THE HLA COMPLEX GROUP 9 GENE MODIFICATION STUDY IN MAJOR PSYCHOSIS

by

Mrinal Pal

A thesis submitted in conformity with the requirements for the degree of
Doctor of Philosophy
Graduate Department of Pharmacology and Toxicology
University of Toronto

THE HLA COMPLEX GROUP 9 GENE MODIFICATION STUDY IN MAJOR PSYCHOSIS

Mrinal Pal

Doctor of Philosophy

Pharmacology and Toxicology

University of Toronto

2015

Abstract

Epigenetic mechanisms can offer new insights into the non-Mendelian features of complex psychiatric disorders. The first epigenome-wide scan in major psychosis, conducted by our group, identified DNA modification differences at HLA complex group 9 (non-protein coding) gene (*HCG9*) in the *post-mortem* brain samples of individuals affected with schizophrenia (SCZ) and bipolar disorder (BPD). In this thesis we present results from bisulfite pyrosequencing based fine mapping of a ~700 bp region of *HCG9* in 1,402 DNA samples from *post-mortem* brain, germline (sperm), and peripheral white blood cells (WBC) of SCZ and BPD patients as well as unaffected controls. We observed significant differences in CpG modification between BPD samples and controls across all tissues and demonstrated the utility of WBC DNA modification density as an epigenetic marker for BPD. We further extended our analysis to full length *HCG9* with bisulfite padlock probe-based deep sequencing in SCZ, BPD and control *post-mortem* brain samples. We identified significant differences in CpH modification between sense and anti-sense

DNA strands and between major psychosis patients and controls.

Next, we characterized genome-wide distribution of 5-methylcytosine (5-mC) and 5-hydroxymethylcytosine (5-hmC) in neural and non-neural tissue from mice and humans. 5-hmC has been recently found to be abundant in brain but its function is poorly understood. We assayed 5-hmC by using glucosylation coupled with restriction enzyme digestion and microarray analysis. We detected enrichment of 5-hmC in genes with synapse-related functions and tissue-specific differential distribution of 5-mC and 5-hmC at the exon-intron boundary. This boundary change was mainly due to 5-hmC in the brain but due to 5-mC in non-neural tissue. Our study suggests a new role for 5-hmC in RNA splicing and synaptic function in the brain. Finally, we determined 5-mC and 5-hmC modification density at *HCG9* CpG6 in *post-mortem* brain tissue and WBCs of SCZ, BPD and control samples. Consistent with our previous report, we identified significant differences between the major psychosis group and controls, which were due to 5-hmC but not 5-mC as it was previously assumed.

Acknowledgements

I would first like to thank my supervisor, Dr. Arturas Petronis for his patience, support and guidance throughout my PhD program. I would also like to thank my PhD committee members, Dr. Albert Wong and Dr. Daniel Mueller for providing their advice during our annual meetings which helped me steer in the right direction. Several people in the lab contributed towards my work and greatly improved my understanding of epigenetics. I would especially like to thank Dr. Tarang Khare and Dr. Rene Cortese who trained me in basic lab techniques and guided me in my initial years and Dr. Zachary Kaminsky, Dr. Shraddha Pai and Dr. Gabriel Oh who helped my understanding of many concepts in statistics. I would also like to thank Dr. Peixin Jia and fellow graduate students in the lab, Sasha Ebrahimi, Andrew Kwan and Richie Jeremian, for their contribution towards my work and for being good friends throughout the years. Additionally, I am thankful to my wife, Aanchal Pal, and my brother, Bharat Pal, this work would not have been possible without their constant love, support and care. I am very thankful to my parents who provided me with the freedom to pursue whichever career and life path I wanted to pursue. I dedicate this thesis to them.

Contributions

For results presented in Section 3.1, I contributed to 8,412 bisulfite pyrosequencing assays with Mamoru Tochigi and Peixin Jia. Zachary Kaminsky and Sun-Chong Wang performed all data analysis. These results were published in the journal Molecular Psychiatry.[1]

For results presented in Section 3.2, Sasha Ebrahimi generated the padlock probe sequences. I generated the sequencing library for bisulfite padlock probe-based sequencing with Sasha Ebrahimi and Aiping Zhang. Sun-Chong Wang generated cytosine modification density data by performing assessment and alignment of sequence reads. I performed all data analysis presented in this section. These results were published in the journal Schizophrenia Bulletin.[2]

For results presented in Section 3.3, I performed quantitative PCR experiments for validation of glucosylation and restriction enzyme-based enrichment of 5-hmC. Thin layer chromatography and artificial oligonucleotide experiments were performed by Edita Kriukiene and Zita Liutkeviciute. I also conducted 774 restriction enzyme-based 5-hmC and 5-mC+5hmC enrichment and microarray experiments with Tarang Khare and Peixin Jia. MBD-based 5-mC enrichment and Helicos single molecule sequencing experiments were performed by Tarang Khare. SAHA treated cell lines were generated by Menghang Xia and Raymond Tice. All data analysis was performed by Shraddha Pai, Karolis Koncevicius, Sun-Chong Wang, Philipp Kapranov and Rafal Kustra. These results were published in the journal Nature Structural and Molecular Biology.[3]

I performed all experiments and data analysis for results presented in Section 3.4.

Table of Contents

List of Tables

List of Figures

List of Appendices

List of Abbreviations

3' UTR	3' untranslated region
5-caC	5-carboxylcytosine
5-fC	5-formylcytosine
5-hmC	5-hydroxymethylcytosine
5-mC	5-methylcytosine
ADHD	Attention-deficit/hyperactive disorder
ANK3	Ankyrin3
ANOVA	Analysis of variance
AP site	Apurinic/apyramidinic site
AS	Angelman syndrome
ATP	Adenosine triphosphate
A^{vy}	Agouti viable yellow
BER	Base excision repair
BGT	β-glucosyltransferase
BPD	Bipolar disorder
BS	Bisulfite
BSA	Bovine serum albumin
BSPP-seq	Bisulfite padlock probe-based sequencing
CACNA1C	L-type voltage-gated calcium channel, alpha 1C subunit
CACNB2	L-type voltage-gated calcium channel, beta 2 subunit
CGI	CpG island
CNV	Copy Number Variation
DAVID	Database for annotation, visualization and integrated discovery
DNA	Deoxyribonucleic acid
DNMT	DNA methyltransferase
DSM	Diagnostic and statistical manual of mental disorders
dNTP	Deoxynucleotide triphospate
DZ	Dizygotic
EEJ	Exon-exon junction
ESC	Embryonic stem cell
EST	Expressed sequence tag
FDR	False discovery rate
GAD1	Glutamate decarboylase 1
GCTA	Genome complex trait analysis
gDNA	Genomic DNA
GEO	Gene expression omnibus
glu*Msp*I	Glucosylated gDNA digested with *Msp*I
GO	Gene ontology
GWAS	Genome-wide association studies
H3K4me3	Histone 3 lysine 4 trimethylation

H3K9me2	Histone 3 lysine 9 demethylation
HAT	Histone acetyltransferase
HCG9	HLA complex group 9 (non-protein coding)
HDAC	Histone deacetylase
HDM	Histone demethylase
HLA	Human leukocyte antigen
HMT	Histone methyltransferase
HPA axis	Hypothalamic-pituitary-adrenal axis
IGF2	Insulin-like growth factor 2
iPSC	Induced pluripotent stem cell
MAT	Model-based analysis of tiling arrays
MBD	Methyl-binding domain
McL	Harvard Brain Tissue Resource Center, McLean Hospital
MDD	Major depressive disorder
MECP2	Methyl-CpG-binding protein 2
MGI	Mouse genome informatics
^{mod}C	Modified cytosine
mRNA	messenger RNA
MZ	Monozygotic
NCAN	Neurocan
NGS	Next generation sequencing
NR3C1	Nuclear receptor subfamily 3, group C, member 1
ODZ4	Odz, Odd Oz/Ten-M homolog 4 (Drosophila)
OR	Odds ratio
ORA	Overrepresentation analysis
PBEF1	Nicotinamide phosphoribosyltransferase
PCR	Polymerase chain reaction
PFC	Prefrontal cortex
PGC	Psychiatric genomics consortium
PWS	Prader-Willi syndrome
qPCR	Quantitative PCR
RBC	Red blood cell
RDoC	Research domain criteria
RNA	Ribonucleic acid
ROC curve	Receiver operating characteristic curve
RYR2	Ryanodine receptor 2
SAHA	Suberoylanilide hydroxamic acid
SCZ	Schiophernia
SMRI	Stanley Medical Research Institute
SMS	Single molecule sequencing
SNP	Single nucleotide polymorphism
TDG	Thymine-DNA glycosylase
TET	Ten-eleven translocation

TLC	Thin layer chromatography
WBC	White blood cell
WGBS	Whole genome bisulfite sequencing
WMW test	Wilcoxon-Mann-Whitney test
ZNF804A	Zinc finger protein 804A
α-KG	α- ketoglutrate

Chapter 1. Introduction

1.1. Statement of the problem

Family, twin, and adoption studies have shown that both heritable and non-heritable risk factors contribute to psychiatric diseases. DNA sequence variants identified in large genome-wide association studies of schizophrenia and bipolar disorder can explain only a small fraction of heritable predisposition to these diseases. Similarly, epidemiological studies experience significant problems in uncovering specific and causal exogenous risk factors.

Along with genetic and environmental causes, a third group of etiological factors, namely acquired and inherited epigenetic misregulation, could play a role in major psychosis. Epigenetics refers to heritable and potentially reversible modifications of DNA and histone proteins that regulate gene expression and other genomic functions without changing the underlying DNA sequence. Putative epigenetic misregulation can offer new insights into a number of non-Mendelian complexities of major psychosis such as fluctuating course, discordance of monozygotic twins, sex- and parent of origin-effects. However, epigenetic studies in human diseases are at an early stage and pose new challenges. First, epigenetic profiles are tissue specific and investigation of the affected tissue is the most appropriate choice for the disease of interest. For psychiatric disorders, the shortage of available brain tissues represents a significant limiting factor for epigenetic studies. Second, sample size consideration for epigenetic studies depends on the effect size of identified risk factors which is yet unclear in the absence of precedent studies. Third, multiple layers of DNA and histone modifications have been identified and it is unclear which is most relevant to disease risk. Lastly, major technological advancements in recent years have enabled successful integration of genome-wide microarray and deep sequencing based approaches for epigenetic analysis. The choice of an appropriate method depends on the amount of genomic coverage and the resolution, with only few methods that provide the right balance between the two and can be effectively applied to populational epigenomic studies in a cost effective manner. We dealt with all these questions in the detailed DNA modification study of human leukocyte antigen (HLA) complex group 9 (non-protein coding) gene (*HCG9*) in major psychosis, and our solutions and results are provided in this thesis.

1.2. Major psychosis

Schizophrenia (SCZ) and bipolar disorder (BPD) are two related forms of severe mental

disorders, which are together termed "major psychosis". The lifetime prevalence of SCZ and BPD is estimated at 1% and 3% in the general population, respectively.[4-6] SCZ usually affects people in young adulthood with patients suffering from cognitive deficits along with an individual combination of positive and negative symptoms.[7] Positive symptoms mainly consist of delusions, hallucinations, disorganised speech and behaviour along with thought disorder while negative symptoms consist of alogia, anhedonia, asociality, avolition and blunted affect.[8,9] The severity of symptoms depends on the disease stage and can change over time. Remission is often incomplete with persistent negative and positive symptoms observed in 30% of patients.[10] Impaired cognitive performance, mainly observed in memory, attention and executive task, forms the core feature of SCZ and is also observed in BPD patients.[11-14] BPD is characterized by unusual and extreme shifts in mood with symptoms including social withdrawal, melancholia and loss of energy during depressive episodes, which are also observed in SCZ. Manic episodes include elation, increased energy, irritability, rapid speech and decreased need for sleep; hallucination and delusion are often observed during manic episodes.

The Diagnostic and Statistical Manual of Mental Disorders (DSM) is a widely used classification system for clinical diagnoses of psychiatric disorders and to guide in treatment strategies.[15] However, due to lack of objective diagnostic tests the DSM classifications are largely descriptive and depend on the quantity and quality of the signs and symptoms. Moreover, multiple signs and symptoms overlap between disorder categories with accumulating evidence of genetic overlap between disorders (discussed in detail in Section 1.2.1). These findings challenge the validity of the currently used classification system in clinical practice and research. The Research Domain Criteria (RDoC) is a new initiative started by the National Institute of Mental Health (USA) which seeks to establish a nosology for mental disorders based on behavioural dimensions and neurobiological measures.[16] The primary objective of the RDoC initiative is to investigate basic neurobiological mechanism at multiple levels (for example genetic, cellular, neuronal circuits, and behaviour) and how their dysfunction contributes to disease development. By focusing on core psychological functions the RDoC initiative can help us better understand homogenous symptoms across several disorders and potentially develop targeted treatment options.

1.2.1. Genetic approaches in major psychosis

Over the past several decades the majority of research has focused on elucidating the role of genetic factors in the etiology of complex psychiatric disorders. Indeed, it has been determined

that genetic factors play an important role in development of major psychosis. Epidemiological studies using twin, family and adoption study design established the large heritable component of major psychosis with heritability estimates of 82-85% for SCZ and 79% for BPD.[17-19] In recent years, genome-wide association studies (GWAS) have identified several polymorphisms, ranging from large chromosomal rearrangements to small DNA sequence variations in SCZ and BPD. International collaborative efforts, such as the Psychiatric Genomics Consortium (PGC), have enabled GWAS studies of unprecedented size, therefore providing sufficient power to detect risk variants of small effect size. The recently published landmark paper by the Schizophrenia Working Group of PGC (2014) compared 36,989 cases and 113,075 controls and identified 108 SCZ risk loci of which 83 were not previously described.[20] Consistent with previous studies,[21-23] risk loci overlapping genes encoding calcium channels, and proteins involved in synaptic plasticity and glutamatergic neurotransmission were identified. Meta-analysis of 11 BPD GWAS studies with a combined sample size of 11,974 cases and 51,792 controls confirmed the significant association of the alpha 1C subunit of the L-type voltage-gated calcium channel *(CACNA1C)* and identified a novel risk locus located in the human homologue of a Drosophila pair-rule gene *odz* (*ODZ4*).[24] However, variants of common single nucleotide polymorphisms (SNPs) to date, show modest effect sizes (relative risk < 1.3) indicating that multiple SNPs may act together to influence risk of psychiatric disorders.

High-risk family studies suggested that some etiological factors are common between SCZ, BPD and major depressive disorder (MDD). Recent meta-analysis conducted by Rasic *et al.* (2014) compared the risk of BPD, SCZ and MDD in offspring of parents with these disorders.[25] The study not only revealed a 32% probability for the offspring to develop one of these disorders but also provided evidence for cross-disorder risk. For example, offspring of SCZ or BPD patients had significantly higher risk of developing MDD, SCZ and BPD. Moreover, it has been shown that risk alleles that are over-represented in SCZ cases are also over represented in BPD.[26] A recent study published by the Cross-Disorder Group of the PGC identified four SNPs that were significantly associated with SCZ, BPD and MDD.[27] Cross-disorder association between SCZ and BPD cases was also previously reported for SNPs located in ankyrin3 (*ANK3*), zinc finger protein (*ZNF804A*), neurocan (*NCAN*), *CACNA1C*, and beta 2 subunit of the L-type voltage-gated calcium channel (*CACNB2*).[28-33] A few studies have also demonstrated increased load of rare copy number variants (CNVs) (present in < 1% of the population) in multiple developmental and psychiatric disorders, including SCZ, autism spectrum disorder (ASD), attention-deficit/hyperactivity disorder (ADHD) and intellectual disability.[34-39] In contrast, low

burden of rare CNVs has been identified in BPD although *de novo* CNVs and singleton deletions occur at higher rates compared to controls.[40-44] Microdeletions and duplications identified in SCZ (reviewed in[45]) have been reported for genomic loci that contain genes vital for neuronal migration, synaptic signalling, myelination and neurotransmitter metabolism.

Despite the high heritability estimates of SCZ and BPD in twin and family studies, GWAS can only explain a small fraction of the genetic variation in liability to major psychosis. Yang *et al.* (2011) developed the genome complex trait analysis (GCTA) method which looks beyond single SNP associations by using an additive model for all SNPs to estimate the genetic variance for a complex trait.[46] Applying the GCTA method, Lee *et al.* (2013) estimated that common variants contribute 23% and 25% to the variation in liability to SCZ and BPD, respectively.[47] Multiple explanations have been provided for this missing heritability also termed the "dark matter" of GWAS, in the sense that one is sure it exists and can detect its influence but cannot 'see' it. It has been proposed that missing heritability can be due to a large number of genetic variants, with small effect sizes, that are yet to be discovered; rare variants, with possibly large effect sizes, which mostly remain undetected due to current technology that focuses on variants present in > 1-5% of the population; inability to accurately determine gene-gene interactions due to low power; and interaction of environmental risk factors with genetic background.

1.2.2. Environmental factors implicated in major psychosis

Several studies have found association of environmental risk factors to major psychosis. The most replicated finding has been the dose-response association of SCZ with urban environment after adjusting for genetic risk factors.[48-50] Furthermore, a reduction in risk of psychotic outcome was observed in longitudinal studies after changing the environmental exposure, for example, moving from urban to rural environment during childhood.[51] Stressful events during early childhood including physical, sexual and emotional abuse have been shown to predict depressive symptoms in adulthood.[52] Cannabis use has been shown to result in exaggerated psychotic response in individuals at high genetic risk for psychosis.[53] Multitude of specific adverse prenatal environmental exposures including prenatal maternal nutritional deficiency and stress, maternal serum lead and homocysteine levels, low and high neonatal vitamin D, rhesus incompatibility, prenatal toxoplasmosis, and pregnancy and birth complications have been reported to increase risk of adult psychotic outcomes.[54]

Given these findings it has been proposed that gene-environment interactions play a vital

role in the etiology of major psychosis. The phenotypic outcome can result from risk alleles that are sensitive to environmental exposure or, vice versa, environmental risk factors acting on an individual's genotype resulting in predisposition to major psychosis.[55,56] Furthermore, gene-environment interactions may explain missing heritability, since familial shared environmental factors are included in heritability estimates of twin studies but excluded from genetic studies of unrelated individuals. This notion has been challenged by studies that show monozygotic twins reared apart and together show similar correlations for various behavioural characteristics.[57,58] Although recent studies have attempted to discover gene-environment interactions these only provide statistical associations with limited attempts at replication and minimal insight into the molecular and cellular mechanisms that underlie such interactions.[59-61]

1.2.3. Beyond the traditional “genes and environment” paradigm

Complex diseases, unlike simple Mendelian traits, display irregular modes of inheritance, discordance of MZ twins, sexual dimorphism, parent-of-origin effects, environmental influence, and fluctuating disease course, among other non-Mendelian features. These peculiarities of complex disorders can be explained by the concept of "epigenetic misregulation" that looks beyond the conventional etiological paradigm of "DNA + environment" and "DNA x environment". Epigenetic regulation is tightly coupled to a large array of genetic and genomic functions and, therefore, epigenetic studies of normal and diseased brains may shed new insights into the molecular etiopathogenesis of SCZ, BPD and other complex psychiatric diseases. The epigenetic theory does not deny the putative role of DNA sequence variation in complex diseases, but rather suggests that research into epigenetic and genetic factors must be conducted in parallel to better understand the molecular etiology of complex psychiatric disease.

1.3. Epigenetics

Epigenetics refers to heritable, but reversible changes in DNA modification (modification of cytosine residue) and chromatin structure (post-translational modifications of histone proteins) that regulate various genomic functions, including gene expression, without altering the underlying DNA sequence. Two types of epigenetic mechanisms are detailed below: namely, DNA and histone modification.

1.3.1. DNA modification

For mammalian genomes, DNA modification has been conventionally referred to as

methylation of a cytosine residue within CpG dinucleotides. However, recently multiple forms of cytosine modification have been discovered in the mammalian genome within CpG and CpH dinucleotides (where H refers to A, T or C), these are discussed below.

1.3.1.1. Cytosine modification in CpG context

In the mammalian genome there are four different types of known cytosine modifications: 5-methylcytosine (5-mC), 5-hydroxymethylcytosine (5-hmC), 5-formylcytosine (5-fC) and 5-carboxylcytosine (5-caC). 5-mC is by far the most well characterized form of DNA modification. CpG sites are asymmetrically distributed between GC rich and poor regions of the mammalian genome and approximately 70-80% are modified.[62] However, the pattern of cytosine modification varies between different tissue and cell types.[63,64] DNA regions that contain a high frequency of CpG sites, also known as CpG islands (CGIs), are preferentially located near the transcription start sites of the house keeping genes and are largely free of 5-mC due to the abundance of GC-rich transcription factor-binding sites.[65] Presence of 5-mC at CGIs located within gene promoters prevents transcription factor binding due to steric hindrance and higher order nucleosome compaction resulting in repression of gene transcription. Beyond transcriptional regulation, 5-mC has been implicated in multiple biological processes including transposable element suppression, X-chromosome inactivation and genomic imprinting.[62,66,67]

5-mC is formed by covalent attachment of a methyl (CH_3) group to C5 position of a cytosine residue by the DNA methyltransferase (DNMT) family of enzymes; DNMT1, DNMT2, DNMT3A, DNMT3B and DNMT3L (Figure 1.1).[68] DNMT1 preferentially binds to hemimethylated cytosine at CpG dinucleotides and functions as a maintenance methyltransferase during replication.[69] DNMT3A and 3B act as *de novo* enzymes by establishing methylation marks on unmodified cytosines during epigenetic reprogramming of germ cells and after fertilization.[70,71] DNMT3L lacks catalytic activity but facilitates methylation by increasing DNMT3A and 3B binding ability to methyl groups.[68,72] DNMT2 primarily functions as RNA methyltransferase and also displays weak DNA methyltransferase activity.[73,74]

The recent rediscovery of 5-hmC in mammals demonstrates that covalent DNA modifications are more dynamic than previously assumed.[75-77] Relative to other tissues, 5-hmC is particularly enriched in the brain and embryonic stem cells (ESCs), as observed in mice and humans, accounting for 0.4 - 0.7% of all cytosines compared to 10% for 5-mC.[76-78] 5-hmC is generated by the oxidation of 5-mC by the ten-eleven translocation (TET) family of enzymes (TET 1, 2 or 3) and acts as an intermediate in functional and active DNA demethylation (Figure

1.1).[79,80] TET mediated oxidation of 5mC requires α-ketoglutrate (α-KG), molecular oxygen and iron as cofactors.[81] α-KG is a key intermediate in the Krebs cycle and is produced by the mitochondrial enzyme isocitrate dehydrogenase by oxidation of isocitrate. Recent evidence suggests that TET enzymes can further catalyze oxidation of 5-hmC to 5-fC and 5-caC, which is subsequently replaced with an unmodified cytosine by thymine-DNA glycosylase (TDG) and the base excision and repair (BER) pathway.[82,83] However, if 5-hmC is an intermediate to 5-fC and 5-caC, an intriguing question may be posed: why are 5-hmC levels 10-100 fold larger than 5-fC or 5-caC?[83,84] In the mouse brain 5-hmC is enriched within genes and appears to increase with increasing transcription levels suggesting a biological role for 5-hmC.[85,86] Furthermore, in ESCs enhancers are enriched in 5-hmC whereas, CpG island rich promoters are devoid of 5-hmC.[87]

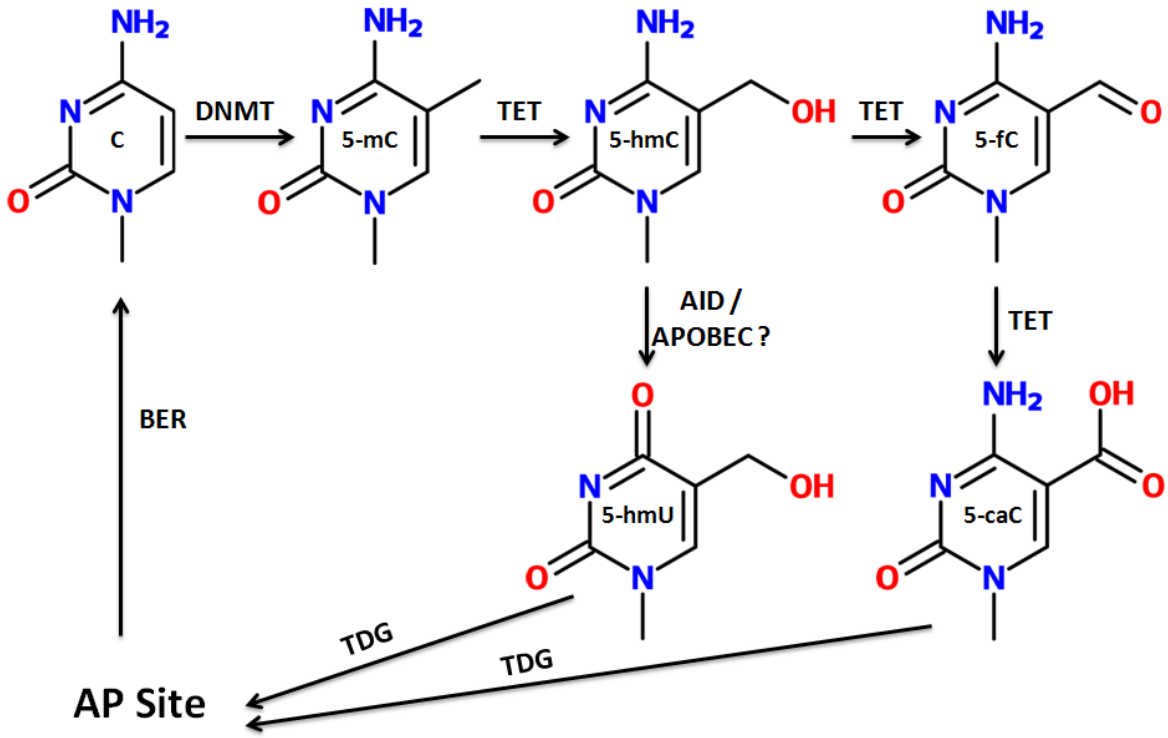

Figure 1.1. Active demethylation of 5-mC in mammalian cells.
Three DNA methyltransferase (DNMT) proteins (DNMT 1, DNMT 3A and DNMT 3B) generate 5-mC by covalent transfer of a methyl group from the donor S-adenosyl methionine (SAM) to cytosine residue. The 5-methyl group can be oxidized by ten-eleven translocation (TET) family of enzymes (TET 1, TET 2 and TET 3) to generate 5-hmC. Further oxidation of 5-hmC by TET results in production of 5-fC and 5-caC which is recognized and subsequently removed by thymine-DNA glycosylase resulting in a apurinic/apyramidinic site (AP site). AP sites form substrates for the base excision and repair (BER) pathway which replace it with an unmodified cytosine. Alternatively, it has been proposed that deamination of 5-hmC to 5-hydroymethyluracil (5-hmU) occurs via activation induced deaminase (AID) and apolipoprotein B mRNA editing enzyme (APOBEC) triggering TDG followed by reintroduction of cytosine by BER.[88] However, this remains controversial since 5-hmU might result from oxidation of thymine by TET and recombinant AID shows no activity on 5-hmC *in vitro*.[89,90]

1.3.1.2. Cytosine modification in non-CpG context

Modification of cytosine within CpH dinucleotide is referred to as non-CpG modification and is frequently found in plants.[91,92] More recently it has also been described in a few mammalian cell types including ESCs, induced pluripotent stem cells (iPSCs), and brain cells.[93-102] Human DNA methyltransferases DNMT3A/B and DNMT3L (Dnmt3a/b and Dnmt3l in mice), responsible for *de novo* methylation at CpG sites, are also linked to CpH methylation. It has been proposed that sustained expression of these proteins, particularly in non-mitotic cells such as neurons, results in CpH methylation.[99,102] *In vitro* studies with DNMT3A suggest its substrate specificity decreases in the order CpG > CpA > CpT >CpC which correlates with relative modification density observed at these sites in mammalian cells.[97,103] *In vivo* experiments on mouse ESCs revealed significant reduction in CpH modification after Dnmt3a, Dnmt3b and Dnmt3l knockout while no changes in the modification profile were observed after Dnmt1 knockout.[97,104] Similar experiments conducted on human ESC (hESC) lines showed 33% and 82% reduction in CpA modification after DNMT3A and DNMT3B knockdown, respectively; while no changes were observed in CpG modification density.[96] Experiments conducted by Aoki *et al.* (2001) on purified recombinant Dnmt3a and Dnmt3b demonstrated that these enzymes were able to methylate CpA to 7% and 28% of CpG methylation level, respectively.[105] CpH sites are distributed asymmetrically across the genome and the inability to bind DNMT1 indicates that there is less likely to be a mechanism for maintenance of CpH methylation. Therefore, such marks would have to be re-established after every cell division.

In hESCs, CpH modification accounts for 20-25% of all cytosine modification and decreases progressively with each stage of differentiation.[93,94] Laurent *et al.* (2010) compared genome-wide CpH methylation profiles in three human cell lines representing progressive stages of differentiation: hESCs, fibroblastic differentiated derivative of hESCs and fully differentiated derivative of peripheral mononuclear cells.[94] Highest level of CpH modification was observed in hESCs (accounting for 20% of all cytosine modification) with gradual decline in more differentiated cell lines (<10% of all cytosine modification in fully differentiated monocyte cell line). CpH methylation occurs predominantly at CpA dinucleotide accounting for 10% of all cytosine modification in ESCs declining to 2% in fully differentiated monocyte cell line.[94] Ziller *et al.* (2011) confirmed this finding by interrogating 76 genome-wide modification density maps across multiple pluripotent and differentiated human cell types.[96] The authors found a gradual decline in CpH modification with progression of differentiation with very low or absence of CpH modification in fully differentiated somatic cells. Interestingly, reoccurrence of CpH

modification pattern has been reported after reprogramming of somatic cells to iPSCs.[93,98]

With respect to CpH modification in differentiated cells, brain tissue is a notable exception with modification levels comparable to that of ESCs, accounting for 25% of all cytosine modification.[101] The similarities end here with both ESCs and brain tissue showing reproducible levels of CpH modification across multiple individuals but at distinct set of loci in both cell types.[102] Moreover, gene body CpH modification correlates with increasing steady state mRNA levels in ESCs while inverse correlation was observed in the brain.[100] These observations extend the role of CpH modification in guiding genomic function exclusively in pluripotent cells to differentiated cells.

1.3.2. Histone modification

Histone proteins and associated factors in the nucleus help form chromatin, which refers to the highly condensed packaging of genomic DNA (gDNA) in the nucleus.[106] Nucleosome forms the basic unit of chromatin with a small stretch of DNA (147 base pairs) wrapped around an octamer of pairs of core histone proteins (H2A, H2B, H3 and H4), and connected with other nucleosomes via linker histones (H1).[107,108] Post-translational modifications at the level of individual histone proteins and the nucleosome, along with other epigenetic mechanisms, provide chromatin with dynamic structural configurations that vary between closed or inactive (heterochromatin) to those that are open and active (euchromatin). This chromatin architecture is crucial for a diverse array of cellular processes including maintenance and fidelity of cell type specific transcription and translation, DNA replication and repair, genomic stability, long range interaction between gene regulatory elements and establishment of gene-dosage effects.[106,109-114]

Covalent post-translational modifications occur primarily at the amino-terminal tails of histone proteins at specific amino-acid residues (glutamate, threonine, arginine, lysine, serine) and include methylation, phophorylation, acetylation, SUMOylation, ubiquitylation, ADP-ribosylation, proline isomerization and deimination.[106] At a defined genomic locus, complementary histone modifications together form a combinatorial "histone code". Addition or removal of these modifications are catalyzed by an array of histone-modifying enzymes, such as histone deacetylases (HDACs), acetyltransferases (HATs), demethylases (HDMs), and methyltransferases (HMTs). These are often part of large multi-protein complexes which may include transcriptional proteins to regulate local transcriptional profiles.[106] For example, high gene expression is strongly linked with histone acetylation, whereas, different combinations of histone phosphorylation and methylation might result in gene repression or activation depending

on the residue on which the modification is present.[115,116] Histone and DNA modification profiles at a chromosomal locus do not operate in isolation, in fact, these are intricately linked to one another with substantial cross-talk between the two. For example, modified CpG sites can be recognized by methyl-binding domain (MBD) proteins, such as methyl-CpG-binding protein 2 (MeCP2), which recruit large protein complexes containing HMTs and HDACs resulting in downregulation of gene expression.[117,118]

1.4. The epigenetic model of major psychosis

There are three fundamental concepts that establish the epigenetic model for major psychosis and complex diseases in general.

1. The epigenetic status of a gene is highly dynamic and regulates the location and timing of gene expression. Epigenetic makeup is influenced by an organism's developmental program, internal and/or external environment and also by stochastic processes in the cell nucleus.[119-122]
2. Some epigenetic marks display partial meiotic stability i.e. they can be relayed from parent to offspring along with DNA sequence, which may account for trans-generational epigenetic heritability.[123-125]
3. The normal functioning of a cell is critically linked to both epigenetic factors and DNA sequence. A shortcoming in either of these could be equally detrimental to a cell, tissue and an individual.[70,126-129]

The epigenetic model of psychiatric disorders assumes that epigenetic misregulation at a specific gene (or genes) occurs during gametogenesis or embryogenesis. Owing to the dynamic nature of epigenetic states, this pre-epimutation(s) can be influenced by multiple pre- and post-natal factors, such as hormones, external environment, or stochastic events, which are tolerated by the cell to a certain extent. The diseased condition is caused only once a critical threshold of epigenetic misregulation is achieved beyond which the cell or tissue is unable to function normally. Thus it could take decades to achieve the threshold of epigenetic misregulation, and therefore only some predisposed individuals may develop the disease. The epigenetic model provides a new opportunity to explain a series of molecular, clinical and epidemiological findings in major psychosis.

1.4.1 Discordance of monozygotic twins

Epidemiological studies in identical (monozygotic, MZ) twins and fraternal (dizygotic,

DZ) twins have been used to demonstrate the genetic component of a trait or disease. Higher disease concordance in MZ twins compared to DZ twins argue for contribution of genetic factors. For BPD and SCZ, concordance rates between MZ twins are estimated at 67% and 50%, respectively, while DZ twins display 20% and 4.1% concordance, respectively.[19,130] Phenotypic differences in MZ twins has been traditionally interpreted as evidence for environmental interactions yet no study has shown consistent differences for such factors acting in discordant twins.[131-134]

These observations can be explained in terms of epigenetic differences that spread across the epigenome of MZ twin pairs.[135] Kaminsky *et al.* (2009) hypothesized that the observed epigenetic differences in MZ twins arise primarily due to stochastic reasons during mitotic transmission of epigenetic profiles in somatic cells.[135] The authors also provide evidence for epigenetic heritability, as DZ twins exhibited significantly larger DNA modification differences compared to MZ twins. These findings are unlikely to result from differences in DNA sequence; a more plausible explanation could be that DZ twins originate from epigenetically different zygotes whereas MZ twins share the same epigenome at the time of blastocyst splitting. Epigenomic differences in MZ twins discordant for SCZ and BPD have been documented in literature[136-138] and may offer a new explanation for full or partial discordance (i.e. variability in disease severity, age-of-onset, and drug response).

Female MZ twin discordance might also result from skewed X chromosome inactivation for X-linked recessive diseases.[139] In females, X chromosome inactivation is a dosage compensatory mechanism with karyo-typically normal males and is achieved by a combination of several levels of epigenetic regulation which include, DNA and histone modification, and non-coding Xist RNA expression. The selective criteria behind inactivation of either one of the two X chromosomes is not known but appears to be non-random and in some cases skewed to discriminate against harmful effects on cell growth.[140,141] Skewed X chromosome inactivation has been observed in approximately 50% of female carriers for X-linked mental retardation with the two X chromosomes exhibiting an activation ratio of 80:20% or higher.[142] In addition, Rosa *et al.* (2008) observed greater epigenetic variation between maternal and paternal X chromosome alleles in female BPD discordant versus non-discordant twin pairs, suggesting involvement of X-linked loci in BPD twin discordance.[138]

1.4.2. Sexual dimorphism

Sexual dimorphism is another non-Mendelian feature of complex disorders and refers to differential susceptibility to a disease in males and females.[143] Men are more often affected with SCZ compared to women (1.4:1 male to female ratio) and also exhibit an earlier age of onset, more deficit symptoms, more pre- and peri-natal complications, and poorer response to antipsychotic medication.[144-146] For BPD, higher incidence of cyclothymia, rapid cycling and mixed states are observed in women while early-onset BPD is more prevalent in men.[147,148] Such differences were earlier thought to be linked with genetic risk factors on sex chromosomes. This is plausible due to the possible role of skewed X chromosome inactivation discussed above (section 1.4.1), however, linkage and association studies also implicate autosomal genes to have sex effects.[149] For example, a GWAS study conducted by Wang *et al.* (2011) found significant association of two SNPs located within ryanodine receptor (*RYR2*, 1q43) gene and four SNPs located near nicotinamide phosphoribosyltransferase (*PBEF1*, 7q22.3) gene in SCZ females but not males ($p = 2 \times 10^{-6}$ and 4.5×10^{-6}, respectively).[150] Sex effects are commonly observed in other complex psychiatric disorders as well.[151] Linkage studies conducted by Abkevich *et al.* (2003) and Zubenko *et al.* (2003) in families with history of MDD provide evidence for male only linkage on chromosome 12q22-q23.2 and female only linkage at 2q33-35, respectively.[152,153] These gender-specific effects are usually attributed to sex hormones due to their crucial role in various regulatory processes and association with disease states.[149]

Sexual dimorphism observed for autosomal genes can be explained by epigenetic changes brought about by sex hormones. Of particular interest are the estrogen and androgen receptors that regulate gene expression by recruiting large protein complexes containing histone modifying enzymes, such as HDACs, HMTs, and HATs, in response to ligand binding. Changes in target gene architecture in turn affects accessibility of transcription factors and RNA polymerase II resulting in activation or repression of gene expression.[154,155] Furthermore, epigenetic changes might arise due to hormonal regulation of DNMT enzymes. For instance, female sex steroid hormones have been shown to regulate DNMT3A, DNMT3B, and DNMT1 expression during menstrual cycle.[156,157]

Gonadal steroid hormones affect many brain sex-specific features during peri-natal development. *In utero* production of testosterone occurs in substantially greater amounts in the testes compared to the female ovaries.[158] Developing neurons convert testosterone to estradiol which mediates sexual brain differentiation, masculinization by activation of estrogen receptor alpha and defeminization by estrogen receptor-beta.[159] Additionally, such changes are brought

about in a tissue-specific manner. Kolodkin and Auger (2011) demonstrated higher expression of Dnmt3a in amygdala of female post-natal day 1 rats compared to males, but not in preoptic area or hypothalamus.[160] Prior injection with gonadal steroid hormones (dihydrotestosterone or estradiol) abolished female specific Dnmt3a expression and no sex differences were observed after post-natal day 10. Amygdala plays a central role in social/emotional integration in humans and abnormal amygdalar development has been documented in multiple neurodevelopmental disorders, including SCZ, autism, childhood BPD, and anxiety disorder.[161-164] Taken together these studies indicate that risk alleles identified in association studies might exert their effect by hormone-induced epigenetic alterations.

1.4.3. Parent-of-origin effects

Parent-of-origin effects refer to sex-dependent transmission of a disorder from the parent to the offspring. Goldstein *et al.* (2011) found significantly higher incidence of SCZ in the sons (19%) of affected mothers compared to the daughters (9.5%). In contrast, in the case of affected fathers, incidence of SCZ was higher in daughters (15.2%) compared to the sons (3.1%).[165] Many other studies have reported maternal-specific transmission of BPD in familial cases.[166-170] The traditional genetic explanation is that sex chromosomes account for the observed parent-of-origin effects. However, linkage studies of BPD have failed to detect risk loci on the X chromosome while, only three risk loci (relative risk < 1.08) have been conclusively identified in SCZ.[20,171,172] The epigenetic phenomenon of genomic imprinting can offer insight into this non-Mendelian feature of SCZ and BPD.

Genomic imprinting refers to the parent gender specific epigenetic silencing and monoallelic expression of either paternal or maternal allele. For instance, only the paternal allele of the insulin-like growth factor 2 (IGF2) gene, is expressed while the maternal allele of the H19 gene (located in the same imprinting cluster as *IGF2*) is expressed. Imprinted genes are often present in clusters which involve a few regulatory elements. Differential epigenetic regulation of these elements in paternal and maternal alleles dictates gene expression in a tissue- and developmental stage-specific manner.[173,174] The epigenetic modifications at imprinted clusters are faithfully maintained throughout an organism's development and are only erased and restated during germ cell development.[125,175,176]

Paternal and maternal genes may dominate over each other in different brain regions. In mice, brain regions showing paternal influence are abundant in the hypothalamus and septum areas, which mediate instinctual behavior, such as feeding, mating, and aggression.[177,178]

Maternal influence is observed in areas related to cognition, which may be significant given that cognition is impaired in BPD and SCZ patients.[179,180] Furthermore, imprinting syndromes often exhibit psychiatric comorbidities. An imprinting region, located at chromosome 15q11-13, is involved in the etiology of Prader–Willi and Angelman syndrome (PWS and AS). AS, or "happy puppet" syndrome, is marked by epigenetic downregulation of maternal genes. Affected individuals exhibit hyperactivity and attention-seeking behavior in infancy, with high incidence of autism. In contrast, PWS is attributed to downregulation of paternal genes in the same chromosomal region, which features extremely placid, undemanding behavior in infancy. PWS affected individuals often display a high incidence of psychosis with depression.[181]

1.4.4. Effect of environmental factors and epigenetic heritability

A growing body of literature suggests that epigenetic changes can be mediated by multitude of environmental factors either at specific loci or genome-wide (reviewed in [182]). Adverse environmental effects in early post-natal life, including neglect or abuse, have been associated with SCZ and BPD.[183] Several animal studies have shown strong relationship between variability in maternal care and alterations in epigenetic profiles of genes involved in the hypothalamic-pituitary-adrenal (HPA) axis function and hippocampus related learning and memory processes. Weaver *et al.* (2004) showed how differences in rat post-natal maternal care resulted in an increase in the glucocorticoid receptor gene (*NR3C1*) methylation in the offspring, which was associated with reduced *NR3C1* expression and behavioral changes that persisted into adulthood.[119] Subsequent animal studies showed that maternal care induced changes in DNA methylation and gene expression can be reversed by methyl donor supplementation in early post-natal life.[184] Glucocorticoid receptors play a central role in the regulation of the HPA axis and *NR3C1* expression has been shown to be down-regulated in hippocampal and cortical brain regions of SCZ patients.[185,186] Furthermore, increase in *NR3C1* promoter methylation is associated with childhood maltreatment severity in healthy subjects and childhood abuse in suicide victims.[187,188] Therefore both animal and human studies indicate that adverse events in early life might result in persistent epigenetic marks associated with psychotic disorders.

Independent retrospective studies have shown high prevalence of SCZ in the offspring of mothers exposed to famine during pregnancy, suggesting an association between pre-natal stress and SCZ.[189-192] Moreover, aberrant epigenetic modification at the imprinted *IGF2* locus was also observed in these offspring.[190,193] Experiments in animal models have shown how epigenetic regulation can mediate gene-environment interaction within the viable yellow agouti (A^{vy}) inbred

mouse strain.[194,195] The wild-type agouti gene encodes for mice coat color and insertion of a transposable element, containing a cryptic promoter, upstream of the transcription start site results in methylation dependent expression of the gene. Supplementation of maternal diet with methyl donors, including folic acid, methionine, and Vitamin B12, resulted in an increase in transposable element methylation and brown coat color phenotype in the offspring.[195,196] While the A^{vy} loci provides evidence for maternal epigenetic inheritance, paternal and maternal epigenetic inheritance has been observed at the murine $Axin^{FU}$ loci,[197] responsible for the kinked-tail phenotype. Such metastable epialleles have not yet been identified in humans although several epigenetic studies in MZ twins provide indirect evidence for epigenetic heritability[135,198-200] (discussed in section 1.4.1). Transgenerational epigenetic traits as mediators of environmental risk factors could, in part, explain missing heritability observed in GWAS, however to account for phenotypic heritability such traits would have to be stable during epigenetic reprogramming across several generations. Transgenerational studies conducted in plants show stability of methylation profiles across eight generations, such data is yet unavailable for humans.[201]

Lastly, use of recreational drugs, such as methamphetamine, are also known to influence *Dnmt1* expression resulting in altered DNA modification profiles and its prolonged use led to behavioral changes.[202] Furthermore, mood stabilizers valproic acid and lithium, first line treatment for BPD patients, have also been shown to modulate changes in DNA and histone modifications.[203-206]

1.5. Tools and techniques for epigenetic studies

Over the past decade numerous approaches have been introduced for investigation of DNA and histone modification patterns. While multiple techniques were designed to interrogate small genetic regions they have now been scaled for genome-wide epigenetic analysis with microarray or next generation sequencing (NGS) platforms. Histone modifications are primarily investigated through enrichment based techniques that utilize antibodies specific to a known post-translational histone modification. Histone bound gDNA pulldown by chromatin immunoprecipitation (ChIP) is subsequently interrogated on either microarrays (ChIP-chip) or by sequencing (ChIP-seq).[207]

Molecular techniques investigating DNA modification can be divided into three main categories: 1) affinity enrichment of modified cytosines, 2) methylation-sensitive restriction endonuclease digestion of gDNA and 3) bisulfite conversion.[208,209] Choice of appropriate method depends upon the amount of genomic coverage and resolution required, which in turn governs

the experimental cost. Affinity enrichment of 5-mC is either achieved by DNA immunoprecipitation with 5-mC specific antibody or by use of MBD protein.[210,211] The enriched fragments, which contain at least one modified cytosine, are then either interrogated on microarrays or by deep sequencing. Since within the enriched fragment the exact location or number of modified cytosines cannot be determined, resolution of affinity based assays depends greatly upon CpG density and fragment size.[212] For restriction endonuclease based assays, resolution is limited by the number of target cut sites in the genome, however this can be improved by digestion of gDNA with multiple restriction enzymes.[213] Interrogation of enriched DNA fragments on a microarray based platform is highly cost effective although only provides qualitative differences between epigenetic profiles.

Absolute modification density can be determined at single base-pair resolution by bisulfite (BS) treatment of gDNA followed by sequencing. BS converts unmodified cytosines to uracils, which after PCR are coded as thymines while 5-mC remains unchanged (Figure 1.2). Traditionally, bisulfite sequencing was limited to single locus analysis with the region of interest amplified by PCR primers binding to the BS converted gDNA followed by sequencing.[214] With the adaptation of whole genome sequencing, bisulfite treated gDNA can be used for genome-wide fine mapping of 5-mC.[215,216] However some caveats of whole genome bisulfite sequencing (WGBS) still remain, such as the inability to distinguish between 5-mC and 5-hmC which, unlike 5-caC and 5-fC, are resistant to bisulfite conversion (Figure 1.2).[217,218]

Furthermore, due to the reduced complexity of the genome upon bisulfite treatment, minimum 30X sequencing depth must be achieved for reliable quantification of modification density at a single nucleotide resolution. This barrier significantly increases the cost of WGBS and limits its utility in populational studies. Theoretically, 1000X sequence coverage will be required to reliably estimate rare cytosine modifications, 5-caC and 5-fC.[219] It has also been argued that 70 to 80% of WGBS reads do not offer any essential information due to either lack of CpG dinucleotides within the sequence reads or due to presence of uninformative CpG sites with constitutive modification profiles across several cell and tissue types.[220] This is not necessarily true for epigenetic studies of the brain where abundant CpH modification has recently been found.[99-102] The cost burden of WGBS can be overcome by combining capture-based methods with bisulfite deep sequencing, such as Illumina's Infinium BeadChIP technology and bisulfite padlock probe-based sequencing (BSPP-seq).[221-223] Several hundred thousand sequence- and strand-specific padlock probes can be multiplexed to target bisulfite converted gDNA. Once the probes bind to DNA they are amplified and subsequently sequenced

using NGS platforms for reliable estimation of CpG and CpH modification density.

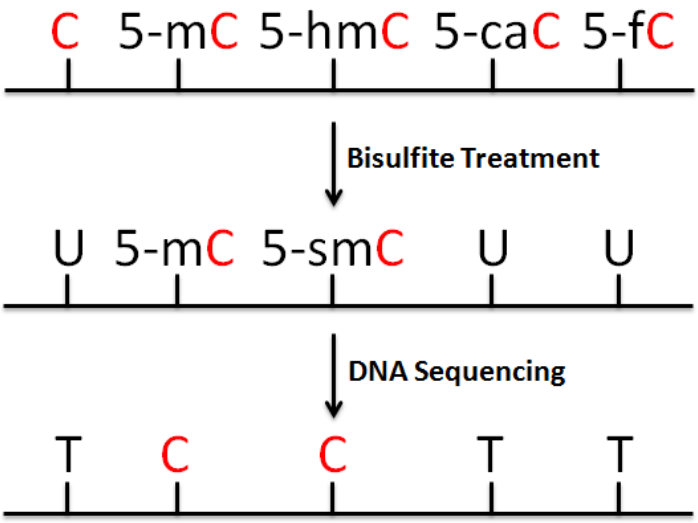

Figure 1.2. Bisulfite conversion based mapping of modified and unmodified cytosines. Bisulfite treatment of gDNA readily forms a 5,6-dihydro-6-sulfonyl adduct at unmodified cytosines (C) which converts to uracil (U) after hydrolytic deamination and is read as thymine (T) during DNA sequencing. 5-methylcytosine (5-mC) is resistant to bisulfite promoted deamination and is therefore read as C. Bisulfite treatment at 5-hydroxymethylcytosine (5-hmC) results in hydrolytically stable 5-sulfonylmethylcytosine (5-smC) which is also read as C and is indistinguishable from 5-mC during DNA sequencing. In contrast, the oxidation products of 5-hmC, 5-formylcytosine (5-fC) and 5-carboxylcytosine (5-caC) are read as T during sequencing due to bisulfite induced conversion to C and subsequently U.

1.6. Epigenetic studies in major psychosis

Early epigenetic studies of SCZ and BPD investigated DNA modification in candidate genes such as catechol-O-methyltransferase, sex-determining region Y (SRY)-box 10, reelin, serotonin receptor 1A, and forkhead box P2.[224-228] Although significant differences were found between cases and controls, the findings were not always replicated.[229,230] The first epigenome-wide study in major psychosis was conducted by Mill *et al.* (2008) investigating 12,000 GC-rich regions, including CGIs, in *post-mortem* prefrontal cortex and germline samples from major psychosis and controls.[231] Aberrant DNA modification patterns were detected for genes involved in GABAergic and glutamatergic neurotransmission pathways, which have previously been implicated in major psychosis.[232,233] The study also identified sex specific differences between affected and unaffected samples with BPD males showing significant differences in genes involved in stress response while genes implicated in brain development were identified in BPD and SCZ females. Interestingly, using partial-correlation network analysis the authors showed reduced epigenetic modularity in brain and germline BPD samples suggesting a systemic epigenetic dysfunction in affected individuals. Another study conducted by Dempster *et al.*

(2011) investigated over 27,000 CpG regions in MZ twin pairs discordant for SCZ or BPD.[137] Disease associated DNA modification differences were identified in several loci of which genes involved in pathways directly relevant to neurodevelopment and psychiatric disorder were statistically overrepresented. Additionally aberrant DNA modification may result from changes in expression levels of DNA methyltransferases or availability of methyl group donors, both of which have been reported to be altered in major psychosis.[234-236]

Differences in histone modification profiles have also been observed in SCZ and BPD. Huang *et al.* (2007) found reduced levels of histone 3 lysine 4 trimethylation (H3K4me3) at the glutamate decarboylase 1 (GAD1) locus in the prefrontal cortex of female SCZ samples.[237] H3K4me3 is associated with active transcription as it promotes an open chromatin state. The authors subsequently observed low expression of GAD1 in SCZ females which is consistent with previous findings in affected individuals.[238] Gavin *et al.* (2009) reported increased levels of the repressive chromatin mark histone 3 lysine 9 demethylation (H3K9me2) in WBCs from living SCZ patients and significant correlation of early disease onset with higher levels of H3K9me2.[239] Additionally over-expression of histone deacetylase 1 enzymes, which promote open chromatin by removing the acetyl group from the terminal tail of histone 3 and 4, has been reported in SCZ brain studies.[240,241] Together, these studies suggest that perturbations in histone regulatory enzymes and modification patterns may alter gene expression in affected brain and contribute to the risk of major psychosis.

1.7. Goals of the study

The epigenome-wide microarray-based scan conducted by our group (detailed in section 1.6) identified several dozen DNA modification differences in *post-mortem* brain samples from individuals affected with major psychosis,[231] and one of the promising targets was the human leukocyte antigen (HLA) complex group 9 (non-protein coding) gene (*HCG9*). Our first objective was to validate these findings using bisulfite pyrosequencing based mapping of a 700 bp 5' region of *HCG9* overlapping the microarray probe in the *post-mortem* prefrontal cortex cohort used for the microarray scan. We also attempted to replicate our primary findings in an independent *post-mortem* prefrontal cortex sample. In the event of validation, we were also interested to determine if pathological *HCG9* modification differences can be detected in peripheral white blood cells (WBC) and germline, the tissues that are not directly involved in psychiatric disease but that are easily accessible and can be used in clinical applications.

Our second objective was to determine the utility of BSPP-seq which has several

advantages over traditional pyrosequencing based assays such as strand- and allele-specific modification profiling in both CpG and CpH context. Several padlock probes can be multiplexed to reliably interrogate large number of target loci in a short time with limited funds. We applied this approach to map cytosine modification density at *HCG9* (plus 1Kb upstream and downstream of the gene) in a cohort of *post-mortem* prefrontal cortex samples from major psychosis and unaffected controls.

Traditional bisulfite based mapping of DNA modification is unable to differentiate between 5-mC and 5-hmC.[217] Our third objective was to characterize the genome-wide distribution of 5-mC and 5-hmC in a variety of neuronal and non-neuronal tissues from mice and humans and investigate their respective roles. 5-hmC was recently rediscovered in the mammalian genome and is abundant in the brain.[76-78] Multiple studies have shown that 5-hmC is an intermediate step in DNA demethylation catalysed by TET enzymes (Figure 1.2).[79,80] However, it is not known if DNA demethylation completely accounts for the enrichment of 5-hmC in the brain and points towards a functional role of 5-hmC in the genome. We assayed 5-hmC using a restriction enzyme-based method and subsequent interrogation on tiling microarrays.

Lastly, our fourth objective was to identify differences in 5-mC and 5-hmC modification density at *HCG9* in *post-mortem* prefrontal cortex and WBC samples from major psychosis and control samples.

Chapter 2. Materials and Methods

2.1. Samples

A summary of the demographic information of samples used in this study is shown in Table 2.1 and Table 2.2. For bisulfite pyrosequencing based *HCG9* analysis, *post-mortem* brain tissue from prefrontal, parietal, and occipital cortices and corpus callosum from individuals with BPD (N = 34), SCZ (N = 35), and matched controls (N = 35) was provided by the Stanley Medical Research Institute (USA) brain-array collection. Detailed information of the SMRI cohort was described elsewhere.[242] Brain prefrontal cortex samples were also provided by the Harvard Brain Tissue Resource Center, McLean Hospital (USA) (N = 34 BPD, 30 SCZ, 50 control). Peripheral white blood cell (WBC 1) samples were provided by Dr. James Kennedy (N = 100 BPD, 50 SCZ, 100 control), a second WBC sample set (WBC 2) from GlaxoSmithKline (USA) (N = 270 BPD, 282 control), and sperm samples collected by our laboratory (N = 29 BPD, 14 SCZ, 30 control). Diagnoses had been made according to the Diagnostic and Statistical Manual of Mental Disorders, Fourth Edition (DSM-IV). For BSPP-seq of *HCG9* we used *post-mortem* prefrontal cortex tissue samples provided by the Harvard Brain Tissue Resource Center, McLean Hospital (USA) (N = 27 BPD, 22 SCZ, 32 control; Table 2.1). For the sample set used in BSPP-seq analysis of *HCG9* we were able to retrieve information regarding post-mortem neuropathological findings which was included in our data analysis.

For microarray based genome-wide mapping of 5-mC and 5-hmC, we used a cohort of *post-mortem* prefrontal cortex tissue samples from the Stanley Medical Research Institute (USA) and the Harvard Brain Tissue Resource Center, McLean Hospital (USA) (N = 28 BPD, 30 SCZ, 28 control; Table 2.1). Human liver samples (N = 13 control) were obtained from commercial tissue banks (Curline (USA) and Cambridge Bioscience (UK)). Liver samples were age- and sex-matched to brain samples. Male C57BL/6J mouse brain tissues (frontal cortex and remaining brain) and other organs (liver, pancreas, kidney, and heart) were obtained from 8 week to 18 month old mice (Table 2.2). Transformed human B-lymphocyte cell line (GM10851, Coriell Cell Repositories) and mouse neuronal cell line (mHypoA-2/24) were also included in this study.

For 5-mC and 5-hmC mapping of *HCG9*, CpG 6 we used a cohort of *post-mortem* prefrontal cortex tissue samples from the Stanley Medical Research Institute (USA) and the Harvard Brain Tissue Resource Center, McLean Hospital (USA) (N = 45 BPD, 25 SCZ, 53 control; Table 2.1) and peripheral white blood cell samples provided by Dr. James Kennedy (N = 30 BPD and 30 control, Table 2.1). *Post-mortem* brain tissue from two control individuals (Sex:

Male, Age: 44.5±3.5 years) was provided by Dr. Stephen Kish for brain regions BA38, BA7b, BA8, BA17, BA23, cerebellar cortex, caudate, corpus callosum (rostral), globus pallidus (internal), internal capsule (caudal and rostral), putamen, corpus callosum (caudal), insular, lateral thalamic nucleus, amygdala, medial pulvinar thalamic nucleus and hippocampal Ammon's horn.

		N	Age (years)	Gender (M:F)	PMI (h)
***Post-mortem* brains**					
SMRI	BPD	34	45.4 ± 10.7	16:18	37.9 ± 18.6
	SCZ	35	42.6 ± 8.47	26:9	31.4 ± 15.5
	Control	35	44.2 ± 7.58	26:9	29.4 ± 12.9
McL	BPD	34	61.4 ± 18.9	14:20	22.2 ± 7.19
	SCZ	30	59.8 ± 13.3	20:10	22.8 ± 6.30
	Control	50	58.4 ± 15.6	30:20	20.9 ± 4.82
Peripheral white blood cells					
Toronto	BPD	100	42.1 ± 10	34:66	-
	SCZ	50	43.3 ± 11.1	25:25	-
	Control	100	41.7 ± 10.9	34:66	-
GSK	BPD	270	44.5 ± 12.6	121:149	-
	Control	282	27.8 ± 14	136:146	-
Germline (Sperm)					
	BPD	29	40.2 ± 11.6	-	-
	SCZ	14	37.4 ± 11.2	-	-
	Control	30	37.7 ± 10.3	-	-
Liver					
	Control	13	54.3 ± 9.8	6:7	5 ± 2.2

Table 2.1. Sample information for the human tissue dataset.
Post-mortem brain tissue were provided from the Stanley Medical Research Institute (SMRI) and the Harvard Brain Tissue Resource Center, McLean Hospital (McL). Peripheral white blood cell samples were provided by Dr. James Kennedy (Toronto) and GlaxoSmithKline (GSK). Sperm samples were collected by our lab and liver samples were obtained from commercial tissue banks.

Animal num.	Age	Brain	Heart	Liver	Pancreas	Kidney
1	8wk	71B	72H	73L	74P	75K
2	24mos	136B	137H	138L	139P	140K
3	18 mos	121B	-	-	-	-
4	8wk	91B	1H	61L	121P	31K
5	8wk	92B	2H	62L	122P	32K
6	8wk	95B	-	-	-	-
7	8wk	93B	3H	63L	123P	33K
8	8wk	51B	52H	53L	54P	55P

Table 2.2. Sample information for the mouse tissue dataset.
All animals were adult male inbred C57/BL6 strain mice; in several instances, multiple tissue samples were collected from the same animal. Not shown in table: An independent set of mouse brains was separated into frontal cortex and the remainder (including brain stem and cerebellum; 8-week old mice; N = 15).

2.2. DNA extraction

Tissue samples were lysed in 750 µl lysis buffer (35mM EDTA, 10mM Tris-Cl (pH 8) and 1% SDS) with 70 µl of proteinase K (20mg/ml) and incubated at 50°C overnight in a thermo shaker. For sperm samples, 100 µl of 1 M β-mercaptoethanol was added to the lysis buffer. For whole blood samples DNA was purified from WBCs after removal of red blood cells (RBCs). 500 µl of whole blood was mixed with 1 ml of RBC lysis buffer (0.01 M Tris-HCl (pH 7.6), 320 mM sucrose, 5 mM MgC12, 1% Triton X 100) followed by centrifugation at 7000 rpm for 2 min. The supernatant was discarded and the procedure was repeated twice to remove all hemoglobin from the pellet (WBCs).

Equal volume of phenol:chloroform:isoamylalcohol (25:24:1 v/v) was added to each lysed tissue sample and placed on a rotator at room temperature for 5 min for complete mixing. Samples were then centrifuged at 14000 rpm for 5 min and the supernatant was collected in a separate tube. This step was repeated three times, once with phenol:chloroform:isoamylalcohol and twice with chloroform:isoamylalcohol (24:1 v/v). DNA was then precipitated with isopropanol followed by two washes with 70% ethanol. DNA pellets were then air dried and dissolved in 10mM Tris-Cl (pH 8.5) and checked for quality and quantity using spectrophotometry and agarose gel analysis.

2.3. Whole genome amplification

40 ng of gDNA was mixed with 2 µL of 10x phi29 reaction buffer and 2 µL of exo-resistant random primer (Thermo Scientific, USA) in a 17 µL reaction. The mixture was incubated at 95°C for 5 min and then gradually cooled to 30°C (1°C/15 sec). 0.5 µL dNTP mix

(10 mM), 0.5 µL BSA (10 mg/ml), 1 µL phi29, and 1 µL inorganic pyrophosphatase (Thermo Scientific, USA) were added to the reaction. The mixture was then incubated for 6 hrs at 30°C. High molecular weight amplification products were observed on an agarose gel, and the amplicons were purified using standard ethanol precipitation.

2.4. Bisulfite treatment

Bisulfite treatment was performed by use of a standard protocol.[243] In brief, ~500ng of gDNA was denatured in 0.3 M NaOH for 20 min at 42°C. After the addition of freshly prepared saturated sodium metabisulfite (Sigma) and 10mM hydroquinone (Sigma) solution, samples were subjected to a 5 hr incubation at 55°C under exclusion of light. The samples were then purified and desulfonated with 0.3 M NaOH by using a Montage PCR96 96-well filtration plate (Millipore).

2.5. Pyrosequencing

Four primer sets were designed to amplify the exon 1 and adjacent intron 1 region of the *HCG9* gene (Table 2.3). Within these regions, six pyrosequencing assays were designed. In total, 28 CpGs in *HCG9* were covered. PCR amplifications were performed with a standard PCR protocol in 50 µl volume reactions containing 2 µl of sodium-bisulfite-treated DNA, 0.15-0.5 µM primers, and master mix containing JumpStart or Platinum Taq DNA polymerase (Sigma Aldrich and Invitrogen, respectively). After agarose gel electrophoresis to ensure successful amplification and specificity, PCR amplicons were processed for pyrosequencing analysis according to the manufacturer's standard protocol (Qiagen). To validate the assay method, various proportions (0, 25, 50, 75, and 100%) of mixtures of CpGenome Universal Methylated and Unmethylated DNA (Millipore) were processed with the same protocol. Before proceeding with the clinical samples, it was confirmed that there are linear correlations between stepwise increments of methylated DNA and the measured methylation by pyrosequencing analysis in all CpG sites.

2.6. Genotyping

Genotyping of SNPs rs17180353, rs2071568, rs373472, rs58031868, rs6903753, rs690402, rs9260832, rs9278524, rs400488, rs422640, rs9278523 and rs1128306 was performed by pyrosequencing. SNP rs1128306 genotypes were confirmed by resequencing in selected samples including the SMRI brain, McL brain, and WBC 2 samples. PCR was performed by

using 10 ng of gDNA, 0.5 µM primers listed in Table 2.3, and Taq DNA polymerase (New England BioLabs). PCR products were purified using the QIAquick PCR purification kit (Qiagen) and processed for sequencing by the Center for Applied Genomics. Sequence alignment was performed by using ClustalW2 (http://www.ebi.ac.uk/Tools/clustalw2/index.html).

Primer name	Sequence
A. Pyrosequencing	
HCG9-5'-F	GGGGGTTATGAGAAAGGAAGT
HCG9-5'-R	Biotin-CTAAAACCCTATCCTCTCCCTA
HCG9-5'-PYRO	GGAAAGAATTTTGGGAA
HCG9-F	Biotin-GGATTTTAGGGAGAGGATAGGG
HCG9-R	CCCCACCCCCTACACTTT
HCG9-PYRO1	CTAAACTATTCCTATAAATAACATT
HCG9-PYRO2	CCTCACCTCTCCTCC
HCG9-3'A-F	GAGAGTAAGTGTAAGAAGAGATT
HCG9-3'A-R	Biotin-CTTATAATCCCAACTACTCAAAAAA
HCG9-3'A-PYRO1	GGGGTGGGGGTGATG
HCG9-3'A-PYRO2	AAGTAGGGTTGAGGAGTG
HCG9-3'B-F	Biotin-AGGATTGAAAAGAGATTGAAAAGT
HCG9-3'B-R	TCATTATCTAATAAAAATTAATACCAAATA
HCG9-3'B-PYRO1	ATAACAAAATTATTCTAACC
B. Resequencing	
HCG9-RSEQ-F	GAGCAGTCGCAGGAAGAATC
HCG9-RSEQ-R	CAGCCCTGCTTTTCAGTCTC
HCG9-RSEQ-S	CAGTCGCAGGAAGAATCCT

Table 2.3. Pyrosequencing primers used in *HCG9* modification analysis.

2.7. Enrichment of padlock probe targets

Figure 2.1 illustrates the bisulfite padlock probe-based sequencing (BSPP-seq) methodology. gDNA from each sample was treated with bisulfite which converts 5-caC and 5-fC modified and unmodified cytosines to uracil while 5-mC and 5-hmC modified cytosines are left unconverted. Strand- and sequence-specific padlock probes anneal to bisulfite converted gDNA and were extended to form circularized DNA which was subsequently amplified using primers containing the sequencing adapters and barcodes (sample identifiers). The resulting libraries from all samples were pooled and sequenced using a next-generation sequencing platform.

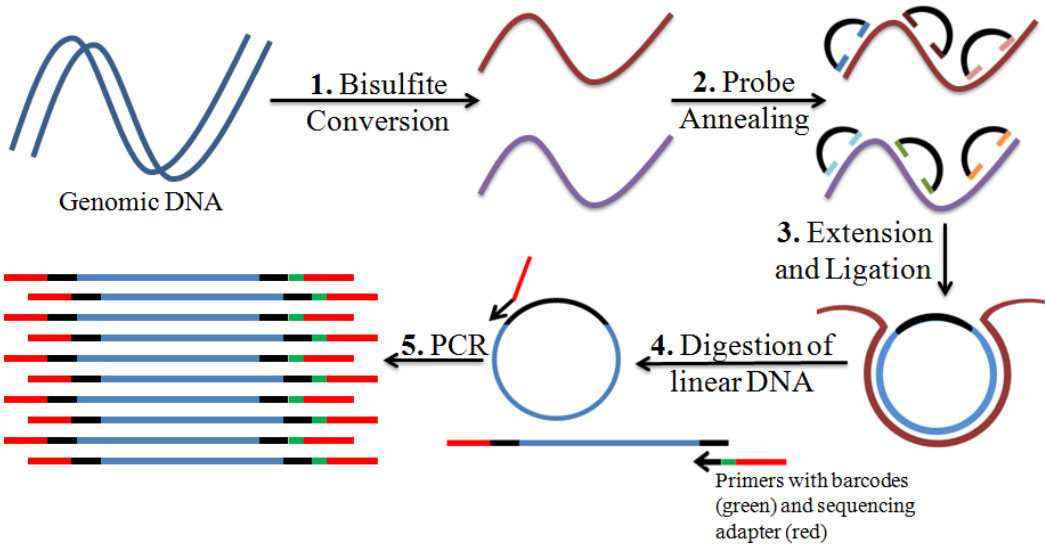

Figure 2.1. Graphical representation of bisulfite padlock probe-based sequencing (BSPP-seq).
1) gDNA is treated with sodium bisulfite which converts 5-caC and 5-fC modified and unmodified cytosines to uracil while 5-mC and 5-hmC modified cytosines remain unchanged. Bisulfite conversion of gDNA also results in loss of complementarity of the two DNA strands. 2) Padlock probes containing two locus and strand specific annealing arms, connected by a common linker sequence, bind to converted DNA. 3) The 3' probe arm is extended and ligated at the 5'end to form circularized DNA. 4) Remaining linear DNA is digested with exonuclease. 5) The captured circularized DNA is amplified using PCR primers containing barcodes (identifiers) and sequenced.

2.7.1. Padlock probe design

Padlock probe design was carried out using ppDesigner.[223] The algorithm was run twice, with both sense and antisense strand of the full length *HCG9*, plus 1kb upstream and downstream of the gene. The padlock probes were then manually interrogated for presence of potentially confounding CpG dinucleotides within the last 5 nucleotides of the annealing arms. In cases when adjustments were not possible, CpG dinucleotide was replaced with a CpR, where R is a mixture of G and A, corresponding to a complementary base pair of a methylated / hydroxymethylated cytosine or an unmodified cytosine (converted to uracil upon bisulfite conversion and then to thymine after amplification), respectively. Similar adjustments were made for probes overlapping SNPs in the annealing arms. Probes found binding to repetitive elements were removed and a total of 57 probes were then synthesized with 5' phosphorylation modification (IDT, USA).

2.7.2. Padlock probe DNA capture and sequencing

400 ng of sodium bisulfite treated gDNA was mixed with 6 µL of the padlock probe mixture (Appendix 1) containing 0.5 pM of each probe and 1x Ampligase buffer to a final

volume of 25 µL. DNA was denatured at 95°C for 10 min and probe hybridization was performed by gradual drop of temperature at 1°C/min to 55°C, followed by incubation at 55°C, 50°C, and 45°C for 5 hr each with 1°C/min gradual drops in between. For circularization of annealed probes, 2.5 µL of HLN buffer (20 µL of 10 mM dNTP, 100 µL of 10x Ampligase buffer, 500 U Ampligase, 200 µL of Hemo KlenTaq (NEB, USA) in 1 mL total volume) was added, and the reaction was incubated at 45°C for 5 hrs, followed by a gradual increase (1°C/min) to 50°C, 5 hrs at 50°C, a gradual increase to 55°C, and 5 hrs at 55°C. This was followed by enzyme inactivation at 94°C for 2 min. To digest any remaining linear DNA after circularization, 3 µL of exo mix (200 U exonuclease III and 10 U exonuclease I (NEB, USA) was added followed by incubation at 37°C for 1 hr and enzyme inactivation at 94°C for 2 min. This was repeated three times.

2 µL of circularized DNA was mixed with 1 µL of 10 µM pp_Famp primer (5' AATGATACGGCGACCACCGAGATCTACACCACTCTCAGATGTTATCGAGGTCCGAC '3), 1 µL of 10 µM barcoded pp_Ramp primer (Appendix 2), and 25 µL of 2X Phusion High-Fidelity PCR Master Mix (GC Buffer) in a 50 µL reaction. Thermo-cycling conditions were as follows: 98°C for 30 sec; 35 cycles, each 98°C for 10 sec, 58°C for 30 sec, 72°C for 10 sec; 72°C for 5 min. The desired product size of ~300bp was confirmed on a 2% agarose gel. All PCR amplicons were pooled and purified with Ampure XP beads (Beckman Coulter, USA) using manufacturer's protocol. The PCR library was sequenced using MiSeq platform (Illumina, USA). Following sequencing primers were used, SolSeq6.3.3 (Read1) 5' TACACCACTCTCAGATGTTATCGAGGTCCGAC '3, SolSeqV6.3.2r (Read2) 5 ' GCTAGGAACGATGAGCCTCCAAC '3 and AmpR6.3IndSeq (IndexRead) 5' GTTGGAGGCTCATCGTTCCTAGC '3. We used 60% PhiX spike-in to compensate for low base-pair diversity inherent to bisulfite conversion.

2.8. Enrichment of 5-mC and 5-hmC modified cytosines in genomic DNA

Figure 2.2 illustrates our enrichment method for genomic 5-mC and 5-hmC. T4-phage enzyme β-glucosyltransferase (BGT) transfers a glucose molecule specifically to the hydroxymethyl group of 5-hmC, thus rendering it resistant to digestion by the methylation insensitive *Msp*I enzyme at the $C^{hm}CGG$ target site.[244,245] 5-hmC is thus estimated by differential resistance to *Msp*I-digestion with and without glucosylation of gDNA. *Hpa*II (targets the same site, CCGG) cannot cut $C^{m}CGG$ or $C^{hm}CGG$, and conceptually its difference with *Msp*I digestion

is a measure of both 5-mC and 5-hmC. Subtraction of the 5-hmC estimate from the *Hpa*II-based estimate therefore measures 5-mC.

2.8.1. Glucosylation and endonuclease restriction digestion

From each tissue, two 1 µg samples of gDNA were sheared to a 200 bp fragment length using a Covaris S2 sonifier (Covaris, USA). Sheared DNA was end-filled in the presence of T4 DNA polymerase (5 U) at 11ºC for 20 min and purified using PN buffer from the QIAquick Nucleotide Removal Kit (Qiagen). The purified 1 µg of blunt-ended DNA was ligated with A1-25 adapters (10 µM) (A1 = 5'AGTTACATCTTGTAGTCAGTCTCCA 3', A25 = 5' TGGAGACTGACTACAAGAT 3') in the presence of T4 DNA ligase (10 U) at 22ºC for 3 h, inactivated at 65ºC for 10 min, and purified with the QIAquick Nucleotide Removal Kit (Qiagen, USA). One third (~300ng) of the purified DNA was treated with 200 µM uridine-5'-diphospho-α -D-glucose (UDP-Glc, Sigma, USA), 80 ng BGT[246] in 100 mM Tris-HCl (pH 8.0) and 25 mM $MgCl_2$ for 3 h at 37ºC, and purified again. Thereafter, 200 ng of glucosylated DNA and 200 ng of untreated DNA were subjected to *Msp*I digestion in a final volume of 20 µl for 12 hrs at 37ºC, followed by enzyme inactivation at 80ºC for 10 min. The same protocol was performed on 200 ng of untreated DNA using *Hpa*II. Following the observation that some 5-hmC may be resistant to *Msp*I digestion,[247] we used 10 units of enzyme (50 fold surplus of the enzyme over the DNA substrate digested for 12 hrs).

2.8.2. PCR amplification and microarray hybridization

Restriction enzyme-digested DNA fragments were amplified with an adaptor primer (5' AGTTACATCTTGTAGTCAGTCTCCA '3), and dUTP was included in the dNTP mix as specified by Affymetrix. Two rounds of PCR amplifications were performed to achieve the required DNA amount for tiling array hybridization. PCR cycling for the 1st round of amplification was performed on the restriction enzyme-digested gDNA sample. The second round of amplification was done on 1/10th of the 1st PCR template; both rounds of amplification used the same PCR cycling conditions (i.e. 95ºC for 1 min, followed by 15 cycles of 94ºC for 15 sec, 65ºC for 30 sec and 1 min at 72ºC, with an extension of 5 sec in each subsequent cycle). The amplicons were then purified using QIAquick 96 PCR Purification Kit (Qiagen, USA) and checked for quality and quantity on a NanoDrop 2000 spectrophotometer (Thermo Fisher Scientific, USA). Nine micrograms of PCR amplicons were fragmented to 50–100 bp using uracil DNA glucosylase enzyme, which cleaves DNA at incorporated dUTP (GeneChip® WT

Double-Stranded DNA Terminal Labeling Kit, Affymetrix, USA). Fragments were end-labeled according to the manufacturer's instructions. Prior to labelling, 1 µL of fragmented DNA was analyzed on a Bioanalyzer using the DNA1000 Chip (Agilent Technologies, USA) to check the uniformity of the fragmented products. Individual samples were hybridized on a separate Gene Chip of Human or Mouse Tiling 2.0R array for 16 hrs at 45ºC (Table 2.4)

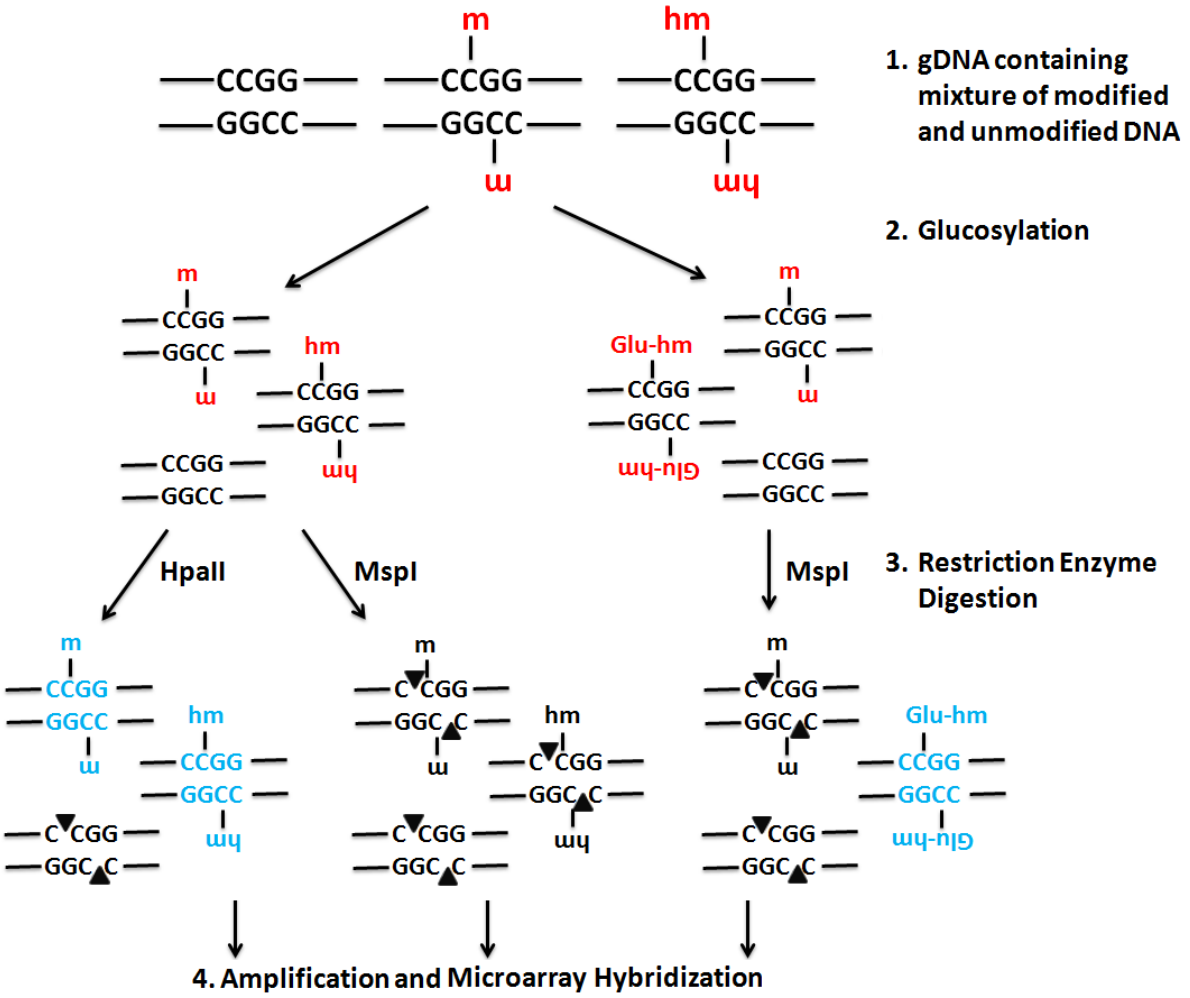

Figure 2.2. Restriction endonuclease based enrichment of genomic 5-mC and 5-hmC.
1) gDNA, containing mixture of 5-mC and 5-hmC modified and unmodified fragments, is sheared and ligated to PCR adapters. 2) The sample is then split in two parts and one half is subjected to glucosylation by T4-phage β-glucosyltransferase (BGT) enzyme. BGT specifically adds a glucose molecule to hydroxymethyl group of 5-hmC rendering it resistant to digestion by methylation insensitive restriction endonuclease, *Msp*I. 3) Glucosylated DNA is subjected to digestion with *Msp*I (target site CCGG) and the other half is subjected to digestion with *Msp*I and methylation sensitive restriction endonuclease, *Hpa*II (same target site, CCGG). 4) Fragments left uncut after treatment with *Hpa*II and *Msp*I (fragments in blue) are subsequently amplified using adapter specific primers and hybridized to tiling microarrays. Differential digestion of gDNA with and without glucosylation measures 5-hmC. Whereas, difference in digestion of gDNA with *Hpa*II and *Msp*I estimates total modification (5-mC + 5-hmC). Subtraction of the 5-hmC estimate from the *Hpa*II-based estimate therefore measures 5-mC.

Dataset	Num. biological samples	Num. arrays (biological + technical replicates)	Array name, chromosomes	Num. CCGG probes [a]
Human				
Brain, BA10 (Stanley & Maclean, controls)	28	84	E: 5,7,16	69,252
Brain, BA10 (Stanley & Maclean, controls)	28 [b]	84	F: 8,11,12	67,225
Brain, BA10 (Stanley & Maclean, psychosis)	54	162	E: 5,7,16	69,252
Brain, BA10 (Stanley), for MBD	6 [b]	12	E: 5,7,16	69,252
Exp 2, Liver (Curline & Cambridge Biosci.)	13	39	E: 5,7,16	69,252
Exp 2, Brain (Stanley, controls)	12 [b]	36	E: 5,7,16	69,252
B-lymphocyte cell line	24	72	G: 10,13,14,17	76,102
Mouse				
Multiple organs (brain, liver, pancreas, kidney, heart)	32	141	A, G: 1,9,10,13,14,19	130,314 [c]
Neuronal cell line (mHypoA-2/24)	6	54	B: 2, X,Y	46,892
Frontal cortex	15	45	A: 1,9,19	69,052
Brain, non-frontal cortex	15	45	A: 1,9,19	69,052
Total	**187**	**774**		

Table 2.4. Sample, array, and probe count for all datasets analyzed in the current study.
Not included is the sample count for Helicos validation, which used 3 technical replicates of a single human brain. [a] After excluding repeat overlaps; b) Biological samples excluded from running total as already previously counted; c) this probe count is lower than that in the synapse-related analyses because this analysis has an additional filter: probes where the chromosomal and Affymetrix probe sequence did not both contain a target site were excluded.

2.9. Production of a 31-mer DNA duplex containing modified cytosines at a CCGG target site

Equal molar amounts (150µM) of complementary single-stranded oligonucleotides (5’ TGACCCACGCTCGCC '3 and 3’ACTGGGTGCGAGCGGGCCTCTATTTAATACA '5) were annealed in water by heating at 95°C for 5 min, followed by slow cooling (1°C/min) to room temperature. Annealed DNA (5 µM) was supplemented with dGTP, dTTP, dATP and dCTP, dmCTP or dhmCTP (Bioline, USA) (1mM each) and Klenow Fragment (0.15 U / µl, Thermo Fisher Scientific, USA), and incubated in Klenow reaction buffer at 37°C for 40 min to produce duplexes containing cytosine (C), 5-mC or 5-hmC at the target site, respectively. 1 µM of duplex oligonucleotide with 5-hmC, 200 µM UDP-Glc (Sigma, USA) and 0.04 µg BGT were incubated

for 1.5 hrs at 37°C in buffer (15 µl, 100 mM Tris-HCl pH 8.0, 25 mM magnesium chloride). Then, 2 µl of Tango buffer, 1 µl (10 U) of *Msp*I (Thermo Fisher Scientific, USA) and 2 µl of water was added to the reaction, and incubation was continued for 1.5 hrs. Samples were supplemented with 1/6 of 6x loading dye solution and analyzed by 15% polyacrylamide gel electrophoresis.

2.10. Production of a 89-mer DNA duplex containing modified cytosines at a CCGG target site

Complementary forward and reverse strands of an oligonucleotide with a single CCGG site containing 5-mC, 5-hmC or cytosine were ordered from IDT, USA. Forward and reverse strands with 5-caC and 5-fC modification at internal cytosine of the target site, respectively, were ordered from Trilink Biotechnologies, USA. Forwards strand sequence: 5' CAGTGAATTCGGCACCACAGATGCAGTGACCGGAGTCATTGCCAACTCTGCAGGAG AGCAAGGTGCTGTCTATAGGTGGATCCAAGTCA '3. Different combination of modified duplex oligonucleotides were generated by combining equal molar (100µM) amounts of differentially modified strands, heating at 95°C for 5 min and then gently cooling to room temperature (1°C/min). 1µM of duplex oligonucleotide was digested with 10U of *Msp*I in 1x Tango buffer and incubation at 37°C for 12 hrs. The products were analyzed by 12% polyacrylamide gel electrophoresis.

2.11. Production of 200 bp 5-hmC modified DNA fragment

A 200 bp DNA fragment containing one CCGG site for qPCR analysis was generated by PCR from mouse gDNA with primers 5’ GCATCCTGGAGATTGTGGGCAACATC^{hm}CGG ′3 (IBA, Germany) and 5’GCCCATGTCGCTGTG ′3 (Metabion, Germany). Incremental amounts (0, 10, 20, 50 and 100%) of 5-hmC modified fragment was mixed with unmodified fragment of same sequence before enzymatic BGT glucosylation in the presence of UDP-G (section 2.8). PCR products were subsequently subjected to *Msp*I restriction hydrolysis for 16 hrs. Real-time PCR experiments were performed with a Rotor-Gene™6000 real-time PCR system (Corbett Research) using Maxima™SYBR Green qPCR Master Mix (Fermentas); 0.3 mM primers were used in each reaction in a final volume of 25 µl. The amplification program was set as: 95°C for 10 minutes, 40 cycles for 15 s, 60°C for 1 minutes, and a melt curve analysis step at the end to check the specificity of the PCR product. Data were analyzed by Rotor-Gene™6000 real-time PCR software.

2.12. Thin-layer chromatography quantification of total genomic 5-mC, 5-hmC and C at CCGG sites

gDNA (40 ng) was digested with *Msp*I (Thermo Fisher Scientific, USA) endonuclease for 2 hrs at 37°C and dephosphorylated with 0.1 U of FastAp (Thermo Fisher Scientific, USA) for 1 hr at 37°C. Enzymes were inactivated by heating at 75°C for 10 min. 5'-end-labelling of DNA fragments was carried out with 4 U of T4 Polynucleotide Kinase (Thermo Fisher Scientific, USA) in the presence of 3.3 µCi of [γ33P]-ATP (Hartmann Analytic, Germany) at 37°C for 10 min in T4 Polynucleotide Kinase reaction buffer, followed by enzyme inactivation at 90°C for 3 min. Labelled fragments were ethanol-precipitated using sodium acetate pH 7.0 (3 M) as part of a standard protocol. Air-dried pellets were dissolved in 4 µL Lambda Exonuclease buffer and incubated with 2.5 U Lambda Exonuclease at 37°C for 2 hrs (Fermentas). Aliquots of hydrolysate (3 replicates) were spotted on PEI cellulose plates (PEI Cellulose F, 20 x 20 cm, Merck) and chromatographed by eluting with isobutyric acid/water/conc. ammonia (66:17:4, v/v/v). Plates were dried, autoradiographed to phosphorimager screens and analyzed with a FLA-5100 scanner and MultiGauge software (Fujifilm). Ratios of C, 5-mC and 5-hmC were calculated after subtracting corresponding gel density values from control experiments. Note that methylation of repetitive elements was quantified by TLC, while repeat-overlapping probes were excluded from the microarray data analysis; this difference could partially account for the discrepancy between these two methods.

2.13. Quantitative polymerase chain reaction

BGT-treated and -untreated DNA was subjected to *Msp*I digestion, as described previously. In addition, DNA was digested by *Hpa*II and an undigested control was used. Locus-specific quantitative PCR (qPCR) was performed using 10 ng gDNA, 1µM reverse and forward primer pair, and 1x SYBR® Green PCR Master Mix (Applied Biosystems, USA) on the 7500 Real-Time PCR System (Applied Biosystems, USA), according to the manufacturer's recommendations (melting temperature of 60ºC). Primer sequences tested for qPCR are listed in Table 2.5. Each sample was performed twice in duplicate and the corresponding *Ct* values were obtained from the 7500 System SDS Software v1.3.1 (Applied Biosystems, USA). All primer pairs flanked either one or two *Msp*I target sites (CCGG) (target primer pairs). One primer pair did not flank an *Msp*I target site and was used as an internal control (reference primer pair, *ref*). The efficiency (*E*) of each primer pair was calculated from the slope of regression line obtained by plotting *Ct* values against varying DNA concentration. Δ*Ct* , percent DNA uncut and 5-mC,

5-hmC modification density were calculated from the following formula, where *Hpa*II and *Msp*I represent gDNA digested with respective restriction endonucleases and glu*Msp*I represents glucosylated gDNA digested with *Msp*I.

$$\Delta Ct = Ct_{meanUndigested} - Ct_{meanDigested}$$

$$\% \text{ DNA Uncut} = \frac{(E_{target})^{\Delta Ct(target)}}{(E_{ref})^{\Delta Ct(ref)}} \times 100$$

$$\% \text{ 5-mC+5-hmC} = \textit{Hpa}\text{II}_{\% \text{ DNA Uncut}} - \textit{Msp}\text{I}_{\% \text{ DNA Uncut}}$$

$$\% \text{ 5-hmC} = \text{glu}\textit{Msp}\text{I}_{\% \text{ DNA Uncut}} - \textit{Msp}\text{I}_{\% \text{ DNA Uncut}}$$

$$\% \text{ 5-mC} = \textit{Hpa}\text{II}_{\% \text{ DNA Uncut}} - \text{glu}\textit{Msp}\text{I}_{\% \text{ DNA Uncut}}$$

Given locus name	Forward Primer	Reverse Primer	Amplicon Length (bp)	# *Msp*I Target sites
MR1	TGCTGCTCGATG CACAGGT	CATCTTCACCT TCCTGCTGAG	86	2
MR2	CAGTCTCTCCCT GCACACAG	GTGGCATGGTC TGGTTTCC	87	1
MR3	CAGTGCCTTAGG CCTCTCTC	CTTGGTCTGCC ATCTTCTGG	93	1
MR4	GAAAGGTGAGC TCCCTGAAG	AAGCAAACGC TGGCTGAG	93	2
DS1	CCAAATTATAAG ACAGATGCCTAG	GCCAACTTCTG TGAAACTACACT	120	1
DS2	TGAGTAGTCATG ACCCCTTTC	CCAGGGTGTAA CATGAATAGGA	91	1
DS3	CTCTTTGGTTCA ACTGGTCCA	CTCTCAGAATCC CAACCAGGA	83	1
DS4	GCCCTTGACTGC CTCCTTAT	TTCCAGGACCCT AAAAAGCTC	133	1
DS5	CTTGCCTGGTCA GATGACAG	TGCCTCTCCATC TAGCATCC	77	1
DS6	TGGTTCTTTACC CCATTAGTCATA	CAGGGATCTGA TGTGCCATAT	104	1
DS7	ACCCACCTGTGT AAGCCTGT	AGGAACTCAGG AGAGCAGGT	135	1
HCG9, CpG6	TGGAGCTGAGAA CACGCGGA	CGAGGGCCCTT CCAGCTCT	113	1
Ref	AATCCTGCCACC TCAGCCTC	GTAAATGACAG CTGGACGTGG	71	0

Table 2.5. Primers used for quantitative PCR.

2.14. Helicos single molecule sequencing (SMS)

Micrococcal nuclease digestion was used to fragment gDNA to a median size of 500 bp and to reduce 3' hydroxyl end at the DNA fragments, where the latter served as the starting end for SMS. 5 µg of gDNA was treated with 1 U of micrococcal nuclease enzyme (NEB) and the reaction was stopped by adding 10 µl of 0.5 M EDTA (in excess) in a time series. A small aliquot was then checked on 1% agarose gel and samples with median fragment size of 500 bp were column purified with buffer PN (QIAquick Nucleotide Removal Kit columns, Qiagen). Glucosylation and control treatments were performed as described before and 200ng of each glucosylated or non-glucosylated treated DNA was subjected to 10 U of restriction enzyme

digestions respectively at 37ºC for 8 hrs, and inactivated at 80ºC for 20 min.

10 ng of each digested product, quantified by Quant-iTTM PicoGreen dsDNA Reagent Kit (Invitrogen), was then processed for Helicos sequencing. In brief, 10ng of DNA was heat denatured at 95ºC for 5 min prior to 3' end labeling with 5 U of terminal transferase (NEB) in presence of 200 µmoles of dATP (Roche) and 5 mmoles of $CoCl_2$ (NEB) in 20 µl reaction volume at 37ºC for 1 hr, and then inactivated at 70ºC for 10 min. Fragments were biotinylated by repeating the terminal transferase enzymatic reaction step in the presence of 100 µmoles of biotin labeled ddATP (Perkin Elmer) instead of dATP in a reaction volume of 30 µl. These processed samples were then sent to the Helicos sequencing service facility (www.helicosbio.com; USA).

2.15. Treatment of suberoylanilide hydroxamic acid (SAHA) on human B-lymphocyte cells

Transformed human B-lymphocyte cells (GM10851, Coriell Cell Repositories) were treated with the histone deacteylase inhibitor SAHA. Prior to the experiment, a cell viability assay (ATP luminescence assay; Cell Titer-Glo; Promega) for SAHA was conducted by titrating different SAHA concentrations. The maximum concentration of SAHA that induced minimal cytotoxicity (e.g., not more than a 10 % decrease in ATP levels on the cytotoxicity concentration response curve) is referred to as IC10 (0.1 µM), while the other concentrations used were 1/5th (0.02 µM) and 1/10th (0.01 µM) of the maximum concentration. SAHA concentrations were dissolved in DMSO (Fisher Scientific). To assess the influence of SAHA on 5-hmC DNA modification, B-lymphocytes cells cultured at 37^0C in 6-well plates were exposed to SAHA for 30 or 72 hrs. A comparable cell confluence was attained for each time point by plating 1×10^6 cells in 4 mL of culture media (RPMI 1640 with 1 % l-glutamine (Invitrogen) supplemented with 15 % FCS (USDA tested (Hyclone)) for the 30 hrs time point and 0.3×10^6 cells in 2.4 mL for the 72 hrs time point. For the 30 hrs time point cells, each of the 3 compound concentrations or vehicle (DMSO, with less than 0.4 % DMSO/well) were added at 5X in 1 mL culture media, while for the 72 hrs time point each of 3 compound concentrations or vehicle were initially added at 5X in 0.6 mL culture media and then at 24 and 48 hrs time points, 1X compound concentration or vehicle in 1 mL media was added to each well. Triplicates were performed for each respective treatment and cells were harvested for gDNA extraction. gDNA was isolated with phenol chloroform and isopropanol precipitation and glucosylation, restriction enzyme digestion and analysis on tiling microarray were performed as described before.

2.16. Data analysis

Data analysis was performed with R and SPSS (IBM Corp., USA). Non-parametric tests were used for statistical comparisons if data distribution was determined to be non-Gaussian (Anderson-Darling test, $p < 0.05$ or Shapiro-Wilk test, $p < 0.05$). Unless otherwise mentioned, results are presented as mean ± s.e.m. Since there is genetic overlap between BPD and SCZ and familial aggregation of the two disorders,[25-27] some of our analysis was performed in the combined samples. Furthermore, since bisulfite sequencing is unable to differentiate between 5-mC and 5-hmC (Figure 1.2), DNA modification density determined with bisulfite sequencing in this thesis refers to 5-mC+5-hmC modification density

2.16.1. BSPP-seq analysis

The *HCG9* reference sequence was obtained from the UCSC Genome Browser (*hg19*). Each sequencing read was trimmed to remove low quality bases (Q <20) and short reads (< 80bp). To be certain that mapping was independent of cytosine modification status, all "C"s in the reference and read sequence were replaced with "T"s before mapping. Mapping of sequenced reads was performed for both sense and anti-sense strands with a maximum mismatch of two nucleotides. After sequence alignment was complete, number of "C"s and "T"s were tallied at CpG and CpH sites for each strand within the *HCG9* region. Modification status, measured as the ratio of modified cytosine (^{mod}C) to cytosine was determined by $^{mod}C/C$ ratio= $N_C/(N_C + N_T)$, where N_C is the number of counts for cytosine and N_T is the number of counts for thymine.

Bioconductor[248] packages "ShortRead"[249] and "Biostrings"[250] were used for input and quality assessment of sequence reads and alignment to reference genome, respectively. R package "FactomineR"[251] was used for PCA analysis. Cross sample range for biological samples and technical replicates was calculated as maximum modification at a single cytosine minus minimum modification ($^{mod}C/C$ ratio (Max-Min)). Nucleosome occupancy scores were generated for each nucleotide in the investigated *HCG9* region using nucleosome prediction algorithms.[252,253]

2.16.2. Microarray data preprocessing and normalization

We distinguished between three types of probes on the whole-genome tiling microarray: “target probes”, which contain the restriction site CCGG, “flanking probes”, which do not contain the recognition sequence but could lie up to 200 bp upstream or downstream of a target probe (sheared DNA fragments were of average 200 bp), and “non-target probes” that neither

contain a target sequence nor are within 200 bp (on either side) of the target sequence. Non-target probes are unaffected by enzymatic cleavage, and therefore, can be used as a baseline for array normalization. After comparing different array preprocessing algorithms (section 3.3.2), we chose to use a sequence-based algorithm after Potter *et al* (2008).[254]

Non-target probes were first trimmed to proportionally match target probes in GC content. The probe sequence-based affinity model (equation (1), the "Potter" model) was applied to non-target probes. The fitted value was subtracted from raw intensities of all probes, resulting in normally-distributed probe-level intensities. In equation (1), α corrects for baseline chip-level intensity differences, β represents the total number of each nucleotide, γ and θ for position of each nucleotide. Each chip was individually normalized. All downstream analyses were carried out at the single-probe level (i.e. without windowing or peak-calling) and exclusively on target probes (henceforth referred to simply as 'probes').

$$\hat{y} = \alpha + \sum_{j\epsilon\{A,C,G,T\}} \beta_j n_j + \sum_{j\epsilon\{A,C,G\}} \gamma_j \sum_{k=1}^{25} I(b_k = j) + \sum_{j\epsilon\{A,C,G\}} \theta_j \Big(\sum_{k=1}^{25} I(b_k = j) \Big)^2 \quad (1)$$

Values for various DNA modifications were generated by computing the log-ratios of base channels of a given biological sample (restriction enzyme treatments are indicated by corresponding names in parenthesis; all values are log2-transformed):

$$5hmC = log_2(gDNA_{glu}(MspI)) - log_2(gDNA_{ref}(MspI)) \quad (2)$$

$$5mC = log_2(gDNA_{ref}(HpaII)) - log_2(gDNA_{glu}(MspI)) \quad (3)$$

$$5hmC + 5mC = log_2(gDNA_{ref}(HpaII)) - log_2(gDNA_{ref}(MspI)) \quad (4)$$

5-hmC was measured as the log-difference between glucosylated DNA and native DNA following treatment with *Msp*I. The sum of all the DNA modifications was measured as the log-difference between *Hpa*II-treated unglucosylated DNA and *Msp*I-treated unglucosylated DNA. 5-mC was estimated by the difference between all DNA modifications and 5-hmC.

Following normalization, probes overlapping repeats were excluded (RepeatMasker, simple repeats and segmental duplications; genomic annotations from UCSC genome browser, build *mm8* for mouse, *hg18* for human). Information on arrays and probe counts are listed in Table 2.4. Unless otherwise specified, analysis was done using R software and BioConductor.[248]

For the exon-intron boundary analysis, probes were required to have a 'CCGG' in both the Affymetrix probe sequence and the chromosomal sequence downloaded from UCSC; probes that did not meet this criterion were excluded.

2.16.3. Identification of 5-hmC enriched genes in mouse brain

Probes were filtered for those overlapping genes on either strand (RefSeq genes (*refGene*) downloaded from UCSC, *mm8*; range between *txStart* and *txEnd* columns). Probes overlapping multiple genes (gene defined by MGI gene symbol) were excluded, as were genes containing exactly one probe, resulting in 73,461 probes over 4,357 genes (range = 2 to 238 probes per gene, mean = 17). A repeated-measures ANOVA was conducted on probes in each gene, with tissue as a between-groups factor ('Brain' = 11 samples, 'Other' = 36 samples (9 each of liver, heart, kidney, pancreas)), and the biological sample as a within-groups error term. Probes within a gene were not averaged across samples, nor collapsed within a gene. Gene-wise nominal p-values were adjusted using the Benjamini-Hochberg method of false discovery rate (FDR) correction(BH; $\alpha = 0.01$).[255]

2.16.4. Identification 5-hmC enriched intergenic regions in mouse brain

We identified differential 5-hmC in intergenic regions using probe-wise linear regression. Intergenic probes were defined as probes which did not overlap any RefSeq genes on either strand; 60,721 probes met this criterion. A probe-wise linear regression was conducted, with the regressor being an indicator variable of tissue type ['Brain','Other']c (*lmFit* from the R package limma).[256] The fit was first moderated using Empirical Bayes shrinkage (eBayes), and nominal p-values were adjusted using Benjamini-Hochberg FDR.

2.16.5. Functional annotation analysis of 5-hmC rich genes

Gene Ontology (GO) overrepresentation analysis (ORA) was done using DAVID (Database for Annotation, Visualization and Integrated Discovery);[257] for the background gene set, we used the 5,925 RefSeq IDs associated with the 4,357 genes (defined by MGI symbols) for which tests were performed. The foreground consisted of genes (MGI symbols) identified as enriched based on gene-wise tests. GO-related databases (GOTERM_CC_FAT, GOTERM_BP_FAT, and GOTERM_MF_FAT) were chosen for annotation databases.

DAVID also identifies 'clusters of annotation terms' with member genes that share annotation terms more than expected by chance. In part, this aggregation serves to combine

individual terms into groups potentially representing biological pathways. The 'Classification Stringency' parameter was set to "High" (Default = "Medium") to create smaller clusters with greater overlaps in annotation terms. The Enrichment Score (ES) of an annotation cluster is the geometric mean of the exponents of p-values associated with individual member annotation terms in the cluster.[257]

2.16.6. Categorization of genes by brain cell type

The list of genes with cell-type specific enrichment scores was downloaded from the Supplementary Online Material accompanying a dataset of steady-state mRNA levels in FACS-sorted brain cell populations.[258] Genes with relative mRNA enrichment > 5.0 were called as being enriched in a particular cell-type. Genes with > 20.0 enrichment were deemed to be cell-type specific (after analyses and threshold set in the source paper).

2.16.7. Analysis of genes with particular GO terms, for mouse and human brain

The list of all mouse (or human) genes mapped to specific GO terms was downloaded from the AmiGO Gene Ontology browser (release date 2011-05-07, AmiGO version 1.8, download date 2011-05-13 (mouse), 2011-05-15 (human)). Gene association files were downloaded for GO:0045202 ("synapse"), GO:0044456 ("synapse part"), GO:0007155 ("cell adhesion"), and GO:0005886 ("plasma membrane") (filter for species *Mus musculus* (or *Homo sapiens*); GO evidence codes not filtered). Genic probes were defined as those that overlapped RefSeq genes on at least one strand (*refGene* table from UCSC genome browser, *hg18* for human, *mm8* for mouse). Genes on interrogated tiling arrays were divided into those that were mapped to the GO term being analyzed, and those that were not. Within each group of genes, individual probes were first averaged (mean) across samples in the tissue group (e.g. brain). Probes were not averaged across a gene. GC content of each probe was obtained using the probe sequence provided in the Affymetrix array annotation (bpmap) file.

2.16.8. Relating DNA modifications to mRNA levels with transcriptomic data

We used a previously-published dataset (GSE10246)[259] that measured steady-state mRNA levels in a variety of adult mouse tissues. Normalized expression values were downloaded in series matrix format from the Gene Expression Omnibus,[260] and analyzed in R using the BioConductor package GEOquery.[261] Array annotation was downloaded from Bioconductor ("mouse4302.db"). Probes were averaged across samples within a tissue, and then

averaged within RefSeq IDs.

The transcriptomic dataset was validated prior to use. Samples were subjected to unsupervised hierarchical clustering (distance = Pearson's correlation, clustering method = "ward"), and the cluster heatmap was manually examined to establish that tissues with similar developmental origin were grouped into closer subtrees than tissues from different cellular lineages (heatmap visualization done in Seurat).[262] Further spot checks were done for individual genes with a known characteristic expression pattern (e.g. *Nanog*, a transcription factor expressed in embryonic stem cells, is expected to be relatively overexpressed in ES cell lines and underexpressed in differentiated tissues). RNA samples were separated into brain (["cerebral_cortex_prefrontal", "cerebral_cortex", "cerebellum"], N = 6 arrays), liver, heart, kidney and pancreas (2 arrays each). For each tissue, genes (RefSeq IDs) were stratified into deciles, based on mRNA level.

Separately, in our dataset of DNA modification estimates, samples were grouped by tissue (brain = 11, [liver, kidney, heart, pancreas] = 9 arrays each). For each tissue, probes were first averaged across samples and then within a gene, resulting in one value per RefSeq ID. Genes were binned according to their mRNA expression decile (previous paragraph), and the average quantity of 5-mC or 5-hmC in each decile was computed.

2.16.9. Calculation of exon-intron boundary differential

Boundary regions were defined as regions at a certain distance on both sides of an exon start or end (RefSeq genes, internal exons only). Where multiple exons had the same extent ([chromosome, start,end] location), only one such exon was retained. For each probe, the distance from the modifiable C (of C<u>C</u>GG) was computed to the nearest exon (or intron) boundary; this value is the boundary distance of the probe (a value of zero was exonic in these calculations). Statistics were computed on probes that lay ≤ 5 or 20 bp on either side of the junction. The result at a distance of 20 bp reflected the general trend observed at the boundary, while that at 5 bp reflected immediate cross-boundary change. A linear mixed-effects model was used to test probe intensity differences between the exonic and intronic side of the junction, using junction side (*junctionSide*=['Exon','Intron']) as the fixed-effects term, and sample (*Sample* in eqn 5,6) as random-effects terms (*lmer4* package in R). For datasets that used multiple array types, array type (*Array* in eqn. 5, 6) was used as an additional random-effects term. ANOVA was used to determine whether the data better fit the null model:

$$Intensity = 1 + (Array + Sample) + residual \quad (5)$$

or the alternative model, which took into account the side of the junction on which the probe occurred (*junctionSide*)):

$$Intensity = 1 + junctionSide + (Array + Sample) + residual \quad (6)$$

Tests with P-value < 0.025 were deemed significant.

The Wilcoxon-Mann-Whitney test (WMW test), a more common choice for testing difference in medians, would have been an inappropriate choice to compare exonic and intronic intensities. Our data contained multiple measurements per sample, violating the assumption of independence required by the WMW.

2.16.10. Identification of cassette exons for exon inclusion analysis.

RNAseq from liver and cortex were analyzed to identify cassette exons and their respective inclusion levels.[263] To identify cassette exons, first, all available human expressed sequence tags (ESTs) and mRNA sequences were mapped to the human genome (*hg19*) using SIM4. The information on intron-exon structures was then merged with Ensembl annotation (release 65). From this database, a bowtie library of exon-exon junction (EEJ) sequences was generated by combining every possible splicing donor and acceptor within each gene. RNAseq from liver and cortex was mapped to this library using bowtie with –m 1 –v 2 parameters. Reads were trimmed to 50 nucleotides and reads mapping to the genome were previously discarded (since EEJs must not exist in the genome). A minimum of eight mapped nucleotides were required at each of the two exons in a given EEJ. The outputs were then parsed to identify cassette exons (exons that are either included or fully excluded from the transcripts), by examining exons that have associated reads maps to (i) both EEJs, supporting the inclusion of the exon (constitutive upstream exon (C1)-cassette exon (A) or A-constitutive downstream exon (C2)) and (ii) the EEJ for the exclusion of the exon (C1-C2). Genome coordinates were converted to build *hg18* (liftOver, UCSC genome browser) prior to the analysis with DNA modification arrays.

The inclusion level of an exon was defined as the proportion of gene transcripts in which the exon was present: % Exon Inclusion = 100*(sum(CiA) + sum(ACj)) / ((sum(CiA) + sum(ACj) + 2*(sum(CiC2) sum(C1Cj))), where Ci is any possible splicing donor upstream the alternative exon (including C1), and Cj any possible splicing acceptor downstream the

alternative exon (including C2). Exons with multiple acceptor or donor splice sites or not supported by RNA-Seq were excluded. We defined an exon as 'alternatively-spliced' if the inclusion level was less than 80%, and as 'constitutive' if inclusion was 100 %. DNA methylation was assessed in those exons that overlapped 'target probes' (probes containing CCGG) (probe counts in Table 2.4). Methylation levels of exons were computed at the whole-exon level and at d = 20 bp away from the boundary. Sample-averaged median probe intensities were compared between probes overlapping constitutive and alternative exons (two-tailed WMW test, $\alpha = 0.05$).

2.16.11. Helicos single molecule sequencing (SMS) analysis.

Three technical replicates of the same human brain DNA sample were processed for glucosylation and respective restriction digestion with *Msp*I enzyme with or without glucosylation treatment, and with *Hpa*II enzyme on non-glucosylated gDNA. Data from all three runs were pooled for analysis, after each run had been separately normalized using the corresponding number of non-target reads. SMS reads were trimmed for leading "T" homopolymers, filtered for reads with a minimal length of 25 bases after trimming as well as for other standard Helicos quality metrics using a suite of Helicos tools available at: http://open.helicosbio.com/mwiki/index.php/Releases. Alignments to the *hg18* version of the human genome were conducted with indexDPgenomic software freely available on the Helicos website (http://open.helicosbio.com/mwiki/index.php/Releases). The sequence reads were aligned using a minimum normalized score of 4.3. Only uniquely-mapped reads were considered for the present analysis.

Reads with a 5' coordinate $\leq$ 3 bp from a target sequence (CCGG) were defined as target reads. Reads with a 5' coordinate $\geq$ 200 bp away from a CCGG sequence were used to normalize the read count. Junction distance of target reads were computed as for the microarray analysis, using the coordinate of modifiable cytosine (underlined "C" in "C<u>C</u>GG") of the read-associated target site. Raw reads were first aggregated by junction distance (e.g. distance to exon start/end or intron start/end) respectively for both channels (unglucosylated or glucosylated DNA samples). Aggregated reads were normalized by non-target read count and scaled relative to the number of reads in the channel with non-glucosylated DNA. Percent 5-hmC was computed as the fold-difference in reads from the glucosylated channel, relative to those in the non-glucosylated channel. The proportion of reads arising from CCGG target sequences was greater in the non-glucosylated DNA sample, compared to the glucosylated DNA sample (Appendix 6).

This is expected since higher levels of digestion will generate more DNA fragments with 3' hydroxyl ends, a prerequisite for Helicos single molecule sequencing.

*Hpa*II digestion resulted in more reads than expected from previous estimates of total DNA modification in the average mammalian cell (Appendix 6). One possibility is that the *Hpa*II enzyme generates single strand nicks in the modified DNA, which remained undetected in earlier studies that estimated total DNA methylation within the genome; this observation requires further investigation. For this study, only *Msp*I digestion was taken in account, as it has identical restriction conditions for glucosylated and for non-glucosylated DNA.

Chapter 3. Results

3.1. Bisulfite pyrosequencing of *HCG9*

Our epigenome-wide study in major psychosis using the 12K CpG island microarray revealed a number of DNA modification differences associated with major psychosis including a region overlapping with the HLA complex group 9 (non-protein coding) gene (*HCG9*) (fold change -0.42, FDR p=0.03).[231] In this work we validated our microarray finding by bisulfite pyrosequencing based fine mapping of *HCG9* first exon and a GC-rich region located 3' of the first exon that extends into the 1st intron (Figure 3.1). *HCG9* is located on chromosome 6p21.33 and is of particular interest as recent GWAS studies detected that chromosome 6p21.3-22 was associated with SCZ and BPD.[26,264,265] *HCG9* is a member of the PERB11 gene family and three splice variants of the gene have been identified (Figure 3.1). While the function of *HCG9* is still unknown it has been suggested to play a role in stress response and cell death.[266]

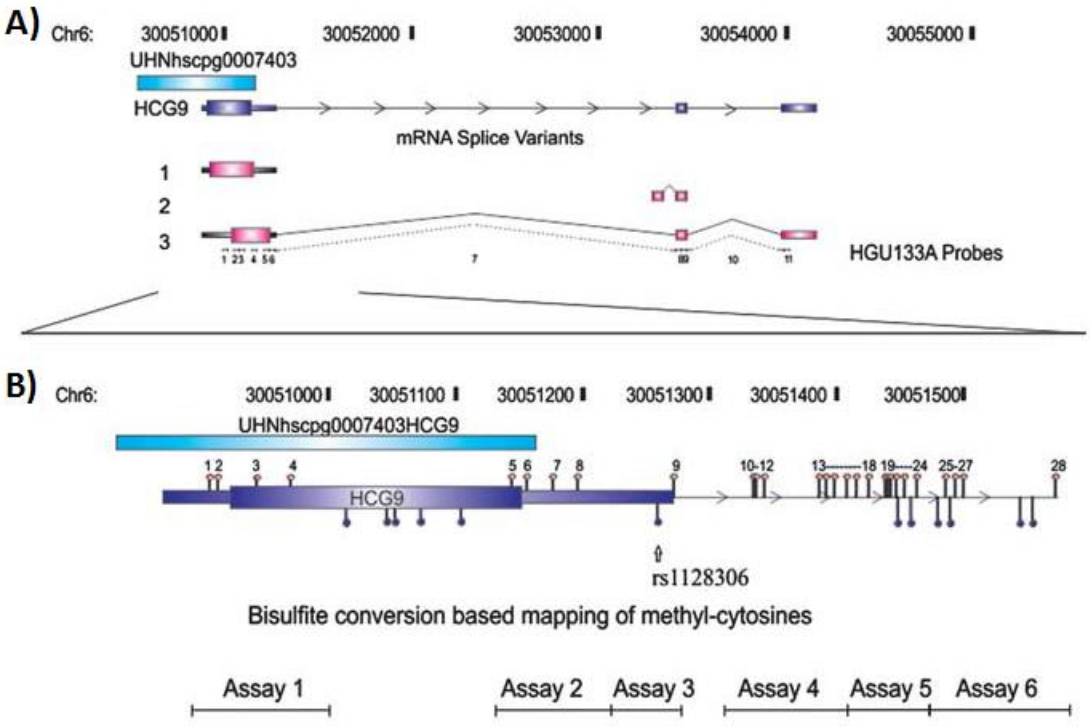

Figure 3.1. The investigated ~700bp *HCG9* region.
A diagram of the *HCG9* gene located on Chr 6p21.33 (A) and magnified region (B) subjected to fine mapping of modified cytosines. Representations are to scale based on the human genome build (NCBI 36) *hg18*. The position of the UHN microarray probe (UHNhscpg0007403) that revealed DNA modification differences in major psychiatric disease in our microarray analysis is shown in light blue. Reference sequence genes are depicted in dark blue while mRNA splice variants 1-3 for *HCG9* as determined by AceView (http://www.ncbi.nlm.nih.gov/IEB/Research/Acembly/) are in pink. Relative gene expression probes from the *Affymetrix* HGU133A microarray relating to this gene are depicted in A. Assay 1-6 in B

represent the regions subjected to sodium bisulfite modification coupled with pyrosequencing. CpG dinucleotides subjected to DNA modification analysis are represented by numbered red lollipops while un-interrogated CpGs and those omitted due to overlap with a dbSNP citation are shown as upside-down blue lollipops

After treatment with sodium bisulfite, each gDNA sample was subjected to six pyrosequencing assays (Figure 3.1). In order to maximize our chances of finding functionally relevant DNA modification differences, we evaluated the mean DNA modification density per individual for all adjacent CpG positions and for all possible window sizes of CpG 1-28 resulting in 406 independent tests. For each of the 406 tested windows, a non-parametric Mann-Whitney U test was performed on these mean *HCG9* modification values between the affected and unaffected groups. P-values obtained from all tests were subjected to Benjamini-Hochberg FDR correction for multiple testing. This method was first used to evaluate DNA modification status in the SMRI brain cohort (see Table 2.1 for details) in attempts to validate the microarray data. In SCZ studies testing the prefrontal cortex of *post-mortem* brains, most results were non-significant (a single CpG window at CpGs13-15 that displayed higher methylation in SCZ as compared with controls; SCZ = 17.4 ± 2.67, Control = 10.4 ± 1.34, p = 0.031). SCZ samples in WBC 2 and germline were also tested with few significant results; and therefore, we decided to focus our efforts solely on BPD.

In BPD studies, numerous *HCG9* regions showed significantly lower DNA modification in the prefrontal cortex of BPD patients as compared to controls (Figure 3.2). The average DNA modification density across CpGs 1-16 was lower in BPD than controls (BPD = 21.7 ± 2.16, Control = 25.5 ± 1.21, Mann-Whitney U test, p = 0.033). Within this region, CpG 3 displayed the largest effect (BPD = 35.6 ± 5.22, Control = 50.4 ± 4.48, Mann-Whitney U test, p = 0.018; Figure 3.2). The second region of interest was within intron 1, where CpGs 23-24 displayed the highest significance for differential modification between the BPD and control groups (BPD = 10.7 ± 3.28, Control 15.2 ± 1.60, Mann-Whitney U test, p = 0.0021; Figure 3.2). None of the p values survived FDR correction.

Given that brain is an extremely complex organ, both morphologically and functionally, we were interested if the detected BPD differences can be detected in other brain regions. In BPD studies using the SMRI brain cohort, no significant disease associations were observed in occipital cortex or corpus callosum. In the parietal cortex, a single window at CpGs 11-12 displayed lower DNA modification density than controls (BPD = 5.89 ± 1.63, Control = 11.8 ± 2.43, Mann-Whitney U test, p = 0.012). Despite being slightly less modified at CpGs 1-16, (the

region that displayed a lower BPD signal in the prefrontal cortex sample) no significant differences were observed in parietal (BPD = 23.8 ± 1.51, Control = 25.5 ± 1.37, Mann-Whitney U test, p = 0.34), occipital (BPD = 26.2 ± 0.92, Control = 27.1 ± 0.79, Mann-Whitney U test, p = 0.48), or corpus callosum brain regions (BPD= 28.8 ± 1.18, Control = 29.8 ± 1.9, Mann-Whitney U test, p = 0.71). These observations suggest that in BPD patients prefrontal cortex exhibited the strongest disease specific epigenetic differences.

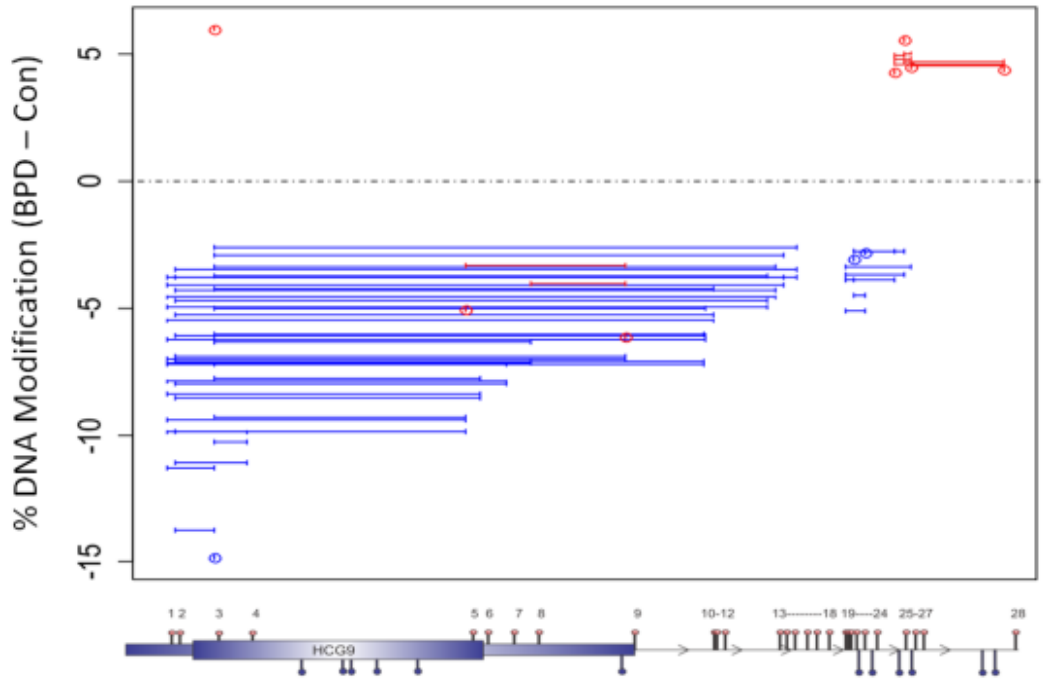

Figure 3.2. *HCG9* modification differences in prefrontal cortex of the brain samples of BPD patients and controls.
Each line shows a region of consecutive CpGs where the mean DNA modification differs significantly between BPD patient and control groups (Blue: SMRI cohort, Red: McL cohort). Circles represent single CpGs. Data are shown only for those windows with $p < 0.05$ (Mann-Whitney U test). The relative genomic positions of CpGs are depicted below.

We next attempted to validate our previous findings in an independent prefrontal cortex cohort (McL, see Table 2.1 for details). The McL brain sample also showed a significant difference between BPD patients and controls at CpG 3, however, in the opposite direction compared to SMRI brain sample (BPD = 37.1 ± 1.36, Control = 31.2 ±1.72, Mann-Whitney U test, p = 0.011; Figure 3.2). A similarly opposite trend was detected at CpGs 23-26 (BPD = 16.3 ± 1.12, Control = 13.7 ± 0.64, Mann-Whitney U test, p = 0.077). A region displaying consistent low modification density in both brain cohorts was identified in the McL cohort between CpGs 5-9 (BPD = 20.5 $\pm$ 1.45, Control = 23.9 $\pm$ 1.15, Mann-Whitney U test, p = 0.044) with CpG 5

displaying the largest effect (BPD = 24.9 $\pm$ 1.74, Control = 29.9 $\pm$ 1.55, Mann-Whitney U test, p = 0.036; Figure 3.2).

3.1.1. Analysis of the discrepancies between the SMRI and McL brain cohorts

In order to understand the reasons for the discrepancies in the two brain collections we investigated the effects of age, medication, and DNA sequence differences on *HCG9* modification density.

3.1.1.1. *HCG9* modification increases with age in controls.

The McL cohort represents significantly older individuals than the SMRI cohort (SMRI = 44.7 yr $\pm$ 1.12 yr, McL = 60.1 yr $\pm$ 1.70 yr; Student's T-test, $p = 2.4 \times 10^{-11}$). In fact, 50% of individuals in the McL sample were individuals over 70 yrs old. For control individuals in each cohort, we correlated the age at death with the average DNA modification density in each of the 406 CpG window combinations. In both cohorts, we detected an increase in DNA modification with age. This phenomenon appears to be fairly generalized over all 28 CpGs in the SMRI cohort (Spearman's rho = 0.30, p = 0.037), with CpGs 7-13 displaying the most significant effect (Spearman's rho = 0.59, p = 0.00039). McL control DNA modification density was positively correlated over CpGs 10-28 (Spearman's rho=0.30, p=0.037), with the strongest effect observed at CpGs 23-24 (Spearman's rho = 0.47, $p = 7.8 \times 10^{-4}$). In a combined analysis of both brain cohorts, CpGs ranging from 3-28 displayed significant positive correlations with age (Figure 3.3A).

In the McL cohort, exclusion of individuals older than 70 yr (remaining BPD patients N = 17, and controls N = 35) eliminated the significantly higher levels of DNA modification in the intronic region at CpGs 23-26 in BPD patients shown earlier. Additionally, we plotted the average DNA modification difference in the most age discrepant groups (30-40 yr *vs.* 70-80 yr) for all CpG windows (Figure 3.3B). DNA modification in younger BPD patients was less than controls, resembling the SMRI results. Together these results corroborate the hypothesis that age differences are at least partially responsible for the discrepant initial findings between the two brain cohorts. Some of our earlier findings, however, cannot be explained by the detected trend of increasing *HCG9* modification in the aging brain. The younger SMRI control group exhibited a higher density of modified cytosines at some positions compared to the older McL control group (SMRI Control = 50.44 $\pm$ 4.48, McL Control = 31.16 $\pm$ 1.72 at CpG 3, Mann-Whitney U test, $p = 1.52 \times 10^{-5}$).

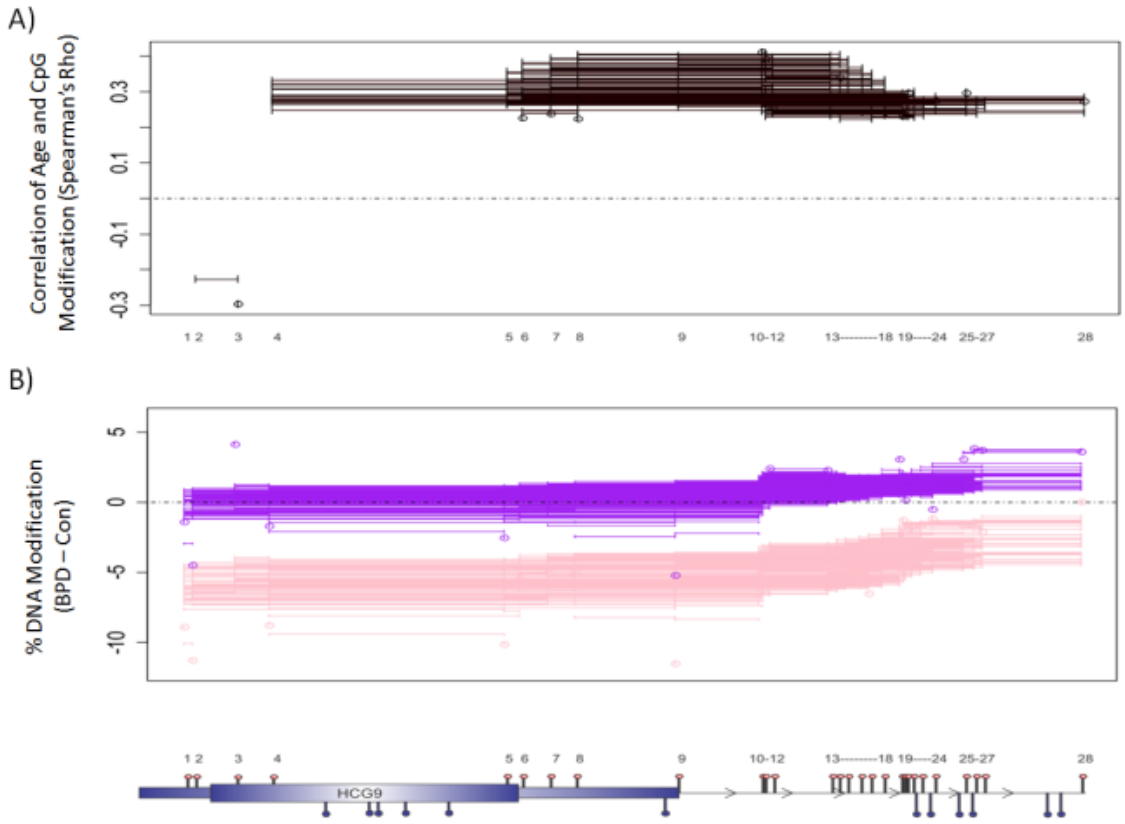

Figure 3.3. Age related increase in *HCG9* modification.
A) Spearman correlation coefficients (rho) for age at death and average DNA modification in control individuals from SMRI and McL brain cohorts combined (N=84); each line shows the corresponding window of CpGs. Only windows with p ≤0.05 are depicted. B) Average difference in DNA modification density between younger (purple: 30-40yrs) and older (pink: 70-80yrs) BPD and control group individuals (McL cohort only). Data shown for all possible CpG window variants.

3.1.1.2. Effect of medication on *HCG9* modification.

In the SMRI cohort (medication information for the McL sample was not available), correlation of lifetime antipsychotic measures (defined as fluphenazine equivalents in milligrams) with average age corrected CpG methylation across the 28 represented CpGs revealed a marginally significant association at a single CpG 14 (Spearmans' r = 0.34, P = 0.05), which became insignificant after correction for multiple testing. These results suggest that antipsychotic medications do not affect DNA methylation at HCG9.

3.1.1.3. Single nucleotide polymorphism rs1128306 is associated with DNA modification differences.

The third factor that may have contributed to discrepancies between the SMRI and McL cohorts is related to the effect of DNA sequence variation on DNA modification status. Using sequencing data obtained for the *HCG9* region, we evaluated 12 single nucleotide

polymorphisms (SNPs) in the region for evidence of allele-specific modification. We detected no DNA modification difference between alleles of SNPs rs17180353, rs2071568, rs373472, rs58031868, rs6903753, rs690402, rs9260832 and rs9278524, while SNPs rs400488, rs422640, rs9278523 and rs1128306 demonstrated significant association with mean DNA modification (Kruskal-Wallis $\chi 2 = 13.26$, Bonferroni corrected $p = 0.0024$ in all cases). These four SNPs were in perfect linkage disequilibrium, and so only rs1128306 (G/A) was used as a marker of genetic polymorphism within this region. DNA modification at each individual CpG position was compared between the GG and GA genotypes (numbers of AA homozygotes were too small to investigate separately and therefore such were included in the GA group) (Appendix 3, Figure 3.4). In the two combined brain samples, CpG 8 displayed the most significant effect (GG minus GA = -10.8 $\pm$ 1.37, Mann-Whitney U test, $p = 2.5 \times 10^{-11}$), while CpG 5 displayed the largest difference between alleles (GG minus GA = -16.8 $\pm$ 2.3, Mann-Whitney U test, $p = 1.7 \times 10^{-9}$). In all tests, modification density for GG homozygotes was significantly lower than GA heterozygotes (Appendix 3).

As the rs1128306 allele A in heterozygotes GA is associated with a higher density of modified cytosines, we tested to see if the SMRI and McL cohorts had differing case and control allele frequencies for this allele. The BPD group in the SMRI cohort has less allele A possessing individuals as compared to that of the McL cohort (allele A frequency: SMRI Control = 0.4, SMRI BPD = 0.16, McL Control = 0.4, McL BPD = 0.26, $\chi 2 = 31.2$, $p = 0.075$). The higher proportion of BPD cases containing the allele A in the McL disease group could be another reason for the observed higher DNA modification in BPD patients in this cohort.

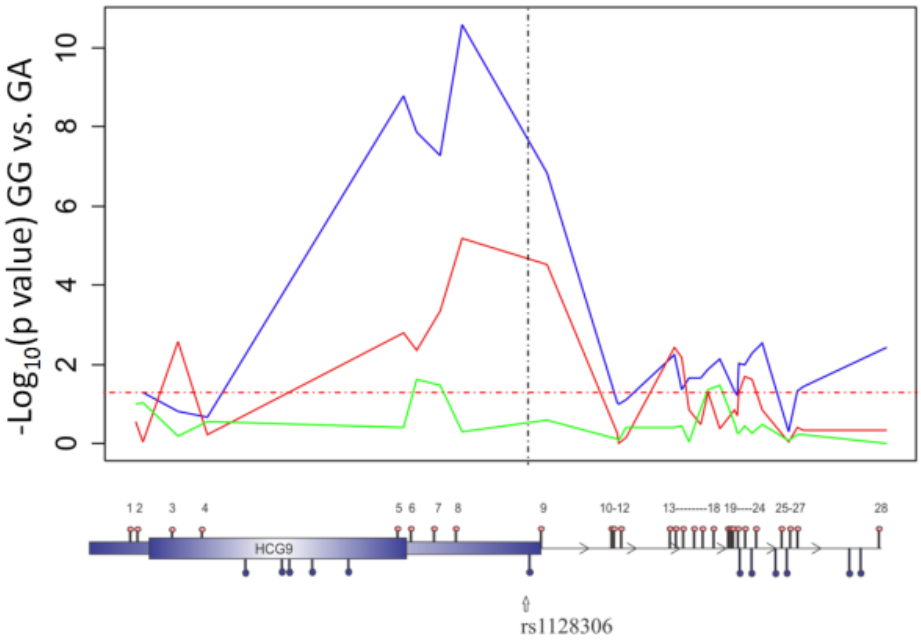

Figure 3.4. Significant differences in DNA modification between rs1128306 SNP genotypes in control samples.
Line plots are used for combined brain sample (blue), WBC (red), and germline (green) to display the effect of the rs1128306 over genomic distance. The position of rs1128306 is depicted by a vertical dashed black line and the negative $\log_{10}$ of $p < 0.05$ is depicted as a horizontal dashed red line.

3.1.2. Combined analysis of SMRI and McL brain cohorts

We eliminated the effects of age and the rs1128306 genotype through linear model transformation and performed non-parametric Mann-Whitney U tests on the corrected modification density across all 406 windows of the overall brain sample (combined SMRI and McL cohorts). This analysis was equivalent to performing an ANOVA test with age and genotype as covariates, with the exception that it was a more conservative non-parametric test. Despite the differences previously identified, these two brain cohorts displayed numerous consistent results in the exon 1 and exon 1 3' UTR region ranging between CpGs 1-10 (BPD = -2.0 ± 1.22 , Control = 1.68 ± 0.86 , Mann-Whitney U test, $p = 0.026$) with CpG 5 displaying the largest effect (BPD= -2.3 ± 1.52 , Control = 1.86 ± 1.13 , Mann-Whitney U test, $p = 0.05$) (Appendix 4). Negative mean DNA modification values are the result of data correction.

3.1.3. *HCG9* modification density in white blood cells

In addition to the brain samples, we investigated the *HCG9* modification profiles using the same approach in peripheral white blood cells (WBC) (Table 2.1). Prior to disease analysis, we evaluated the effect of age on DNA modification in controls from the WBC cohort. Similarly to the brain studies, DNA modification increased with age over a majority of CpGs in the WBC

cohort with significant correlations observed in windows between CpGs 8-28, with the most significant effect observed at CpGs 14-20 (Spearman's rho = 0.3, p = 0.00012). Another similarity with the brain findings was that GG homozygotes at rs1128306 displayed significantly lower modification density compared to GA heterozygotes in the control population (Appendix 3, Figure 3.4), and CpG 8 displayed the most significant difference (GG minus GA= -1.43 $\pm$ 0.56, Mann-Whitney U test, $p = 6.8 \times 10^{-6}$).

After correction for age and rs1128306 genotype, we evaluated DNA modification associations with BPD (Appendix 4). CpGs within the region of CpGs 5-8 displayed lower modification density in BPD compared to controls (BPD = -0.39 $\pm$ 0.14, Control = 0.35 $\pm$ 0.21, Mann-Whitney U test, p = 0.00059) with CpG 6 displaying the most significant difference (BPD = -0.57 $\pm$ 0.23, Control = 0.52 $\pm$ 0.27, Mann-Whitney U test, $p = 2.3 \times 10^{-5}$, FDR p = 0.062). Additionally, a significant difference was observed at CpGs 20-24 (BPD = -0.15 $\pm$ 0.23, Control = 0.15 $\pm$ 0.2, Mann-Whitney U test, p = 0.035), which is a similar region to the most significant finding in the SMRI cohort. All of these identified differences are located in the same region and same direction as those identified in the brain studies.

To gain confidence that significant DNA modification differences in the region of CpGs 5-8 were independent of SNP and age effects, we stratified the entire WBC sample (N = 752) into four groups consisting of rs1128306 GG homozygotes and GA heterozygotes split into the age groups above and below the mean of 42 yr and performed the sliding window-based comparison of *HCG9* modification in BPD *vs*. control groups. In all four groups, the most consistent difference between BPD patients and controls was a lower DNA modification in the region of CpGs 5-8 (Figure 3.5).

In order to understand the BPD predictive value of DNA modification changes in peripheral WBCs, we performed a logistic regression analysis on all 406 CpG combinations, modeling the effect of DNA modification on disease, controlling for age and rs1128306 genotype. Mean DNA modification density at CpGs 6-9 displayed the most significant effect (OR = 1.07, $p = 4.6 \times 10^{-3}$).

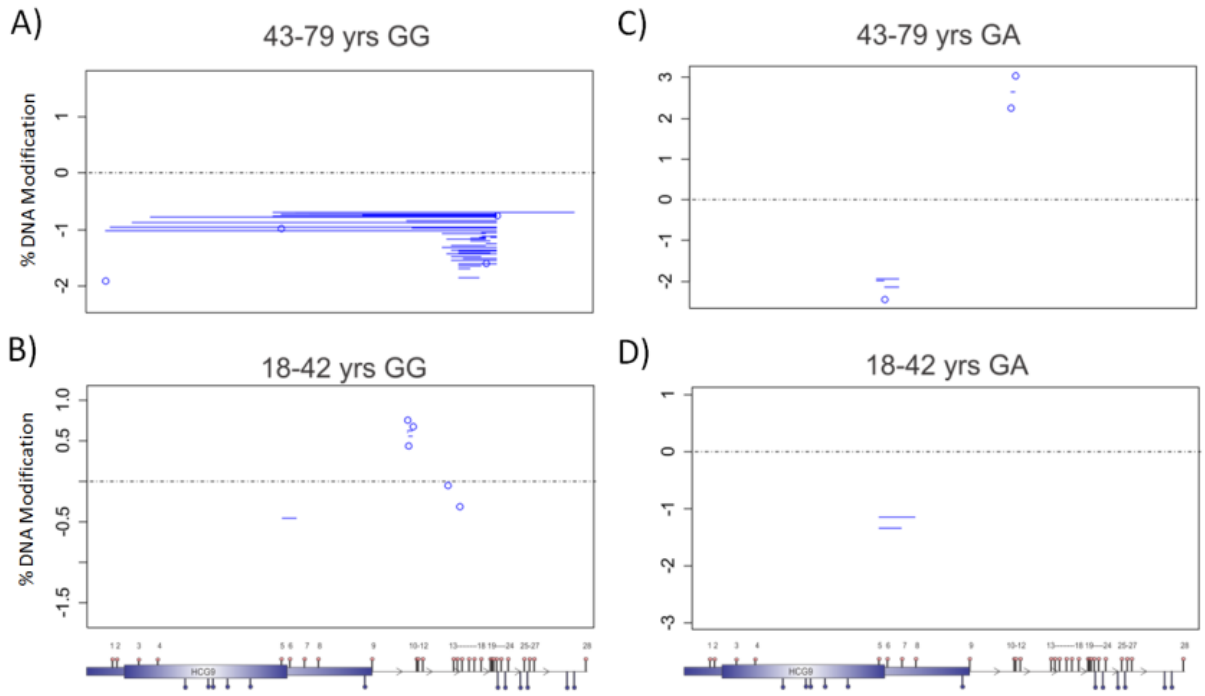

Figure 3.5. *HCG9* modification differences in WBC samples.
To gain confidence that significant DNA modification differences in the region of CpGs 5-8 were independent of SNP and age effects, we stratified the entire WBC sample (N = 752) into four groups consisting of rs1128306 GG homozygotes above (A) and below (B) the mean age and GA heterozygotes above (C) and below (D) the mean age of 42 yrs. Lines representing the mean DNA modification difference between BPD patients and control individuals are plotted for the WBC cohorts for only those CpG combinations with a $p < 0.05$ as determined by Mann-Whitney U test. In all four groups, the most consistent difference between BPD patients and controls was a lower DNA modification in the region of CpGs 5-8.

Next, we evaluated the logistic regression model for its ability to predict a diagnosis of BPD in two ways. As the WBC sample was contributed from two separate sources (see Table 2.1), WBC 1 and WBC 2, we first generated a logistic regression model based on age, rs1128306 genotype and DNA modification density at CpGs 5-8 in the larger cohort (WBC 2) and used it to predict the probability for each individual from WBC 1 to be a case or a control. We compared the predicted probabilities with the true diagnoses over a range of probability thresholds to generate a receiver operating characteristic (ROC) curve based on the calculation of the true positive and false positive rates. The second method was to create our logistic regression model on a randomly selected 90% of the combined WBC sample and to generate a ROC curve based on the remaining 10% consisting of 38 BPD patients and 37 controls. The logistic regression model generated from WBC 2 produces an area under the ROC curve (A prime) value of 0.69, while that based on the prediction of the randomly selected 10% of the sample is A prime = 0.72 (Figure 3.6). This value means that, given one BPD and one control sample, these models have 69% and 72% probability, respectively, of assigning a higher value to the BPD case.

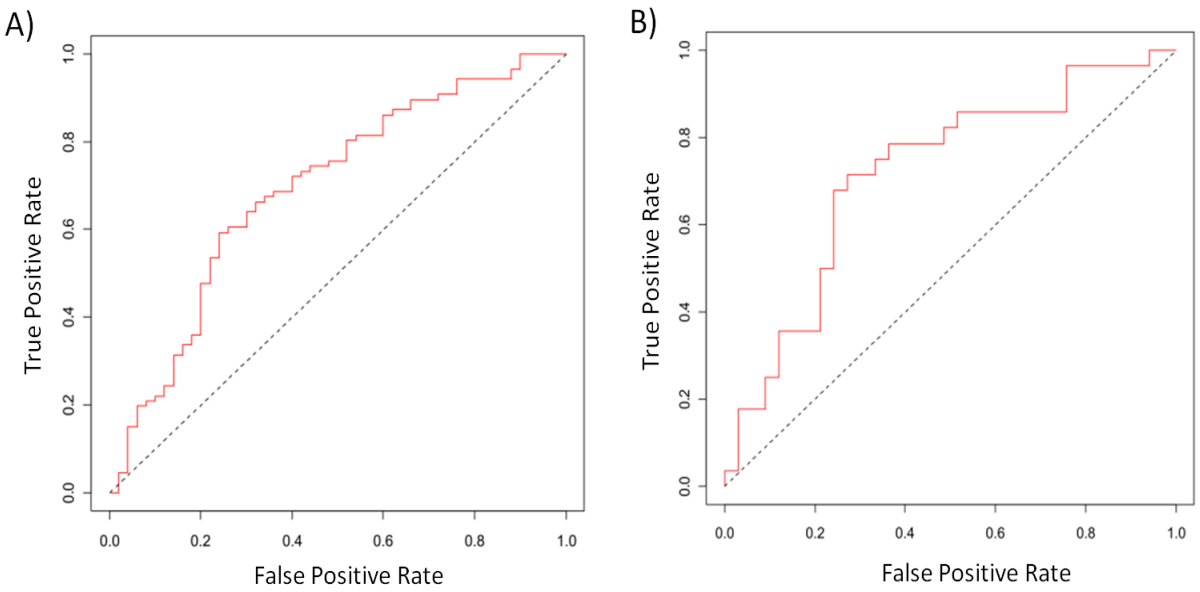

Figure 3.6. BPD predictive model characteristics.
Receiver operating characteristic curves generated for the prediction of the WBC 1 cohort and a randomly selected 10% of the combined WBC cohorts using logistic regression models were generated using the mean DNA modification density at CpGs 5-8, rs1128306 genotype, and age in the WBC 2 (A) cohorts and 90% of the WBC sample (B).

3.1.4. *HCG9* modification density in the germline

In the germline samples, the epigenetic effects of the rs1128306 genotype were consistent with the other tissues; significant differences between GG homozygotes and GA heterozygotes were observed in the region of CpGs 6-7 and CpG 17 (Appendix 3, Figure 3.4). In the germline, unlike the brain and WBC, older age was associated with lower *HCG9* modification density. For example: Spearman's correlation was rho= -0.65 (p = 0.021) for CpGs 4-20, rho =-0.4 (p = 0.033) for CpGs 1, rho = -0.41 (p = 0.035) for CpGs 11-12, and rho = -0.39 (p = 0.036) for CpG 16 (none of these tests survived correction for multiple testing). The opposite age effects in germline and WBC suggests that germline modification differences in BPD patients were genuine and not some artefacts of WBC contamination, which may reach 5% of the total cell count in semen samples. After correction for rs1128306 genotype and age a significantly lower DNA modification density was detected at CpG 5 (BPD = -0.68 $\pm$ 0.44, Control = 0.65 $\pm$ 0.54, Mann-Whitney U test, p = 0.028).

3.1.5. *HCG9* modification density across multiple tissues.

In the process of this project, we noticed that despite variable density of modified cytosines at each specific position *HCG9* modification profiles were similar across the brain, WBC, and germline samples, which suggests that germline *HCG9* epigenetic pattern was

partially retained in the somatic tissues. As expected, the average *HCG9* modification profile across controls in each tissue correlated with the germline (Figure 3.7, Table 3.1).

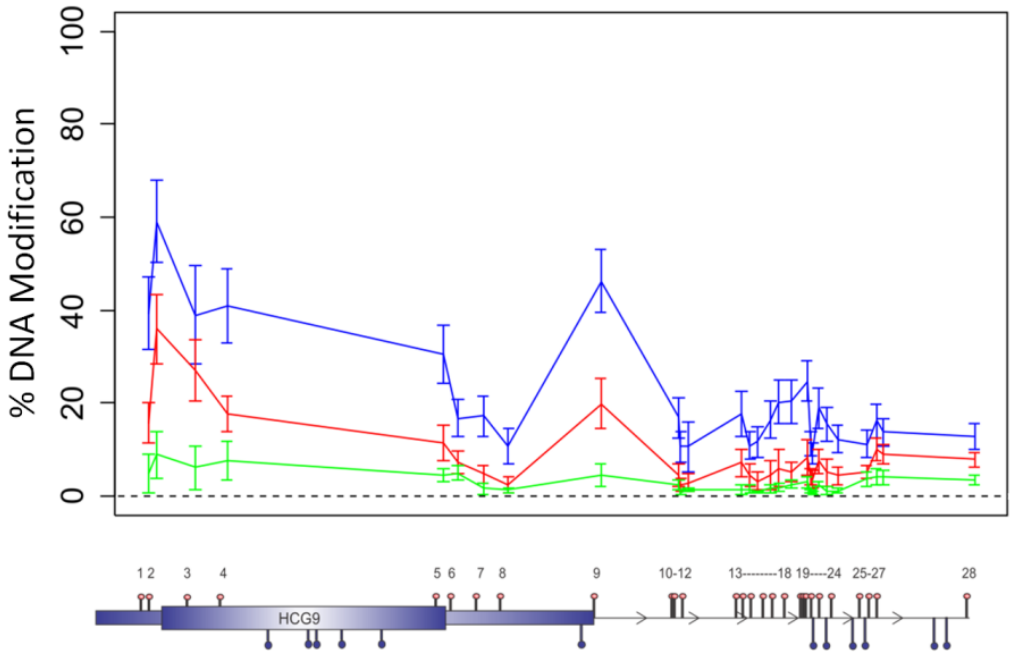

Figure 3.7. Conserved *HCG9* DNA modification patterns across multiple tissues.
Mean DNA modification for the control groups for each of the 28 CpGs for the brain (blue), WBC (red), and germline (green) cohorts are plotted.

Tissue	Correlation SMRI PFC (Rho)	P-value	Correlation Germline (rho)	P-value
SMRI parietal	0.83	1.40×10^{-06}	-0.72	2.70×10^{-05}
SMRI corpus callosum	0.72	0.00021	-0.9	3.50×10^{-06}
SMRI occipital	0.74	1.20×10^{-05}	-0.76	6.60×10^{-06}
SMRI PFC	1	0	-0.8	2.00×10^{-06}
McL PFC	0.65	0.00027	-0.59	0.0013
WBC	-0.77	3.60×10^{-06}	0.74	1.40×10^{-05}
Germline	-0.8	2.00×10^{-06}	1	-

Table 3.1. Comparison of control DNA modification profiles with germline and brain tissue.

We next investigated the relative risk to disease of genotype- and age- independent DNA modification effects using the entire multi-tissue sample. To this end, for brain, WBC, and germline samples we performed a logistic regression, modeling disease as a function of DNA modification, while correcting for the effects of rs1128306 genotype and age. We evaluated the

cumulative effects using a random effects meta-analysis to get a summary statistic for each of the 406 CpG windows. Lower DNA modification at CpGs 5-9 represented the most significant effect on disease risk (OR = 1.24, p = 0.0011), with CpG 8 demonstrated the highest odds ratio (OR = 1.3, p = 0.018; Figure 3.8). The results of this analysis suggest that, independent of age and the rs1128306 genotype, lower DNA modification at *HCG9* at CpGs 5-9 is associated with BPD across the brain, WBC, and germline.

Interesting to note, that when rs1128306 polymorphism alone was evaluated in BPD evidence for association was observed for the combined brain sample and WBC sample; however, allele G exerted an increased risk in the combined brain (OR = 2.45, p = 0.017) but protective effect in WBC (OR = 0.72, p = 0.044, respectively). No association was observed in the germline sample or an analysis of all samples combined.

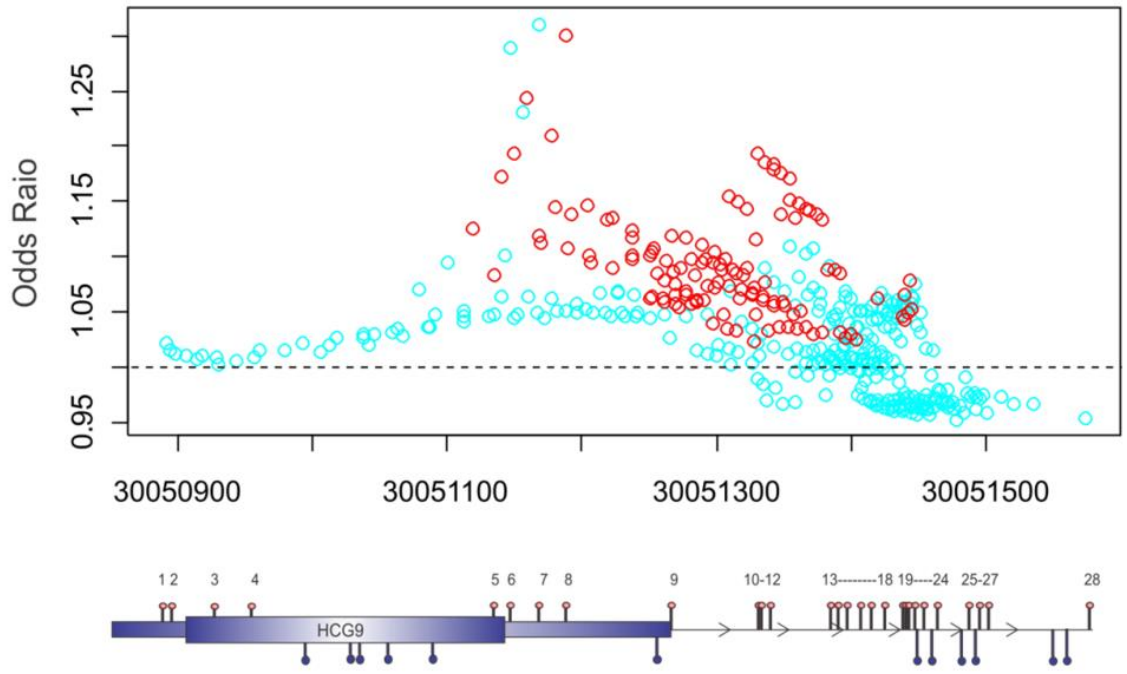

Figure 3.8. Combined sample analysis of BPD risk by genomic location.
The Odds ratio for BPD risk due solely to DNA modification differences in BPD patients and controls is plotted relative to the middle genomic position of each of the 406 tested CpG windows. The rs1128306 SNP and age were modeled as covariates in a logistic regression model and odds ratios due to DNA modification density were combined with a random effects meta- analysis for all tissues. The middle genomic positions of CpG windows where DNA modification significantly contributes to risk are plotted in red.

3.1.6. Putative pathological roles of *HCG9* in BPD

As the pathological role of intronic modification differences at *HCG9* in BPD are not completely clear, we sought to elucidate any relationship between DNA modification differences

and *HCG9* steady state mRNA levels. We analyzed three gene expression data sets (Studies 2, 3, and 7) from the SMRI genomics database (https://www.stanleygenomics.org/) that were performed on the *Affymetrix* HGU133A gene expression arrays, which contain 11 probes covering *HCG9* (Figure 3.1). A limited number of McL cohort individuals (N = 34) also contained steady state mRNA data from the same array platform, and we performed *HCG9* mRNA analyses on the combined SMRI and McL brain dataset. As there are indications that intronic chromatin modifications may be involved in alternative splicing, we were interested to ascertain steady state mRNA levels for individual *HCG9* splice variants. In order to do this, we subjected all data to background correction and quantile normalization and took the mean $\log_{10}$ probe value for each of the 11 probes spanning *HCG9*. Probes 10 and 11 (probe group 1) uniquely bind mRNA from splice variant 3, while probes 1-4 (probe group 2) bind mRNA from variants 1 and 3. After averaging these two probe groups, we determined the levels of *HCG9* splice variant 1 by taking the residuals of a linear model between probe groups 1 and 2, statistically subtracting splice variant 3 from the combined 1 and 3 measures. A significantly higher ratio of *HCG9* mRNA splice variant 3/1 was detected in BPD compared to controls (BPD = 1.45 $\pm$ 2.17, Control = -1.77$\pm$ 15.6, Mann-Whitney U test, p = 0.034).

Since the maximal DNA modification differences in BPD were detected around CpG 5 in the two brain cohorts, and CpGs 5-9 represented the most significant effect on disease risk in the total sample of the three tissues, we hypothesized that epigenetic modifications may be contributing to splicing decisions in this region which corresponds to the overlapping splice variants 1 and 3. The ratio of splice variant 3 over splice variant 1 was tested using the DNA modification sliding window. CpG 6 (i.e. next to CpG 5) displayed a significant negative correlation with splice variant ratios (Spearman's rho = -0.23, p = 0.029; Figure 3.9) although did not survive correction for multiple testing. The largest correlation was detected for CpG 25 (Spearman's rho = -0.44, p = 5.33×10^{-5}, FDR p = 0.022; Figure 3.9), which is interesting as CpGs within this region displayed the most significant difference between BPD and controls in the SMRI brain cohort reported above. Interestingly, both CpG 6 and CpG 25 corresponded to peak mean histone occupancy scores as predicted by a nucleosome prediction algorithm[252] (Figure 3.9). Two additional nucleosome prediction algorithms return similar results, adding confidence that nucleosomes occupancy these positions. An algorithm developed by Mavrich *et al.* (2008)[267] mapped tentative nucleosome centers 50 bp downstream of CpG 6 and ranging from 50 bp downstream to 9 bp upstream of CpG 25. An alternative program[268] mapped nucleosome centers 50 bp upstream of CpG 6 and 7 bp downstream of CpG 25. As the amount of DNA that

coils around a single histone is ~150 bp, CpGs 5-8 and 24-27 can be reasonably associated with a single histone peak. Consistently with our initial observation of lower modification density at *HCG9* in BPD patients, these data suggest that as *HCG9* DNA modification decreases in BPD, there is an increase in the ratio of splice variant 3 to splice variant 1 in the brain.

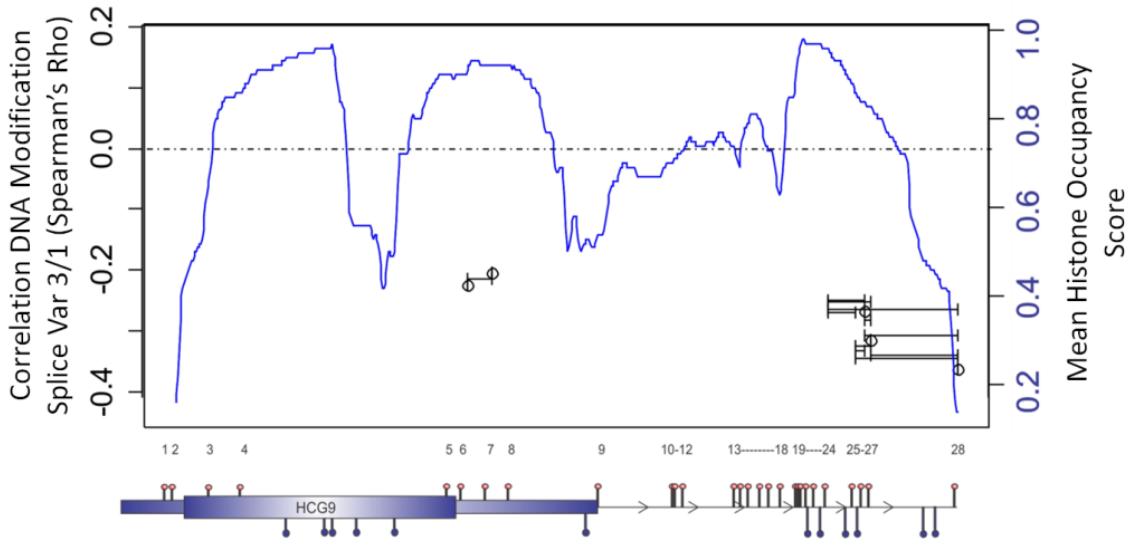

Figure 3.9. DNA modification *vs. HCG9* steady state mRNA levels.
Non-parametric correlations between DNA modification and the ratio of *HCG9* splice variants 3 and 1 (windows with $p \leq 0.05$) (black, Y axis left), and mean histone occupancy scores as modeled by a sequence based histone prediction program (blue, Y axis right).

3.2. Bisulfite padlock probe-based sequencing (BSPP-seq) of *HCG9*

Note: The following credit line is included for results presented in this section as part of the licencing agreement with Oxford University Press.

Pal *et al.*, High Precision DNA Modification Analysis of *HCG9* in Major Psychosis, Schizophrenia Bulletin, 2015, pii: sbv079 [Epub ahead of print] PubMed PMID: 26078387, by permission of Oxford University Press

For a more detailed analysis of *HCG9* we used the recently developed BSPP-seq method in a cohort of *post-mortem* prefrontal cortex tissue from patients affected with SCZ, BPD and unaffected controls (N = 81, for details refer to section 2.1 and Table 2.1). BSPP-seq offers several advantages over traditional bisulfite pyrosequencing, including strand- and allele-specific quantification of modification density in CpG and CpH context.

Padlock probes were designed for the total length of ~5kb, which includes the full length *HCG9,* plus 1kb upstream and downstream of the gene. Not all probes, however, generated sufficient number of reads and we only considered CpG and CpH sites with read depth ≥ 30X in at least 10 samples in each group of patients and controls. Sites that did not meet these selection criteria were removed from further analysis. Each DNA sample was then characterized by 96,470 ± 8,733 (mean ± s.e.m) read counts (total counts per sample are shown in Figure 3.10). The reads, which met the above criteria, covered a total of 34 of 149 CpG sites and 379 of 2,364 CpH sites. Figure 3.11 shows genomic distribution of investigated CpG and CpH sites. On average, each sample generated reliable DNA modification estimates for 21 CpG sites and 213 CpH sites.

To estimate the bisulfite conversion rate, we used DNA devoid of any modifications. This control was prepared using whole genome amplification of two unrelated gDNA samples, which were subjected to bisulfite conversion and deep sequencing using the same set of padlock probes. Cytosine conversion to thymine rate was 99.34%, indicating highly efficient bisulfite conversion. We also observed greater biological variability than technical variability in our dataset. Technical replicates of 8 samples showed higher pairwise correlation compared to biological samples (Pearson's r = 0.82 ± 0.06 and 0.77 ± 0.02, respectively) and cross sample range was significantly higher for biological samples compared to technical replicates (0.31 ± 0.03 and 0.13 ± 0.01, respectively; Mann-Whitney U test, $p = 4.8 \times 10^{-5}$).

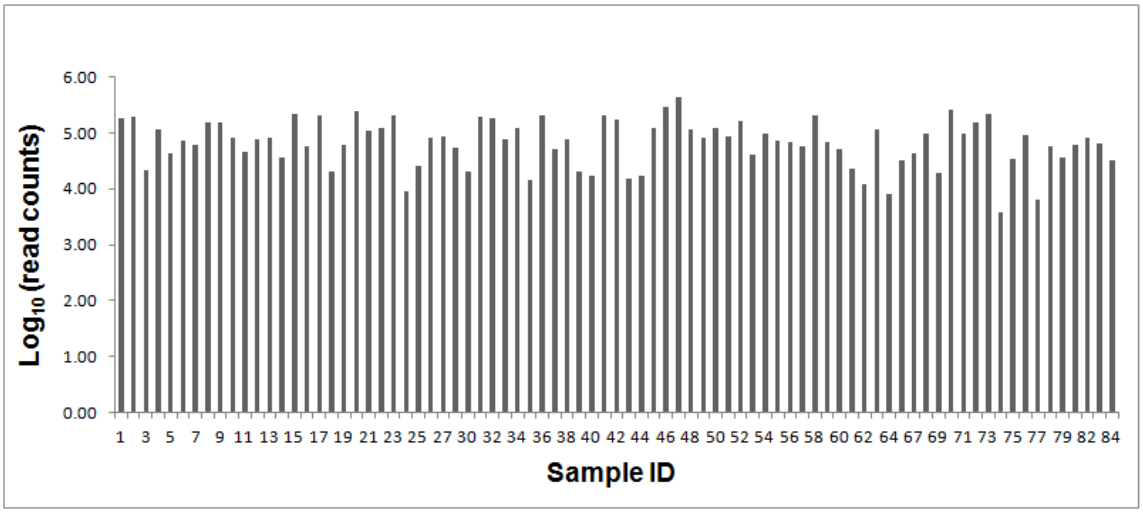

Figure 3.10. Total reads obtained per sample after alignment of unique reads.
X-axis represents Sample I.D. number for each tissue sample, and y-axis represents the Log_{10} of read counts.

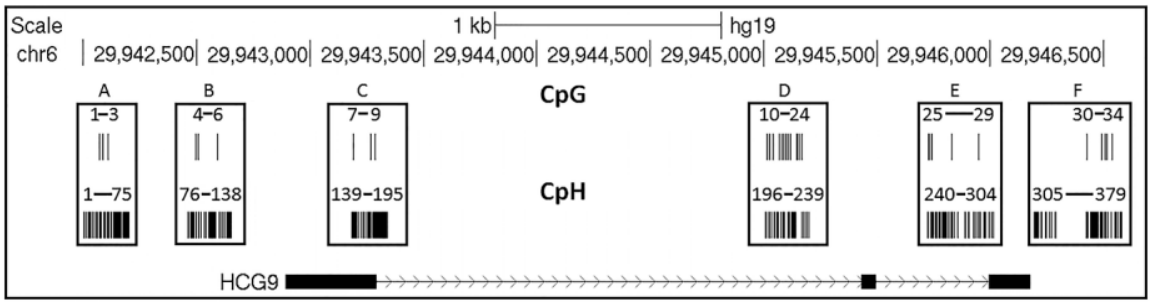

Figure 3.11. The finely mapped DNA modification regions of *HCG9* (UCSC Genome browser, version hg19).
Black bars represent investigated CpG (N = 34) and CpH (N = 379) sites while the numbers represent their relative position. The interrogated sequence was further broken down into 6 smaller regions (A-F).

3.2.1. CpH modification differs in the sense and anti-sense *HCG9* strands

One of the important advantages of bisulfite sequencing is the ability to investigate DNA modification profiles separately on the sense and anti-sense strand. Mean CpG modification was found to be similar for sense strand compared to anti-sense strand (^{mod}C/C ratio; 0.83 ± 0.18 and 0.85 ± 0.13, respectively; Wilcoxon signed ranks test, p = 0.6). This was also true for SCZ, BPD and control groups when analyzed separately.

Multiple recent publications have reported CpH modification in adult human and mouse brain.[96,99,100,102] In our dataset, we observed mean CpH modification density of 1.69%, although 17 CpH sites far exceeded this average with modification density greater than 10%. Since CpH sites are asymmetric on the two DNA strands and cannot be compared individually, we analyzed modification density across all CpH sites to estimate differences between sense and anti-sense

strands. We found significantly lower modification for sense strand compared to the antisense strand (^{mod}C/C ratio; 0.01 ± 0.002 and 0.02 ± 0.002, respectively; Wilcoxon signed ranks test, p = $7.7x10^{-7}$). We suspected a potential bias in our estimation of strand differences could be due to greater number of CpH sites on anti-sense strand (N = 289) compared to sense strand (N = 90). To address this potential issue, we compared CpH modification density from regions E and F (Figure 3.11) with identical number of CpH sites (N = 42) on each strand. Consistent with the earlier observation, sense strand exhibited significantly lower degree of DNA modification compared to anti-sense strand (^{mod}C/C ratio; 0.014 ± 0.003 and 0.04 ± 0.003, respectively, p = $4.1x10^{-7}$). As shown in Figure 3.12, significantly lower modification density for sense strand *vs.* anti-sense strand was similarly observed for control, BPD and SCZ groups ($p \leq 0.011$).

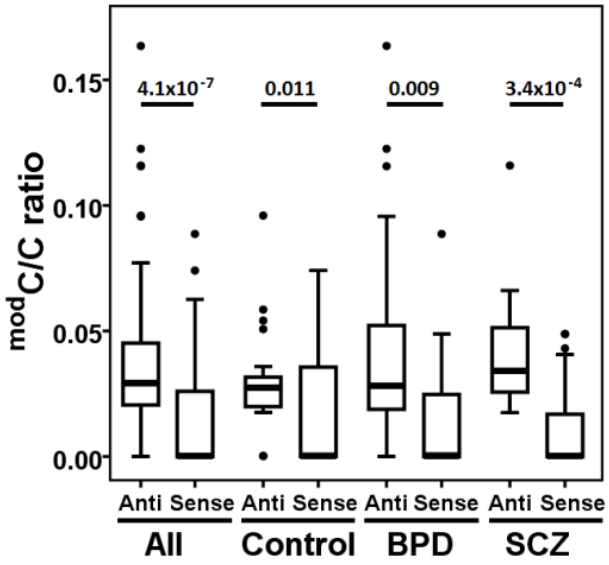

Figure 3.12. Strand bias in CpH modification.
Boxplots represent modification density for sense and anti-sense strand for region E and F (chr6: 29,945,722 - 815 and 29,946,419 - 576, respectively) with equal number of CpH sites on both strands. P-values were obtained from Wilcoxon signed ranks test and are shown above boxplots.

3.2.2. *HCG9* CpH modification is enriched in CpA dinucleotides and nucleosome linker regions

To further explore the biological roles of CpH modification we identified modified CpH sites (≥2% density of ^{mod}C present in at least 10 individuals). Consistent with previous findings,[102,269] multiple sequence alignment[270] of modified CpH sites showed CpH modification occurs predominantly at CpA dinucleotide of *HCG9* in brain gDNA (Figure 3.13A). We next determined CpA dinucleotide frequency in both strands since it could account for modification differences between anti-sense and sense strands. Occurrence of CpA dinucleotide was similar in

both strands (6.2% for anti-sense and 6.8% for sense strand), despite significantly higher CpH modification in the anti-sense strand.

Next, we compared the selected modified CpH sites to the remaining unmodified CpH sites for nucleosome occupancy using two different algorithms,[252,253] which assign a score for each nucleotide ranging between 0 (linker region) and 1 (nucleosome center). Both algorithms predicted a significantly lower mean nucleosome occupancy score (Mann-Whitney U test, $p < 0.01$), indicating that modified CpH sites were predominantly present in linker regions (Figure 3.13B).

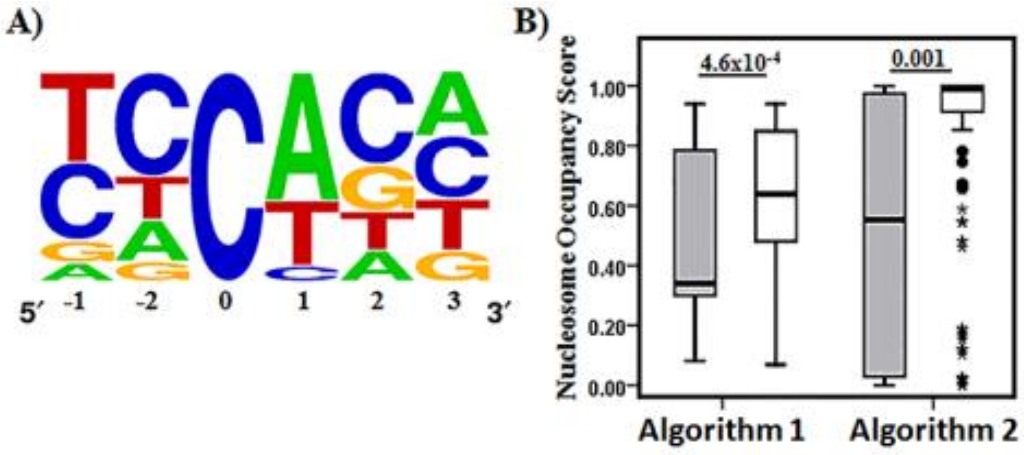

Figure 3.13. Comparison of modified and unmodified CpH sites.
A) Multiple sequence alignment[270] of modified CpH sites (i.e. CpH sites with ≥2% modification density in at least 10 individuals) with modified cytosine at position 0. B) Nucleosome occupancy score determined for modified (grey bars) and unmodified (white bars) CpH sites by nucleosome occupancy prediction algorithm 1[252] and algorithm 2.[253] Significantly lower score for modified CpH sites indicate they are present in nucleosome void regions of *HCG9*. P-values above boxplots were obtained from Mann-Whitney U test.

3.2.3. *HCG9* DNA modification differences in major psychosis and control brains

To identify significant differences in DNA modification between cases and controls, we eliminated effects of age, sex, PMI, and neuropathological differences through linear model transformation followed by non-parametric Mann-Whitney U test. Mean CpH modification density at selected CpH sites was significantly higher in major psychosis samples compared to controls (^{mod}C/C ratio, 0.016 ± 0.01 and -0.024 ± 0.008, respectively, $p = 0.006$; negative mean DNA modification densities are a result of linear model correction). Similar results were also obtained when SCZ and BPD groups were analyzed separately (^{mod}C/C ratio; SCZ = 0.022 ± 0.01 *vs.* control = -0.015 ± 0.008, $p = 0.01$; BPD = 0.022 ± 0.02 *vs.* control = -0.019 ± 0.008; $p = 0.02$). In the analysis of individual cytosines (CpGs 1-34, Figure 3.11), we detected significant

differences in DNA modification at CpG 29, located in the second intron of *HCG9*, between BPD group and controls (^{mod}C/C ratio; 0.10 ± 0.01 and -0.09 ± 0.05, respectively, Bonferroni corrected p = 0.004). Although BPD and controls were not matched for sex we did not find any significant association between sex and CpH or CpG 29 modification density.

To determine how CpH modification levels change at *HCG9* with age, we divided the brain samples into three age groups (i.e. < 40 yr, 41-60 yr and > 60 yr) and compared the CpH modification density at selected CpH sites amongst the groups. For BPD patients, we found significant DNA modification differences between different age groups (Kruskal-Wallis test, p = 0.027) with sharp decline in CpH modification from age group < 40 years to 41-60 years (Figure 3.14). In SCZ patients a similar decline in CpH modification for age groups < 40 years to 41-60 years was observed, however, sample numbers were too low for statistical comparison. No significant changes were observed in control samples (p = 0.85). Significant age induced decline of CpH modification in BPD was also observed after eliminating the effects of sex, PMI, and neuropathological differences (^{mod}C/C ratio; <40years = 0.09 ± 0.05, 41-60years = -0.01 ± 0.01, >60years = -0.002 ± 0.016, Kruskal-Wallis test, p = 0.04).

Epigenetic profiles vary significantly between neuronal and glia cells[269] and inter-individual variation in brain cellular composition may generate false positive epigenetic differences.[271] For a subset of BPD and control brain samples (N = 19 BPD, 15 control) we had access to data on the proportions of neurons and glia, which were estimated using a cell epigenotype specific model. DNA modification at selected CpH sites and CpG 29 were found to be significantly higher in BPD samples compared to controls after correction for neuron and glia proportions, age, sex, PMI, and neuropathological differences (^{mod}C/C ratio; CpH: BPD = 0.020 ± 0.016, controls = -0.026 ± 0.016, p = 0.03; CpG 29: BPD = 0.10 ± 0.02, controls = -0.11 ± 0.05, Bonferroni corrected p = 0.002). Similar to the age effects detected in the whole BPD sample, BPD subsample showed age-dependent differences in CpH modification density after correction for neuron and glia proportions, sex, PMI, and neuropathological differences (<40years = 0.13 ± 0.06, 41-60years = -0.001 ± 0.014, >60years = -0.02 ± 0.006, Kruskal-Wallis test, p = 0.01). No significant differences were identified in control samples (p = 0.6).

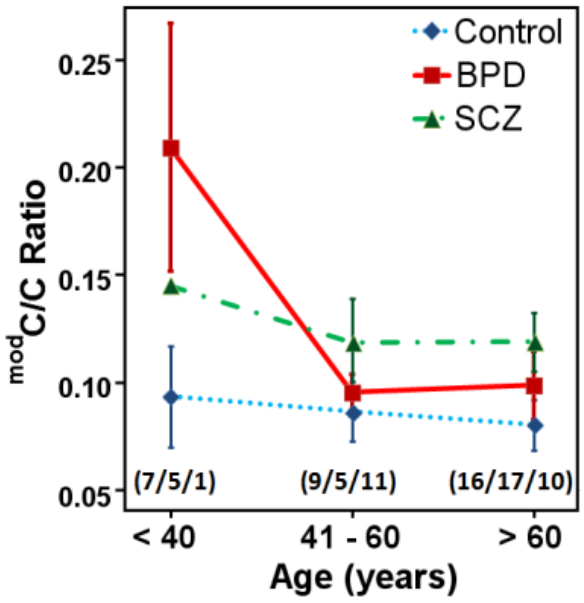

Figure 3.14. Age-stratified changes in *HCG9* CpH modification density in the brain. Line plots represent mean CpH modification density for control (blue), BPD (red) and SCZ (green) individuals for different age groups. Numbers of samples in each age group are presented in brackets as (controls/BPD/SCZ) and error bars represent s.e.m. Most dramatic difference was observed in BPD age group younger than 40 years compared to the older BPD and SCZ patients as well as controls in all age categories.

3.2.4. Hierarchical clustering of *HCG9* modification in major psychosis

Comparison of Euclidean distances between samples revealed higher variation in DNA modification for major psychosis compared to control individuals (3.5 ± 0.03 and 3.2 ± 0.04, respectively; Student's t test, $p = 1.9 \times 10^{-10}$). Variability in epigenetic profiles was further investigated using hierarchical clustering and principle component analysis (PCA) of *HCG9*. Hierarchical clustering of region A (Figure 3.11) identified a subgroup of major psychosis samples, which clustered together and formed a separate clade from the rest of the samples (Figure 3.15A). Figure 3.15B shows separation of the identified clusters in three dimensional space based on the first three principal components. As expected, we found CpH modification to be a significant driving factor behind the variance explained by the first principle component (Fisher's combined probability test using FDR corrected Pearson's correlation p-value between PC1 and CpH modification; $p = 6.5 \times 10^{-5}$). The separation of clusters was not associated with age, sex or neuropathological changes. Lack of clinical information and relatively small brain sample size prohibited further exploration of associations between clusters and phenotypic peculiarities of major psychosis.

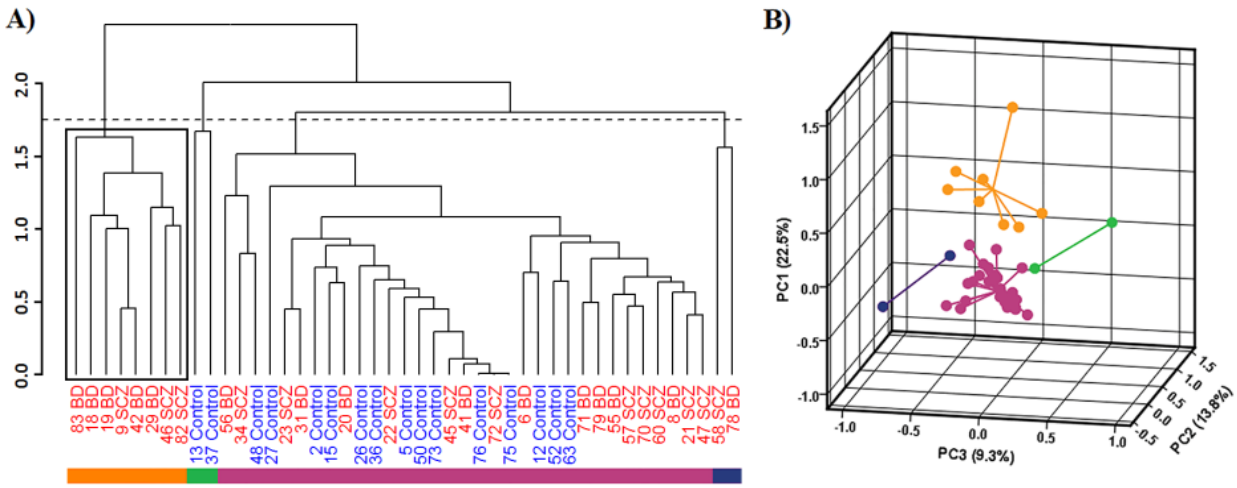

Figure 3.15. Inter-individual variation of modification density in major psychosis and control samples for *HCG9* region A.

Samples with no missing value for 3 CpG and 75 CpH sites in region A were included. A) Dendrogram obtained from hierarchical clustering for individual samples (SCZ and BPD in red and control in blue). Bottom color bar represents clusters identified at the cut height indicated by the horizontal dashed line. A subset of major psychosis samples (black box) formed a separate cluster from the remaining samples. B) Three dimensional representation of first three principal components. Each dot represents an individual while dot colors represent the cluster to which they belong. The lines connecting each dot merge at the geometric center of each cluster. PC1, PC2 and PC3 are shown on the y-, x-, and z- axis, respectively, and variance explained by each principal component is represented in brackets.

3.3. Distribution of 5-mC and 5-hmC across multiple mammalian tissue

Traditional methods of assaying DNA modifications have either been unable to differentiate between 5-mC and 5-hmC (e.g. bisulfite mapping) or have been specific for 5-mC (e.g. antibodies against 5-mC). Relative to other tissues, 5-hmC is particularly enriched in the brain, as observed in mice and humans.[77,78] We mapped both 5-mC and 5-hmC in a variety of neuronal and non-neuronal tissues from mice and humans to investigate their respective roles.

Our method for 5-hmC enrichment relies on transfer of a glucose molecule specifically to the hydroxymethyl group of 5-hmC by T4-phage β-glucosyltransferase (BGT) enzyme, thus rendering it resistant to digestion by the methylation insensitive *Msp*I enzyme at the C^{hm}CGG target site[244,245] (Figure 2.2). 5-hmC is thus estimated by differential resistance to *Msp*I-digestion with and without glucosylation of gDNA. *Hpa*II (targets the same site, CCGG) cannot cut C^{m}CGG or C^{hm}CGG, and conceptually its difference with *Msp*I digestion is a measure of both 5-mC and 5-hmC. Subtraction of the 5-hmC estimate from the *Hpa*II-based estimate therefore measures 5-mC.

3.3.1. Validation of 5-hmC assay

We performed three groups of control experiments to demonstrate the validity of using BGT to estimate the quantity of 5-hmC. First, glucosylation treatment was investigated on a 31-mer DNA duplex that contained 5-hmC modification (Figure 3.16A). Second, we determined the influence of the glucosylation treatment on unmethylated (C) and on methylated cytosines (5-mC). This was performed on whole genome PCR-amplified DNA that had lost all genomic modifications. The glucosylation and restriction digestion procedure was then applied to either whole-genome amplified (WGA, unmethylated genome) or SssI methyltransferase treated WGA DNA (fully methylated genome). Quantitative PCR (qPCR) was used to estimate the % modifications (5-mC or 5-hmC) present at specific loci (Figure 3.16B). These two control experiments showed that there is no influence of the BGT glucosylation procedure on 5-mC or on unmethylated cytosines, and that it is specific for 5-hmC. As a third control, we evaluated the linearity of the measure of 5-hmC by employing the BGT-based procedure in a model system (Figure 3.16C). Incremental amounts of a 5-hmC modified DNA fragment containing one *Msp*I site was mixed with unmodified DNA fragment of the same sequence, glucosylated with BGT followed by *Msp*I digestion and qPCR analysis. Increasing amount of 5-hmC modified fragment corresponds linearly with decrease in *Msp*I digestion.

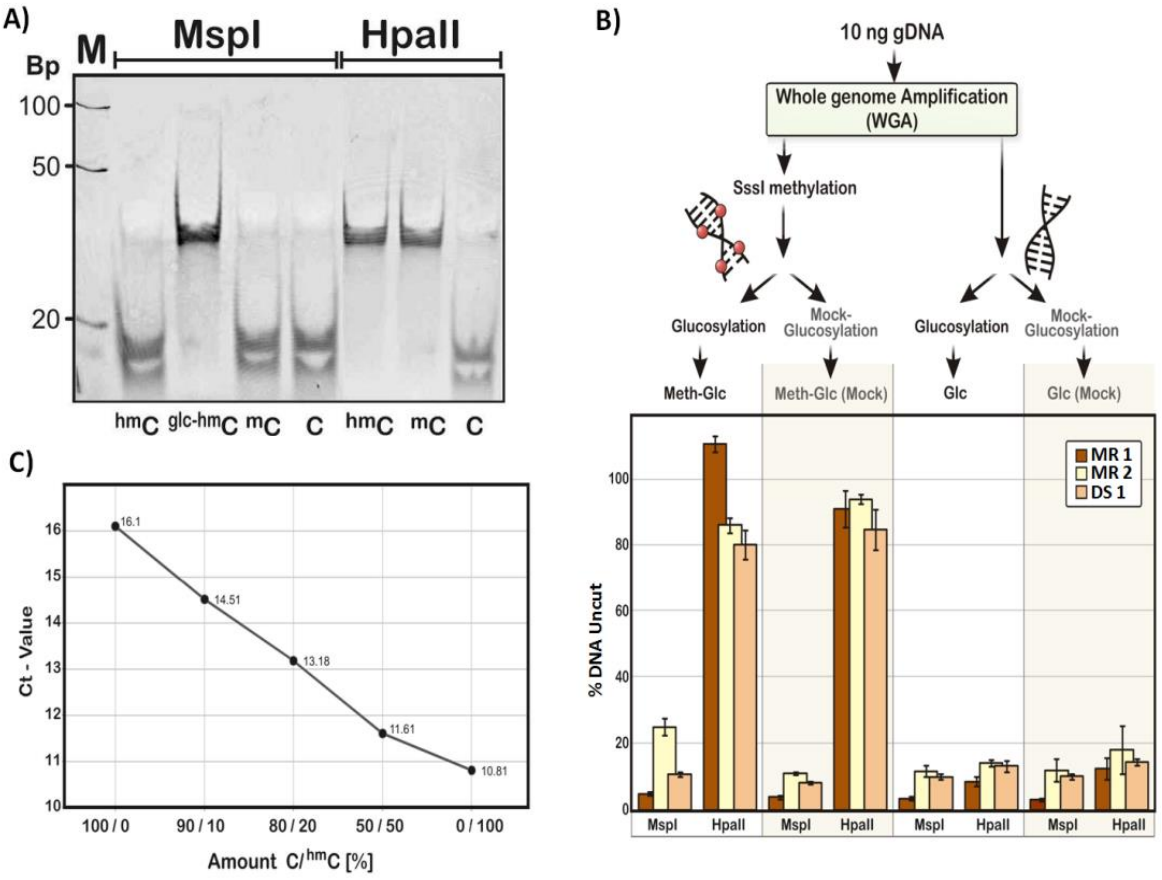

Figure 3.16. Validation for BGT-glucosylation as an assay for 5-hmC measurement.
A) Effect of glucosylation treatment on a 31-mer DNA duplex containing 5-hmC, 5-mC or C on one strand of a CCGG target site. Cleavage by *Msp*I is only blocked by glucosylation of the 5-hmC residue (lanes 1–4); *Hpa*II digestion is inhibited when either modification, 5-mC or 5-hmC (lanes 5–7). B) % DNA modification present at 3 different loci (three bars in a series) as measured by qPCR following restriction enzyme digestion. The experimental design is shown in the upper panel. Each sample gDNA was amplified using phi29 in order to erase all modifications on the gDNA, and was then divided in half. One half was completely methylated using SssI methyltransferase while the other half was not modified. Methylated DNA and unmodified DNA were further divided in half, and one part was glucosylated while other part was mock-glucosylated. From each treatment group DNA was subjected to restriction enzyme digestion with *Msp*I or *Hpa*II followed by qPCR. Presence of methylation restricts *Hpa*II digestion; however *Msp*I digestions occur at methylated cytosine or unmethylated cytosine on both templates i.e. glucosylated or unglucosylated DNA template. C) Standard curve for 5-hmC estimates from qPCR. A 200 bp DNA fragment containing one 5-hmC-modified *Msp*I site ($C^{hm}CGG$) was spiked in different amounts into a quantity of unmodified DNA of the same sequence (x-axis). The total DNA was subjected to BGT-glucosylation and subsequent treatment *Msp*I followed by qPCR. The threshold cycle (Ct) values of the corresponding DNA mixtures are shown as inset.

3.3.2. Selection of microarray normalization algorithm

We first investigated various methods of array preprocessing to identify the algorithm best suited to analyze DNA modification data on tiling arrays. We considered quantile normalization and two variants of probe-sequence based normalization. Quantile normalization, a conventional choice, results in every microarray having the same overall intensity distribution,

an assumption that may be invalid in cases where microarrays represent different tissues and interrogate modifications that may vary several-fold in magnitude among them,[209] for instance, 5-hmC is higher in brain than in other tissues.[77,78,272] Moreover, it does not explicitly correct for probe sequence-based affinity bias, a known issue in tiling arrays. We considered MAT (model-based analysis of tiling arrays)[273] and an alternative sequence-based normalization scheme with fewer parameters (the "Potter" algorithm).[254] We then correlated single-probe intensities normalized using each algorithm with 11 arbitrary selected loci on which we performed qPCR (Section 2.13, Table 2.5). The Potter algorithm showed the highest correlation with qPCR estimates. It was originally unclear if targets analyzed at the single-probe level had a smaller measurement bias than those analyzed by averaging probe intensities in a window surrounding the target. We therefore correlated digestion efficacies from qPCR experiments with microarray intensities measured at the single-probe level, and using rectangular or distance-weighted windows. Both types of windows were tested at longer (~340 bp, microarray amplicon size) and shorter (~100 bp, average size of qPCR amplicon) lengths. Single-probe intensities showed the strongest correlation with qPCR estimates (Table 3.2, Figure 3.17A). Windowed probe averages in glucosylated samples had dramatically lower correlations with qPCR estimates, relative to single probe measurements (Correlations: Single probe = 0.52, 100 bp rectangular window = 0.03, 100 bp distance-weighted window = 0.17). We concluded that single-probe estimates provided the best balance between bias and precision for these data, and analyzed our data at the single probe level.

Despite the increased variance in single probe estimates, biological variability across samples significantly exceeded variability in technical replicates. gDNA from two human brain samples was used to create two sets of technical replicates. Each DNA sample was split six ways, and six technical replicates were generated for *Msp*I-treated gDNA (*Msp*I-gDNA). These technical replicates were compared to six biological replicates, using *Msp*I-treated gDNA from six individual human brain samples (samples randomly chosen from full set of 28 used in the study). DNA was hybridized onto Affymetrix 2.0R human whole-genome tiling arrays (Array E: chr 5, 7, 16), generating a total of 24 arrays. Arrays were normalized using the selected Potter algorithm and probes were extracted for chromosome 5 (27,546 probes). For each of the three sets (two technical replicate sets, and one set of biological replicates), we computed the sample range of individual probe intensities. The probe-wise range in technical replicates (presumably owing to technical variation) was subtracted from that in biological replicates (Figure 3.17B), and the shift in range was tested using a one-sample t-test ($\alpha = 0.05$)

Average probe intensities had the relative magnitude expected from the three treatments: *Msp*I-digested non-glucosylated gDNA had the lowest intensity (mean ± SD = -1.45 ± 1.00; 134,521 target probes), followed by *Msp*I-digested glucosylated gDNA (-1.35 ± 1.00); *Hpa*II-digested non-glucosylated gDNA had the highest intensity (-0.81 ± 1.03). Negative values reflect the digestion (underrepresentation) of target sites relative to the baseline of undigested sequences.

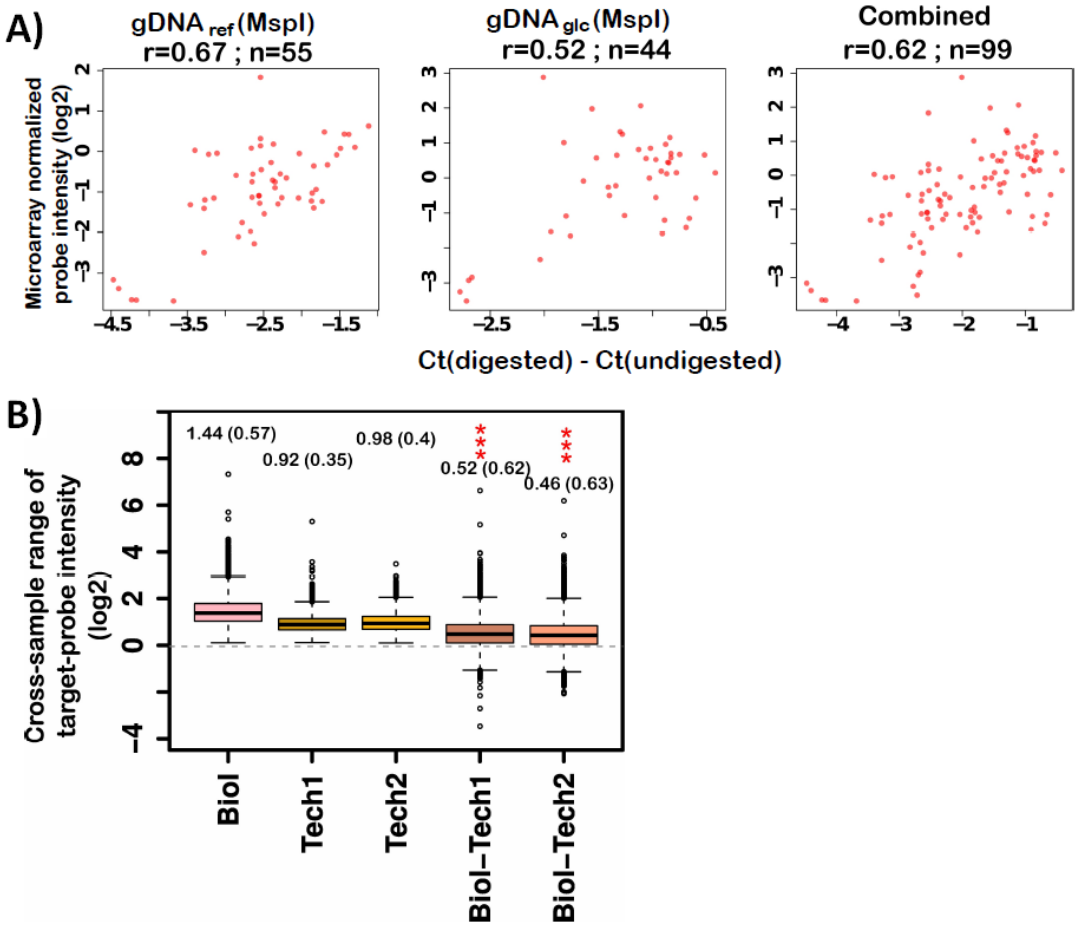

Figure 3.17. Microarray-based validation of 5-hmC assay.
A) Correlation of digestion efficacy as measured by quantitative PCR (qPCR) and by microarray. Microarray single probe intensity (y-axis) is plotted with qPCR measures (Ct value; x-axis) at 11 arbitrarily-selected loci (Table 2.5). Loci for qPCR have the property that the target site (C<u>C</u>GG) lay directly on a microarray probe. For each locus, DNA from 4-5 individuals was separately qPCR-amplified. Shown are correlations for (left) changes in unglucosylated gDNA following *Msp*I digestion (gDNA (*Msp*I)), (middle) changes in glucosylated gDNA following *Msp*I digestion (glc-gDNA (*Msp*I)), and (right) data from both conditions combined. The Ct has an inverse relationship with the amount of DNA fragment at the start of qPCR; i.e. a greater Ct value reflects a lower starting template to be PCR-amplified. Each dot shows individual-level (not sample-averaged) data; n denotes number of data points and r is the correlation coefficient. B) Microarray analysis results in biological variability that exceeds technical variability. Each boxplot shows the distribution of the range of target-probe intensities. Data is shown for *Msp*I-treated unglucosylated gDNA. Six biological replicates ("Biol", pink) were compared to each of two technical replicates ("Tech1" and "Tech2"; brown, orange). Each dot measures the cross-sample range (max - min) intensity for target probes on human chromosome 5 (27,546 probes). Mean (sd)

shown above each boxplot. The range of probe intensities is greater for biological replicates ("Biol - Tech1" (dark salmon); "Biol - Tech2" (light salmon)); one-sample t-test in both cases results in p-values $< 10^{-16}$.

Type of array measurement	gDNA (*Msp*I)	Glu-gDNA (*Msp*I)	Combined
# data points	55	44	99
Single probe	0.67	0.52	0.62
Probe-averaged window (340 bp)	0.02	– 0.29	0.09
Weighted window (340 bp)	0.20	– 0.15	0.27
Probe-averaged window (100bp)	0.42	0.03	0.39
Weighted window (100bp)	0.51	0.17	0.48

Table 3.2. Correlation of sequence quantity at 11 loci, as measured by quantitative PCR and by microarrays.
Correlation deteriorates dramatically for glucosylated DNA, when window-based averaging is used in arrays. The reason for this drop in correlation is not understood. Based on these results we decided to analyze arrays at the single probe level, without window-based averaging.

3.3.3. Characterization of 5-mC and 5-hmC in adult mouse tissues

Using thin layer chromatography (TLC) and intensities from microarray probes, we verified that 5-hmC levels were the highest in mouse brain gDNA, compared to liver, kidney, pancreas and heart (Table 3.3, Figure 3.18). This finding is consistent with previous reports.[77,78] To investigate the origin of increased levels of 5-hmC in the brain, we identified genes and intergenic regions with significantly different 5-hmC in the mouse brain compared to other tissues. Of 134,521 probes that overlapped the non-repetitive genome (six chromosomes), 73,461 overlapped exactly one gene (defined by Mouse Genome Informatics (MGI) symbol; 2–238 probes per gene, median = 16.8). Out of 4,357 genes tested, 730 had different 5-hmC levels in the brain relative to the other tissues (repeated-measures analysis of variance (ANOVA), $p < 10^{-2}$, false discovery rate (FDR) $Q < 0.05$). All but one gene had higher 5-hmC, and lower 5-mC, levels in the brain. Even within the brain, these genes contained above average 5-hmC levels ($p < 10^{-4}$, bootstrapping (no replacement), median 5-hmC of randomly-sampled genes (R = 10,000)). Separately, we identified 83 differential intergenic probes that also had significantly higher 5-hmC levels in the brain (probe-wise linear regression, total of 60,721 intergenic probes, $Q < 0.05$).

We further explored possible associations of genic 5-hmC and steady-state mRNA levels in the mouse dataset. We related 5-hmC and 5-mC intensity levels from our microarrays to tissue-specific mRNA levels from a previously-published dataset.[259] In three out of five investigated tissues, 5-hmC showed a significant increase in genes with higher transcription (Figure 3.19).

This trend was also observed in the brain, but did not reach statistical significance (p = 0.17, linear regression); the same was true for the 730 genes identified to have higher 5-hmC levels (p = 0.06, bootstrapping randomly-sampled genes (R=10,000)). In contrast, genic levels of 5-mC significantly decreased with increasing mRNA levels, in all five tissues (Figure 3.19).

	5-hmC % (relative to 5-mC)	
Sample	**Mean**	**SD**
Human		
Brain	18.6	2.6
Mouse		
Brain	13.6	0.1
Liver	8.8	0.3
Kidney	6.0	0.1
Heart	5.1	0.4
Pancreas	2.8	0.8

Table 3.3. Thin layer chromatography quantification of 5-hmC.
5-hmC was estimated in 20 human *post-mortem* brain gDNA samples (age range 34 - 85 years). Mouse: Brain, heart, kidney, and liver samples were obtained from a 24-month animal, and the pancreas sample from an 8-week old mouse (male). Standard deviation (SD) from 3 technical replicates per sample.

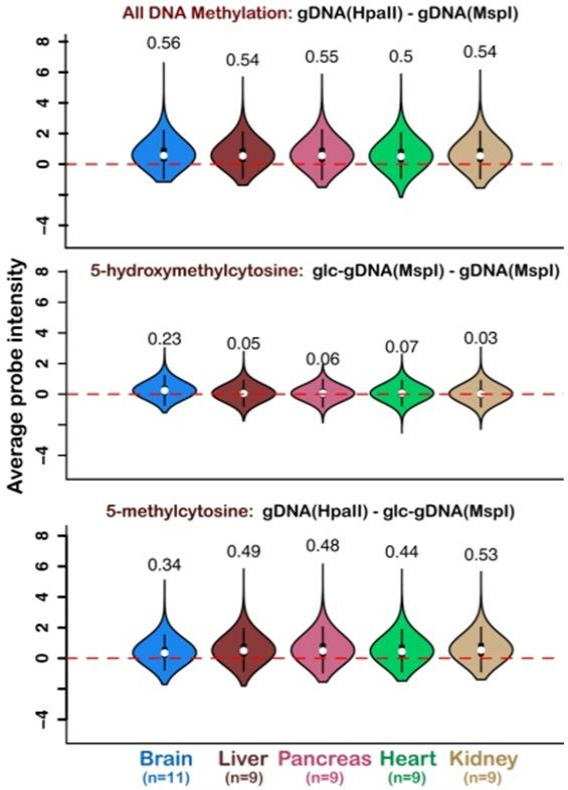

Figure 3.18. Distribution of single-probe intensities in different mouse tissues.
Each violin shows the distribution of normalized intensities for target probes (n=134,521 across six chromosomes) (median shown above violin). Each probe is averaged across all samples in the tissue group. Relative to unglucosylated gDNA treated with *Msp*I, the differential increase in (Top) *Hpa*II-

treated unglucosylated DNA estimates all DNA modifications, while (Middle) *Msp*I-treated glucosylated DNA estimates 5-hmC. The difference of the first two quantities estimates 5-mC (Bottom). 5-hmC in the brain (blue, middle panel) is the highest of all tissue groups.

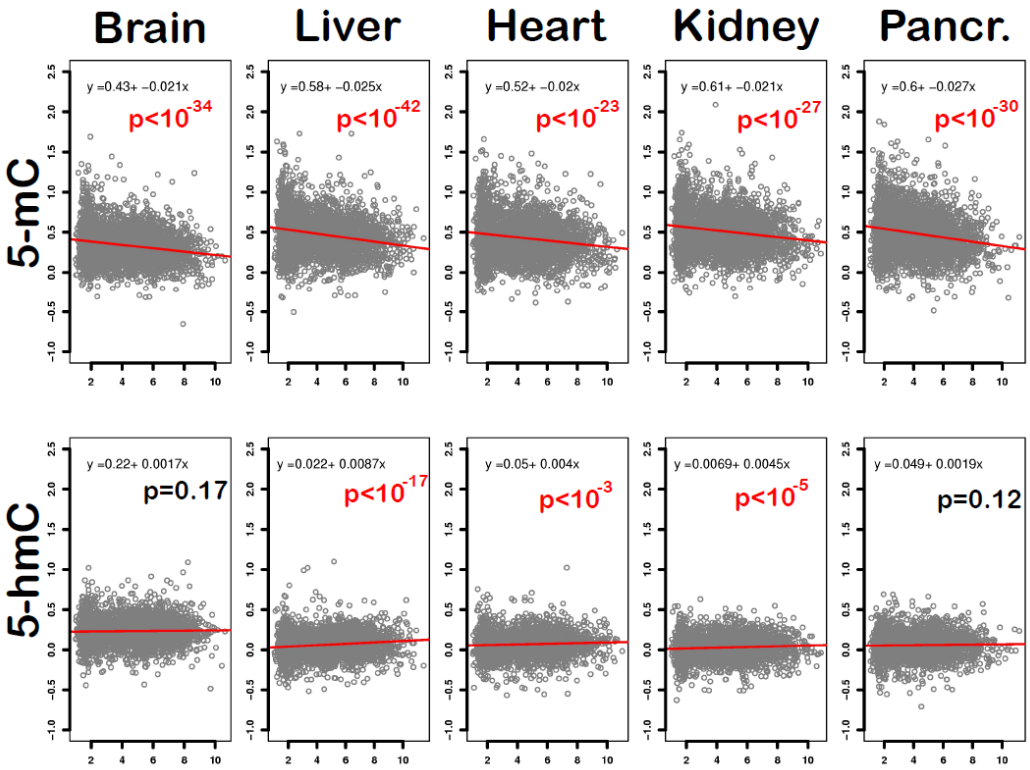

Figure 3.19. Linear regression of steady-state mRNA levels with mean genic intensity of DNA modifications.
Plots show gene-averaged (mean) mRNA levels (x-axis) against averaged probe intensities for corresponding genes (y-axis); genes defined by RefSeq ID (top: 5-mC, bottom: 5-hmC). Regression line shown in red, P-values are for slope ($\alpha = 0.005$), with significant p-values shown in red. The inverse relationship of genic 5-mC and gene expression levels were consistently found in all tissues investigated. The relationship of 5-hmC and transcription levels was significant only in some non-neuronal tissues, although the slight upward trend is visible in all cases.

3.3.4. Synaptic protein genes enriched for 5-hmC in mouse brain

To determine if the 730 genes obtained above had unifying cellular functions, we performed a functional overrepresentation analysis (ORA) (DAVID, background set of 4,357 genes tested). We observed that these genes were statistically overrepresented in 8 Gene Ontology (GO) terms (total 156 candidate terms), which in turn were associated with synapse function (Table 3.4). The top overrepresented terms were "cell adhesion" ($Q = 1.3\text{x}10^{-6}$), "plasma membrane" ($Q = 8.7\text{x}10^{-5}$), and "synapse" ($Q = 3.5\text{x}10^{-4}$). These genes also had functional annotation clusters pertaining to ion channel activity, Rho GTPase signaling, and neuronal development (Appendix 5). Interestingly, some of these 5-hmC rich genes showed transcript enrichment or specificity for non-neuronal brain cells. We identified the cell-type specificity of

the 730 brain-enriched genes using a list from a previously published transcriptomic dataset of individual brain cell populations.[258] Using this list, 25 of the brain-enriched genes (3 %) have enriched transcription in astrocytes, 21 (3 %) in oligodendrocytes, and 57 (8 %) in neurons. By this definition, a few genes were also specific to the cell type (astrocytes: *Rfx4*, *Gli3*; oligodendrocytes: *Elovl7*, *Cpm*; neurons: *Nts, Syt1*, *Nrg3, Trhde, Kcnc2*, *Clstn2*).

We tested to see if the increase in brain 5-hmC observed in the 730 identified genes generalized to the entire functional class of synapse-related genes. We compared average 5-hmC in all genes mapped to the synapse-relevant GO terms to that in genes outside these categories.[274] Genes within each tested category ("synapse" and "synapse part", "cell adhesion", "plasma membrane") had significantly higher 5-hmC levels (e.g. 5-hmC for "synapse", N = 3,258 probes) compared to genes in other categories (N = 70,203; p = 3.2×10^{-20}, two-tailed Wilcoxon-Mann-Whitney (WMW) test; Figure 3.20A). This effect persisted after controlling for GC content of the probe sequence, a parameter that could artificially influence probe intensity (Figure 3.20B).

Term	Count	%	*P*	*Q* (BH)
GO:0022610, biological adhesion	53	7.3	6.3×10^{-10}	1.3×10^{-6}
GO:0007155, cell adhesion	53	7.3	6.3×10^{-10}	1.3×10^{-6}
GO:0005886, plasma membrane	156	21.5	2.4×10^{-7}	8.7×10^{-5}
GO:0045202, synapse	33	4.5	1.9×10^{-6}	3.5×10^{-4}
GO:0030054, cell junction	41	5.6	1.2×10^{-5}	1.4×10^{-3}
GO:0044459, plasma membrane part	90	12.4	5.6×10^{-5}	5.0×10^{-3}
GO:0044456, synapse part	23	3.2	1.7×10^{-4}	1.2×10^{-2}
GO:0016337, cell-cell adhesion	22	3.0	3.7×10^{-5}	3.9×10^{-2}

Table 3.4. Statistically overrepresented Gene Ontology categories in genes enriched for 5-hmC in the mouse brain, relative to that in other tissues.

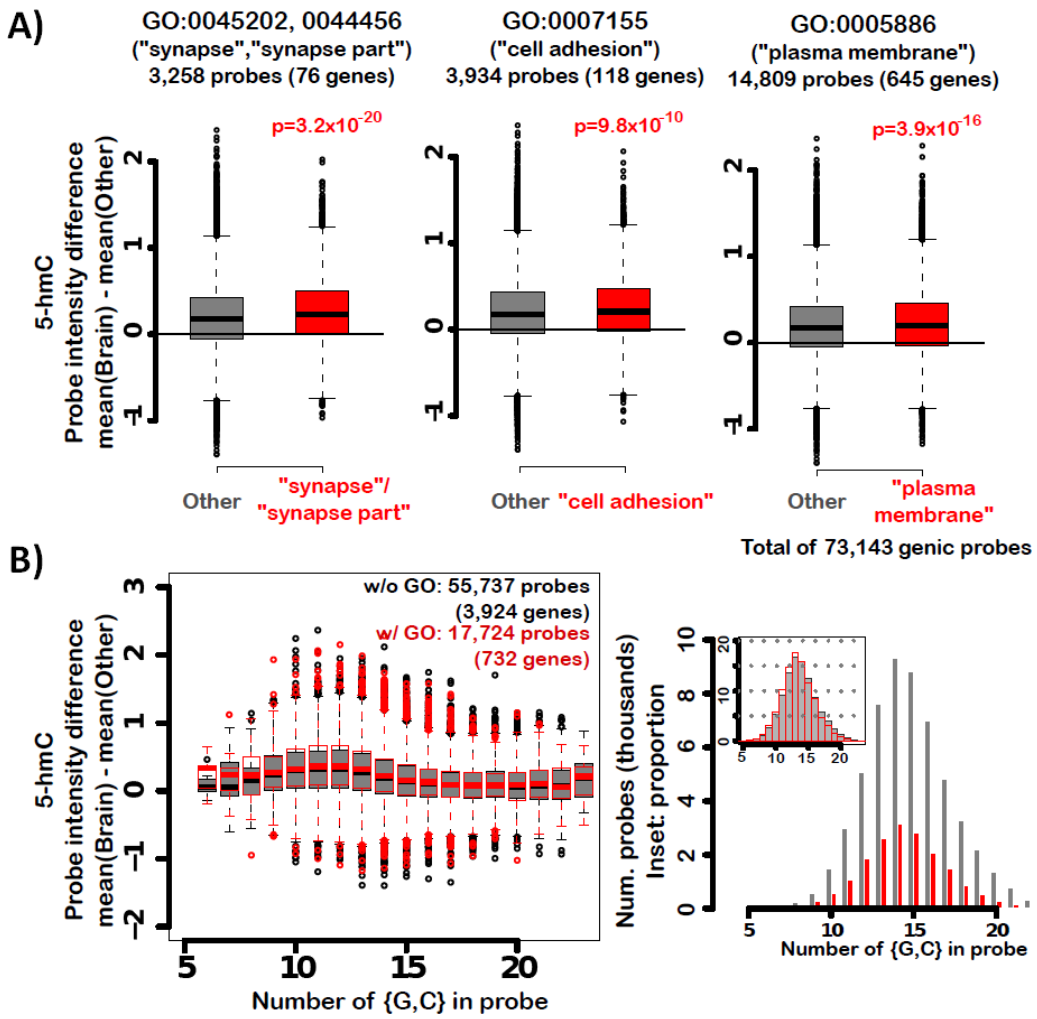

Figure 3.20. 5-hmC in the adult mouse brain is higher in genes mapped to synapse-related categories, compared to that in genes outside these categories.
A) Probes in genes mapped to each Gene Ontology (GO) category (red) had greater cross-tissue differences (Brain-Other) than those in other genes (gray). The GO categories tested were the top three categories overrepresented in brain 5-hmC rich genes (Table 3.4); this result generalizes the observation in enriched genes to all genes in these categories. Each dot measures the difference in probe intensity between brain samples and samples from other tissues; probes were not averaged within genes. (P-values from two-tailed WMW test, $\alpha = 0.016$). B) Probe-level differences persist even after probe stratification by GC content. This panel shows probes combined for all three GO terms tested in (a) (red), compared to other probes (gray). (Left): Increase in 5-hmC levels is evident, particularly in strata with most probes ($9 \leq GC \leq 16$). This increase is more pronounced for individual GO categories (not shown). (Right): Number of probes in each GC-stratum, (inset: probe proportions).

3.3.5. Enrichment of 5-hmC in synapse genes extends to human brain

We extended the comparison of 5-hmC enrichment in synapse-related genes to human brain by assaying gDNA from 28 human *post-mortem* brains. Human gDNA samples were processed the same way as the mouse gDNA samples (Section 2.8), and interrogated on Affymetrix 2.0R human whole genome tiling microarrays E and F (Table 2.4 for probe counts). Similar to the mouse brain, the human frontal cortex had higher 5-hmC levels within genes that

mapped to the GO terms “synapse”, “synapse part”, “cell adhesion” and “plasma membrane” (Figure 3.21).

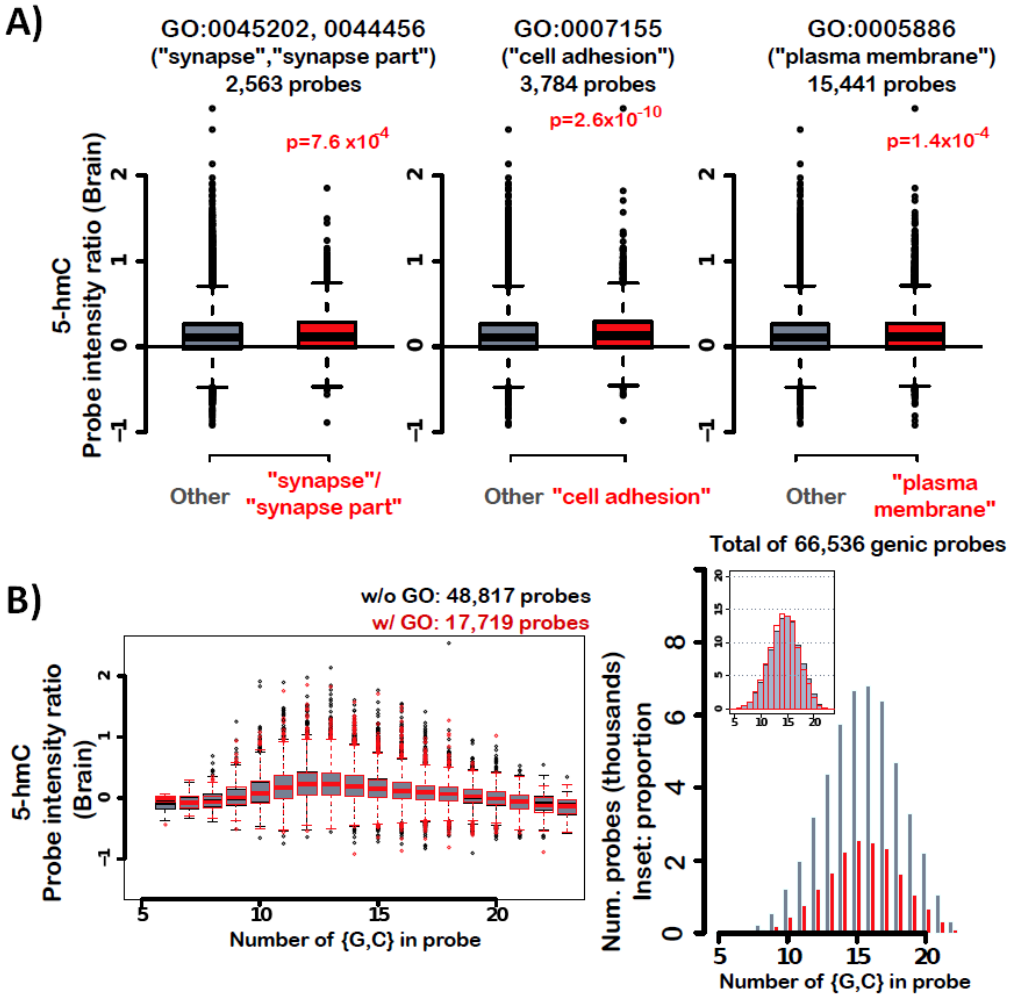

Figure 3.21. 5-hmC in the adult human brain is higher in genes mapped to synapse-related categories, compared to that in genes outside these categories.
A) Probes in genes mapped to each GO category (red) had higher density of 5-hmC than those in other genes (gray). The GO categories tested were the same as those tested in the mouse (Table 3.4). Each dot measures probe-level 5-hmC (sample-averaged). Probes were not averaged within genes. B) Probe-level intensities with probe stratification by GC content. This panel shows probes combined for all three GO terms (red, "w/ GO"), compared to other probes (gray, "w/o GO"). (Left): 5-hmC intensity (Right): Number of probes in each GC-stratum (Inset: probe proportions).

3.3.6. 5-hmC marks the exon-intron boundary in the human brain

Several studies performed before the re-discovery of 5-hmC have noted a change in the density of modified cytosines at the exon-intron boundary of genes, with the density being higher on the exonic side.[94,275] These studies, however, used bisulfite sequencing and therefore did not differentiate between 5-mC and 5-hmC. As our restriction enzyme-based assay provided single

nucleotide resolution mapping of DNA modifications, we compared densities of each modification on either side of the exon-intron boundary (linear mixed-effects model, Section 2.16.9). Distances to the boundary were measured from the external 'C' of the target sequence (CCGG). DNA modification differences at exon-intron boundaries are reported for two regions: one immediately adjacent to the boundary (cumulative for the first 5 nucleotides, d = 5) and one that captures the general peri-boundary trend (cumulative for the first 20 nucleotides, d = 20). Relevant parameters (e.g. probe count) and statistics for all datasets are in Table 3.5 and overall probe intensities for various DNA modifications in all datasets are in Table 3.6.

Levels of all DNA modifications changed at the exon-intron boundary in human frontal cortex samples (N = 28 brain samples; Affymetrix tiling microarrays E and F, covering six chromosomes). Consistent with previous findings, we found higher densities of all modifications in exons relative to introns (d = 5, $p = 2.8 \times 10^{-7}$; and d = 20, $p = 2.8 \times 10^{-40}$; p-values from linear mixed-effects model unless otherwise stated; mean differences are not reported in the text as they are in arbitrary units of microarray intensity, and are not by themselves biologically meaningful). Most of this cross-boundary change in modification density was attributable to 5-hmC (d = 5, $p = 2.3 \times 10^{-6}$; d = 20, $p = 2.8 \times 10^{-20}$; Table 3.5, Figure 3.22A). In contrast, 5-mC showed no changes closer to the boundary (d = 5, p = 0.57) and relatively smaller exonic increases than 5-hmC for longer peri-boundary distances (d = 20, $p = 9.1 \times 10^{-6}$; Table 3.5, Figure 3.22A). The change in 5-hmC, relative to 5-mC, was most evident in the first 5–10 bp from the boundary. For longer distances (up to 50 bp tested), both modifications showed robust increases in exons relative to introns (Figure 3.22A).

We also mapped DNA modifications in brain samples from patients diagnosed with schizophrenia and bipolar disorder, collectively termed "major psychosis" (N = 54 brain samples (frontal cortex); Affymetrix tiling microarray E). Consistent with the findings in the control brain set reported above, gDNA from brains of psychosis patients also showed a predominant change in 5-hmC at the exon-intron boundary, at both distances (Table 3.5, Figure 3.22B).

To ascertain if the cross-boundary change was unique to brain tissue in humans, we analyzed human liver samples (N = 13) in parallel with age- and sex-matched frontal cortex samples (N = 12). Here again, brain gDNA samples showed a larger change in 5-hmC than in 5-mC (Table 3.5, Figure 3.22C). In contrast, liver gDNA samples showed a predominant change in 5-mC, both at the boundary (d = 5, $p = 4.8 \times 10^{-6}$) and longer distances (d = 20, $p = 4.8 \times 10^{-12}$; Table 3.5, Figure 3.22D).

				5-hydroxymethylcytosine (5-hmC)			5-methylcytosine (5-mC)		
Biological context	***n*** [1]	**# E** [2]	**# I** [2]	**E – I** [3]	**95%CI**	***P*** [4]	**E – I** [3]	**95%CI**	***P*** [4]
d = 5 bp from exon-intron boundary									
Human									
Brain BA10, Controls	28	762	43	0.10	[0.06, 0.15]	2.3×10^{-6}	0.02	[-0.03, 0.06]	0.57
Brain BA10, Psychosis	54	239	26	0.11	[0.06, 0.14]	6.8×10^{-9}	- 0.01	[-0.05, 0.05]	0.06
Brain BA10, Controls (exp2)	12	339	26	0.06	[-0.04, 0.17]	1.4×10^{-2}	0.13	[0.05, 0.24]	2.60×10^{-4}
Liver (exp2)	13	339	26	0.02	[-0.05, 0.12]	0.30	0.16	[0.09, 0.26]	4.80×10^{-6}
Mouse									
Frontal cortex	15	393	11	0.00	[-0.12, 0.18]	0.35	0.22	[0.14, 0.34]	9.20×10^{-4}
Rest of the brain	15	393	11	0.05	[-0.07, 0.22]	0.60	0.11	[-0.04, 0.10]	6.40×10^{-2}
Liver, Pancreas, Heart, Kidney	24	504	15	-0.01	[-0.13, 0.08]	0.25	0.11	[0.04, 0.24]	3.10×10^{-2}
Neuronal cell line (mouse)	18	285	10	0.07	[-0.04, 0.21]	0.72	0.12	[-0.05, 0.28]	2.80×10^{-2}
B-lymphocyte cell line (human)	6	400	13						
Vehicle- treated (0 µM SAHA)	6	400	13	-0.03	[-0.11, 0.06]	0.95	0.21	[0.02, 0.45]	0.40
$1/10^{th}$ of IC-10 (0.01 µM SAHA)	6	400	13	0.07	[-0.13, 0.24]	0.35	0.10	[0.02, 0.45]	0.27
$1/5^{th}$ of IC-10 (0.02 µM SAHA)	6	400	13	0.09	[-0.11, 0.20]	0.97	0.05	[-0.26, 0.24]	0.95
IC-10 (0.10 µM SAHA)	6	400	13	-0.07	[-0.26, 0.08]	0.44	0.06	[-0.27, 0.26]	0.84
d= 20 bp from exon-intron boundary									
Human									
Brain BA10, Controls	28	2,445	811	0.05	[0.03, 0.06]	2.80×10^{-20}	0.02	[0.01, 0.04]	9.10×10^{-6}
Brain BA10, Psychosis	54	1,148	424	0.04	[0.03, 0.06]	9.20×10^{-20}	0.01	[0.00, 0.03]	2.60×10^{-2}
Brain BA10, Controls (exp2)	12	1,148	424	0.07	[0.05, 0.08]	1.50×10^{-12}	0.05	[0.03, 0.07]	4.60×10^{-10}
Liver (exp2)	13	1,148	424	0.01	[-0.01, 0.03]	3.10×10^{-2}	0.05	[0.03, 0.07]	4.80×10^{-12}

				5-hydroxymethylcytosine (5-hmC)			5-methylcytosine (5-mC)		
Biological context	***n*** [1]	**# E** [2]	**# I** [2]	**E – I** [3]	**95%CI**	***P*** [4]	**E – I** [3]	**95%CI**	***P*** [4]
d=20 bp from the exon-intron boundary									
Mouse									
Frontal cortex	15	1,244	309	0.06	[0.04, 0.09]	7.50×10^{-6}	-0.01	[-0.03, 0.01]	0.84
Rest of the brain	15	1,244	309	0.06	[0.04, 0.09]	3.00×10^{-8}	0.01	[-0.02, 0.04]	0.33
Liver, Pancreas, Heart, Kidney	24	1,613	422	0.03	[0.01, 0.04]	5.70×10^{-3}	0.08	[0.06, 0.10]	3.40×10^{-19}
Neuronal cell line (mouse)	18	857	203	0.03	[-0.00, 0.06]	4.50×10^{-2}	0.13	[0.10, 0.17]	1.60×10^{-17}
B-lymphocyte cell line (human)									
Vehicle- treated (0 µM SAHA)	6	1,301	418	0.02	[-0.01, 0.05]	0.275	0.08	[0.01, 0.12]	2.70×10^{-5}
$1/10^{th}$ of IC-10 (0.01 µM SAHA)	6	1,301	418	0.06	[-0.01, 0.08]	0.0144	0.09	[0.05, 0.15]	6.10×10^{-8}
$1/5^{th}$ of IC-10 (0.02 µM SAHA)	6	1,301	418	0.05	[0.02, 0.10]	8.8×10^{-4}	0.05	[0.02, 0.11]	4.00×10^{-2}
IC-10 (0.10 µM SAHA)	6	1,301	418	0.05	[0.02, 0.09]	3.6×10^{-4}	0.06	[0.00, 0.09]	7.20×10^{-4}

Table 3.5. Statistics on DNA modification changes at cumulative distances (d = 5 and d = 20) from the exon-intron boundary.
[1] Number of biological samples; [2] Number of measurements for boundary side = (# probes * *n*); [3] Median increase in exons, relative to introns (value > 0 indicate higher exonic levels); [4] P-values from linear mixed-effects model.

			5-mC+5-hmC		5-hmC		5-mC	
	Num. arrays	**Probes per array**	**Median**	**IQR**	**Median**	**IQR**	**Median**	**IQR**
Human								
Controls	56	136,477	0.26	1.02	0.11	0.99	0.14	0.97
Psychosis	54	69,252	0.34	1.00	0.11	1.01	0.23	0.97
Exp2, Brain	12	69,252	0.45	0.97	0.23	0.86	0.23	0.81
Exp2, Liver	13	69,252	0.40	0.98	0.04	0.84	0.37	0.90
Mouse								
Frontal cortex	15	69,052	0.56	1.02	0.19	0.91	0.37	0.90
Brain, non-frontal cortex	15	69,052	0.59	1.02	0.18	0.90	0.41	0.93
Non-brain organs	36	130,314	0.57	1.10	0.05	0.89	0.51	1.05
Cell lines								
Human B-lymphocyte	24	76,102	0.36	0.99	-0.07	0.88	0.44	0.98
Mouse neuronal	18	46,892	0.47	1.05	0.02	0.93	0.47	1.06

Table 3.6. Median probe intensities for DNA modifications for all datasets used in the exon-intron boundary analysis.
Probes were pooled across all samples and arrays without aggregation before median and interquantile range (IQR) were computed. Consistent with previous literature, 5-hmC intensities are higher in tissues sampled from the brain, relative to those sampled from other tissues (e.g. human liver, mouse non-brain organs). Negligible 5-hmC was detected in the two cell lines tested.

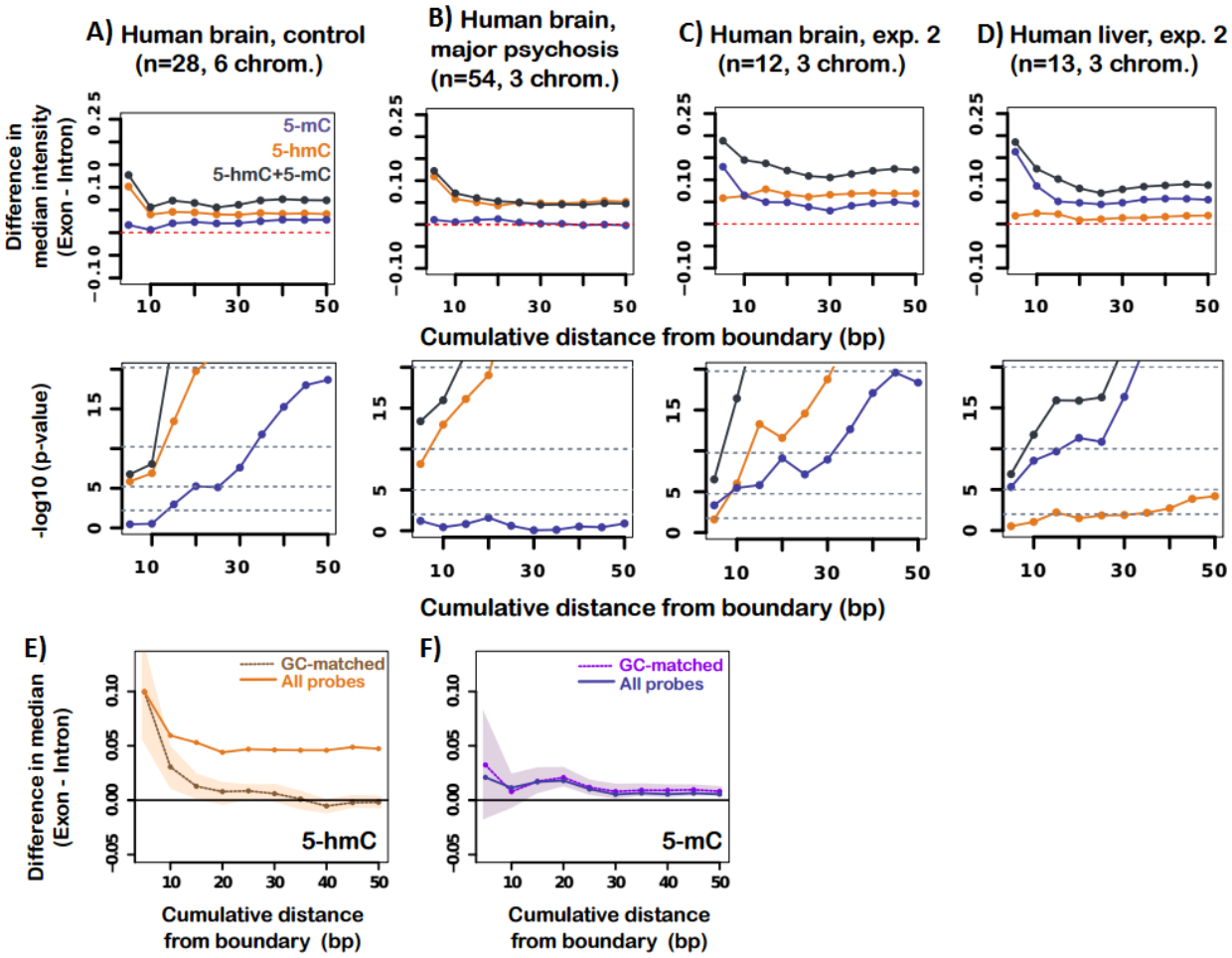

Figure 3.22. Exonic increase in DNA modifications in human tissues.
A-D show cross-boundary changes in DNA modifications in human tissues, for various cumulative distances (d = 5-50 bp). Data are shown for A) human brains without diagnosis of mental illness (N = 28; 6 chromosomes), B) human brains from individuals diagnosed with major psychosis (N = 54, 3 chromosomes), and for an independent experiment on age- and sex-matched C) brain (N = 12, 3 chromosomes) and D) liver samples (N = 13, 3 chromosomes). In each case, the top panel shows median exonic increase in DNA modifications at various cumulative distances from the exon-intron boundary, and the bottom panel shows corresponding p-values from statistical comparison of exonic and intronic probe intensities (linear mixed-effects model). In brain samples, exonic increase in all DNA modifications (black) is predominantly mirrored by changes in 5-hmC (orange), and to a lesser extent in 5-mC (purple); in the liver, this pattern is reversed. E-F show exon-intron peri-boundary differential after probes on either side of the boundary are matched for GC content, at various cumulative distances from the boundary (100 iterations of matching; trend lines show median, shaded areas show the range between the 5th and 95th percentile of differences). E) Following GC-matching, exonic increase in 5-hmC levels are notable at d = 5 and persist up to 20 bp in the peri-boundary region. F) The relatively modest change in 5-mC persists after GC-matching. At d = 5, zero lies within the range of GC-matched values. Following GC-matching, 5-hmC and 5-mC values are similar for peri-boundary distances greater than 30bp. It is unclear at present whether this similarity is due to loss of statistical power from GC-matching.

3.3.7. 5-mC and 5-hmC boundary changes validated in human brain

We validated the findings in the human frontal cortex using three independent methods. First, to estimate the impact of GC differences in exons and introns, we compared probes representing introns to those representing exons, after matching for GC/AT content and distance from boundary (100 iterations to sample greater number of exonic probes). The overall trends of 5-hmC and 5-mC for matched probes were similar to those for unmatched probes (Figure 3.22E and F). 5-hmC cross-boundary differences were significant ($p < 0.05$) for distances starting at d = 5 up to d = 30, while 5-mC differences became significant only at d = 15.

Second, we assayed 5-hmC by SMS (Helicos Biosciences Corp., Cambridge MA), which does not require PCR and microarray hybridization, thus eliminating artifacts due to DNA sequence composition.[276] gDNA from one frontal cortex sample was *Msp*I-digested with or without prior glucosylation, followed by polyA end-labeling, and sequencing (Appendix 6). Direct sequencing confirmed a sharp intronic decrease in 5-hmC at the exon-intron boundary (one-tailed WMW test, median exonic increase in 5-hmC = + 4.6% for d = 20, $p = 9.2 \times 10^{-3}$; Figure 3.23). We were unable to accurately measure exon-intron differences for d = 5 due to a small number of reads 4 bp into the intron. However, the exonic increase in 5-hmC is evident for cumulative distances larger than 10 bp.

The third approach compared DNA enriched in 5-mC to a fraction depleted in 5-mC, separated using MethylMiner Methylated DNA Enrichment kit (Invitrogen). The kit employs a methyl-binding domain (MBD) to enrich for genomic 5-mC fraction. Each fraction was then subjected to the treatment as in Figure 2.2. We expected that if 5-mC levels in the brain actually changed across the exon-intron boundary, then we would detect a greater cross-boundary change in the 5-mC rich fraction relative to the depleted fraction. We did not observe such a change at the exon-intron boundary; in fact, intronic levels of 5-mC were slightly higher (d = 20, $p = 0.05$; Figure 3.24).

The above three control experiments collectively show that in the human frontal cortex, the change in 5-hmC at the exon-intron boundary exceeds that of 5-mC.

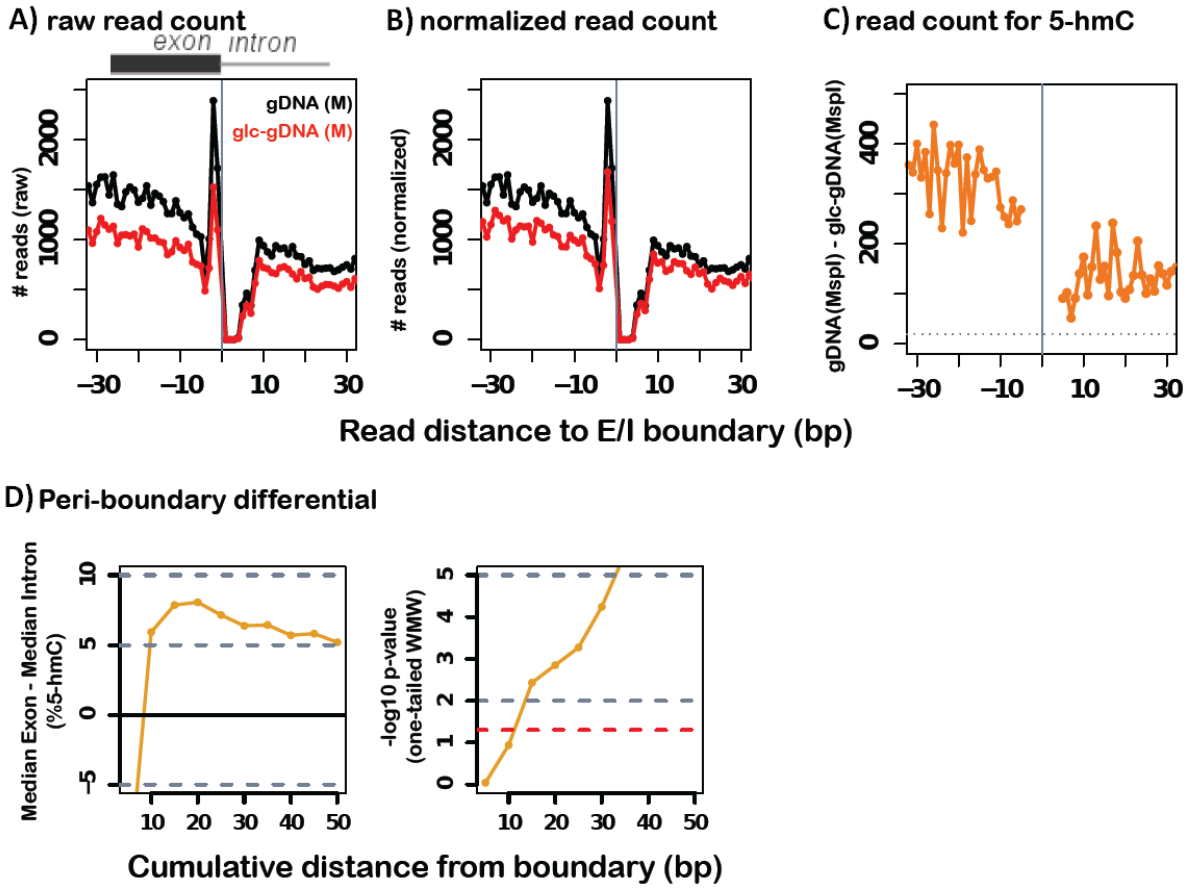

Figure 3.23. 5-hmC changes measured at exon-intron boundary using single molecule sequencing (SMS).

One brain sample (cortex Brodmann Area 10; female, age at death 49 years, no diagnosis of brain disease) was analyzed (3 replicates) for DNA modifications by single molecule sequencing. 5-hmC was estimated as the percent difference in read count of non-glucosylated DNA and that in glucosylated DNA, following *Msp*I restriction digestion. In all panels, the x-axis shows the distance from the second cytosine in the target site (CCGG) of a read generated by a CCGG sequence, relative to an exon-intron boundary. The y-axis shows: A) raw read count, B) read count normalized by reads in unglucosylated channel, and C) the difference in reads from restriction-digested DNA with and without glucosylation. D) shows exonic increase in % 5-hmC at various cumulative distances from the boundary. The x-axis is the cumulative distance (in bp) from the second cytosine of a target read to an exon-intron boundary. Left: Percent difference in reads obtained with and without glucose protection of *Msp*I sites. Right: P-values from comparison of exonic and intronic % 5-hmC at distances corresponding to the left graph (one-tailed Wilcoxon Mann Whitney test, N = Distance (in bp) from the boundary).

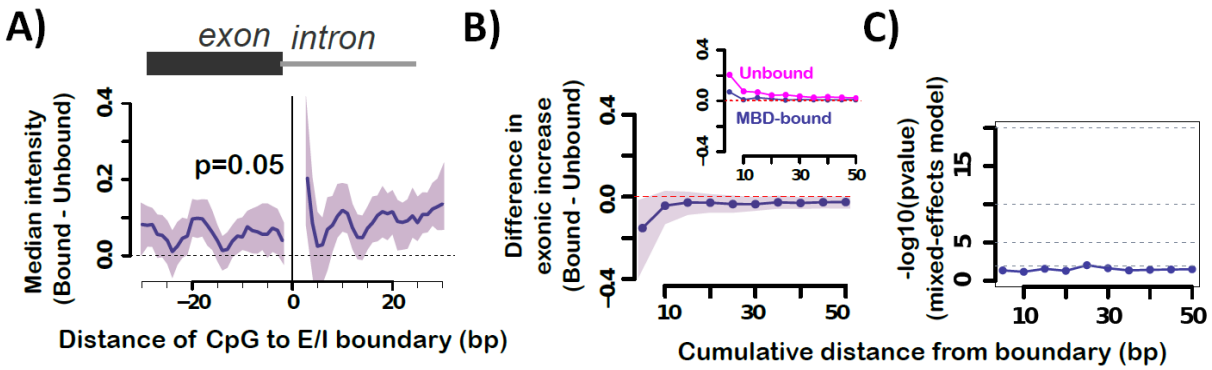

Figure 3.24. Validation of cross-boundary change in brain samples as being due to 5-hmC. Exon-intron boundary comparisons in DNA enriched for 5-mC, relative to that depleted in 5-mC. Sample gDNA from one human brain (N = 6 technical replicates) was separated into a 5-mC rich fraction ('bound', Methyl Miner Kit (Invitrogen)) and the rest ('unbound'). Both fractions were then analyzed for 5-mC following glucosylation and restriction enzyme treatment (as in Figure 2.2) on human tiling array chip E (chr 5,7,16). A) Junction detail showing the relative increase in 5-mC in the MBD-bound fraction. Compared to the unbound fraction, the bound fraction shows an intronic increase at the boundary. Shaded regions show bootstrapped (R = 1,000) 95 % CI. B and C show changes in the bound fraction at various cumulative distances from the boundary (Inset: bound and unbound fractions). B) Relative intensity; dip at 5 bp indicates intronic increase in 5-mC at the boundary. C) Informal p-values of cross-boundary comparisons. Shading shows 95 % CI from separately bootstrapping exonic and intronic values prior to subtraction (R = 1,000).

3.3.8. DNA modification at exon-intron boundaries in mouse tissues

To determine if the change in 5-hmC at exon-intron boundaries is unique to the human frontal cortex, we examined DNA in multiple mouse organs, including the mouse brain. Mouse brain samples (8 weeks old male C57/BL6) were split into frontal cortex and the remainder (including cerebellum and brain stem) (N = 15 samples per group, Affymetrix mouse tiling microarray A). The frontal cortex was analyzed separately to match the brain region investigated in human samples. Consistent with patterns in human brain, at d = 20 from the exon-intron boundary, the change in levels of DNA modifications were mainly due to 5-hmC in both frontal cortex ($p = 7.5 \times 10^{-6}$, linear mixed-effects model) and in the rest of the brain ($p = 3.0 \times 10^{-8}$; Figure 3.25A and B, Table 3.5). In these contexts, the change in 5-hmC was not evident at the immediate boundary (e.g. p = 0.35 at d = 5 in frontal cortex), but rather was gradual over a region adjacent to the boundary. In contrast, 5-mC exhibited a more dramatic immediate cross-boundary change ($p = 9 \times 10^{-4}$ at d = 5; Figure 3.25A and B, Table 3.5).

For non-neural organs, we used the mouse dataset described previously in this work (Table 2.2; 36 samples from liver, pancreas, kidney, and heart, microarrays A and G). In agreement with the markings seen in the human liver, mouse organs of non-neural origin primarily had changes in cross-boundary densities of 5-mC (d = 5, p = 0.03), with a non-significant change in 5-hmC (d = 5, p = 0.25; Figure 3.25C, Table 3.5). Although both types of

DNA modifications show relatively higher exonic densities at the larger peri-boundary regions (d = 20), the magnitude of change in 5-mC ($p = 3.4x10^{-19}$) exceeded that of 5-hmC ($p = 5.7x10^{-3}$; Figure 3.25C, Table 3.5).

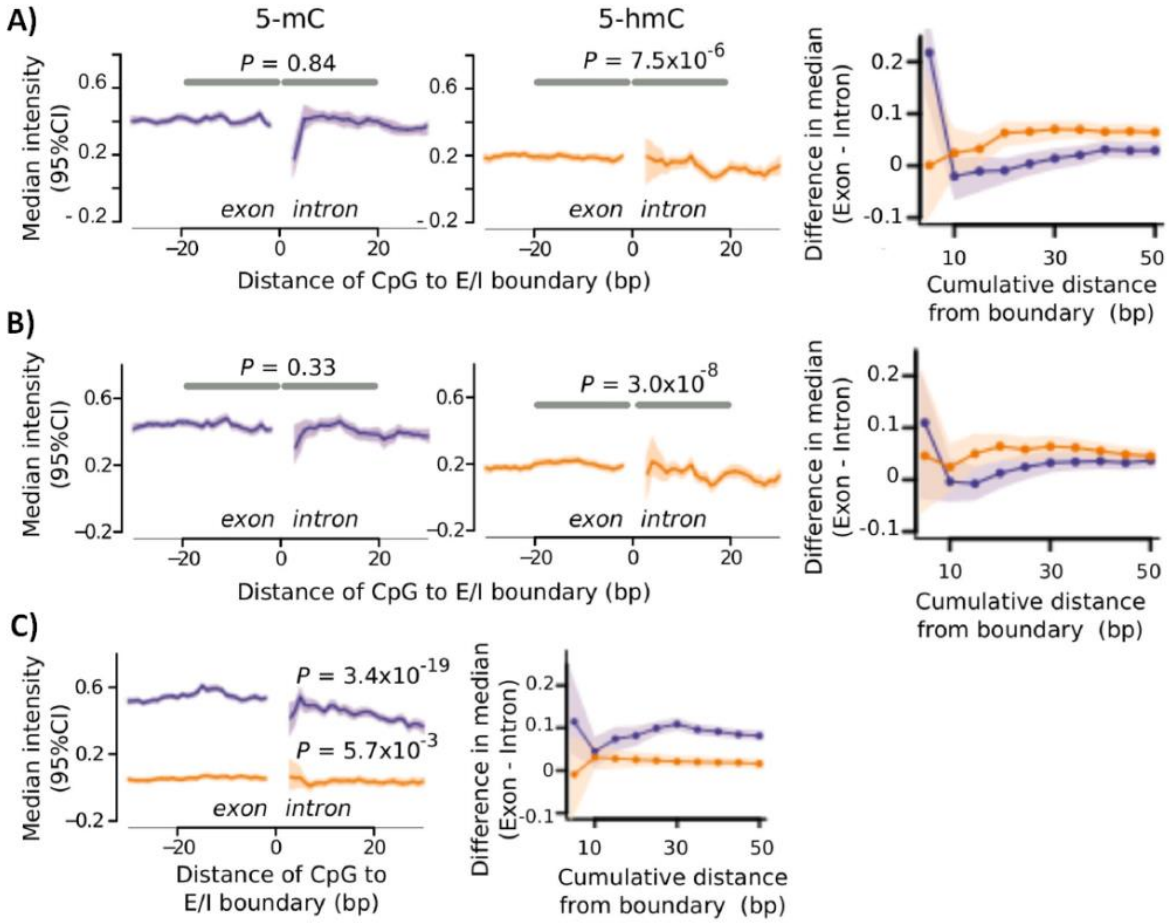

Figure 3.25. 5-hmC marks exon-intron boundaries in the mouse brain, but not in mouse organs of non-neural origin.
Peri-boundary modifications in DNA in the mouse brain: A) frontal cortex, and B) remainder (N = 15 animals per group, 3 chromosomes). Left, middle: Median intensities around the exon-intron boundary. Right: Relative exonic increase at various cumulative distances from the boundary. C) DNA modifications in non-neural mouse organs (N = 24 animals, 6 chromosomes; DNA from heart, liver, pancreas and kidney). Shading shows 95%CI from bootstrapping exonic and intronic intensities separately.

3.3.9. 5-mC and 5-hmC at the exon-intron boundary in human and mouse cell lines

We next assayed a murine neuronal cell line which had undetectable (< 1%) global levels of 5-hmC measured by TLC (mHypoA-2/24; N = 18 samples,). The microarray analysis, which is more sensitive to region-specific differences than TLC, detected a marginally significant change in 5-hmC at d = 20 ($p = 0.045$, linear mixed-effects model) and a highly significant difference in 5-mC ($p = 1.6x10^{-17}$; Figure 3.26, Table 3.5).

We also examined boundary differences in a B-lymphocyte dataset where cells were treated with suberoylanilide hydroxamic acid (SAHA). Histone deacetylase (HDAC) inhibitors like SAHA may promote DNA demethylation,[277] and have also been shown to induce changes in pre-mRNA splicing.[278] We examined whether SAHA induces changes in DNA modifications at the exon-intron junction. Vehicle-treated B lymphocytes (no SAHA) showed no 5-hmC exon-intron differences at d = 20 (p = 0.27); however, SAHA-treated cells showed an increase in the cross-boundary differences in a manner roughly corresponding to increasing doses of SAHA (p = 0.01, 8.8×10^{-4}, and 3.6×10^{-4} for 0.01 µM, 0.02 µM and 0.1 µM of SAHA, respectively; Table 3.5)

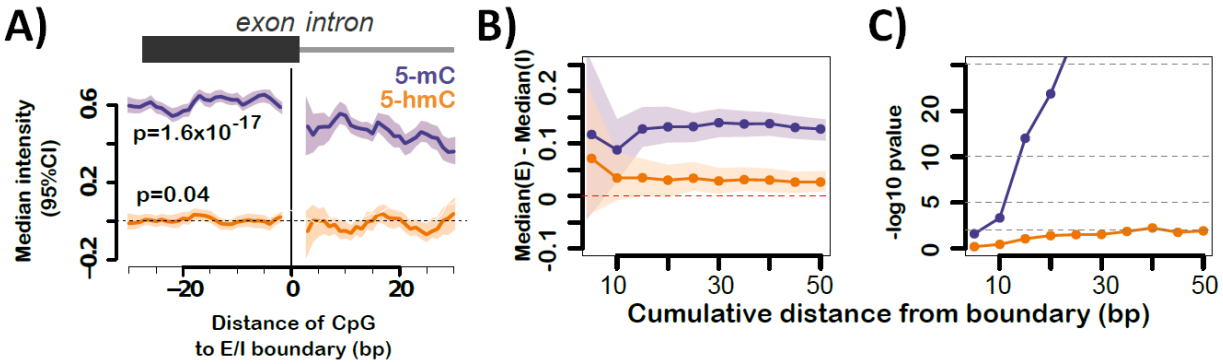

Figure 3.26. DNA modifications at exon-intron boundary in a mouse neuronal cell line with low global levels of 5-hmC.
This cell line (mHypoA-2/24) has negligible amounts of 5-hmC (thin-layer chromatography, not shown; microarray, Table 3.6) (N = 24, 3 chromosomes). Consistent with globally undetectable levels of 5-hmC, the main change at the exon-intron boundary was that in 5-mC rather than 5-hmC. A) Probe intensities in the region immediately around the exon-intron boundary. B) Median exonic increase at various cumulative distances from the boundary; C) Informal p-values (linear mixed-effects model) for exon-intron comparisons.

3.3.10. Exonic 5-hmC and exon inclusion levels

Changes in DNA modifications at exon-intron boundaries may impact exon recognition and exon inclusion levels. To investigate this possibility, we measured 5-hmC as a function of exon inclusion levels, using RNAseq data for human frontal cortex and liver.[263] Exon inclusion levels were measured as the proportion of transcripts that include a given exon; exons were classified as being alternatively-spliced (exon included in ≤ 80 % of gene transcripts) or constitutive (exon included in 100 % of gene transcripts). We observed that, at d = 20, the median 5-hmC density was lower in alternatively-spliced exons, relative to constitutive exons (p = 0.05, two-tailed WMW test; see Table 3.7 for probe counts and exon sampling), but was not different for 5-mC (d = 20, p = 0.96); number of probes at d = 5 were too small for a meaningful

comparison. The 5-hmC effect was even stronger at the level of the whole exon ($p = 8.4x10^{-5}$); here a marginal change was observed in 5-mC levels ($p = 0.025$, Figure 3.27A). The increase in 5-hmC was still borderline significant after probes in alternative and constitutive exons were matched for GC content (1000 iterations, 197 probes per group; geometric mean of $p = 0.07$ for 5-hmC and $p = 0.08$ for 5-mC). In the liver, neither 5-hmC nor 5-mC was different among the two types of exons (Figure 3.27B).

Separately, we found that average exonic 5-hmC in intron-less or single exon genes was lower than that in genes with multiple exons. For this analysis we used tiling array data for control human brains over six chromosomes (chips E and F, 30 samples, sample median used for each probe; RefSeq gene definitions were used; within each category, duplicate bases were collapsed). We found that exonic 5-hmC in single-exon genes was statistically different from that in multi-exon genes (median, intronless = 0.06, multi-exon = 0.073; $p = 0.026$, two-tailed WMW; N: intronless = 1,256 probes, multi-exon = 22,500 probes). This finding is consistent with an exonic change in 5-hmC due to gene splicing.

Collectively, the results suggest that the density of 5-hmC proximal to splice sites and within exons could impact splicing and affect exon inclusion levels in the mammalian brain.

	Brain (six chromosomes)		**Liver (three chromosomes)**	
	Constitutive exons	**Alternative exons**	**Constitutive exons**	**Alternative exons**
# exons on array chromosomes	5,862	980	1,048	118
Whole exon				
exons with probes	1,010	95	177	15
probes	1,234	199	224	18
d<=20 from boundary				
exons with probes	349	30	71	4
probes	358	31	73	4

Table 3.7. Exonic probe count of RNAseq data from human liver and brain (cortex).
RNAseq data was obtained from Brawand *et al.* (2011).[263]

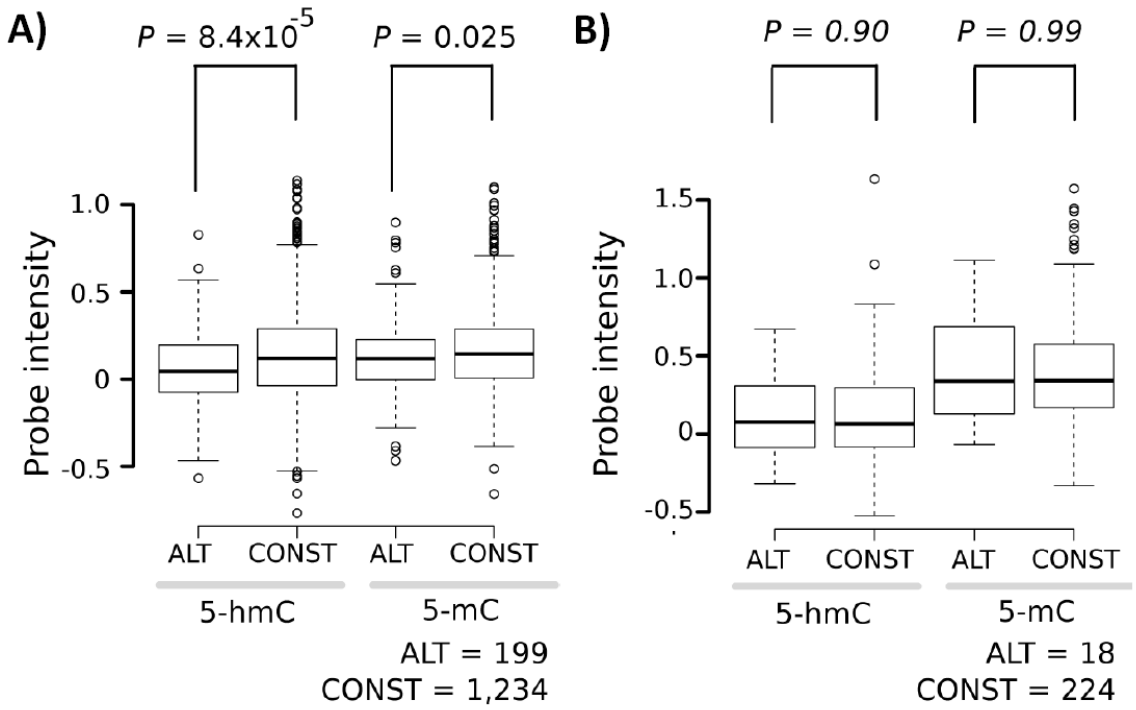

Figure 3.27. Constitutive exons have higher 5-hmC levels than alternatively-spliced exons in human brain.

Comparison of probe-level intensities of each modification in alternatively-spliced (ALT) and constitutive (CONST) exons. Comparisons were made for A) human brain and B) human liver. P-values are from two-tailed WMW ($\alpha = 0.05$). Probe count shown under boxplots for each panel.

3.4. 5-mC and 5-hmC at *HCG9* CpG6

In our bisulfite pyrosequencing based study of *HCG9* (section 3.1), lower DNA modification at CpGs 5-9 represented the most significant effect on disease risk across brain, WBC and sperm tissue (OR = 1.24, p = 0.001; Figure 3.8). We determined 5-mC and 5-hmC modification density at *HCG9* CpG6 (contains *Msp*I/*Hpa*II target site (CCGG)) in a cohort of *post-mortem* brain samples (N = 45 BPD, 25 SCZ, 53 control) and WBC samples (N = 30 BPD and 30 control) (Table 2.1) by quantitative real-time PCR (Section 2.13).

Consistent with previous results (Figure 3.18, Table 3.3), we found significantly higher 5-hmC in brain tissue compared to WBC (18.76 ± 0.56, 0.75 ± 0.04, respectively, Mann-Whitney U test, $p = 1.8x10^{-40}$). We also found significantly higher 5-mC density in brain versus WBC (13.77 ± 1.01, 6.33 ± 0.47, respectively, Mann-Whitney U test, $p = 6.5x10^{-6}$; Figure 3.28).

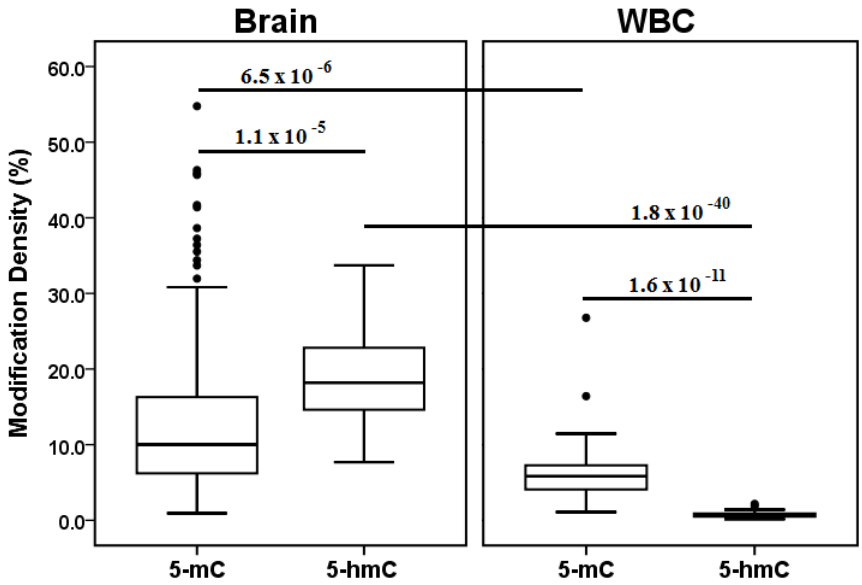

Figure 3.28. Distribution of 5-mC and 5-hmC density at *HCG9* CpG6 in brain and WBC samples.
5-mC and 5-hmC density was significantly higher in brain tissue compared to WBCs. P-values were obtained from Mann-Whitney U test and are indicated above boxplots. Within brain tissue, 5-hmC density was significantly higher than 5-mC. The opposite was true for WBC samples. P-values were obtained from Wilcoxon signed rank test and are indicated above boxplots.

3.4.1. 5-mC and 5-hmC density at *HCG9* CpG6 in brain and WBCs

In control brain samples, age significantly correlated with 5-hmC density (Spearman's rho = 0.36, p = 0.008) but not with 5-mC (Spearman's rho = -0.14, p = 0.32). Consistent with previous results (Appendix 3) total modification density (5-mC + 5-hmC) for rs1128306 SNP was lower for GG genotype compared to GA genotype but did not reach significance (30.9 ± 1.2,

34.01 ± 1.52, respectively, Mann-Whitney U test, p = 0.09). Also, we did not find any significant differences for 5-mC or 5-hmC density between rs1128306 SNP genotypes. We next tested association of 5-mC and 5-hmC density with rs603753 SNP (G/A) which is 93bp upstream of rs1128306 and is in perfect linkage disequilibrium with rs6904029 (r^2 = 1 and *D'*= 1) and rs2071568 (r^2 = 1 and *D'*= 1) but not with rs1128306 (r^2 = 0.08 and *D'*= 1). For control brain samples, 5-hmC density was significantly different between rs6903753 SNP AA and GA genotypes (17.25 ± 0.96 and 22.9 ± 1.3, respectively, Mann-Whitney U Test, p= 0.002). Numbers of GG homozygotes were too small to investigate separately and therefore those were included in the GA group. Marginally significant differences were identified for 5-mC modification density between rs6903753 SNP genotypes (AA = 10.3 ± 1.6 and GG / GA = 14.4 ± 2, p= 0.06). In our dataset, the minor allele frequency (MAF) of rs6903753, rs6904029 and rs2071568 was 0.3, 0.3 and 0.31 respectively, which is consistent with the reported MAF of 0.37, 0.37 and 0.36, respectively, for these SNPs (1000 genomes; dbSNP NCBI).

We eliminated the effects of age and rs6903753 genotype by linear model transformation and performed non-parametric Mann-Whitney U test on 5-hmC and 5-mC modification density between cases and controls. For the combined major psychosis group 5-hmC density was significantly lower compared to control samples (-1.12 ± 0.63 and 1.39 ± 0.77, respectively, Mann-Whitney U test, p = 0.04), while no significant differences were found for 5-mC density (p = 0.8). Similarly, we observed lower 5-hmC modification density in SCZ and BPD groups when analyzed separately, however these results were only marginally significant possibly due to loss in sample size (SCZ = -1.9 ± 1.07 *vs.* control = 0.87 ± 0.76, p = 0.07; BPD = -1.25 ± 0.8 *vs.* control = 0.97 ± 0.77; p = 0.1).

In contrast to brain tissue, age was significantly correlated with 5-mC density (Spearman's rho = 0.36, p = 0.05) in BPD WBC samples. Whereas, consistent with brain samples, significant differences in 5-hmC density between AA and GA genotypes at rs6903753 were identified for control WBC samples (0.62 ± 0.13 and 0.91 ± 0.02, respectively, p = 0.05). After correction for age and rs6903753 genotype no significant differences between 5-mC (p = 0.6) and 5-hmC (p = 0.35) density were detected in WBC samples of BPD and control individuals.

3.4.2. 5-mC and 5-hmC density at *HCG9* CpG6 across different brain regions

Next we determined variability in 5-mC and 5-hmC density for 10 different brain regions obtained from two unrelated individuals. As shown in Figure 3.29, we found substantial variation

in 5-mC and 5-hmC density across different brain regions. Consistent with our previous results (Figure 3.28), 5-mC density was lower than 5-hmC density across all brain regions. To determine if variability in modification density across different brain regions correlates with *HCG9* expression we obtained normalized mRNA values for 6 unrelated individuals across 8 different brain regions from Allen brain atlas. Log_2 normalized probe intensities for each brain region were averaged across 6 individuals and correlated with *HCG9* CpG6 5-mC and 5-hmC density. We found significant negative correlation of 5-hmC density with steady state *HCG9* mRNA levels (Spearman's rho = -0.81, p = 0.015), however, no significant results were obtained for 5-mC density (Spearman's rho = -0.45, p = 0.26).

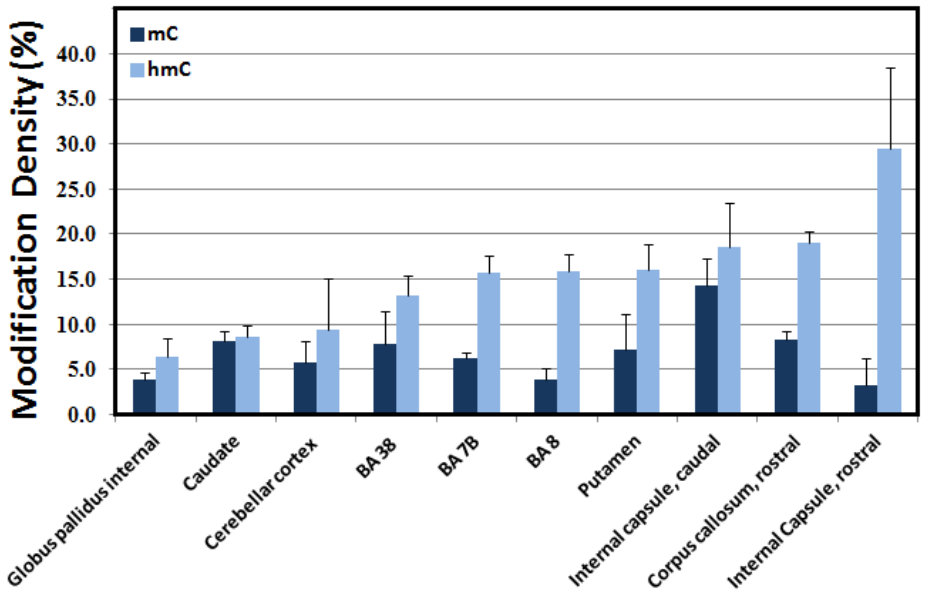

Figure 3.29. 5-mC and 5-hmC density varies significantly across different brain regions. Multiple brain tissue from different brain regions were obtained for two unrelated individuals. 5-mC and 5-hmC density was estimated for *HCG9* CpG6.

3.4.3. Incomplete *Msp*I digestion of brain gDNA

We noticed that a fraction of brain gDNA was resistant to *Msp*I but the enzyme nearly completely cut WBC gDNA (5.56 ± 0.49 and 0.31 ± 0.05, respectively, Mann Whitney U test, p = 4.3×10^{-27}; Figure 3.30A). This phenomenon was interesting given that *Msp*I was previously shown (Figure 3.16A) to result in complete digestion of 5-mC and 5-hmC modified CpG dinucleotide within the CCGG sequence. Moreover, we also detected strong correlation between amount of DNA uncut after *Msp*I and *Hpa*II digestion (Spearman's rho = 0.78, p = 3.2×10^{-26}; Figure 3.30B), indicating brain gDNA resistance to *Msp*I digestion also extends to *Hpa*II digestion.

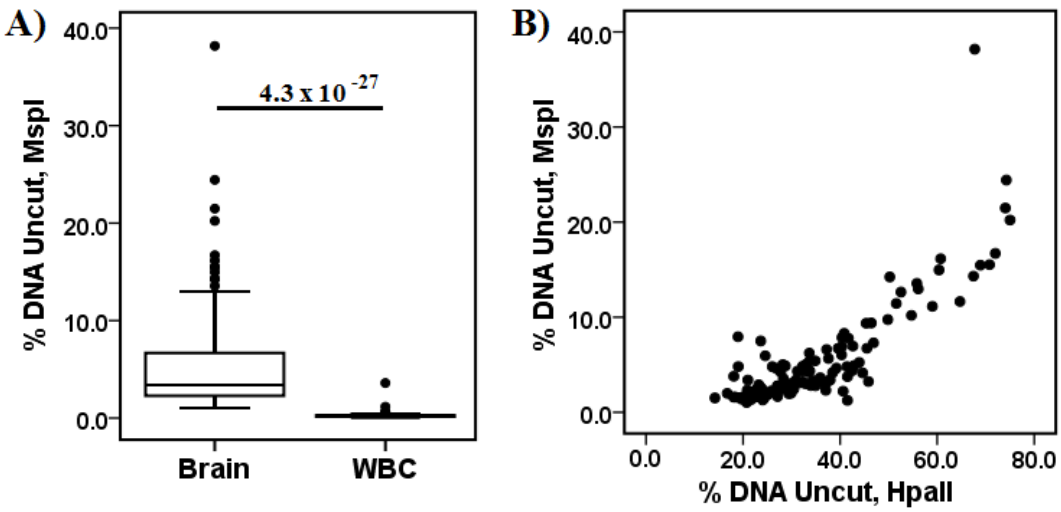

Figure 3.30. Incomplete *Msp*I digestion of brain gDNA at *HCG9* CpG6.
A) % DNA uncut after *Msp*I treatment is significantly higher for brain gDNA (N = 123) compared to WBC (N = 60). P-value is provided above boxplots. B) Significant correlation of % DNA uncut after *Msp*I and *Hpa*II digestion in brain gDNA (Spearman's rho = 0.78, $p = 3.2 \times 10^{-26}$, N = 123).

To determine if incomplete *Msp*I digestion extends to other loci across the genome we determined the amount of DNA uncut after *Msp*I digestion across 11 loci (Table 2.5) in 4 unrelated prefrontal cortex samples. There was significant variability in *Msp*I digestion across the investigated 11 loci (% DNA uncut by *Msp*I, range = 5.6 ± 0.75 - 37.9 ± 3.36, Kruskal-Wallis test, $p = 2.3 \times 10^{-5}$; Figure 3.31A). We next estimated amount of DNA uncut across 18 different brain regions for 5 different genomic loci and found substantial variability in *Msp*I digestion (Figure 3.31B)

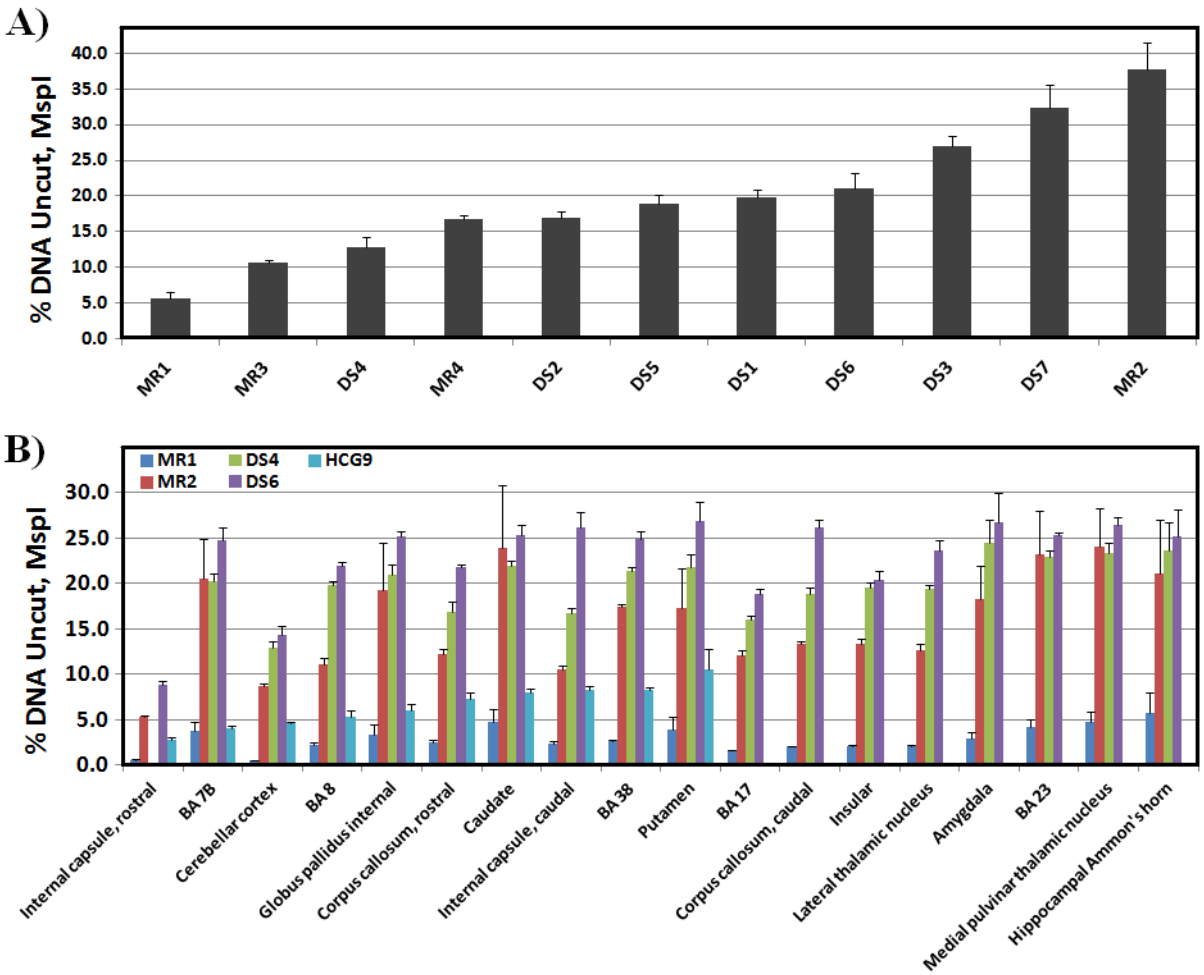

Figure 3.31. *Msp*I digestion varies significantly across different brain regions and across different genomic loci.
% DNA uncut after *Msp*I digestion was estimated for a) prefrontal cortex tissue (N = 4) for 11 genomic loci and b) 18 and 10 different brain regions for 4 loci and *HCG9* CpG 6, respectively. Different brain regions were obtained from 2 unrelated individuals (N = 2). There was significant variability in *Msp*I digestion across the investigated 11 loci in prefrontal cortex gDNA ($p = 2.3x10^{-5}$) and across different brain regions for 5 different genomic loci.

3.4.4. Incomplete *Msp*I digestion as a proxy for 5-fC and 5-caC modification

We hypothesized that somatic DNA sequence variation at the CCGG site might abolish the recognition sequence and result in incomplete *Msp*I digestion. To test this hypothesis we performed *Msp*I digestion on purified PCR products from 4 prefrontal cortex gDNA samples. PCR removes all DNA modification enabling us to investigate dependence of *Msp*I digestion strictly on gDNA sequence. *Msp*I digestion was very efficient on 6 different PCR fragments (Table 2.5) compared to gDNA which indicated the absence of sequence variation at the investigated CCGG sites (Figure 3.32A). Modification of external cytosine within CCGG sequence might also abolish *Msp*I digestion,[279,280] however our padlock probe-based deep sequencing analysis of *HCG9* (Section 3.2) did not reveal any significant amount of modification

(0.06%) at cytosine residue directly upstream of CpG6. Furthermore, the consensus sequence GGCCGG has been shown to inhibit *Msp*I digestion,[281,282] however sequence alignment of all investigated loci did not identify presence of GGCCGG sequence instead "G" was overrepresented on the 3' and 5' end of CCGG (Figure 3.32B).

Next, we interrogated the effect of 5-hmC oxidation products (5-fC and 5-caC) on *Msp*I digestion. To this end, we generated double stranded artificial oligonucleotides (89bp) with a single CCGG site containing different modification on the forward and reverse strand at internal cytosine. *Msp*I was unable to digest 5-caC modified oligonucleotides irrespective of the modification present on the other strand and was able to partially digest 5-fC modified oligonucleotides (Figure 3.32C).

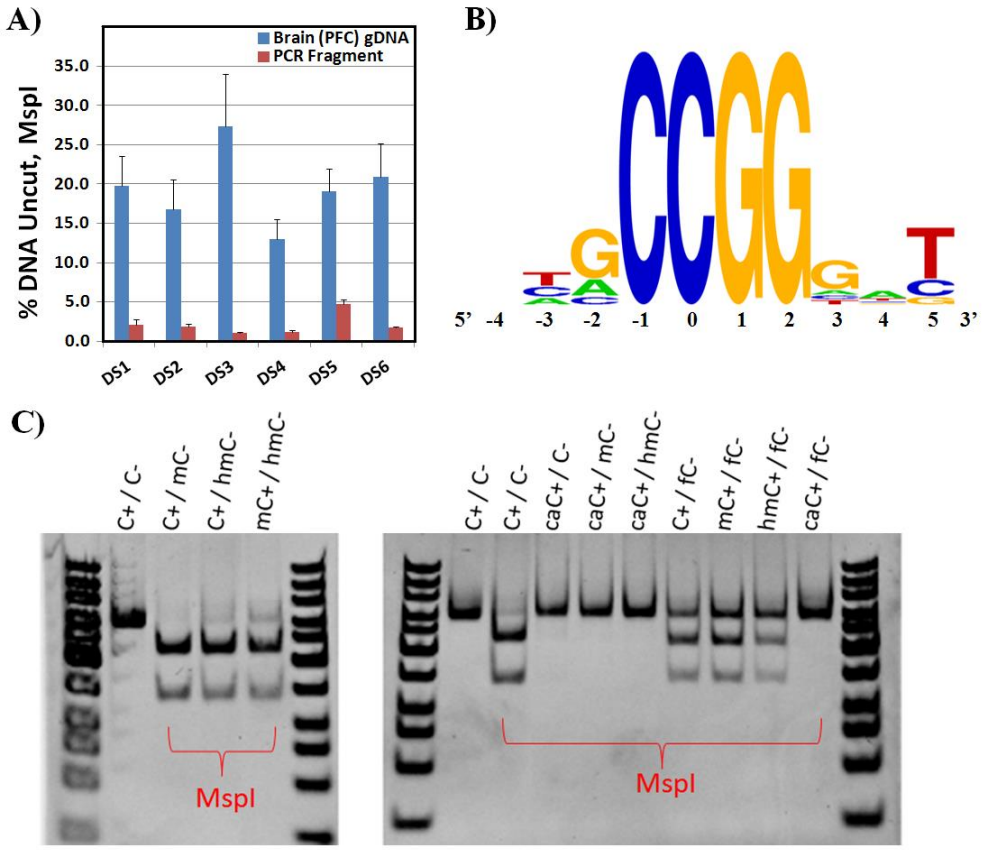

Figure 3.32. Incomplete *Msp*I digestion in the presence of 5-hmC oxidative products (5-fC and 5-caC).
A) *Msp*I digestion of PCR amplified fragments and gDNA from 4 prefrontal cortex samples. *Msp*I completely cut PCR amplified fragments but not brain gDNA indicating incomplete *Msp*I digestion is not due to somatic DNA sequence variation. B) Sequence alignment of all loci interrogated for *Msp*I digestion efficiency show overrepresentation of "G" on the 5' and 3' end of CCGG sequence. C) 89bp artificial oligonucleotide, containing a single CCGG site differentially modified on the plus (+) and minus (-) strand (shown above each lane), was digested with *Msp*I and run on 12.5% PAGE. *Msp*I digestion shows incomplete digestion of 5-caC and 5-fC modified oligonucleotides.

Chapter 4. Discussion, conclusions and future directions

4.1. Bisulfite pyrosequencing of *HCG9*

The goal of this study was to validate previously identified DNA modification differences in major psychosis and control individuals at *HCG9*.[231] To this end, we finely mapped the 5' end of *HCG9* with a bisulfite pyrosequencing based assay in a cohort of *post-mortem* brain tissues, WBC, and the germline. Our analysis revealed consistently lower *HCG9* modification density in BPD patients compared to controls, across all investigated tissues. We also observed that *HCG9* modification patterns are associated with DNA sequence variation and are age dependent. While in brain and WBC, *HCG9* modification density increases with age, the opposite is true for the germline. These findings suggest that with increasing paternal age, the probability of contributing low DNA modification in sperm upon fertilization increases and may have relevance to epidemiological observations of higher incidence of BPD amongst children of older fathers.[283-285]

The rs1128306 SNP at *HCG9* also was a factor which contributed to DNA modification differences across individuals. More specifically, allele A exhibited a higher density of modified cytosines in the surrounding vicinity of CpG 8 $\pm$ ~65 bp in each direction in comparison to the alternative allele G. This was observed across all three investigated tissues. While we identified associations of this SNP with BPD between the combined brain sample and WBC sample, alone, the sample sizes of these individual cohorts is likely too small to identify true SNP associations with BPD and may explain why allele G appeared to exert a risk inducing and protective effect in the respective cohorts. These results highlight the possibility that DNA polymorphisms associated with disease, such as those identified in GWAS studies, may be acting synergistically with epigenetic misregulation.[286,287] This is a particularly interesting proposal in the light of recent findings suggesting that 6p21.3-6p22 (*HCG9* maps to NCBI 36, 6p22.3) has been implicated by several large GWAS studies in major psychiatric disease.[26,264,265] rs1128306 SNP was a strong marker for a portion of the epigenetic variation in the region and, along with other SNPs, could be linked to the markers tested by the GWAS studies, suggesting that a portion of the disease association signal coming out of the MHC could represent this and possibly other epigenetically misregulated regions.

Some of the *HCG9* modification differences in BPD affected individuals and controls cannot be explained by any of the above listed factors. Opposite to the expected overall age-dependent increase in modification, the SMRI control group exhibited a higher density of

modified cytosines at some positions compared to the McL group (e.g. 50% and 31% at CpG 3), despite the fact that the former group consisted of younger individuals than the latter. One interpretation is that age effects may be bi-directional, even within a short DNA fragment. Alternatively, we cannot exclude the possibility that this difference in the two brain cohorts was due to the small sample size, which was not sufficient to reliably estimate age dynamics at each individual CpG dinucleotide but which would disappear if additional brain tissues were investigated.

The cause *vs.* effect relationship between disease and DNA modification is not readily evident since epigenetic patterns can be influenced by disease state, treatment or other events related to the pathological process. To address this issue we analyzed non-brain tissues which are unlikely to be involved in the disease process. Both, WBC and sperm, 'mirrored' most of the brain findings which argues for (although does not prove) an etiological role of DNA modification differences in BPD. Furthermore, the results of germline studies allow us to infer the possibility that *HCG9* epimutation may represent one of the heritable epigenetic risk factors in BPD. DNA modification is known to be subject to changes during both gametogenesis and after fertilization in humans and mice; however, numerous single locus examples exist where epigenetic alterations in the parental generation are passed to the offspring.[128,197,288] It is important to keep this finding in perspective, however, as the epigenetic difference was identified in the sperm of men affected with BPD and not the fathers of BPD patients. We can only suggest that the identified *HCG9* modification profiles are to some extent meiotically stable and thus may have the potential to survive successive epigenetic resetting and be passed to the next generation.

The identification of an epigenetic difference in BPD detectable in WBCs holds promise for identification of predictive biomarkers for the disease. Although area under the ROC curves demonstrated only a mildly predictive value for disease, given that *HCG9* was selected from an interrogation of only ~1% of the epigenome,[231] the *HCG9* epigenetic biomarker for BPD stands as a proof of principle that the epigenomic studies of multiple tissues may result in clinical applications. Identification of biomarkers should be far more successful if the candidate genes or regions are selected from the front-runners of the scan of the entire epigenome.

Functional implications of the detected epigenetic difference at CpGs 5-9, just downstream of the first coding exon, on *HCG9* activity and its role in BPD are not fully clear and this requires a dedicated study. Although the regulatory role of DNA modification is most characteristically recognized in its role in the gene promoter,[289-292] in the case of *HCG9* and

BPD, the disease- related mechanism can be different. One possibility is that DNA modification at CpGs 5-9 may be contributing to alternative splicing scenarios of the *HCG9* mRNA transcript. According to AceView (http://www.ncbi.nlm.nih.gov/IEB/Research/Acembly/index.html), *HCG9* has three mRNA splice variants, only one of which appears to be functionally relevant. There is a growing body of evidence that chromatin conformation and histone modifications help direct aspects of co-transcriptional splicing.[293,294] In our experiments, we found correlations between DNA modification and *HCG9* splice variant ratios at peaks of predicted histone occupancy, suggesting that our DNA modification levels may be a reflection of associated histone modification status and that epigenetic alterations in these positions are important for alternative splicing mechanisms.

4.2. Bisulfite padlock probe-based sequencing (BSPP-seq) of *HCG9*

We detected significant modification differences in both CpH and CpG dinucleotides of the *HCG9* region between major psychosis and control individuals. Our study underlines the importance of mapping modified cytosines in a strand specific manner. Many psychiatric epigenetic studies have not considered the potential asymmetry of DNA modification thus far. This may result in missing both disease epimutations and information about the molecular mechanisms of epigenetic misregulation in diseased brains. Our findings of CpH modification differences in sense and anti-sense strands of *HCG9* are consistent with earlier studies on mammalian ESCs.[93,295] Similar to the stem cell studies, we also detected CpH modification predominantly at CpA dinucleotides. The functional role of CpH modification is still unclear but recent studies have shown that gene body CpH modification is inversely proportional to the abundance of the associated transcripts.[100,101]

Several studies have shown that genome-wide accumulation of CpH modification is strongly linked to activity of DNA methyltrasferases DNMT3A/B and DNMTL.[101,104,296] Higher CpH modification density in major psychosis samples may be due to upregulation of *DNMT3A* in the prefrontal cortex of psychosis patients compared to controls.[235] Although limited to one gene, the sharp decline of CpH modification observed after 40 years of age in BPD samples is consistent with the observation that some aging BPD patients demonstrate diminished psychotic symptoms and significant improvements in the latter parts of their lives.[297] Identification of epigenetic markers, which exhibit significant differences in young BPD patients but eventually return to control levels, may help to understand the impact of aging brain on the origin of major psychosis.

Our hierarchical clustering analysis, although based on a single gene and therefore of limited scope, offers a molecular approach for division of the complex phenotype of major psychosis into more clinically and etiologically homogenous subgroups. It has been generally accepted that SCZ (amongst many other psychiatric diseases) represents a "fuzzy cluster" of syndromes which are heterogeneous from clinical, pathological, and etiological point of view.[298] Individual epigenetic profiles may help us identify disease subtypes that cannot be differentiated clinically or using other molecular approaches such as genetic linkage and association studies or proteomic approaches. To this end, our cluster and PCA analysis based on modification density at CpG and CpH sites resulted in delineation of groups of psychosis samples with different *HCG9* modification profiles.

We have also identified certain limitations of the padlock based approach that could be

improved in future studies. The investigated 1.4kb region of *HCG9* translates to approximately 33% of probes that were initially designed. The regions that were not interrogated could have resulted from degradation of DNA during bisulfite reaction and insufficient specificity of some probes. Furthermore, uneven read depth might result from unequal amplification efficiency of probes and possible non-ideal reaction conditions caused by the replacement of the discontinued Stoffel fragment polymerase with Hemo KlenTaq polymerase. Further improvement of the technology is required to obtain even coverage across probe set.

Finally, although bisulfite sequencing is a powerful approach, interrogation of CpH modification density requires careful design of primer sequences. When designing PCR primers for bisulfite converted DNA it is usually assumed that cytosines within CpH dinucleotides are not modified and therefore converted to uracils. This may create bias for amplification in the presence of modified CpH sites. Similarly, the instances of probe arms overlapping with SNPs and CpG sites should be minimized to limit potential annealing biases. Therefore, for reliable estimation of modification levels, combination of different, overlapping padlock probes interrogating the same target region should be utilized.

4.3. Distribution of 5-mC and 5-hmC across multiple mammalian tissue

In this study, we adapted a 5-hmC detection strategy that uses glucosylation-induced resistance to restriction enzymes for tiling microarray-based mapping of 5-hmC. The glucosylation-based detection of 5-hmC has been successfully used in other studies.[85,86,299] Our microarray-based quantification of DNA modifications (verified by TLC) showed significantly higher levels of 5-hmC in brain as compared to other tissues, which is consistent with previous reports.[77,78] We identified a large number of genes with higher densities of 5-hmC in the brain compared to those in heart, liver, kidney, and pancreas. More generally, we discovered a trend where 5-hmC in the gene body increases with increased transcription of the corresponding gene, while 5-mC decreases. This is consistent with the observation that 5-hmC prevents the binding of transcriptional repressor proteins and is found within actively-transcribed genes.[85,300-302] Interestingly, the association between genic 5-hmC and mRNA levels was weakest in the brain ($p = 0.17$), suggesting that in addition to the regulation of gene activity 5-hmC may have other functions in the cell.

Functional annotation analysis revealed a statistical overrepresentation of terms pertaining to synaptic plasticity in genes enriched for 5-hmC in the brain. In particular, annotation clusters were composed of protein groups involved in distinct aspects of synaptic remodelling: ion channels, members of the Rho GTPase signalling pathway, and axon guidance molecules. Notably, most genes in these clusters encode proteins that are functionally located at the plasma membrane, rather than being cytosolic. The observation that 5-hmC is overrepresented in the genes controlling synaptic plasticity may shed a new light on the epigenetics of learning and memory. Adult animals with conditional double knockouts of genes encoding DNA methyltransferase 1 and 3b (*Dnmt1* and *Dnmt3b*) in the cerebral cortex showed learning and memory defects in hippocampal-dependent learning tasks, suggesting that changes in DNA modification occur in post-mitotic neurons.[303] Fear conditioning, an experimental paradigm for emotional learning, results in the demethylation and transcriptional activation of *reelin* (*RELN*).[304] Recent mappings of 5-hmC in the brain have identified generation of 5-hmC and demethylation in response to neuronal activity, but the degree to which the two correspond is not completely clear.[88,305] In the absence of DNA replication in post-mitotic neurons, it is possible that loci undergoing gene reactivation via DNA demethylation accumulate 5-hmC over time. Experiments that map 5-hmC over the course of multiple gene reactivations will explore whether, in the context of synaptic activity, this base marks a stable epigenetic state, or if it is an intermediate in cycles including complete demethylation (i.e. conversion to unmodified

cytosines).

In addition to transcriptional regulation, it is possible that 5-hmC and 5-mC may impact the process of pre-mRNA splicing. Our finding of boundary changes in all modified cytosines replicates earlier observations that used bisulfite modification and sequencing.[94,275,306] The separation of the two DNA modifications in our study showed that 5-hmC, rather than 5-mC, accounts for most of the density difference at the immediate exon-intron boundary in human frontal cortex samples. This finding was validated by three control experiments: comparison of GC/AT matched probes, Helicos single molecule sequencing, and methyl-binding domain (MBD)-enriched 5-mC mapping. While the mouse brain samples show a predominant change in 5-hmC at longer peri-boundary distances, the main change within the first 10 bp of the boundary was in 5-mC.

The non-neural tissues investigated included human liver and four mouse organs (liver, pancreas, heart and kidney). While these tissues had measurable levels of genomic 5-hmC density by TLC, modification changes at the exon-intron boundary were mainly due to 5-mC. In contrast to the pattern observed in the brain, peri-boundary changes in 5-hmC were relatively minor in these tissues. These findings suggest that any splicing-related functions or effects of DNA modification in non-neural organs are mediated mainly by 5-mC rather than 5-hmC. Separately, we found that B-lymphocytes show increased 5-hmC differences at the exon-intron boundary upon treatment with SAHA. SAHA is a chemotherapeutic agent that acts as a HDAC inhibitor and may promote DNA demethylation.[277] As the dose used in this study is comparable to those in plasma of patients treated for cancer with SAHA,[307] 5-hmC changes at the exon-intron boundary may also be expected to take place *in vivo.* Hence, it is also possible that such changes in 5-hmC in response to HDAC inhibition could also contribute to alternative splicing.[278]

Previous studies have reported that exons are enriched in the histone modifications H3K36me3, H3K4me3, and H3K27me2, relative to flanking intronic regions.[308-310] These modifications can recruit splicing regulators to impact alternative splicing of nascent transcripts.[311] While the resolution of histone maps is limited by the size of nucleosomal DNA (147 bp), our DNA modification studies have identified changes in modification levels within 20 bp (and in some cases, within 5 bp) surrounding the exon-intron boundary. This precision argues for a specific effect at the exon-intron boundary as well as a possible difference between exons and introns as a whole, as described in earlier studies.[305] Our finding that 5-hmC densities are lower in alternatively-spliced exons, relative to constitutive exons, complements the observation

that H3K36me3 is less enriched in alternatively-spliced exons.[308,309,312] DNA methylation has been shown to modulate exon inclusion levels by influencing rate of transcript elongation in B-lymphocytes. Lack of DNA methylation can promote exon inclusion by causing ‘pauses’ in RNA Polymerase II (RNA PolII)-mediated elongation, and may also affect RNA PolII elongation-dependent changes in alternative splicing by affecting the binding of the transcriptional repressor CTCF.[313] There may also be other mechanisms by which DNA methylation could influence exon inclusion levels, for example, through the recruitment of splicing factors via methyl binding proteins.

Our findings suggest that tissue-specific distributions of 5-hmC or 5-mC at the exon-intron boundary, and within genes, may simultaneously influence both transcription and splicing. The direction of the influence remains unclear, as transcription and splicing may also affect epigenetic DNA modifications, and mechanisms for cross-talk likely exist between epigenetic regulation, splicing, and transcription.[314,315]

4.4. 5-mC and 5-hmC at *HCG9* CpG6

In this study we investigated differences in 5-mC and 5-hmC modification density, at previously investigated *HCG9* CpG6, between major psychosis and control individuals. Consistent with our previous results (Section 3.3), we found enrichment of 5-hmC in brain tissue but not in WBC. Also, our observation of increasing 5-hmC at *HCG9* with age in brain tissue is in agreement with existing literature.[305] Furthermore, our previously identified association between modification density at *HCG9* CpG6 with rs1128306 SNP proved to be invalid, in both brain and WBC, and instead, we found significant correlation with rs6903753. These results indicate that associations between genetic polymorphisms and DNA modification might change depending on the specific type of cytosine modification (i.e. 5-mC+5-hmC, 5-mC or 5-hmC). Further investigation is required in a larger cohort as this inconsistency could be attributed to a small sample size. Lastly, in brain tissue, significantly lower modification density at *HCG9* CpG6 in major psychosis versus control individuals was due to 5-hmC modification and not total modification (5mC+5hmC) as determined previously (Section 3.1), while no significant differences were reported for WBC. Our results highlight the importance of segregating different types of DNA modifications which may serve different functions in the genome and therefore, their impact on psychiatric disorders should be interpreted accordingly. We emphasize that these differences were identified in a much smaller cohort of brain and WBC samples compared to our previous study (Section 3.1), and therefore the results need extensive validation in large samples.

We also measured 5-mC and 5-hmC density across several different brain regions from two control individuals. Our results indicate that 5-hmC density at *HCG9* CpG6 is not only enriched in brain tissue but changes significantly across these brain regions, which is consistent with previous reports of changes in 5-hmC density across several brain regions.[78] Furthermore, significant correlation of 5-hmC density with steady state mRNA level of *HCG9* suggests that 5-hmC may be involved in regulation of *HCG9* transcription.

An intriguing finding from this work was the incomplete digestion of brain gDNA by *Msp*I. This finding has previously been reported in the literature,[280-282] although not thoroughly investigated. Our results show that the efficiency of *Msp*I digestion varies across different regions in the genome, individuals and brain regions. One possible explanation for incomplete *Msp*I digestion is existence of non-B DNA structures (for example, cruciform DNA, Z-DNA, G-quadruplex and slipped DNA) which are distributed in a non-random manner across the genome.[316] Formation of these DNA structures depends on DNA sequence and are primarily found overlapping with repetitive regions of the genome. However, the observed incomplete

digestion of *Msp*I was sequence independent and tissue-specific (present only in the brain but not in WBC). Furthermore, *Msp*I digestion of artificial oligonucleotides is partially and completely blocked by the presence of 5-fC and 5-caC, respectively. These results indicate that estimation of undigested DNA after *Msp*I digestion might serve as a proxy for 5-hmC oxidation products, 5-fC and 5-caC, and *HCG9* CpG 6 might be a site for active DNA demethylation. Further investigation with methods that can reliably detect 5-fC and 5-caC is required to validate these findings.

4.5. Conclusions

In this study we thoroughly interrogated *HCG9* modification density at CpG and CpH sites and established its role as a potent epigenetic risk factor in major psychosis. Consistent epigenetic differences observed between affected and unaffected samples across multiple tissues strongly argue for a causal link between aberrant *HCG9* modification and major psychosis. We identified significant associations with age and DNA sequence variation which, if ignored, have the potential to result in false positive or false negative findings. Furthermore, we also demonstrated the predictive utility of epigenetic markers to differentiate between diseased states. Identification and extensive validation of such epigenetic markers can have a profound effect on accurate diagnosis and management of major psychosis, including early clinical intervention.

Additionally, we also demonstrate the utility of a recently developed BSPP-seq technology in populational studies of complex disorders. This approach offers a new opportunity to interrogate multiple samples while WGBS remains cost prohibitive. With BSPP-seq one can theoretically interrogate ~ 80,000 probes at minimum 50x coverage in 384 samples with the Illumina HiSeq 2500 platform, while taking advantage of all the benefits associated with WGBS, such as strand- and allele-specific estimation of modification densities at CpG and CpH sites, at one-tenth of the cost associated with WGBS. We also employed novel analytical strategies which favour replication across different tissues rather than direct statistical evidence for isolated analysis of each tissue. By using a sliding window based method and performing statistical tests on all possible adjacent CpG combinations, we identified overlaps across tissues and replication cohorts that we would have otherwise missed. This approach may become exemplary in the primary studies of DNA modification differences when neither the number of cytosines involved, nor the multiple ways of their interaction is known. The study highlights the need of methodological principles for epigenomic and epigenetic studies of complex disease that are quite different from the ones used in DNA sequence-based approaches.

We also interrogated genome-wide distribution of 5-mC and 5-hmC in neural and non-neural mammalian tissue. We observed tissue-specific distribution of 5-hmC and 5-mC within genes and at the exon-intron boundary which may simultaneously influence both transcription and splicing. 5-hmC in concert with other regulators of splicing, including histone modifications, may help determine splice-site choice in mammalian brain; in turn regulating neuronal function. Since all current observations are of a correlative nature, it is equally possible that 5-hmC changes at splice junction sites are secondary to other signals deciding splice-site choice, or that the two influence each other. While our work was limited to 5-hmC

modification within *Msp*I/*Hpa*II target sites (CCGG) further research should involve fine-mapping of 5-hmC and 5-mC around the splice junction with more sensitive techniques to clarify the mechanistic role of 5-hmC at the junction, particularly in neuronal tissue.

Lastly, our analysis of 5-mC and 5-hmC at *HCG9* CpG6 revealed significant difference in 5-hmC modification density between major psychosis and control individuals. CpG6 is located ~70bp upstream of the splice site and DNA modification (5-mC+5hmC) at this site was observed to be correlated with *HCG9* splice variants. Given the significant role of 5-hmC in brain mRNA splicing and its abundance at *HCG9* CpG6, it is plausible to hypothesize that 5-hmC may play a role in splicing at *HCG9*. Fine mapping of 5-hmC and 5-mC across *HCG9* combined with mRNA splice variant analysis will reveal its potential contribution in *HCG9* mRNA splicing in affected and unaffected brain tissue. We also observed significant variability in *Msp*I digestion in brain gDNA but not in WBCs. Our experiments indicate that inefficient *Msp*I digestion might result from presence of 5-hmC oxidation products, 5-fC and 5-caC, which may accumulate in differing amounts depending upon the investigated genomic locus and brain region. Although further validation is required with techniques that directly investigate these rare forms of DNA modification, our data suggests active DNA demethylation might be an ongoing process in the brain.

4.6. Recommendations for future studies

Our *HCG9* modification study represents one of the most thoroughly investigated epigenetic risk factors in major psychosis. Despite the comprehensive DNA modification mapping of *HCG9* several aspects can be improved and new important epigenetic avenues added in future epigenetic studies of major psychiatric disease. In this section, we discuss issues related to the brain cellular heterogeneity, requirements for the brain samples, differentiation of inherited epigenetic risk factors from the acquired ones, mapping of the various subtypes of cytosine modifications, and several other developments.

Epigenetic profiles, unlike DNA sequence, are cell type specific and tissue from primary disease sites is required.[317] However, brain exhibits significant cellular heterogeneity with approximately half of the bulk tissue composed of non-neuronal (glial) cells that exhibit different epigenetic profiles compared to neurons.[318,319] This situation is further complicated by different neuronal and glial cell subtypes which exhibit diverse gene expression profiles, function and morphology.[258,320-322] As a result of this cellular heterogeneity, epigenetic differences between affected and unaffected individuals may be "diluted" by the cells which are not related to the disease process. For example, serotonergic neurons are implicated in MDD while dopaminergic and glutamatergic neurons are involved in SCZ.[323-325] Furthermore, epigenetic differences observed in affected individuals compared to controls might in fact be due to differing ratios of cellular subtypes. This is of particular relevance to psychiatric disorders where significant disease associated morphological changes in brain have been detected.[326-330] For example, in the prefrontal cortex decrease in neuronal size and increase in neuronal density has been reported in SCZ while significant reduction in oligodendrocytes has been reported in SCZ, MDD and BPD.[329,331,332] In our *HCG9* studies we have controlled for this bias to some extent by using *in silico* methods.[271] The current gold standard approach is separation of the neuronal and glial fractions using antibody enrichment methods.[333] An ideal design, however, would be to use laser capture microdissection to isolate and interrogate each neuronal cell type individually.

There is considerable lack of brain tissue for SCZ and BPD with accompanying detailed clinical information such as, ethnicity, *post-mortem* interval, brain pH, age of disease onset, lifetime history of illness, operational criteria applied to diagnosis, comorbidities and presence of other neuropsychiatric diseases, treatment history, antipsychotic use, cause of death, and use of alcohol, cigarettes and recreational drugs. Relevant clinical information is critical for epigenetic studies to help distinguish between causative epimutations and those associated with clinical intervention (use of antipsychotics) and other environmental or disease related confounding

factors (recreational drug use).

Another problem with human brain studies is that of small sample size, which may result in relatively low power to detect significant differences between groups. In reality, adequately powered epigenomic studies of *post-mortem* human brain may not be feasible, especially assuming small absolute DNA modification changes in diseased individuals. For example, at $\alpha = 10^{-6}$ (the low α allows for correction of multiple testing in epigenome-wide studies) a sample size of 400 cases and controls is required to detect 10% difference in modification density with 80% power whereas, a sample size of 800 is required for 7% difference.[334] To overcome these issues in the epigenetic brain studies, we added a number of other more accessible tissues from living patients, such as WBCs and germline (sperm), where inter-individual differences in modification profiles have been shown to correlate with that of brain tissue.[335] Inclusion of multiple tissues dramatically improves sample size, and identification of consistent epigenetic differences across several tissues makes a strong case for a causal relationship, as these epigenetic marks likely arose early during pre-natal development before tissue differentiation, or were inherited from the parental germ cells.

To further understand the role of inherited epigenetic risk factors for brain disease, affected familial samples can be investigated. Such effort may turn productive in establishing trans-generational inheritance of etiological epigenetic marks and can also offer new insights into "missing heritability" (discussed in Sections 1.2.2 and 1.4.4). Paternal and maternal transmission of epigenetic states have been identified in inbred (and, therefore, presumably isogenic) mice.[197,288,336] Additionally, inclusion of samples from affected monozygotic and dizygotic twin pairs can also help establish heritability of epigenetic marks.[135,199,200] Meanwhile, epigenetic differences in discordant monozygotic twins may help to identify environment- and stochastically-induced epigenetic risk factors.[337]

Several approaches for interrogation of the epigenome have been described in this thesis, including, bisulfite pyrosequencing, BSPP-seq, and restriction enzyme-based enrichment of methylated and hydroxymethylated cytosines and interrogation on tiling array. Tiling arrays provide comprehensive coverage of the non-repetitive genome, however the enrichment method is limited to investigation of CpG sites located within the restriction enzyme recognition sequence. Meanwhile, BSPP- seq offers complete coverage of CpG and CpH sites but does not differentiate between 5-mC and 5-hmC. Recent advances in bisulfite sequencing enable simultaneous detection of 5-mC and 5-hmC[338,339] and should be utilized in combination with padlock probes to estimate allele- and strand-specific modification density in CpG and CpH

context. Several hundred thousand padlock probes can be designed to simultaneously investigate hundreds of samples for regions previously identified to be differentially modified in major psychosis[231] and/or major psychosis-associated loci identified from GWAS studies, to interrogate genetic-epigenetic interaction in disease. Furthermore, multiple sets of several thousand padlock probes can be designed to investigate the most variable CpG sites in the human genome (5.6 million dynamic CpG sites) with no *a priori* knowledge of disease specific differentially modified regions.[220] Quantification of 5-fC and 5-caC density and distribution across the genome still remains a challenge due to low abundance of these modifications (10 to 100 fold lower than 5-hmC). However, recent studies have been published that quantify 5-fC and 5-caC at single base pair resolution in embryonic stem cells.[340,341] These techniques should be properly assessed and implemented in future studies.

Altered mRNA levels in brain tissue have been consistently reported in SCZ, BPD and other psychiatric diseases.[342-346] DNA and histone modifications are intricately linked to gene expression and have been shown to bring about alterations in mRNA levels in response to a variety of factors.[347-349] Therefore, DNA modification studies can be further complemented with analysis of histone modification status combined with traditional quantification of mRNA and protein levels to decipher gene regulatory networks. Of particular interest are methylated forms of histone 3 at lysines 4, 9 and 27 since these marks are relatively stable in *post-mortem* tissue and are able to differentiate between inactive and active chromatin states.[350]

Lastly, functional consequence of 5-hmC density at the exon-intron boundary and its effects on pre-mRNA splicing should be evaluated in a model system. This can be achieved with the use of exon-trapping technique which utilizes a vector with a strong promoter driving expression of a multi-exon "minigene".[351] The region of interest is cloned in the intron of the vector derived gene and transfected in mammalian cells. After several hours, RNA is isolated from transfected cells and inclusion of the region of interest in vector RNA is evaluated with reverse transcriptase PCR using primers in vector exons flanking the insertion site. Since we require specific modification at the exon-intron boundary, a genomic fragment of an exon flanked by introns containing the splice donor and acceptor sites should be artificially synthesized to contain either 5-mC, 5-hmC or unmodified cytosine at exonic CpG sites up to 50bp upstream from the exon-intron boundary. Furthermore, the synthesized oligo must be from a genomic fragment previously characterized to be alternatively spliced in the minigene assay (for example, exon10 of human muscle specific receptor tyrosine kinase gene).[352] The cloned vector containing specific modification marks should be transfected in neuronal and non-

neuronal cell lines to investigate the role of the observed tissue specific accumulation of 5-hmC at exon-intron boundary.

References

1. Kaminsky Z, Tochigi M, Jia P, et al. A multi-tissue analysis identifies HLA complex group 9 gene methylation differences in bipolar disorder. *Molecular Psychiatry.* 2012;17(7):728-740.
2. Pal M, Ebrahimi S, Oh G, et al. High Precision DNA Modification Analysis of HCG9 in Major Psychosis. *Schizophrenia Bulletin.* 2015.
3. Khare T, Pai S, Koncevicius K, et al. 5-hmC in the brain is abundant in synaptic genes and shows differences at the exon-intron boundary. *Nature Structural & Molecular Biology.* 2012;19(10):1037-1043.
4. McGrath J, Saha S, Chant D, Welham J. Schizophrenia: a concise overview of incidence, prevalence, and mortality. *Epidemiologic Reviews.* 2008;30:67-76.
5. Jablensky A. Schizophrenia: recent epidemiologic issues. *Epidemiologic Reviews.* 1995;17(1):10-20.
6. Merikangas KR, Akiskal HS, Angst J, et al. Lifetime and 12-month prevalence of bipolar spectrum disorder in the National Comorbidity Survey replication. *Archives of General Psychiatry.* 2007;64(5):543-552.
7. Kendler KS, Gallagher TJ, Abelson JM, Kessler RC. Lifetime prevalence, demographic risk factors, and diagnostic validity of nonaffective psychosis as assessed in a US community sample. The National Comorbidity Survey. *Archives of General Psychiatry.* 1996;53(11):1022-1031.
8. Crow TJ. Molecular pathology of schizophrenia: more than one disease process? *British Medical Journal.* 1980;280(6207):66-68.
9. Andreasen NC, Arndt S, Alliger R, Miller D, Flaum M. Symptoms of schizophrenia. Methods, meanings, and mechanisms. *Archives of General Psychiatry.* 1995;52(5):341-351.
10. Hasan A, Falkai P, Wobrock T, et al. World Federation of Societies of Biological Psychiatry (WFSBP) Guidelines for Biological Treatment of Schizophrenia, part 1: update 2012 on the acute treatment of schizophrenia and the management of treatment resistance. *World Journal of Biological Psychiatry.* 2012;13(5):318-378.
11. Hoff AL, Harris D, Faustman WO, et al. A neuropsychological study of early onset schizophrenia. *Schizophrenia Research.* 1996;20(1-2):21-28.
12. Hoff AL, Svetina C, Shields G, Stewart J, DeLisi LE. Ten year longitudinal study of neuropsychological functioning subsequent to a first episode of schizophrenia. *Schizophrenia Research.* 2005;78(1):27-34.
13. Heinrichs RW, Zakzanis KK. Neurocognitive deficit in schizophrenia: a quantitative review of the evidence. *Neuropsychology.* 1998;12(3):426-445.
14. Torres IJ, Boudreau VG, Yatham LN. Neuropsychological functioning in euthymic bipolar disorder: a meta-analysis. *Acta Psychiatrica Scandinavica. Supplementum.* 2007(434):17-26.
15. *Diagnostic and statistical manual of mental disorders: DSM-5.* 5th ed. Washington, D.C.: American Psychiatric Association; 2013.
16. Cuthbert BN. Research Domain Criteria: toward future psychiatric nosologies. *Dialogues in Clinical Neuroscience.* 2015;17(1):89-97.
17. Cardno AG, Owen MJ. Genetic relationships between schizophrenia, bipolar disorder, and schizoaffective disorder. *Schizophrenia Bulletin.* 2014;40(3):504-515.
18. Taylor L, Faraone SV, Tsuang MT. Family, twin, and adoption studies of bipolar disease. *Curr Psychiatry Rep.* 2002;4(2):130-133.

19. Shih RA, Belmonte PL, Zandi PP. A review of the evidence from family, twin and adoption studies for a genetic contribution to adult psychiatric disorders. *Int Rev Psychiatry.* 2004;16(4):260-283.
20. Schizophrenia Working Group of the Psychiatric Genomics Consortium. Biological insights from 108 schizophrenia-associated genetic loci. *Nature.* 2014;511(7510):421-427.
21. Purcell SM, Moran JL, Fromer M, et al. A polygenic burden of rare disruptive mutations in schizophrenia. *Nature.* 2014;506(7487):185-190.
22. Fromer M, Pocklington AJ, Kavanagh DH, et al. De novo mutations in schizophrenia implicate synaptic networks. *Nature.* 2014;506(7487):179-184.
23. Kirov G, Pocklington AJ, Holmans P, et al. De novo CNV analysis implicates specific abnormalities of postsynaptic signalling complexes in the pathogenesis of schizophrenia. *Molecular Psychiatry.* 2012;17(2):142-153.
24. Psychiatric GWAS Consortium Bipolar Disorder Working Group. Large-scale genome-wide association analysis of bipolar disorder identifies a new susceptibility locus near ODZ4. *Nature Genetics.* 2011;43(10):977-983.
25. Rasic D, Hajek T, Alda M, Uher R. Risk of mental illness in offspring of parents with schizophrenia, bipolar disorder, and major depressive disorder: a meta-analysis of family high-risk studies. *Schizophrenia Bulletin.* 2014;40(1):28-38.
26. Purcell SM, Wray NR, Stone JL, et al. Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. *Nature.* 2009;460(7256):748-752.
27. Cross-Disorder Group of the Psychiatric Genomics Consortium. Identification of risk loci with shared effects on five major psychiatric disorders: a genome-wide analysis. *Lancet.* 2013;381(9875):1371-1379.
28. Green EK, Grozeva D, Jones I, et al. The bipolar disorder risk allele at CACNA1C also confers risk of recurrent major depression and of schizophrenia. *Molecular Psychiatry.* 2010;15(10):1016-1022.
29. Ferreira MA, O'Donovan MC, Meng YA, et al. Collaborative genome-wide association analysis supports a role for ANK3 and CACNA1C in bipolar disorder. *Nature Genetics.* 2008;40(9):1056-1058.
30. O'Donovan MC, Craddock N, Norton N, et al. Identification of loci associated with schizophrenia by genome-wide association and follow-up. *Nature Genetics.* 2008;40(9):1053-1055.
31. Williams HJ, Norton N, Dwyer S, et al. Fine mapping of ZNF804A and genome-wide significant evidence for its involvement in schizophrenia and bipolar disorder. *Molecular Psychiatry.* 2011;16(4):429-441.
32. Cichon S, Muhleisen TW, Degenhardt FA, et al. Genome-wide association study identifies genetic variation in neurocan as a susceptibility factor for bipolar disorder. *American Journal of Human Genetics.* 2011;88(3):372-381.
33. Muhleisen TW, Mattheisen M, Strohmaier J, et al. Association between schizophrenia and common variation in neurocan (NCAN), a genetic risk factor for bipolar disorder. *Schizophrenia Research.* 2012;138(1):69-73.
34. Walsh T, McClellan JM, McCarthy SE, et al. Rare structural variants disrupt multiple genes in neurodevelopmental pathways in schizophrenia. *Science.* 2008;320(5875):539-543.
35. Xu B, Roos JL, Levy S, van Rensburg EJ, Gogos JA, Karayiorgou M. Strong association of de novo copy number mutations with sporadic schizophrenia. *Nature Genetics.* 2008;40(7):880-885.

36. Pinto D, Pagnamenta AT, Klei L, et al. Functional impact of global rare copy number variation in autism spectrum disorders. *Nature.* 2010;466(7304):368-372.
37. Williams NM, Zaharieva I, Martin A, et al. Rare chromosomal deletions and duplications in attention-deficit hyperactivity disorder: a genome-wide analysis. *Lancet.* 2010;376(9750):1401-1408.
38. Williams NM, Franke B, Mick E, et al. Genome-wide analysis of copy number variants in attention deficit hyperactivity disorder: the role of rare variants and duplications at 15q13.3. *American Journal of Psychiatry.* 2012;169(2):195-204.
39. Cooper GM, Coe BP, Girirajan S, et al. A copy number variation morbidity map of developmental delay. *Nature Genetics.* 2011;43(9):838-846.
40. Grozeva D, Kirov G, Ivanov D, et al. Rare copy number variants: a point of rarity in genetic risk for bipolar disorder and schizophrenia. *Archives of General Psychiatry.* 2010;67(4):318-327.
41. Grozeva D, Kirov G, Conrad DF, et al. Reduced burden of very large and rare CNVs in bipolar affective disorder. *Bipolar Disord.* 2013;15(8):893-898.
42. McQuillin A, Bass N, Anjorin A, et al. Analysis of genetic deletions and duplications in the University College London bipolar disorder case control sample. *European Journal of Human Genetics.* 2011;19(5):588-592.
43. Malhotra D, McCarthy S, Michaelson JJ, et al. High frequencies of de novo CNVs in bipolar disorder and schizophrenia. *Neuron.* 2011;72(6):951-963.
44. Zhang D, Cheng L, Qian Y, et al. Singleton deletions throughout the genome increase risk of bipolar disorder. *Molecular Psychiatry.* 2009;14(4):376-380.
45. Doherty JL, Owen MJ. Genomic insights into the overlap between psychiatric disorders: implications for research and clinical practice. *Genome Medicine.* 2014;6(4):29.
46. Yang J, Lee SH, Goddard ME, Visscher PM. GCTA: a tool for genome-wide complex trait analysis. *American Journal of Human Genetics.* 2011;88(1):76-82.
47. Lee SH, Ripke S, Neale BM, et al. Genetic relationship between five psychiatric disorders estimated from genome-wide SNPs. *Nature Genetics.* 2013;45(9):984-994.
48. McGrath J, Saha S, Welham J, El Saadi O, MacCauley C, Chant D. A systematic review of the incidence of schizophrenia: the distribution of rates and the influence of sex, urbanicity, migrant status and methodology. *BMC Medicine.* 2004;2:13.
49. Krabbendam L, van Os J. Schizophrenia and urbanicity: a major environmental influence--conditional on genetic risk. *Schizophrenia Bulletin.* 2005;31(4):795-799.
50. Kelly BD, O'Callaghan E, Waddington JL, et al. Schizophrenia and the city: A review of literature and prospective study of psychosis and urbanicity in Ireland. *Schizophrenia Research.* 2010;116(1):75-89.
51. Pedersen CB, Mortensen PB. Evidence of a dose-response relationship between urbanicity during upbringing and schizophrenia risk. *Archives of General Psychiatry.* 2001;58(11):1039-1046.
52. Shapero BG, Black SK, Liu RT, et al. Stressful life events and depression symptoms: the effect of childhood emotional abuse on stress reactivity. *Journal of Clinical Psychology.* 2014;70(3):209-223.
53. D'Souza DC, Abi-Saab WM, Madonick S, et al. Delta-9-tetrahydrocannabinol effects in schizophrenia: implications for cognition, psychosis, and addiction. *Biological Psychiatry.* 2005;57(6):594-608.
54. Cannon M, Jones PB, Murray RM. Obstetric complications and schizophrenia: historical and meta-analytic review. *American Journal of Psychiatry.* 2002;159(7):1080-1092.

55. van Os J, Rutten BP, Poulton R. Gene-environment interactions in schizophrenia: review of epidemiological findings and future directions. *Schizophrenia Bulletin.* 2008;34(6):1066-1082.
56. European Network of Schizophrenia Networks for the Study of Gene-Environment Interactions. Schizophrenia aetiology: do gene-environment interactions hold the key? *Schizophrenia Research.* 2008;102(1-3):21-26.
57. Bouchard TJ, Jr., McGue M. Genetic and environmental influences on human psychological differences. *Journal of Neurobiology.* 2003;54(1):4-45.
58. Bouchard TJ, Jr., Lykken DT, McGue M, Segal NL, Tellegen A. Sources of human psychological differences: the Minnesota Study of Twins Reared Apart. *Science.* 1990;250(4978):223-228.
59. Modinos G, Iyegbe C, Prata D, et al. Molecular genetic gene-environment studies using candidate genes in schizophrenia: a systematic review. *Schizophrenia Research.* 2013;150(2-3):356-365.
60. Iyegbe C, Campbell D, Butler A, Ajnakina O, Sham P. The emerging molecular architecture of schizophrenia, polygenic risk scores and the clinical implications for GxE research. *Social Psychiatry and Psychiatric Epidemiology.* 2014;49(2):169-182.
61. Decoster J, van Os J, Myin-Germeys I, De Hert M, van Winkel R. Genetic variation underlying psychosis-inducing effects of cannabis: critical review and future directions. *Current Pharmaceutical Design.* 2012;18(32):5015-5023.
62. Varriale A. DNA methylation, epigenetics, and evolution in vertebrates: facts and challenges. *International Journal of Evolutionary Biology.* 2014;2014:475981.
63. Moore LD, Le T, Fan G. DNA methylation and its basic function. *Neuropsychopharmacology.* 2013;38(1):23-38.
64. Serrano L, Vazquez BN, Tischfield J. Chromatin structure, pluripotency and differentiation. *Experimental Biology and Medicine (Maywood, N.J.).* 2013;238(3):259-270.
65. Deaton AM, Bird A. CpG islands and the regulation of transcription. *Genes and Development.* 2011;25(10):1010-1022.
66. Chaligne R, Heard E. X-chromosome inactivation in development and cancer. *FEBS Letters.* 2014;588(15):2514-2522.
67. Adalsteinsson BT, Ferguson-Smith AC. Epigenetic control of the genome-lessons from genomic imprinting. *Journal of Neuroscience.* 2014;5(3):635-655.
68. Jin B, Li Y, Robertson KD. DNA methylation: superior or subordinate in the epigenetic hierarchy? *Genes & Cancer.* 2011;2(6):607-617.
69. Probst AV, Dunleavy E, Almouzni G. Epigenetic inheritance during the cell cycle. *Nature Reviews: Molecular Cell Biology.* 2009;10(3):192-206.
70. Li E. Chromatin modification and epigenetic reprogramming in mammalian development. *Nature Reviews: Genetics.* 2002;3(9):662-673.
71. Saitou M, Kagiwada S, Kurimoto K. Epigenetic reprogramming in mouse pre-implantation development and primordial germ cells. *Development.* 2012;139(1):15-31.
72. Bird A. DNA methylation patterns and epigenetic memory. *Genes and Development.* 2002;16(1):6-21.
73. Goll MG, Kirpekar F, Maggert KA, et al. Methylation of tRNAAsp by the DNA methyltransferase homolog Dnmt2. *Science.* 2006;311(5759):395-398.
74. Schaefer M, Pollex T, Hanna K, et al. RNA methylation by Dnmt2 protects transfer RNAs against stress-induced cleavage. *Genes and Development.* 2010;24(15):1590-1595.

75. Penn NW, Suwalski R, O'Riley C, Bojanowski K, Yura R. The presence of 5-hydroxymethylcytosine in animal deoxyribonucleic acid. *Biochemical Journal.* 1972;126(4):781-790.
76. Tahiliani M, Koh KP, Shen Y, et al. Conversion of 5-methylcytosine to 5-hydroxymethylcytosine in mammalian DNA by MLL partner TET1. *Science.* 2009;324(5929):930-935.
77. Kriaucionis S, Heintz N. The nuclear DNA base 5-hydroxymethylcytosine is present in Purkinje neurons and the brain. *Science.* 2009;324(5929):929-930.
78. Globisch D, Munzel M, Muller M, et al. Tissue distribution of 5-hydroxymethylcytosine and search for active demethylation intermediates. *PloS One.* 2010;5(12):e15367.
79. Ito S, D'Alessio AC, Taranova OV, Hong K, Sowers LC, Zhang Y. Role of Tet proteins in 5mC to 5hmC conversion, ES-cell self-renewal and inner cell mass specification. *Nature.* 2010;466(7310):1129-1133.
80. Branco MR, Ficz G, Reik W. Uncovering the role of 5-hydroxymethylcytosine in the epigenome. *Nature Reviews: Genetics.* 2012;13(1):7-13.
81. Mariani CJ, Madzo J, Moen EL, Yesilkanal A, Godley LA. Alterations of 5-hydroxymethylcytosine in human cancers. *Cancers.* 2013;5(3):786-814.
82. He YF, Li BZ, Li Z, et al. Tet-mediated formation of 5-carboxylcytosine and its excision by TDG in mammalian DNA. *Science.* 2011;333(6047):1303-1307.
83. Ito S, Shen L, Dai Q, et al. Tet proteins can convert 5-methylcytosine to 5-formylcytosine and 5-carboxylcytosine. *Science.* 2011;333(6047):1300-1303.
84. Pfaffeneder T, Hackner B, Truss M, et al. The discovery of 5-formylcytosine in embryonic stem cell DNA. *Angewandte Chemie. International Ed. In English.* 2011;50(31):7008-7012.
85. Jin SG, Wu X, Li AX, Pfeifer GP. Genomic mapping of 5-hydroxymethylcytosine in the human brain. *Nucleic Acids Research.* 2011;39(12):5015-5024.
86. Song CX, Szulwach KE, Fu Y, et al. Selective chemical labeling reveals the genome-wide distribution of 5-hydroxymethylcytosine. *Nature Biotechnology.* 2011;29(1):68-72.
87. Yu M, Hon GC, Szulwach KE, et al. Base-resolution analysis of 5-hydroxymethylcytosine in the mammalian genome. *Cell.* 2012;149(6):1368-1380.
88. Guo JU, Su Y, Zhong C, Ming GL, Song H. Hydroxylation of 5-methylcytosine by TET1 promotes active DNA demethylation in the adult brain. *Cell.* 2011;145(3):423-434.
89. Pfaffeneder T, Spada F, Wagner M, et al. Tet oxidizes thymine to 5-hydroxymethyluracil in mouse embryonic stem cell DNA. *Nature Chemical Biology.* 2014;10(7):574-581.
90. Nabel CS, Jia H, Ye Y, et al. AID/APOBEC deaminases disfavor modified cytosines implicated in DNA demethylation. *Nature Chemical Biology.* 2012;8(9):751-758.
91. Zemach A, McDaniel IE, Silva P, Zilberman D. Genome-wide evolutionary analysis of eukaryotic DNA methylation. *Science.* 2010;328(5980):916-919.
92. Feng S, Cokus SJ, Zhang X, et al. Conservation and divergence of methylation patterning in plants and animals. *Proceedings of the National Academy of Sciences of the United States of America.* 2010;107(19):8689-8694.
93. Lister R, Pelizzola M, Dowen RH, et al. Human DNA methylomes at base resolution show widespread epigenomic differences. *Nature.* 2009;462(7271):315-322.
94. Laurent L, Wong E, Li G, et al. Dynamic changes in the human methylome during differentiation. *Genome Research.* 2010;20(3):320-331.
95. Chen PY, Feng S, Joo JW, Jacobsen SE, Pellegrini M. A comparative analysis of DNA methylation across human embryonic stem cell lines. *Genome Biology.* 2011;12(7):R62.
96. Ziller MJ, Muller F, Liao J, et al. Genomic distribution and inter-sample variation of non-CpG methylation across human cell types. *PLoS Genetics.* 2011;7(12):e1002389.

97. Ramsahoye BH, Biniszkiewicz D, Lyko F, Clark V, Bird AP, Jaenisch R. Non-CpG methylation is prevalent in embryonic stem cells and may be mediated by DNA methyltransferase 3a. *Proceedings of the National Academy of Sciences of the United States of America.* 2000;97(10):5237-5242.
98. Lister R, Pelizzola M, Kida YS, et al. Hotspots of aberrant epigenomic reprogramming in human induced pluripotent stem cells. *Nature.* 2011;471(7336):68-73.
99. Xie W, Barr CL, Kim A, et al. Base-resolution analyses of sequence and parent-of-origin dependent DNA methylation in the mouse genome. *Cell.* 2012;148(4):816-831.
100. Lister R, Mukamel EA, Nery JR, et al. Global epigenomic reconfiguration during mammalian brain development. *Science.* 2013;341(6146):1237905.
101. Guo JU, Su Y, Shin JH, et al. Distribution, recognition and regulation of non-CpG methylation in the adult mammalian brain. *Nature Neuroscience.* 2014;17(2):215-222.
102. Varley KE, Gertz J, Bowling KM, et al. Dynamic DNA methylation across diverse human cell lines and tissues. *Genome Research.* 2013;23(3):555-567.
103. Gowher H, Jeltsch A. Enzymatic properties of recombinant Dnmt3a DNA methyltransferase from mouse: the enzyme modifies DNA in a non-processive manner and also methylates non-CpG [correction of non-CpA] sites. *Journal of Molecular Biology.* 2001;309(5):1201-1208.
104. Arand J, Spieler D, Karius T, et al. In vivo control of CpG and non-CpG DNA methylation by DNA methyltransferases. *PLoS Genetics.* 2012;8(6):e1002750.
105. Aoki A, Suetake I, Miyagawa J, et al. Enzymatic properties of de novo-type mouse DNA (cytosine-5) methyltransferases. *Nucleic Acids Research.* 2001;29(17):3506-3512.
106. Kouzarides T. Chromatin modifications and their function. *Cell.* 2007;128(4):693-705.
107. Godde JS, Ura K. Cracking the enigmatic linker histone code. *Journal of Biochemistry.* 2008;143(3):287-293.
108. Kamakaka RT, Biggins S. Histone variants: deviants? *Genes and Development.* 2005;19(3):295-310.
109. Chen ES, Zhang K, Nicolas E, Cam HP, Zofall M, Grewal SI. Cell cycle control of centromeric repeat transcription and heterochromatin assembly. *Nature.* 2008;451(7179):734-737.
110. Downs JA, Nussenzweig MC, Nussenzweig A. Chromatin dynamics and the preservation of genetic information. *Nature.* 2007;447(7147):951-958.
111. Grewal SI, Jia S. Heterochromatin revisited. *Nature Reviews: Genetics.* 2007;8(1):35-46.
112. Berger SL. The complex language of chromatin regulation during transcription. *Nature.* 2007;447(7143):407-412.
113. Ito T. Role of histone modification in chromatin dynamics. *Journal of Biochemistry.* 2007;141(5):609-614.
114. Ruthenburg AJ, Li H, Patel DJ, Allis CD. Multivalent engagement of chromatin modifications by linked binding modules. *Nature Reviews: Molecular Cell Biology.* 2007;8(12):983-994.
115. Wade PA, Pruss D, Wolffe AP. Histone acetylation: chromatin in action. *Trends in Biochemical Sciences.* 1997;22(4):128-132.
116. Bartova E, Krejci J, Harnicarova A, Galiova G, Kozubek S. Histone modifications and nuclear architecture: a review. *Journal of Histochemistry and Cytochemistry.* 2008;56(8):711-721.
117. Lachner M, Jenuwein T. The many faces of histone lysine methylation. *Current Opinion in Cell Biology.* 2002;14(3):286-298.

118. Lachner M, Sengupta R, Schotta G, Jenuwein T. Trilogies of histone lysine methylation as epigenetic landmarks of the eukaryotic genome. *Cold Spring Harbor Symposia on Quantitative Biology.* 2004;69:209-218.
119. Weaver IC, Cervoni N, Champagne FA, et al. Epigenetic programming by maternal behavior. *Nature Neuroscience.* 2004;7(8):847-854.
120. Ching TT, Maunakea AK, Jun P, et al. Epigenome analyses using BAC microarrays identify evolutionary conservation of tissue-specific methylation of SHANK3. *Nature Genetics.* 2005;37(6):645-651.
121. Riggs AD, Xiong Z, Wang L, LeBon JM. Methylation dynamics, epigenetic fidelity and X chromosome structure. *Novartis Foundation Symposium.* 1998;214:214-225; discussion 225-232.
122. Eckhardt F, Lewin J, Cortese R, et al. DNA methylation profiling of human chromosomes 6, 20 and 22. *Nature Genetics.* 2006;38(12):1378-1385.
123. Klar AJ. Propagating epigenetic states through meiosis: where Mendel's gene is more than a DNA moiety. *Trends in Genetics.* 1998;14(8):299-301.
124. Rakyan V, Whitelaw E. Transgenerational epigenetic inheritance. *Current Biology.* 2003;13(1):R6.
125. Reik W, Dean W, Walter J. Epigenetic reprogramming in mammalian development. *Science.* 2001;293(5532):1089-1093.
126. Kaneda M, Okano M, Hata K, et al. Essential role for de novo DNA methyltransferase Dnmt3a in paternal and maternal imprinting. *Nature.* 2004;429(6994):900-903.
127. Chan TL, Yuen ST, Kong CK, et al. Heritable germline epimutation of MSH2 in a family with hereditary nonpolyposis colorectal cancer. *Nature Genetics.* 2006;38(10):1178-1183.
128. Oates NA, van Vliet J, Duffy DL, et al. Increased DNA methylation at the AXIN1 gene in a monozygotic twin from a pair discordant for a caudal duplication anomaly. *American Journal of Human Genetics.* 2006;79(1):155-162.
129. Suter CM, Martin DI, Ward RL. Germline epimutation of MLH1 in individuals with multiple cancers. *Nature Genetics.* 2004;36(5):497-501.
130. Cardno AG, Marshall EJ, Coid B, et al. Heritability estimates for psychotic disorders: the Maudsley twin psychosis series. *Archives of General Psychiatry.* 1999;56(2):162-168.
131. Deb-Rinker P, Klempan TA, O'Reilly RL, Torrey EF, Singh SM. Molecular characterization of a MSRV-like sequence identified by RDA from monozygotic twin pairs discordant for schizophrenia. *Genomics.* 1999;61(2):133-144.
132. Deb-Rinker P, O'Reilly RL, Torrey EF, Singh SM. Molecular characterization of a 2.7-kb, 12q13-specific, retroviral-related sequence isolated by RDA from monozygotic twin pairs discordant for schizophrenia. *Genome.* 2002;45(2):381-390.
133. Lavrentieva I, Broude NE, Lebedev Y, et al. High polymorphism level of genomic sequences flanking insertion sites of human endogenous retroviral long terminal repeats. *FEBS Letters.* 1999;443(3):341-347.
134. Vincent JB, Kalsi G, Klempan T, et al. No evidence of expansion of CAG or GAA repeats in schizophrenia families and monozygotic twins. *Human Genetics.* 1998;103(1):41-47.
135. Kaminsky ZA, Tang T, Wang SC, et al. DNA methylation profiles in monozygotic and dizygotic twins. *Nature Genetics.* 2009;41(2):240-245.
136. Kuratomi G, Iwamoto K, Bundo M, et al. Aberrant DNA methylation associated with bipolar disorder identified from discordant monozygotic twins. *Molecular Psychiatry.* 2008;13(4):429-441.

137. Dempster EL, Pidsley R, Schalkwyk LC, et al. Disease-associated epigenetic changes in monozygotic twins discordant for schizophrenia and bipolar disorder. *Human Molecular Genetics.* 2011;20(24):4786-4796.
138. Rosa A, Picchioni MM, Kalidindi S, et al. Differential methylation of the X-chromosome is a possible source of discordance for bipolar disorder female monozygotic twins. *American Journal of Medical Genetics. Part B: Neuropsychiatric Genetics.* 2008;147b(4):459-462.
139. Loat CS, Asbury K, Galsworthy MJ, Plomin R, Craig IW. X inactivation as a source of behavioural differences in monozygotic female twins. *Twin Research.* 2004;7(1):54-61.
140. Tsuchiya KD, Willard HF. Chromosomal domains and escape from X inactivation: comparative X inactivation analysis in mouse and human. *Mammalian Genome.* 2000;11(10):849-854.
141. Belmont JW. Genetic control of X inactivation and processes leading to X-inactivation skewing. *American Journal of Human Genetics.* 1996;58(6):1101-1108.
142. Plenge RM, Stevenson RA, Lubs HA, Schwartz CE, Willard HF. Skewed X-chromosome inactivation is a common feature of X-linked mental retardation disorders. *American Journal of Human Genetics.* 2002;71(1):168-173.
143. Ostrer H. Sex-based differences in gene transmission and gene expression. *Lupus.* 1999;8(5):365-369.
144. Castle DJ, Wessely S, Murray RM. Sex and schizophrenia: effects of diagnostic stringency, and associations with and premorbid variables. *British Journal of Psychiatry.* 1993;162:658-664.
145. Aleman A, Kahn RS, Selten JP. Sex differences in the risk of schizophrenia: evidence from meta-analysis. *Archives of General Psychiatry.* 2003;60(6):565-571.
146. Abel KM, Drake R, Goldstein JM. Sex differences in schizophrenia. *Int Rev Psychiatry.* 2010;22(5):417-428.
147. Kennedy N, Boydell J, Kalidindi S, et al. Gender differences in incidence and age at onset of mania and bipolar disorder over a 35-year period in Camberwell, England. *American Journal of Psychiatry.* 2005;162(2):257-262.
148. Braunig P, Sarkar R, Effenberger S, Schoofs N, Kruger S. Gender differences in psychotic bipolar mania. *Gender Medicine.* 2009;6(2):356-361.
149. Kaminsky Z, Wang SC, Petronis A. Complex disease, gender and epigenetics. *Annals of Medicine.* 2006;38(8):530-544.
150. Wang KS, Liu X, Zhang Q, Aragam N, Pan Y. Genome-wide association analysis of age at onset in schizophrenia in a European-American sample. *American Journal of Medical Genetics. Part B: Neuropsychiatric Genetics.* 2011;156B(6):671-680.
151. Kigar SL, Auger AP. Epigenetic mechanisms may underlie the aetiology of sex differences in mental health risk and resilience. *Journal of Neuroendocrinology.* 2013;25(11):1141-1150.
152. Abkevich V, Camp NJ, Hensel CH, et al. Predisposition locus for major depression at chromosome 12q22-12q23.2. *American Journal of Human Genetics.* 2003;73(6):1271-1281.
153. Zubenko GS, Maher B, Hughes HB, 3rd, et al. Genome-wide linkage survey for genetic loci that influence the development of depressive disorders in families with recurrent, early-onset, major depression. *American Journal of Medical Genetics. Part B: Neuropsychiatric Genetics.* 2003;123b(1):1-18.
154. Fu M, Rao M, Wang C, et al. Acetylation of androgen receptor enhances coactivator binding and promotes prostate cancer cell growth. *Molecular and Cellular Biology.* 2003;23(23):8563-8575.

155. Fu M, Wang C, Zhang X, Pestell RG. Acetylation of nuclear receptors in cellular growth and apoptosis. *Biochemical Pharmacology.* 2004;68(6):1199-1208.
156. Yamagata Y, Asada H, Tamura I, et al. DNA methyltransferase expression in the human endometrium: down-regulation by progesterone and estrogen. *Human Reproduction.* 2009;24(5):1126-1132.
157. van Kaam KJ, Delvoux B, Romano A, D'Hooghe T, Dunselman GA, Groothuis PG. Deoxyribonucleic acid methyltransferases and methyl-CpG-binding domain proteins in human endometrium and endometriosis. *Fertility and Sterility.* 2011;95(4):1421-1427.
158. Reyes FI, Boroditsky RS, Winter JS, Faiman C. Studies on human sexual development. II. Fetal and maternal serum gonadotropin and sex steroid concentrations. *Journal of Clinical Endocrinology and Metabolism.* 1974;38(4):612-617.
159. McCarthy MM. Estradiol and the developing brain. *Physiological Reviews.* 2008;88(1):91-124.
160. Kolodkin MH, Auger AP. Sex difference in the expression of DNA methyltransferase 3a in the rat amygdala during development. *Journal of Neuroendocrinology.* 2011;23(7):577-583.
161. Phelps EA, LeDoux JE. Contributions of the amygdala to emotion processing: from animal models to human behavior. *Neuron.* 2005;48(2):175-187.
162. Shaw P, Lawrence EJ, Radbourne C, Bramham J, Polkey CE, David AS. The impact of early and late damage to the human amygdala on 'theory of mind' reasoning. *Brain.* 2004;127(Pt 7):1535-1548.
163. Schumann CM, Bauman MD, Amaral DG. Abnormal structure or function of the amygdala is a common component of neurodevelopmental disorders. *Neuropsychologia.* 2011;49(4):745-759.
164. Amaral DG, Bauman MD, Schumann CM. The amygdala and autism: implications from non-human primate studies. *Genes Brain Behav.* 2003;2(5):295-302.
165. Goldstein JM, Cherkerzian S, Seidman LJ, et al. Sex-specific rates of transmission of psychosis in the New England high-risk family study. *Schizophrenia Research.* 2011;128(1-3):150-155.
166. Kato T, Winokur G, Coryell W, Keller MB, Endicott J, Rice J. Parent-of-origin effect in transmission of bipolar disorder. *American Journal of Medical Genetics.* 1996;67(6):546-550.
167. Kornberg JR, Brown JL, Sadovnick AD, et al. Evaluating the parent-of-origin effect in bipolar affective disorder. Is a more penetrant subtype transmitted paternally? *Journal of Affective Disorders.* 2000;59(3):183-192.
168. Lambert D, Gill M. Evaluation of parent-of-origin effect in bipolar affective disorder relating to susceptibility loci on chromosome 18. *Bipolar Disord.* 2002;4 Suppl 1:31-32.
169. Lan TH, Beaty TH, DePaulo JR, McInnis MG. Parent-of-origin effect in the segregation analysis of bipolar affective disorder families. *Psychiatric Genetics.* 2007;17(2):93-101.
170. Stine OC, Xu J, Koskela R, et al. Evidence for linkage of bipolar disorder to chromosome 18 with a parent-of-origin effect. *American Journal of Human Genetics.* 1995;57(6):1384-1394.
171. Hamshere ML, Schulze TG, Schumacher J, et al. Mood-incongruent psychosis in bipolar disorder: conditional linkage analysis shows genome-wide suggestive linkage at 1q32.3, 7p13 and 20q13.31. *Bipolar Disord.* 2009;11(6):610-620.
172. Mahon PB, Payne JL, MacKinnon DF, et al. Genome-wide linkage and follow-up association study of postpartum mood symptoms. *American Journal of Psychiatry.* 2009;166(11):1229-1237.

173. Kishino T. Imprinting in neurons. *Cytogenetic and Genome Research.* 2006;113(1-4):209-214.
174. Crespi B. Genomic imprinting in the development and evolution of psychotic spectrum conditions. *Biological Reviews of the Cambridge Philosophical Society.* 2008;83(4):441-493.
175. Hajkova P, Erhardt S, Lane N, et al. Epigenetic reprogramming in mouse primordial germ cells. *Mechanisms of Development.* 2002;117(1-2):15-23.
176. Sasaki H, Matsui Y. Epigenetic events in mammalian germ-cell development: reprogramming and beyond. *Nature Reviews: Genetics.* 2008;9(2):129-140.
177. Keverne EB, Martel FL, Nevison CM. Primate brain evolution: genetic and functional considerations. *Proceedings: Biological Sciences.* 1996;263(1371):689-696.
178. Keverne EB, Fundele R, Narasimha M, Barton SC, Surani MA. Genomic imprinting and the differential roles of parental genomes in brain development. *Brain Research: Developmental Brain Research.* 1996;92(1):91-100.
179. Barrett SL, Mulholland CC, Cooper SJ, Rushe TM. Patterns of neurocognitive impairment in first-episode bipolar disorder and schizophrenia. *British Journal of Psychiatry.* 2009;195(1):67-72.
180. Schouws SN, Comijs HC, Stek ML, et al. Cognitive impairment in early and late bipolar disorder. *American Journal of Geriatric Psychiatry.* 2009;17(6):508-515.
181. Knoll JH, Nicholls RD, Magenis RE, et al. Angelman syndrome: three molecular classes identified with chromosome 15q11q13-specific DNA markers. *American Journal of Human Genetics.* 1990;47(1):149-154.
182. Sutherland JE, Costa M. Epigenetics and the environment. *Annals of the New York Academy of Sciences.* 2003;983:151-160.
183. Schmitt A, Malchow B, Hasan A, Falkai P. The impact of environmental factors in severe psychiatric disorders. *Frontiers in Neuroscience.* 2014;8:19.
184. Weaver IC, Champagne FA, Brown SE, et al. Reversal of maternal programming of stress responses in adult offspring through methyl supplementation: altering epigenetic marking later in life. *Journal of Neuroscience.* 2005;25(47):11045-11054.
185. Wand GS, Oswald LM, McCaul ME, et al. Association of amphetamine-induced striatal dopamine release and cortisol responses to psychological stress. *Neuropsychopharmacology.* 2007;32(11):2310-2320.
186. Webster MJ, Knable MB, O'Grady J, Orthmann J, Weickert CS. Regional specificity of brain glucocorticoid receptor mRNA alterations in subjects with schizophrenia and mood disorders. *Molecular Psychiatry.* 2002;7(9):985-994, 924.
187. Tyrka AR, Price LH, Marsit C, Walters OC, Carpenter LL. Childhood adversity and epigenetic modulation of the leukocyte glucocorticoid receptor: preliminary findings in healthy adults. *PloS One.* 2012;7(1):e30148.
188. McGowan PO, Sasaki A, D'Alessio AC, et al. Epigenetic regulation of the glucocorticoid receptor in human brain associates with childhood abuse. *Nature Neuroscience.* 2009;12(3):342-348.
189. Altschuler EL. Schizophrenia and the Chinese famine of 1959-1961. *JAMA.* 2005;294(23):2968; discussion 2968-2969.
190. Kyle UG, Pichard C. The Dutch Famine of 1944-1945: a pathophysiological model of long-term consequences of wasting disease. *Current Opinion in Clinical Nutrition and Metabolic Care.* 2006;9(4):388-394.
191. St Clair D, Xu M, Wang P, et al. Rates of adult schizophrenia following prenatal exposure to the Chinese famine of 1959-1961. *JAMA.* 2005;294(5):557-562.

192. Xu MQ, Sun WS, Liu BX, et al. Prenatal malnutrition and adult schizophrenia: further evidence from the 1959-1961 Chinese famine. *Schizophrenia Bulletin.* 2009;35(3):568-576.
193. Elias SG, Keinan-Boker L, Peeters PH, et al. Long term consequences of the 1944-1945 Dutch famine on the insulin-like growth factor axis. *International Journal of Cancer.* 2004;108(4):628-630.
194. Dolinoy DC, Weidman JR, Waterland RA, Jirtle RL. Maternal genistein alters coat color and protects Avy mouse offspring from obesity by modifying the fetal epigenome. *Environmental Health Perspectives.* 2006;114(4):567-572.
195. Cooney CA, Dave AA, Wolff GL. Maternal methyl supplements in mice affect epigenetic variation and DNA methylation of offspring. *Journal of Nutrition.* 2002;132(8 Suppl):2393s-2400s.
196. Wolff GL, Kodell RL, Moore SR, Cooney CA. Maternal epigenetics and methyl supplements affect agouti gene expression in Avy/a mice. *FASEB Journal.* 1998;12(11):949-957.
197. Rakyan VK, Chong S, Champ ME, et al. Transgenerational inheritance of epigenetic states at the murine Axin(Fu) allele occurs after maternal and paternal transmission. *Proceedings of the National Academy of Sciences of the United States of America.* 2003;100(5):2538-2543.
198. Fraga MF, Ballestar E, Paz MF, et al. Epigenetic differences arise during the lifetime of monozygotic twins. *Proceedings of the National Academy of Sciences of the United States of America.* 2005;102(30):10604-10609.
199. Heijmans BT, Kremer D, Tobi EW, Boomsma DI, Slagboom PE. Heritable rather than age-related environmental and stochastic factors dominate variation in DNA methylation of the human IGF2/H19 locus. *Human Molecular Genetics.* 2007;16(5):547-554.
200. Ollikainen M, Smith KR, Joo EJ, et al. DNA methylation analysis of multiple tissues from newborn twins reveals both genetic and intrauterine components to variation in the human neonatal epigenome. *Human Molecular Genetics.* 2010;19(21):4176-4188.
201. Johannes F, Porcher E, Teixeira FK, et al. Assessing the impact of transgenerational epigenetic variation on complex traits. *PLoS Genetics.* 2009;5(6):e1000530.
202. Numachi Y, Yoshida S, Yamashita M, et al. Psychostimulant alters expression of DNA methyltransferase mRNA in the rat brain. *Annals of the New York Academy of Sciences.* 2004;1025:102-109.
203. Guidotti A, Dong E, Kundakovic M, Satta R, Grayson DR, Costa E. Characterization of the action of antipsychotic subtypes on valproate-induced chromatin remodeling. *Trends in Pharmacological Sciences.* 2009;30(2):55-60.
204. Dong E, Chen Y, Gavin DP, Grayson DR, Guidotti A. Valproate induces DNA demethylation in nuclear extracts from adult mouse brain. *Epigenetics.* 2010;5(8):730-735.
205. Kwon B, Houpt TA. Phospho-acetylation of histone H3 in the amygdala after acute lithium chloride. *Brain Research.* 2010;1333:36-47.
206. McColl G, Killilea DW, Hubbard AE, Vantipalli MC, Melov S, Lithgow GJ. Pharmacogenetic analysis of lithium-induced delayed aging in Caenorhabditis elegans. *Journal of Biological Chemistry.* 2008;283(1):350-357.
207. Soldi M, Bremang M, Bonaldi T. Biochemical systems approaches for the analysis of histone modification readout. *Biochimica et Biophysica Acta.* 2014;1839(8):657-668.
208. Bock C. Analysing and interpreting DNA methylation data. *Nature Reviews: Genetics.* 2012;13(10):705-719.

209. Laird PW. Principles and challenges of genomewide DNA methylation analysis. *Nature Reviews: Genetics.* 2010;11(3):191-203.
210. Serre D, Lee BH, Ting AH. MBD-isolated Genome Sequencing provides a high-throughput and comprehensive survey of DNA methylation in the human genome. *Nucleic Acids Research.* 2010;38(2):391-399.
211. Hsu HK, Weng YI, Hsu PY, Huang TH, Huang YW. Detection of DNA methylation by MeDIP and MBDCap assays: an overview of techniques. *Methods in Molecular Biology.* 2014;1105:61-70.
212. Nair SS, Coolen MW, Stirzaker C, et al. 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.* 2011;6(1):34-44.
213. Schumacher A, Kapranov P, Kaminsky Z, et al. Microarray-based DNA methylation profiling: technology and applications. *Nucleic Acids Research.* 2006;34(2):528-542.
214. Clark SJ, Harrison J, Paul CL, Frommer M. High sensitivity mapping of methylated cytosines. *Nucleic Acids Research.* 1994;22(15):2990-2997.
215. Cokus SJ, Feng S, Zhang X, et al. Shotgun bisulphite sequencing of the Arabidopsis genome reveals DNA methylation patterning. *Nature.* 2008;452(7184):215-219.
216. Lister R, O'Malley RC, Tonti-Filippini J, et al. Highly integrated single-base resolution maps of the epigenome in Arabidopsis. *Cell.* 2008;133(3):523-536.
217. Huang Y, Pastor WA, Shen Y, Tahiliani M, Liu DR, Rao A. The behaviour of 5-hydroxymethylcytosine in bisulfite sequencing. *PloS One.* 2010;5(1):e8888.
218. Booth MJ, Branco MR, Ficz G, et al. Quantitative sequencing of 5-methylcytosine and 5-hydroxymethylcytosine at single-base resolution. *Science.* 2012;336(6083):934-937.
219. Pastor WA, Aravind L, Rao A. TETonic shift: biological roles of TET proteins in DNA demethylation and transcription. *Nature Reviews: Molecular Cell Biology.* 2013;14(6):341-356.
220. Ziller MJ, Gu H, Muller F, et al. Charting a dynamic DNA methylation landscape of the human genome. *Nature.* 2013;500(7463):477-481.
221. Bibikova M, Barnes B, Tsan C, et al. High density DNA methylation array with single CpG site resolution. *Genomics.* 2011;98(4):288-295.
222. Deng J, Shoemaker R, Xie B, et al. Targeted bisulfite sequencing reveals changes in DNA methylation associated with nuclear reprogramming. *Nature Biotechnology.* 2009;27(4):353-360.
223. Diep D, Plongthongkum N, Gore A, Fung HL, Shoemaker R, Zhang K. Library-free methylation sequencing with bisulfite padlock probes. *Nature Methods.* 2012;9(3):270-272.
224. Abdolmaleky HM, Cheng KH, Faraone SV, et al. Hypomethylation of MB-COMT promoter is a major risk factor for schizophrenia and bipolar disorder. *Human Molecular Genetics.* 2006;15(21):3132-3145.
225. Iwamoto K, Bundo M, Yamada K, et al. DNA methylation status of SOX10 correlates with its downregulation and oligodendrocyte dysfunction in schizophrenia. *Journal of Neuroscience.* 2005;25(22):5376-5381.
226. Grayson DR, Jia X, Chen Y, et al. Reelin promoter hypermethylation in schizophrenia. *Proceedings of the National Academy of Sciences of the United States of America.* 2005;102(26):9341-9346.
227. Carrard A, Salzmann A, Malafosse A, Karege F. Increased DNA methylation status of the serotonin receptor 5HTR1A gene promoter in schizophrenia and bipolar disorder. *Journal of Affective Disorders.* 2011;132(3):450-453.

228. Tolosa A, Sanjuan J, Dagnall AM, Molto MD, Herrero N, de Frutos R. FOXP2 gene and language impairment in schizophrenia: association and epigenetic studies. *BMC Medical Genetics.* 2010;11:114.
229. Tochigi M, Iwamoto K, Bundo M, et al. Methylation status of the reelin promoter region in the brain of schizophrenic patients. *Biological Psychiatry.* 2008;63(5):530-533.
230. Dempster EL, Mill J, Craig IW, Collier DA. The quantification of COMT mRNA in post mortem cerebellum tissue: diagnosis, genotype, methylation and expression. *BMC Medical Genetics.* 2006;7:10.
231. Mill J, Tang T, Kaminsky Z, et al. Epigenomic profiling reveals DNA-methylation changes associated with major psychosis. *American Journal of Human Genetics.* 2008;82(3):696-711.
232. Benes FM, Berretta S. GABAergic interneurons: implications for understanding schizophrenia and bipolar disorder. *Neuropsychopharmacology.* 2001;25(1):1-27.
233. Coyle JT. The GABA-glutamate connection in schizophrenia: which is the proximate cause? *Biochemical Pharmacology.* 2004;68(8):1507-1514.
234. Veldic M, Guidotti A, Maloku E, Davis JM, Costa E. In psychosis, cortical interneurons overexpress DNA-methyltransferase 1. *Proceedings of the National Academy of Sciences of the United States of America.* 2005;102(6):2152-2157.
235. Zhubi A, Veldic M, Puri NV, et al. An upregulation of DNA-methyltransferase 1 and 3a expressed in telencephalic GABAergic neurons of schizophrenia patients is also detected in peripheral blood lymphocytes. *Schizophrenia Research.* 2009;111(1-3):115-122.
236. Guidotti A, Ruzicka W, Grayson DR, et al. S-adenosyl methionine and DNA methyltransferase-1 mRNA overexpression in psychosis. *Neuroreport.* 2007;18(1):57-60.
237. Huang HS, Matevossian A, Whittle C, et al. Prefrontal dysfunction in schizophrenia involves mixed-lineage leukemia 1-regulated histone methylation at GABAergic gene promoters. *Journal of Neuroscience.* 2007;27(42):11254-11262.
238. Torrey EF, Barci BM, Webster MJ, Bartko JJ, Meador-Woodruff JH, Knable MB. Neurochemical markers for schizophrenia, bipolar disorder, and major depression in postmortem brains. *Biological Psychiatry.* 2005;57(3):252-260.
239. Gavin DP, Rosen C, Chase K, Grayson DR, Tun N, Sharma RP. Dimethylated lysine 9 of histone 3 is elevated in schizophrenia and exhibits a divergent response to histone deacetylase inhibitors in lymphocyte cultures. *Journal of Psychiatry and Neuroscience.* 2009;34(3):232-237.
240. Sharma RP, Grayson DR, Gavin DP. Histone deactylase 1 expression is increased in the prefrontal cortex of schizophrenia subjects: analysis of the National Brain Databank microarray collection. *Schizophrenia Research.* 2008;98(1-3):111-117.
241. Benes FM, Lim B, Matzilevich D, Walsh JP, Subburaju S, Minns M. Regulation of the GABA cell phenotype in hippocampus of schizophrenics and bipolars. *Proceedings of the National Academy of Sciences of the United States of America.* 2007;104(24):10164-10169.
242. Torrey EF, Webster M, Knable M, Johnston N, Yolken RH. The stanley foundation brain collection and neuropathology consortium. *Schizophrenia Research.* 2000;44(2):151-155.
243. Kaminsky Z, Petronis A. Methylation SNaPshot: a method for the quantification of site-specific DNA methylation levels. *Methods in Molecular Biology.* 2009;507:241-255.
244. Huang LH, Farnet CM, Ehrlich KC, Ehrlich M. Digestion of highly modified bacteriophage DNA by restriction endonucleases. *Nucleic Acids Research.* 1982;10(5):1579-1591.

245. Josse J, Kornberg A. Glucosylation of deoxyribonucleic acid. III. alpha- and beta-Glucosyl transferases from T4-infected Escherichia coli. *Journal of Biological Chemistry.* 1962;237:1968-1976.
246. Lariviere L, Kurzeck J, Aschke-Sonnenborn U, Ruger W, Morera S. Crystallization and preliminary crystallographic study of a ternary complex between the T4 phage beta-glucosyltransferase, uridine diphosphoglucose and a DNA fragment containing an abasic site. *Acta Crystallographica. Section D: Biological Crystallography.* 2002;58(Pt 9):1484-1486.
247. Ichiyanagi K. Inhibition of *Msp*I cleavage activity by hydroxymethylation of the CpG site: a concern for DNA modification studies using restriction endonucleases. *Epigenetics.* 2012;7(2):131-136.
248. Gentleman RC, Carey VJ, Bates DM, et al. Bioconductor: open software development for computational biology and bioinformatics. *Genome Biology.* 2004;5(10):R80.
249. Morgan M, Anders S, Lawrence M, Aboyoun P, Pages H, Gentleman R. ShortRead: a bioconductor package for input, quality assessment and exploration of high-throughput sequence data. *Bioinformatics.* 2009;25(19):2607-2608.
250. Pages H, Aboyoun P, Gentleman R, DebRoy S. Biostrings: String objects representing biological sequences, and matching algorithms.R package version 2.36.1.
251. Le S, Josse J, Husson F. FactoMineR: An R Package for Multivariate Analysis. *Journal of Statistical Software.* 2008;25(1):1-18.
252. Kaplan N, Moore IK, Fondufe-Mittendorf Y, et al. The DNA-encoded nucleosome organization of a eukaryotic genome. *Nature.* 2009;458(7236):362-366.
253. Xi L, Fondufe-Mittendorf Y, Xia L, Flatow J, Widom J, Wang JP. Predicting nucleosome positioning using a duration Hidden Markov Model. *BMC Bioinformatics.* 2010;11:346.
254. Potter DP, Yan P, Huang TH, Lin S. Probe signal correction for differential methylation hybridization experiments. *BMC Bioinformatics.* 2008;9:453.
255. Benjamini Y, Yekutieli D. The control of the false discovery rate in multiple testing under dependency. *Annals of Statistics.* 2001;29:1165–1188.
256. Ritchie ME, Phipson B, Wu D, et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Research.* 2015;43(7):e47.
257. Huang da W, Sherman BT, Lempicki RA. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nature Protocols.* 2009;4(1):44-57.
258. Cahoy JD, Emery B, Kaushal A, et al. A transcriptome database for astrocytes, neurons, and oligodendrocytes: a new resource for understanding brain development and function. *Journal of Neuroscience.* 2008;28(1):264-278.
259. Lattin JE, Schroder K, Su AI, et al. Expression analysis of G Protein-Coupled Receptors in mouse macrophages. *Immunome Research.* 2008;4:5.
260. Edgar R, Domrachev M, Lash AE. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Research.* 2002;30(1):207-210.
261. Davis S, Meltzer PS. GEOquery: a bridge between the Gene Expression Omnibus (GEO) and BioConductor. *Bioinformatics.* 2007;23(14):1846-1847.
262. Gribov A, Sill M, Luck S, et al. SEURAT: visual analytics for the integrated analysis of microarray data. *BMC Medical Genomics.* 2010;3:21.
263. Brawand D, Soumillon M, Necsulea A, et al. The evolution of gene expression levels in mammalian organs. *Nature.* 2011;478(7369):343-348.
264. Shi J, Levinson DF, Duan J, et al. Common variants on chromosome 6p22.1 are associated with schizophrenia. *Nature.* 2009;460(7256):753-757.
265. Stefansson H, Ophoff RA, Steinberg S, et al. Common variants conferring risk of schizophrenia. *Nature.* 2009;460(7256):744-747.

266. Seidl C, Port M, Apostolidis C, et al. Differential gene expression triggered by highly cytotoxic alpha-emitter-immunoconjugates in gastric cancer cells. *Investigational New Drugs.* 2010;28(1):49-60.
267. Mavrich TN, Ioshikhes IP, Venters BJ, et al. A barrier nucleosome model for statistical positioning of nucleosomes throughout the yeast genome. *Genome Research.* 2008;18(7):1073-1083.
268. Gabdank I, Barash D, Trifonov EN. FineStr: a web server for single-base-resolution nucleosome positioning. *Bioinformatics.* 2010;26(6):845-846.
269. Zhu H, Chang Q, Gao Y, et al. Differences in DNA methylation between human neuronal and glial cells are concentrated in enhancers and non-CpG sites. *Nature Neuroscience.* 2014;42(1):109-127.
270. Crooks GE, Hon G, Chandonia JM, Brenner SE. WebLogo: a sequence logo generator. *Genome Research.* 2004;14(6):1188-1190.
271. Guintivano J, Aryee MJ, Kaminsky ZA. A cell epigenotype specific model for the correction of brain cellular heterogeneity bias and its application to age, brain region and major depression. *Epigenetics.* 2013;8(3):290-302.
272. Nestor CE, Ottaviano R, Reddington J, et al. Tissue type is a major modifier of the 5-hydroxymethylcytosine content of human genes. *Genome Research.* 2012;22(3):467-477.
273. Johnson WE, Li W, Meyer CA, et al. Model-based analysis of tiling-arrays for ChIP-chip. *Proceedings of the National Academy of Sciences of the United States of America.* 2006;103(33):12457-12462.
274. Balsa-Canto E, Banga JR. AMIGO, a toolbox for advanced model identification in systems biology using global optimization. *Bioinformatics.* 2011;27(16):2311-2313.
275. Chodavarapu RK, Feng S, Bernatavichute YV, et al. Relationship between nucleosome positioning and DNA methylation. *Nature.* 2010;466(7304):388-392.
276. Hart C, Lipson D, Ozsolak F, et al. Single-molecule sequencing: sequence methods to enable accurate quantitation. *Methods in Enzymology.* 2010;472:407-430.
277. Ou JN, Torrisani J, Unterberger A, et al. Histone deacetylase inhibitor Trichostatin A induces global and gene-specific DNA demethylation in human cancer cell lines. *Biochemical Pharmacology.* 2007;73(9):1297-1307.
278. Hnilicova J, Hozeifi S, Duskova E, Icha J, Tomankova T, Stanek D. Histone deacetylase activity modulates alternative splicing. *PloS One.* 2011;6(2):e16727.
279. Tardy-Planechaud S, Fujimoto J, Lin SS, Sowers LC. Solid phase synthesis and restriction endonuclease cleavage of oligodeoxynucleotides containing 5-(hydroxymethyl)-cytosine. *Nucleic Acids Research.* 1997;25(3):553-559.
280. McClelland M, Nelson M, Raschke E. Effect of site-specific modification on restriction endonucleases and DNA modification methyltransferases. *Nucleic Acids Research.* 1994;22(17):3640-3659.
281. Busslinger M, deBoer E, Wright S, Grosveld FG, Flavell RA. The sequence GGCmCGG is resistant to MspI cleavage. *Nucleic Acids Research.* 1983;11(11):3559-3569.
282. Keshet E, Cedar H. Effect of CpG methylation on Msp I. *Nucleic Acids Research.* 1983;11(11):3571-3580.
283. Frans EM, Sandin S, Reichenberg A, Lichtenstein P, Langstrom N, Hultman CM. Advancing paternal age and bipolar disorder. *Archives of General Psychiatry.* 2008;65(9):1034-1040.
284. Menezes PR, Lewis G, Rasmussen F, et al. Paternal and maternal ages at conception and risk of bipolar affective disorder in their offspring. *Psychological Medicine.* 2010;40(3):477-485.

285. Crow TJ. Mutation and psychosis: a suggested explanation of seasonality of birth. *Psychological Medicine.* 1987;17(4):821-828.
286. Tycko B. Allele-specific DNA methylation: beyond imprinting. *Human Molecular Genetics.* 2010;19(R2):R210-220.
287. Meaburn EL, Schalkwyk LC, Mill J. Allele-specific methylation in the human genome: implications for genetic studies of complex disease. *Epigenetics.* 2010;5(7):578-582.
288. Morgan HD, Sutherland HG, Martin DI, Whitelaw E. Epigenetic inheritance at the agouti locus in the mouse. *Nature Genetics.* 1999;23(3):314-318.
289. Dulac C. Brain function and chromatin plasticity. *Nature.* 2010;465(7299):728-735.
290. Ehrlich M, Wang RY. 5-Methylcytosine in eukaryotic DNA. *Science.* 1981;212(4501):1350-1357.
291. Brown DD. Gene expression in eukaryotes. *Science.* 1981;211(4483):667-674.
292. Razin A, Riggs AD. DNA methylation and gene function. *Science.* 1980;210(4470):604-610.
293. Schwartz S, Ast G. Chromatin density and splicing destiny: on the cross-talk between chromatin structure and splicing. *EMBO Journal.* 2010;29(10):1629-1636.
294. Sims RJ, 3rd, Millhouse S, Chen CF, et al. Recognition of trimethylated histone H3 lysine 4 facilitates the recruitment of transcription postinitiation factors and pre-mRNA splicing. *Molecular Cell.* 2007;28(4):665-676.
295. Guo W, Chung WY, Qian M, Pellegrini M, Zhang MQ. Characterizing the strand-specific distribution of non-CpG methylation in human pluripotent cells. *Nucleic Acids Research.* 2014;42(5):3009-3016.
296. Shirane K, Toh H, Kobayashi H, et al. Mouse oocyte methylomes at base resolution reveal genome-wide accumulation of non-CpG methylation and role of DNA methyltransferases. *PLoS Genetics.* 2013;9(4):e1003439.
297. Depp CA, Jeste DV. Bipolar disorder in older adults: a critical review. *Bipolar Disord.* 2004;6(5):343-367.
298. Keshavan MS, Nasrallah HA, Tandon R. Schizophrenia, "Just the Facts" 6. Moving ahead with the schizophrenia concept: from the elephant to the mouse. *Schizophrenia Research.* 2011;127(1-3):3-13.
299. Davis T, Vaisvila R. High sensitivity 5-hydroxymethylcytosine detection in Balb/C brain tissue. *J Vis Exp.* 2011(48).
300. Jin SG, Kadam S, Pfeifer GP. Examination of the specificity of DNA methylation profiling techniques towards 5-methylcytosine and 5-hydroxymethylcytosine. *Nucleic Acids Research.* 2010;38(11):e125.
301. Kondo E, Gu Z, Horii A, Fukushige S. The thymine DNA glycosylase MBD4 represses transcription and is associated with methylated p16(INK4a) and hMLH1 genes. *Molecular and Cellular Biology.* 2005;25(11):4388-4396.
302. Valinluck V, Tsai HH, Rogstad DK, Burdzy A, Bird A, Sowers LC. Oxidative damage to methyl-CpG sequences inhibits the binding of the methyl-CpG binding domain (MBD) of methyl-CpG binding protein 2 (MeCP2). *Nucleic Acids Research.* 2004;32(14):4100-4108.
303. Feng J, Zhou Y, Campbell SL, et al. Dnmt1 and Dnmt3a maintain DNA methylation and regulate synaptic function in adult forebrain neurons. *Nature Neuroscience.* 2010;13(4):423-430.
304. Miller CA, Sweatt JD. Covalent modification of DNA regulates memory formation. *Neuron.* 2007;53(6):857-869.
305. Szulwach KE, Li X, Li Y, et al. 5-hmC-mediated epigenetic dynamics during postnatal neurodevelopment and aging. *Nature Neuroscience.* 2011;14(12):1607-1616.

306. Hodges E, Smith AD, Kendall J, et al. High definition profiling of mammalian DNA methylation by array capture and single molecule bisulfite sequencing. *Genome Research.* 2009;19(9):1593-1605.
307. Kelly WK, O'Connor OA, Krug LM, et al. Phase I study of an oral histone deacetylase inhibitor, suberoylanilide hydroxamic acid, in patients with advanced cancer. *Journal of Clinical Oncology.* 2005;23(17):3923-3931.
308. Kolasinska-Zwierz P, Down T, Latorre I, Liu T, Liu XS, Ahringer J. Differential chromatin marking of introns and expressed exons by H3K36me3. *Nature Genetics.* 2009;41(3):376-381.
309. Schwartz S, Meshorer E, Ast G. Chromatin organization marks exon-intron structure. *Nature Structural & Molecular Biology.* 2009;16(9):990-995.
310. Tilgner H, Nikolaou C, Althammer S, et al. Nucleosome positioning as a determinant of exon recognition. *Nature Structural & Molecular Biology.* 2009;16(9):996-1001.
311. Luco RF, Pan Q, Tominaga K, Blencowe BJ, Pereira-Smith OM, Misteli T. Regulation of alternative splicing by histone modifications. *Science.* 2010;327(5968):996-1000.
312. Andersson R, Enroth S, Rada-Iglesias A, Wadelius C, Komorowski J. Nucleosomes are well positioned in exons and carry characteristic histone modifications. *Genome Research.* 2009;19(10):1732-1741.
313. Shukla S, Kavak E, Gregory M, et al. CTCF-promoted RNA polymerase II pausing links DNA methylation to splicing. *Nature.* 2011;479(7371):74-79.
314. Irimia M, Blencowe BJ. Alternative splicing: decoding an expansive regulatory layer. *Current Opinion in Cell Biology.* 2012;24(3):323-332.
315. Luco RF, Allo M, Schor IE, Kornblihtt AR, Misteli T. Epigenetics in alternative pre-mRNA splicing. *Cell.* 2011;144(1):16-26.
316. Zhao J, Bacolla A, Wang G, Vasquez KM. Non-B DNA structure-induced genetic instability and evolution. *Cellular and Molecular Life Sciences.* 2010;67(1):43-62.
317. Kundaje A, Meuleman W, Ernst J, et al. Integrative analysis of 111 reference human epigenomes. *Nature.* 2015;518(7539):317-330.
318. Azevedo FA, Carvalho LR, Grinberg LT, et al. Equal numbers of neuronal and nonneuronal cells make the human brain an isometrically scaled-up primate brain. *Journal of Comparative Neurology.* 2009;513(5):532-541.
319. Iwamoto K, Bundo M, Ueda J, et al. Neurons show distinctive DNA methylation profile and higher interindividual variations compared with non-neurons. *Genome Research.* 2011;21(5):688-696.
320. Zhang Y, Chen K, Sloan SA, et al. An RNA-sequencing transcriptome and splicing database of glia, neurons, and vascular cells of the cerebral cortex. *The Journal of Neuroscience.* 2014;34(36):11929-11947.
321. Luo L, Salunga RC, Guo H, et al. Gene expression profiles of laser-captured adjacent neuronal subtypes. *Nature Medicine.* 1999;5(1):117-122.
322. Schitine C, Nogaroli L, Costa MR, Hedin-Pereira C. Astrocyte heterogeneity in the brain: from development to disease. *Frontiers in Cellular Neuroscience.* 2015;9:76.
323. De Berardis D, Conti CM, Serroni N, et al. The effect of newer serotonin-noradrenalin antidepressants on cytokine production: a review of the current literature. *International Journal of Immunopathology and Pharmacology.* 2010;23(2):417-422.
324. Moncrieff J. A critique of the dopamine hypothesis of schizophrenia and psychosis. *Harvard Review of Psychiatry.* 2009;17(3):214-225.
325. Inta D, Monyer H, Sprengel R, Meyer-Lindenberg A, Gass P. Mice with genetically altered glutamate receptors as models of schizophrenia: a comprehensive review. *Neuroscience and Biobehavioral Reviews.* 2010;34(3):285-294.

326. Nifosi F, Toffanin T, Follador H, et al. Reduced right posterior hippocampal volume in women with recurrent familial pure depressive disorder. *Psychiatry Research.* 2010;184(1):23-28.
327. Ekman CJ, Lind J, Ryden E, Ingvar M, Landen M. Manic episodes are associated with grey matter volume reduction - a voxel-based morphometry brain analysis. *Acta Psychiatrica Scandinavica.* 2010;122(6):507-515.
328. Cotter D, Hudson L, Landau S. Evidence for orbitofrontal pathology in bipolar disorder and major depression, but not in schizophrenia. *Bipolar Disord.* 2005;7(4):358-369.
329. Uranova NA, Vostrikov VM, Orlovskaya DD, Rachmanova VI. Oligodendroglial density in the prefrontal cortex in schizophrenia and mood disorders: a study from the Stanley Neuropathology Consortium. *Schizophrenia Research.* 2004;67(2-3):269-275.
330. Matigian N, Windus L, Smith H, et al. Expression profiling in monozygotic twins discordant for bipolar disorder reveals dysregulation of the WNT signalling pathway. *Molecular Psychiatry.* 2007;12(9):815-825.
331. Rajkowska G, Selemon LD, Goldman-Rakic PS. Neuronal and glial somal size in the prefrontal cortex: a postmortem morphometric study of schizophrenia and Huntington disease. *Archives of General Psychiatry.* 1998;55(3):215-224.
332. Selemon LD, Rajkowska G, Goldman-Rakic PS. Abnormally high neuronal density in the schizophrenic cortex. A morphometric analysis of prefrontal area 9 and occipital area 17. *Archives of General Psychiatry.* 1995;52(10):805-818; discussion 819-820.
333. Weyer A, Schilling K. Developmental and cell type-specific expression of the neuronal marker NeuN in the murine cerebellum. *Journal of Neuroscience Research.* 2003;73(3):400-409.
334. Rakyan VK, Down TA, Balding DJ, Beck S. Epigenome-wide association studies for common human diseases. *Nature Reviews: Genetics.* 2011;12(8):529-541.
335. Davies MN, Volta M, Pidsley R, et al. Functional annotation of the human brain methylome identifies tissue-specific epigenetic variation across brain and blood. *Genome Biology.* 2012;13(6):R43.
336. Blewitt ME, Vickaryous NK, Paldi A, Koseki H, Whitelaw E. Dynamic reprogramming of DNA methylation at an epigenetically sensitive allele in mice. *PLoS Genetics.* 2006;2(4):e49.
337. Bell JT, Spector TD. A twin approach to unraveling epigenetics. *Trends in Genetics.* 2011;27(3):116-125.
338. Booth MJ, Ost TW, Beraldi D, et al. Oxidative bisulfite sequencing of 5-methylcytosine and 5-hydroxymethylcytosine. *Nature Protocols.* 2013;8(10):1841-1851.
339. Yu M, Hon GC, Szulwach KE, et al. Tet-assisted bisulfite sequencing of 5-hydroxymethylcytosine. *Nature Protocols.* 2012;7(12):2159-2170.
340. Booth MJ, Marsico G, Bachman M, Beraldi D, Balasubramanian S. Quantitative sequencing of 5-formylcytosine in DNA at single-base resolution. *Nature Chemistry.* 2014;6(5):435-440.
341. Wu H, Wu X, Shen L, Zhang Y. Single-base resolution analysis of active DNA demethylation using methylase-assisted bisulfite sequencing. *Nature Biotechnology.* 2014;32(12):1231-1240.
342. Akbarian S, Ruehl MG, Bliven E, et al. Chromatin alterations associated with down-regulated metabolic gene expression in the prefrontal cortex of subjects with schizophrenia. *Archives of General Psychiatry.* 2005;62(8):829-840.
343. Konradi C, Eaton M, MacDonald ML, Walsh J, Benes FM, Heckers S. Molecular evidence for mitochondrial dysfunction in bipolar disorder. *Archives of General Psychiatry.* 2004;61(3):300-308.

344. Middleton FA, Mirnics K, Pierri JN, Lewis DA, Levitt P. Gene expression profiling reveals alterations of specific metabolic pathways in schizophrenia. *Journal of Neuroscience.* 2002;22(7):2718-2729.
345. Knable MB, Barci BM, Webster MJ, Meador-Woodruff J, Torrey EF. Molecular abnormalities of the hippocampus in severe psychiatric illness: postmortem findings from the Stanley Neuropathology Consortium. *Molecular Psychiatry.* 2004;9(6):609-620, 544.
346. Haroutunian V, Katsel P, Dracheva S, Stewart DG, Davis KL. Variations in oligodendrocyte-related gene expression across multiple cortical regions: implications for the pathophysiology of schizophrenia. *International Journal of Neuropsychopharmacology.* 2007;10(4):565-573.
347. Duman RS, Newton SS. Epigenetic marking and neuronal plasticity. *Biological Psychiatry.* 2007;62(1):1-3.
348. Buckley NJ. Analysis of transcription, chromatin dynamics and epigenetic changes in neural genes. *Progress in Neurobiology.* 2007;83(4):195-210.
349. Abel T, Zukin RS. Epigenetic targets of HDAC inhibition in neurodegenerative and psychiatric disorders. *Current Opinion in Pharmacology.* 2008;8(1):57-64.
350. Akbarian S, Huang HS. Epigenetic regulation in human brain-focus on histone lysine methylation. *Biological Psychiatry.* 2009;65(3):198-203.
351. Wapenaar MC, Den Dunnen JT. Exon trapping. Application of a large-insert multiple-exon-trapping system. *Methods in Molecular Biology.* 2001;175:201-215.
352. Nasrin F, Rahman MA, Masuda A, Ohe K, Takeda J, Ohno K. HnRNP C, YB-1 and hnRNP L coordinately enhance skipping of human MUSK exon 10 to generate a Wnt-insensitive MuSK isoform. *Scientific Reports.* 2014;4:6841.
353. Khare T, Pal M, Petronis A. Understanding bipolar disorder: the epigenetic perspective. *Current Topics in Behavioral Neurosciences.* 2011;5:31-49.

List of publications

Pal M, Ebrahimi S, Oh G, Khare T, Zhang A, Kaminsky ZA, Wang SC, Petronis A. High Precision DNA Modification Analysis of *HCG9* in Major Psychosis. **Schizophrenia Bulletin.** 2015 Jun 15. pii: sbv079. [Epub ahead of print] PubMed PMID: 26078387.

Oh G, Wang SC, **Pal M**, Chen ZF, Khare T, Tochigi M, Ng C, Yang YA, Kwan A, Kaminsky ZA, Mill J, Gunasinghe C, Tackett JL, Gottesman II, Willemsen G, de Geus EJ, Vink JM, Slagboom PE, Wray NR, Heath AC, Montgomery GW, Turecki G, Martin NG, Boomsma DI, McGuffin P, Kustra R, Petronis A. DNA modification study of major depressive disorder: beyond locus-by-locus comparisons. **Biological Psychiatry.** 2015 Feb 1;77(3):246-55.

Khare T, Pai S, Koncevicius K, **Pal M**, Kriukiene E, Liutkeviciute Z, Irimia M, Jia P, Ptak C, Xia M, Tice R, Tochigi M, Moréra S, Nazarians A, Belsham D, Wong AH, Blencowe BJ, Wang SC, Kapranov P, Kustra R, Labrie V, Klimasauskas S, Petronis A. 5-hmC in the brain is abundant in synaptic genes and shows differences at the exon-intron boundary. **Nature Structural and Molecular Biology.** 2012 Oct;19(10):1037-43.

Kaminsky Z, Tochigi M, Jia P, **Pal M**, Mill J, Kwan A, Ioshikhes I, Vincent JB, Kennedy JL, Strauss J, Pai S, Wang SC, Petronis A. A multi-tissue analysis identifies HLA complex group 9 gene methylation differences in bipolar disorder. **Molecular Psychiatry.** 2012 Jul;17(7):728-40.

Khare T, **Pal M**, Petronis A. Understanding bipolar disorder: the epigenetic perspective. **Current Topics in Behavioral Neuroscience.** 2011;5:31-49.

Copyright Acknowledgement

Multiple sections of this thesis were taken from previously published papers authored and co-authored by me.[1-3,353] Permissions obtained for reproducing the published work are attached below. For Khare *et al.* (2011),[353] Kaminsky *et al.* (2012)[1] and Khare *et al.* (2012);[3] permissions were also obtained from the first authors, Dr. Zachary Kaminsky and Dr. Tarang Khare, and from Dr. Arturas Petronis, last author of all papers.

OXFORD UNIVERSITY PRESS LICENSE
TERMS AND CONDITIONS

Jun 19, 2015

This is a License Agreement between mrinal pal ("You") and Oxford University Press ("Oxford University Press") provided by Copyright Clearance Center ("CCC"). The license consists of your order details, the terms and conditions provided by Oxford University Press, and the payment terms and conditions.

All payments must be made in full to CCC. For payment instructions, please see information listed at the bottom of this form.

License Number	3652660608518
License date	Jun 19, 2015
Licensed content publisher	Oxford University Press
Licensed content publication	Schizophrenia Bulletin
Licensed content title	High Precision DNA Modification Analysis of HCG9 in Major Psychosis:
Licensed content author	Mrinal Pal, Sasha Ebrahimi, Gabriel Oh, Tarang Khare, Aiping Zhang, Zachary A. Kaminsky, Sun-Chong Wang, Arturas Petronis
Licensed content date	06/15/2015
Type of Use	Thesis/Dissertation
Institution name	None
Title of your work	the HLA complex group 9 (non-protein coding) gene modification study in major psychosis
Publisher of your work	n/a
Expected publication date	Jan 2016
Permissions cost	0.00 USD
Value added tax	0.00 USD
Total	0.00 USD
Total	0.00 USD
Terms and Conditions	

OXFORD UNIVERSITY PRESS LICENSE TERMS AND CONDITIONS

Jun 19, 2015

This is a License Agreement between mrinal pal ("You") and Oxford University Press ("Oxford University Press") provided by Copyright Clearance Center ("CCC"). The license consists of your order details, the terms and conditions provided by Oxford University Press, and the payment terms and conditions.

All payments must be made in full to CCC. For payment instructions, please see information listed at the bottom of this form.

License Number	3652660500477
License date	Jun 19, 2015
Licensed content publisher	Oxford University Press
Licensed content publication	Schizophrenia Bulletin
Licensed content title	High Precision DNA Modification Analysis of HCG9 in Major Psychosis:
Licensed content author	Mrinal Pal, Sasha Ebrahimi, Gabriel Oh, Tarang Khare, Aiping Zhang, Zachary A. Kaminsky, Sun-Chong Wang, Arturas Petronis
Licensed content date	06/15/2015
Type of Use	Thesis/Dissertation
Institution name	None
Title of your work	the HLA complex group 9 (non-protein coding) gene modification study in major psychosis
Publisher of your work	n/a
Expected publication date	Jan 2016
Permissions cost	0.00 USD
Value added tax	0.00 USD
Total	0.00 USD
Total	0.00 USD
Terms and Conditions	

SPRINGER LICENSE
TERMS AND CONDITIONS

Jun 27, 2015

This is a License Agreement between mrinal pal ("You") and Springer ("Springer") provided by Copyright Clearance Center ("CCC"). The license consists of your order details, the terms and conditions provided by Springer, and the payment terms and conditions.

All payments must be made in full to CCC. For payment instructions, please see information listed at the bottom of this form.

License Number	3657160551525
License date	Jun 27, 2015
Licensed content publisher	Springer
Licensed content publication	Springer eBook
Licensed content title	Understanding Bipolar Disorder: The Epigenetic Perspective
Licensed content author	Tarang Khare
Licensed content date	Jan 1, 2010
Type of Use	Thesis/Dissertation
Portion	Excerpts
Author of this Springer article	Yes and you are the sole author of the new work
Order reference number	None
Title of your thesis / dissertation	the HLA complex group 9 (non-protein coding) gene modification study in major psychosis
Expected completion date	Jan 2016
Estimated size(pages)	160
Total	0.00 USD
Terms and Conditions	

Appendices

Appendix 1. List of padlock probes used in this study.

ID	Padlock Probe Sequence
pp1	AAACTCAAACTAACTTTCAGTTGGAGGCTCATCGTTCCTATTCAGGCAGATGTTAT CGAGGTCCGACTACAAAAAAAACGAACTATAA
pp2	ACTCTATCAAAAATACTTTAAAGTTGGAGGCTCATCGTTCCTATTCAGGCAGATGT TATCGAGGTCCGACCCCTAAAAAAAAAAAAATCAA
pp3	CCCAACTCCAAAAACTCTCTTTGTTGGAGGCTCATCGTTCCTATTCAGGCAGATGTT ATCGAGGTCCGACTAATCTCACCACTACACT
pp4	TAACCAAAATCCAAATTTATTCTGTTGGAGGCTCATCGTTCCTATTCAGGCAGATG TTATCGAGGTCCGACCTTAACCACCACACAAA
pp5	ACTATAATCTCACCACTACGTTGGAGGCTCATCGTTCCTATTCAGGCAGATGTTAT CGAGGTCCGACACATAAAAAATTATTCCCCAA
pp6	TACTTACAAATTTATCCTCCTCTGTTGGAGGCTCATCGTTCCTATTCAGGCAGATGT TATCGAGGTCCGACTTCAAAAAAAATAAAAATCACA
pp7	ATACCAATAATACCACTTACATGTTGGAGGCTCATCGTTCCTATTCAGGCAGATGT TATCGAGGTCCGACTACACACCCACATATCCA
pp8	AACAAAACAAAACCCTTGTTGGAGGCTCATCGTTCCTATTCAGGCAGATGTTATCG AGGTCCGACCGTAACTATTCTTTAAAAAACTA
pp9	ATCCCCCCATCCTTCGTTGGAGGCTCATCGTTCCTATTCAGGCAGATGTTATCGAG GTCCGACAACTCTAAAATAAAAATAATAAACC
pp10	CTAAAATACATAAAAATCAAATACAGTTGGAGGCTCATCGTTCCTATTCAGGCAGA TGTTATCGAGGTCCGACCGCGCCTTCAATAAA
pp11	TCRAATCCTAAAATCAATAAGTTGGAGGCTCATCGTTCCTATTCAGGCAGATGTTA TCGAGGTCCGACATAACCTTCGCTCTAATCTT
pp12	CCCRCAAAAAAACTCACAACGTTGGAGGCTCATCGTTCCTATTCAGGCAGATGTTA TCGAGGTCCGACCTAAAACGCGATCTAAAATC
pp13	TTTCTATATCAATAAACACTTATGTTGGAGGCTCATCGTTCCTATTCAGGCAGATGT TATCGAGGTCCGACCAATTCCCTACTCACCC
pp14	CACACCTACCGCTATAAAAATACAAGTTGGAGGCTCATCGTTCCTATTCAGGCAGA TGTTATCGAGGTCCGACACTACGCTAAACRCC
pp15	ACTAATAAATCCAAAAATCTAAAGTTGGAGGCTCATCGTTCCTATTCAGGCAGATG TTATCGAGGTCCGACTTTCCGTAAACCAAAAC
pp16	TATATAACTCAAAACGCGACGTTGGAGGCTCATCGTTCCTATTCAGGCAGATGTTA TCGAGGTCCGACATCTCCCAAATCTTTTCTATCCA
pp17	TTAAAACAAACTTTACATATTATTGTTGGAGGCTCATCGTTCCTATTCAGGCAGAT GTTATCGAGGTCCGACAACTCAAAACAAAAAAAACTAAA
pp18	AATCTTAAACTTTAAATAAATCGGTTGGAGGCTCATCGTTCCTATTCAGGCAGATG TTATCGAGGTCCGACACTCTCTCGAAAACTTT
pp19	CCCCTCACTACCCTTATGTTGGAGGCTCATCGTTCCTATTCAGGCAGATGTTATCGA GGTCCGACCTAAATACGTTAAAACAAACTTT
pp20	TTCTCCCTATACCAATTCGTTGGAGGCTCATCGTTCCTATTCAGGCAGATGTTATCG AGGTCCGACTCCACATTCACTAATTATTTAT
pp21	TAAAACTAACGAAAATAAAAAAAGTTGGAGGCTCATCGTTCCTATTCAGGCAGAT GTTATCGAGGTCCGACCAAAACCCCAATCCCTA
pp22	CCAACCCCCATTAAACAATAAACCAGTTGGAGGCTCATCGTTCCTATTCAGGCAGA TGTTATCGAGGTCCGACCACCTCGCTACACAT

pp23	ACTTACTCTCCTCACCTCGTTGGAGGCTCATCGTTCCTATTCAGGCAGATGTTATCG AGGTCCGACACTTAACGTTTCTAAAAAATC
pp24	ACTAAAAAACAAAACTAAAAAATGTTGGAGGCTCATCGTTCCTATTCAGGCAGAT GTTATCGAGGTCCGACACTACCTCAACGACAAAAC
pp25	ACCAAAAACTAATACTTCCGTTGGAGGCTCATCGTTCCTATTCAGGCAGATGTTAT CGAGGTCCGACCTAAAATCCCTAAATTCCTAA
pp26	ACAAAACTTCAAAAACCCGTTGGAGGCTCATCGTTCCTATTCAGGCAGATGTTATC GAGGTCCGACATAAACTAAAAATAAAAACAATC
pp27	TTCCTTTCTCATAACCCGTTGGAGGCTCATCGTTCCTATTCAGGCAGATGTTATCGA GGTCCGACCTATAAAACTCCTAACTCTCAAA
pp28	ACCCTACAAATAAATCAAAAAAGTTGGAGGCTCATCGTTCCTATTCAGGCAGATGT TATCGAGGTCCGACCCAAAACTCCCCTTAAAT
pp29	TCCCCACTCCTTCATGTTGGAGGCTCATCGTTCCTATTCAGGCAGATGTTATCGAGG TCCGACCATTCCCAACTATCCCTAACCCTAA
pp30	CCTAAAACCCAACTTTCGTTGGAGGCTCATCGTTCCTATTCAGGCAGATGTTATCG AGGTCCGACAAATAAAAACAACATAAAATCCC
pp31	ATTCTCCCAAAATCACCCCCCCTTGTTGGAGGCTCATCGTTCCTATTCAGGCAGAT GTTATCGAGGTCCGACTCTTAAAAACCCCACC
pp32	ATCCCCTCCTCTAAACCTTGTTGGAGGCTCATCGTTCCTATTCAGGCAGATGTTATC GAGGTCCGACTCACCCCCA CTTCCTTC
pp33	AACACCACTCCTCTCTAAAAACCGTTGGAGGCTCATCGTTCCTATTCAGGCAGATG TTATCGAGGTCCGACCTCCTTAACCCTCATCC
pp34	ATAACCCTAAAAAAACTTAAAATGTTGGAGGCTCATCGTTCCTATTCAGGCAGATG TTATCGAGGTCCGACAAATAACCACAACCCAC
pp35	CACCCTTTAATCCTTACCGTTGGAGGCTCATCGTTCCTATTCAGGCAGATGTTATCG AGGTCCGACATACTAAACCATCCCTCCCTAA
pp36	CAACACTAAATAAAAACTACTGTTGGAGGCTCATCGTTCCTATTCAGGCAGATGTT ATCGAGGTCCGACCCCAAAACAAACATCTACA
pp37	ACTAATAAAAATTAAACCTTAAACGTTGGAGGCTCATCGTTCCTATTCAGGCAGAT GTTATCGAGGTCCGACTCTACCCTCATCCCCT
pp38	CTCCAACCCCTCAACTTGTTGGAGGCTCATCGTTCCTATTCAGGCAGATGTTATCG AGGTCCGACAACAAAAAAAACAACAAAACCCA
pp39	TCCCAAAATTCTTTCCTCTCTATCGTTGGAGGCTCATCGTTCCTATTCAGGCAGATG TTATCGAGGTCCGACTATTCTCAACTCCAAAAC
pp40	ACCTATAATTACTAACCATTCCGTTGGAGGCTCATCGTTCCTATTCAGGCAGATGTT ATCGAGGTCCGACCCGCTCTAACCGAAAACTAA
pp41	TACTTCAATAACGATAATACTGTTGGAGGCTCATCGTTCCTATTCAGGCAGATGTT ATCGAGGTCCGACCCCCACCTTAAAAAAATTA
pp42	TCCTTTACTAAAACTAACACGTTGGAGGCTCATCGTTCCTATTCAGGCAGATGTTA TCGAGGTCCGACACCAAACCTAAACCCTAAAA
pp43	TCTCAAAACTACGACCCGTTGGAGGCTCATCGTTCCTATTCAGGCAGATGTTATCG AGGTCCGACCCTCCAACTCCTTCTACTCTAAA
pp44	CCTACAATCTACGAAAAAAGTTGGAGGCTCATCGTTCCTATTCAGGCAGATGTTAT CGAGGTCCGACCGTTATTCCGTTTCTATATCA
pp45	CCCACTAATTCCAAAATCCGTTGGAGGCTCATCGTTCCTATTCAGGCAGATGTTAT CGAGGTCCGACAAAAACCGAACTACCTTCAAA
pp46	CTCACCCGACTCCGAATCCACGTTGGAGGCTCATCGTTCCTATTCAGGCAGATGTT ATCGAGGTCCGACATCCTCACTAAAAAC
pp47	TCACAAAATACGAAAACACGTTGGAGGCTCATCGTTCCTATTCAGGCAGATGTTAT CGAGGTCCGACCCGACCTCTAAACGATTCTAA
pp48	ATTAAAATCCTCACTAAAAACGTTGGAGGCTCATCGTTCCTATTCAGGCAGATGTT ATCGAGGTCCGACCTCTACATTTAATCCCCAT

pp49	TATTATTCTCACCTTCACCTTTGTTGGAGGCTCATCGTTCCTATTCAGGCAGATGTT ATCGAGGTCCGACCACCCTAACACATCTAAA
pp50	AAAATTCCACGCAATCGTTGGAGGCTCATCGTTCCTATTCAGGCAGATGTTATCGA GGTCCGACCAAAATCAAAATACAAATCTCAAA
pp51	ACCAAACAATACAACTCAGTTGGAGGCTCATCGTTCCTATTCAGGCAGATGTTATC GAGGTCCGACATAATTCCATATATCCTCRTCC
pp52	ACCCCAAAAACTAAAATAAGTTGGAGGCTCATCGTTCCTATTCAGGCAGATGTTAT CGAGGTCCGACATAAAATTCCACGCAATCCTT
pp53	TTTCCCCTACTTAAAATATATATGTTGGAGGCTCATCGTTCCTATTCAGGCAGATGT TATCGAGGTCCGACACAACCTCGAACTCCTA
pp54	ATAAAAATAAACCACACACCCAGTTGGAGGCTCATCGTTCCTATTCAGGCAGATGT TATCGAGGTCCGACTCCCAACTCTAAAAACTA
pp55	TATCACCCAAACTAAAATATAGTTGGAGGCTCATCGTTCCTATTCAGGCAGATGTT ATCGAGGTCCGACAACCCCCTTAAAAAAAAAT
pp56	TAAAAAAAAAAAAACAAAATCCAGTTGGAGGCTCATCGTTCCTATTCAGGCAGAT GTTATCGAGGTCCGACAAAATAAACCACACACC
pp57	TAAAAATATAAAACCCCCTTAAGTTGGAGGCTCATCGTTCCTATTCAGGCAGATGT TATCGAGGTCCGACACTTTATCTCAACCCAAA

Appendix 2. List of barcoded pp_Ramp primers used to multiplex samples during padlock probe capture.

ID	Primer
Ind1	CAAGCAGAAGACGGCATACGAGATCGTGATGCTAGGAACGATGAGCCTCCAAC
Ind2	CAAGCAGAAGACGGCATACGAGATACATCGGCTAGGAACGATGAGCCTCCAAC
Ind3	CAAGCAGAAGACGGCATACGAGATGCCTAAGCTAGGAACGATGAGCCTCCAAC
Ind4	CAAGCAGAAGACGGCATACGAGATTGGTCAGCTAGGAACGATGAGCCTCCAAC
Ind5	CAAGCAGAAGACGGCATACGAGATCACTGTGCTAGGAACGATGAGCCTCCAAC
Ind6	CAAGCAGAAGACGGCATACGAGATATTGGCGCTAGGAACGATGAGCCTCCAAC
Ind7	CAAGCAGAAGACGGCATACGAGATGATCTGGCTAGGAACGATGAGCCTCCAAC
Ind8	CAAGCAGAAGACGGCATACGAGATTCAAGTGCTAGGAACGATGAGCCTCCAAC
Ind9	CAAGCAGAAGACGGCATACGAGATCTGATCGCTAGGAACGATGAGCCTCCAAC
Ind10	CAAGCAGAAGACGGCATACGAGATAAGCTAGCTAGGAACGATGAGCCTCCAAC
Ind11	CAAGCAGAAGACGGCATACGAGATGTAGCCGCTAGGAACGATGAGCCTCCAAC
Ind12	CAAGCAGAAGACGGCATACGAGATTACAAGGCTAGGAACGATGAGCCTCCAAC
Ind13	CAAGCAGAAGACGGCATACGAGATTCATGGGCTAGGAACGATGAGCCTCCAAC
Ind14	CAAGCAGAAGACGGCATACGAGATTGTCTTGCTAGGAACGATGAGCCTCCAAC
Ind15	CAAGCAGAAGACGGCATACGAGATAGGAAGGCTAGGAACGATGAGCCTCCAAC
Ind16	CAAGCAGAAGACGGCATACGAGATAACCCCGCTAGGAACGATGAGCCTCCAAC
Ind17	CAAGCAGAAGACGGCATACGAGATGATGAGGCTAGGAACGATGAGCCTCCAAC
Ind18	CAAGCAGAAGACGGCATACGAGATTGAACTGCTAGGAACGATGAGCCTCCAAC
Ind19	CAAGCAGAAGACGGCATACGAGATTGCGTCGCTAGGAACGATGAGCCTCCAAC
Ind20	CAAGCAGAAGACGGCATACGAGATGACAGGGCTAGGAACGATGAGCCTCCAAC
Ind21	CAAGCAGAAGACGGCATACGAGATGGGTTGGCTAGGAACGATGAGCCTCCAAC
Ind22	CAAGCAGAAGACGGCATACGAGATTCCGAGGCTAGGAACGATGAGCCTCCAAC
Ind23	CAAGCAGAAGACGGCATACGAGATTTTCGAGCTAGGAACGATGAGCCTCCAAC
Ind24	CAAGCAGAAGACGGCATACGAGATGCGAATGCTAGGAACGATGAGCCTCCAAC
Ind25	CAAGCAGAAGACGGCATACGAGATGCAGTAGCTAGGAACGATGAGCCTCCAAC
Ind26	CAAGCAGAAGACGGCATACGAGATTCACGAGCTAGGAACGATGAGCCTCCAAC
Ind27	CAAGCAGAAGACGGCATACGAGATCGCGTAGCTAGGAACGATGAGCCTCCAAC
Ind28	CAAGCAGAAGACGGCATACGAGATGCACCTGCTAGGAACGATGAGCCTCCAAC
Ind29	CAAGCAGAAGACGGCATACGAGATGTTCGTGCTAGGAACGATGAGCCTCCAAC
Ind30	CAAGCAGAAGACGGCATACGAGATCACTAAGCTAGGAACGATGAGCCTCCAAC
Ind31	CAAGCAGAAGACGGCATACGAGATGTGGTGGCTAGGAACGATGAGCCTCCAAC
Ind32	CAAGCAGAAGACGGCATACGAGATCCTTGCGCTAGGAACGATGAGCCTCCAAC
Ind33	CAAGCAGAAGACGGCATACGAGATGTTGCGGCTAGGAACGATGAGCCTCCAAC
Ind34	CAAGCAGAAGACGGCATACGAGATTCAGTTGCTAGGAACGATGAGCCTCCAAC
Ind35	CAAGCAGAAGACGGCATACGAGATCCCGATGCTAGGAACGATGAGCCTCCAAC
Ind36	CAAGCAGAAGACGGCATACGAGATTGCTTGGCTAGGAACGATGAGCCTCCAAC
Ind37	CAAGCAGAAGACGGCATACGAGATTGTAGCGCTAGGAACGATGAGCCTCCAAC
Ind38	CAAGCAGAAGACGGCATACGAGATGCGTGAGCTAGGAACGATGAGCCTCCAAC
Ind39	CAAGCAGAAGACGGCATACGAGATCTCTGGGCTAGGAACGATGAGCCTCCAAC
Ind40	CAAGCAGAAGACGGCATACGAGATCCGTCAGCTAGGAACGATGAGCCTCCAAC
Ind41	CAAGCAGAAGACGGCATACGAGATGTTCCCGCTAGGAACGATGAGCCTCCAAC

Ind42	CAAGCAGAAGACGGCATACGAGATCTTTTCGCTAGGAACGATGAGCCTCCAAC
Ind43	CAAGCAGAAGACGGCATACGAGATGGCACTGCTAGGAACGATGAGCCTCCAAC
Ind44	CAAGCAGAAGACGGCATACGAGATGGATGCGCTAGGAACGATGAGCCTCCAAC
Ind45	CAAGCAGAAGACGGCATACGAGATCGTAGTGCTAGGAACGATGAGCCTCCAAC
Ind46	CAAGCAGAAGACGGCATACGAGATGAAATGGCTAGGAACGATGAGCCTCCAAC
Ind47	CAAGCAGAAGACGGCATACGAGATGGAGAGGCTAGGAACGATGAGCCTCCAAC
Ind48	CAAGCAGAAGACGGCATACGAGATTAACGTGCTAGGAACGATGAGCCTCCAAC
Ind49	CAAGCAGAAGACGGCATACGAGATACACAGGCTAGGAACGATGAGCCTCCAAC
Ind50	CAAGCAGAAGACGGCATACGAGATAAAGGTGCTAGGAACGATGAGCCTCCAAC
Ind51	CAAGCAGAAGACGGCATACGAGATGCGATAGCTAGGAACGATGAGCCTCCAAC
Ind52	CAAGCAGAAGACGGCATACGAGATCGTGTCGCTAGGAACGATGAGCCTCCAAC
Ind53	CAAGCAGAAGACGGCATACGAGATGTAGAAGCTAGGAACGATGAGCCTCCAAC
Ind54	CAAGCAGAAGACGGCATACGAGATGGACGTGCTAGGAACGATGAGCCTCCAAC
Ind55	CAAGCAGAAGACGGCATACGAGATAGTCGAGCTAGGAACGATGAGCCTCCAAC
Ind56	CAAGCAGAAGACGGCATACGAGATGTCTGAGCTAGGAACGATGAGCCTCCAAC
Ind57	CAAGCAGAAGACGGCATACGAGATGAAGGAGCTAGGAACGATGAGCCTCCAAC
Ind58	CAAGCAGAAGACGGCATACGAGATATGCTGGCTAGGAACGATGAGCCTCCAAC
Ind59	CAAGCAGAAGACGGCATACGAGATTCTATCGCTAGGAACGATGAGCCTCCAAC
Ind60	CAAGCAGAAGACGGCATACGAGATATCTGTGCTAGGAACGATGAGCCTCCAAC
Ind61	CAAGCAGAAGACGGCATACGAGATATAGAGGCTAGGAACGATGAGCCTCCAAC
Ind62	CAAGCAGAAGACGGCATACGAGATGCTAAAGCTAGGAACGATGAGCCTCCAAC
Ind63	CAAGCAGAAGACGGCATACGAGATACCAGGGCTAGGAACGATGAGCCTCCAAC
Ind64	CAAGCAGAAGACGGCATACGAGATCCAACTGCTAGGAACGATGAGCCTCCAAC
Ind65	CAAGCAGAAGACGGCATACGAGATAAGGAAGCTAGGAACGATGAGCCTCCAAC
Ind66	CAAGCAGAAGACGGCATACGAGATCCTCCAGCTAGGAACGATGAGCCTCCAAC
Ind67	CAAGCAGAAGACGGCATACGAGATCACGTCGCTAGGAACGATGAGCCTCCAAC
Ind68	CAAGCAGAAGACGGCATACGAGATCATAACGCTAGGAACGATGAGCCTCCAAC
Ind69	CAAGCAGAAGACGGCATACGAGATCCATATGCTAGGAACGATGAGCCTCCAAC
Ind70	CAAGCAGAAGACGGCATACGAGATGAAGTCGCTAGGAACGATGAGCCTCCAAC
Ind71	CAAGCAGAAGACGGCATACGAGATCAAAGAGCTAGGAACGATGAGCCTCCAAC
Ind72	CAAGCAGAAGACGGCATACGAGATTGGCAGGCTAGGAACGATGAGCCTCCAAC
Ind73	CAAGCAGAAGACGGCATACGAGATGAGTCCGCTAGGAACGATGAGCCTCCAAC
Ind74	CAAGCAGAAGACGGCATACGAGATTCGCCAGCTAGGAACGATGAGCCTCCAAC
Ind75	CAAGCAGAAGACGGCATACGAGATAAGTCGGCTAGGAACGATGAGCCTCCAAC
Ind76	CAAGCAGAAGACGGCATACGAGATAATAGGGCTAGGAACGATGAGCCTCCAAC
Ind77	CAAGCAGAAGACGGCATACGAGATACCCGTGCTAGGAACGATGAGCCTCCAAC
Ind78	CAAGCAGAAGACGGCATACGAGATAACACGGCTAGGAACGATGAGCCTCCAAC
Ind79	CAAGCAGAAGACGGCATACGAGATGCTTGGGCTAGGAACGATGAGCCTCCAAC
Ind80	CAAGCAGAAGACGGCATACGAGATTTACCAGCTAGGAACGATGAGCCTCCAAC
Ind81	CAAGCAGAAGACGGCATACGAGATCCAGGTGCTAGGAACGATGAGCCTCCAAC
Ind82	CAAGCAGAAGACGGCATACGAGATCGTTTGGCTAGGAACGATGAGCCTCCAAC
Ind83	CAAGCAGAAGACGGCATACGAGATGACCACGCTAGGAACGATGAGCCTCCAAC
Ind84	CAAGCAGAAGACGGCATACGAGATACAAGAGCTAGGAACGATGAGCCTCCAAC
Ind85	CAAGCAGAAGACGGCATACGAGATACCGCAGCTAGGAACGATGAGCCTCCAAC
Ind86	CAAGCAGAAGACGGCATACGAGATTGGGTAGCTAGGAACGATGAGCCTCCAAC
Ind87	CAAGCAGAAGACGGCATACGAGATATTCCGGCTAGGAACGATGAGCCTCCAAC
Ind88	CAAGCAGAAGACGGCATACGAGATGAATGTGCTAGGAACGATGAGCCTCCAAC

Ind89	CAAGCAGAAGACGGCATACGAGATGCTGATGCTAGGAACGATGAGCCTCCAAC
Ind90	CAAGCAGAAGACGGCATACGAGATAGTGCTGCTAGGAACGATGAGCCTCCAAC
Ind91	CAAGCAGAAGACGGCATACGAGATCAGGGAGCTAGGAACGATGAGCCTCCAAC
Ind92	CAAGCAGAAGACGGCATACGAGATCATGCGGCTAGGAACGATGAGCCTCCAAC
Ind93	CAAGCAGAAGACGGCATACGAGATTGCCTAGCTAGGAACGATGAGCCTCCAAC
Ind94	CAAGCAGAAGACGGCATACGAGATCTATACGCTAGGAACGATGAGCCTCCAAC
Ind95	CAAGCAGAAGACGGCATACGAGATCCGAGTGCTAGGAACGATGAGCCTCCAAC
Ind96	CAAGCAGAAGACGGCATACGAGATACCTGCGCTAGGAACGATGAGCCTCCAAC

Appendix 3. Effects of the rs1128306 genotype on *HCG9* modification.

PFC – combined prefrontal cortex sample from SMRI and McL cohorts; WBC – white blood cell sample. FDR column: empty cells are FDR=1

Tissue	CpG	GG Mean ± SEM	GA Mean ± SEM	P value	FDR P value
PFC	1	37.8 ± 2.21	42.9 ± 2.5	0.0086	8.40E-02
PFC	2	57 ± 2.47	61.2 ± 2.89	0.051	2.70E-01
PFC	3	37.9 ± 3	41.7 ± 3.58	0.16	6.80E-01
PFC	4	40.9 ± 2.30	41.9 ± 2.48	0.22	9.00E-01
PFC	5	23.9 ± 1.09	40.8 ± 2.03	1.70E-09	9.30E-08
PFC	6	12.6 ± 0.56	23.0 ± 1.47	1.30E-08	4.80E-07
PFC	7	13.0 ± 0.75	23.6 ± 1.48	5.10E-08	1.40E-06
PFC	8	6.41 ± 0.39	17.2 ± 1.31	2.50E-11	2.70E-09
PFC	9	40.1 ± 1.68	55.4 ± 1.79	1.40E-07	3.10E-06
PFC	10	15.9 ± 1.14	18.8 ± 1.58	0.089	4.10E-01
PFC	11	9.94 ± 0.86	11.9 ± 1.10	0.1	4.40E-01
PFC	12	8.7 ± 0.64	11.0 ± 1.01	0.076	3.60E-01
PFC	13	16.0 ± 1.24	20.9 ± 1.66	0.0058	7.10E-02
PFC	14	10.0 ± 0.75	12.4 ± 1.11	0.042	2.60E-01
PFC	15	10.6 ± 0.91	13.7 ± 1.1	0.022	1.50E-01
PFC	16	15.1 ± 1.00	18.7 ± 1.46	0.022	1.50E-01
PFC	17	18.7 ± 1.21	22.9 ± 1.7	0.013	1.00E-01
PFC	18	18.7 ± 1.13	23.3 ± 1.86	0.0071	7.80E-02
PFC	19	23.9 ± 1.15	26.9 ± 1.67	0.049	2.70E-01
PFC	20	10.8 ± 0.63	12.6 ± 1.01	0.062	3.10E-01
PFC	21	8.27 ± 0.58	11 ± 0.92	0.0092	8.40E-02
PFC	22	17.4 ± 0.97	22.0 ± 1.77	0.01	8.50E-02
PFC	23	13.8 ± 0.78	18.4 ± 1.53	0.0053	7.10E-02
PFC	24	10.8 ± 0.75	14.8 ± 1.18	0.003	5.50E-02
PFC	25	11.2 ± 0.99	11.1 ± 0.86	0.5	
PFC	26	15.6 ± 1.18	17.4 ± 1.03	0.047	2.70E-01
PFC	27	12.9 ± 0.93	15.1 ± 0.94	0.038	2.50E-01
PFC	28	11.7 ± 0.79	14.6 ± 1.01	0.0037	5.80E-02
WBC	1	15.5 ± 0.6	16.4 ± 0.83	0.27	
WBC	2	36.3 ± 1.04	35.5 ± 1.23	0.92	
WBC	3	26.2 ± 0.9	30.0 ± 1.22	0.0026	0.05718
WBC	4	17.7 ± 0.55	18 ± 0.73	0.6	
WBC	5	11.0 ± 0.45	12.7 ± 0.74	0.0015	0.04124
WBC	6	6.98 ± 0.26	8.29 ± 0.56	0.0043	0.06755
WBC	7	4.33 ± 0.17	5.92 ± 0.57	0.00046	0.01686
WBC	8	2.16 ± 0.15	3.59 ± 0.54	6.80E-06	0.00075
WBC	9	19.1 ± 0.67	22.3 ± 0.84	3.00E-05	0.00165
WBC	10	4.58 ± 0.29	4.32 ± 0.42	0.55	

WBC	11	2.49 ± 0.17	2.39 ± 0.23	1	
WBC	12	2.99 ± 0.24	2.75 ± 0.29	0.7	
WBC	13	6.65 ± 0.33	8.6 ± 0.74	0.0036	0.06598
WBC	14	4.29 ± 0.29	5.22 ± 0.53	0.007	0.09622
WBC	15	3.32 ± 0.23	3.31 ± 0.26	0.15	
WBC	16	4.71 ± 0.43	4.27 ± 0.35	0.34	
WBC	17	5.72 ± 0.43	6.09 ± 0.54	0.052	0.48566
WBC	18	5.33 ± 0.27	5.19 ± 0.36	0.43	
WBC	19	8.28 ± 0.49	8.38 ± 0.6	0.14	
WBC	20	4.07 ± 0.21	4.42 ± 0.34	0.2	
WBC	21	3.36 ± 0.20	4.21 ± 0.47	0.053	0.48566
WBC	22	7.24 ± 0.30	8.7 ± 0.63	0.02	0.24436
WBC	23	4.96 ± 0.41	5.41 ± 0.47	0.025	0.2749
WBC	24	4.36 ± 0.29	4.7 ± 0.37	0.14	
WBC	25	5.26 ± 0.2	5.27 ± 0.34	0.89	
WBC	26	10.4 ± 0.33	9.67 ± 0.52	0.39	
WBC	27	8.96 ± 0.30	8.87 ± 0.35	0.45	
WBC	28	7.74 ± 0.21	8.08 ± 0.42	0.44	
Germline	1	2.23 ± 0.65	7.46 ± 2.86	0.1	
Germline	2	5.46 ± 0.85	12.2 ± 3.51	0.094	
Germline	3	3.43 ± 0.82	8.6 ± 3.35	0.67	
Germline	4	5.91 ± 0.88	9 ± 2.96	0.28	
Germline	5	4.05 ± 0.73	5.03 ± 0.82	0.38	
Germline	6	3.69 ± 0.54	6.32 ± 0.86	0.023	
Germline	7	0.78 ± 0.48	2.58 ± 0.71	0.032	
Germline	8	1.15 ± 0.25	1.34 ± 0.26	0.49	
Germline	9	2.99 ± 0.44	6.43 ± 1.72	0.27	
Germline	10	2.19 ± 0.28	2.8 ± 0.77	0.75	
Germline	11	0.92 ± 0.28	0.84 ± 0.18	0.82	
Germline	12	1.21 ± 0.11	1.54 ± 0.24	0.38	
Germline	13	0.99 ± 0.14	1.88 ± 0.66	0.38	
Germline	14	1.23 ± 0.18	1.93 ± 0.72	0.35	
Germline	15	1.01 ± 0.14	0.86 ± 0.14	0.9	
Germline	16	1.26 ± 0.074	2.04 ± 0.62	0.13	
Germline	17	1.34 ± 0.14	2.50 ± 0.67	0.042	
Germline	18	2.03 ± 0.11	3.08 ± 0.71	0.032	
Germline	19	2.53 ± 0.17	3.50 ± 0.8	0.31	
Germline	20	1.22 ± 0.17	0.92 ± 0.23	0.55	
Germline	21	0.75 ± 0.14	0.64 ± 0.17	0.56	
Germline	22	2.34 ± 0.27	2.78 ± 0.39	0.36	
Germline	23	0.63 ± 0.21	1.42 ± 0.65	0.53	
Germline	24	1.42 ± 0.4	0.98 ± 0.35	0.33	
Germline	25	4.14 ± 0.99	2.84 ± 0.5	0.84	
Germline	26	4.02 ± 1.04	3.53 ± 0.63	0.61	
Germline	27	3.88 ± 0.77	3.68 ± 0.61	0.61	

Germline	28	3.63 ± 0.7	3.22 ± 0.48	0.96

Appendix 4. Corrected DNA methylation differences between BPD and controls.

Data represented in the table are derived from age and rs1128306 genotype corrected DNA methylation levels. SMRI PFC – prefrontal cortex samples from the Stanley Medical Research Institute; McL PFC - prefrontal cortex samples from the McLean Brain Tissue bank; PFC – combined prefrontal cortex sample from SMRI and McL cohorts; WBC – white blood cell sample. Only significant tests are reported. FDR column: empty cells are FDR=1

Tissue	Window Start (CpG)	Window End (CpG)	N BPD	N Con	BPD Mean ± SEM	Con Mean ± SEM	P value	FDR P value
SMRI PFC	1	10	31	33	-3.5 ± 2.14	2.96 ± 1.75	0.019	
SMRI PFC	1	11	31	33	-3 ± 2.11	2.59 ± 1.61	0.028	
SMRI PFC	1	12	31	33	-2.7 ± 2.13	2.23 ± 1.46	0.026	
SMRI PFC	1	13	30	33	-2.4 ± 2.21	2 ± 1.38	0.029	
SMRI PFC	1	14	30	33	-2.2 ± 2.08	1.87 ± 1.29	0.034	
SMRI PFC	1	3	33	34	-5.6 ± 4.35	5.47 ± 3.46	0.024	
SMRI PFC	1	4	33	34	-5 ± 4.35	4.85 ± 3.35	0.03	
SMRI PFC	1	5	33	33	-4.5 ± 3.6	4.58 ± 2.92	0.037	
SMRI PFC	1	6	33	33	-4.1 ± 3.01	4.13 ± 2.52	0.034	
SMRI PFC	1	7	33	33	-3.9 ± 2.63	3.93 ± 2.25	0.027	
SMRI PFC	1	8	33	33	-3.5 ± 2.33	3.58 ± 2	0.030	
SMRI PFC	1	9	33	33	-3.5 ± 2.17	3.50 ± 1.97	0.032	
SMRI PFC	2	10	31	33	-3.7 ± 1.96	2.85 ± 1.67	0.015	
SMRI PFC	2	11	31	33	-3.1 ± 1.94	2.45 ± 1.51	0.019	
SMRI PFC	2	12	31	33	-2.7 ± 1.97	2.07 ± 1.36	0.028	
SMRI PFC	2	13	30	33	-2.4 ± 2.07	1.84 ± 1.28	0.028	
SMRI PFC	2	14	30	33	-2.2 ± 1.94	1.71 ± 1.19	0.034	
SMRI PFC	2	3	33	34	-6.8 ± 4.81	6.57 ± 3.97	0.024	
SMRI PFC	2	4	33	34	-5.5 ± 4.70	5.37 ± 3.53	0.031	
SMRI PFC	2	5	33	33	-4.8 ± 3.67	4.73 ± 2.94	0.029	
SMRI PFC	2	6	33	33	-4.2 ± 2.96	4.16 ± 2.46	0.033	
SMRI PFC	2	7	33	33	-3.9 ± 2.54	3.92 ± 2.17	0.024	
SMRI PFC	2	8	33	33	-3.5 ± 2.22	3.53 ± 1.9	0.027	
SMRI PFC	2	9	33	33	-3.5 ± 2.08	3.45 ± 1.90	0.032	
SMRI PFC	22	24	23	26	-3 ± 2.84	1.30 ± 1.79	0.0086	
SMRI PFC	22	25	17	22	-0.58 ± 2.71	2.41 ± 1.22	0.017	
SMRI PFC	23	23	29	29	-0.88 ± 3.39	0.88 ± 2.02	0.026	
SMRI PFC	23	24	23	26	-2.3 ± 3.27	1.31 ± 1.68	0.0032	
SMRI PFC	23	25	17	22	0.033 ± 2.84	2.02 ± 1.01	0.0092	
SMRI PFC	24	24	25	26	-1.3 ± 3.33	1.26 ± 1.64	0.018	
SMRI PFC	3	10	31	33	-3.1 ± 1.76	2.59 ± 1.53	0.017	
SMRI PFC	3	11	31	33	-2.5 ± 1.76	2.18 ± 1.37	0.019	
SMRI PFC	3	12	31	33	-2.2 ± 1.81	1.78 ± 1.22	0.025	

SMRI PFC	3	13	30	33	-1.9 ± 1.93	1.55 ± 1.14	0.039	
SMRI PFC	3	14	30	33	-1.7 ± 1.80	1.44 ± 1.05	0.035	
SMRI PFC	3	3	33	34	-7.6 ± 5.05	7.38 ± 4.31	0.014	
SMRI PFC	3	4	33	34	-5.3 ± 4.85	5.18 ± 3.58	0.021	
SMRI PFC	3	5	33	33	-4.4 ± 3.45	4.66 ± 2.76	0.029	
SMRI PFC	3	6	33	33	-3.7 ± 2.67	3.97 ± 2.25	0.032	
SMRI PFC	3	7	33	33	-3.5 ± 2.27	3.71 ± 1.97	0.029	
SMRI PFC	3	8	33	33	-3.1 ± 1.96	3.29 ± 1.73	0.036	
SMRI PFC	3	9	33	33	-3.1 ± 1.89	3.23 ± 1.76	0.049	
McL PFC	25	25	32	49	2.50 ± 1.36	-1.6 ± 0.64	0.012	
McL PFC	25	26	32	49	2.92 ± 1.47	-1.9 ± 0.72	0.014	
McL PFC	25	27	32	49	2.88 ± 1.5	-1.9 ± 0.73	0.02	
McL PFC	25	28	31	49	2.74 ± 1.49	-1.9 ± 0.7	0.018	
McL PFC	26	26	32	49	3.34 ± 1.68	-2.2 ± 0.82	0.013	
McL PFC	26	27	32	49	3.07 ± 1.60	-2 ± 0.79	0.021	
McL PFC	26	28	31	49	2.81 ± 1.53	-1.9 ± 0.73	0.017	
McL PFC	27	27	32	49	2.8 ± 1.59	-1.8 ± 0.78	0.036	
McL PFC	27	28	31	49	2.86 ± 1.50	-1.8 ± 0.71	0.015	
McL PFC	28	28	32	49	2.8 ± 1.38	-1.8 ± 0.68	0.0055	
McL PFC	3	3	34	49	3.93 ± 1.28	-2.7 ± 1.72	0.004	
McL PFC	3	4	34	49	1.85 ± 1.17	-1.3 ± 1.18	0.047	
PFC	1	10	65	81	-2.0 ± 1.21	1.68 ± 0.87	0.021	
PFC	1	11	65	81	-1.7 ± 1.18	1.47 ± 0.8	0.035	
PFC	1	8	67	82	-1.9 ± 1.31	1.61 ± 0.97	0.05	
PFC	1	9	67	82	-2 ± 1.27	1.69 ± 0.98	0.038	
PFC	2	10	65	81	-2.0 ± 1.14	1.60 ± 0.84	0.018	
PFC	2	11	65	81	-1.7 ± 1.11	1.38 ± 0.77	0.024	
PFC	2	12	65	81	-1.4 ± 1.11	1.12 ± 0.71	0.043	
PFC	2	9	67	82	-2 ± 1.23	1.66 ± 0.95	0.044	
PFC	28	28	56	76	2.34 ± 1.37	-1.7 ± 0.61	0.026	
PFC	3	10	65	81	-1.7 ± 1.07	1.41 ± 0.8	0.031	
PFC	3	11	65	81	-1.4 ± 1.04	1.18 ± 0.73	0.042	
PFC	4	10	65	81	-1.5 ± 0.98	1.31 ± 0.73	0.049	
PFC	4	9	67	82	-1.9 ± 1.07	1.54 ± 0.82	0.047	
PFC	5	5	68	82	-2.3 ± 1.52	1.87 ± 1.13	0.044	
PFC	6	7	68	82	-1.4 ± 0.95	1.16 ± 0.73	0.043	
PFC	6	8	68	82	-1.2 ± 0.88	1.03 ± 0.66	0.034	
PFC	7	7	68	82	-1.6 ± 1.04	1.36 ± 0.8	0.042	
PFC	7	8	68	82	-1.3 ± 0.9	1.07 ± 0.67	0.026	
PFC	7	9	68	82	-1.8 ± 1.12	1.48 ± 0.84	0.049	
WBC	1	18	226	260	-0.48 ± 0.22	0.2 ± 0.23	0.05	
WBC	1	20	224	259	-0.5 ± 0.21	0.20 ± 0.22	0.035	
WBC	1	21	224	258	-0.5 ± 0.21	0.22 ± 0.21	0.028	
WBC	1	22	222	257	-0.49 ± 0.21	0.21 ± 0.21	0.024	
WBC	1	23	212	244	-0.53 ± 0.21	0.27 ± 0.22	0.0094	0.96

WBC	1	24	197	212	-0.5 ± 0.22	0.42 ± 0.23	0.0042	0.74
WBC	1	25	157	204	-0.26 ± 0.24	0.38 ± 0.23	0.046	
WBC	1	26	157	204	-0.29 ± 0.24	0.37 ± 0.22	0.038	
WBC	1	27	157	204	-0.30 ± 0.23	0.35 ± 0.22	0.036	
WBC	1	28	145	190	-0.28 ± 0.24	0.4 ± 0.22	0.03	
WBC	10	10	295	329	0.13 ± 0.37	-0.12 ± 0.28	0.036	
WBC	10	11	292	328	0.24 ± 0.25	-0.23 ± 0.18	0.006	0.84
WBC	10	12	291	328	0.28 ± 0.25	-0.26 ± 0.18	0.0055	0.81
WBC	11	11	292	328	0.37 ± 0.23	-0.33 ± 0.16	9.6e-05	0.085
WBC	11	12	291	328	0.36 ± 0.26	-0.33 ± 0.16	0.00025	0.13
WBC	12	12	291	328	0.36 ± 0.37	-0.32 ± 0.22	0.00098	0.32
WBC	15	24	249	259	-0.25 ± 0.21	0.26 ± 0.24	0.045	
WBC	16	23	269	300	-0.3 ± 0.21	0.25 ± 0.24	0.048	
WBC	16	24	249	259	-0.27 ± 0.22	0.28 ± 0.25	0.034	
WBC	17	24	249	259	-0.26 ± 0.24	0.25 ± 0.24	0.032	
WBC	18	24	249	259	-0.21 ± 0.25	0.18 ± 0.22	0.038	
WBC	19	24	249	259	-0.17 ± 0.26	0.16 ± 0.22	0.047	
WBC	2	20	224	259	-0.49 ± 0.21	0.19 ± 0.21	0.037	
WBC	2	21	224	258	-0.49 ± 0.2	0.21 ± 0.21	0.03	
WBC	2	22	222	257	-0.49 ± 0.2	0.19 ± 0.20	0.027	
WBC	2	23	212	244	-0.52 ± 0.20	0.25 ± 0.21	0.011	
WBC	2	24	197	212	-0.49 ± 0.21	0.39 ± 0.22	0.0053	0.81
WBC	2	26	157	204	-0.29 ± 0.23	0.33 ± 0.21	0.043	
WBC	2	27	157	204	-0.30 ± 0.22	0.32 ± 0.21	0.045	
WBC	2	28	145	190	-0.29 ± 0.23	0.38 ± 0.22	0.029	
WBC	20	24	249	259	-0.15 ± 0.23	0.15 ± 0.2	0.035	
WBC	3	10	233	261	-0.58 ± 0.25	0.19 ± 0.26	0.048	
WBC	3	17	227	260	-0.47 ± 0.18	0.12 ± 0.19	0.042	
WBC	3	18	226	260	-0.46 ± 0.18	0.13 ± 0.18	0.037	
WBC	3	19	226	260	-0.46 ± 0.18	0.12 ± 0.18	0.041	
WBC	3	20	224	259	-0.48 ± 0.18	0.14 ± 0.18	0.023	
WBC	3	21	224	258	-0.47 ± 0.17	0.16 ± 0.18	0.019	
WBC	3	22	222	257	-0.47 ± 0.18	0.15 ± 0.18	0.017	
WBC	3	23	212	244	-0.51 ± 0.18	0.19 ± 0.19	0.007	0.86
WBC	3	24	197	212	-0.49 ± 0.19	0.31 ± 0.2	0.0026	0.61
WBC	3	26	157	204	-0.28 ± 0.21	0.25 ± 0.19	0.041	
WBC	3	27	157	204	-0.29 ± 0.2	0.24 ± 0.19	0.038	
WBC	3	28	145	190	-0.29 ± 0.21	0.3 ± 0.19	0.02	
WBC	3	7	233	261	-0.58 ± 0.29	0.25 ± 0.3	0.048	
WBC	3	8	233	261	-0.55 ± 0.25	0.20 ± 0.25	0.036	
WBC	3	9	233	261	-0.66 ± 0.26	0.24 ± 0.28	0.04	
WBC	4	10	233	261	-0.6 ± 0.2	0.12 ± 0.21	0.0092	0.96
WBC	4	11	231	260	-0.52 ± 0.19	0.09 ± 0.19	0.012	
WBC	4	12	230	260	-0.49 ± 0.18	0.053 ± 0.17	0.017	
WBC	4	13	228	260	-0.5 ± 0.17	0.058 ± 0.17	0.018	

WBC	4	14	228	260	-0.47 ± 0.17	0.074 ± 0.17	0.019	
WBC	4	15	228	260	-0.46 ± 0.16	0.08 ± 0.16	0.016	
WBC	4	16	227	260	-0.45 ± 0.16	0.073 ± 0.16	0.022	
WBC	4	17	227	260	-0.48 ± 0.15	0.075 ± 0.16	0.013	
WBC	4	18	226	260	-0.47 ± 0.15	0.094 ± 0.16	0.013	
WBC	4	19	226	260	-0.47 ± 0.16	0.081 ± 0.16	0.018	
WBC	4	20	224	259	-0.48 ± 0.16	0.11 ± 0.16	0.0088	0.96
WBC	4	21	224	258	-0.47 ± 0.15	0.12 ± 0.16	0.0071	0.86
WBC	4	22	222	257	-0.47 ± 0.16	0.11 ± 0.16	0.0069	0.86
WBC	4	23	212	244	-0.52 ± 0.17	0.14 ± 0.17	0.003	0.61
WBC	4	24	197	212	-0.52 ± 0.17	0.24 ± 0.18	0.0010	0.32
WBC	4	25	157	204	-0.28 ± 0.19	0.19 ± 0.18	0.036	
WBC	4	26	157	204	-0.32 ± 0.19	0.19 ± 0.18	0.025	
WBC	4	27	157	204	-0.32 ± 0.18	0.18 ± 0.17	0.027	
WBC	4	28	145	190	-0.32 ± 0.19	0.24 ± 0.18	0.014	
WBC	4	4	234	266	-0.51 ± 0.51	0.45 ± 0.5	0.041	
WBC	4	5	233	262	-0.62 ± 0.35	0.23 ± 0.35	0.040	
WBC	4	6	233	261	-0.73 ± 0.26	0.15 ± 0.26	0.0082	0.95
WBC	4	7	233	261	-0.63 ± 0.21	0.14 ± 0.21	0.0042	0.74
WBC	4	8	233	261	-0.57 ± 0.18	0.11 ± 0.18	0.0027	0.61
WBC	4	9	233	261	-0.7 ± 0.21	0.17 ± 0.23	0.0052	0.81
WBC	5	24	244	251	-0.23 ± 0.17	0.17 ± 0.17	0.028	
WBC	5	6	293	324	-0.54 ± 0.23	0.48 ± 0.31	0.0029	0.61
WBC	5	7	293	324	-0.44 ± 0.18	0.39 ± 0.24	0.0011	0.32
WBC	5	8	293	324	-0.39 ± 0.14	0.35 ± 0.21	0.00059	0.26
WBC	5	9	293	324	-0.45 ± 0.18	0.42 ± 0.25	0.027	
WBC	6	24	245	251	-0.21 ± 0.17	0.15 ± 0.16	0.03	
WBC	6	6	294	324	-0.57 ± 0.23	0.52 ± 0.27	2.3e-05	0.061
WBC	6	7	294	324	-0.40 ± 0.16	0.37 ± 0.20	6.7e-05	0.085
WBC	6	8	294	324	-0.35 ± 0.11	0.32 ± 0.19	0.00015	0.10
WBC	6	9	294	324	-0.44 ± 0.17	0.41 ± 0.23	0.026	
Germline	5	5	27	28	-0.68 ± 0.44	0.65 ± 0.54	0.028	

Appendix 5. Functional annotation clusters for 5-hmC enriched brain genes (DAVID).

Each cluster represents a group of genes with significant overlap in annotation terms. The Enrichment Score of a cluster is the geometric mean of the exponents of the P-values associated with all the member terms in a cluster. The low P-values of individual GO terms are a trade-off for identifying clusters where genes had greater overlap in annotation terms (DAVID classification stringency = "High"). Using the default setting would have identified clusters with higher enrichment scores but lower overlap.

Functional annotation category	Annotation terms in cluster	*P* of term	*Q* of term
Annotation Cluster 1	**Enrichment Score: 2.79**		
GOTERM_MF_FAT	GO:0005216~ion channel activity	5.4×10^{-4}	0.16
GOTERM_MF_FAT	GO:0022838~substrate specific channel activity	8.4×10^{-4}	0.16
GOTERM_MF_FAT	GO:0015267~channel activity	1.5×10^{-3}	0.18
GOTERM_MF_FAT	GO:0022803~passive transmembrane transporter activity	1.5×10^{-3}	0.18
GOTERM_MF_FAT	GO:0022836~gated channel activity	2.6×10^{-3}	0.21
GOTERM_MF_FAT	GO:0005261~cation channel activity	6.1×10^{-3}	0.32
Annotation Cluster 2	**Enrichment Score: 2.42**		
GOTERM_MF_FAT	GO:0005089~Rho guanyl-nucleotide exchange factor activity	1.0×10^{-3}	0.15
GOTERM_MF_FAT	GO:0005088~Ras guanyl-nucleotide exchange factor activity	2.2×10^{-3}	0.21
GOTERM_BP_FAT	GO:0035023~regulation of Rho protein signal transduction	2.4×10^{-2}	0.78
Annotation Cluster 3	**Enrichment Score: 2.05**		
GOTERM_BP_FAT	GO:0048666~neuron development	4.3×10^{-3}	0.48
GOTERM_BP_FAT	GO:0031175~neuron projection development	1.2×10^{-2}	0.65
GOTERM_BP_FAT	GO:0030030~cell projection organization	1.4×10^{-2}	0.68

Appendix 6. Read counts from Helicos single-molecule sequencing.

Target reads are reads where the 5' end lies within ± 3bp of a CCGG site. Non-target reads are reads where the 5' end lies outside ±200bp of a CCGG site.

	Channel	Aligned	Target (T)	Non-target (NT)	(T/(T+NT)) * 100
Replicate 1	Undigested	408,987	1,341	276,142	0.5
	gDNA (*Msp*I)	2,761,844	1,083,483	662,416	62.1
	gDNA (*Hpa*II)	1,407,105	444,298	511,772	46.5
	glc-gDNA (*Msp*I)	1,231,589	421,646	356,355	54.2
Replicate 2	Undigested	1,301,480	4,766	850,546	0.6
	gDNA (*Msp*I)	2,224,007	829,158	593,030	58.3
	gDNA (*Hpa*II)	2,840,191	918,979	1,004,470	47.8
	glc-gDNA (*Msp*I)	3,166,749	1,185,353	833,898	58.7
Replicate 3	Undigested	2,395,206	5,758	1,663,120	0.3
	gDNA (*Msp*I)	2,166,860	646,267	620,031	51.0
	gDNA (*Hpa*II)	2,115,006	662,905	534,374	55.4
	glc-gDNA (*Msp*I)	1,487,507	374,791	469,684	44.4