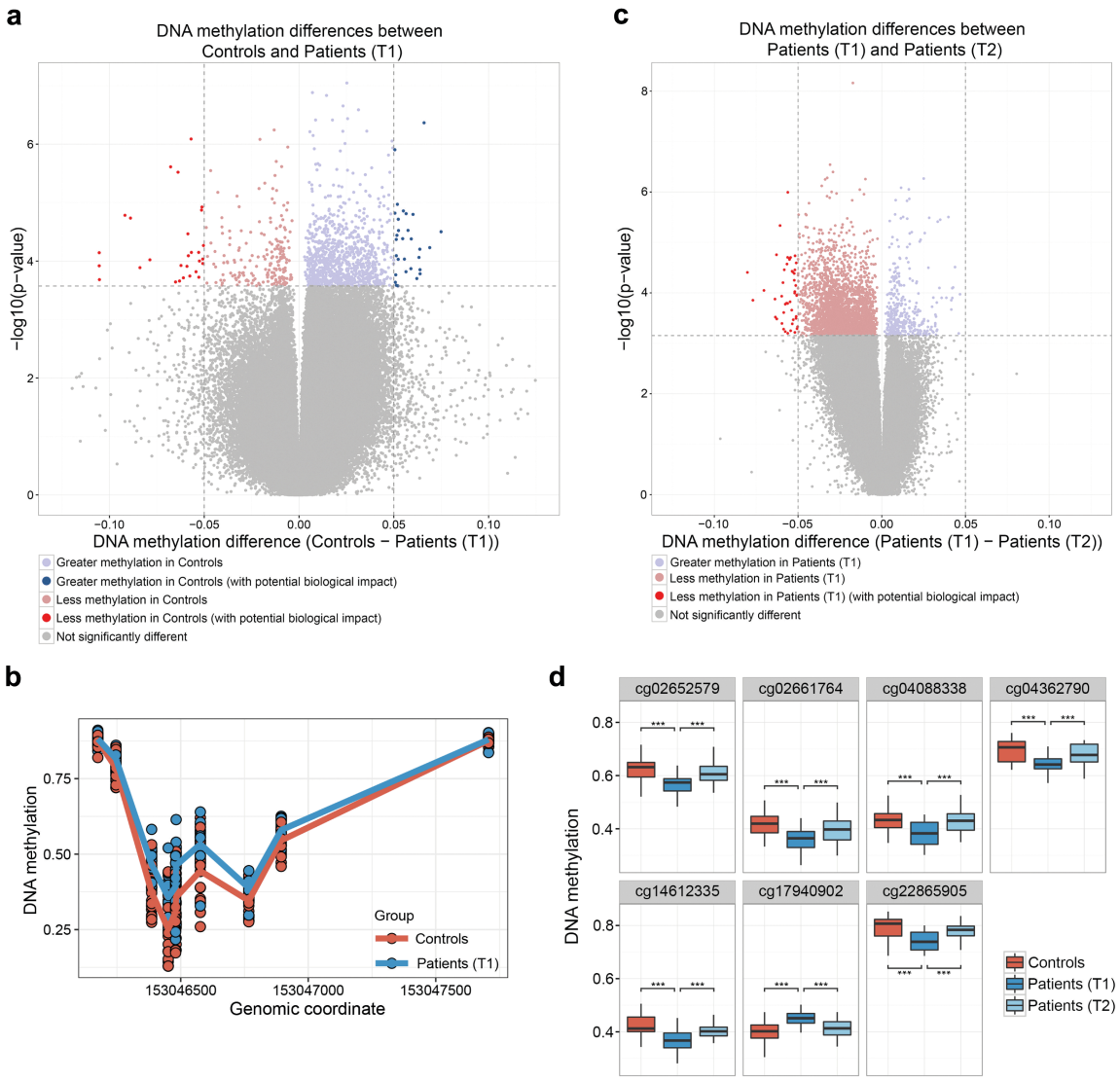


Figure 4.1 Differential sites and regions identified in the 450 K array analyses.

(a) Volcano plot depicting differences in DNAm levels between controls and patient (T1) for each probe in the corrected 450 K dataset (indicated on X axis) against FDR (indicated on Y axis, on –log10 scale). Dashed horizontal line denotes FDR threshold of 0.1 while dashed vertical lines denote DNAm difference thresholds of −0.05 and 0.05, respectively. (b) Differential DNAm detected by DMRcate in the promoter region of the SRPK3 gene (chrX:153, 046, 386–153, 046, 482). (c) Volcano plot depicting differences in DNAm levels between patients (T1) and patients (T2) as described in panel (a). (d) DNAm levels of seven sites which show reversion of DNAm post-treatment.***Indicate an FDR < 0.001.

a) Differentially methylated sites between Controls and Patients (T1)							
Probe ID	Gene	Region	Average beta Controls	Average beta Patients (T1)	Δ-beta	*P*-Value	BH-adjusted *P*-value
cg18752527*	*HECW2*	intragenic	0.342	0.276	0.066	4.30E-07	0.0213
cg08109624		intergenic	0.760	0.817	-0.057	8.15E-07	0.0234
cg10168086		intergenic	0.535	0.484	0.051	1.24E-06	0.0256
cg07280807*		intergenic	0.755	0.822	-0.068	2.44E-06	0.0366
cg12173150		intergenic	0.321	0.385	-0.064	3.02E-06	0.0370
cg01059398	*TNFSF10*	intragenic	0.261	0.209	0.052	1.07E-05	0.0627
cg17940902	*HLA-DMA*	promoter	0.399	0.450	-0.051	1.19E-05	0.0640
cg22778903	*MX2*	intragenic	0.304	0.355	-0.051	1.34E-05	0.0666
cg14612335	*SKIL*	promoter	0.423	0.368	0.055	1.38E-05	0.0666
cg11580026		intergenic	0.600	0.549	0.051	1.51E-05	0.0691
b) Differentially methylated sites between Patients (T1) and Patients (T2)							
Probe ID	Gene	Region	Average beta Patients (T1)	Average beta Patients (T2)	Δ-beta	*P*-Value	BH-adjusted *P*-value
cg15500907	*LAMA4*	intragenic	0.485	0.542	-0.056	1.01E-06	0.0323
cg05266321	*CCR2*	intragenic	0.545	0.606	-0.061	4.63E-06	0.0487
cg13279700	*C6orf10*	intragenic	0.481	0.544	-0.063	1.76E-05	0.0561
cg14054990	*KRTAP19-5*	promoter	0.431	0.482	-0.052	1.84E-05	0.0565
cg21049302		intergenic	0.466	0.522	-0.056	1.98E-05	0.0565
cg17022548	*NRG2*	intragenic	0.204	0.258	-0.054	1.99E-05	0.0565
cg22472360	*TRIO*	intragenic	0.514	0.569	-0.055	2.09E-05	0.0569
cg07920414	*RIMS3*	intragenic	0.438	0.493	-0.055	2.18E-05	0.0572
cg04088338		intergenic	0.378	0.429	-0.051	2.54E-05	0.0590
cg12240358	*HOMER2*	intragenic	0.462	0.519	-0.057	2.68E-05	0.0590

Table 4.2 Top 10 differentially methylated sites a) between controls and patients (T1) and b) between patients (T1) and patients (T2).
Probe IDs marked with an asterisk were validated by pyrosequencing. Abbreviations: Average beta, mean methylation values (%); Benjamini-Hochberg (BH) adjusted *P*-value.

4.3.3 Treatment-related alterations in T-cell DNAm profiles

To identify differentially methylated sites potentially playing an important role in alcohol withdrawal and early recovery in AD patients, we compared genome-wide T-cell DNAm profiles of patients before (T1) and after 3 weeks (T2) of participation in an alcohol treatment program. Using paired testing in our site-specific analyses, we identified 48 differentially methylated sites between patients (T1) and patients (T2), all of which showed increased methylation at T2 ranging from 5 to 12% difference (FDR < 0.1, Δ-beta > 5%) (Figure 4.1C, Supplementary Table

4.3). The top 10 hits are listed in Table 4.2B. Utilizing the same threshold as before, we did not observe any DMRs in patients before and after treatment.

4.3.4 Post-treatment reversion of differentially methylated sites

To examine whether AD-associated DNAm is influenced by a 3-week alcohol treatment program, we assessed DNAm levels in patients post-treatment at the 59 sites identified in the analysis comparing controls and patients (T1). After the treatment (T2), the DNAm levels of 7 out of 59 sites reverted back to a level where they no longer significantly differed from controls (Figure 4.1D). Based on paired testing, we determined that these 7 sites were indeed differentially methylated between patients (T1) and patients (T2). Moreover, 32 CpG sites showed a trend to revert back, though not significant at an FDR < 0.1. The DNAm levels of the remaining 20 sites did not change from T1 to T2.

4.3.5 Assessment of mean global DNAm differences between groups

Given the unidirectional change in our site-specific analysis of patients before and after treatment, particularly at AD-associated sites which showed post-treatment reversion, we next examined if this trend was related to AD-associated differences in mean global DNAm. Here we defined mean global DNAm as the calculated average of DNAm values across all sites in each sample. We found that although the result was only nominally statistically significant, prior to the alcohol treatment (T1), mean global DNAm was lower in patients compared to controls (P = 0.048, Mann-Whitney U test). However, at the end of treatment (T2), global DNAm of the patients approximated the levels seen in controls and no longer differed significantly from controls (Figure 4.2A). This finding was consistent with the unidirectional differences, in that all significant sites between patients before and after treatment showed increased methylation at T2 in the site-specific analysis, and supported the observed post-treatment reversion of AD-associated sites. Of note, these differences in mean global DNAm are unlikely to be driven by batch effects or other sources of technical variation due to the fact that all samples were run in a randomized manner on the same set of arrays.

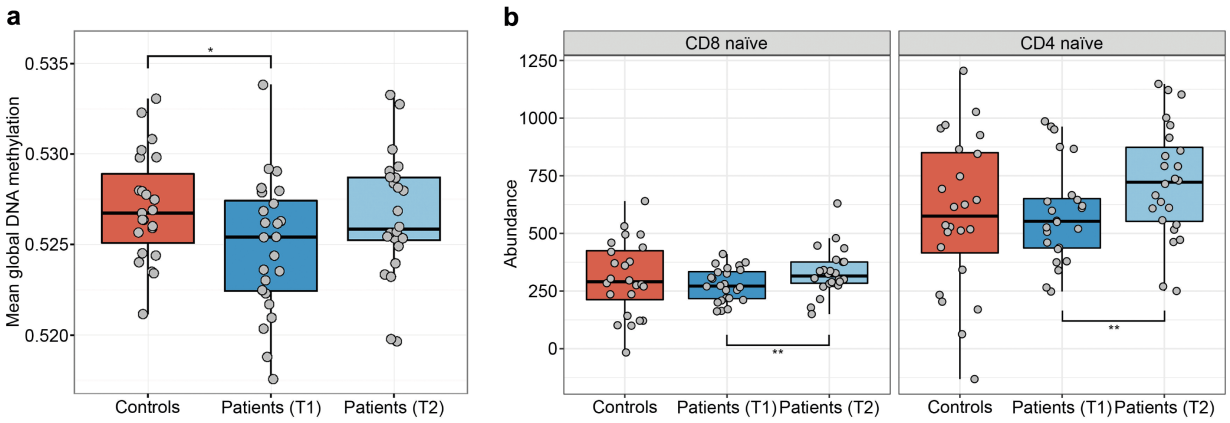


Figure 4.2 Mean global DNAm differences and naïve T-cell subtype differences between groups.

(a) Patients (T1) showed significantly decreased mean global DNAm levels compared to controls (P = 0.048, Mann-Whitney U test). Differences between controls vs. patients (T2) and patients (T1) vs. patients (T2) were not significant. (b) Abundance levels of naïve CD8+ and CD4+ T-cells were predicted using an advanced blood DNA methylation age prediction tool. Both naïve T-cell subtypes significantly increased post-treatment in patients (**Indicates an FDR < 0.01, Wilcoxon signed-rank test) but were not significantly different between controls and patients at either time point.

4.3.6 Differences in naïve T-cell subtype abundances between groups

To evaluate if there were differences in underlying T-cell subtypes between the groups, we estimated abundance measures of additional blood cell subsets using an advanced blood analysis option for an epigenetic clock prediction tool on our T-cell 450 K profiles (322). We observed that the predicted abundance levels of both $CD4^+$ and $CD8^+$ naïve T-cell subsets significantly increased post-treatment in AD patients (FDR < 0.01, Wilcoxon signed rank test) (Figure 4.2B). However, the abundance of these naïve T-cell subtypes was not statistically significantly different between controls and patients at either time point.

4.3.7 Validation of AD-associated differential DNAm by pyrosequencing

To verify the results from the 450 K dataset, we selected two top-ranking differentially methylated sites between controls and patients (T1) (cg18752527 in the *HECW2* gene and cg07280807 in an intergenic region) for validation using pyrosequencing as an independent readout of DNAm measures. We additionally validated two promoter CpGs of *SRPK3* (cg16529483 and cg24496423) since differential methylation in the *SRPK3* gene region was found to be a robust finding in our DMRcate analyses. We were able to confirm significant differences between controls and patients (T1) at all 4 sites, as shown in Figure 4.3A (Student's

t-test, FDR < 0.01). Although Bland-Altman plots showed a general bias for lower methylation levels measured by pyrosequencing (Supplementary Figure 4.3), the correlation in measurements between the two methodologies was highly concordant for all 4 sites (Spearman's correlation r_s > 0.7, FDR < 0.001) (Supplementary Figure 4.3).

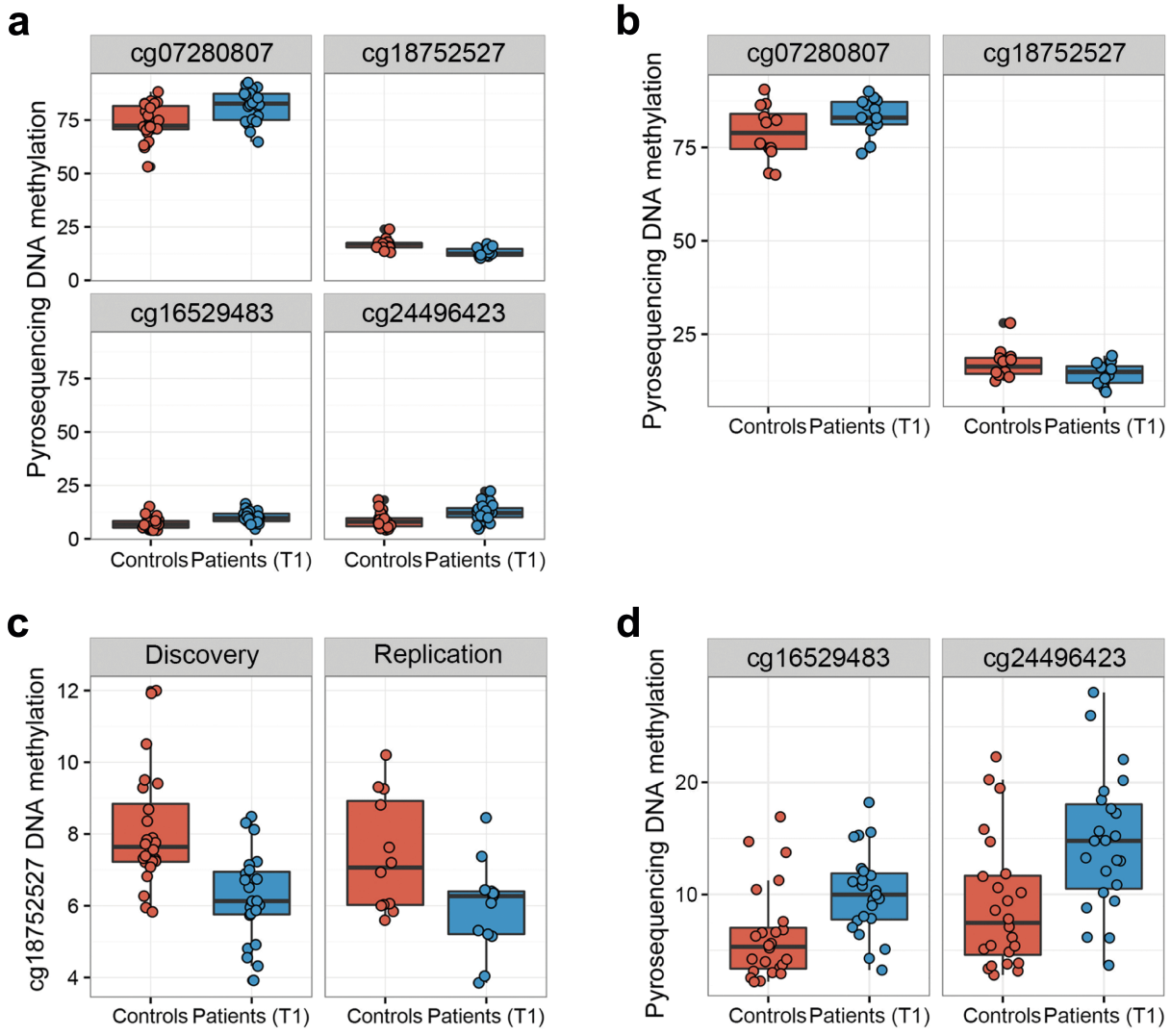


Figure 4.3 Validation and replication of top-ranking hits by pyrosequencing.

(a) Boxplots showing differences in DNAm levels of discovery cohort T-cell samples as measured by pyrosequencing (FDR < 0.01, Student's t-test). (b) Two top-ranked hits (cg07280807 and cg18752527) were verified as being differentially methylated in T-cell samples of the replication cohort (FDR < 0.05, one-sided t-test). (c) Verification of differential methylation of cg18752527 (HECW2) in the discovery (left) and the replication cohort (right) in DNA derived from whole blood (FDR < 0.05, two-sided t-test). (d) Verification of cg16529483 and cg24496423 (SRPK3) differential methylation in the discovery cohort in DNA derived from whole blood (FDR < 0.01, two-sided t-test).

4.3.8 Replication of AD-associated differential DNAm in an independent cohort

To further test the robustness of our EWAS findings, we analysed the previously mentioned 4 sites in T-cells of an independent replication cohort by pyrosequencing. The two top-ranking hits, cg07280807 in an intergenic region and cg18752527 in *HECW2*, were differentially methylated in the replication cohort (FDR < 0.05, one-sided t-test) (Figure 4.3B). However, the two sites within the *SRPK3* promoter region (cg16529483 and cg24496423) did not replicate in this cohort, likely due to insufficient power with the low sample size in this cohort, but showed a trend in the same direction as in the discovery cohort.

4.3.9 Analysis of differential DNAm in whole blood

To identify sites that are not only differentially methylated in T-cells, but also in whole blood DNA, we sought to reproduce our most robust EWAS findings from T-cells in whole blood DNA samples of both our discovery and replication cohorts. Therefore, we analysed DNAm of the 4 previously mentioned sites in whole blood samples by pyrosequencing. We observed differential methylation of cg18752527 in the intragenic region of *HECW2* between controls and patients (T1) in both cohorts (FDR < 0.05, Student's t-test) (Figure 4.3C). Furthermore, similar to the findings from T-cells, the two sites within the *SRPK3* promoter region (cg16529483 and cg24496423) were differentially methylated in whole blood samples of the discovery cohort (Figure 4.3D), but not of the replication cohort. We found that differential DNAm of cg07280807 did not replicate in whole blood of either cohort. Using a previous 450 K dataset of purified blood cell types, we confirmed that the DNAm status of cg18752527 in *HECW2* was highly associated with T-cells, along with NK cells, suggesting that the DNAm differences we measured in whole blood were driven, in part, by T-cells (P = 7.6E-15, ANOVA) (Supplementary Figure 4.4) (458). The DNAm statuses of the two sites in the *SRPK3* promoter were not associated with any specific cell type (Supplementary Figure 4.4).

4.4 Discussion

By analysing genome-wide DNAm profiles of purified $CD3^+$ T-cells using the Illumina 450 K array, we found 59 CpG sites to be differentially methylated in a group of 24 alcohol dependent patients compared to 23 closely matched healthy controls. These site-specific hits showed considerable overlap to detected DMRs, suggesting that the results were not contingent

on the analytical approach used. Furthermore, we discovered 48 sites that were differentially methylated between AD patients at the time of hospital admission (T1) and after 3 weeks (T2) of participation in an alcohol treatment program and showed a reversion of some of the AD-associated sites post-treatment. In addition, we were able to validate four of the top-ranking AD-associated hits by pyrosequencing, and replicate two of them in an independent cohort. Finally, we found the top-ranked hits in *HECW2* (cg18752527) and *SRPK3* (cg16529483 and cg24496423) to be differentially methylated in whole blood, signifying the potential relevance of these associations in other blood cell types. To our knowledge, this is the first study to assess and replicate alcohol-associated differential DNAm in purified T-cells and to assess DNAm variation that may be related to early recovery from AD in closely matched human population cohorts.

EWAS pose an excellent hypothesis-free opportunity to identify as yet undiscovered disease-associated genes. Our EWAS findings of AD-associated differential DNAm revealed both site-specific and regional differences between patients before treatment and matched controls in a clinically relevant cell type. The observed bi-directional patterns of changes are consistent with previous evidence of AD-associated differential DNAm (432, 434, 436, 438, 442) However, our findings derived from T-cells did not overlap with previously reported associations of AD with DNAm (432, 434, 436, 438, 442). This can at least in part be explained by the use of heterogeneous biological material (i.e. whole blood, PBMCs), differences in the cohorts used or in the strategies applied to match patients and controls as well as by varying methodologies for DNAm measurement, with reduced or discordant coverage of CpG sites in previous studies compared to the present study (432, 434, 436, 442). However, our top-ranking hits in *HECW2* and *SRPK3* might contribute to reveal mechanisms that may play a role in AD. HECW2 is a HECT-type E3 ubiquitin ligase involved in the cellular stress response (459, 460). This finding is in line with previous evidence for the role of epigenetic regulation of cellular stress response genes in AD, such as *GDAP1*, which was identified in a previous EWAS and subsequently replicated in whole blood samples derived from an independent cohort (432, 441). However, *GDAP1* did not come up in this present analysis using DNA isolated from purified T-cells. Presumably, the previously described differential methylation of *GDAP1* in whole blood is driven by another cell type other than T-cells. *SRPK3* encodes a serine/arginine protein kinase and is essential for the development of the skeletal muscle (461). It was shown that the drosophila homolog *SRPK79D* plays an important role in the function of synapses (462).

Although an association between *SRPK3* and the nervous system in humans has not been described so far, the high homology between SRPK79D and SRPK3 (65%) makes an as yet uncharacterized role in the nervous system possible.

In addition to the assessment of AD-associated differential DNAm in T-cells prior to alcohol treatment, we also examined treatment-related site-specific alterations in DNAm by comparing DNAm profiles in T-cells of patients before (T1) and after a 3-week alcohol treatment (T2). Our findings include numerous sites in which DNAm in patients (T2) reverts back to levels comparable to those observed in controls. More specifically, we showed post-treatment DNAm reversion (at 7 sites) or partial reversion (at 32 sites) back to control levels. These findings confirm the results of a previous pilot study, which also showed reversion of DNAm after a short term alcohol treatment program (432). Other epigenetic studies in human populations investigating the effect of short-term treatments, including exercise or dietary interventions, on DNAm of relevant tissues have identified similar numbers of site-specific DNAm changes with a comparable magnitude of effect sizes to our findings (463, 464).

Based on our assessment of mean global DNAm, measured as averaged methylation across all interrogated CpGs, we found that global DNAm levels were significantly lower in patients prior to the alcohol treatment compared to controls. Following alcohol treatment, the mean global DNAm of patients no longer differed significantly from controls. These results are in accordance with the unidirectionality of our treatment-related hits, with all significant sites exhibiting increased DNAm after treatment, and with our site-specific findings that numerous AD-associated CpGs exhibited post-treatment reversion to levels comparable to controls. The reduction in mean global DNAm observed in AD patients is supported by previous studies, which also demonstrated decreased methylation (437, 438). It has been hypothesized that such alcohol-associated decreases in global DNAm are attributed to the lack of methionine adenosyl transferase regulation in AD patients (427, 465). However, in contrast, earlier studies have postulated that due to the higher homocysteine levels in AD patients, global DNAm patterns should be elevated, although such associations have not been confirmed (466, 467). The lack of consensus in regard to alterations in alcohol-related global DNAm measures highlights the need for further investigation into the biological mechanisms underlying global DNAm patterns in AD patients.

Using bioinformatic predictions from our T-cell DNAm profiles, we observed a significant increase in naive $CD4^+$ and $CD8^+$ T-cell subsets post-treatment, which is consistent with evidence of decreased frequencies of these naïve T-cell subtypes due to chronic AD and a resultant restoration of peripheral T-cell numbers following short-term alcohol abstinence (445–447). These findings, along with known effects of alcohol dependence on T-cell homeostasis, proliferation and activation, highlight the importance of understanding alcohol-related effects on T-cell-specific biology, particularly in the context of AD pathophysiology and treatment, of which our study serves as the first to profile such AD-associated changes on the T-cell epigenome (447, 468).

In order to verify that our results are robust and largely reflective of potential biological variation as opposed to technical variation, we took a number of precautions in our analyses, including I) constraining our hits to sites with DNAm differences greater than 5% between groups in order to increase the likelihood of biological relevance, II) confirming 450 K measures by pyrosequencing and III) validating top-ranked hits by pyrosequencing in an independent replication cohort. Although we observed a general bias between the two methodologies, in which the pyrosequencing measures were lower than 450 K values, there was high concordance of measures between the two methods and we were still able to detect significant differences in DNAm between groups, signifying the strength of our results. Moreover, we were able to confirm three top-ranking hits from purified T-cells in whole blood, further strengthening the robustness of our findings and highlighting their potential importance in AD.

It is important to note that our study had a few inherent limitations. Firstly, using bioinformatic cell type predictions, we detected notable levels (up to 32%) of cellular contamination in our bead-purified T-cell samples. This is consistent with previous work which confirmed the presence of cellular heterogeneity in samples even after purification using cell surface markers (469). We removed cell heterogeneity using a regression-based method, thereby ensuring inter-individual differences in cell composition were normalized in our dataset prior to DNAm analyses. Secondly, our analyses were limited by a rather small sample size. To work around this limitation, we utilized a relaxed FDR threshold in the differential methylation modelling to capture more potentially biologically relevant sites and focused on validating and replicating our top-ranked hits to ensure these results were robust. Although we were able to validate the hits within the *SRPK3* promoter by pyrosequencing in T-cell and whole blood

samples of the discovery cohort, we could not replicate the differential DNAm of *SRPK3* in our second cohort, unlike our findings for *HECW2*. This probably results from insufficient statistical power due to the low sample size of the replication cohort. We acknowledge that the small samples size analysed in our study could also hinder successful validation of our results in future studies. The phenomenon of non-replication could also be observed in previous transcriptome-wide studies in human populations of AD patients and control individuals, where the overlap between the individual studies was fairly small (470, 471). However, by technically validating and replicating our results in a second cohort, we made an attempt to reduce the risk of false-positive findings to a minimum. Despite these efforts, our results should be verified in a larger cohort spanning different populations to confirm the associations for *HECW2* and *SRPK3*. So far, neither *HECW2* nor *SRPK3* were among top-ranked hits in transcriptome-wide studies. Therefore, functional data is required to investigate the interplay of DNAm, transcription and functioning of these genes related to AD. Thirdly, we cannot rule out that the DNAm differences between the patients before (T1) and after treatment (T2) may be due to stochastic temporal DNAm variation, although previous work in blood has revealed minimal evidence of temporal variation in the majority of 450 K probes across a 9 month period (472). In addition, differences in DNAm could also be due to direct influences of acute ethanol intoxication, which has been shown to have an effect on transcriptome regulation (470, 471). We tried to circumvent this limitation by only including subjects who had their last drink in a narrow time frame of 1.2 ± 0.6 days. Additionally, the 20 CpG sites which did not change from pre- to post-treatment could potentially be differentially methylated due to chronic alcohol exposure and not due to early withdrawal. To clarify this issue, future longitudinal studies are warranted. Finally, we cannot disregard the potential influence of genetic variation on our differentially methylated CpG sites. However, we attempted to reduce genetic heterogeneity in our cohort by using only Caucasian participants.

In conclusion, we report that AD is associated with lower mean global DNAm and with differential DNAm of specific sites in $CD3^+$ T-cells. Additionally, we were able to identify changes in DNAm related to alcohol treatment in patients. These changes include the reversion of AD-associated DNAm alterations at certain sites to levels comparable to controls. Validation of our top-ranking associations by pyrosequencing and replication of our top-ranked hits in a second independent cohort strongly supports the robustness of our results. Finally, we show that

the differential methylation of *HECW2* and *SRPK3* is not only present in T-cells, but also in whole blood, indicating that *HECW2* and *SRPK3* are likely robust findings which should be followed up in future studies.

Chapter 5: Integration of DNA methylation patterns and genetic variation in human pediatric tissues help inform EWAS design and interpretation

5.1 Background and rationale

Epigenome-wide association studies (EWASs) are becoming increasingly popular, due to their potential to enhance our understanding of the determinants of health and disease, including potential early life embedding of experiences and exposures on later life outcomes (10, 16, 20–23, 25). DNA methylation (DNAm), which involves the covalent attachment of a methyl group to a cytosine primarily at CpG dinucleotides, is the most well studied epigenetic mark in human populations due to its relative stability and ease of measurement on quantitative array-based methods (26, 27). To date, EWASs have identified differential DNAm across a broad range of contexts including disease states, genetic background and environmental exposures, thereby providing evidence for the potential contribution of DNAm to mediating gene-by-environment (GxE) interactions (6, 25, 473).

Given that tissue specificity is an integral feature of epigenetic biology, as different tissues and cell types acquire distinct epigenomes across differentiation, the selection of tissue source is a key consideration in the careful design and interpretation of EWAS analyses (56, 474, 475). The collection of a disease-relevant, target tissue allows for the direct assessment of epigenetic associations which may be implicated in the underlying phenotypic or disease biology. In certain cases, readily accessible peripheral tissues may represent the target tissue; for example, use of PBMCs for the investigation of DNAm associations to immune or inflammatory phenotypes (20, 315, 317, 476). However, in many cases, the target tissue, such as brain, muscle, adipose tissue, among others, may be impossible or extremely difficult to collect from living individuals or at sufficient quality for analysis from postmortem samples (21). Easily accessible peripheral tissues are therefore often used in human epigenetic studies for biomarker discovery in lieu of target tissues that are difficult to collect. This is particularly pertinent to pediatric cohorts in which biopsy specimens with invasive collection procedures or postmortem samples are less common than in adult populations. As such, more readily accessible tissues with minimally invasive collection procedures, such as cord blood, saliva, buccal epithelium cells (BECs) or peripheral blood mononuclear cells (PBMCs), are widely used tissue source materials for early life EWASs. The use of pediatric tissues in DNAm analyses is further complicated by

the fact that widespread alterations occur in tissue-specific DNAm patterns during development, thereby conferring additional complexity in the selection of appropriate source material for early life DNAm studies.

Currently, a major focus in human epigenetic research is to elucidate the tissue specificity of DNAm patterns with respect to individual CpGs as well as inter-individual variation within a single tissue (52, 53, 59, 477). At a population level, a number of studies have examined the concordance of DNAm patterns across multiple tissues (57, 59, 68, 292, 295, 317, 396). Findings have shown that beyond tissue-specific differences in absolute DNAm measures, inter-individual DNAm variability also varies by tissue type. For example, previous work by our group has shown that BECs have greater DNAm variability over matched PBMCs at both the genome-wide level and at individual CpGs (317). Moreover, CpG sites with higher DNAm variability tend to be more correlated between matched tissues (57, 59, 68, 292). Although these results provide important insights into the comparability of DNAm measures across matched tissues, the analyses to date have been conducted in adult tissues, thereby limiting their relevance to DNAm profiles from pediatric samples. As previous studies have demonstrated that developmental changes in blood DNAm patterns tend to be more pronounced and occur more rapidly in childhood, the examination DNAm concordance and variability in pediatric tissues represents a fundamental step in our understanding of EWAS associations from pediatric peripheral tissues (52, 478).

Genetic variation represents an additional contributor to DNAm patterns in tissues, with genetic influences accounting for nearly 20-80% of DNAm variance within a tissue (84–88, 479). Methylation quantitative trait loci (mQTL), sites at which DNAm is associated with genetic variation, are present across the genome and are often consistent across tissues, ancestral populations and developmental stage (43, 74–76). Notably, genetically influenced sites of inter-individual DNAm variation, which can co-occur across tissues, may be biologically informative. For example, allele-specific DNAm of the FK605 binding protein 5 (*FKBP5*) gene, which has been associated with risk of developing stress-related psychiatric disorders, responds to glucocorticoid stimulation in a similar way in peripheral blood cells and neuronal progenitor cells (73). Moreover, previous work demonstrated that inter-individual DNAm within a single tissue is largely attributed to GxE effects (89). As such, delineating the contribution of genetic influences to tissue-specific DNAm may help clarify the interpretation of EWAS associations.

Given that early life development brings about sizable changes to DNAm patterns, it is necessary to specifically address questions concerning DNAm variability and concordance between peripheral tissues, as well as genetic influences on early life DNAm patterns, in childhood (52, 478) . To this end, we used matched samples of two commonly used peripheral tissues in EWAS, PBMCs and BECs, from two independent early life cohorts of different ages, in order to a) understand differences in inter-individual variability and concordance of DNAm between these tissues in pediatric samples and b) determine genetic contributions to these patterns at the site-specific level. Our results showed that genome-wide DNAm variability is tissue-specific with BECs exhibiting greater inter-individual DNAm variability over PBMCs. Moreover, we found that highly variable CpGs were more likely to be positively correlated between matched tissues and enriched for DNAm sites under genetic influence. Collectively, these findings highlight a number of potential insights and considerations for the appropriate design and interpretation of EWAS analyses performed in commonly used peripheral tissues of pediatric samples.

5.2 Materials and Methods

5.2.1 Study cohorts and tissue samples

Matched tissues were obtained from a subset of two separate pediatric cohorts. Specifically, a subset of samples from the previously described C3ARE (Cleaning, Carrying, Changing, Attending, Reading and Expressing) cohort were collected from 16 individuals (8 females; 50%) aged 3-5 years (age range: 3.6-4.2 years (BEC) and 4.5-5.2 years (PBMC)) from Vancouver, British Columbia (480). The GECKO cohort samples (Gene Expression Collaborative Kids Only) comprised of 79 individuals (36 females; 46%) aged 6-13 years (age range: 6-11 years (BEC) and 7-13 years (PBMC)) also from Vancouver, British Columbia. Birth dates were not available for all GECKO participants; age in years was recorded at the BEC sample collections. In both cohorts, the majority of BEC samples were collected at the first visit and PBMCs were collected at a later date. In the C3ARE cohort, follow-up visits ranged from 7 days to 1.5 years, with three pairs of matched BECs and PBMCs being collected on the same day. In the GECKO cohort, the follow-up visits at which peripheral blood was collected ranged from 6 months to 2.3 years after the initial visit. Demographic descriptors of both cohorts are provided in Table 5.1. All experimental procedures were conducted in accordance to institutional

review board policies approved by the joint University of British Columbia and Children and Women's Hospital Ethics board (Certificates: H07-01317 and H07-02773). Written informed consent was obtained from a parent or legal guardian and assent was obtained from each child before study participation. For both cohorts, BECs were collected using the Isohelix Buccal Swabs (Cell Projects Ltd., Kent, UK) and stabilized with Isohelix Dri-Capsules for storage at room temperature prior to DNA extraction, as previously described (147). Whole blood was collected into Vacutainer® CPT™ Cell Preparation Tubes (Becton, Dickinson and Company, NJ, USA) and PBMCs were isolated following centrifugation, washing and resuspension into R10 media (Sigma-Aldrich, MO, USA), as previously described (481). PBMC pellets were frozen and stored at -80°C until DNA extraction.

5.2.2 DNA isolation and DNAm arrays

Genomic DNA from stabilized buccal samples was isolated using Isohelix Buccal DNA Isolation Kits (Cell Projects Ltd., Kent, UK) and was purified and concentrated using DNA Clean & Concentrator (Zymo Research, CA, USA). Genomic DNA was extracted from PBMC pellets using the DNeasy kit (Qiagen, MD, USA). DNA yield and purity was assessed using a Nanodrop ND-1000 (Thermo Fisher Scientific, MA, USA). Bisulfite conversion of DNA (750 ng) was performed using the Zymo Research EZ DNA Methylation Kit (Zymo Research, CA, USA). Samples were subsequently randomized and 160 ng of bisulfite-converted DNA was applied to the Illumina Infinium HumanMethylation450K Beadchip (450K) array, as per manufacturer's protocols (Illumina, CA, USA) (129).

5.2.3 DNAm array data quality control and normalization

Data from each cohort were analyzed separately. Specifically, raw intensity values from the DNAm arrays were imported into Illumina GenomeStudio V2011.1 software and subjected to initial quality control checks for array staining, extension and bisulfite conversion followed by color correction and background adjustment using control probes contained on the 450K array. Data were exported from GenomeStudio as beta values which represent the estimated DNAm level based on a ratio of intensities between methylated and unmethylated alleles, with beta values ranging from 0 (unmethylated) to 1 (fully methylated). Subsequent processing and analysis were performed in R Version 3.2.1 (http://www.r-project.org). Profiles from 65 probes

targeting single nucleotide polymorphisms (SNPs) were used to ensure matched tissue samples originated from the same individual. The 65 SNP probes were subsequently filtered out of the dataset. Since the cohorts were not equally matched for sex, we removed sex chromosome probes (11,648) from both datasets. Additional probe filtering was performed in which poor performing probes including those with detection p-values greater than 0.01 or probes with missing beta values in more than 2% of samples were removed (14,400 C3ARE, 13,374 GECKO). Re-annotation of the Illumina 450K array was used to filter probes that are known to be polymorphic at the target CpG (324). Probes which have non-specific *in silico* binding to the sex chromosomes were also removed (324). Final probe count after quality control probe filtering was 429,494 probes for C3ARE and 430,581 probes for GECKO. Following quality control processing, quantro determined quantile normalization to be inappropriate as the global DNAm distributions between the two distinct tissues were highly differential (382). Beta Mixture Quantile dilation (BMIQ) normalization was performed to remove differences between Type I and Type II probes on the 450K array, yielding normalized DNAm (383).

5.2.4 Cell-type correction of DNAm data

The effects of cellular heterogeneity on DNAm measures were removed from PBMC and BEC samples in both cohorts. Specifically, blood cell type proportions were estimated for the PBMC samples using an established blood deconvolution method (63, 455). Given that no cell deconvolution algorithm for buccal tissues exists and that buccal swabs, like saliva, are predominantly composed of BECs and leukocytes, we used a saliva-based deconvolution method which was designed to predict these cell types from underlying DNAm patterns (68, 482). Predicted cell proportions from both PBMC and BEC tissues were used to normalize cellular heterogeneity within each tissue using a regression-based strategy (456)(Supplementary Figure 5.1). Principal component analysis (PCA) was subsequently used to confirm that the correlation of estimated cell-type proportions to DNAm variance within a tissue were minimal in the corrected 450K datasets (data not shown).

5.2.5 Assessment of cross-tissue correlation, tissue-specific variability and tissue-specific differences in DNAm data

Prior to subsequent DNAm analyses, the corrected 450K datasets were filtered down to overlapping probes (419,507) between the GECKO and C3ARE cohorts. Probewise cross-tissue Spearman's correlations were calculated on beta values between the matched PBMC and BEC tissues. Inter-individual variability of each CpG was calculated as the range between the 10th and 90th percentile beta values for each CpG, referred to as "reference range" (372). This method captures variability across the bulk of samples while being largely robust to outlier samples.

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

Paired Wilcoxon signed-rank tests were used to compare global differences in reference range between matched BEC and PBMC samples. Fligner-Killeen tests were used to compare probewise variability differences in each of the cohorts. Using previously published methods, we aimed to identify informative sites between BECs and PBMCs, which we defined as CpGs that are both variable across individuals and highly correlated between both tissues (292). To identify informative sites, we first subset each cohort down to CpGs with a reference range greater than 0.10 in both tissues. We subsequently ran a beta mixture model on Spearman correlation rho values generating two Gaussian distributions, which separated out a group of highly concordant CpGs (Supplementary Figure 5.2). The Spearman rho distributions in this set of highly correlated CpGs was used to define a threshold correlation coefficient, the cutoff being two standard deviations lower than the mean of the distribution. In the GECKO cohort rho > 0.47 was determined as the threshold and in the C3ARE cohort, rho > 0.32 was determined as the threshold. We also set a minimum reference range of 0.05 in both tissues to exclude CpGs with little inter-individual variation.

Finally, we identified CpGs which were differentially methylated between tissues by running Wilcoxon signed-rank tests across all probes in the C3ARE, GECKO and GECKOsub

datasets. For all tests, the resulting p-values were adjusted using the Benjamini-Hochberg (BH) false discovery rate (FDR) method (332). CpGs which passed an FDR < 0.05 and an effect size threshold, delta beta > 0.05, independently in all three datasets, C3ARE, GECKO and GECKOsub, were classified as "differential sites".

5.2.6 SNP genotyping arrays

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

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

Quality control pre-preprocessing of Illumina Infinium PsychChip data was performed separately for each cohort according to recommended guidelines (483). Specifically, SNPs with a low 10th percentile GenCall score or with a low average GenCall score were filtered out. Additionally, SNP probes located on mitochondrial DNA, on sex chromosomes or without chromosome labels were removed. After probe filtering, final SNP probe counts for the C3ARE and GECKO datasets were 550,200 and 547,662, respectively. To test for difference in genetic ancestry between the two cohorts, we ran all samples in PCA, using the 542,699 SNPs called for every individual in both processed datasets. Genetic ancestry was not found to differ significantly between the cohorts (Supplementary Figure 5.3), as determined by Wilcoxon ranked sum test of GECKO vs C3ARE in PC1 scores ($p = 0.8$) and PC2 scores ($p = 0.4$). Therefore, genetic ancestry was not considered in further analyses.

5.2.8 Cis-mQTL analyses

We ran *cis*-mQTL analyses in each cohort separately, using GECKO as the discovery cohort and C3ARE as the validation cohort. In the GECKO cohort, PsychChip data were filtered after quality control to remove any SNP probes containing missing values in 5% of all samples, leaving 560,770 SNPs. In addition, SNPs with a minor allele frequency less than 5% or not in Hardy-Weinberg equilibrium were removed. Remaining SNPs (249,835) were then numerically coded, as 1, 2, or 3, for correlational analyses. Therefore, all SNPs used in mQTL analyses were directly measured on array, rather than generated through imputation. CpGs with a reference range of less than 5% were removed from mQTL analysis; this was performed separately in each tissue, leaving 131,706 CpGs in PBMCs and 210,784 CpGs in BECs. Finally, SNP-CpG pairs less than 5 kb apart were tested as mQTL using Spearman correlations. We selected a 5kb window as previous mQTL analyses using whole genome bisulfite sequencing data reported that associations between SNP–CpG pairs are more likely to be causal within a 5 kb window (43, 89, 484). Pairs with FDR ≤ 0.05 and DNAm change per allele $\geq 2.5\%$ were designated as *cis*-mQTL hits and followed up for validation in the C3ARE cohort. For validation testing in the C3ARE samples, SNP-CpG pairs were further filtered to exclude those with SNPs that were a) not present in the filtered C3ARE PsychChip data or b) monomorphic or had less than 2 heterozygotes in the C3ARE samples. The mQTL analyses were repeated in the C3ARE data. SNP-CpG pairs with FDR ≤ 0.05 and DNAm change per allele $\geq 2.5\%$ were designated as validated *cis*-mQTL hits and followed up in subsequent analyses.

5.2.9 Representation of identified sites in published EWAS findings

In order to relate our results to published EWAS findings performed in pediatric cohorts, we selected five published studies which used the 450K array to measure DNAm profiles in pediatric BECs or peripheral blood. Specifically, these studies examined DNAm variation associated with puberty, aging in early life, childhood psychotic symptoms, fetal alcohol spectrum disorder and autism spectrum disorder (147, 148, 294, 485, 486). For each study, we downloaded the list of probes reported as significant and matched these probes to sites, which we identified as: 1) informative sites, 2) differential sites and/or 3) *cis*-mQTL-associated CpGs. For one study, in which differentially methylated regions (DMRs) were reported, we downloaded the dataset (Accession # GSE50759) and extracted individual probes underlying the DMRs (148).

5.3 Results

5.3.1 Study cohorts and DNAm data processing

To explore the tissue-specific DNAm patterns of pediatric PBMCs and BECs, we used subsets from two independent human cohorts, GECKO and C3ARE, both of which contained matched samples from healthy children from the Lower Mainland Vancouver area. In GECKO, individuals ranged in age from 6 to 11 years at time of BEC collection (median = 8.8) and 7 to 13 years at time of PBMC collection (median = 10.3). Of the GECKO study samples (n = 79), 46% were female (n = 36). In C3ARE (n = 16), individuals ranged in age from 3 to 5 years at time of BEC collection (median = 4.5) and 4 to 5 years at time of PBMC collection (median = 5.1) and 50% were female (n = 8) (Table 5.1). DNAm data, as measured across ~485,000 CpGs by the 450K array, were filtered down to overlapping 419,507 sites which passed independent quality control measures in both cohorts. Each 450K dataset was normalized to remove probe type differences and adjusted for cell-type heterogeneity in each tissue using established bioinformatic correction methods (63, 68, 383, 455, 456). We used these corrected 450K data of matched PBMC and BEC samples from both cohorts to assess inter-individual DNAm variability, DNAm concordance across tissues and genetic influence on DNAm, in order to gain insight into DNAm variation in these commonly used pediatric peripheral tissues.

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

Table 5.1 Sample characteristics for C3ARE and GECKO cohorts

5.3.2 BECs had significantly greater inter-individual DNAm variability than PBMC

As inter-individual DNAm variability within a tissue likely relates to the potential effect sizes that are detectable in EWAS analyses, we were interested in assessing tissue-specific DNAm variability. To this end, we first interrogated the global differences in inter-individual DNAm variability between PBMC and BEC samples, following *in silico* correction for cell type differences in each tissue. We used reference range as a measure of DNAm variability as

opposed to absolute range in order to minimize potential skewing by outlier values and non-normal DNAm values at individual CpGs, as previously described (292, 487). Within each cohort, BEC DNAm had a significantly greater reference range than PBMC DNAm (Figure 5.1A; Wilcoxon signed-rank test, all p-values = 2.2×10^{-16}). In GECKO, the median reference range, measured in beta values, was 1.9% higher in BECs (5.2%) than in PBMCs (3.3%). Similarly, in C3ARE, the median reference range was 1.6% higher BECs (3.6%) than in PBMCs (2.0%). The difference in reference range was not dependent on sample size, as demonstrated by the consistency between GECKO and GECKOsub, the GECKO cohort randomly subsampled to the sample size of C3ARE (n = 16) 100 times. These differences in median reference range were modest but consistent in both cohorts. In addition, tissue-specific differences in DNAm variability were observed at individual CpGs, as determined by a Fligner-Killeen test of each site. In GECKO, 217,091 probes had significantly greater variability in BEC at FDR ≤ 0.05, while only 32,350 probes were more variable in PBMC. Similarly, in the C3ARE cohort, 127,472 probes had greater variability in BECs (FDR ≤ 0.05) and 8,183 probes in PBMCs (FDR ≤ 0.05; Figure 5.1B-C). Collectively, 85% of C3ARE probes (108,498) with greater variability in BEC were also found in the GECKO cohort to have greater BEC variability. These 108,498 CpGs were enriched for those with high inter-individual BEC variability in both cohorts (10,000 permutations, p-value $< 1 \times 10^{-4}$). As well, 84% of C3ARE probes (6,840) with greater variability in PBMCs, were also more variable in PBMCs in the GECKO cohort; similarly, this subset was enriched for CpGs with high PBMC variability in both cohorts (10,000 permutations, p-value $< 1 \times 10^{-4}$). These findings suggested that BEC DNAm was consistently more variable than PBMC DNAm across both cohorts.

Apart from tissue-specific differences in reference range, we also observed a cohort-specific difference in DNAm variability. Specifically, CpGs in GECKO had a significantly greater reference range than C3ARE CpGs in both tissues (Wilcoxon rank sum test, p-value = 2.2×10^{-16}). In BECs, the median reference range was 1.6% higher in GECKO than C3ARE and in PBMCs, it was greater by 1.3%. This difference remained significant when GECKOsub was used in lieu of GECKO (Wilcoxon rank sum test, $p = 2.2 \times 10^{-16}$), suggesting that these cohort-specific DNAm variability differences occurred irrespective of sample size.

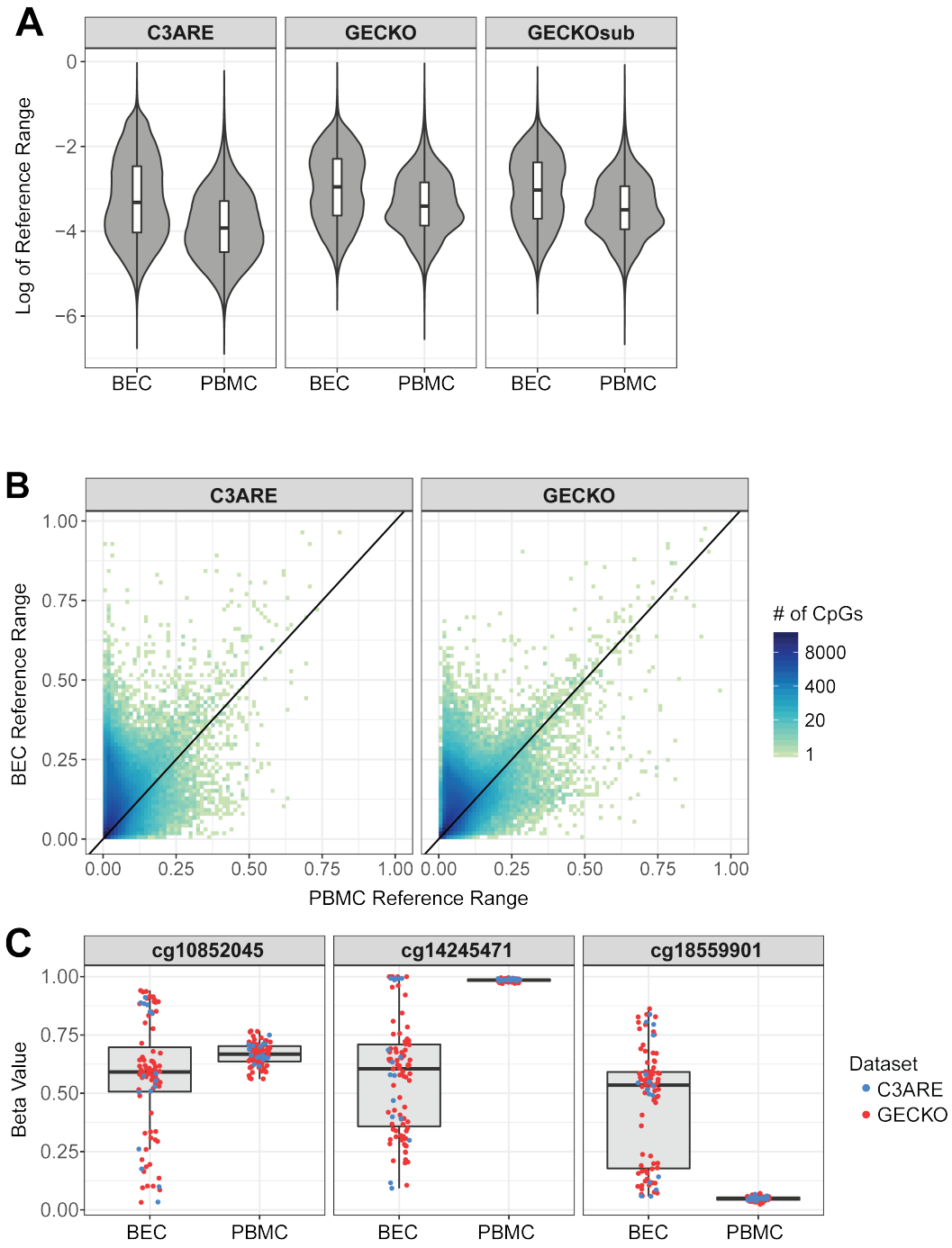


Figure 5.1 BEC DNAm was consistently more variable than PBMC DNAm at the genome-wide and probe-wise level.

A) Distribution of reference range in C3ARE, GECKO and GECKOsub, showing a significantly great variability in BEC vs. PBMC (Wilcoxon $p < 2.2x10^{-16}$ in each cohort). B) Scatterplot of PBMC versus BEC reference range in each cohort. C) Three examples of CpGs with the greatest reference range difference between tissues. Individuals from the GECKO cohort are shown in red and individuals from C3ARE are shown in blue.

5.3.3 Variable CpGs were more highly correlated between tissues

Taking advantage of the matched tissue design of our cohorts, we evaluated whether DNAm variation in one tissue reflected DNAm variation in the other. We performed probewise Spearman's correlations between paired BEC and PBMC samples for the C3ARE, GECKO and GECKOsub datasets, respectively (Supplementary Figure 5.4). Using multiple reference range thresholds to capture increasingly variable CpGs, as previously described, we observed progressively greater enrichment of highly positively correlated CpGs, irrespective of sample size (Figure 5.2A and Supplementary Table 5.1) (292). This suggested that CpGs with greater variability were more likely to be correlated between these tissues.

We next sought to investigate DNAm concordance and variability at individual CpGs. Specifically, we aimed to identify "informative sites", which we defined as CpGs that are both variable across individuals and highly correlated between BECs and PBMCs (292). Such CpGs may be predictive of PBMC DNAm when measured in BECs or vice versa, while reflecting potentially relevant biological variability. Using a previously described method, we defined a correlation coefficient threshold for informative CpGs in each cohort (292). To be classified as informative, i.e. concordant and variable, a CpG was required to have a reference range $\geq 5\%$ in both tissues and meet the predetermined minimum correlation coefficient between tissues of 0.47 in GECKO samples and 0.32 in C3ARE samples. Overlapping CpGs that met these criteria in both cohorts resulted in a set of 8,140 informative sites. Of note, we observed a greater than expected by chance overlap (3682 out of 8140 sites, 45%, 10,000 permutations, $p < 1x10^{-4}$) between our set of informative sites and informative CpGs previously identified between matches samples from adult brain and blood tissues (292). Visualization of our six most correlated informative sites revealed continuous distributions of positively correlated DNAm values between the tissues, as expected (Figure 5.2B). However, the most variable informative sites exhibited discrete distributions with 2 to 3 distinct clusters, suggesting that these CpGs may be enriched for CpGs which are likely under genetic influence (Figure 5.2B).

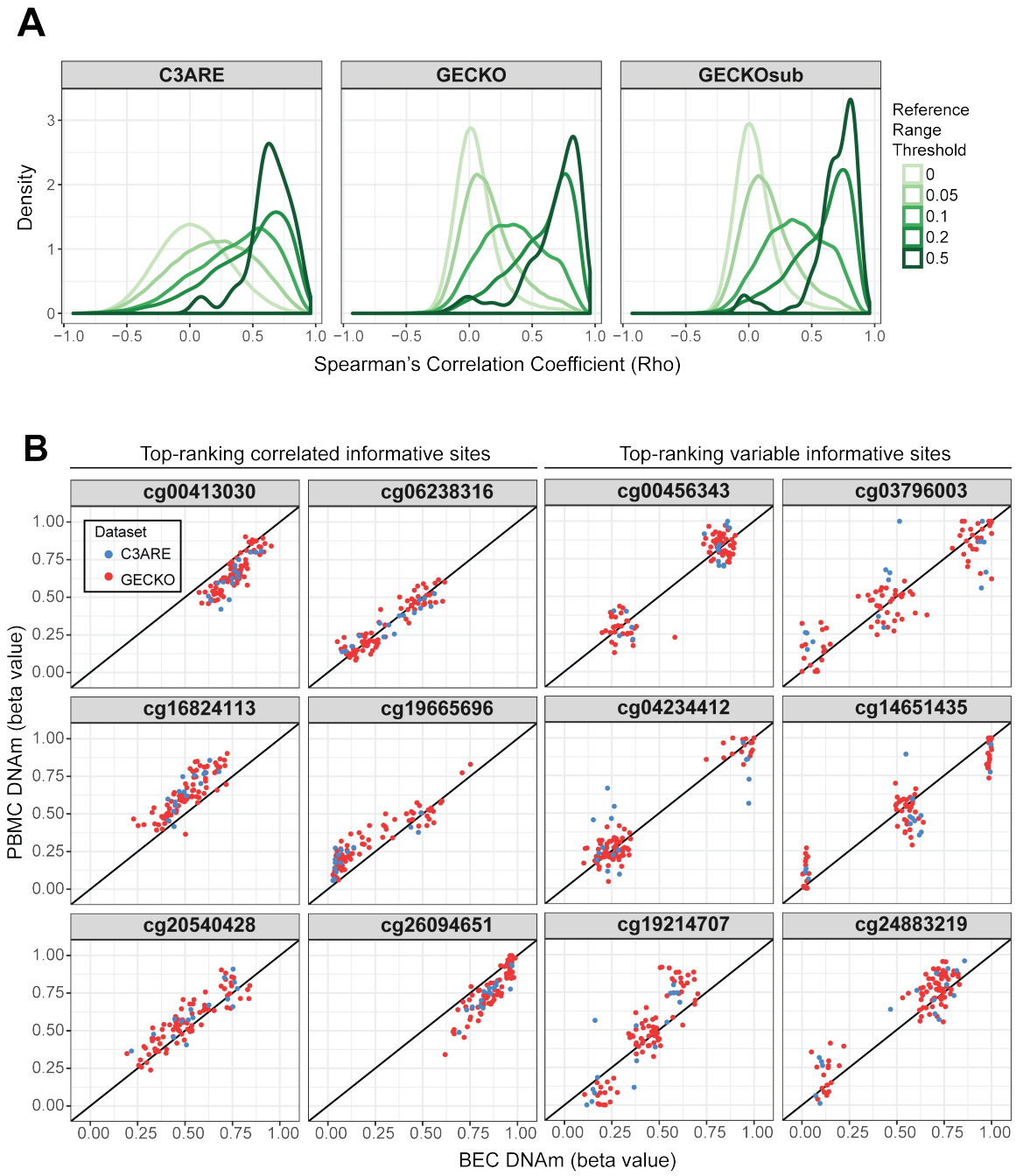


Figure 5.2 Variable CpGs were more highly correlated between tissues.

A) Density distribution plots of Spearman's correlation rho between matched PBMCs and BECs across C3ARE, GECKO and GECKOsub datasets showing progressively greater enrichment of highly positively correlated CpGs at increasing reference range thresholds. Reference range thresholds were set along a sliding scale with cut-offs at 0, 0.05, 0.1, 0.2 and 0.5 (depicted by gradient of green lines). B) Scatterplots of BEC DNAm versus PBMC DNAm for a representative set of informative sites (defined as CpGs that are both variable across individuals and highly correlated between BECs and PBMCs). Top-ranking correlated informative sites (shown in the left two columns) exhibited continuous distributions. In contrast, top-ranking variable informative sites (shown in the right two columns) exhibited discrete distributions, suggesting that these Cps may be under genetic influence. C3ARE samples are shown in blue while GECKO samples are shown in red.

5.3.4 Genetic variation contributed to tissue concordance

In order to determine the influence of local genetic variation on inter-individual DNAm variability and concordance of DNAm signal across matched peripheral tissues, we identified *cis*-mQTL in both BEC and PBMC samples, respectively. Briefly, CpGs were filtered by DNAm variability (reference range ≥ 0.05) in their respective tissues and were correlated against all SNPs within a 5kb window, a window size previously demonstrated to enrich for mQTLs that are more likely to be functionally linked to proximal CpGs (43, 89, 484). As the GECKO cohort had a larger sample size as compared to C3ARE and was therefore more adequately powered for *cis*-mQTL detection, the GECKO samples were used as the discovery cohort. In GECKO, a total of 165,591 unique SNP-CpG pairs in PBMC and 261,739 unique SNP-CpG pairs in BEC were interrogated for associations between DNAm and allelic variation; this included 145,222 SNP-CpG pairs tested in both tissues. A total of 10,521 PBMC-specific, 11,886 BEC-specific and 6,359 shared-tissue significant *cis*-mQTL were identified in GECKO (FDR ≤ 0.05 and DNAm change per allele ≥ 2.5%) and were selected for validation testing in C3ARE (Figure 5.3A).

After quality control processing and variability filtering of the C3ARE DNAm and genotyping data, 16,138 and 17,563 SNP-CpG pairs could be tested for validation in PBMCs and BECs, respectively. This resulted in a total of 1,871 PBMC-specific, 3,705 BEC-specific and 1,097 shared-tissue validated *cis*-mQTL (FDR ≤ 0.05 and DNAm change per allele ≥ 2.5%), which exhibited highly consistent effect sizes between GECKO and C3ARE cohorts (Spearman rho = 0.92, $p = 2.2 \times 10^{-16}$) (Figure 5.3A-B). The overlap between validated *cis*-mQTL between tissues was greater than expected by chance (10,000 permutations, p-value < 1×10^{-4}) (Figure 5.3A and Supplementary Figure 5.5). This suggested that genetic influences contribute to co-variation between tissues. Finally, we found a significant overlap of our PBMC-specific and shared-tissue *cis*-mQTL with previously published mQTL hits from 7-year-old whole blood samples (1810 out of 2968 sites, 61%, 10,000 permutations, $p < 1 \times 10^{-4}$), further supporting our mQTL findings (77).

We next sought to characterize the 6,673 validated *cis*-mQTL in terms of their genomic localization and functional features. Firstly, the 4,980 unique CpGs associated with the validated *cis*-mQTL showed a greater than expected by chance enrichment in intergenic regions and were depleted in intragenic and north shelf regions (2-4 kb upstream of CpG islands) (Supplementary

Figure 5.6A, FDR ≤ 0.05). In addition, we found that CpGs associated with shared-tissue *cis*-mQTLs exhibited a greater than expected by chance enrichment of informative CpGs (687 out of 812 unique CpGs in shared-tissue *cis*-mQTLs, 85%, 10,000 permutations, $p < 1x10^{-4}$), further substantiating that site-specific DNAm correlation between tissues are influenced, in part, by genetic variation (Supplementary Figure 5.6B).

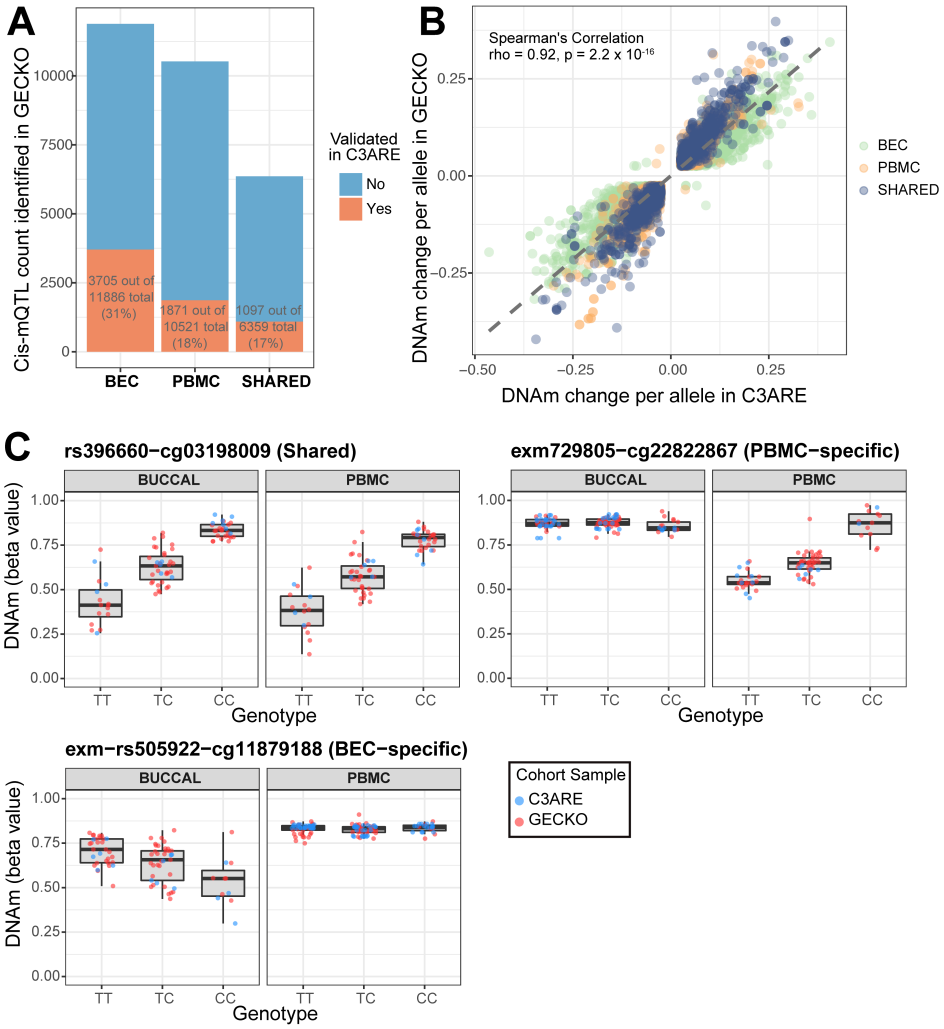


Figure 5.3 Independently validated *cis*-mQTL were more likely to be shared across tissues than expected by chance.

A) Stacked bar plot representing number of *cis*-mQTL identified in GECKO discovery cohort (shown in blue) and number of *cis*-mQTL validated in C3ARE cohort (shown in red) in either BECs, PBMCs or shared across both tissues. (B) Scatterplot of DNAm change per allele in GECKO versus C3ARE across all validated *cis*-mQTL shows mQTL effect sizes (measured as DNAm change per allele) were highly consistent across cohorts (BEC-specific, PBMC-specific and shared-tissue mQTL shown in different colours). C) Boxplots of genotype versus DNAm for representative examples of a shared-tissued (top left), BEC-specific (top right) and a PBMC-specific (bottom) validated *cis*-mQTL. C3ARE samples are shown in blue while GECKO samples are shown in red.

5.3.5 Tissue-specific differential DNAm was consistent across cohorts

Taking further advantage of our matched tissue design, we subsequently assessed differential DNAm between PBMCs and BECs at individual CpGs for both cohorts. In the GECKO samples, 36% of CpGs (150,647) were differentially methylated between matched BECs and PBMCs (Wilcoxon signed rank test; FDR ≤ 0.05 and delta beta ≥ 0.05). The number of significant differentially methylated sites were not greatly affected by sample size differences as GECKOsub had similar findings with 36% of sites exhibiting differential DNAm (149,094 CpGs, with 148767 sites overlapping with GECKO). Similarly, in C3ARE, 38% of CpGs (157,992) were significantly differentially methylated (Wilcoxon signed rank test; FDR ≤ 0.05 and delta beta ≥ 0.05). The overwhelming majority of these CpGs (139,662) were differentially methylated in the same direction in GECKO, GECKOsub and C3ARE (Figure 5.4). Of these sites, 102,203 (73%) had greater average DNAm in PBMCs and 37,459 (27%) had greater average DNAm in BECs, suggesting PBMC DNA was more highly methylated as compared to BEC DNA.

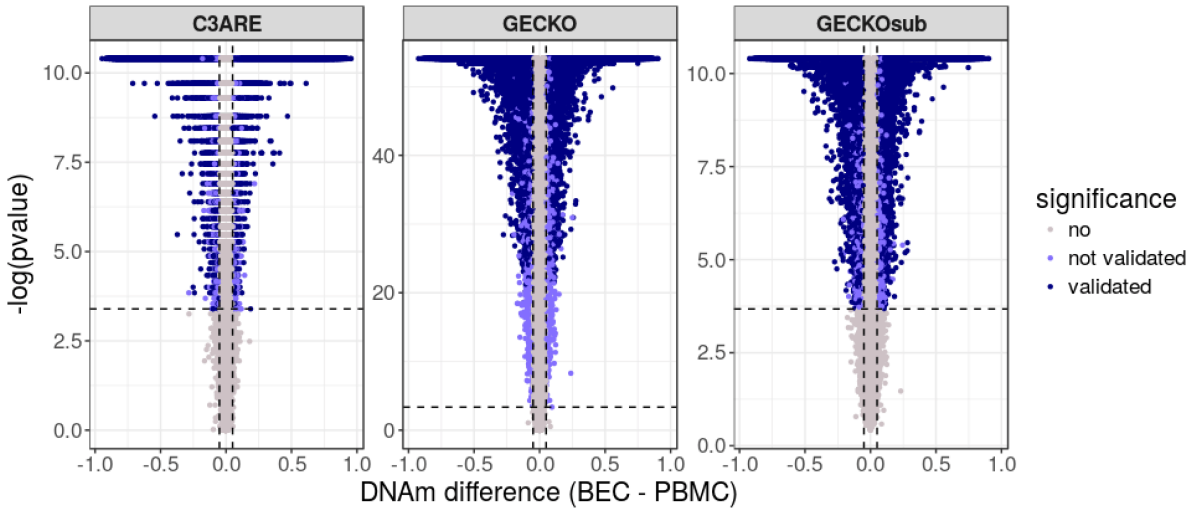


Figure 5.4 Tissue-specific differential DNA methylation was consistent across cohorts.

Volcano plots of differential methylation analysis (run using a paired Wilcoxon signed rank test) between BEC and PBMC tissues for C3ARE, GECKO and GECKOsub datasets. Vertical lines represent an effect size threshold of > 0.05 for absolute mean difference between tissues (BEC - PBMC) and the horizontal line represents the nominal p-value corresponding to an FDR < 0.05 in each cohort. CpGs in dark purple met the effect size and significance cut-offs independently in all three datasets (139,662 CpGs).

5.3.6 Differentially methylated sites were common in published EWAS findings

To provide a granular categorization of CpGs measured on the 450K array, we overlapped CpGs that were identified as a) informative (ie variable across individuals and correlated between BECs and PBMCs) (8,140), b) differentially methylated between matched tissues (139,662), or c) under genetic influence (4,980; i.e. number of unique CpGs associated with validated *cis*-mQTL) across both GECKO and C3ARE cohorts. Of all CpGs found in *cis*-mQTL, 17.7% were informative and 76.2% were differentially methylated (Figure 5.5A). However, in CpGs associated with cross-tissue *cis*-mQTL (812 unique CpGs in total), 84.6% were informative and 58.8% were differentially methylated. As expected, CpGs found in cross-tissue mQTL were enriched for informative sites, as compared to all *cis*-mQTL-associated CpGs (PBMC-specific, BEC-specific, or cross-tissue) (Figure 5.5A).

We then applied this categorization scheme to previously reported EWAS findings performed in pediatric BEC or PBMC tissues to provide an example of how the classification of CpGs can aid in the interpretation of such studies. We selected five published studies that used the 450K array in pediatric BECs or peripheral blood to assess DNAm variation associated with puberty, aging in early life, childhood psychotic symptoms, fetal alcohol spectrum disorder and autism spectrum disorder (147, 148, 294, 485, 486). By implementing our CpG classification scheme on their respective list of significant EWAS hits, we found that *cis*-mQTLs accounted for 0.02-13.5% of significant CpGs reported in these five studies. Differentially methylated CpGs comprised the most represented type of CpG across all 5 studies with only one study demonstrating an overlap of 24.3% with our identified informative sites (Figure 5.5B; Supplementary Table 5.2) (148). This suggested that while most identified EWAS associations may be distinct to the tissue in which they were examined, in some instances, these associations may be reflected across multiple tissues and/or under genetic influence. Finally, we tabulated our CpGs classifications across all 419,507 450K probes assessed in our study in order to serve as a resource for researchers wishing to compare their own EWAS results. Collectively, these findings reveal the importance of considering DNAm variability and correlation between tissues, as well as genetic influences on these patterns, when interrogating and interpreting EWAS findings from pediatric peripheral tissues.

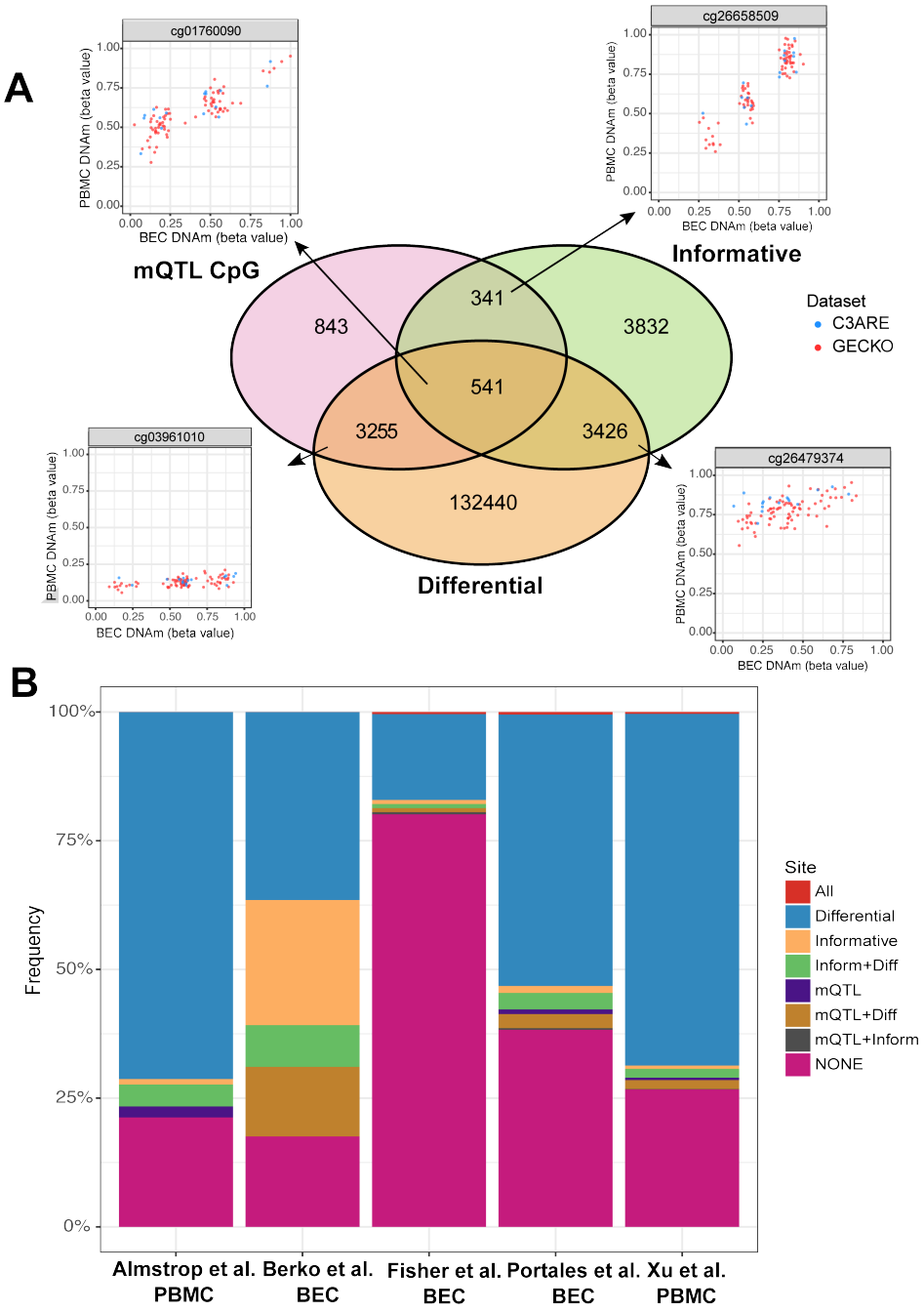


Figure 5.5 Overlap and representation of identified CpGs in previously published pediatric EWAS findings.

A) Venn diagram of CpGs identified as informative, differentially methylated between tissues, or underlying our set of validated *cis*-mQTL. Scatterplots display three representative CpGs from the pairwise intersections between categories. B) Stacked bar plot showing proportion of CpGs of each defined category represented in significant CpGs of various pediatric EWAS publications in BECs or PBMCs. (All = all categories; Differential = differentially methylated between tissues; Informative = informative CpG; Inform + Diff = informative and differential; mQTL = CpG associated with mQTL; mQTL+Diff = mQTL CpG and differential; mQTL+Inform = mQTL CpG and informative; None = not in any of the listed categories).

5.4 Discussion

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

Our tissue-specific DNAm findings highlighted the importance of tissue selection when designing an EWAS. To a large extent, EWAS tissue selection in early life cohorts is guided by two factors. Firstly, ease of collection is particularly important in this age range and may restrict tissue availability. Buccal swabs are less invasive than intravenous puncture, and the latter contributes to participation refusal in pediatric cohorts (488). Secondly, the relevance of the tissue to the phenotype or exposure being tested represents an important consideration for all EWAS analyses, irrespective of age. As peripheral blood represents a circulating tissue with broad immune and inflammatory functions, it might be more relevant to a wider range of health phenotypes than BECs. However, a somewhat competing hypothesis posits that tissues that arise from the same germ layer are more epigenetically similar and thus might be a preferred choice for surrogate tissue selection (489). For example, in comparison to blood, it has been proposed that BEC DNAm may more closely reflect brain DNAm than blood DNAm, as both derive from the ectodermal germ layer (148, 295). Adding to the complexity of this issue, we found that BEC DNAm had significantly greater inter-individual variability than PBMC DNAm at the genome-wide level and at the site-specific level, a finding consistent with adult BECs and PBMCs (317). Having a higher proportion of variable CpGs might be desirable for EWAS analyses as testing any tissue with little inter-individual DNAm variation would naturally limit effect sizes. From this perspective, BECs might represent a more appropriate choice of peripheral tissue for

population-based epigenetic studies over PBMCs. However, it is worth noting that the higher proportion of variable CpGs in BECs may, to some extent, be attributed to the increased diversity of cell types or residual cellular heterogeneity in BECs over PBMCs.

Taking advantage of our matched sample design, we were able to rigorously interrogate the extent of correlation between DNAm signatures of BECs and PBMCs. CpGs with greater variability were more likely to be correlated between matched tissues, as best exemplified by the 8,140 informative sites we identified. These may serve the purpose of aiding in the inference of unmeasured PBMC or BEC DNAm (when the other tissue is measured) as well as for prioritization of sites for cross-tissue replication. In the latter case, cross-tissue replication typically involves the generation of candidate gene lists in accessible tissues for validation in less available tissues, such as post-mortem samples, an approach which can boost confidence in identified associations (490–492). Although there was a substantial overlap (45%) between our informative sites and those previously published in matched adult blood and brain tissues, we found only 1.9% of total measured CpGs to be informative by our measures and thresholds as compared to 9.7% found in the previous analyses from our laboratory (292). These quantitative differences might have a number of reasons, with the most likely being that the blood-brain informative sites were identified using a single cohort, our blood-buccal informative sites were filtered down to sites that were common across both GECKO and C3ARE cohorts.

Integration of genetic and epigenetic information may further clarify the relative contribution of genetic and environmental factors on inter-individual DNAm variability. We found that genetic variation contributed to both inter-individual DNAm variation within a tissue, as well as common DNAm variation between tissues. This is in line with previous findings that show that many – but not all – mQTL have consistent effects across tissues and human populations (75–77, 79, 493). It is currently unclear why in our matched design we observed more BEC-specific mQTL as compared to PBMC-specific or cross tissue mQTL. The most likely explanation is that BECs contained more validated *cis*-mQTL due to greater inter-individual DNAm variability. It is also tempting to speculate that allelic variation contributes more strongly to DNAm in BECs over PBMCs, as blood DNAm may need to be more plastic and responsive due to the role of cells in the immune system. For example, changes in genome-wide transcriptional programs and DNAm profiles are observed in response to an inflammatory

stimulus in blood leukocytes, which could be incongruent with a high degree of fixed, genetically-driven DNAm patterns in these cells (494–496).

As touched upon in several recent reviews, genetic contribution to DNAm might be more prominent in shaping the DNA methylome than initially anticipated, and thus affect the analysis and interpretation of EWAS findings (25, 497). To illustrate this, we tested for the presence of our categorized CpGs in published EWAS findings. Notably, we found CpGs associated with autism spectrum disorder to contain the highest proportion of *cis*-mQTL. While there might be a number of reasons for this, it is possible that the proportion of genetically-influenced CpGs found in an EWAS may be proportional to the heritability of the phenotype under examination, although such hypotheses will require rigorous testing in large cohorts across a diverse spectrum of phenotypes with and without heritable contributions. Furthermore, it is difficult to discern whether having a high proportion of mQTL in EWAS analyses is favourable or not. Previous work has shown the majority of variably methylated regions are best described by an interaction of both genetic and environmental factors (89, 492). As such, any mQTL CpGs found in an EWAS would require further investigation for potential gene by environment interactions.

It is worth noting that our study had a few inherent limitations. Firstly, in both GECKO and C3ARE cohorts, PBMCs were collected from individuals at a slightly later time point than BECs, resulting in an age-related difference (0 – 1.5 years for C3ARE; 0.5 – 2.3 years for GECKO) between matched tissues, which may have affected analyses of DNAm variability. However, we anticipate that age-related differences in DNAm variability are relatively small compared to tissue-specific differences as our findings are consistent with previous work performed on age-matched tissues in adults (317). Another limitation was the relatively small sample size of our cohorts, which may have inflated type II error rates. We also chose to not assess distal genetic effects on DNAm (ie *trans*-mQTL) due to the increased multiple testing burden, but rather prioritized *cis*-mQTL as previous work has suggested these may be more functionally linked to nearby CpGs (43, 89, 484). As well, previous work in blood has shown that the proportion of DNAm variance explained by *trans*-mQTL is much lower than that of *cis*-mQTL (77). For these reasons, we examined SNPs that were directly measured and not imputed within a 5 kb window. Future work using large cohorts will be required to clarify the contribution of distal genetic variants to DNAm in other peripheral tissues.

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

Chapter 6: Conclusion

6.1 Dissertation summary and intersecting features

The body of work presented in this dissertation highlights the strengths and challenges related to the use of different tissue sources in genome-wide DNAm studies of various neuropathological phenotypes. Furthermore, this dissertation provides a basic framework for methodological and biological considerations in the design, analysis and interpretation of DNAm studies of neurobiological diseases.

Beginning with a neurodegenerative disease phenotype, I analyzed genome-wide DNAm profiles from age-matched HD and control postmortem cortex and liver tissues in order to investigate the role of DNAm in HD pathogenesis and tissue-specific *HTT* expression. Although there was minimal evidence of HD-associated DNAm alterations at queried sites after correction for cell heterogeneity, I found that DNAm may be correlated to the age of disease onset in cortex tissues. In contrast, comparison of matched cortex and liver samples revealed numerous site-specific DNAm differences between tissues in the *HTT* gene region, including a novel differentially methylated CTCF binding site in the *HTT* proximal promoter. Overall, these results suggested that DNAm may, in part, contribute to tissue-specific *HTT* transcription through differential CTCF occupancy. Moreover, these findings demonstrated the utility of postmortem tissues in the elucidation of gene regulatory mechanisms, which may be implicated in disease pathogenesis or serve as potential therapeutic targets.

The limited availability of postmortem human brain tissues has prompted the generation of brain-derived *in vitro* cell models for molecular experimentation and preclinical drug testing. In this context, I next assessed DNAm variation in a cell culture system of a neurological malignancy, specifically comparing genome-wide DNAm measures from GBM tumours and matched primary neurosphere cultures enriched for BTIC populations. I detected widespread differences between paired BTICs and tumours in their global DNAm profiles, including a homeobox-enriched differential DNAm signature comprising multiple genes from the HOXA, HOXC and HOXD clusters, as well as other HOX gene family members. Beyond DNAm levels, I also observed differences in DNAm variability, finding that BTIC DNAm was more variable than tumours and that increasingly variable CpGs were more likely to be positively correlated between matched samples. Despite these differences, the relationship between DNAm and gene

expression at a key prognostic marker, the *MGMT* promoter region, was consistent between BTICs and tumours, signifying that BTICs may, to some extent, conserve certain regions of the transcriptional regulatory circuitry from their parental tumours. Together, these findings provided a starting framework for the evaluation of genome-wide DNAm landscapes in BTICs in comparison to bulk tumour tissue. Furthermore, as these *in vitro* model systems are heavily used in drug testing, such in-depth molecular characterizations may help interpret drug screen results and strategically select target biological processes (498).

Moving beyond analyses in brain-derived samples, I subsequently assessed DNAm variation related to a neuropsychiatric addiction disorder, alcohol dependence (AD), in purified $CD3^+$ T-lymphocytes, a relevant blood cell type whose relative subtype abundance and activity have been linked to alcohol consumption (445, 447, 499). Specifically, I examined longitudinal measures of DNAm from a cohort of alcohol dependent patients undergoing a clinical intervention, along with closely matched healthy controls. I identified numerous differentially methylated CpG sites comparing patients prior to treatment with healthy controls and was able to confirm a subset of those sites in additional analyses for differentially methylated regions. Comparing patients before and after the alcohol treatment program revealed another unique set of differentially methylated CpG sites. Additionally, I found that the mean global DNAm was significantly lower in patients prior to treatment compared to controls, but reverted back to levels similar to controls after treatment. Following verification of top-ranked hits by pyrosequencing and replication in an independent cohort, I confirmed differential DNAm of *HECW2* and *SRPK3* in whole blood, demonstrating the potential relevance of the identified associations as blood-based biomarkers. Overall, these results captured DNAm variation in a disease-relevant blood cell type of AD and implicate *HECW2* and *SRPK3* DNAm as promising blood-based candidates to follow up in future studies. Moreover, this work underscored the potential for longitudinal monitoring of disease-related DNAm associations in available peripheral tissues, particularly in the context of clinical interventions.

The use of readily accessible peripheral tissues is particularly prevalent in pediatric cohorts in which postmortem samples or biopsy specimens with invasive collection procedures are less common than in adult population (500, 501). In my final data chapter, I present a systematic comparison of genome-wide DNAm patterns between matched pediatric BECs and PBMCs, two of the most widely used peripheral tissues in human epigenetic studies.

Specifically, I assessed DNAm variability, cross-tissue DNAm concordance and genetic determinants of DNAm across two independent early life cohorts encompassing different ages. Drawing on methods for DNAm variability analysis used in Chapter 2, I found that BECs had greater inter-individual DNAm variability over PBMCs and that highly variable CpGs were more likely to be positively correlated between the matched tissues. These sites were enriched for CpGs under genetic influence, suggesting that a substantial proportion of DNAm co-variation between tissues could be attributed to genetic variation. Finally, I demonstrated the relevance of these findings to human epigenetic studies by categorizing CpGs from published DNAm association studies of pediatric BECs and peripheral blood. Taken together, these results highlighted a number of important considerations and practical implications in the design and interpretation of EWAS analyses performed in pediatric peripheral tissues.

A common theme across all studies presented in this dissertation is the implementation of cross-tissue comparisons to investigate the potential role of DNAm associations in different neurobiological disease domains. Specifically, these cross-tissue analyses between matched samples offer a number of advantages and opportunities to gain new potential insights into DNAm variation in neuropathological conditions or other complex phenotypes in general. Firstly, this approach may help identify novel tissue-specific gene regulatory mechanisms, as demonstrated in Chapter 2 by the use of postmortem cortex-liver comparisons to delineate the potential contribution of differential DNAm and CTCF occupancy on tissue-specific *HTT* gene expression. Secondly, such cross-tissue analyses may be informative in the evaluation of primary culture models relative to their parental tissue source, as exemplified in Chapter 3 in which DNAm profiles between BTIC lines and parental GBM tumours were contrasted to reveal unique DNAm differences and similarities between the sample types. Thirdly, DNAm comparisons using peripheral tissues may allow for repeated or long-term DNAm assessments or the discovery of disease biomarkers, as illustrated in Chapter 4 by the detection AD-associated DNAm alterations in patients' T-cells during clinical intervention treatment and confirmation of these DNAm associations in whole blood. Finally, comparative analyses using peripheral tissues may bolster cross-tissue replication potential in DNAm studies, particularly if informative sites with high cross-tissue correlation can be determined, as identified between matched pediatric BECs and PBMCs in Chapter 5 (490–492). Taken together, the comparison of DNAm profiles between matched tissues may not only help clarify the tissue-specificity of DNAm variation in

health and disease, but may also illuminate new strategies to track such variation across the lifespan.

6.2 Limitations and caveats

Although the work presented herein offers new potential insights and tissue-specific approaches to investigate DNAm associations to neurobiological diseases, there are a number of methodological limitations and conceptual caveats which should be taken into consideration when interpreting these findings. These considerations, which are largely consistent between studies, include potential technical confound in DNAm signal by DNA hydroxymethylation, limited coverage of the methylome by the Illumina 450K array, potential biological confounds by cellular composition, genetic variation or environmental factors and the correlative nature of these findings.

6.2.1 DNAhm signal confound in bisulfite-converted DNAm measures

Given that DNA bisulfite conversion leads to deamination of unmethylated cytosine residues to uracil, leaving 5mC or 5hmC marks intact, the resultant bisulfite-converted DNAm readouts represent a mixed signal of both 5mC and 5hmC modifications (502). As all studies in this dissertation used bisulfite-converted genomic DNA for Illumina 450K or pyrosequencing analyses, it is not possible to distinguish between 5mC and 5hmC levels in the DNAm data. This technical confound may have more problematic implications in certain tissues and disease contexts over others. For example, although 5hmC content varies dramatically across different tissues, DNAhm levels are highest in brain, accounting for 25% of all modified CpG dinucleotides in the frontal cortex, the target tissue used in the HD DNAm study in Chapter 2 (98, 116, 353). Variation in DNAhm may also be particularly pertinent to brain tumour biology as previous studies have reported that a commonly mutated gene in gliomas, *IDH1/2*, can indirectly impair TET-mediated catalysis of 5hmC and that pattern-specific loss of 5hmC is associated with poor survival in GBM patients (503, 504). Furthermore, as previous work has suggested that 5hmC production is required for tumorigenicity of GBM-derived BTICs, it is possible that DNAhm variation represents an additional contributing factor to the epigenetic discordance between GBM tumours and matched BTICs observed in Chapter 3 (505). Consequently, follow-up analyses using alternative biochemical treatment of genomic DNA,

such as TAB or oxBS conversion, are needed to resolve 5hmC from 5mC levels in bisulfite-converted DNAm measures (121, 122).

6.2.2 Limited methylome coverage of Illumina 450K array

The Illumina Infinium DNAm arrays are extensively used in EWAS analyses as a relatively fast and cost-effective method to perform high-throughput DNAm profiling across large sample sets (506). However, these microarray-based platforms are restricted by both the number and specificity of probes, thereby limiting their genomic coverage. For example, the Illumina 450K array, which was exclusively used for all EWAS analyses in this dissertation, interrogates less than 2% of all DNAm sites genome-wide (129). This coverage is relatively sparse for distal regulatory elements such as enhancers and generally biased towards CpG-dense promoter regions, which have limited inter-individual and inter-tissue variation (52, 85, 507). Moreover, removal of 450K probes due to cross-reactivity, non-specific hybridization and binding to polymorphic loci can further reduce the number of usable probes (324, 508). As a result, it is possible that disease-relevant DNAm variation exists at sites beyond those queried in the reported analyses for which more comprehensive DNAm quantification technologies, such as the newly-released Illumina EPIC array, may be warranted (130). Additionally, sequencing-based techniques, including RRBS, targeted Methyl-Capture Sequencing (MC-Seq) and the gold standard, whole genome bisulfite sequencing (WGBS), offer higher genomic coverage with greater density-per-region over the 450K array, although each approach may have relative drawbacks in terms of cost, reproducibility and feasibility (126, 506, 509, 510).

6.2.3 Potential confounding by cellular heterogeneity

Due to their role in establishing and maintaining cellular identity, epigenetic marks including DNAm exhibit substantial cell type specificity. Interindividual differences in cell composition within complex human tissues can often confound EWAS analyses by masking true associations or altering DNAm signatures to give rise to spurious associations (20–23, 25, 371, 497). In the context of the studies presented here, two broad approaches were employed to address cellular heterogeneity: (1) cell type adjustment methods of DNAm measures from bulk tissue; and (2) isolation of target cell(s) for epigenetic interrogation. In the first approach, used for DNAm analyses of brain, blood and buccal tissues (Chapters 2, 4 and 5, respectively),

bioinformatic deconvolution algorithms were applied to estimate cell type proportions of bulk tissue based on underlying reference profiles from isolated cells (ie neuronal versus nonneuronal populations predicted in cortex tissue in Chapter 2). In the second approach, which was used to obtain $CD3^+$ T-lymphocytes in Chapter 4, desired cell type(s) were sorted or purified from heterogeneous tissue. Both approaches relied, either directly or indirectly, on cell isolation techniques using cell surface markers. However, emerging evidence from single-cell transcriptomic studies have revealed much finer molecular resolution than cell surface marker classifications, suggesting that surface marker expression may be inadequate at capturing higher levels of continuity between cellular subtypes in bulk tissue (511, 512). Futhermore, previous work, along with my own analyses in purified T-cell fractions (Chapter 4), have confirmed the presence of cellular heterogeneity in samples following purification using cell surface markers, indicating that even after purification techniques, the resultant pool of cells is composed of multiple epigenomes (termed a "meta-epigenome")(469). In the context of the DNAm comparisons between BTICs and parental GBM tumours (Chapter 3), "reference-free" methods which correct for the effects of cell composition without actually predicting cell proportions can be used to deconvolute 'tumoural purity' from solid tumours, although it is unclear whether such algorithms would be suitable for primary cell cultures such as BTICs (404, 411). Taken together, it is possible, and likely, that the full spectrum of cell-type diversity has not have been accounted for in the examined tissues and that reported EWAS findings, may, in part, reflect residual cell-type variation (25). As such, future work, likely invoking single-cell assays, is needed to allow for fine-grained estimation and correction for cell heterogeneity in EWAS associations.

6.2.4 Accounting for genetic variation and ethnicity

Ethnicity differences may influence DNAm patterns, in part through population-specific genetic influences on DNAm as well as through culturally associated differences in lifestyle, diet, or habitat (76, 497, 513, 514). In the first case, genetic influences on DNAm can vary across different populations and ethnic groups as supported by the identification of population-specific mQTLs and the fact that more than half of the differential CpG methylation detected between two populations can be primarily explained by local genetic variation (43, 76, 89, 484, 514–517). Although genotyping data was used to rule out imbalances in genetic ancestry groups in Chapter 5, unaccounted genetic differences in population structure may still be present in EWAS findings

of the other studies. In addition, ethnicity-related patterns in DNAm variation may arise from cultural and environmental commonalities. For instance, a study of two African populations showed distinct DNAm signatures of historical lifestyle and current habitats, in which the former was associated with developmental processes while the latter was related to cellular and immune functions (513). In support of these results, a more recent study in diverse Hispanic populations reported that self-identified ethnicity and genetically-determined ancestry have unique associations to DNAm in whole blood (514). Overall, these observations suggest that ethnicity differences are associated with DNAm patterns either directly through ancestry-related genetic variation or indirectly through culturally-dependent, environmental factors. Future studies, which use genotyping data or statistical methods to infer ancestry information from DNAm data, are needed to clarify genetic- and ethnicity-based differences in DNAm variation, particularly in the context of disease phenotypes (518).

6.2.5 Accounting for potential environmental covariates

Environmental factors and past exposures have been widely implicated with DNAm variation and therefore represent an additional source of potential confounding in EWAS analyses (4, 5, 519). For example, both current and prenatal exposure to cigarette smoke has been reproducibly associated with DNAm alterations across numerous tissues including PBMCs, BECs, cord blood and placenta, although accumulating evidence suggests that blood-based DNAm associations to tobacco exposure may be confounded by changes in cell composition (17–19, 439, 520–525). While smoking habits were matched between AD patients and controls in Chapter 4, smoking behavior was not accounted for in analyses of other disease contexts, including the HD study in Chapter 2 and the GBM study in Chapter 3, thereby potentially confounding identified DNAm associations.

Pharmacological treatments and medication history may also represent unaccounted sources of inter-individual variation which may influence DNAm. For instance, dose-dependent inhibition of DNMT1 activity and global reduction of mC content have been associated with disulfiram administration, a drug used to treat chronic alcoholism (526). Although information on treatment history was not available for our AD patient cohort, it is possible that prior or ongoing disulfiram use may, in part, explain the decrease in global DNAm levels of T-cells in AD patients over controls (Chapter 4). Moreover, in certain cases, treatment-induced DNAm

changes may not be associated with exposure to the drug itself but rather through secondary effects induced by drug treatment (ie hypermutation and induction of genomic instability). For example, GBM recurrence following treatment with a DNA alkylating chemotherapeutic, temozolomide (TMZ), often results in drug-resistant tumours which carry mutations in mismatch repair (MMR) genes and thus, exhibit a hypermutated phenotype (527). The degree to which TMZ-induced hypermutation affects DNAm in GBM tumours and primary BTIC cultures remains to be further investigated (Chapter 3).

Finally, from a broader perspective, there has been increasing efforts to explore linkages between epigenetic variation and an individual's 'exposome', defined as all exposures to which an individual is subjected from conception to death (528, 529). The exposome comprises specific external factors (ie chemical toxicants, pollutants, infectious agents, etc), internal processes (ie hormones, gut microflora, oxidative stress, metabolism, etc) and more general environmental influences (ie socioeconomic status, psychological stress, etc) (528). Beyond the type of exposure, the dose, duration and the developmental timing of the exposure may also play an important role in potentially influencing epigenetic processes and overall disease susceptibility (529). Specifically, the relationship between different exposures and DNAm variation in neurobiological disease risk remains largely undefined and represents an active area of epigenetic epidemiological research.

6.2.6 Correlation versus causation

Although the data presented in this dissertation implicates DNAm variation in disease risk and progression across different neuropathological domains, these findings are merely correlational and do not imply causal mechanisms. Moreover, it is difficult to resolve the functional implications of these associations, as it is not yet clear whether DNAm regulates transcription or if it is a result thereof, making interpretations of reverse causation highly possible (23, 25). In order to test causality of DNAm associations, a number of different approaches are relevant. Firstly, advanced statistical methods can be used such as Mendelian randomization, which employs genetic polymorphisms as instruments to strengthen causal inference from observational data (530, 531). Secondly, the establishment of mechanistic causes would require experimental manipulation and testing using cellular and animal models. In this context, emerging technologies such as CRISPR fused to chromatin modifiers may be useful

tools for targeted epigenome editing (532). Thirdly, the implementation of prospective cohort designs, including intervention studies, will afford temporally-ordered data collection across critical developmental periods. Together, these approaches may help clarify the appropriateness of causal interpretations from DNAm studies in neurobiological diseases.

6.3 Considerations for future EWAS analyses of neurobiological disease phenotypes

Substantial progress has been made to date in characterizing DNAm variation associated with various brain-related pathologies. However, moving forward, a number of considerations should be taken into account when designing and interpreting EWAS analyses of neurobiological disease phenotypes, as outlined below.

6.3.1 Diagnostic or phenotypic heterogeneity

In terms of diagnostic or phenotypic heterogeneity, many of the neurobiological phenotypes of interest in human DNAm studies exist as syndromes that display a wide range of clinical symptoms and varying degrees of each diagnostic characteristic. For example, alcoholism (examined in Chapter 4) can be categorized into two broad groups: Type I, which is characterized by a late onset of dependence, low prevalence and familial alcoholism and a milder course, and Type II, which is described by early onset of dependence, high familial alcoholism in fathers, frequent antisocial personality and heightened intensity of alcohol-related problems (533, 534). To reduce such phenotypic variability, well-characterized endophenotypes are being developed to clarify descriptive diagnostic criteria into more stable phenotypes with clear genetic linkages (535, 536). In addition, the inclusion of an illness comparison group can help discern if an epigenetic association is linked to a particular disorder, as opposed to a generalized effect of disease in the brain(537, 538). For instance, a previous study delineated shared versus distinct epigenetic signatures between the neurodegenerative disease modalities in postmortem brain tissues from individuals with Alzheimer's disease, dementia with Lewy bodies, PD, and Alzheimer-like neurodegenerative profile associated with Down syndrome (262). Similarly, in the context of GBM (analyzed in Chapter 3), the inclusion of DNAm patterns and mutational profiles of epigenetic factors helped refine molecular subtype classifications of GBM tumours across the age spectrum (226, 227). These approaches, along with well-defined inclusion and exclusion criteria, can help limit phenotypic heterogeneity in epigenetic study cohorts.

6.3.2 Effect sizes

Effect size in DNAm studies, often calculated as the mean difference in proportion of methylated DNA alleles, is largely dictated by the method and location of the DNAm measurement. In reality, a particular CpG is present only twice within a given cell, on each allele of a homologous chromosome pair, so DNAm can only be 0% (unmethylated), 50% (hemi-methylated) or 100% (methylated). Current methods of DNAm assessment generally involve simultaneous quantification of DNAm from a pool of cells within a given sample. Thus, a site-specific change in overall DNAm of a biological sample may reflect changes in DNAm at a small subset of cells, thereby representing unaccounted cellular mosaicism (23). Moreover, the magnitude of the DNAm effect sizes can, in part, be attributed to the location of DNAm measurement in the methylome. For example, CGIs in promoters typically exhibit low DNAm levels, with little inter-tissue and inter-individual variability, as compared to distal regulatory elements such as enhancers, where more dynamic DNAm variation occurs (52). At present, a key methodological challenge in EWAS analyses is to distinguish true signal from noise in order to gain more accurate estimates of DNAm effect size. The establishment of an a priori threshold for DNAm change (ie 5%) may be used in an effort to decrease the likelihood of false positives and prioritize findings that may have biological significance. However, these effect size thresholds, if used at all, can be arbitrary and difficult to standardize across different contexts. For example, the magnitude of DNAm change expected for cancer and cross-tissue comparisons may be higher than DNAm differences for broader, noisier and less tangible exposures (ie early life adversity). Importantly, the functional consequence of an observed change in DNAm remains unclear, although a number of DNAm findings related to environmental exposures have been linked to transcriptional alterations and replicated in independent cohorts (539). Taken together, careful consideration of DNAm effect sizes may help reduce false discoveries, facilitate replication potential and illuminate functional relevance in the interpretation of EWAS analyses.

6.3.3 Data sharing and integration

The development of publicly available resources and repositories for sharing molecular data have become a high priority, motivated, in part, by the difficulty in obtaining high-quality human tissue samples, particularly from less accessible tissues such as brain. Large-scale consortiums, such as the National Institutes of Health (NIH) Roadmap Epigenomics Consortium,

the International Human Epigenome Consortium (IHEC), The Cancer Genome Atlas (TCGA), the Encyclopedia of DNA Elements (ENCODE) and PsychENCODE, have initiated large-scale projects to generate human reference epigenomes in a wide range of tissues and cell-types across different developmental stages and disease contexts (53, 54, 335, 360, 540) . These efforts not only allow for the development of tools and resources for the larger scientific community to fuel replication efforts, but also facilitate multidisciplinary collaborations to intersect expertise and resources. It is worth mentioning that public access to data can be a sensitive issue, as privacy protections and ethical issues can be associated participant metadata (541). In such cases, appropriate consent procedures need to be enforced so that data can be properly anonymized and shared with the broader community (541). In cases where certain data types (ie genetic information) cannot be shared publicly, collaborations between individual researchers can help to ensure that provisions for data access are made openly and responsibly (497, 541).

Finally, these types of data-sharing initiatives and collaborations are important resources for integrative analyses aimed at combining molecular data types, including different epigenetic marks, genetic variation, RNA expression, protein levels, and chromosomal conformation (542). For example, recent integration of DNAm profiles and histone marks from the NIH Roadmap Epigenomics Consortium and ENCODE projects revealed distinct chromatin signatures of DNA hypomethylation in aging and various tumour types, suggesting that the role of DNAm as a molecular link between aging and cancer is more complex than previously thought (543). Such integrative analyses are gaining traction due to their potential to provide insights into robust biological variation in the human brain, particularly as they pertain to complex disease states (544, 545).

6.4 Future directions

EWAS analyses in neurobiological disease are rapidly evolving to provide novel opportunities to characterize molecular features of brain-related function and pathology. Specifically, promising advancements in this field are occurring along a number of distinct avenues. Firstly, the development of new methodologies for single-cell epigenomic profiling has the potential to transform our understanding of cellular subtype diversity in heterogeneous tissues, as well as more deeply interrogate the epigenetic regulatory landscape in individual cell types (546, 547). For example, the advent of a single-cell method for parallel chromatin

accessibility, transcriptome and DNAm profiling revealed novel associations between the three molecular layers in murine embryonic stem cells (548). Although such techniques have yet to be applied in the context of human tissues, they provide promising new options for investigating cell-specific gene regulatory mechanisms.

Secondly, given that DNAm is mitotically heritable and may serve as a mechanism for cellular memory, there has been increasing interest in the potential for epigenetic inheritance across generations. While this possibility is certainly intriguing, the appropriate examination of this phenomenon is complex and complicated by conceptual misunderstandings. For example, the distinction between intragenerational and intergenerational (or transgenerational) inheritance is often misinterpreted. In the former scenario of intragenerational inheritance, offspring DNAm patterns are altered by exposure affecting the maternal in utero environment. As the germ line cells of the F2 generation can be affected in the pregnant F0 female, these effects may be plausibly observed for up to two generations (549). By contrast, transgenerational inheritance requires evidence of transmission across three generations in the females (550). As there are currently no human cohorts which meet this level of multigenerational transmission, there is, at present, a lack of unequivocal evidence to support transgenerational epigenetic inheritance in humans. Consequently, this potential phenomenon remains a compelling area of research that warrants further investigation.

Finally, the translation of epigenetic findings towards clinical applications, particularly in the context of personalized medical care, represents a powerful new paradigm in neurobiological health research (551). In the case of DNAm associations for which evidence of mechanistic involvement from target brain tissue is available, understanding the actual molecular mechanism is important for identifying novel therapeutics or treatments. In the absence of mechanistic evidence, DNAm findings may provide a solid foundation for biomarker discovery in peripheral tissues. These biomarkers can be used as tools for early disease diagnosis and monitoring as well as for stratifying patient populations towards more targeted interventions or assessing treatment response. However, given the complexity and the dynamic nature of the methylome, caution is warranted in regards to the selection of appropriate tissue source(s), accounting for potential technical or biological confounds, balanced interpretation of findings and integration with other –omic data types. Taken together, these efforts may provide compelling insights into the gene-

regulatory architecture underlying neurobiological disease susceptibility and pathogenesis as well as offer novel strategies to monitor or mitigate such disease-related variation.

Bibliography

1. Bird A (2002) DNA methylation patterns and epigenetic memory. *Genes Dev* 16(1):6–21.
2. Henikoff S, Greally JM (2016) Epigenetics, cellular memory and gene regulation. *Curr Biol* 26(14):R644–R648.
3. Kim M, Costello J (2017) DNA methylation: An epigenetic mark of cellular memory. *Exp Mol Med* 49(4). doi:10.1038/emm.2017.10.
4. Feil R, Fraga MF (2012) Epigenetics and the environment: emerging patterns and implications. *Nat Rev Genet* 13(2):97–109.
5. Marsit CJ (2015) Influence of environmental exposure on human epigenetic regulation. *J Exp Biol* 218(1):71–79.
6. Meaney MJ (2010) Epigenetics and the biological definition of gene x environment interactions. *Child Dev* 81(1):41–79.
7. Mitchell A, Roussos P, Peter C, Tsankova N, Akbarian S (2014) *The Future of Neuroepigenetics in the Human Brain* (Elsevier Inc.). 1st Ed. doi:10.1016/B978-0-12-800977-2.00008-5.
8. Bakulski KM, Halladay A, Hu VW, Mill J, Fallin MD (2016) Epigenetic Research in Neuropsychiatric Disorders: the "Tissue Issue." *Curr Behav Neurosci Reports* 3(3):264–274.
9. Sweatt JD (2013) The emerging field of neuroepigenetics. *Neuron* 80(3):624–632.
10. Boyce WT, Kobor MS (2014) Development and the epigenome: the 'synapse' of gene-environment interplay. *Dev Sci* 18(1):1–23.
11. Waddington C (1968) Towards a theoretical biology. *Nature* 218(5141):525–527.
12. Christopher MA, Kyle SM, Katz DJ (2017) Neuroepigenetic mechanisms in disease. *Epigenetics Chromatin* 10(1). doi:10.1186/s13072-017-0150-4.
13. Bird A (2007) Perceptions of epigenetics. *Nature* 447(7143):396–398.
14. Greally JM (2018) A user's guide to the ambiguous word "epigenetics." *Nat Rev Mol Cell Biol*. doi:10.1038/nrm.2017.135.
15. Schübeler D, Schuebeler D, Schübeler D (2015) Function and information content of DNA methylation. *Nature* 517(7534):321–326.
16. Mill J, Heijmans BT (2013) From promises to practical strategies in epigenetic epidemiology. *Nat Rev Genet* 14(8):585–594.
17. Joubert BR, et al. (2012) 450K epigenome-wide scan identifies differential DNA methylation in newborns related to maternal smoking during pregnancy. *Environ Health Perspect* 120(10):1425–1431.
18. Breitling LP, Yang R, Korn B, Burwinkel B, Brenner H (2011) Tobacco-smoking-related differential DNA methylation: 27K discovery and replication. *Am J Hum Genet* 88(4):450–457.
19. Monick MM, et al. (2012) Coordinated changes in AHRR methylation in lymphoblasts and pulmonary macrophages from smokers. *Am J Med Genet B Neuropsychiatr Genet* 159B(2):141–151.
20. Michels KB, et al. (2013) Recommendations for the design and analysis of epigenome-wide association studies. *Nat Methods* 10(10):949–955.

21. Chadwick LH, et al. (2015) New insights and updated guidelines for epigenome-wide association studies. *Neuroepigenetics* 1(C):14–19.

22. Rakyan VK, Down TA, Balding DJ, Beck S (2011) Epigenome-wide association studies for common human diseases. *Nat Rev Genet* 12(8):529–541.

23. Birney E, Smith GD, Greally JM (2016) Epigenome-wide Association Studies and the Interpretation of Disease -Omics. *PloS Genet* 12(6):e1006105-9.

24. Teschendorff AE, Relton CL (2018) Statistical and integrative system-level analysis of DNA methylation data. *Nat Rev Genet* 19(3):129–147.

25. Lappalainen T, Greally JM (2017) Associating cellular epigenetic models with human phenotypes. *Nat Rev Genet* 18(7):441–451.

26. Jones PA (2012) Functions of DNA methylation: islands, start sites, gene bodies and beyond. *Nat Rev Genet* 13(7):484–492.

27. Bock C (2012) Analysing and interpreting DNA methylation data. *Nat Rev Genet* 13(10):705–719.

28. Jones PA, Takai D (2001) The role of DNA methylation in mammalian epigenetics. *Science (80-)* 293(5532):1068–1070.

29. Illingworth RS, Bird AP (2009) CpG islands--'a rough guide'. *Fed Eur Biochem Soc Lett* 583(11):1713–1720.

30. Weber M, et al. (2007) Distribution, silencing potential and evolutionary impact of promoter DNA methylation in the human genome. *Nat Genet* 39(4):457–466.

31. Gardiner-Garden M, Frommer M (1987) CpG islands in vertebrate genomes. *J Mol Biol* 196(2):261–282.

32. Ulahannan N, Greally JM (2015) Genome-wide assays that identify and quantify modified cytosines in human disease studies. *Epigenetics Chromatin* 8(1). doi:10.1186/1756-8935-8-5.

33. Saxonov S, Berg P, Brutlag DL (2006) A genome-wide analysis of CpG dinucleotides in the human genome distinguishes two distinct classes of promoters. *Proc Natl Acad Sci U S A* 103(5):1412–1417.

34. Bestor TH (2000) The DNA methyltransferases of mammals. *Hum Mol Genet* 9(16):2395–2402.

35. Robertson KD, et al. (1999) The human DNA methyltransferases (DNMTs) 1, 3a and 3b: Coordinate mRNA expression in normal tissues and overexpression in tumors. *Nucleic Acids Res* 27(11):2291–2298.

36. Okano M, Xie S, Li E (1998) Cloning and characterization of a family of novel mammalian DNA (cytosine-5) methyltransferases. *Nat Genet* 19(3):219–220.

37. Hervouet E, Peixoto P, Delage-Mourroux R, Boyer-Guittaut M, Cartron PF (2018) Specific or not specific recruitment of DNMTs for DNA methylation, an epigenetic dilemma. *Clin Epigenetics*. doi:10.1186/s13148-018-0450-y.

38. Lam LL, et al. (2012) Factors underlying variable DNA methylation in a human community cohort. *Proc Natl Acad Sci U S A* 109 Suppl(Supplement 2):17253–17260.

39. Edgar R, Tan PPC, Portales-Casamar E, Pavlidis P (2014) Meta-analysis of human methylomes reveals stably methylated sequences surrounding CpG islands associated with high gene expression. *Epigenetics Chromatin* 7(1):28.

40. Irizarry RA, et al. (2008) Comprehensive high-throughput arrays for relative methylation

(CHARM). *Genome Res* 18(5):780–790.

41. Baubec T, Schübeler D, Schuebeler D (2014) Genomic patterns and context specific interpretation of DNA methylation. *Curr Opin Genet Dev* 25(1):85–92.

42. Tate PH, Bird AP (1993) Effects of DNA methylation on DNA-binding proteins and gene expression. *Curr Opin Genet Dev* 3(2):226–231.

43. Gutierrez-Arcelus M, et al. (2013) Passive and active DNA methylation and the interplay with genetic variation in gene regulation. *Elife* 2(2):e00523–e00523.

44. Jones MJ, Fejes AP, Kobor MS (2013) DNA methylation, genotype and gene expression: who is driving and who is along for the ride? *Genome Biol* 14(7):126.

45. Shukla S, et al. (2011) CTCF-promoted RNA polymerase II pausing links DNA methylation to splicing. *Nature* 479(7371):74-U99.

46. Maunakea AK, et al. (2010) Conserved role of intragenic DNA methylation in regulating alternative promoters. *Nature* 466(7303):253–257.

47. Maunakea AK, Chepelev I, Cui K, Zhao K (2013) Intragenic DNA methylation modulates alternative splicing by recruiting MeCP2 to promote exon recognition. *Cell Res* 23(11):1256–1269.

48. Lev Maor G, Yearim A, Ast G (2015) The alternative role of DNA methylation in splicing regulation. *Trends Genet* 31(5):274–280.

49. Cordaux R, Batzer MA (2009) The impact of retrotransposons on human genome evolution. *Nat Rev Genet* 10(10):691–703.

50. Donnelly SR, Hawkins TE, Moss SE (1999) A conserved nuclear element with a role in mammalian gene regulation. *Hum Mol Genet* 8(9):1723–1728.

51. Smith ZD, Meissner A (2013) DNA methylation: roles in mammalian development. *Nat Rev Genet* 14(3):204–220.

52. Ziller MJ, et al. (2014) Charting a dynamic DNA methylation landscape of the human genome. *Nature* 500(7463):477–481.

53. Kundaje A, et al. (2015) Integrative analysis of 111 reference human epigenomes. *Nature* 518(7539):317–330.

54. Stunnenberg HG, Consortium TIHE, Hirst M (2016) The International Human Epigenome Consortium: A Blueprint for Scientific Collaboration and Discovery. *Cell* 167(5):1145–1149.

55. K L, et al. (2014) DNA methylome profiling of human tissues identifies global and tissue-specific methylation patterns. *Genome Biol* 15:1–14.

56. Yuen RKC, et al. (2011) Extensive epigenetic reprogramming in human somatic tissues between fetus and adult. *Epigenetics Chromatin* 4(1):7.

57. 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.

58. Irizarry RA, et al. (2009) The human colon cancer methylome shows similar hypo- and hypermethylation at conserved tissue-specific CpG island shores. *Nat Genet* 41(2):178–186.

59. Farré P, et al. (2015) Concordant and discordant DNA methylation signatures of aging in human blood and brain. *Epigenetics Chromatin* 8(1):19.

60. Bock C, et al. (2012) DNA Methylation Dynamics during In Vivo Differentiation of Blood and Skin Stem Cells. *Mol Cell* 47(4):633–647.

61. McGregor K, et al. (2016) An evaluation of methods correcting for cell-type heterogeneity in DNA methylation studies. *Genome Biol* 17(1). doi:10.1186/s13059-016-0935-y.

62. Kaushal A, et al. (2017) Comparison of different cell type correction methods for genome-scale epigenetics studies. *BMC Bioinformatics* 18(1). doi:10.1186/s12859-017-1611-2.

63. Houseman EA, et al. (2012) DNA methylation arrays as surrogate measures of cell mixture distribution. *BMC Bioinformatics* 13(1):86.

64. Houseman EA, Molitor J, Marsit CJ (2014) Reference-free cell mixture adjustments in analysis of DNA methylation data. *Bioinformatics* 30(10):1431–1439.

65. Houseman EA, et al. (2016) Reference-free deconvolution of DNA methylation data and mediation by cell composition effects. *BMC Bioinformatics* 17(1). doi:10.1186/s12859-016-1140-4.

66. Rahmani E, et al. (2016) Sparse PCA corrects for cell type heterogeneity in epigenome-wide association studies. *Nat Methods* 13(5):443–445.

67. Kaushal A, Zhang H, Karmaus WJJ, Wang JSL (2015) Which methods to choose to correct cell types in genome-scale blood-derived DNA methylation data? *BMC Bioinformatics* 16(15). doi:10.1186/1471-2105-16-S15-P7.

68. Smith AK, et al. (2015) DNA extracted from saliva for methylation studies of psychiatric traits: Evidence tissue specificity and relatedness to brain. *Am J Med Genet Part B Neuropsychiatr Genet* 168(1):36–44.

69. Guintivano J, Aryee MJ, Kaminsky ZA (2013) A cell epigenotype specific model for the correction of brain cellular heterogeneity bias and its application to age, brain region and major depression. *Epigenetics* 8(3):290–302.

70. Montaño CM, et al. (2013) Measuring cell-type specific differential methylation in human brain tissue. *Genome Biol* 14(8):R94.

71. Darmanis S, et al. (2015) A survey of human brain transcriptome diversity at the single cell level. *Proc Natl Acad Sci* 112(23):7285–7290.

72. Rutter M (2010) Gene-environment interplay. *Depress Anxiety* 27(1):1–4.

73. Klengel T, et al. (2012) Allele-specific FKBP5 DNA demethylation mediates gene-childhood trauma interactions. *Nat Neurosci* 16(1):33–41.

74. Heyn H, et al. (2013) DNA methylation contributes to natural human variation. *Genome Res* 23(9):1363–1372.

75. Smith AK, et al. (2014) Methylation quantitative trait loci (meQTLs) are consistently detected across ancestry, developmental stage, and tissue type. *BMC Genomics* 15(1):145.

76. Fraser HB, Lam LL, Neumann SM, Kobor MS (2012) Population-specificity of human DNA methylation. *Genome Biol* 13(2):R8–R8.

77. Gaunt TR, et al. (2016) Systematic identification of genetic influences on methylation across the human life course. *Genome Biol* 17(1):61.

78. Zhang D, et al. (2010) Genetic Control of Individual Differences in Gene-Specific Methylation in Human Brain. *Am J Hum Genet* 86(3):411–419.

79. Gibbs JR, et al. (2010) Abundant Quantitative Trait Loci Exist for DNA Methylation and Gene Expression in Human Brain. *Plos Genet* 6(5):29.

80. Gamazon ER, et al. (2013) Enrichment of cis-regulatory gene expression SNPs and methylation quantitative trait loci among bipolar disorder susceptibility variants. *Mol Psychiatry* 18(3):340–346.

81. Hannon E, et al. (2015) Methylation QTLs in the developing brain and their enrichment in schizophrenia risk loci. *Nat Neurosci* 19(1):48–54.

82. Jaffe AE, et al. (2016) Mapping DNA methylation across development, genotype and schizophrenia in the human frontal cortex. *Nat Neurosci* 19(1):40–47.

83. Ng B, et al. (2017) An xQTL map integrates the genetic architecture of the human brain's transcriptome and epigenome. *Nat Neurosci* 20:1418.

84. Bell JT, et al. (2011) DNA methylation patterns associate with genetic and gene expression variation in HapMap cell lines. *Genome Biol* 12(1):R10.

85. Grundberg E, et al. (2013) Global Analysis of DNA Methylation Variation in Adipose Tissue from Twins Reveals Links to Disease-Associated Variants in Distal Regulatory Elements. *Am J Hum Genet* 93(5):876–890.

86. Gertz J, et al. (2011) Analysis of DNA Methylation in a Three-Generation Family Reveals Widespread Genetic Influence on Epigenetic Regulation. *PLOS Genet* 7(8):e1002228.

87. Chen L, et al. (2016) Genetic Drivers of Epigenetic and Transcriptional Variation in Human Immune Cells. *Cell* 167(5):1398–1414.e24.

88. Cheung WA, et al. (2017) Functional variation in allelic methylomes underscores a strong genetic contribution and reveals novel epigenetic alterations in the human epigenome. *Genome Biol* 18(1):50.

89. Teh AL, et al. (2014) The effect of genotype and in utero environment on interindividual variation in neonate DNA methylomes. *Genome Res* 24(7):1064–1074.

90. Sharpley CF, Palanisamy SKA, Glyde NS, Dillingham PW, Agnew LL (2014) An update on the interaction between the serotonin transporter promoter variant (5-HTTLPR), stress and depression, plus an exploration of non-confirming findings. *Behav Brain Res* 273:89–105.

91. Yokochi T, Robertson KD (2002) Preferential Methylation of Unmethylated DNA by Mammalian de NovoDNA Methyltransferase Dnmt3a. *J Biol Chem* 277(14):11735–11745.

92. Guo JU, et al. (2014) Distribution, recognition and regulation of non-CpG methylation in the adult mammalian brain. *Nat Neurosci* 17(2):215–222.

93. Ramsahoye BH, et al. (2000) Non-CpG methylation is prevalent in embryonic stem cells and may be mediated by DNA methyltransferase 3a. *Proc Natl Acad Sci* 97(10):5237–5242.

94. Lister R, et al. (2009) Human DNA methylomes at base resolution show widespread epigenomic differences. *Nature* 462(7271):315–322.

95. Laurent L, et al. (2010) Dynamic changes in the human methylome during differentiation. *Genome Res* 20(3):320–331.

96. Ziller MJ, et al. (2011) Genomic Distribution and Inter-Sample Variation of Non-CpG Methylation across Human Cell Types. *PloS Genet* 7(12):e1002389-15.

97. Lister R, et al. (2011) Hotspots of aberrant epigenomic reprogramming in human induced pluripotent stem cells. *Nature* 471(7336):68.

98. Lister R, et al. (2013) Global Epigenomic Reconfiguration During Mammalian Brain Development. *Science (80-)* 341(6146):1237905.

99. Kinde B, Gabel HW, Gilbert CS, Griffith EC, Greenberg ME (2015) Reading the unique DNA methylation landscape of the brain: Non-CpG methylation, hydroxymethylation, and MeCP2. *Proc Natl Acad Sci* 112(22):6800–6806.

100. Meyer KD, et al. (2012) Comprehensive Analysis of mRNA Methylation Reveals Enrichment in 3′ UTRs and near Stop Codons. *Cell* 149(7):1635–1646.

101. Dominissini D, et al. (2013) Topology of the human and mouse m6A RNA methylomes revealed by m6A-seq. *Nature* 485(7397):201–206.

102. Koziol MJ, et al. (2015) Identification of methylated deoxyadenosines in vertebrates reveals diversity in DNA modifications. *Nat Struct Mol Biol* 23(1):24–30.

103. Meyer KD, Jaffrey SR (2016) Expanding the diversity of DNA base modifications with N6-methyldeoxyadenosine. *Genome Biol*:1–4.

104. Ooi SKT, Bestor TH (2008) The colorful history of active DNA demethylation. *Cell* 133(7):1145–1148.

105. Wu H, Zhang Y (2014) Reversing DNA Methylation: Mechanisms, Genomics, and Biological Functions. *Cell* 156(1–2):45–68.

106. Tognini P, Napoli D, Pizzorusso T (2015) Dynamic DNA methylation in the brain: a new epigenetic mark for experience-dependent plasticity. *Front Cell Neurosci* 9:611–671.

107. Ma DK, et al. (2009) Neuronal Activity-Induced Gadd45b Promotes Epigenetic DNA Demethylation and Adult Neurogenesis. *Science (80-)* 323(5917):1074–1077.

108. Bhutani N, Burns DM, Blau HM (2011) DNA Demethylation Dynamics. *Cell* 146(6):866–872.

109. Tahiliani M, et al. (2009) Conversion of 5-Methylcytosine to 5-Hydroxymethylcytosine in Mammalian DNA by MLL Partner TET1. *Science (80-)* 324(5929):930–935.

110. Santiago M, Antunes C, Guedes M, Sousa N, Marques CJ (2014) TET enzymes and DNA hydroxymethylation in neural development and function — How critical are they? *Genomics* 104(5):334–340.

111. Ito S, et al. (2010) Role of Tet proteins in 5mC to 5hmC conversion, ES-cell self-renewal and inner cell mass specification. *Nature* 466(7310):1129–1133.

112. He YF, et al. (2011) Tet-mediated formation of 5-carboxylcytosine and its excision by TDG in mammalian DNA. *Science (80-)*. Available at: http://science.sciencemag.org/content/333/6047/1303.short.

113. Nabel CS, et al. (2012) AID/APOBEC deaminases disfavor modified cytosines implicated in DNA demethylation. *Nat Chem Biol* 8(9):751–758.

114. Sadakierska-Chudy A, Kostrzewa RM, Filip M (2014) A Comprehensive View of the Epigenetic Landscape Part I: DNA Methylation, Passive and Active DNA Demethylation Pathways and Histone Variants. *Neurotox Res* 27(1):84–97.

115. Pfeifer GP, Kadam S, Jin S-G (2013) 5-hydroxymethylcytosine and its potential roles in development and cancer. *Epigenetics Chromatin* 6(1):10.

116. Kriaucionis S, Heintz N (2009) The nuclear DNA base 5-hydroxymethylcytosine is present in Purkinje neurons and the brain. *Science (80-)* 324(5929):929–930.

117. Wang T, et al. (2012) Genome-wide DNA hydroxymethylation changes are associated with neurodevelopmental genes in the developing human cerebellum. *Hum Mol Genet* 21(26):5500–5510.

118. Alaghband Y, Bredy TW, Wood MA (2016) The role of active DNA demethylation and Tet enzyme function in memory formation and cocaine action. *Neurosci Lett* 625:40–46.

119. Harrison A, Parle-McDermott A (2011) DNA Methylation: A Timeline of Methods and Applications. *Front Genet* 2:74.

120. Sun Z, Cunningham J, Slager S, Kocher J-P (2015) Base resolution methylome profiling: considerations in platform selection, data preprocessing and analysis. *Epigenomics* 7(5):813–828.

121. Yu M, et al. (2012) Tet-assisted bisulfite sequencing of 5-hydroxymethylcytosine. *Nat Protoc* 7(12):2159.

122. Booth MJ, et al. (2013) Oxidative bisulfite sequencing of 5-methylcytosine and 5-hydroxymethylcytosine. *Nat Protoc* 8(10):1841–1851.

123. Meissner A, et al. (2005) Reduced representation bisulfite sequencing for comparative high-resolution DNA methylation analysis. *Nucleic Acids Res* 33(18):5868–5877.

124. Boyle P, et al. (2012) Gel-free multiplexed reduced representation bisulfite sequencing for large-scale DNA methylation profiling. *Genome Biol* 13(10):R92–R92.

125. Bock C, et al. (2010) Quantitative comparison of genome-wide DNA methylation mapping technologies. *Nat Biotechnol* 28(10):1106–1114.

126. Kurdyukov S, Bullock M (2016) DNA methylation analysis: choosing the right method. *Biology (Basel)* 5(1):3.

127. Bibikova M, et al. (2006) High-throughput DNA methylation profiling using universal bead arrays. *Genome Res* 16(3):383–393.

128. Bibikova M, Le J, Barnes B, Saedinia-Melnyk S (2009) Genome-wide DNA methylation profiling using Infinium® assay. *Epigenomics* 1(1):177–200.

129. Bibikova M, et al. (2011) High density DNA methylation array with single CpG site resolution. *Genomics* 98(4):288–295.

130. Moran S, Arribas C, Esteller M (2016) Validation of a DNA methylation microarray for 850,000 CpG sites of the human genome enriched in enhancer sequences. *Epigenomics* 8(3):389–399.

131. Pälmke N, Santacruz D, Walter J (2011) Comprehensive analysis of DNA-methylation in mammalian tissues using MeDIP-chip. *Methods* 53(2):175–184.

132. Taiwo O, et al. (2012) Methylome analysis using MeDIP-seq with low DNA concentrations. *Nat Protoc* 7(4):617–636.

133. Nair SS, et al. (2011) Comparison of methyl-DNA immunoprecipitation (MeDIP) and methyl-CpG binding domain (MBD) protein capture for genome-wide DNA methylation analysis reveal CpG sequence coverage bias. *Epigenetics* 6(1):34–44.

134. Zahir FR, Brown CJ (2011) Epigenetic Impacts on Neurodevelopment: Pathophysiological Mechanisms and Genetic Modes of Action. *Pediatr Res* 69(5):92R–100R.

135. Amir RE, et al. (1999) Rett syndrome is caused by mutations in X-linked MECP2, encoding methyl-CpG-binding protein 2. *Nat Genet* 23:185–188.

136. Hansen RS, et al. (1999) The DNMT3B DNA methyltransferase gene is mutated in the ICF

immunodeficiency syndrome. *Proc Natl Acad Sci U S A* 96(25):14412–14417.

137. Ehrlich M, Jackson K, Weemaes C (2006) Immunodeficiency, centromeric region instability, facial anomalies syndrome (ICF). *Orphanet J Rare Dis* 1:2.

138. Knoll JH, et al. (1989) Angelman and Prader-Willi syndromes share a common chromosome 15 deletion but differ in parental origin of the deletion. *Am J Med Genet* 32(2):285–290.

139. Karin B (2010) Prader–Willi syndrome and Angelman syndrome. *Am J Med Genet Part C Semin Med Genet* 154C(3):365–376.

140. Rapin I (1997) Autism. *N Engl J Med* 337(2):97–104.

141. Samaco RC, Hogart A, LaSalle JM (2004) Epigenetic overlap in autism-spectrum neurodevelopmental disorders: MECP2 deficiency causes reduced expression of UBE3A and GABRB3. *Hum Mol Genet* 14(4):483–492.

142. Lintas C, Sacco R, Persico AM (2016) Differential methylation at the RELN gene promoter in temporal cortex from autistic and typically developing post-puberal subjects. *J Neurodev Disord*:1–12.

143. Nagarajan RP, Hogart AR, Gwye Y, Martin MR, LaSalle JM (2014) Reduced MeCP2 Expression is Frequent in Autism Frontal Cortex and Correlates with Aberrant MECP2 Promoter Methylation. *Epigenetics* 1(4):172–182.

144. Nardone S, et al. (2014) DNA methylation analysis of the autistic brain reveals multiple dysregulated biological pathways. *Transl Psychiatry* 4(9):e433.

145. Ladd-Acosta C, et al. (2014) Common DNA methylation alterations in multiple brain regions in autism. *Mol Psychiatry* 19(8):862–871.

146. Laufer BI, et al. (2015) Associative DNA methylation changes in children with prenatal alcohol exposure. *Epigenomics* 7(August):1–16.

147. Portales-Casamar E, et al. (2016) DNA methylation signature of human fetal alcohol spectrum disorder. *Epigenetics Chromatin* 9(25):81–101.

148. Berko ER, et al. (2014) Mosaic epigenetic dysregulation of ectodermal cells in autism spectrum disorder. *PloS Genet* 10(5):e1004402–e1004402.

149. Radtke KM, et al. (2015) Epigenetic modifications of the glucocorticoid receptor gene are associated with the vulnerability to psychopathology in childhood maltreatment. *Transl Psychiatry* 5:e571.

150. Unternaehrer E, et al. (2015) Childhood maternal care is associated with DNA methylation of the genes for brain-derived neurotrophic factor (BDNF) and oxytocin receptor (OXTR) in peripheral blood cells in adult men and women. *Stress* 18(4):451–461.

151. Naumova O, et al. (2017) *Aberrant DNA Methylation in Lymphocytes of Children with Neurodevelopmental Disorders* doi:10.1134/S1022795417110072.

152. Hodyl NA, Roberts CT, Bianco-Miotto T (2016) Cord Blood DNA Methylation Biomarkers for Predicting Neurodevelopmental Outcomes. *Genes (Basel)* 7(12):117.

153. Lester BM, Marsit CJ (2018) Epigenetic mechanisms in the placenta related to infant neurodevelopment. *Epigenomics* 10(3):321–333.

154. Nestler EJ, et al. (2016) Epigenetic Basis of Mental Illness. *Neurosci* 22(5):447–463.

155. Bredy TW, Sun YE, Kobor MS (2010) How the epigenome contributes to the development of

psychiatric disorders. *Dev Psychobiol* 52(4):331–342.

156. Connor CM, Akbarian S (2008) DNA methylation changes in schizophrenia and bipolar disorder. *Epigenetics* 3(2):55–58.

157. Fullard JF, et al. (2016) Understanding the genetic liability to schizophrenia through the neuroepigenome. *Schizophr Res*:1–10.

158. Klengel T, Binder EB (2015) Epigenetics of Stress-Related Psychiatric Disorders and Gene × Environment Interactions. *Neuron* 86(6):1343–1357.

159. Abdolmaleky HM, et al. (2006) Hypomethylation of MB-COMT promoter is a major risk factor for schizophrenia and bipolar disorder. *Hum Mol Genet* 15(21):3132–3145.

160. Grayson DR, et al. (2005) Reelin promoter hypermethylation in schizophrenia. *Proc Natl Acad Sci U S A* 102(26):9341–9346.

161. Kaminsky Z, et al. (2012) A multi-tissue analysis identifies HLA complex group 9 gene methylation differences in bipolar disorder. *Mol Psychiatry* 17(7):728–740.

162. Pal M, et al. (2015) High Precision DNA Modification Analysis of HCG9in Major Psychosis. *Schizophr Bull*:sbv079-8.

163. Abdolmaleky HM, et al. (2005) Hypermethylation of the reelin (RELN) promoter in the brain of schizophrenic patients: A preliminary report. *Am J Med Genet Part B Neuropsychiatr Genet* 134B(1):60–66.

164. Abdolmaleky HM, et al. (2011) Epigenetic dysregulation of HTR2A in the brain of patients with schizophrenia and bipolar disorder. *Schizophr Res* 129(2):183–190.

165. Carrard A, Salzmann A, Malafosse A, Karege F (2011) Increased DNA methylation status of the serotonin receptor 5HTR1A gene promoter in schizophrenia and bipolar disorder. *J Affect Disord* 132(3):450–453.

166. Kaminsky Z, et al. (2015) DNA methylation and expression of KCNQ3 in bipolar disorder. *Bipolar Disord* 17(2):150–159.

167. Kordi-Tamandani DM, Sahranavard R, Torkamanzehi A (2012) DNA methylation and expression profiles of the brain-derived neurotrophic factor (BDNF) and dopamine transporter (DAT1) genes in patients with schizophrenia. *Mol Biol Rep* 39(12):10889–10893.

168. Jin L, Yoshida T, Ho R, Owens GK, Somlyo A V (2009) The actin-associated protein Palladin is required for development of normal contractile properties of smooth muscle cells derived from embryoid bodies. *J Biol Chem* 284(4):2121–2130.

169. Tamura Y, Kunugi H, Ohashi J, Hohjoh H (2007) Epigenetic aberration of the human REELIN gene in psychiatric disorders. *Mol Psychiatry* 12(6):593–600.

170. Tochigi M, et al. (2008) Methylation Status of the Reelin Promoter Region in the Brain of Schizophrenic Patients. *Biol Psychiatry* 63(5):530–533.

171. Dempster EL, Mill J, Craig IW, Collier DA (2006) The quantification of COMT mRNA in post mortem cerebellum tissue: diagnosis, genotype, methylation and expression. *BMC Med Genet* 7:10.

172. Mill J, et al. (2008) Epigenomic Profiling Reveals DNA-Methylation Changes Associated with Major Psychosis. *Am J Hum Genet* 82(3):696–711.

173. Dong E, Ruzicka WB, Grayson DR, Guidotti A (2015) DNA-methyltransferase1 (DNMT1) binding to CpG rich GABAergic and BDNF promoters is increased in the brain of schizophrenia

and bipolar disorder patients. *Schizophr Res* 167(0):35–41.

174. Chen C, et al. (2014) Correlation between DNA methylation and gene expression in the brains of patients with bipolar disorder and schizophrenia. *Bipolar Disord* 16(8):790–799.

175. Pidsley R, et al. (2014) Methylomic profiling of human brain tissue supports a neurodevelopmental origin for schizophrenia. *Genome Biol* 15(10):411–473.

176. Wockner LF, et al. (2014) Genome-wide DNA methylation analysis of human brain tissue from schizophrenia patients. *Transl Psychiatry* 4(1):e339-8.

177. Numata S, Ye T, Herman M, Lipska BK (2014) DNA methylation changes in the postmortem dorsolateral prefrontal cortex of patients with schizophrenia. *Front Genet* 5:280.

178. Ruzicka W, Subburaju S, FM B (2015) Circuit- and diagnosis-specific dna methylation changes at γ-aminobutyric acid–related genes in postmortem human hippocampus in schizophrenia and bipolar disorder. *JAMA Psychiatry* 72(6):541–551.

179. Gagliano SA, et al. (2016) REPOR T Allele-Skewed DNA Modification in the Brain: Relevance to a Schizophrenia GWAS. *Am J Hum Genet* 98(5):956–962.

180. Sabunciyan S, et al. (2012) Genome-Wide DNA Methylation Scan in Major Depressive Disorder. *PLoS One* 7(4):e34451-9.

181. Nemoda Z, et al. (2015) Maternal depression is associated with DNA methylation changes in cord blood T lymphocytes and adult hippocampi. *Transl Psychiatry* 5(4):e545–e545.

182. Keller S, et al. (2011) TrkB gene expression and DNA methylation state in Wernicke area does not associate with suicidal behavior. *J Affect Disord* 135(1–3):400–404.

183. Gross JA, et al. (2017) Gene-body 5-hydroxymethylation is associated with gene expression changes in the prefrontal cortex of depressed individuals. *Transl Psychiatry* 7(5):e1119-8.

184. Wang F, Xu H, Zhao H, Gelernter J, Zhang H (2016) DNA co-methylation modules in postmortem prefrontal cortex tissues of European Australians with alcohol use disorders. *Sci Rep* 6:19430.

185. Taqi MM, et al. (2011) Prodynorphin CpG-SNPs associated with alcohol dependence: elevated methylation in the brain of human alcoholics. *Addict Biol* 16(3):499–509.

186. Kim D, et al. (2016) Psychological factors and DNA methylation of genes related to immune/inflammatory system markers: the VA Normative Aging Study. *BMJ Open* 6(1). Available at: http://bmjopen.bmj.com/content/6/1/e009790.abstract.

187. Kahl KG, et al. (2016) Altered DNA methylation of glucose transporter 1 and glucose transporter 4 in patients with major depressive disorder. *J Psychiatr Res* 76:66–73.

188. Dempster EL, et al. (2014) Genome-wide Methylomic Analysis of Monozygotic Twins Discordant for Adolescent Depression. *Biol Psychiatry* 76(12):977–983.

189. Montano C, et al. (2016) Association of DNA methylation differences with schizophrenia in an epigenome-wide association study. *JAMA Psychiatry* 73(5):506–514.

190. Hannon E, et al. (2016) An integrated genetic-epigenetic analysis of schizophrenia: evidence for co-localization of genetic associations and differential DNA methylation. *Genome Biol* 17(1):176.

191. Cravo M, et al. (1997) DNA methylation and subclinical vitamin deficiency of folate, pyridoxal-phosphate and vitamin B12 in chronic alcoholics. *Clin Nutr* 16(1):29–35.

192. Hillemacher T, et al. (2009) Promoter specific methylation of the dopamine transporter gene is

altered in alcohol dependence and associated with craving. *J Psychiatr Res* 43(4):388–392.

193. Hillemacher T, et al. (2009) Epigenetic regulation and gene expression of vasopressin and atrial natriuretic peptide in alcohol withdrawal. *Psychoneuroendocrinology* 34(4):555–560.

194. Stefan B, et al. (2006) Epigenetic DNA Hypermethylation of the HERP Gene Promoter Induces Down-regulation of Its mRNA Expression in Patients With Alcohol Dependence. *Alcohol Clin Exp Res* 30(4):587–591.

195. Bönsch D, Lenz B, Kornhuber J, Bleich S (2005) DNA hypermethylation of the alpha synuclein promoter in patients with alcoholism. *Neuroreport* 16(2):167–170.

196. Gaudet F, et al. (2003) Induction of Tumors in Mice by Genomic Hypomethylation. *Science (80-)* 300(5618):489 LP-492.

197. Eden A, Gaudet F, Waghmare A, Jaenisch R (2003) Chromosomal Instability and Tumors Promoted by DNA Hypomethylation. *Science (80-)* 300(5618):455 LP-455.

198. Jones PA, Baylin SB (2002) The fundamental role of epigenetic events in cancer. *Nat Rev Genet* 3:415.

199. Herman JG, Baylin SB (2003) Gene Silencing in Cancer in Association with Promoter Hypermethylation. *N Engl J Med* 349(21):2042–2054.

200. Nebbioso A, Tambaro FP, Dell'Aversana C, Altucci L (2018) Cancer epigenetics: Moving forward. *PLOS Genet* 14(6):e1007362.

201. Zhou W, et al. (2018) DNA methylation loss in late-replicating domains is linked to mitotic cell division. *Nat Genet* 50(4):591–602.

202. Feinberg AP, Ohlsson R, Henikoff S (2006) The epigenetic progenitor origin of human cancer. *Nat Rev Genet* 7:21.

203. Feinberg AP, Koldobskiy MA, Göndör A (2016) Epigenetic modulators, modifiers and mediators in cancer aetiology and progression. *Nat Rev Genet* 17:284.

204. Mack SC, Hubert CG, Miller TE, Taylor MD, Rich JN (2015) An epigenetic gateway to brain tumor cell identity. *Nat Neurosci* 19:10.

205. Hovestadt V, et al. (2014) Decoding the regulatory landscape of medulloblastoma using DNA methylation sequencing. *Nature* 510:537.

206. Henrich KO, et al. (2016) Integrative genome-scale analysis identifies epigenetic mechanisms of transcriptional deregulation in unfavorable neuroblastomas. *Cancer Res* 76(18):5523–5537.

207. Olsson M, Beck S, Kogner P, Martinsson T, Carén H (2016) Genome-wide methylation profiling identifies novel methylated genes in neuroblastoma tumors. *Epigenetics* 11(1):74–84.

208. Gómez S, et al. (2015) DNA methylation fingerprint of neuroblastoma reveals new biological and clinical insights. *Genomics Data* 5:360–363.

209. Mack SC, et al. (2014) Epigenomic alterations define lethal CIMP-positive ependymomas of infancy. *Nature* 506:445.

210. Fouse SD, Costello JF (2009) Epigenetics of neurological cancers. *Future Oncol* 5(10):1615–1629.

211. Noushmehr H, et al. (2010) Identification of a CpG Island Methylator Phenotype that Defines a Distinct Subgroup of Glioma. *Cancer Cell* 17(5):510–522.

212. Paul Y, Mondal B, Patil V, Somasundaram K (2017) DNA methylation signatures for 2016 WHO

classification subtypes of diffuse gliomas. *Clin Epigenetics* 9(1):32.

213. Cadieux B, Ching T-TT, VandenBerg SR, Costello JF (2006) Genome-wide hypomethylation in human glioblastomas associated with specific copy number alteration, methylenetetrahydrofolate reductase allele status, and increased proliferation. *Cancer Res* 66(17):8469–8476.

214. Martinez R, et al. (2009) A microarray-based DNA methylation study of glioblastoma multiforme. *Epigenetics* 4(4):255–264.

215. Baeza N, Weller M, Yonekawa Y, Kleihues P, Ohgaki H (2003) *PTEN methylation and expression in glioblastomas* doi:10.1007/s00401-003-0748-4.

216. Nakamura M, Yonekawa Y, Kleihues P, Ohgaki H (2001) Promoter hypermethylation of the RB1 gene in glioblastomas. *Lab Investig* 81(1):77–82.

217. Fei G, et al. (2008) PDCD4 gene silencing in gliomas is associated with 5′CpG island methylation and unfavourable prognosis. *J Cell Mol Med* 13(10):4257–4267.

218. Zhang Z, et al. (2008) Promoter Hypermethylation-mediated Inactivation of LRRC4 in Gliomas. *BMC Mol Biol* 9(1):99.

219. Yi JM, et al. (2008) Abnormal DNA Methylation of CD133 in Colorectal and Glioblastoma Tumors. *Cancer Res* 68(19):8094–8103.

220. Mueller W, et al. (2007) Downregulation of RUNX3 and TES by hypermethylation in glioblastoma. *Oncogene* 26(4):583–593.

221. Pangeni RP, et al. (2018) Genome-wide methylomic and transcriptomic analyses identify subtype-specific epigenetic signatures commonly dysregulated in glioma stem cells and glioblastoma. *Epigenetics*. doi:10.1080/15592294.2018.1469892.

222. Esteller M, et al. (2000) Inactivation of the DNA-Repair Gene *MGMT* and the Clinical Response of Gliomas to Alkylating Agents. *N Engl J Med* 343(19):1350–1354.

223. Hegi ME, et al. (2004) Clinical Trial Substantiates the Predictive Value of O-6-Methylguanine-DNA Methyltransferase Promoter Methylation in Glioblastoma Patients Treated with Temozolomide. *Clin Cancer Res* 10(6):1871–1874.

224. Hegi ME, et al. (2005) *MGMT* Gene Silencing and Benefit from Temozolomide in Glioblastoma. *N Engl J Med* 352(10):997–1003.

225. Wick W, et al. (2014) MGMT testing—the challenges for biomarker-based glioma treatment. *Nat Rev Neurol* 10:372.

226. Sturm D, et al. (2012) Hotspot Mutations in H3F3A and IDH1 Define Distinct Epigenetic and Biological Subgroups of Glioblastoma. *Cancer Cell* 22(4):425–437.

227. Sturm D, et al. (2014) Paediatric and adult glioblastoma: multiform (epi)genomic culprits emerge. *Nat Rev Cancer* 14(2):92–107.

228. Capper D, et al. (2018) DNA methylation-based classification of central nervous system tumours. *Nature* 555:469.

229. Landgrave-Gómez J, Mercado-Gómez O, Guevara-Guzmán R (2015) Epigenetic mechanisms in neurological and neurodegenerative diseases . *Front Cell Neurosci* 9:58.

230. Lee J, Hwang YJ, Kim KY, Kowall NW, Ryu H (2013) Epigenetic Mechanisms of Neurodegeneration in Huntington's Disease. *Neurotherapeutics* 10(4):664–676.

231. Villar-Menéndez I, et al. (2013) Increased 5-Methylcytosine and Decreased 5-

Hydroxymethylcytosine Levels are Associated with Reduced Striatal A2AR Levels in Huntington's Disease. *Neuromolecular Med* 15(2):295–309.

232. Flanagan JM, et al. (2006) Intra- and Interindividual Epigenetic Variation in Human Germ Cells. *Am J Hum Genet* 79(1):67–84.

233. Reik W, Maher ER, Morrison PJ, Harding AE, Simpson SA (1993) Age at onset in Huntington's disease and methylation at D4S95. *J Med Genet* 30(3):185–188.

234. Horvath S, et al. (2016) Huntington' s disease accelerates epigenetic aging of human brain and disrupts DNA methylation levels. 8(7):1496–1523.

235. Mastroeni D, McKee A, Grover A, Rogers J, Coleman PD (2009) Epigenetic Differences in Cortical Neurons from a Pair of Monozygotic Twins Discordant for Alzheimer's Disease. *PLoS One* 4(8):e6617–e6617.

236. Chouliaras L, et al. (2013) Consistent decrease in global DNA methylation and hydroxymethylation in the hippocampus of Alzheimer's disease patients. *Neurobiol Aging* 34(9):2091–2099.

237. Lashley T, et al. (2015) Alterations in global DNA methylation and hydroxymethylation are not detected in Alzheimer's disease. *Neuropathol Appl Neurobiol* 41(4):497–506.

238. Coppieters N, et al. (2014) Global changes in DNA methylation and hydroxymethylation in Alzheimer's disease human brain. *Neurobiol Aging* 35(6):1334–1344.

239. Bradley-Whitman MA, Lovell MA (2013) Epigenetic changes in the progression of Alzheimer's disease. *Mech Ageing Dev* 134(10):10.1016/j.mad.2013.08.005-10.1016/j.mad.2013.08.00.

240. Condliffe D, et al. (2014) Cross-region reduction in 5-hydroxymethylcytosine in Alzheimer's disease brain. *Neurobiol Aging* 35(8):1850–1854.

241. Silva PN, et al. (2014) Analysis of HSPA8 and HSPA9 mRNA expression and promoter methylation in the brain and blood of Alzheimer's disease patients. *J Alzheimer's Dis* 38(1):165–170.

242. Iwata A, et al. (2014) Altered CpG methylation in sporadic Alzheimer's disease is associated with APP and MAPT dysregulation. *Hum Mol Genet* 23(3):648–656.

243. Yu L, et al. (2015) Association of Brain DNA Methylation in SORL1, ABCA7, HLA-DRB5, SLC24A4, and BIN1With Pathological Diagnosis of Alzheimer Disease. *JAMA Neurol* 72(1):10–15.

244. Foraker J, et al. (2015) The APOE Gene is Differentially Methylated in Alzheimer's Disease. *J Alzheimer's Dis* 48(3):745–755.

245. Siegmund KD, et al. (2007) DNA Methylation in the Human Cerebral Cortex Is Dynamically Regulated throughout the Life Span and Involves Differentiated Neurons. *PLoS One* 2. doi:10.1371/journal.pone.0000895.

246. Barrachina M, Ferrer I (2009) DNA Methylation of Alzheimer Disease and Tauopathy-Related Genes in Postmortem Brain. *J Neuropathol Exp Neurol* 68(8):880–891.

247. Chibnik LB, et al. (2015) Alzheimer's loci: epigenetic associations and interaction with genetic factors. *Ann Clin Transl Neurol* 2(6):636–647.

248. Watson CT, et al. (2016) Genome-wide DNA methylation profiling in the superior temporal gyrus reveals epigenetic signatures associated with Alzheimer's disease. *Genome Med* 8:5.

249. De Jager PL, et al. (2014) Alzheimery's disease pathology is associated with early alterations in

brain DNA methylation at ANK1, BIN1, RHBDF2 and other loci. *Nat Neurosci* 17(9):1156–1163.

250. Lunnon K, et al. (2014) Cross-tissue methylomic profiling strongly implicates a role for cortex-specific deregulation of ANK1 in Alzheimer's disease neuropathology. *Nat Neurosci* 17(9):1164–1170.

251. Chang L, et al. (2014) Elevation of Peripheral BDNF Promoter Methylation Links to the Risk of Alzheimer's Disease. *PLoS One* 9(11):e110773.

252. Nagata T, et al. (2015) Association between DNA methylation of the BDNF promoter region and clinical presentation in Alzheimer's disease. *Dement Geriatr Cogn Dis Extra* 5(1):64–73.

253. Tannorella P, et al. (2015) Methylation analysis of multiple genes in blood DNA of Alzheimer's disease and healthy individuals. *Neurosci Lett* 600:143–147.

254. Jowaed A, Schmitt I, Kaut O, Wüllner U (2010) Methylation regulates alpha-synuclein expression and is decreased in Parkinson's disease patients' brains. *J Neurosci* 30(18):6355 LP-6359.

255. Matsumoto L, et al. (2010) CpG Demethylation Enhances Alpha-Synuclein Expression and Affects the Pathogenesis of Parkinson's Disease. *PLoS One* 5(11):e15522–e15522.

256. de Boni L, et al. (2015) DNA methylation levels of α-synuclein intron 1 in the aging brain. *Neurobiol Aging* 36(12):3334.e7-3334.e11.

257. Pieper HC, et al. (2008) Different methylation of the TNF-alpha promoter in cortex and substantia nigra: Implications for selective neuronal vulnerability. *Neurobiol Dis* 32(3):521–527.

258. Desplats P, et al. (2011) Alpha-Synuclein Sequesters Dnmt1 from the Nucleus: a novel mechanism for epigenetic alterations in Lewy body diseases. *J Biol Chem* 286(11):1–8.

259. (IPDGC) IPDGC, (WTCCC2) WTCCC 2 (2011) A Two-Stage Meta-Analysis Identifies Several New Loci for Parkinson's Disease. *PLoS Genet* 7(6):e1002142–e1002142.

260. Kaut O, Schmitt I, Wüllner U (2012) Genome-scale methylation analysis of Parkinson's disease patients' brains reveals DNA hypomethylation and increased mRNA expression of cytochrome P450 2E1. *Neurogenetics* 13(1):87–91.

261. Masliah E, Dumaop W, Galasko D, Desplats P (2013) Distinctive patterns of DNA methylation associated with Parkinson disease: Identification of concordant epigenetic changes in brain and peripheral blood leukocytes. *Epigenetics* 8(10):1030–1038.

262. Sanchez-Mut J V., et al. (2016) Human DNA methylomes of neurodegenerative diseases show common epigenomic patterns. *Transl Psychiatry* 6(1):e718-8.

263. Wey H-Y, et al. (2016) Insights into neuroepigenetics through human histone deacetylase PET imaging. *Sci Transl Med* 8(351):351ra106 LP-351ra106.

264. Samarasekera N, et al. (2018) Brain banking for neurological disorders. *Lancet Neurol* 12(11):1096–1105.

265. Pidsley R, Mill J (2011) Epigenetic Studies of Psychosis: Current Findings, Methodological Approaches, and Implications for Postmortem Research. *BPS* 69(2):146–156.

266. Palmer-Aronsten B, Sheedy D, McCrossin T, Kril J (2016) An International Survey of Brain Banking Operation and Characterization Practices. *Biopreserv Biobank* 14(6):464–469.

267. Harrison PJ, et al. (1995) The relative importance of premortem acidosis and postmortem interval for human brain gene expression studies: selective mRNA vulnerability and comparison with their encoded proteins. *Neurosci Lett* 200(3):151–154.

268. Tomita H, et al. (2004) Effect of agonal and postmortem factors on gene expression profile: Quality control in microarray analyses of postmortem human brain. *Biol Psychiatry* 55(4):346–352.

269. Li JZ, et al. (2004) Systematic changes in gene expression in postmortem human brains associated with tissue pH and terminal medical conditions. *Hum Mol Genet* 13(6):609–616.

270. Nagy C, et al. (2015) Effects of postmortem interval on biomolecule integrity in the brain. *J Neuropathol Exp Neurol* 74(5):459–469.

271. Gibbons HM, Dragunow M (2010) Adult human brain cell culture for neuroscience research. *Int J Biochem Cell Biol* 42(6):844–856.

272. Allen DD, et al. (2005) Cell lines as in vitro models for drug screening and toxicity studies. *Drug Dev Ind Pharm* 31(8):757–768.

273. Harry GJ, Tiffany-Castiglioni E (2005) Evaluation of neurotoxic potential by use of in vitro systems. *Expert Opin Drug Metab Toxicol* 1(4):701–13.

274. Reynolds BA, Weiss S (1992) Generation of neurons and astrocytes from isolated cells of the adult mammalian central nervous system. *Science (80-)* 255(5052):1707 LP-1710.

275. Morshead CM, et al. (1994) Neural stem cells in the adult mammalian forebrain: A relatively quiescent subpopulation of subependymal cells. *Neuron* 13(5):1071–1082.

276. Weiss S, et al. (1996) Multipotent CNS stem cells are present in the adult mammalian spinal cord and ventricular neuroaxis. *J Neurosci* 16(23):7599–609.

277. Reynolds BA, Rietze RL (2005) Neural stem cells and neurospheres—re-evaluating the relationship. *Nat Methods* 2(5):333–336.

278. Singh SK, et al. (2004) Identification of human brain tumour initiating cells. *Nature* 432(7015):396–401.

279. Galli R, et al. (2004) Isolation and characterization of tumorigenic, stem-like neural precursors from human glioblastoma. *Cancer Res* 64(19):7011–7021.

280. Lee J, et al. (2006) Tumor stem cells derived from glioblastomas cultured in bFGF and EGF more closely mirror the phenotype and genotype of primary tumors than do serum-cultured cell lines. *Cancer Cell* 9(5):391–403.

281. Vescovi AL, Galli R, Reynolds BA (2006) Brain tumour stem cells. *Nat Rev Cancer* 6(6):425–436.

282. Rahman M, et al. (2015) Neurosphere and adherent culture conditions are equivalent for malignant glioma stem cell lines. *Anat Cell Biol* 48(1):25–35.

283. Lancaster MA, et al. (2013) Cerebral organoids model human brain development and microcephaly. *Nature* 501:373.

284. Kelava I, Lancaster MA (2016) Dishing out mini-brains: Current progress and future prospects in brain organoid research. *Dev Biol* 420(2):199–209.

285. Luo C, et al. (2016) Cerebral Organoids Recapitulate Epigenomic Signatures of the Human Fetal Brain. *Cell Rep* 17(12):3369–3384.

286. Varley KE, et al. (2013) Dynamic DNA methylation across diverse human cell lines and tissues. 555–567.

287. Quail DF, Joyce JA (2018) The Microenvironmental Landscape of Brain Tumors. *Cancer Cell*

31(3):326–341.

288. Di Giorgio FP, Boulting GL, Bobrowicz S, Eggan KC (2008) Human Embryonic Stem Cell-Derived Motor Neurons Are Sensitive to the Toxic Effect of Glial Cells Carrying an ALS-Causing Mutation. *Cell Stem Cell* 3(6):637–648.

289. Keverne EB, Pfaff DW, Tabansky I (2015) Epigenetic changes in the developing brain: Effects on behavior. *Proc Natl Acad Sci* 112(22):6789–6795.

290. Bock C (2009) Epigenetic biomarker development. *Epigenomics* 1(1):99–110.

291. Davies MN, et al. (2012) Functional annotation of the human brain methylome identifies tissue-specific epigenetic variation across brain and blood. *Genome Biol* 13(6):R43.

292. Edgar RD, Jones MJ, Meaney MJ, Turecki G, Kobor MS (2017) BECon: A tool for interpreting DNA methylation findings from blood in the context of brain. *Transl Psychiatry* 7(8):e1187-10.

293. Sugawara H, et al. (2011) Hypermethylation of serotonin transporter gene in bipolar disorder detected by epigenome analysis of discordant monozygotic twins. *Transl Psychiatry* 1(7):e24-7.

294. Fisher HL, et al. (2015) Methylomic analysis of monozygotic twins discordant for childhood psychotic symptoms. *Epigenetics* 10(11):1014–1023.

295. Lowe R, et al. (2014) Buccals are likely to be a more informative surrogate tissue than blood for epigenome-wide association studies. *Epigenetics* 8(4):445–454.

296. Pringsheim T, et al. (2012) The incidence and prevalence of Huntington's disease: A systematic review and meta-analysis. *Mov Disord* 27(9):1083–1091.

297. Myers RH (2004) Huntington's disease genetics. *Neurotherapeutics* 1(2):255–262.

298. Rinaldi C, et al. (2012) Predictors of Survival in a Huntington's Disease Population from Southern Italy. *Can J Neurol Sci* 39(01):48–51.

299. MacDonald ME, et al. (1993) A novel gene containing a trinucleotide repeat that is expanded and unstable on Huntington's disease chromosomes. *Cell* 72(6):971–983.

300. Kremer B, et al. (1994) A worldwide study of the Huntington's disease mutation. The sensitivity and specificity of measuring CAG repeats. *N Engl J Med* 330(20):1401–1406.

301. Rubinsztein DC, et al. (1996) Phenotypic characterization of individuals with 30-40 CAG repeats in the Huntington disease (HD) gene reveals HD cases with 36 repeats and apparently normal elderly individuals with 36-39 repeats. *Am J Hum Genet* 59(1):16–22.

302. Stine OC, et al. (1993) Correlation between the onset age of Huntington's disease and length of the trinucleotide repeat in IT-15. *Hum Mol Genet* 2(10):1547–1549.

303. Duyao M, et al. (1993) Trinucleotide repeat length instability and age of onset in Huntington's disease. *Nat Genet* 4(4):387–392.

304. Collaborative TUS-V, et al. (2004) Venezuelan kindreds reveal that genetic and environmental factors modulate Huntington's disease age of onset. *Proc Natl Acad Sci* 101(10):3498–3503.

305. Andresen JM, et al. (2006) The relationship between CAG repeat length and age of onset differs for Huntington's disease patients with juvenile onset or adult onset. *Ann Hum Genet* 71(3):295–301.

306. Bečanović K, et al. (2015) A SNP in the HTT promoter alters NF-κB binding and is a bidirectional genetic modifier of Huntington disease. *Nat Neurosci* 18(6):807–816.

307. Van Raamsdonk JM, et al. (2007) Testicular degeneration in Huntington disease. *Neurobiol Dis*

26(3):512–520.

308. Li S-H, et al. (1993) Huntington's disease gene (IT15) is widely expressed in human and rat tissues. *Neuron* 11(5):985–993.

309. Dixon KT, Cearley JA, Hunter JM, Detloff PJ (2004) Mouse Huntington's Disease Homolog mRNA Levels: Variation and Allele Effects. *Gene Expr* 11(5–6):221–231.

310. Zuccato C, et al. (2007) Widespread disruption of repressor element-1 silencing transcription factor/neuron-restrictive silencer factor occupancy at its target genes in Huntington's disease. *J Neurosci* 27(26):6972–6983.

311. Boutell JM, et al. (1999) Aberrant Interactions of Transcriptional Repressor Proteins with the Huntington's Disease Gene Product, Huntingtin. *Hum Mol Genet* 8(9):1647–1655.

312. Steffan JS, et al. (2000) The Huntington's disease protein interacts with p53 and CREB-binding protein and represses transcription. *Proc Natl Acad Sci* 97(12):6763–6768.

313. Shimohata T, et al. (2000) Expanded polyglutamine stretches interact with TAF(II)130, interfering with CREB-dependent transcription. *Nat Genet* 26(1):29–36.

314. Dunah AW, et al. (2002) Sp1 and TAFII130 Transcriptional Activity Disrupted in Early Huntington's Disease. *Science (80-)* 296(5576):2238–2243.

315. Hatchwell E, Greally JM (2007) The potential role of epigenomic dysregulation in complex human disease. *Trends Genet* 23(11):588–595.

316. Wan J, et al. (2015) Characterization of tissue-specific differential DNA methylation suggests distinct modes of positive and negative gene expression regulation. *BMC Genomics* 16(1):49.

317. Jiang R, et al. (2015) Discordance of DNA methylation variance between two accessible human tissues. *Sci Rep* 5(1):2877–2878.

318. Ng CW, et al. (2013) Extensive changes in DNA methylation are associated with expression of mutant huntingtin. *Proc Natl Acad Sci U S A* 110(6):2354–2359.

319. G. Vonsattel JP, DiFiglia M (1998) Huntington Disease. *J Neuropathol Exp Neurol* 57(5):369–384.

320. Hedreen JC, Peyser CE, Folstein SE, Ross CA (1991) Neuronal loss in layers V and VI of cerebral cortex in Huntington's disease. *Neurosci Lett* 133(2):257–261.

321. Langbehn DR, Brinkman RR, Falush D, Paulsen JS, Hayden MR (2004) A new model for prediction of the age of onset and penetrance for Huntington's disease based on CAG length. *Clin Genet* 65(4):267–277.

322. Horvath S (2013) DNA methylation age of human tissues and cell types. *Genome Biol* 14(10):R115–R115.

323. R Development Core Team R, R DCT (2008) *R: A Language and Environment for Statistical Computing* (R Foundation for Statistical Computing, Vienna, Austria.) doi:10.1007/978-3-540-74686-7.

324. Price EM, et al. (2013) Additional annotation enhances potential for biologically-relevant analysis of the Illumina Infinium HumanMethylation450 BeadChip array. *Epigenetics Chromatin* 6(1):1–4.

325. Du P, Kibbe WA, Lin SM (2008) lumi: a pipeline for processing Illumina microarray. *Bioinformatics* 24(13):1547–1548.

326. Maksimovic J, Gordon L, Oshlack A (2012) SWAN: Subset-quantile within array normalization

for Illumina Infinium HumanMethylation450 BeadChips. *Genome Biol* 13(6):R44–R44.

327. Johnson WE, Li C, Rabinovic A (2007) Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics* 8(1):118–127.

328. Jones MJ, et al. (2013) Distinct DNA methylation patterns of cognitive impairment and trisomy 21 in Down syndrome. *BMC Med Genomics* 6:58.

329. Jones MJ, Islam SA, Edgar RD, Kobor MS (2015) Adjusting for Cell Type Composition in DNA Methylation Data Using a Regression-Based Approach. *Methods Mol Biol*:1–8.

330. Jaffe AE, et al. (2012) Bump hunting to identify differentially methylated regions in epigenetic epidemiology studies. *Int J Epidemiol* 41(1):200–209.

331. Smyth GK (2004) Linear models and empirical Bayes methods for assessing differential expression in microarray experiments. *Stat Appl Genet Mol Biol* 3(1):Article3-Article3.

332. Benjamini Y, Hochberg Y (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J R Stat Soc Ser B* 57(1):289–300.

333. Orr M, Liu P (2009) Sample size estimation while controlling false discovery rate for microarray experiments using the ssize. fdr package. *R J* 1(May):47–53.

334. Du P, et al. (2010) Comparison of Beta-value and M-value methods for quantifying methylation levels by microarray analysis. *BMC Bioinformatics* 11(1):587.

335. Feingold EA, et al. (2004) The ENCODE (ENCyclopedia of DNA Elements) Project. *Science (80-)* 306(5696):636–640.

336. Portales-Casamar E, et al. (2007) PAZAR: A framework for collection and dissemination of cis-regulatory sequence annotation. *Genome Biol* 8(10). doi:10.1186/gb-2007-8-10-r207.

337. Mathelier A, et al. (2014) JASPAR 2014: An extensively expanded and updated open-access database of transcription factor binding profiles. *Nucleic Acids Res* 42(D1). doi:10.1093/nar/gkt997.

338. Mathelier A, et al. (2015) Cis-regulatory somatic mutations and gene-expression alteration in B-cell lymphomas. *Genome Biol* 16(1). doi:10.1186/s13059-015-0648-7.

339. Wasserman WW, Sandelin A (2004) Applied bioinformatics for the identification of regulatory elements. *Nat Rev Genet* 5(4):276–287.

340. Wiegreffe C, et al. (2015) Bcl11a (Ctip1) controls migration of cortical projection neurons through regulation of Sema3c. *Neuron* 87(2):311–325.

341. Leavitt BR, et al. (2001) Wild-type huntingtin reduces the cellular toxicity of mutant huntingtin in vivo. *Am J Hum Genet* 68(2):313–324.

342. Warby SC, et al. (2009) CAG expansion in the Huntington disease gene is associated with a specific and targetable predisposing haplogroup. *Am J Hum Genet* 84(3):351–366.

343. Reiner A, Dragatsis I, Dietrich P (2011) Genetics and neuropathology of Huntington's disease. *Int Rev Neurobiol*:325–372.

344. Hannum G, et al. (2013) Genome-wide methylation profiles reveal quantitative views of human aging rates. *Mol Cell* 49(2):359–367.

345. Jones MJ, Goodman SJ, Kobor MS (2015) DNA methylation and healthy human aging. *Aging Cell* 14(6):924–932.

346. Vandesompele J, Kubista M, Pfaffl MMW (2009) Reference gene validation software for

improved normalization. *Real-time PCR Curr Technol Apllications* 4:47–64.

347. Eckhardt F, et al. (2006) DNA methylation profiling of human chromosomes 6, 20 and 22. *Nat Genet* 38(12):1378–1385.

348. Maurano MTT, et al. (2015) Role of DNA methylation in modulating transcription factor occupancy. *Cell Rep* 12(7):1184–1195.

349. Ong C-T, Corces VG (2014) CTCF: an architectural protein bridging genome topology and function. *Nat Rev Genet* 15(4):234–246.

350. Vonsattel J-P, et al. (1985) Neuropathological classification of Huntington's disease. *J Neuropathol Exp Neurol* 44(6):559–577.

351. Wang H, et al. (2012) Widespread plasticity in CTCF occupancy linked to DNA methylation. *Genome Res* 22(9):1680–1688.

352. Tsai PC, Bell JT (2015) Power and sample size estimation for epigenome-wide association scans to detect differential DNA methylation. *Int J Epidemiol* 44(4):1429–1441.

353. Wen L, et al. (2014) Whole-genome analysis of 5-hydroxymethylcytosine and 5-methylcytosine at base resolution in the human brain. *Genome Biol* 15(3):R49-17.

354. Engel N, West AG, Felsenfeld G, Bartolomei MS (2004) Antagonism between DNA hypermethylation and enhancer-blocking activity at the H19 DMD is uncovered by CpG mutations. *Nat Genet* 36(8):883–888.

355. Stupp R, et al. (2005) Radiotherapy plus concomitant and adjuvant temozolomide for glioblastoma. *N Engl J Med* 352(10):987–996.

356. Mason WP, et al. (2007) Canadian recommendations for the treatment of glioblastoma multiforme. *Curr Oncol* 14(3):110–117.

357. Stommel JM, et al. (2007) Coactivation of receptor tyrosine kinases affects the response of tumor cells to targeted therapies. *Science (80-)*. Available at: http://science.sciencemag.org/content/early/2007/09/13/science.1142946.abstract.

358. Nathanson DA, et al. (2014) Targeted therapy resistance mediated by dynamic regulation of extrachromosomal mutant EGFR DNA. *Science (80-)* 343(6166):72 LP-76.

359. Sottoriva A, et al. (2013) Intratumor heterogeneity in human glioblastoma reflects cancer evolutionary dynamics. *Proc Natl Acad Sci U S A* 110(10):4009 LP-4014.

360. Network TCGAR (2008) Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature* 455:1061.

361. Parsons DW, et al. (2008) An Integrated Genomic Analysis of Human Glioblastoma Multiforme. *Science (80-)* 321(5897):1807 LP-1812.

362. Verhaak RGW, et al. (2010) An integrated genomic analysis identifies clinically relevant subtypes of glioblastoma characterized by abnormalities in PDGFRA, IDH1, EGFR and NF1. *Cancer Cell* 17(1):98–110.

363. Valent P, et al. (2012) Cancer stem cell definitions and terminology: the devil is in the details. *Nat Rev Cancer* 12(11):767–775.

364. Zhou B-BS, et al. (2009) Tumour-initiating cells: challenges and opportunities for anticancer drug discovery. *Nat Rev Drug Discov* 8:806.

365. Jackson M, Hassiotou F, Nowak A (2014) Glioblastoma stem-like cells: at the root of tumor

recurrence and a therapeutic target. *Carcinogenesis* 36(2):bgu243.

366. Liau BB, et al. (2017) Adaptive chromatin remodeling drives glioblastoma stem cell plasticity and drug tolerance. *Cell Stem Cell* 20(2):233–246.e7.

367. Kelly JJP, et al. (2009) Proliferation of human glioblastoma stem cells occurs independently of exogenous mitogens. *Stem Cells* 27(8):1722–1733.

368. Davis B, et al. (2016) Comparative genomic and genetic analysis of glioblastoma-derived brain tumor-initiating cells and their parent tumors. *Neuro Oncol* 18(3):350–360.

369. Lan X, et al. (2017) Fate mapping of human glioblastoma reveals an invariant stem cell hierarchy. *Nature* 549:227.

370. Jones PA, Baylin SB (2007) The epigenomics of cancer. *Cell* 128(4):683–692.

371. Heijmans BT, Mill J (2012) Commentary: The seven plagues of epigenetic epidemiology. *Int J Epidemiol* 41(1):74–78.

372. Lemire M, et al. (2015) Long-range epigenetic regulation is conferred by genetic variation located at thousands of independent loci. *Nat Commun* 6:6326.

373. Baylin SB, Jones PA (2011) A decade of exploring the cancer epigenome - biological and translational implications. *Nat Rev Cancer* 11(10):726–734.

374. Hansen KD, et al. (2011) Increased methylation variation in epigenetic domains across cancer types. *Nat Genet* 43(8):768.

375. Lai RK, et al. (2014) Genome-wide methylation analyses in glioblastoma multiforme. *PLoS One* 9(2):e89376.

376. Kurscheid S, et al. (2015) Chromosome 7 gain and DNA hypermethylation at the HOXA10 locus are associated with expression of a stem cell related HOX-signature in glioblastoma. *Genome Biol* 16(1):16.

377. Murat A, et al. (2008) Stem cell-related "self-renewal" signature and high epidermal growth factor receptor expression associated with resistance to concomitant chemoradiotherapy in glioblastoma. *J Clin Oncol* 26(18):3015–3024.

378. Gallo M, et al. (2013) A tumorigenic MLL-homeobox network in human glioblastoma stem cells. *Cancer Res* 73(1):417–427.

379. Lee E-J, et al. (2015) Identification of global DNA methylation signatures in glioblastoma-derived cancer stem cells. *J Genet Genomics* 42(7):355–371.

380. Cusulin C, et al. (2015) Precursor states of brain tumor initiating cell lines are predictive of survival in xenografts and associated with glioblastoma subtypes. *Stem Cell Reports* 5(1):1–9.

381. Nguyen SA, et al. (2014) Novel MSH6 mutations in treatment-naïve glioblastoma and anaplastic oligodendroglioma contribute to temozolomide resistance independently of MGMT promoter methylation. *Clin Cancer Res* 20(18):4894 LP-4903.

382. Hicks SC, Irizarry RA (2015) quantro: a data-driven approach to guide the choice of an appropriate normalization method. *Genome Biol*:1–8.

383. Teschendorff AE, et al. (2013) A beta-mixture quantile normalization method for correcting probe design bias in Illumina Infinium 450 k DNA methylation data. *Bioinformatics* 29(2):189–196.

384. Huang DW, Sherman BT, Lempicki RA (2008) Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat Protoc* 4(1):44–57.

385. Huang DW, Sherman BT, Lempicki RA (2009) Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Res* 37(1):1–13.

386. Ashburner M, et al. (2000) Gene Ontology: tool for the unification of biology. *Nat Genet* 25:25.

387. Esteller M, Hamilton SR, Burger PC, Baylin SB, Herman JG (1999) Inactivation of the DNA repair gene O(6)-methylguanine-DNA methyltransferase by promoter hypermethylation is a common event in primary human neoplasia. *Cancer Res* 59(4):793–797.

388. Sciuscio D, et al. (2011) Extent and patterns of MGMT promoter methylation in glioblastoma- and respective glioblastoma-derived spheres. *Clin Cancer Res* 17(2):255–266.

389. Bady P, et al. (2012) MGMT methylation analysis of glioblastoma on the Infinium methylation BeadChip identifies two distinct CpG regions associated with gene silencing and outcome, yielding a prediction model for comparisons across datasets, tumor grades, and CIMP-status. *Acta Neuropathol* 124(4):547–560.

390. Mazor T, et al. (2017) Clonal expansion and epigenetic reprogramming following deletion or amplification of mutant *IDH1*. *Proc Natl Acad Sci* 114(40):10743–10748.

391. Heyn H, et al. (2016) Epigenomic analysis detects aberrant super-enhancer DNA methylation in human cancer. *Genome Biol* 17(1):11.

392. de Souza CF, et al. (2018) A distinct DNA methylation shift in a subset of glioma CpG island methylator phenotypes during tumor recurrence. *Cell Rep* 23(2):637–651.

393. Lun ATL, Smyth GK (2014) De novo detection of differentially bound regions for ChIP-seq data using peaks and windows: controlling error rates correctly. *Nucleic Acids Res* 42(11):e95–e95.

394. Robinson MD, et al. (2014) Statistical methods for detecting differentially methylated loci and regions. *Front Genet*. doi:10.3389/fgene.2014.00324.

395. Wu L, et al. (2016) Aberrant promoter methylation of cancer-related genes in human breast cancer. *Oncol Lett* 12(6):5145–5155.

396. Slieker RC, et al. (2013) Identification and systematic annotation of tissue-specific differentially methylated regions using the Illumina 450k array. *Epigenetics Chromatin* 6(1):26.

397. Qian XC, Brent TP (1997) Methylation hot spots in the 5' flanking region denote silencing of the O6-methylguanine-DNA methyltransferase gene. *Cancer Res* 57(17):3672–3677.

398. Watts GS, et al. (1997) Methylation of discrete regions of the O6-methylguanine DNA methyltransferase (MGMT) CpG island is associated with heterochromatinization of the MGMT transcription start site and silencing of the gene. *Mol Cell Biol* 17(9):5612–5619.

399. Abdel-Fattah R, et al. (2006) Differential expression of HOX genes in neoplastic and non-neoplastic human astrocytes. *J Pathol* 209(1):15–24.

400. Gaspar N, et al. (2010) MGMT-independent temozolomide resistance in pediatric glioblastoma cells associated with a PI3-kinase-mediated HOX/stem cell gene signature. *Cancer Res* 70(22):9243–9252.

401. Costa BM, et al. (2010) Reversing HOXA9 oncogene activation by PI3K inhibition: epigenetic mechanism and prognostic significance in human glioblastoma. *Cancer Res* 70(2):453–462.

402. Pearson JC, Lemons D, McGinnis W (2005) Modulating Hox gene functions during animal body patterning. *Nat Rev Genet* 6:893.

403. Sanai N, Alvarez-Buylla A, Berger MS (2005) Neural Stem Cells and the Origin of Gliomas. *N Engl J Med* 353(8):811–822.

404. Zhang N, et al. (2015) Predicting tumor purity from methylation microarray data. *Bioinformatics* 31(21):3401–3405.

405. Meacham CE, Morrison SJ (2013) Tumour heterogeneity and cancer cell plasticity. *Nature* 501(7467):328–337.

406. Chen J, et al. (2012) A restricted cell population propagates glioblastoma growth after chemotherapy. *Nature* 488:522.

407. Li Z, et al. (2009) Hypoxia-inducible factors regulate tumorigenic capacity of glioma stem cells. *Cancer Cell* 15(6):501–513.

408. Bao S, et al. (2006) Glioma stem cells promote radioresistance by preferential activation of the DNA damage response. *Nature* 444:756.

409. Bhat KPL, et al. (2013) Mesenchymal differentiation mediated by NF-κB promotes radiation resistance in glioblastoma. *Cancer Cell* 24(3):10.1016/j.ccr.2013.08.001.

410. Patel AP, et al. (2014) Single-cell RNA-seq highlights intratumoral heterogeneity in primary glioblastoma. *Science (80-)* 344(6190):1396 LP-1401.

411. Zheng X, Zhang N, Wu H-J, Wu H (2017) Estimating and accounting for tumor purity in the analysis of DNA methylation data from cancer studies. *Genome Biol* 18(1):17.

412. Vidal E, et al. (2017) A DNA methylation map of human cancer at single base-pair resolution. *Oncogene* 36:5648.

413. Figueroa ME, et al. (2010) DNA methylation signatures identify biologically distinct subtypes in acute myeloid leukemia. *Cancer Cell* 17(1):13–27.

414. F.J. KT, et al. (2012) Low values of 5-hydroxymethylcytosine (5hmC), the "sixth base," are associated with anaplasia in human brain tumors. *Int J Cancer* 131(7):1577–1590.

415. Song C-X, et al. (2010) Selective chemical labeling reveals the genome-wide distribution of 5-hydroxymethylcytosine. *Nat Biotechnol* 29:68.

416. Orr BA, Haffner MC, Nelson WG, Yegnasubramanian S, Eberhart CG (2012) Decreased 5-hydroxymethylcytosine is associated with neural progenitor phenotype in normal brain and shorter survival in malignant glioma. *PLoS One* 7(7):e41036.

417. Antequera F, Boyes J, Bird A (1990) High levels of de novo methylation and altered chromatin structure at CpG islands in cell lines. *Cell* 62(3):503–514.

418. Meissner A, et al. (2008) Genome-scale DNA methylation maps of pluripotent and differentiated cells. *Nature* 454(7205):766–770.

419. Heyn H, et al. (2014) Linkage of DNA methylation quantitative trait loci to human cancer risk. *Cell Rep* 7(2):331–338.

420. Britton A, O'Neill D, Bell S (2016) Underestimating the alcohol content of a glass of wine: the implications for estimates of mortality risk. *Alcohol Alcohol* 51(5):609–614.

421. Young-Wolff KC, Enoch M-A, Prescott CA (2011) The influence of gene–environment interactions on alcohol consumption and alcohol use disorders: A comprehensive review. *Clin Psychol Rev* 31(5):800–816.

422. Enoch M-A (2012) The influence of gene-environment interactions on the development of alcoholism and drug dependence. *Curr Psychiatry Rep* 14(2):150–158.

423. Wall TL, Luczak SE, Hiller-Sturmhöfel S (2016) Biology, genetics, and environment: underlying

factors Influencing alcohol metabolism. *Alcohol Res* 38(1):59–68.

424. Choy M-K, et al. (2010) Genome-wide conserved consensus transcription factor binding motifs are hyper-methylated. *BMC Genomics* 11(1):519.

425. Irvine RA, Lin IG, Hsieh C-L (2002) DNA methylation has a local effect on transcription and histone acetylation. *Mol Cell Biol* 22(19):6689–6696.

426. Sandoval J, et al. (2011) Validation of a DNA methylation microarray for 450,000 CpG sites in the human genome. *Epigenetics* 6(6):692–702.

427. Varela-Rey M, Woodhoo A, Martinez-Chantar M-L, Mato JM, Lu SC (2013) Alcohol, DNA methylation, and cancer. *Alcohol Res* 35(1):25–35.

428. Jaffe AE, Irizarry RA (2014) Accounting for cellular heterogeneity is critical in epigenome-wide association studies. *Genome Biol* 15(2). doi:10.1186/gb-2014-15-2-r31.

429. Chu M, et al. (2008) Inferring relative numbers of human leucocyte genome replications. *Br J Haematol* 141(6):862–871.

430. Liu J, Morgan M, Hutchison K, Calhoun VD (2010) A Study of the influence of sex on genome wide methylation. *PLoS One* 5(4):e10028.

431. Zhang FF, et al. (2011) Significant differences in global genomic DNA methylation by gender and race/ethnicity in peripheral blood. *Epigenetics* 6(5):623–629.

432. Philibert RA, et al. (2014) A pilot examination of the genome-wide DNA methylation signatures of subjects entering and exiting short-term alcohol dependence treatment programs. *Epigenetics* 9(9):1212–1219.

433. Philibert RA, et al. (2008) The relationship of 5HTT (SLC6A4) methylation and genotype on mRNA expression and liability to major depression and alcohol dependence in subjects from the Iowa Adoption Studies. *Am J Med Genet Part B Neuropsychiatr Genet* 147B(5):543–549.

434. Zhang R, et al. (2013) Genome-wide DNA methylation analysis in alcohol dependence. *Addict Biol* 18(2):392–403.

435. Heberlein A, et al. (2011) Epigenetic down regulation of nerve growth factor during alcohol withdrawal. *Addict Biol* 18(3):508–510.

436. Ruggeri B, et al. (2015) Association of protein phosphatase PPM1G with alcohol use disorder and brain Activity during behavioral control in a genome-wide methylation analysis. *Am J Psychiatry* 172(6):543–552.

437. Zhu Z-ZZ, et al. (2010) Predictors of global methylation levels in blood DNA of healthy subjects: a combined analysis. *Int J Epidemiol* 41(1):126–139.

438. Liu C, et al. (2018) A DNA methylation biomarker of alcohol consumption. *Mol Psychiatry* 23(2):422–433.

439. Shenker NS, et al. (2013) Epigenome-wide association study in the European Prospective Investigation into Cancer and Nutrition (EPIC-Turin) identifies novel genetic loci associated with smoking. *Hum Mol Genet* 22(5):843–851.

440. Zhang H, Gelernter J (2016) Review: DNA methylation and alcohol use disorders: Progress and challenges. *Am J Addict* 26(5):502–515.

441. Brückmann C, Di Santo A, Karle KN, Batra A, Nieratschker V (2016) Validation of differential GDAP1 DNA methylation in alcohol dependence and its potential function as a biomarker for disease severity and therapy outcome. *Epigenetics* 11(6):456–463.

442. Hagerty SL, et al. (2016) An exploratory association study of alcohol use disorder and DNA methylation. *Alcohol Clin Exp Res* 40(8):1633–1640.

443. Hillemacher T, et al. (2015) DNA methylation of the LEP gene is associated with craving during alcohol withdrawal. *Psychoneuroendocrinology* 51:371–377.

444. Nieratschker V, et al. (2012) Epigenetic alteration of the dopamine transporter gene in alcohol-dependent patients is associated with age. *Addict Biol* 19(2):305–311.

445. Pasala S, Barr T, Messaoudi I (2015) Impact of alcohol abuse on the adaptive immune system. *Alcohol Res* 37(2):185–97.

446. Trannesen H, Andersen JR, Pedersen AE, Kaiser AH (1990) Lymphopenia in Heavy Drinkers - Reversibility and Relation to the Duration of Drinking Episodes. *Ann Med* 22(4):229–231.

447. Cook RT, et al. (1991) Activated CD-8 cells and HLA DR expression in alcoholics without overt liver disease. *J Clin Immunol* 11(5):246–253.

448. Porretta E, Happel KI, Teng XS, Ramsay A, Mason CM (2011) The Impact of Alcohol on BCG-Induced Immunity Against Mycobacterium tuberculosis. *Alcohol Clin Exp Res* 36(2):310–317.

449. Guillemin C, et al. (2014) DNA Methylation Signature of Childhood Chronic Physical Aggression in T Cells of Both Men and Women. *PLoS One* 9(1):e86822.

450. Saunders JB, Aasland OG, Babor TF, De La Fuente JR, Grant M (1993) Development of the Alcohol Use Disorders Identification Test (AUDIT): WHO collaborative project on early detection of persons with harmful alcohol consumption-II. *Addiction* 88(6):791–804.

451. Mercier C, et al. (1992) Profiles of alcoholics according to the SCL-90-R: a confirmative study. *Int J Addict* 27(11):1267–1282.

452. Moak DH, Anton RF, Latham PK (1998) Further validation of the obsessive-compulsive drinking scale (OCDS). *Am J Addict* 7(1):14–23.

453. Esposito EA, et al. (2016) Differential DNA methylation in peripheral blood mononuclear cells in adolescents exposed to significant early but not later childhood adversity. *Dev Psychopathol* 28(4pt2):1385–1399.

454. De Souza RAG, et al. (2016) DNA methylation profiling in human Huntington's disease brain. *Hum Mol Genet* 25(10):2013–2030.

455. Koestler DC, et al. (2013) Blood-based profiles of DNA methylation predict the underlying distribution of cell types: A validation analysis. *Epigenetics* 8(8):816–826.

456. Jones MJ, Islam SA, Edgar RD, Kobor MS (2017) Adjusting for Cell Type Composition in DNA Methylation Data Using a Regression-Based Approach. *Population Epigenetics: Methods and Protocols* (Springer New York, New York, NY), pp 99–106.

457. Peters TJ, et al. (2014) De novo identification of differentially methylated regions in the human genome. *Epigenetics Chromatin* 8(1):6.

458. Reinius LE, et al. (2012) Differential DNA methylation in purified human blood cells: implications for cell lineage and studies on disease susceptibility. *PLoS One* 7(7):e41361.

459. Miyazaki K, et al. (2003) A novel HECT-type E3 ubiquitin ligase, NEDL2, stabilizes p73 and enhances its transcriptional activity. *Biochem Biophys Res Commun* 308(1):106–113.

460. Yang A, et al. (2000) p73-deficient mice have neurological, pheromonal and inflammatory defects but lack spontaneous tumours. *Nature* 404(6773):99–103.

461. Nakagawa O (2005) Centronuclear myopathy in mice lacking a novel muscle-specific protein kinase transcriptionally regulated by MEF2. *Genes Dev* 19(17):2066–2077.

462. Nieratschker V, et al. (2009) Bruchpilot in ribbon-like axonal agglomerates, behavioral defects, and early death in SRPK79D kinase mutants of Drosophila. *PLoS Genet* 5(10):e1000700.

463. Nitert MD, et al. (2012) Impact of an exercise intervention on DNA methylation in skeletal muscle from first-degree relatives of patients with type 2 diabetes. *Diabetes* 61(12):3322–3332.

464. Rönn T, et al. (2013) A six month exercise intervention influences the genome-wide DNA methylation pattern in human adipose tissue. *PLoS Genet* 9(6):e1003572.

465. French SW (2013) Epigenetic events in liver cancer resulting from alcoholic liver disease. *Alcohol Res* 35(1):57–67.

466. Bönsch D, et al. (2004) Homocysteine associated genomic DNA hypermethylation in patients with chronic alcoholism. *J Neural Transm* 111(12):1611–6.

467. Semmler A, et al. (2015) Alcohol abuse and cigarette smoking are associated with global DNA hypermethylation: Results from the German Investigation on Neurobiology in Alcoholism (GINA). *Alcohol* 49(2):97–101.

468. Cho BK, Rao VP, Ge Q, Eisen HN, Chen J (2000) Homeostasis-stimulated proliferation drives naive T cells to differentiate directly into memory T cells. *J Exp Med* 192(4):549–556.

469. Wijetunga NA, et al. (2014) The meta-epigenomic structure of purified human stem cell populations is defined at cis-regulatory sequences. *Nat Commun* 5:5195.

470. Farris SP, Arasappan D, Hunicke-Smith S, Harris RA, Mayfield RD (2014) Transcriptome organization for chronic alcohol abuse in human brain. *Mol Psychiatry* 20(11):1438–1447.

471. Zhang H, et al. (2014) Differentially co-expressed genes in postmortem prefrontal cortex of individuals with alcohol use disorders: influence on alcohol metabolism-related pathways. *Hum Genet* 133(11):1383–1394.

472. Shvetsov YB, et al. (2014) Intraindividual variation and short-term temporal trend in DNA methylation of human blood. *Cancer Epidemiol Biomarkers Prev* 24(3):490–497.

473. Jirtle RL, Skinner MK (2007) Environmental epigenomics and disease susceptibility. *Nat Rev Genet* 8:253.

474. Byun H-M, et al. (2009) Epigenetic profiling of somatic tissues from human autopsy specimens identifies tissue- and individual-specific DNA methylation patterns. *Hum Mol Genet* 18(24):4808–4817.

475. Laird PW (2010) Principles and challenges of genomewide DNA methylation analysis. *Nat Rev Genet* 11(3):191–203.

476. Li Y, et al. (2010) The DNA methylome of human peripheral blood mononuclear cells. *PLoS Biol* 8(11):e1000533.

477. Spiers H, et al. (2015) Methylomic trajectories across human fetal brain development. *Genome Res* 25(3):338–352.

478. Alisch RS, et al. (2012) Age-associated DNA methylation in pediatric populations. *Genome Res* 22(4):623–632.

479. Bell JT, et al. (2012) Epigenome-wide scans identify differentially methylated regions for age and age-related phenotypes in a healthy ageing population. *PLOS Genet* 8(4):189–200.

480. Moore S, et al. (2017) *Epigenetic correlates of neonatal contact in humans* doi:10.1017/S0954579417001213.

481. Miller GE, et al. (2009) Low early-life social class leaves a biological residue manifested by decreased glucocorticoid and increased proinflammatory signaling. *Proc Natl Acad Sci U S A* 106(34):14716–14721.

482. Eipel M, et al. (2016) Epigenetic age predictions based on buccal swabs are more precise in combination with cell type-specific DNA methylation signatures. *Aging (Albany NY)* 8(5):1034–1044.

483. Guo Y, et al. (2014) Illumina human exome genotyping array clustering and quality control. *Nat Protoc* 9:2643.

484. Banovich NE, et al. (2014) Methylation QTLs are associated with coordinated changes in transcription factor binding, histone modifications, and gene expression levels. *Plos Genet* 10(9):e1004663.

485. Almstrup K, et al. (2016) Pubertal development in healthy children is mirrored by DNA methylation patterns in peripheral blood. *Sci Rep* 6:28657.

486. Xu C-J, et al. (2017) The emerging landscape of dynamic DNA methylation in early childhood. *BMC Genomics* 18(1):25.

487. McEwen LM, et al. (2017) Differential DNA methylation and lymphocyte proportions in a Costa Rican high longevity region. *Epigenetics Chromatin* 10(1):21.

488. Dlugos DJ, Scattergood TM, Ferraro TN, Berrettinni WH, Buono RJ (2005) Recruitment rates and fear of phlebotomy in pediatric patients in a genetic study of epilepsy. *Epilepsy Behav* 6(3):444–446.

489. Lin X, et al. (2017) Choice of surrogate tissue influences neonatal EWAS findings. *BMC Med* 15(1):211.

490. Teschendorff AE, et al. (2015) Correlation of smoking-associated DNA methylation changes in buccal cells with DNA methylation changes in epithelial cancer. *JAMA Oncol* 1(4):476–485.

491. Houtepen LC, et al. (2016) Genome-wide DNA methylation levels and altered cortisol stress reactivity following childhood trauma in humans. *Nat Commun* 7:10967.

492. Andrews S V, et al. (2017) Cross-tissue integration of genetic and epigenetic data offers insight into autism spectrum disorder. *Nat Commun* 8(1):1011.

493. Gutierrez-Arcelus M, et al. (2015) Tissue-specific effects of genetic and epigenetic variation on gene regulation and splicing. *Plos Genet* 11(1):e1004958.

494. Calvano SE, et al. (2005) A network-based analysis of systemic inflammation in humans. *Nature* 437:1032.

495. Marr AK, et al. (2014) Leishmania donovani infection aauses distinct epigenetic DNA methylation changes in host macrophages. *PLOS Pathog* 10(10):e1004419.

496. Pacis A, et al. (2015) Bacterial infection remodels the DNA methylation landscape of human dendritic cells. *Genome Res* 25(12):1801–1811.

497. Jones MJ, Moore SR, Kobor MS (2017) Principles and challenges of applying epigenetic epidemiology to psychology. *Annu Rev Psychol*. Available at: http://www.annualreviews.org/doi/abs/10.1146/annurev-psych-122414-033653.

498. Quartararo CE, Reznik E, DeCarvalho AC, Mikkelsen T, Stockwell BR (2015) High-throughput

screening of patient-derived cultures reveals potential for precision medicine in glioblastoma. *ACS Med Chem Lett* 6(8):948–952.

499. Cook RT, et al. (1994) Fine T-Cell subsets in alcoholics as determined by the expression of l-Selectin, leukocyte common antigen, and beta-Integrin. *Alcohol Clin Exp Res* 18(1):71–80.

500. Brisson AR, Matsui D, Rieder MJ, Fraser DD (2012) Translational research in pediatrics: tissue sampling and biobanking. *Pediatrics* 129(1):153 LP-162.

501. Holland NT, Pfleger L, Berger E, Ho A, Bastaki M (2005) Molecular epidemiology biomarkers—sample collection and processing considerations. *Toxicol Appl Pharmacol* 206(2):261–268.

502. Huang Y, et al. (2010) The behaviour of 5-hydroxymethylcytosine in bisulfite sequencing. *PLoS One* 5(1):e8888.

503. An J, Rao A, Ko M (2017) TET family dioxygenases and DNA demethylation in stem cells and cancers. *Exp Mol Med* 49:e323.

504. Johnson KC, et al. (2016) 5-Hydroxymethylcytosine localizes to enhancer elements and is associated with survival in glioblastoma patients. *Nat Commun* 7:13177.

505. Takai H, et al. (2014) 5-Hydroxymethylcytosine Plays a Critical Role in Glioblastomagenesis by Recruiting the CHTOP- Methylosome Complex. *CellReports* 9(1):48–60.

506. Teh AL, et al. (2016) Comparison of Methyl-capture Sequencing vs. Infinium 450K methylation array for methylome analysis in clinical samples. *Epigenetics* 11(1):36–48.

507. Allum F, et al. (2015) Characterization of functional methylomes by next-generation capture sequencing identifies novel disease-associated variants. *Nat Commun* 6:7211.

508. Chen Y, et al. (2013) Discovery of cross-reactive probes and polymorphic CpGs in the Illumina Infinium HumanMethylation450 microarray. *Epigenetics* 8(2):203–209.

509. Carmona JJ, et al. (2017) Empirical comparison of reduced representation bisulfite sequencing and Infinium BeadChip reproducibility and coverage of DNA methylation in humans. *npj Genomic Med* 2(1):13.

510. Kacmarczyk TJ, et al. (2018) "Same difference": comprehensive evaluation of four DNA methylation measurement platforms. *Epigenetics Chromatin* 11(1):21.

511. Paul F, et al. (2016) Transcriptional heterogeneity and lineage commitment in myeloid progenitors. *Cell* 164(1):325.

512. Jaitin DA, et al. (2014) Massively parallel single-cell RNA-Seq for marker-free decomposition of tissues into cell types. *Science (80-)* 343(6172):776 LP-779.

513. Fagny M, et al. (2015) The epigenomic landscape of African rainforest hunter-gatherers and farmers. *Nat Commun* 6:10047.

514. Galanter JM, et al. (2017) Differential methylation between ethnic sub-groups reflects the effect of genetic ancestry and environmental exposures. *Elife* 6:e20532–e20532.

515. Hannon E, et al. (2016) Methylation quantitative trait loci in the developing brain and their enrichment in schizophrenia-associated genomic regions. *Nat Neurosci* 19(1):48–54.

516. Van Dongen J, et al. (2016) Genetic and environmental influences interact with age and sex in shaping the human methylome. *Nat Commun* 7.

517. Moen EL, et al. (2013) Genome-wide variation of cytosine modifications between European and African populations and the implications for complex traits. *Genetics* 194(4):987–996.

518. Rahmani E, et al. (2017) Genome-wide methylation data mirror ancestry information. *Epigenetics Chromatin* 10(1):1.

519. Baccarelli A, Bollati V (2009) Epigenetics and environmental chemicals. *Curr Opin Pediatr* 21(2):243–251.

520. Breton C V, et al. (2009) Prenatal Tobacco Smoke Exposure Affects Global and Gene-specific DNA Methylation. *Am J Respir Crit Care Med* 180(5):462–467.

521. Terry MB, et al. (2008) Genomic DNA methylation among women in a multi-ethnic New York City birth cohort. *Cancer Epidemiol Biomarkers Prev* 17(9):10.1158/1055-9965.EPI-08-0312.

522. Gao X, Jia M, Zhang Y, Breitling LP, Brenner H (2015) DNA methylation changes of whole blood cells in response to active smoking exposure in adults: a systematic review of DNA methylation studies. *Clin Epigenetics* 7(1):113.

523. Su D, et al. (2016) Distinct Epigenetic Effects of Tobacco Smoking in Whole Blood and among Leukocyte Subtypes. *PLoS One* 11(12):e0166486.

524. Bauer M, et al. (2015) A varying T cell subtype explains apparent tobacco smoking induced single CpG hypomethylation in whole blood. *Clin Epigenetics* 7(1):81.

525. Bauer M, et al. (2016) Tobacco smoking differently influences cell types of the innate and adaptive immune system---indications from CpG site methylation. *Clin Epigenetics* 8(1):83.

526. Lin J, et al. (2011) Disulfiram Is a DNA demethylating agent and inhibits prostate cancer cell growth. *Prostate* 71(4):333–343.

527. Yip S, et al. (2009) MSH6 mutations arise in glioblastomas during temozolomide therapy and mediate temozolomide resistance. *Clin Cancer Res* 15(14):4622–4629.

528. Wild CP (2012) The exposome: From concept to utility. *Int J Epidemiol* 41(1):24–32.

529. Herceg Z, et al. (2018) Roadmap for investigating epigenome deregulation and environmental origins of cancer. *Int J Cancer* 142(5):874–882.

530. Relton CL, Davey Smith G (2012) Two-step epigenetic Mendelian randomization: a strategy for establishing the causal role of epigenetic processes in pathways to disease. *Int J Epidemiol* 41(1):161–176.

531. Latvala A, Ollikainen M (2016) Mendelian randomization in (epi)genetic epidemiology: an effective tool to be handled with care. *Genome Biol* 17(1):156.

532. Enríquez P (2016) CRISPR-mediated epigenome editing. *Yale J Biol Med* 89(4):471–486.

533. Cloninger CR, et al. (1988) Genetic heterogeneity and the classification of alcoholism. *Adv Alcohol Subst Abuse* 7(3–4):3–16.

534. Room R, Babor T, Rehm J (2005) Alcohol and public health. *Lancet* 365(9458):519–530.

535. Hasler G, Drevets WC, Manji HK, Charney DS (2004) Discovering endophenotypes for major depression. *Neuropsychopharmacology* 29(10):1765–1781.

536. Park S, Gooding DC (2014) Working memory impairment as an endophenotypic marker of a schizophrenia diathesis. *Schizophr Res Cogn* 1(3):127–136.

537. Dorph-Petersen K-A, et al. (2009) Volume and neuron number of the lateral geniculate nucleus in schizophrenia and mood disorders. *Acta Neuropathol* 117(4):369–384.

538. Deming Y, et al. (2016) A potential endophenotype for Alzheimer's disease: cerebrospinal fluid clusterin. *Neurobiol Aging* 37:208.e1-208.e9.

539. Breton C V., et al. (2017) Small-magnitude effect sizes in epigenetic end points are important in children's environmental health studies: The children's environmental health and disease prevention research center's epigenetics working group. *Environ Health Perspect* 125(4):511–526.

540. Consortium TP, et al. (2015) The PsychENCODE project. *Nat Neurosci* 18(12):1707–1712.

541. Francis LP (2014) Genomic knowledge sharing: A review of the ethical and legal issues. *Appl Transl Genomics* 3(4):111–115.

542. Ng B, et al. (2017) Brain xQTL map: integrating the genetic architecture of the human brain transcriptome And epigenome. *Nat Neurosci*:1–23.

543. F. PR, Ramón TJ, F. BG, F. FA, F. FM (2018) Distinct chromatin signatures of DNA hypomethylation in aging and cancer. *Aging Cell* 0(0):e12744.

544. Wu Y, et al. (2018) Integrative analysis of omics summary data reveals putative mechanisms underlying complex traits. *Nat Commun* 9(1):918.

545. Schulz H, et al. (2017) Genome-wide mapping of genetic determinants influencing DNA methylation and gene expression in human hippocampus. *Nat Commun* 8(1):1511.

546. Schwartzman O, Tanay A (2015) Single-cell epigenomics: techniques and emerging applications. *Nat Rev Genet* 16(12):716–726.

547. Clark SJ, Lee HJ, Smallwood SA, Kelsey G, Reik W (2016) Single-cell epigenomics: powerful new methods for understanding gene regulation and cell identity. *Genome Biol* 17:72.

548. Clark SJ, et al. (2018) scNMT-seq enables joint profiling of chromatin accessibility DNA methylation and transcription in single cells. *Nat Commun* 9(1):781.

549. Miska EA, Ferguson-Smith AC (2016) Transgenerational inheritance: Models and mechanisms of non–DNA sequence–based inheritance. *Science (80-)* 354(6308):59 LP-63.

550. van Otterdijk SD, Michels KB (2016) Transgenerational epigenetic inheritance in mammals: how good is the evidence? *Fed Am Soc Exp Biol J* 30:1–9.

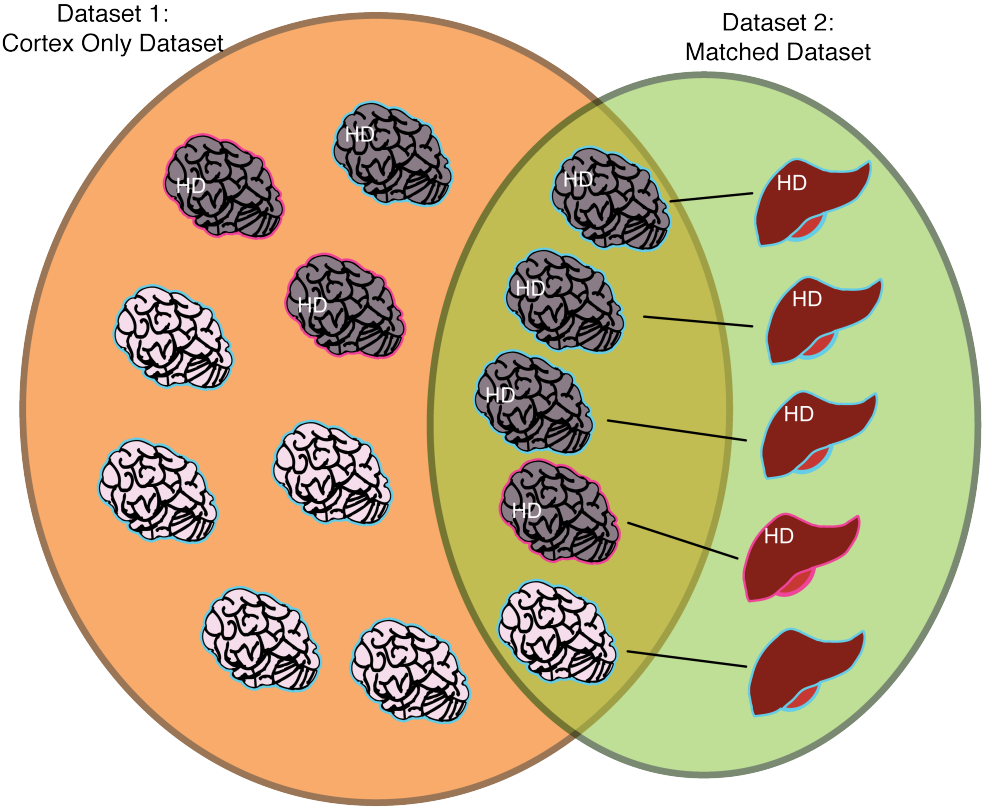
551. Carter AC, et al. (2017) Challenges and recommendations for epigenomics in precision health. *Nat Biotechnol* 35:1128.

Appendices

Appendix A Supplementary Material for Chapter 2

A.1 Supplementary Figures

A

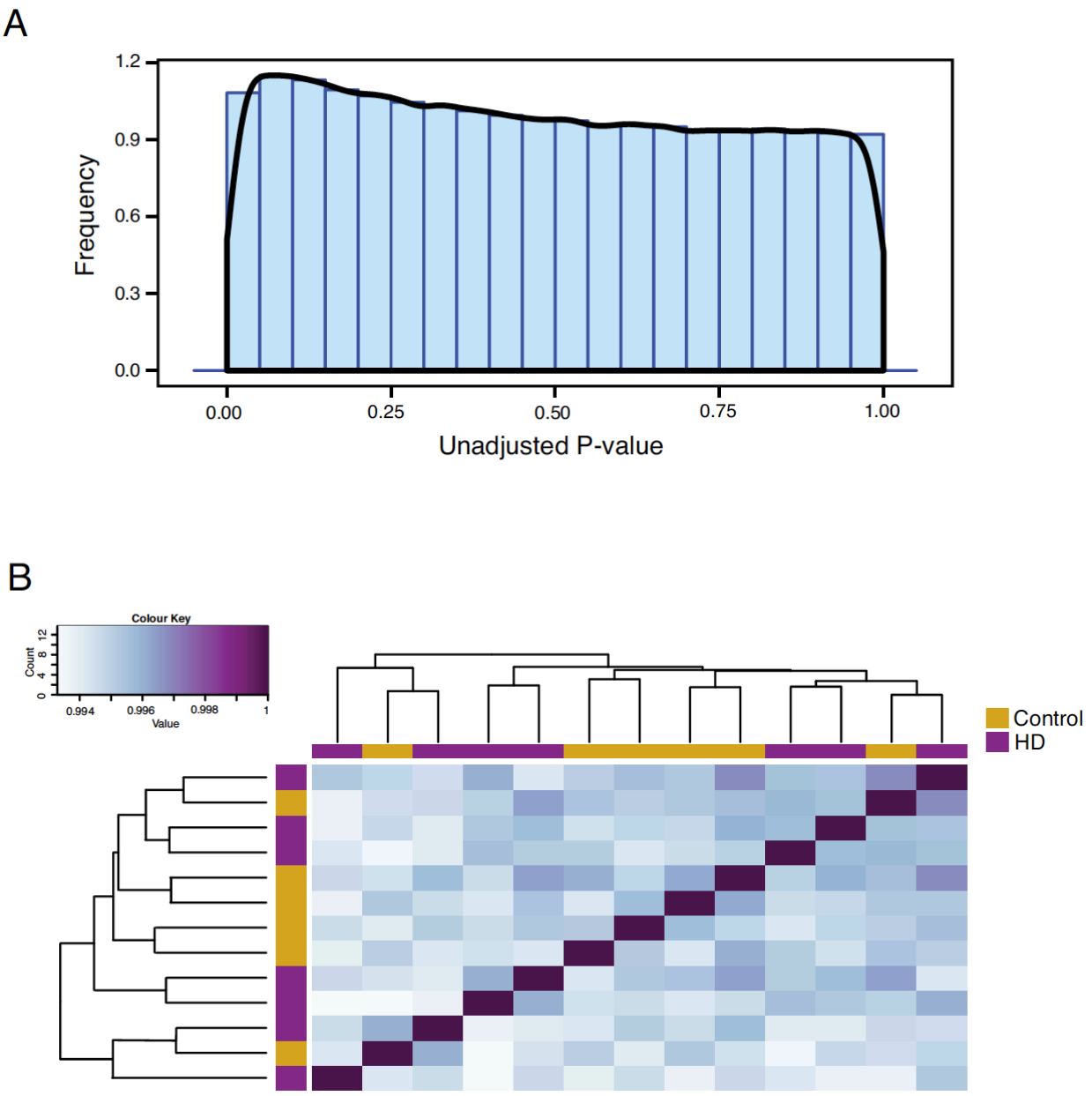


B

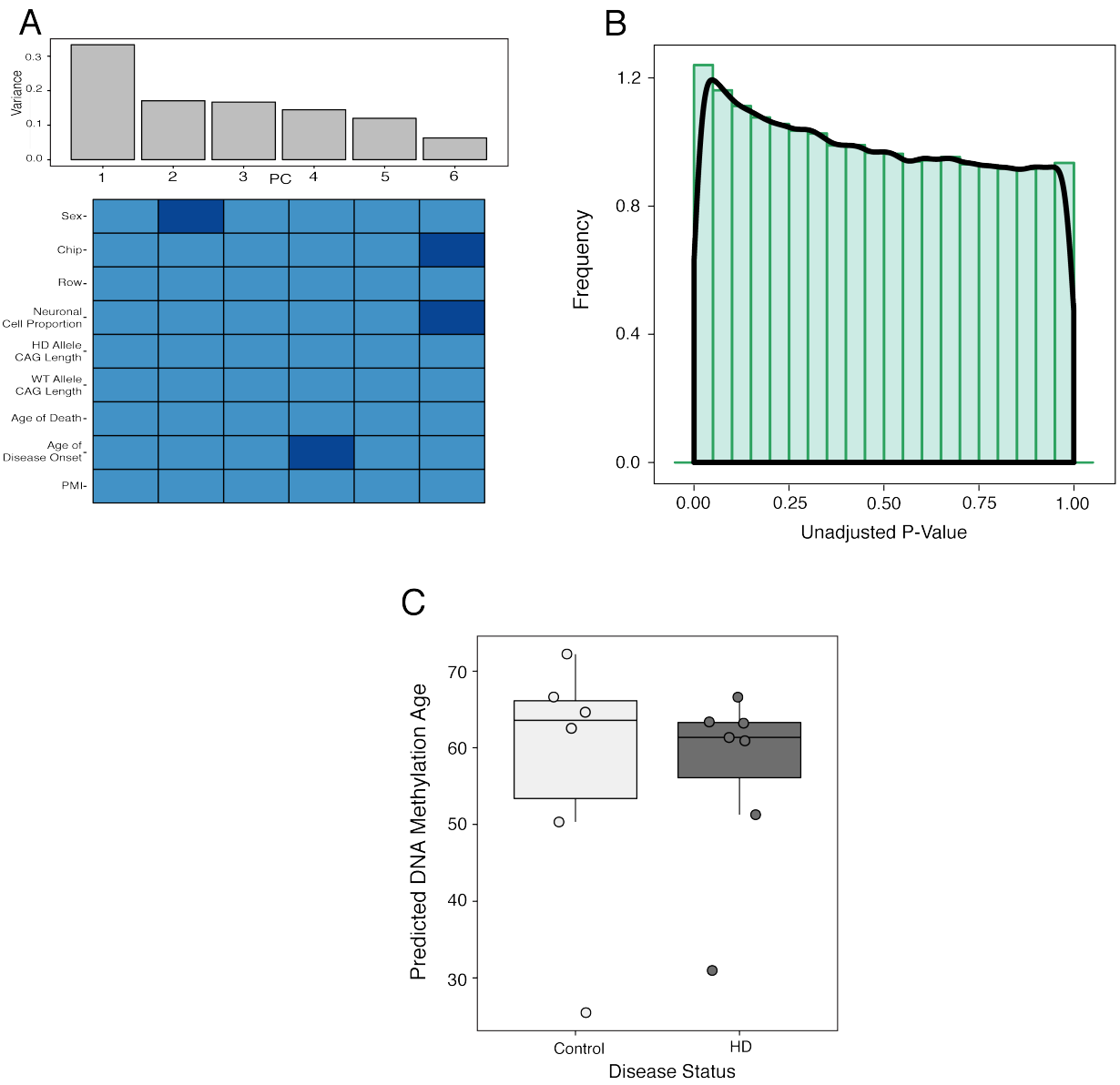
Cortex Only Dataset	N = 13
Sex (M/F)	9/4
Disease Status (HD/Control)	7/6
HTT CAG lenght (avg. repeats) (HD/control)	46/20.4
WT CAG length (avg. repeats)	18.8
Age of Death (avg. years)	62.4
Age of Onset (avg. years)	45.4

Matched Dataset	N = 5 Liver, N = 5 Cortex
Sex (M/F)	4/1
Disease Status (HD/Control)	4/1
HTT CAG lenght (avg. repeats) (HD/control)	49/19
WT CAG length (avg. repeats)	20
Age of Death (avg. years)	56.2
Age of Onset (avg. years)	45.2

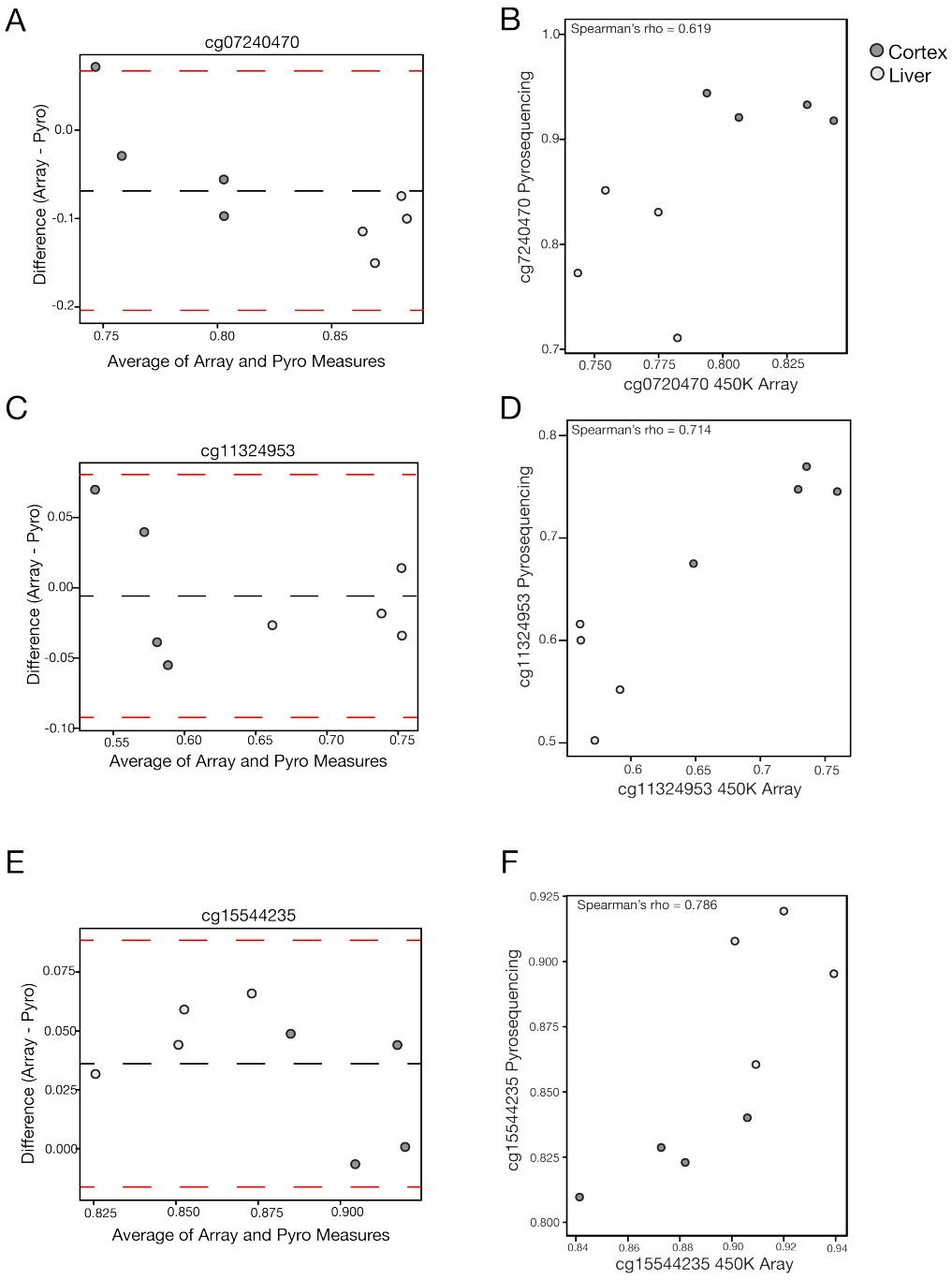
Supplementary Figure 2.1 Schematic of Distribution of Human Samples in Cortex-only and Matched 450K Datasets (A) Venn diagram depicting the distribution of the human samples used in the creation of the cortex only and matched datasets. Blue and pink outlines denote male and female, respectively. (B) Table depicting the sample characteristics for each dataset.



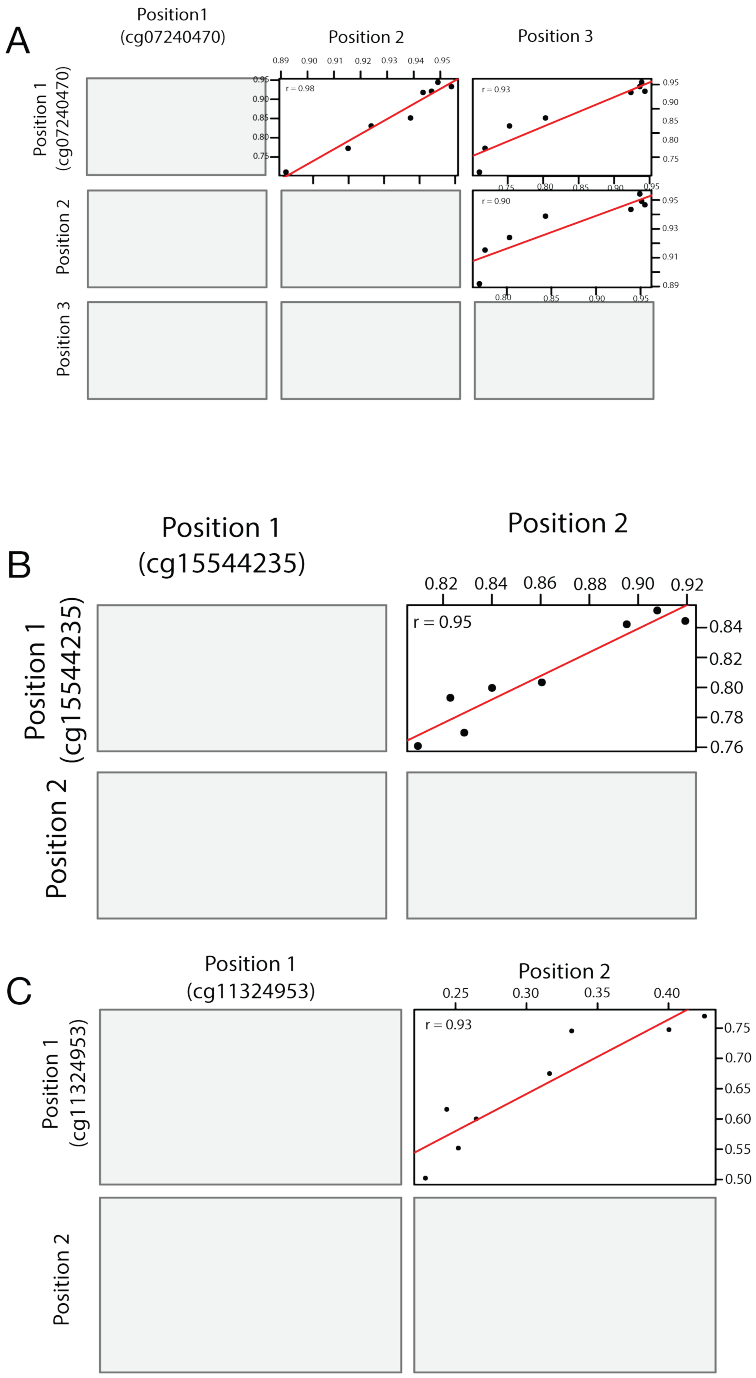
Supplementary Figure 2.2 P-value Distribution of Association to HD Status and Hierarchical Sample Clustering of Genome-wide Cortex Methylation Profiles (A) Graphical representation of uniform p-value distribution indicating the lack association of HD status to cortex DNA methylation profiles. (B) Heatmap of sample-tosample correlations of genome-wide cortex methylation profiles in cortex-only dataset. Purple denotes HD cases while yellow indicates control samples.



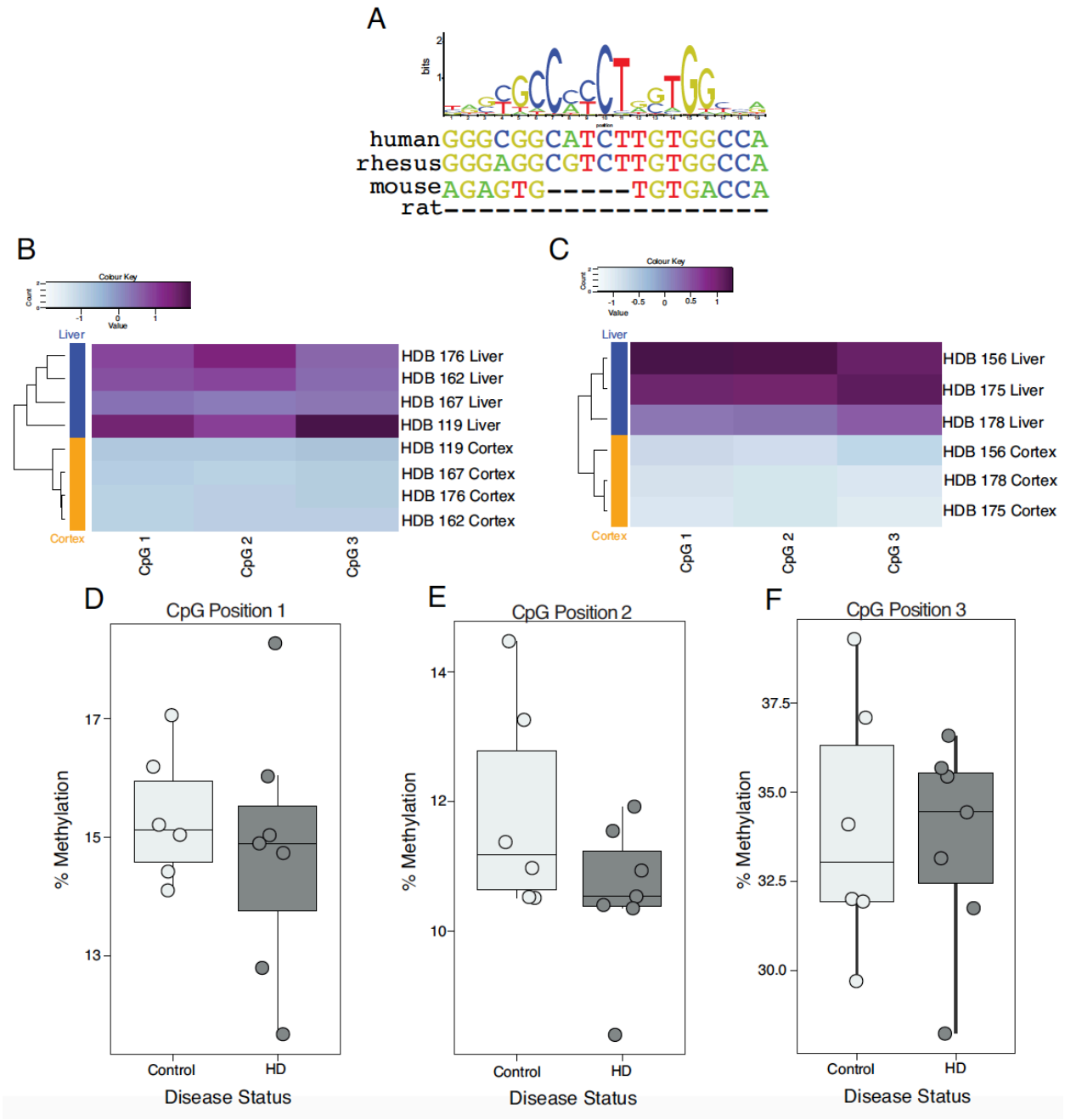
Supplementary Figure 2.3 Association of Age of Disease Onset to HD Cortex Methylation Profiles and DNA Methylation Age Differences Between HD Cases versus Controls. (A) PCA of the HD cortex methylation profiles shows the correlation of known phenotypic and technical variables to PCs (bottom heatmap), each representing an incremental proportion of the variance in the methylation data (top scree plot). Age of onset significantly correlated with PC4 representing 14.5% of the methylation variance in the HD cortex samples. (B) Leftward skewing of p-value distribution for the association of age of onset (adjusted for sex) on DNA methylation in HD cortex samples, signifying an enrichment of low p-values beyond expected by chance. (C) Predicted DNA methylation age analysis revealed no significant differences between HD cases versus controls in the cortex-only dataset (Mann-Whitney U test, ns).



Supplementary Figure 2.4 Comparison of 450K and Pyrosequencing Measures. Bland-Altman plots for (A) cg07240470, (C) cg11324953 and (E) cg15544235 show unbiased agreement between 450K and pyrosequencing measures. Corresponding scatterplots for (B) cg07240470 (D) cg11324953 and (F) cg15544235 show correlation between 450K and pyrosequencing measures (Spearman correlation, ns for cg07240470 and cg11324953, *p < 0.05 for cg15544235).



Supplementary Figure 2.5 Correlation of Neighbouring CpG sites in Each Pyrosequencing Assay. Spearman correlation of adjacent CpG sites underlying pyrosequencing assay for (A) cg07240470, (B) cg11324953 and (C) cg15544235 shows that neighbouring CpG sites are highly correlated (Spearman correlation, ***$p < 0.0001$).



Supplementary Figure 2.6 DNA Methylation of CTCF Site in Original 450K and Validation Individuals and in Cortex-only Dataset (A) CTCF Consensus sequence from JASPAR (above). Conservation of identified CTCF site across rhesus monkey, mouse and rat as identified from the UCSC Genome Browser (below). (B-C) Heatmap representation of CpG methylation of the three identified CpGs used in the CTCF pyrosequencing assay for the original individuals used in the 450K analysis (B) and the validation individuals (C). Yellow denotes cortex samples while blue indicates liver samples. (D-F) Measured DNA methylation for each pyrosequenced CpG underlying CTCF TFBS in HD (coloured in orange) versus control (coloured in green) cortex samples (Mann-Whitney U test, ns).

A.2 Supplementary Tables

Supplementary Table 2.1 RT-qPCR Primers

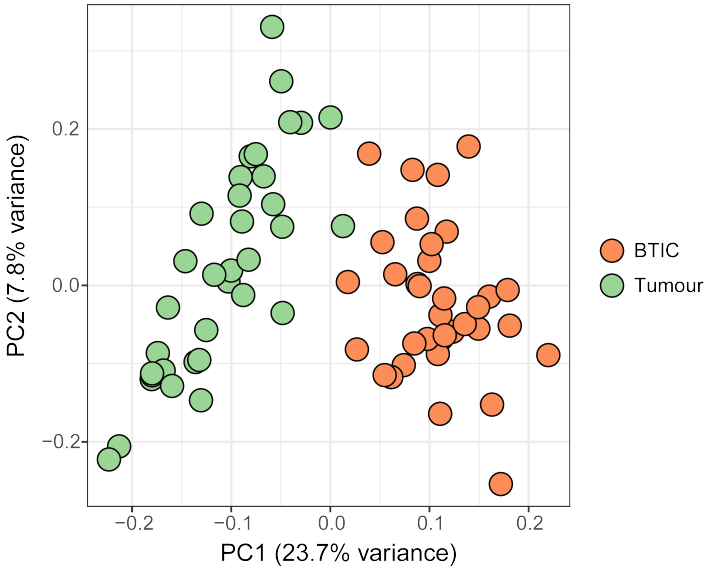
Human Primers	Forward	Reverse
HTT	TCCACCATGCAAGACTCACTTAG	TGGGATTTGACAAGATGAACGT
ActB	AGTACTCCGTGTGGATCGGC	GCTGATCCACATCTGCTGGA
RGAG4	GGACAGCGCCCAACATTG	CTGGCTACCCTTTAGGCAACA
Ddah1	TTTAAGGACTATGCAGTCTCCACAGT	AGCCATGCTGCAGAAACTCTTC
HPRT	TTATGGACAGGACTGAACGTCTTG	GCACACAGAGGGCTACAATGTG
PGK1	CAAATGGAACACGGAGGATAAAG	CTTTACCTTCCAGGAGCTCCAA
Mouse Primers	**Forward**	**Reverse**
Hdh	CATCCTGGAAGCCATTGCA	TTTGTATATCTGAGTCTACTTCCTCCTTTC
ActB	CCAGCCTTCCTTCTTGGGTAT	TGTGTTGGCATAGAGGTCTTTACG
PGK1	CCCCAAGTGGAGGGAAGTACA	TGCCCAGCCGATAGACATC
HPRT	CGTCGTGATTAGCGATGATGA	TCCAAATCCTCGGCATAATGA
18S	AGAAACGGCTACCACATCCAA	GGGTCGGGAGTGGGTAATTT
Csnk2a2	CCACATAGACCTAGATCCACACTTCA	AGGTGCCTGTTCTCACTATGGATAA

Supplementary Table 2.2 Pyrosequencing Primers

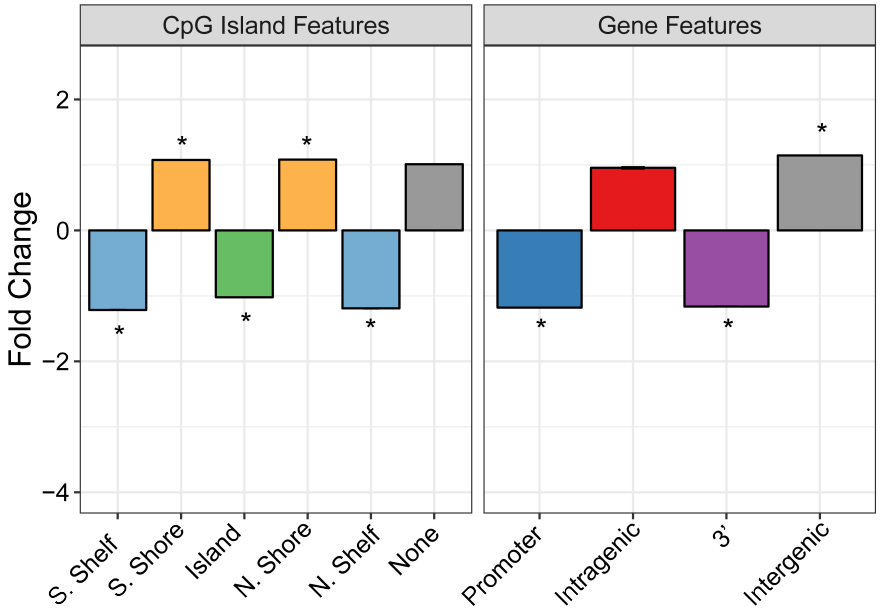
Assay for cg15544235	
Target Sequence Forward Primer	TGGATGTTTTGATGAAGTTAGTTGTTATGT
Target Sequence Reverse Primer	CCCAACTTAACCAACTCCACTT (Biotin)
Sequencing Primer	GTTATGTTGGAGAGGT
Assay for cg0720470	
Target Sequence Forward Primer	TTGATGGGGAGGTTAATTGT
Target Sequence Reverse Primer	ACTTCCTAACTCCTACTATACACT (Biotin)
Sequencing Primer	GAAATAGGAAAAGAGAGATTATTAA
Assay for cg11324953	
Target Sequence Forward Primer	TATAGGTGTAGGGTTTAGTAGTGAGTAGAT (Biotin)
Target Sequence Reverse Primer	CTAACATTTCCCTATCCCCTTCC
Sequencing Primer	CCCTACTTTAAAATTCCTC
Assay for CTCF Site	
Target Sequence Forward Primer	AGGAGGTTTTGGAGATTAGGA
Target Sequence Reverse Primer	CCTACTACCCACAAAAACACTA (Biotin)
Sequencing Primer	AGGGGTGTATTTTTATAAAG

Appendix B Supplementary Material for Chapter 3

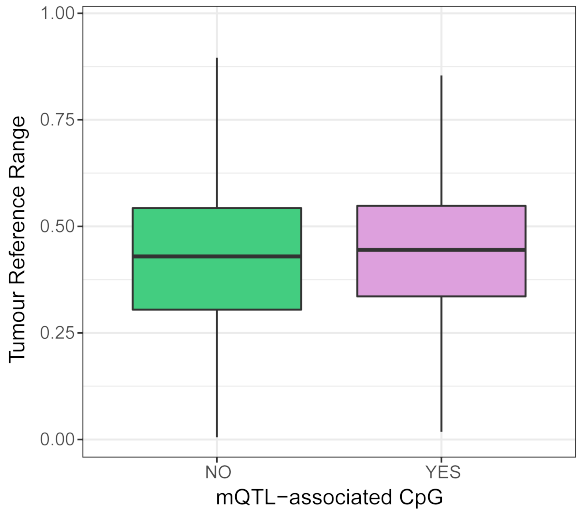
B.1 Supplementary Figures



Supplementary Figure 3.1 Scatterplots of PC1 scores (x-axis) versus PC2 score (y-axis) for BTICs (orange) and matched tumour (green) samples. PC1, which comprised of 23.7% of the total variance in the processed 450K data, was significantly associated with tissue source (Wilcoxon signed-rank test, $p = 5.8 \times 10^{-11}$).



Supplementary Figure 3.2 A) Representation of 131,307 significant differentially variable CpGs across various genomic features. Bars show the fold-change between CpG count in each genomic region and the mean count of randomly selected CpGs in that same genomic feature, from 10,000 iterations. Error bars show standard error. (* denotes significant enrichment or depletion at FDR ≤ 0.05) (S = South; N = North; 3' = 3' Untranslated Region).



Supplementary Figure 3.3 Boxplot of tumour DNAm reference range values of significantly correlated sites between BTICs and GBM tumours, showing modest increase in DNAm reference range for 288 CpGs associated with previously identified mQTL in brain tissues (purple) over 29,716 non-mQTL-associated CpGs (green), although this difference did not reach statistical significance (79).

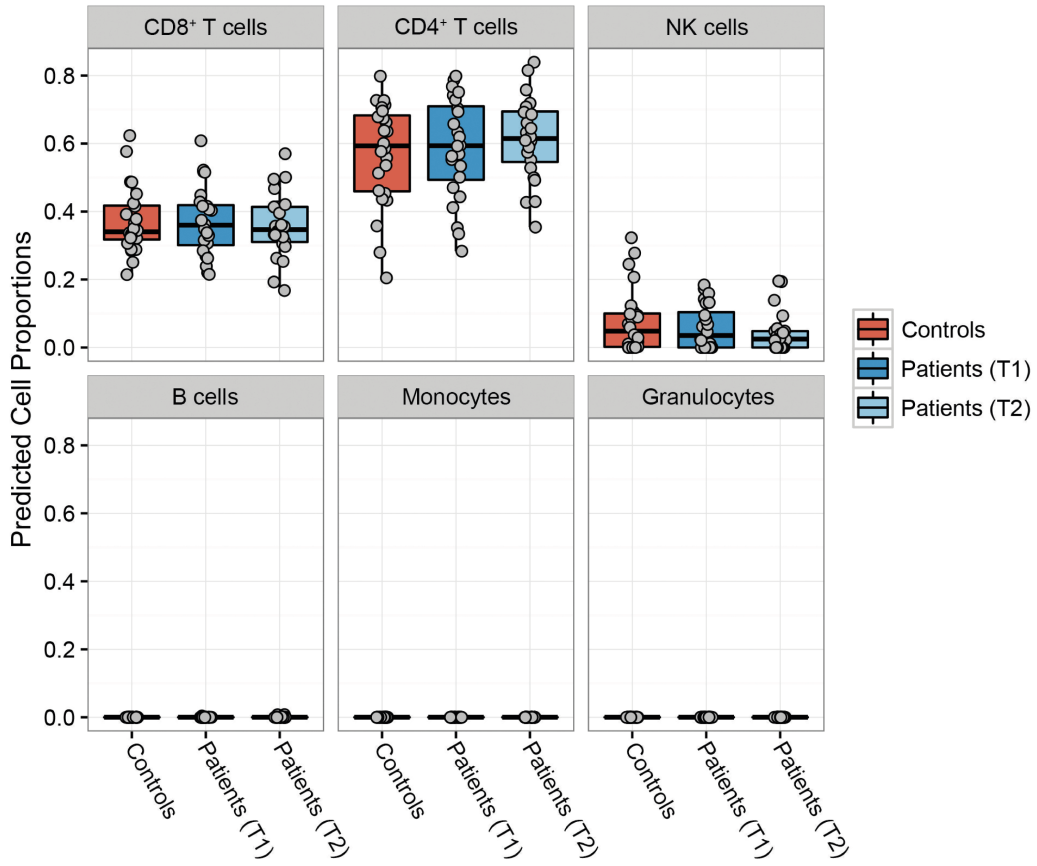
B.2 Supplementary Tables

Supplementary Table 3.1 The number of CpG sites at various thresholds of reference range and Spearman's correlation rho between matched BTICs and GBM tumours

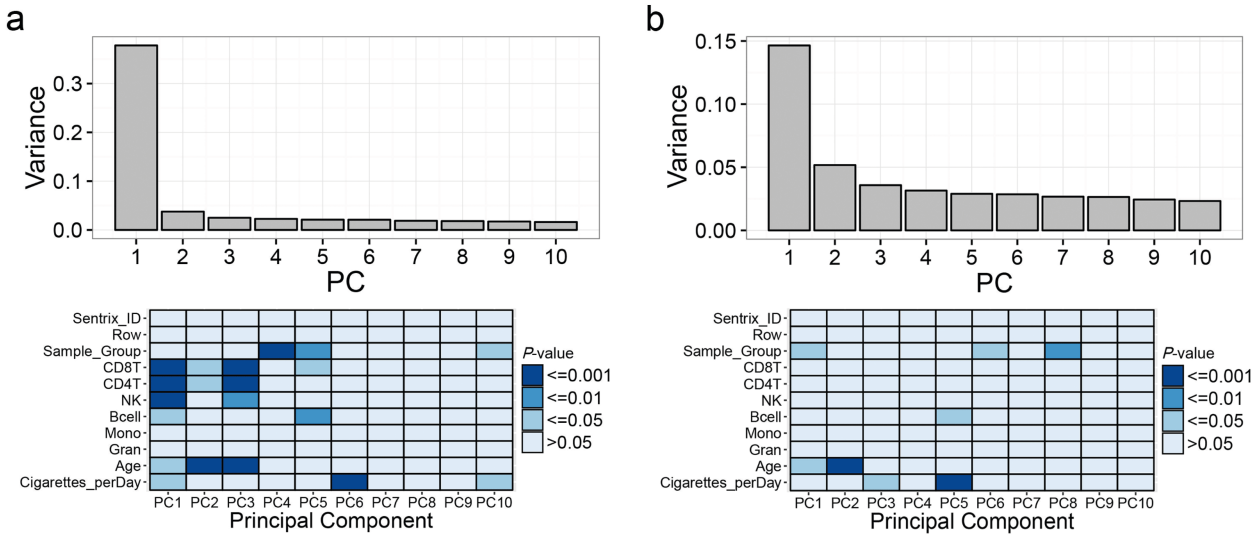
		Tumour Reference Range						
		0	**≥ 0.05**	**≥ 0.1**	**≥ 0.2**	**≥ 0.3**	**≥ 0.4**	**≥ 0.5**
	NA	30,004	29,103	28,122	26,025	22,714	17,047	10,251
Positive	**≥ 0.3**	29,875	29,059	28,099	26,010	22,702	17,040	10,249
Correlation	**≥ 0.6**	9,678	9,580	9,429	9,005	8,229	6,644	4,342
Rho	**≥ 0.9**	19	19	19	19	19	16	15
Negative	**≤ -0.3**	129	44	23	15	12	7	2
Correlation	**≤ -0.6**	7	4	2	1	1	1	0
Rho	**≤ -0.9**	0	0	0	0	0	0	0

Appendix C Supplementary Material for Chapter 4

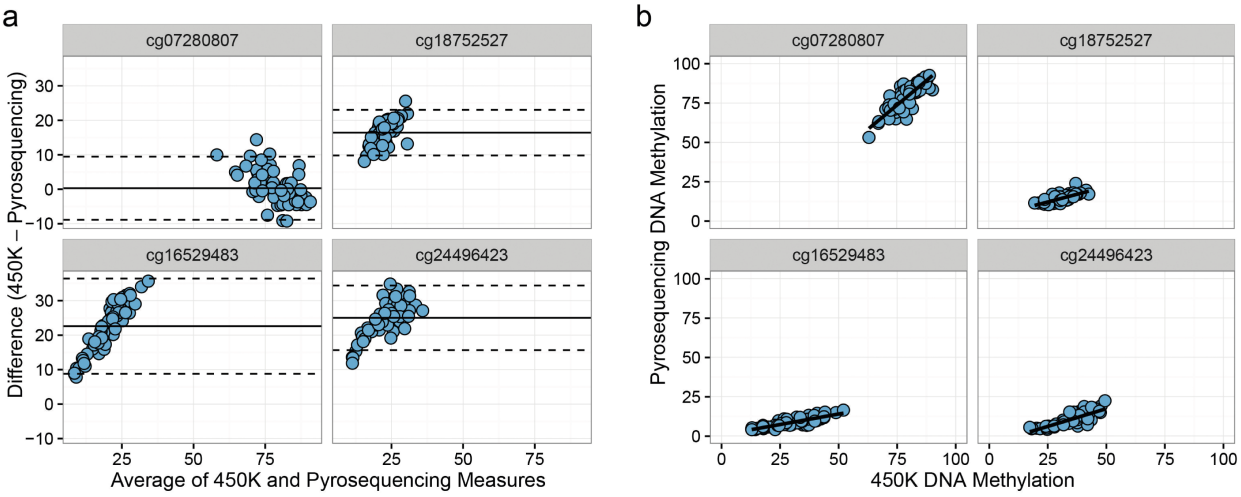
C.1 Supplementary Figures



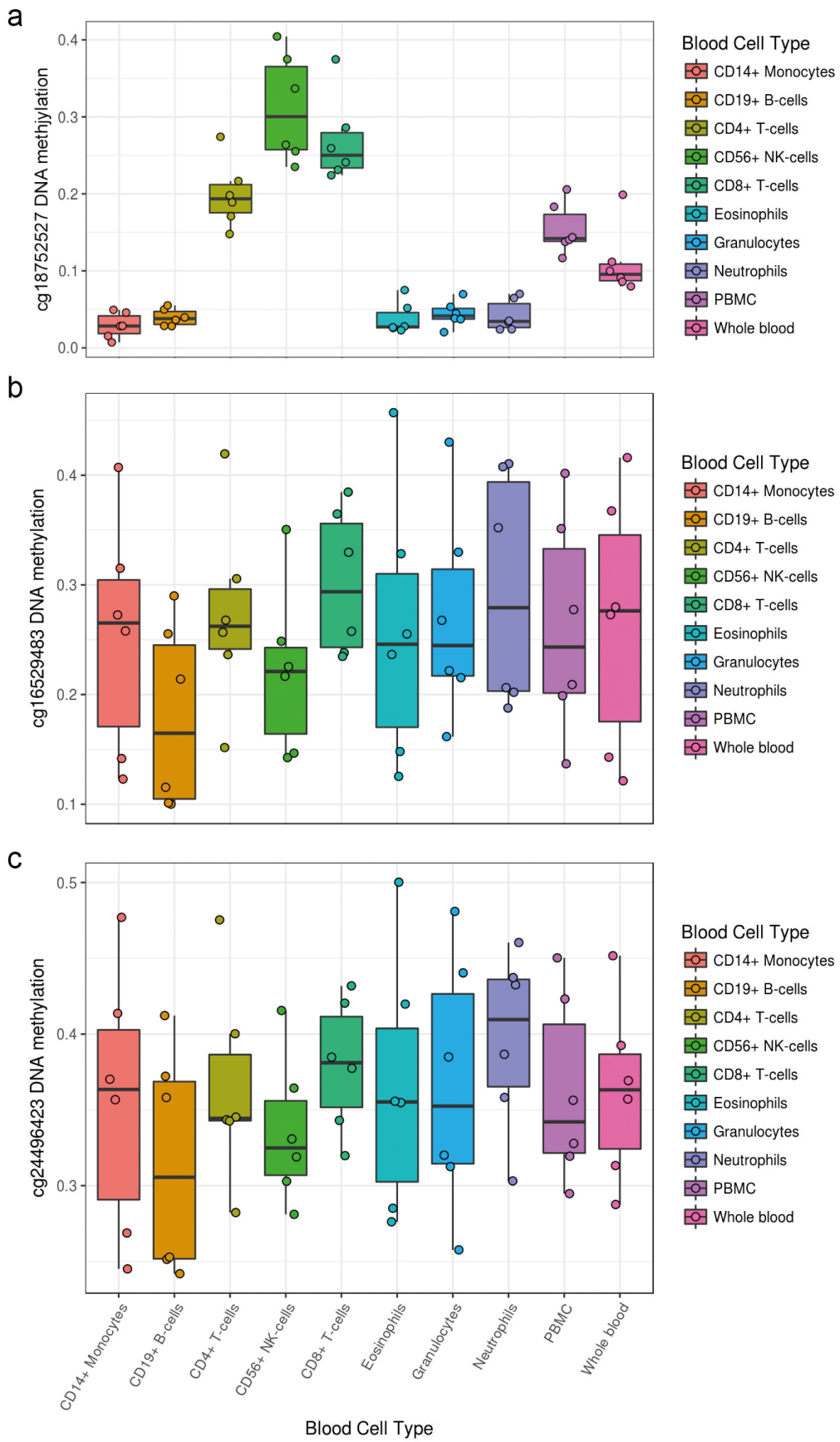
Supplementary Figure 4.1 Estimations of blood cell proportions in samples based on underlying reference DNAm profiles. Estimates were predicted using the Houseman blood cell deconvolution algorithm. There was no significant association between predicted proportions of any cell type and sample group (Mann-Whitney U test for comparison of controls and patients (T1) or controls and patients (T2); Wilcoxon signed-rank test for comparison of matched patients (T1) and patients (T2) samples).



Supplementary Figure 4.2 Principal component analyses before and after regression-based adjustment of the 450K data. PCA showing the correlation of known phenotypic and technical variables to the top 10 principal components, each representing an incremental proportion of the variance in the methylation data. a) Top 10 PCs in unadjusted 450K dataset (representing 60% of the DNAm variance) and b) top 10 PCs in the adjusted 450K dataset (representing 45% of the DNAm variance).



Supplementary Figure 4.3 Correlations between 450K array and pyrosequencing measures. a) Bland-Altman plots for verified CpGs show a slightly biased agreement between 450K dataset and pyrosequencing measures. b) Strong positive correlation between 450K and pyrosequencing measures for cg07280807 (Spearman $r_s = 0.85$, $P = 2E\text{-}16$), cg18752527 ($r_s = 0.71$, $P = 3E\text{-}12$), cg16529483 ($r_s = 0.79$, $P = 4E\text{-}16$) and cg24496423 ($r_s = 0.80$, $P = 2E\text{-}16$).


a
cg18752527 DNA methylation
0.0
0.1
0.2
0.3
0.4
b
cg16529483 DNA methylation
0.1
0.2
0.3
0.4
c
cg24496423 DNA methylation
0.3
0.4
0.5
CD14+ Monocytes
CD19+ B-cells
CD4+ T-cells
CD56+ NK-cells
CD8+ T-cells
Eosinophils
Granulocytes
Neutrophils
PBMC
Whole blood
Blood Cell Type

Supplementary Figure 4.4 Blood cell type associations of 3 examined CpG sites. a) DNA methylation of cg18752527 in the *HECW2* gene was significantly associated with $CD4^+$ and $CD8^+$ T cells, along with NK cells, as determined by differential DNAm testing using a previous 450K dataset of purified blood cell types[17] (P = 7.6E-15, ANOVA). DNA methylation of cg16529483 (b) and cg24496423 (c) in the *SRPK3* gene were not significantly associated with any cell type (P > 0.6, ANOVA).

C.2 Supplementary Tables

Supplementary Table 4.1 Differentially methylated sites between Controls and Patients (T1)

#	Probe ID	Gene	Region	Average beta Controls	Average beta Patients (T1)	Δ-beta	*P*-Value	BH-adjusted *P*-Value
1	cg18752527	HECW2	intragenic	0.342	0.276	0.066	4.30E-07	0.0213
2	cg08109624	None	intergenic	0.760	0.817	-0.057	8.15E-07	0.0234
3	cg10168086	None	intergenic	0.535	0.484	0.051	1.24E-06	0.0256
4	cg07280807	None	intergenic	0.755	0.822	-0.068	2.44E-06	0.0366
5	cg12173150	None	intergenic	0.321	0.385	-0.064	3.02E-06	0.0370
6	cg01059398	TNFSF10	intragenic	0.261	0.209	0.052	1.07E-05	0.0627
7	cg17940902	HLA-DMA	promoter	0.399	0.450	-0.051	1.19E-05	0.0640
8	cg22778903	MX2	intragenic	0.304	0.355	-0.051	1.34E-05	0.0666
9	cg14612335	SKIL	promoter	0.423	0.368	0.055	1.38E-05	0.0666
10	cg11580026	None	intergenic	0.600	0.549	0.051	1.51E-05	0.0691
11	cg12284098	MYOM2	intragenic	0.534	0.477	0.056	1.54E-05	0.0691
12	cg26091609	CTLA4	intragenic	0.578	0.518	0.060	1.59E-05	0.0691
13	cg09768654	SRPK3	promoter	0.374	0.466	-0.092	1.65E-05	0.0691
14	cg06851207	PNMAL1	promoter	0.528	0.617	-0.089	1.84E-05	0.0691
15	cg14702960	None	intergenic	0.742	0.689	0.052	1.92E-05	0.0691
16	cg00449728	MAPRE2	intragenic	0.750	0.693	0.057	2.98E-05	0.0702
17	cg22851561	ELMSAN1	intragenic	0.432	0.380	0.052	3.00E-05	0.0702
18	cg02536838	ANGPT1	promoter	0.605	0.530	0.075	3.14E-05	0.0702
19	cg15841511	None	intergenic	0.729	0.788	-0.059	3.42E-05	0.0706
20	cg24392939	CRYBG3	intragenic	0.562	0.510	0.052	3.62E-05	0.0725
21	cg12761472	CEP85L	promoter	0.621	0.566	0.055	4.13E-05	0.0754
22	cg02652579	SYNGAP1	promoter	0.623	0.563	0.059	4.17E-05	0.0758
23	cg22865905	SNORA69	three_plus	0.794	0.743	0.051	4.26E-05	0.0764
24	cg27201673	PNMAL1	promoter	0.213	0.263	-0.050	5.41E-05	0.0778
25	cg04936619	C17orf75	intragenic	0.314	0.245	0.069	5.88E-05	0.0778
26	cg11121969	PCBP3	promoter	0.691	0.627	0.064	6.26E-05	0.0778
27	cg00246693	ARHGAP42	promoter	0.340	0.393	-0.053	7.10E-05	0.0778
28	cg10399005	None	intergenic	0.776	0.833	-0.057	7.11E-05	0.0778
29	cg16529483	SRPK3	promoter	0.252	0.357	-0.105	7.18E-05	0.0780
30	cg01220513	SH3KBP1	intragenic	0.506	0.454	0.051	8.08E-05	0.0791
31	cg26926002	None	intergenic	0.719	0.777	-0.058	8.10E-05	0.0791
32	cg14544087	MIR155HG	intragenic	0.290	0.227	0.063	8.64E-05	0.0791
33	cg20893919	TRPC3	intragenic	0.703	0.754	-0.051	9.23E-05	0.0801
34	cg18682028	FYCO1	intragenic	0.394	0.338	0.056	9.24E-05	0.0801
35	cg04362790	None	intergenic	0.697	0.644	0.052	9.32E-05	0.0801
36	cg09060654	LIPA	intragenic	0.578	0.656	-0.079	9.51E-05	0.0801

#	Probe ID	Gene	Region	Average beta Controls	Average beta Patients (T1)	Δ-beta	*P*-Value	BH-adjusted *P*-Value
37	cg02451774	NBPF8	intragenic	0.431	0.483	-0.053	9.98E-05	0.0806
38	cg18723276	USP29	promoter	0.723	0.774	-0.051	0.0001	0.0819
39	cg13180722	None	intergenic	0.338	0.401	-0.062	0.0001	0.0830
40	cg12230162	SRPK3	promoter	0.357	0.463	-0.105	0.0001	0.0835
41	cg18890544	None	intergenic	0.846	0.905	-0.059	0.0001	0.0839
42	cg24496423	SRPK3	promoter	0.309	0.393	-0.084	0.0001	0.0854
43	cg02661764	None	intergenic	0.419	0.360	0.059	0.0001	0.0867
44	cg01400671	None	intergenic	0.409	0.345	0.064	0.0001	0.0874
45	cg13609457	None	intergenic	0.577	0.521	0.056	0.0002	0.0897
46	cg25880958	None	intergenic	0.591	0.645	-0.054	0.0002	0.0898
47	cg18376497	INPP4B	intragenic	0.286	0.223	0.064	0.0002	0.0919
48	cg13784312	RAPGEF1	intragenic	0.187	0.136	0.051	0.0002	0.0928
49	cg07135405	MIR1914	three_plus	0.540	0.394	0.146	0.0002	0.0928
50	cg20475486	None	intergenic	0.702	0.759	-0.058	0.0002	0.0936
51	cg11858450	CCDC105	intragenic	0.709	0.762	-0.053	0.0002	0.0940
52	cg05927817	None	intergenic	0.726	0.787	-0.061	0.0002	0.0940
53	cg00306893	None	intergenic	0.737	0.675	0.062	0.0002	0.0940
54	cg10365886	TNXB	intragenic	0.566	0.672	-0.105	0.0002	0.0947
55	cg27503950	None	intergenic	0.633	0.696	-0.063	0.0002	0.0952
56	cg01089001	GALNT18	intragenic	0.317	0.382	-0.065	0.0002	0.0953
57	cg12564698	GAL	three_plus	0.312	0.261	0.051	0.0002	0.0953
58	cg16197188	NRG3	intragenic	0.723	0.672	0.051	0.0003	0.0995
59	cg04088338	None	intergenic	0.430	0.378	0.052	0.0003	0.0999

Abbreviations: Average beta, mean methylation values (%); Benjamini-Hochberg (BH) adjusted P-value.

Supplementary Table 4.2 Top listed hits detected by both site-specific and DMRcate analysis.

Probe ID	Gene	DMR	Position	Average beta Controls	Average beta Patients (T1)	Δ-beta	*P*-Value	FDR
cg16529483	*SRPK3*	chrX:153046175-153047707	153046451	0.252	0.357	-0.105	3.52E-23	5.90E-19
cg24496423	*SRPK3*	chrX:153046175-153047707	153046480	0.309	0.393	-0.084	2.84E-23	4.94E-19
cg12230162	*SRPK3*	chrX:153046175-153047707	153046482	0.357	0.463	-0.105	2.80E-23	4.94E-19
cg09768654	*SRPK3*	chrX:153046175-153047707	153046386	0.374	0.466	-0.092	6.72E-23	1.01E-18
cg18890544		chr1:242220301-242220925	242220538	0.846	0.905	-0.059	1.75E-18	8.88E-15
cg08109624		chr1:242220301-242220925	242220925	0.760	0.817	-0.057	1.69E-19	1.02E-15
cg27503950		chr6:160023581-160024144	160024002	0.633	0.696	-0.063	2.92E-15	6.57E-12
cg09060654	*LIPA*	chr10:90985055-90985062	90985062	0.578	0.656	-0.079	1.96E-07	4.53E-05

Abbreviations: Average beta, mean methylation values (%); FDR, Benjamini-Hochberg False Discovery Rate; DMR, differentially methylated region.

Supplementary Table 4.3 Differentially methylated sites between Patients (T1) and Patients (T2)

#	Probe ID	Gene	Region	Average beta Patients (T1)	Average beta Patients (T2)	Δ-beta	*P*-Value	BH-adjusted *P*-Value
1	cg15500907	LAMA4	intragenic	0.485	0.542	-0.056	1.01E-06	0.0323
2	cg05266321	CCR2	intragenic	0.545	0.606	-0.061	4.63E-06	0.0487
3	cg13279700	C6orf10	intragenic	0.481	0.544	-0.063	1.76E-05	0.0561
4	cg14054990	KRTAP19-5	promoter	0.431	0.482	-0.052	1.84E-05	0.0565
5	cg21049302	None	intergenic	0.466	0.522	-0.056	1.98E-05	0.0565
6	cg17022548	NRG2	intragenic	0.204	0.258	-0.054	1.99E-05	0.0565
7	cg22472360	TRIO	intragenic	0.514	0.569	-0.055	2.09E-05	0.0569
8	cg07920414	RIMS3	intragenic	0.438	0.493	-0.055	2.18E-05	0.0572
9	cg04088338	None	intergenic	0.378	0.429	-0.051	2.54E-05	0.0590
10	cg12240358	HOMER2	intragenic	0.462	0.519	-0.057	2.68E-05	0.0590
11	cg09712306	AURKA	intragenic	0.602	0.660	-0.058	3.48E-05	0.0605
12	cg07939743	None	intergenic	0.289	0.341	-0.052	3.50E-05	0.0605
13	cg00803692	CCR5	promoter	0.370	0.424	-0.054	3.73E-05	0.0620
14	cg10177030	SNORD12	three_plus	0.419	0.472	-0.053	3.85E-05	0.0627
15	cg15439110	None	intergenic	0.444	0.525	-0.080	3.93E-05	0.0628
16	cg20385229	SLIRP	intragenic	0.392	0.444	-0.052	4.13E-05	0.0628
17	cg02393640	LUZP6	intragenic	0.390	0.443	-0.052	5.63E-05	0.0668
18	cg17863551	CD177	promoter	0.419	0.478	-0.059	6.27E-05	0.0670
19	cg15279541	None	intergenic	0.388	0.439	-0.051	7.14E-05	0.0677
20	cg20171999	RRS1	three_plus	0.403	0.474	-0.070	8.93E-05	0.0680
21	cg20559385	None	intergenic	0.428	0.479	-0.052	9.43E-05	0.0680
22	cg21429780	MAML3	intragenic	0.493	0.545	-0.052	0.0001	0.0680
23	cg01482790	HNRNPM	intragenic	0.289	0.339	-0.050	0.0001	0.0681
24	cg20684197	FGF1	intragenic	0.395	0.445	-0.051	0.0001	0.0684
25	cg04279139	MANSC4	promoter	0.410	0.461	-0.051	0.0001	0.0688
26	cg16853860	PSMB9	intragenic	0.272	0.332	-0.060	0.0001	0.0696
27	cg27062514	CTR9	intragenic	0.463	0.526	-0.064	0.0001	0.0721
28	cg09931909	MB21D1	intragenic	0.420	0.497	-0.077	0.0001	0.0735
29	cg13340231	ZNF704	intragenic	0.528	0.583	-0.055	0.0002	0.0751
30	cg10035831	RPTOR	intragenic	0.446	0.503	-0.057	0.0002	0.0753
31	cg13927756	MYO10	intragenic	0.468	0.524	-0.056	0.0002	0.0754
32	cg08749576	None	intergenic	0.627	0.684	-0.058	0.0002	0.0761
33	cg15484808	RPS18	intragenic	0.480	0.534	-0.054	0.0002	0.0811
34	cg12802876	None	intergenic	0.359	0.418	-0.059	0.0002	0.0828
35	cg03548415	None	intergenic	0.422	0.473	-0.051	0.0003	0.0853
36	cg20547015	PPP1CC	intragenic	0.453	0.517	-0.064	0.0003	0.0862
37	cg23214895	None	intergenic	0.569	0.620	-0.051	0.0003	0.0878

#	Probe ID	Gene	Region	Average beta Patients (T1)	Average beta Patients (T2)	Δ-beta	*P*-Value	BH-adjusted *P*-Value
38	cg12478092	CCDC116	promoter	0.510	0.573	-0.063	0.0003	0.0879
39	cg15683542	MIPEP	intragenic	0.694	0.747	-0.053	0.0003	0.0883
40	cg09514545	MIR525	three_plus	0.442	0.501	-0.060	0.0004	0.0908
41	cg01789743	NID1	intragenic	0.499	0.552	-0.053	0.0004	0.0910
42	cg18524114	None	intergenic	0.339	0.389	-0.050	0.0005	0.0933
43	cg04410448	ZC2HC1B	intragenic	0.491	0.541	-0.051	0.0005	0.0949
44	cg13714407	RAPGEF1	intragenic	0.367	0.426	-0.059	0.0005	0.0953
45	cg27367066	None	intergenic	0.455	0.510	-0.054	0.0006	0.0967
46	cg26837708	YBX1	intragenic	0.388	0.445	-0.058	0.0006	0.0967
47	cg14817867	PRPSAP2	intragenic	0.419	0.471	-0.052	0.0006	0.0971
48	cg13598358	PPP1CC	intragenic	0.362	0.418	-0.056	0.0006	0.0978

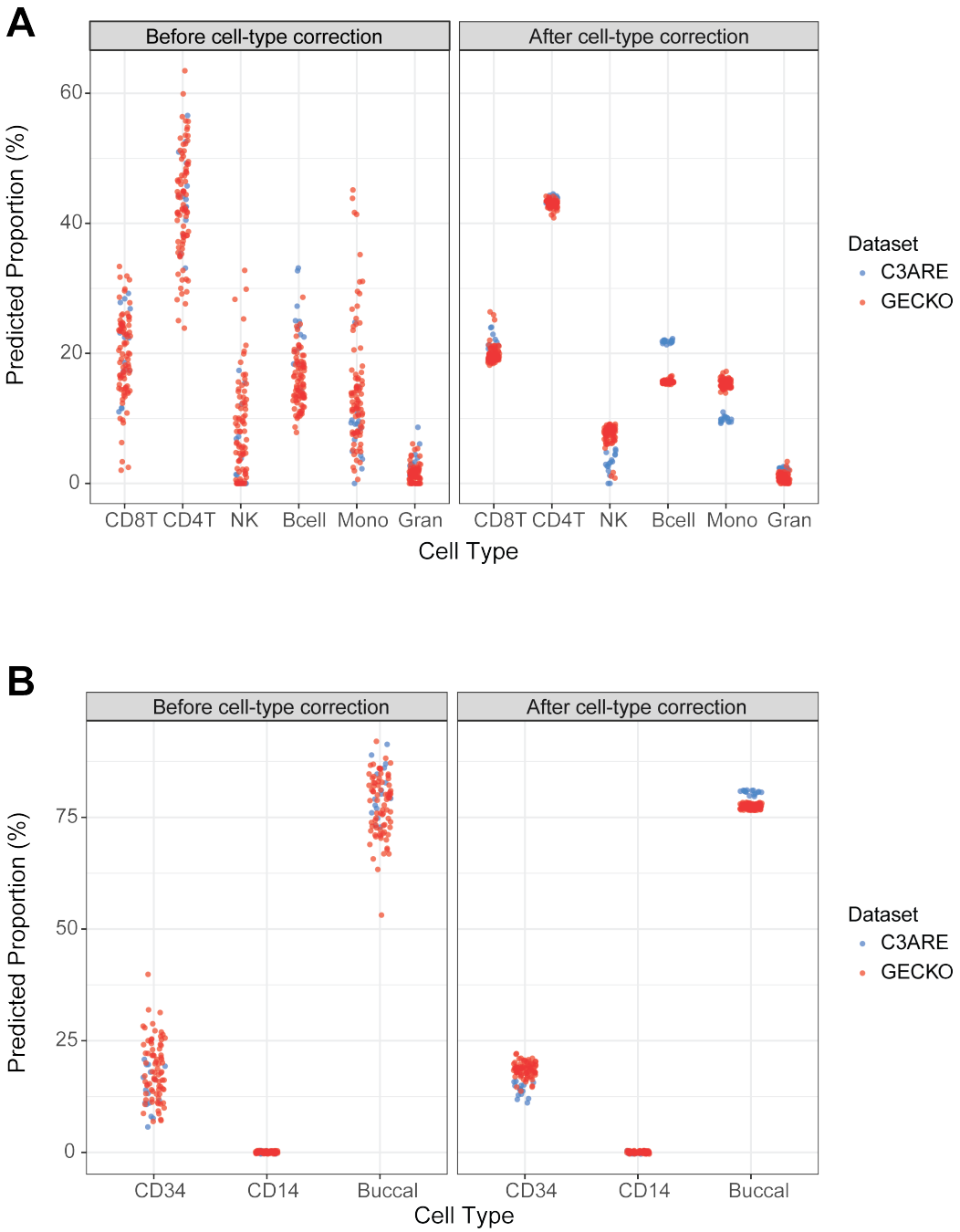
Abbreviations: Average beta, mean methylation values (%); Benjamini-Hochberg (BH) adjusted P-value.

Supplementary Table 4.4 Primers and PCR programs for validation and replication.

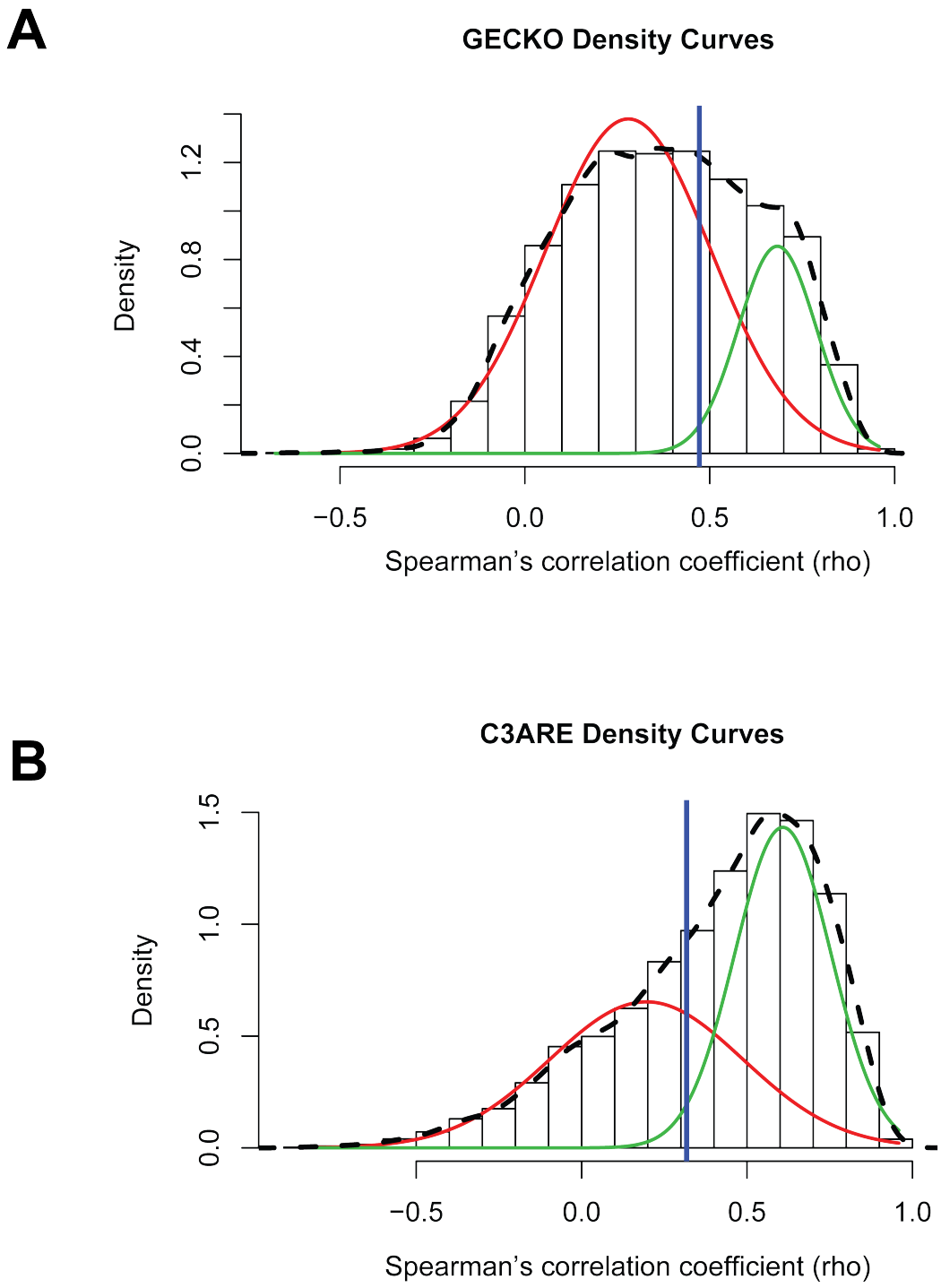
5cg07280807 (intergenic) fwd: 5'-GTTATGGTTGGGTTTTTGGG-3' rev: 5'-Bio-CCTATCTCCTCAAACAAAAACTAAAAA-3' seq: 5'-AGTTAGGGATTATAGTGTAGTTG-3' Amplicon length: 156 bp coordinates: chr14:70,317,178-70,317,333 Note: The amplicon contains 3 CpG sites, of which the third is cg07280807	**PCR program:** 95°C – 15 min ---------- 45 cycles: 94°C – 30 sec 50°C – 30 sec 72°C – 30 sec ---------- 72°C – 10 min 4°C – hold
cg18752527 (*HECW2*) fwd: 5'-GTGTTTGTGGGAATGTTTTTTATA-3' rev: 5'-Bio- CACACTACACTTTCATTTTCTATCAA-3' seq: 5'- TTTTTAGATATATAAATTTTTTTTTT-3' Amplicon length: 135 bp coordinates: chr2:197,132,798-197,132,932	**PCR program:** 95°C – 15 min ---------- 45 cycles: 94°C – 30 sec 50°C – 30 sec 72°C – 30 sec ---------- 72°C – 10 min 4°C – hold
cg16529483 / cg24496423 (*SRPK3*) fwd/seq: 5'-GTTATTTATAAAGGAGGGTGAGATTA-3' rev: 5'-Bio-AACCACTACTCCTATAAAACCCCAC-3' Amplicon length: 85 bp coordinates: chrX:153,046,424-153,046,508 Note: The amplicon contains 5 CpG sites, of which the first is cg16529483 and the fourth is cg24496423. Due to CpG sites in the primer binding area, the primers contain 1 (fwd) and 2 (rev) mismatches, which are highlighted underlined.	**PCR program:** 95°C – 15 min ---------- 45 cycles: 94°C – 30 sec 48°C – 30 sec 72°C – 30 sec ---------- 72°C – 10 min 4°C – hold
Abbreviations: fwd, forward primer; rev, reverse primer; seq, sequencing primer; Bio, biotin-modification; bp, basepair.	

Appendix D Supplementary Material for Chapter 5

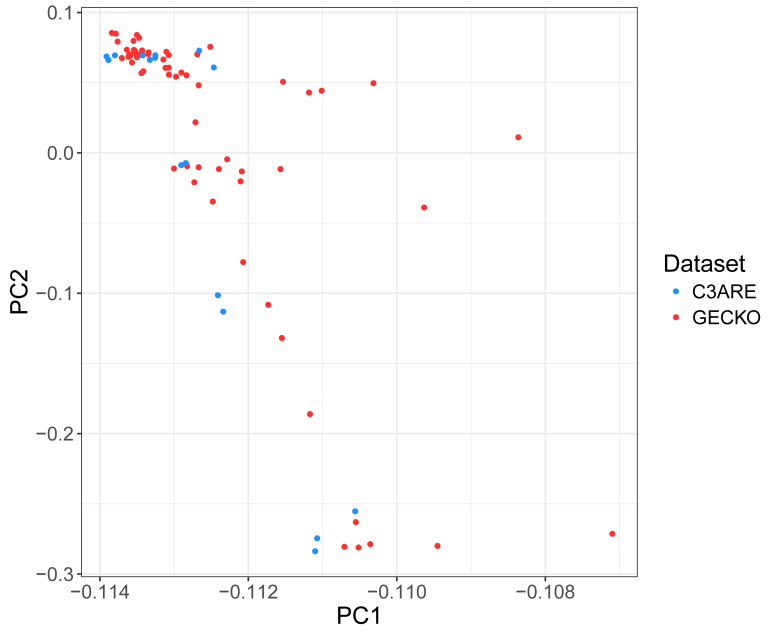
D.1 Supplementary Figures



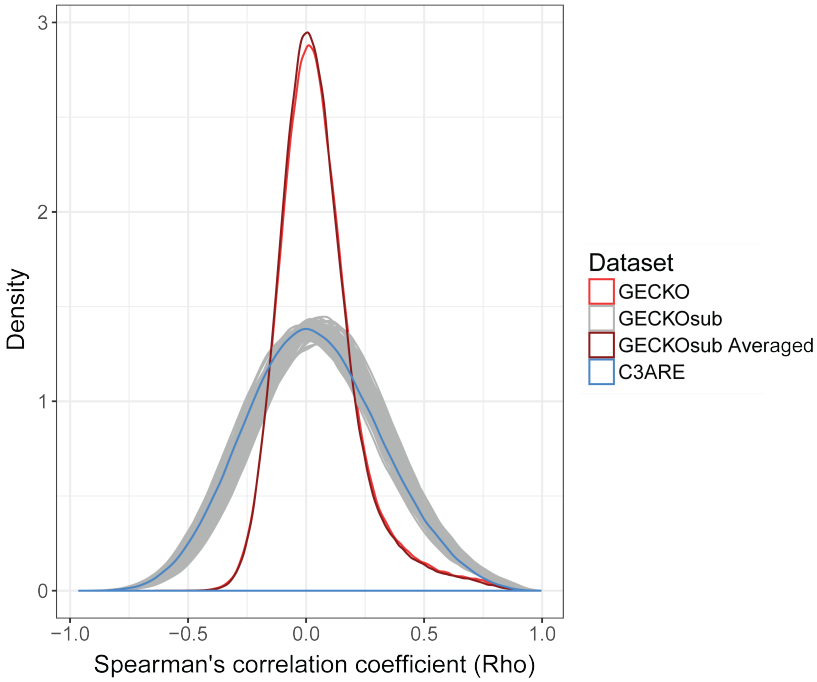
Supplementary Figure 5.1 A) Predicted proportions of cell types in PBMCs for both datasets (Mono = monocytes; Gran = Granulocytes). B) Predicted proportions of cell types in BECs for both datasets.



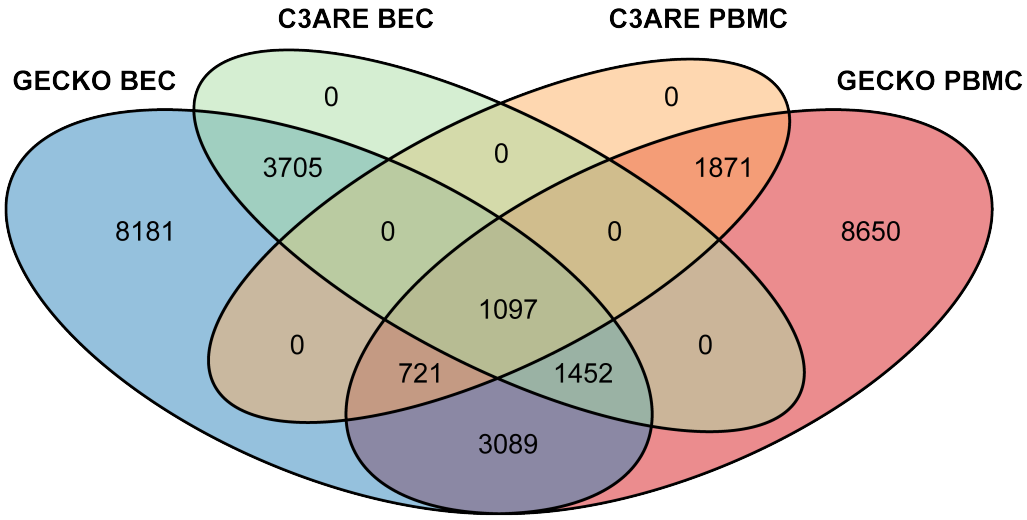
Supplementary Figure 5.2 Beta mixture modeling on Spearman correlation rho values between matched BECs and PBMCs for A) GECKO and b) C3ARE cohorts. The bimodal distribution of Spearman rho values indicated two underlying populations of CpGs, a set of uncorrelated CpGs (shown in red) and a set of right-skewed highly positively correlated CpGs (shown in green). Correlation coefficient threshold for informative CpGs were determined at 2 standard deviations minus the mean of the green Gaussian distribution (GECKO rho = 0.47; C3ARE rho = 0.32).



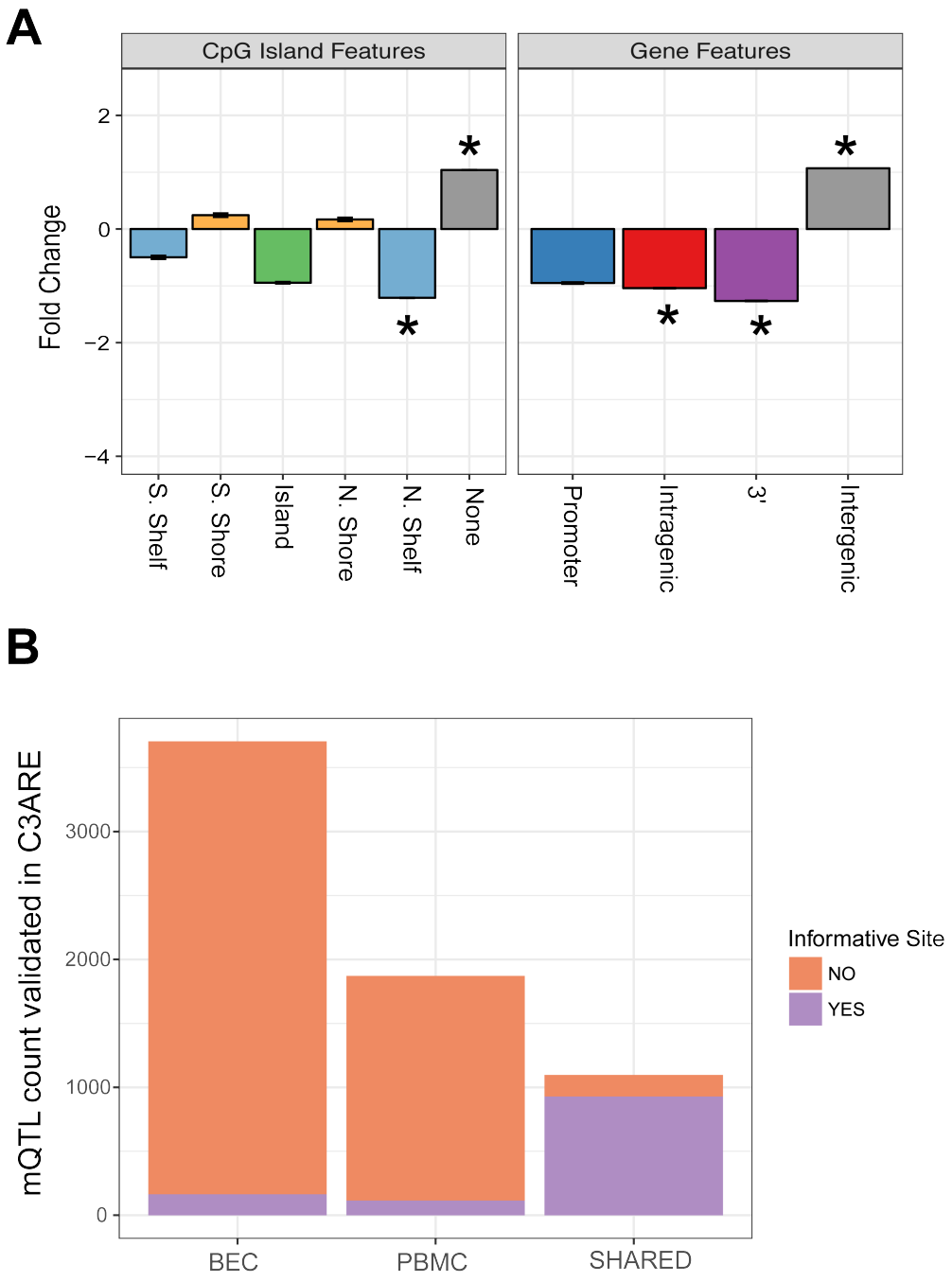
Supplementary Figure 5.3 Principal component analysis of PsychChip genotyping profiles (542,699 SNPs) for C3ARE (shown in blue) and GECKO (shown in red) revealed that genetic ancestry did not differ significantly between the cohorts as determined by Wilcoxon ranked sum test of GECKO versus C3ARE in PC1 scores (p = 0.8) and PC2 scores (p = 0.4).



Supplementary Figure 5.4 Density distribution of Spearman's correlation coefficient (Rho) across 419, 507 CpGs in matched BEC and PBMC tissues for GECKO, GECKOsub, GECKOsub Averaged (mean of 100 trials of GECKOsub) and C3ARE datasets.



Supplementary Figure 5.5 Overlap of *cis*-mQTL identified in matched tissues of both C3ARE and GECKO cohorts, respectively.



Supplementary Figure 5.6 A) Representation of 4,980 CpGs underlying validated *cis*-mQTL across various genomic features. Bars show the fold-change between CpG count in each genomic region and the mean count of randomly selected CpGs in that same genomic feature, from 10,000 iterations. Error bars show standard error. (* denotes significant enrichment or depletion at FDR ≤ 0.05) (S = South; N = North). B) A) Stacked bar plot representing overlap of identified informative sites in BEC-specific, PBMC-specific and shared-tissue validated *cis*-mQTL.

D.2 Supplementary Tables

Supplementary Table 5.1 The number of CpG sites at various thresholds of Spearman's correlation rho and reference range for C3ARE, GECKO and GECKOsub datasets.

C3ARE		PBMC Reference Range				
		0	**≥ 0.05**	**≥ 0.1**	**≥ 0.2**	**≥ 0.5**
Positive Correlation Rho	**0**	419,507	64,204	12,218	1,742	46
	≥ 0.3	74,137	24,542	7,474	1,282	44
	≥ 0.6	10,476	6,489	3,112	747	29
	≥ 0.9	45	45	35	9	0
Negative Correlation Rho	**≤ -0.3**	51,436	4,866	372	40	0
	≤ -0.6	3,300	327	34	6	0
	≤ -0.9	3	0	0	0	0
GECKO		**PBMC Reference Range**				
		0	**≥ 0.05**	**≥ 0.1**	**≥ 0.2**	**≥ 0.5**
Positive Correlation Rho	**0**	419,507	131,227	28,311	3,597	159
	≥ 0.3	333,22	29,158	15,774	3,055	146
	≥ 0.6	6,285	6,174	5,355	1,985	119
	≥ 0.9	41	41	41	38	4
Negative Correlation Rho	**≤ -0.3**	1,557	331	82	7	0
	≤ -0.6	8	8	6	0	0
	≤ -0.9	0	0	0	0	0
GECKOsub		**PBMC Reference Range**				
		0	**≥ 0.05**	**≥ 0.1**	**≥ 0.2**	**≥ 0.5**
Positive Correlation Rho	**0**	419,507	115,404	21,563	2,689	93
	≥ 0.3	30,615	26,385	12,916	2,306	88
	≥ 0.6	5,252	5,172	4,381	1,476	76
	≥ 0.9	11	11	11	10	0
Negative Correlation Rho	**≤ -0.3**	1,357	243	53	5	0
	≤ -0.6	5	5	4	0	0
	≤ -0.9	0	0	0	0	0

Supplementary Table 5.2 The number of CpG sites of each defined category represented in reported significant hits of various pediatric EWAS publications

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