The impact of the environment on DNA methylation in humans and zebrafish

A thesis presented in partial fulfilment
of the requirements for the degree of

Doctor of Philosophy
in
Biology

At the University of Canterbury, Christchurch
New Zealand

Alexandra Noble
2020

Abstract

Noble, Alexandra J., *The impact of the environment on DNA methylation in humans and the zebrafish*

Doctor of Philosophy, December 2020, University of Canterbury, Christchurch, New Zealand.

DNA methylation is a chemical modification to the DNA strand, which can control gene expression. DNA methylation can be modified by the environment. For example, tobacco use substantially alters DNA methylation, and hence DNA methylation therefore provides a route through which the environment can lead to alterations in gene expression. Consequently, alterations to DNA methylation patterns have been associated with disease phenotypes in humans and other mammals. However, the precise role of environmentally-induced DNA methylation changes in the onset of pathological phenotypes is not often clearly defined.

Here, we investigate the response of DNA methylation to two different environmental exposures – adulthood cannabis and *in utero* tobacco exposure. These environmental exposures are important because they are associated with adverse phenotypes – long-term cannabis use, particularly through adolescence, is associated with adverse psychosocial wellbeing. The development of conduct problem (CP, including autism and antisocial behaviour disorder) in childhood and adolescence is associated with exposure to tobacco during development (*in utero*). However, as yet, no studies have explored the role of DNA methylation in the link between these exposures and their associated phenotypic effects.

Therefore, here we first asked whether DNA methylation in a longitudinal human cohort, the Christchurch Health and Development Study (CHDS), was altered in response to long term cannabis exposure, with and without tobacco. Using the Illumina EPIC array, we detected nominal differential DNA methylation in response to cannabis specifically, in genes associated with the following pathways; Cholinergic synapse, glutamatergic synapse and dopaminergic synapse. These observations show a potential mediation between DNA methylation in the observed phenotypic effects of cannabis use.

In order to develop a tool to investigate this association further, we assessed the efficacy of a targeted, high throughput amplicon-based approach, bisulfite-based amplicon sequencing (BSAS), to replicate differential methylation at loci identified via EPIC array. We found that the ability of BSAS to detect equivalent differential methylation was locus-specific, meaning that it has value as a validation and replication tool, but that each locus for validation must be tested before being applied to a large study.

Cannabis use is a contentious issue, mainly because of the debate around its therapeutic but also its psychoactive properties. In order to quantify the impact of both of its main cannabinoids, (-)-trans-Δ^9-tetrahydrocannabinol (THC) and cannabidiol (CBD) were exposed to zebrafish embryos. Following exposure reduced representation bisulfite sequencing (RRBS) was used to quantify their impact of each cannabinoid on DNA methylation. Differential methylation was found in each of the exposure groups, findings demonstrated the greatest number of methylation differences was in the CBD exposure group. CBD DNA methylation differences were found in genes that have roles in neurodevelopment, neurotransmission and behaviour. THC DNA methylation differences on the other hand were found to alter genes with roles in the axon guidance and retinal ganglion pathways, supporting the role of DNA methylation in the biological response to THC. Furthermore, our data revealed a role for both THC and CBD in brain related pathways, indicating that further research is needed to understand the full biological impacts of the two compounds.

Next, to determine if tobacco-induced DNA methylation alterations are important in the link between *in utero* tobacco exposure and the development of CP, here, we applied BSAS to a subset of CHDS participants to assess DNA methylation in *in utero*-exposed individuals compared to non-exposed individuals, with and without CP. We selected a panel of genes with known roles in *in utero* neurodevelopment, and identified differential methylation that was specific to individuals exposed to tobacco during development, who had high CP scores. We imply that developmentally-induced DNA methylation alterations may be playing a role in the development of CP in exposed individuals. To investigate this further, we applied a genome-wide approach (EPIC array) to a larger cohort and identified nominal significance at genes

involved in global developmental delay and neurological disorders, indicating that, in addition to CP, visual impairment may be a phenotypic response to *in utero* tobacco exposure.

Lastly, we discuss whether DNA methylation analysis in whole blood samples is able to predict DNA methylation changes in brain tissue. To answer this question, we used publicly available data of the top lists of differentially methylated CpG sites in blood and brain tissue from individuals with schizophrenia. We found that, the methylation of individual CpG sites did not replicate between tissues, the genes and pathways that have biological relevance to schizophrenia (e.g. mTOR signalling pathway and the mRNA surveillance pathway) were identified in both tissue types, demonstrating the value and applicability of whole blood as a proxy tissue.

Overall, here we demonstrate a role for DNA methylation in the biological response to cannabis, and a link between *in utero* tobacco exposure and development of CP. Further research is required to understand the mechanism through which these changes can contribute to disease.

Acknowledgments

I would like to express my deepest gratitude to my supervisor Dr Amy Osborne. From the very beginning, you have had faith in me, you have continued to show this with your support throughout my entire PhD journey. Moreso, you have offered such integral critiques in the writing of this thesis, for that I am extremely thankful for. You have also taught me how to mourn the loss of data (with a spin and tin, obviously) and move on with the best foot forward, this ensures we always strive for the best. I hope that one day I will be able to have as much faith in someone else, as you have had in me.

I would also like to further acknowledge John Pearson and Martin Kennedy at the University of Otago for their co-supervision. In particular, John, your guidance and expertise in bioinformatics has taught me valuable skills that are imperative for becoming a better scientist. Martin, thank you for overseeing all of the finer details and your valuable suggestions.

I would like to also offer my gratitude to fellow lab members and the people who resided on the 5th floor of Biological Sciences. My PhD would not be the same without the great people I have met along the way, so much so that this place is “the best school in the country”. Rudolf, especially thank you for always ensuring my caffeine levels were met adequately on a daily basis. More so, staff members including Jan McKenzie, Jonathan Hill, Sarah Flanagan, Rennie Bishop and Elissa Cameron and Alison Miller (University of Otago) who all provided help to me throughout my studies.

To Mum, Dad, Josie and Ben thank you for always supporting me and being there when I needed it. You have all maintained such enthusiasm when I talk about my research and for that I am very thankful. Lastly, to the OG Dr.A Noble thank you for helping me with the most important things, such as in a hurry Venn diagrams and emails informing people their statistical packages were wrong. I hope that people will now get us confused and presume I know things about statistics!

Finally, I would like to thank Biological Sciences and the University of Canterbury for my Doctoral Scholarship, being paid was very nice.

Contents

List of Tables

List of Figures

Chapter 1

1. Introduction and outline

Part 1: The Molecular Mechanism of DNA Methylation

1.1.1 From 'epigenotype' to epigenetics

The field of epigenetics began with a series of experiments by Conrad Waddington, in 1942 [1]. Waddington observed that when exposed to heat or shock, the fruit fly, *Drosophila melanogaster,* would respond with the *"development of adaptive character which might itself become so far canalised it continued to appear even when the conditions appear to the previous norm*" [2]. The observation led Waddington to propose the existence of an intermediate and independent link between a gene and the expected phenotype [1], and he coined the term 'epigenotype'. Little did Waddington know, he was actually describing what we refer to now as epigenetics, literally translated as epi- "on top of" genetics- "genes". The meaning has been refined over time, and is now specifically used to describe reversible gene regulation occurring independently of the underlying DNA sequence [3].

In 1957, Waddington proposed the *epigenetic landscape theory*. The influential theory was a way to describe the process of cell-fate determination during the various phases of development in a multicellular organism [4]. It was of high importanance because it provided a way to illustrate the concept that the vast majority of cells within an individual share identical genotypes, yet the diversity of cell end-point is phenomenal [5]. In the theory, a marble (Figure 1.1) at the top of the valley depicts a pluripotent cell, which has the capacity to differentiate into any cell type. The valleys represent the many different trajectories a cell can take while its fate is being determined; essentially, they are pathways for differentiation, cell-fate determination, and tissue development. Each end point has its own unique biological function that is important for all multicellular organisms, and thus the epigenetic landscape shapes the opportunity for a cell to follow a specific pathway to differentiation; as that one cell's role becomes more defined, its gene expression becomes restricted and exhibits a "locked in" state [6], signifying the end of a pluripotent state. A key process that is central to this process of cell differentiation is DNA methylation.

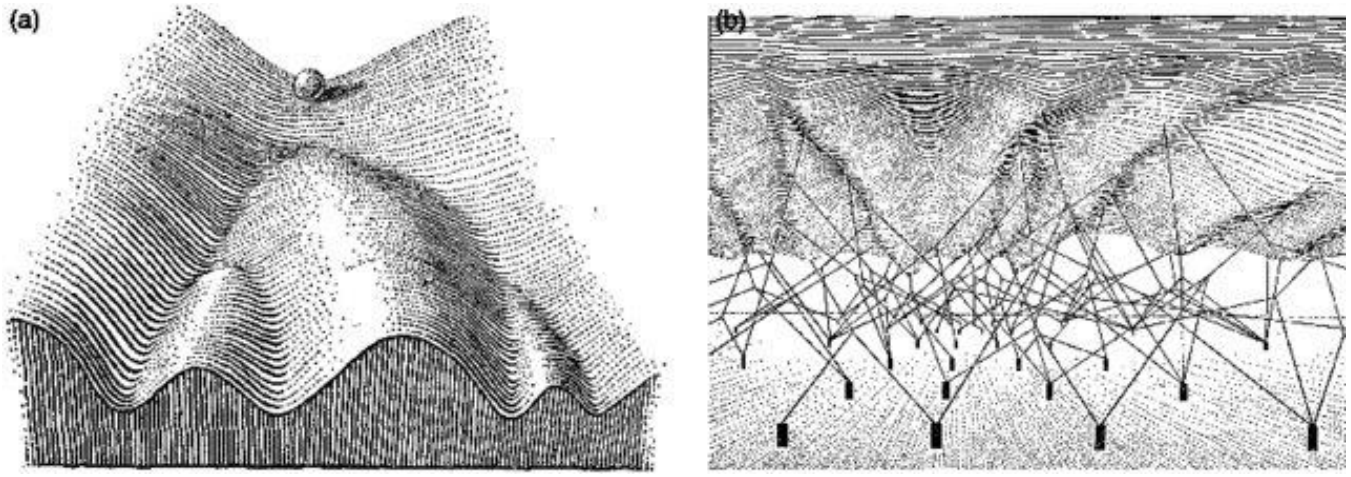


Figure 1.1 Waddington's epigenetic landscape. a) a marble represents a singular pluripotent cell which travels down a route that is shaped by the epigenetic landscape, this ultimately leads to a defined/differentiated state. b) the complexity of the different trajectories which is driving the underlying decisions of the landscape [7]. Permission granted for the use of this image.

1.1.2 Epigenetic regulation via DNA Methylation

There are several ways in which epigenetic processes can cause phenotypic changes, but one of the most well-studied is DNA methylation. DNA methylation is one type of epigenetic modification and it occurs when a methyl group is covalently transferred to the C5 position of the cytosine ring of a DNA molecule by a methyltransferase enzyme (Figure 1.2), which is then termed 5-methylcytosine (m^5C). DNA methylation plays a crucial role in regulating gene expression and normal development [8].

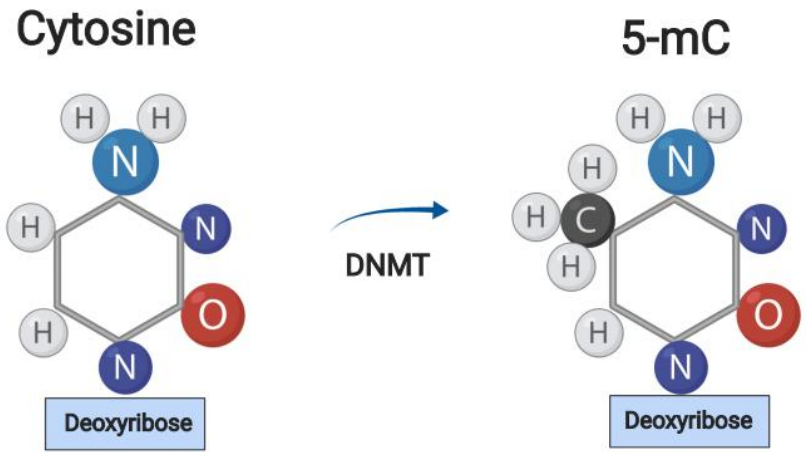


Figure 1.2 Cytosine methylated at the 5' carbon by DNA methyltransferaases resulting in a 5-mC. Made with BioRender.com

Cytosine methylation is present in the DNA of vertebrates, some flowering plants, fungi, invertebrates, protists and bacterial species [9], and is common in all large genomes of eukaryotes [10]. In mammals, approximately 98% of DNA methylation occurs at a CpG dinucleotide, which is the methylated cytosine and the phosphodiester bond that joins the cytosine with an adjacent guanine nucleotide. However differs in embryonic cells, where a quarter of methylation is in a non-CpG context. The difference in DNA methylation context has been hypothesised to be functionally significant; non-CpG methylation around gene bodies in ooytes correlates with the level of expression of corresponding genes, showing context-dependent functional significance of non-CpG methylation [10]. Further, DNA methylation during development is dynamic – extensive epigenetic remodelling must be undertaken during zygote formation, with DNA methylation almost entirely erased after fertilisation, and then re-established in the embryo [11]. Specifically, DNA methylation in the paternal genome (where overall DNA methylation is very high) will undergo demethylation early in zygote formation [12], while the maternal genome, which has relatively lower global methylation levels, undergoes demethylation at a less dynamic pace [13]. The dynamics of demethylation prompts key events in early development, and is essential for life [14, 15].

Once established, DNA methylation can be influenced by the surrounding environment, and factors such as diet, stress and aging can all impact on DNA methylation at CpG residues [16]. Of these environmental factors, age is possibly the most well-studied, with DNA methylation patterns shown to be intrinsically linked to an individual's age. For example, twin studies revealed that younger twins had virtually indistinguishable patterns of DNA methylation, whereas older twins had comparably different patterns [17]. It was hypothesised that the methylation patterns of adult twins differed due to the environmental influences that each individual twin had been exposed to [18]. A further study in monozygotic and non-twin individuals identified 88 CpG sites in and around 80 different genes which drastically changed methylation status in relation to age [19], and further, the DNA methylation status of just 71 CpG sites in the genome can predict an individual's age down to a standard error of 3.9 years [20]. Thus, considering that DNA methylation is dynamic, and can change with age and environmental exposures, there exists the potential for DNA methylation to

serve as a hallmark of individual environmental exposures, and this will be discussed fully in the role of the environment and disease (Part 2).

1.1.3 CpG Islands

As previously stated, 98% of DNA methylation occurs at CpG dinucleotides. The human genome contains ~3 x 10^7 CpG dinucleotides, and each can either be in a methylated or unmethylated state [21]. Groups of CpG sites are known as CpG islands and span 0.5 - 3 kilobases (kb) in length [22, 23]. CpG islands are mathematically defined as sequences exhibiting greater than 55% G+C content, with an observed/expected ratio of 0.65 [24]. CpG islands are associated with the promoter regions of roughly 76% of all human genes [25, 26]; there are over 30,000 CpG islands across the genome, and 21,000 of them lie within the promoter region of genes. Usually, CpG islands at promoters of active genes are unmethylated, which then allows transcription to occur [27]. Conversely, dense promoter methylation via CpG islands can prevent expression of genes that are not necessary for that cell type [24]. DNA methylation can occur also at CpG dinucleotides in the gene body [23] and gene body CpG islands are more likely to become methylated than promoter CpG islands [28]. Methylation both in promoter regions and in gene bodies can impede the transcriptional machinery, preventing the DNA sequence from being read, essentially silencing genes [29], via a reduction in the accumulation of gene transcripts [30].

Functionally, CpG methylation at CpG islands has many roles, both for correct developmental trajectories and also in disease. Of the former category, perhaps the best studied is the way in which DNA methylation contributes to the stability of X chromosome inactivation. X inactivation is the process in which one X chromosome in each cell of a female mammal is completely inactivated during development, to provide dosage compensation in gene expression [31]. Failure of X inactivation can lead to developmental disease [32]. An example from the latter category is methylation at CpG islands within tumour suppressor genes; promoters of tumour suppressor genes, which should be unmethylated to allow tumour suppressor gene expression, may be methylated in cancer cells [33], disrupting gene expression and causing disease. Thus, given that CpG island methylation patterns have been associated with a variety

of diseases, promoter methylation can be interpreted as a “hallmark” or a “biomarker” for disease states [34]

1.1.4 DNA methylation via DNA methyltransferases

DNA methylation is regulated by a family of DNA methyltransferase enzymes (DNMTs): DNMT1, DNMT2, DNMT3a, DNMT3b and DNMT3L. The family of enzymes catalyse cytosine methylation by transferring a methyl group from S-adenosyl-L-methionine (SAM) to deoxycytosine [35]. DNA methyltransferases can largely be split into two subgroups: i) maintenance methyltransferases (Figure 1.3a), and ii) *de novo* methyltransferases (Figure 1.3b) [36].

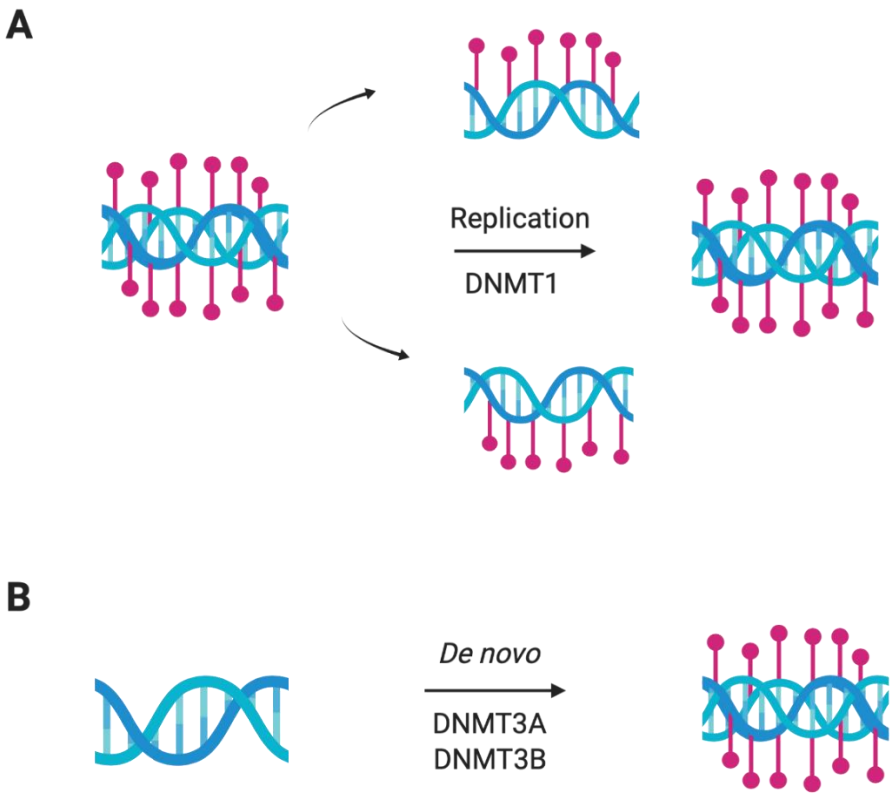


Figure 1.3 Members of the DNMT family. A) DNMT1 is responsible for the maintenance of methylation patterns during cell division, replicating existing CpG signatures to newly synthesised strands of DNA. B) DNMT3A and DNMT3B are known as de novo methyltransferases, and are responsible for new CpG signatures (made with biorender.com).

DNMT1 is the most abundant methyltransferase enzyme in adult cells [37], and it is largely responsible for maintenance of DNA methylation through the cell division cycle [38]. It maintains DNA methylation by copying the methylation pattern from a replicating to a nascent DNA strand [36], thus replicating the CpG signature from parent to daughter strands [39]. DNMT2 is the least understood methyltransferase in terms of its role in DNA methylation, but it is known to have a significant role in methylation of transfer RNA [40], and in *Drosophila*, DNMT2 is the sole cytosine DNA methyltransferase [41].

The *de novo* methyltransferases *dnmt3a* and *dnmt3b* are highly expressed in undifferentiated embryonic cells and then downregulated in adult somatic tissues when studied in mice [42]. They transfer a methyl group to a cytosine residue that is unmethylated, and are mainly active during development [43]. DNMT3L is necessary for the establishment of methylation marks at maternally imprinted loci in developing oocytes [44].

All of the enzymes in the DNMT family have individual but crucial roles, which have been shown to be lethal in mice models if knocked out [10, 43, 45]. Thus, given the importance of DNA methylation as a mechanism, it is crucial that we understand the way in which different environmental factors might influence this key mechanism.

1.1.5 Detecting differential DNA methylation

There are numerous methods for quantifying and analysing DNA methylation (Table 1.1). A common method is bisulfite sequencing [46], which is a technique that can detect DNA methylation at individual CpG sites via a combination of sodium bisulfite treatment and DNA sequencing. Briefly, treatment of DNA with sodium bisulfite converts all non-methylated cytosine residues to uracil using polymerase chain reaction (PCR) (Figure 1.5). It then becomes possible to ‘read’ which cytosines were methylated in the original sample via DNA sequencing, when aligned to an unconverted reference sequence [47].

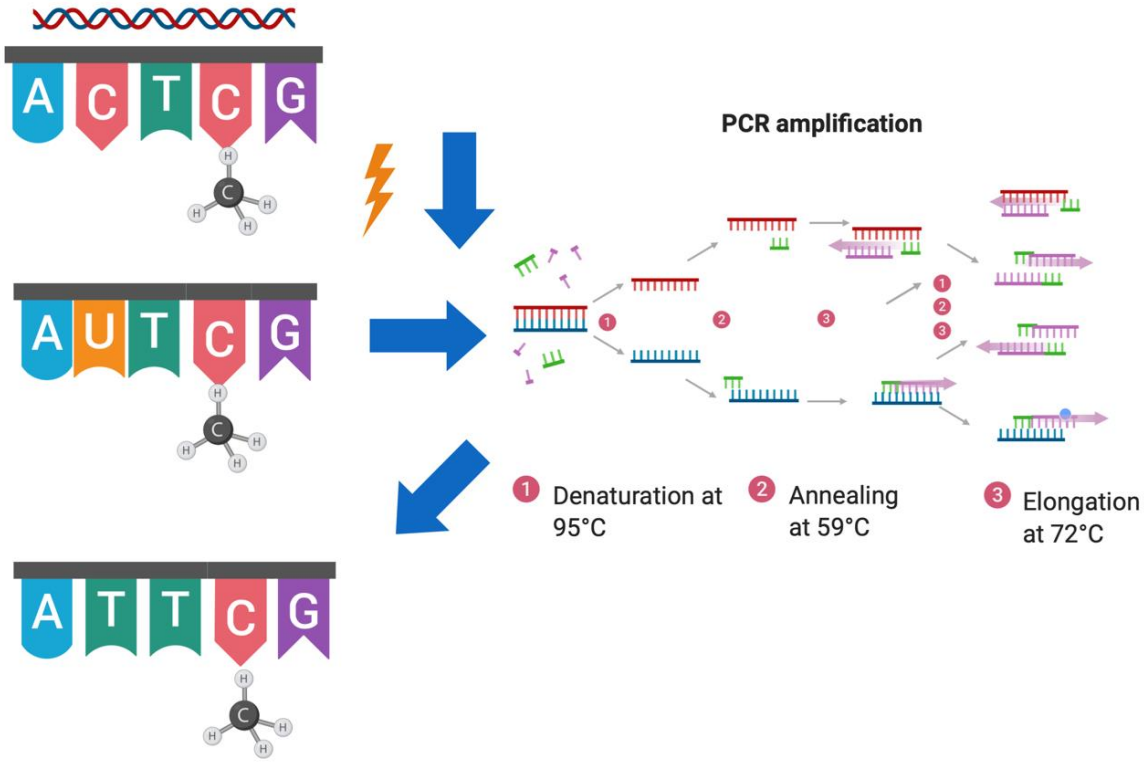


Figure 1.4 The process of bisulfite treatment of DNA to preserve methylated cytosines and chemically modify unmethylated cytosines to tyrosines using the process of PCR amplification. Figure was made with BioRender.

This provides a cost- and time-efficient method of comparing differential methylation between two individuals, or it can be averaged to compare methylation levels between cohorts or populations. Advancements in our ability to quantify and analyse DNA methylation have been driven by next generation sequencing technology, where mass investigation of methylation across the genome can now be achieved. Thus, there are multiple modes through which bisulfite-converted DNA can be quantified. Some include sequencing of the entire genome, some a reduced representation, some sequence amplicons, and some are probe based:

1. Techniques for targeted methylation analysis
    a) **Bisulfite-based amplicon sequencing (BSAS):** This technique uses both bisulfite conversion and PCR amplification of short amplicons of ~250bp to obtain estimates of differential methylation between two populations [48].

Genomic material must be first bisulfite converted before primers for the methylated template amplify the sites of interest. The relative disadvantage to this method is that PCR amplification can be prone to error [49].

b) **Pyrosequencing:** This is a DNA sequencing technique which relies upon the release of pyrophosphate (PPi) during DNA synthesis [50]. Pyrosequencing can be performed after bisulfite conversion of DNA PCR products or in conjunction with long interspersed element-1 (LINE-1) whole genome approach. It relies upon four enzymes: DNA polymerase, ATP sulfurylase, firefly luciferase and apyrase. As the single stranded DNA template is made, each nucleotide is incorporated which coincides with the release of pyrophosphate which triggers ATP sulfurase. Then the firefly luciferases sense light, ultimately produces a light reaction [51]. Thus, Pyrosequencing is known to have quantitative flaws due to the output of sequences generated through fluorescence methods [52].

2. Common technologies used for genome-scale analysis of the methylome are:

a) **Methylation arrays: Illumina EPIC 850K arrays** quantify methylation at 863,904 different CpG sites [53]. Although this is still a small proportion of the total number of CpG sites in the genome (~28 million) it represents a broad distribution of sites that give a specific and robust measurement of methylation at those sites. The technique relies on a probe-based method, which can be expensive.

b) **Methylated DNA immunoprecipitation sequencing (MeDIP-Seq):** This method requires minimal DNA input and so is useful in experiments where DNA yield is limited. Methylated DNA is immunoprecipitated with an antibody raised against a CpG site which is followed by DNA sequencing [54, 55]. The antibody-based selection is biased towards higher CpG density [56] and it has low base resolution (~150 bp), compared to many other techniques which allows for single base resolution [57].

c) **Whole genome bisulfite sequencing (WGBS):** DNA undergoes bisulfite conversion which is then coupled with next generation sequencing technology to obtain large numbers of DNA sequences with methylated cytosine residues converted to uracil. The method has been used frequently and in particular with mapping methylation in human cancers [58, 59]. There

is an extensive literature regarding preparation protocols, sequencing output and interpretation of data [60]. Bisulfite conversion does have its pitfalls, with sequencing biases and overestimation of global methylation [60, 61].

d) **Reduced-representation bisulfite sequencing (RRBS):** This technique utilises a reduced representation of the CpG sites within the genome which equates to around 85% of the CpG islands [62] via sequencing. Since the output of RRBS is sequence-based, RRBS returns more information than the probe-based EPIC array. The technique utilizes the methylation-insensitive restriction enzyme MspI to cut sites within the genome. The cut fragments vary in length between 40 - 220bp [62]. The fragments are then converted using sodium bisulfite and sequenced. Although this technique provides reduced representation of the whole genome, cut sites span most promoter regions which ensures most CpG sites are represented. The approach provides single-nucleotide resolution that is highly sensitive that only requires relatively small amounts of DNA input [62]. For example, clinical tumour samples [63] or samples where little material can be obtained such as organ specific sampling in mice can still assess genome wide methylation This technique is rather intensive both in wet lab work as well as computationally, compared to other methods. Although it is considered to be “whole genome” it is still only a representation of the total number of CpG sites.

e) **Nanopore MinION**: the Oxford Nanopore sequencing system provides real-time, high-throughput, and high read length sequences via a portable sequencing device [64]. It reads a DNA sequence by measuring the changes in electrical conductivity generated as the DNA strands pass through hundreds of nanopores, with sequencing complete in 48 hours. Genome coverage during this period depends on the size of the genome. Larger genomes will need multiple sequencing runs. Due to the pore-based method of sequencing, unmethylated cytosines and methylated cytosines disturb the ion current in distinct ways, enabling differentiation between modified and unmodified cytosines [65, 66]. Allowing for distinct methylation detection in difficult-to-map regions of the genome [67].

Other technologies such as enzyme-linked immunosorbent assay (ELISA), high performance liquid chromatography mass spectrometry (HPLC-MS) and high performance liquid chromatography ultra violet (HPLC-UV) can all quantify total methylation levels within a genome [68] [69, 70]. However, they are not sequence-based and therefore unable to identify specific differentially methylated cytosines nor their precise location within the genome. The capacity to identify the genes (or nearest genes) which display differential methylation is important to this research project, so global methylation techniques will not be discussed here.

Table 1.1 Different methods for detecting DNA methylation

Method	System	Coverage	Starting material	DNA origin	Sensitivity	Specificity	Cost	Reference
Whole genome methods								
Illuminia EPIC array	BS convert/ Bead array	850,000 sites (4% of genome)	0.5- 1 µg	Humans	Very high	Very good	High	[68, 71]
MeDIP	Antibody/ Array	Whole genome	50 ng	Humans	Medium	Medium	High	[54]
WGBS	BS convert/ Sequencing	Whole genome	1-5 µg	Any	High	Good	High	[60]
RRBS	Sequencing	Whole genome	1 µg	Any	High	Good	Medium	[62]
Specific targeted approaches								
PCR based	BS convert/ Sequencing	Gene specific	100 ng	Any	High	Good	Low	[48]
Pyrosequencing + LINE 1	BS convert	Gene specific	1 µg	Any	High	Good	Medium	[72-74]

1.1.6 Choice of tissue sample type in studies of DNA methylation

Given that levels of methylation vary substantially across different tissues [29], tissue sample choice is pivotal, and also frequently debated. It is of particularly importance when investigating diseases which are specific to, or associated with a certain cell type. Ideally, methylation would be measured in tissues of most relevance, but this becomes particularly difficult in human studies and disease of a specific cell type, e.g. the brain or other internal organs [75]; clearly, access to these cells from a live

organism would be impossible. As such, whole blood samples and saliva are the easiest and the least invasive way to obtain a sample.

While DNA methylation does vary between tissue types, whole blood samples have been shown to be a useful proxy tissue in which to assess phenotypically relevant DNA methylation differences. For example, tobacco smoking, which affects the lungs primarily, is associated with methylation changes in DNA of aryl hydrocarbon receptor repressor (*AHRR*) and this effect of tobacco on DNA methylation is seen in whole blood samples in numerous studies across multiple cell types [76-82].

One last limitation of using whole blood as proxy tissues is that they may suffer from tissue heterogeneity – whole blood is made up of multiple cell types, all of which have their own unique DNA methylation pattern. The variation in proportion to different cell types between samples from different individuals may bias or skew estimates of differential DNA methylation. However, bioinformatics tools have been developed to attempt to mitigate tissue heterogeneity as a cofounding variable [75].

Part 2: The role of the environment in disease

1.2.1 Environmental epigenetics

The ability of an organism to sense the environment and adapt its phenotype in response is a key concept in epigenetics [83]. This is particularly pertinent in the current research environment, where mounting evidence suggests that not all biological responses are determined by variation in DNA sequence [84]; it is increasingly clear that differences in methylation patterns within the genome can alter biological responses [85], and we know the environment can have a major influence on epigenetic modifications [86]. For example, alterations to DNA methylation patterns have been associated with nutritional, chemical, physical, and even psychosocial factors (e.g. stress) [3, 87-91]. In fact, methylation can generate epigenetic patterns that are specific to individual environmental factors, serving as an enduring hallmark of exposure to these factors. For example, differential methylation at very precise genomic regions has been identified in heavy alcohol use [92], and tobacco smoking [93].

Epigenetic changes can also occur in response to illicit, recreational and prescribed drugs, and it has been hypothesised that DNA methylation could play a role through addiction responses to such substances [94]. In particular, if we consider here exposure to nicotine via tobacco smoking, while nicotine as a chemical plays minor roles in the diseases caused by smoking (e.g. lung cancer, cardiovascular disease), it has a major role in the development of addiction through the mediation of persistent neuroplasticity [95]. Neuroplasticity is the ability of the brain to form new neural connections and structure in an adult brain [96], and it is associated with addiction. Plasticity is influenced by DNA methyltransferases [95], for example *DNMT3A* and 3B create dynamic changes in DNA methylation of plasticity-relevant genes that are important for learning and memory formation [97]. While the links between DNA methylation, neuroplasticity, and nicotine are not fully understood, it is feasible to suggest that, given the correlation between both nicotine and DNA methylation and neuroplasticity, DNA methyltransferase action could be altered by nicotine consumption, influencing neuroplasticity and addiction. Indeed, studies carried out in mice show an epigenetically mediated effect of early exposure to nicotine on pup

neural structure, that then persisted into adulthood [98], demonstrating that the epigenetic effects of nicotine exposure are lifelong.

Importantly, there is no evidence as yet to suggest that the effect of nicotine on addiction is isolated; given the ability of DNA methylation to respond to environmental factors, it is possible that other illicit and prescribed drugs also affect addiction via epigenetic mechanisms. Addiction itself is a complex disease that has a multitude of contributing factors, in particular environmental, behavioural, and biological; twin studies have revealed that the heritable genetic component which predisposes an individual to a drug addiction could be between 20-50%, with the remaining component due to non-genetic factors [99, 100]. Suggesting a complex relationship between addiction, genetics, and the environment. Therefore, probing the relationship between environmental factors and DNA methylation is required to begin to fully understand the biological effects of the environment on the genome.

1.2.2 Epigenetics and cannabis

The research in this thesis sets out to assess the impact of heavy long term cannabis use on DNA methylation in the human genome. Cannabis was chosen as the initial environmental factor to investigate because the strong interaction between DNA methylation and substances such as tobacco [101] suggests that cannabis may likewise be influencing DNA methylation within the genome.

Cannabis itself is a global public health issue and a growing topic of international controversy due to the debate surrounding its medicinal and therapeutic benefits [102]. Its main psychoactive ingredient is (-)-trans-Δ^9-tetrahydrocannabinol (THC), however the non-psychoactive component, cannabidiol (CBD), is the 2nd largest component of cannabis and is gaining interest as a therapeutic for pain relief [103]. Both THC and CBD target the endocannabinoid system, which plays a role in pathways related to neurodevelopment as well as other organs in the body.

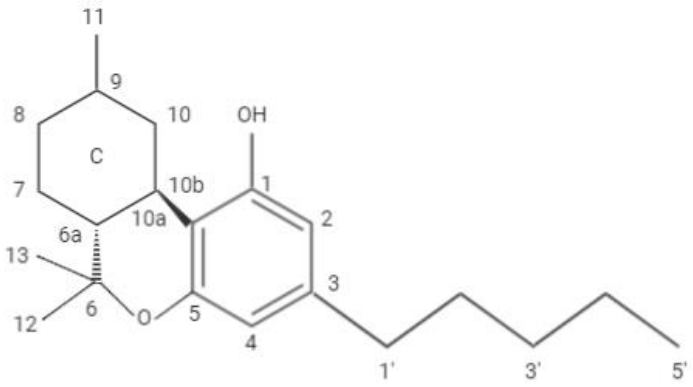
There is strong evidence to show that heavy cannabis usage is associated with increased risk of adverse psychosocial outcomes [104-107]. For example, low

educational achievement, low life satisfaction, inability to form social relationships, and, through co-substance use with other illicit drugs, psychosis in adults, are all associated with cannabis dependency [108-110]. In animal studies, behavioural abnormalities and molecular impairments to the brain have been associated with lifelong cannabis consumption [111, 112]. Importantly, DNA methylation can affect brain function. For example, DNA methylation is involved in behaviour, brain development, learning and memory, drug addiction, depression/bipolar and schizophrenia [103]. Thus, considering the links between recreational substances such as tobacco and alcohol and altered DNA methylation patterns, and given that altered methylation can affect brain development and brain function, we need to rigorously explore the relationship between cannabis and DNA methylation, so that we can better understand the links between cannabis and adverse psychosocial outcomes.

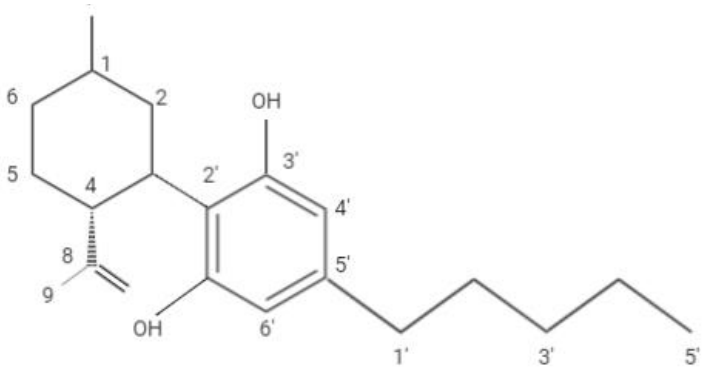
1.2.3 Cannabis

Marijuana (*Cannabis sativa)* is the most commonly used illicit substance in the Western world [113]. According to the World Health Organization (WHO), as of 2014 it is estimated that approximately 5% of the world's population (183 million) use cannabis annually [114]. It continues to be the most widely cultivated, produced and trafficked drug worldwide. There are a variety of ways the plant can be harnessed, each with a range of potencies. In its herbal form, marijuana is the unpurified product which consists of the leaves and stem of the plant. Cannabinoids are produced in the epidermal glands on the leaves, stems and bracts of the plant [115]. Hashish or "hash" is the dried cannabis resin which is compressed from the flowering tops of the cannabis plant. Ingestion of these products is either by smoking, eating, or vaporisation [116]. Cannabis oils are the most concentrated form as they consist of just the cannabinoids from the plant. Users report more addictive behaviours and withdrawal symptoms with high concentrations of THC in oil preparations [117]. Initial experiences with cannabis consist of paranoia, short term memory loss, relaxation, heightened sensory perception, laughter, altered perception of time and an increased appetite [118]. However, lasting impairments of the effects of cannabis in adult users have been well characterised; several studies have shown that deficits in learning,

memory, sustained attention, inability to make decisions and mental processing are all associated with cannabis consumption [118]. The plant itself contains 400 naturally occurring chemicals and of these approximately 100 are cannabinoids, which are the C1, C3 and C5 side chains found in cannabis. The remainder of cannabis components are terpenoids and flavonoids [119]. However, the exact makeup of the plants very much depends on plant genetics, growth conditions, and harvesting.



(-)-Trans-Δ^9-Tetrahydrocannabinol



(-)-Cannabidiol

Figure 1.2 the chemicalstructures of the two manjor compunds found in cannabis, (-)-trans-Δ^9-tetrahydrocannabinol and (-)-cannabidoil.

1.2.4 The endocannabinoid system

Cannabis remains a source of controversy, largely for the strong psychoactive effect of its main cannabinoid component THC. THC binds to cannabinoid receptor 1 (CB1, strongly) and cannabinoid receptor 2 (CB2, less preferentially than CB1). The CB1 receptor is located within the central nervous system (CNS) particularly in the neocortex, hippocampus, basal ganglia, cerebellum, and the brainstem [120, 121], and to a lesser extent in other areas of the body. The CB2 receptor is mainly located outside the CNS and is associated with the immune system.

The endocannabinoid system itself serves various roles within the body: appetite control, sensory processing, metabolism, hormonal regulation, and brain development [122, 123]. Cannabinoid receptors are present in both mammalian and non-mammalian vertebrates [124], suggesting the response of the endocannabinoid receptor to THC and THC-like substances is highly conserved across evolutionary timescales [125].

Stimulation of cannabinoid receptors causes activation of numerous transduction pathways through the inhibition of adenylyl cyclase and the reductions in cyclic AMP [126]. Both CB1 and CB2 receptors regulate the phosphorylation and activation of different members of the mitogen-activated protein kinase (MAPKs), Extracellular signal-regulated kinase-1 and -2 (ERK1/2), p38 MAPK and c-Jun *N*-Terminal kinase (JNK) [126]. CB1 receptors positively couple with K^+ channels and negatively couple with Ca^{2+} channels [126]. The activation of CB1 leads to inhibition of transmitter release thus regulates synaptic function [127].

Endocannabinoids are released from postsynaptic cells and then work their way back across the synapse forming a negative feedback loop [128-130]. As well as having a crucial role in neurotransmission, the endocannabinoid signalling system is also crucial for brain development; it guides cell fate decisions to differentiate between either neuronal (nerve cell) or glial cells (central nervous system- surround neuronal cells) [131, 132] .

CBD is thought to be responsible for the purported therapeutic effects of cannabis [133-135]. However, unlike THC, this component only targets CB2 receptors, and therefore it does not have psychoactive effects of THC, as there are comparatively few

CB2 receptors in the brain. Given the lack of psychoactive effects, and suggested therapeutic benefits of CBD, much current research focusses on removing THC from cannabis cultivars, in an attempt to shape cannabis as a therapeutic drug for treating numerous diseases, for example, epilepsy [136]. However, cannabinoids work in conjunction with one another and display a synergistic effect. Meaning that skewing the ratio of cannabinoids may not provide a therapeutic benefit [137]. Interestingly, over the last five decades, THC to CBD ratios have changed dramatically; in the 1970s, THC concentrations found in cannabis were less than 3%, while current evidence from the Netherlands shows concentrations are at least 20%, and some have even been found to contain 40% THC [138, 139]. High levels of THC are associated with an increased risk of psychosis and, due to the synergistic action of THC and CBD, this is particularly evident when CBD concentrations are low [140].

1.2.5 Offspring environmental exposures *in utero*

The theory that the intrauterine developmental environment can affect disease risk in childhood and into adult life is widely accepted [141]. One such risk factor for disease in later life may be aberrant DNA methylation patterns, induced by environmental exposures *in utero*. For example, exposure to toxins during development can lead to altered DNA methylation in offspring [142-144]. Thus, while we know that DNA methylation is dynamic and that its distribution can change in response to environmental factors, the extent to which these environmental factors can affect the DNA methylation patterns of the developing offspring is not yet clear. Further, just like somatic cells, DNA methylation patterns of adult germ cells can be affected by the environment, raising the possibility that DNA methylation marks that have been altered in germ cells by environmental exposure will be passed onto the next generation [145, 146]. While it is usually the case that most DNA methylation marks are erased during germ cell maturation and early embryonic development, methylation at some CpG sites may persist through this process [147-149], potentially permanently altering offspring DNA methylation patterns. Therefore, there are multiple routes through which the maternal environment can alter offspring DNA methylation, with potential downstream consequences for gene expression and phenotypes.

Differential DNA methylation that occurs during embryogenesis can result in what has become known as metastable epialleles [150]. Metastable epialleles can be generated during the vulnerable time of demethylation and then re-methylation, where DNA methylation patterns are (mostly) erased and re-established. Any environmental exposure at this sensitive time that alters DNA methylation patterns therefore can lead to regions in the genome that are distinctly variable between identical individuals, due to alteration by an environmental stimulus *in utero* [151]. Thus far, the agouti mouse model in which nutritional alterations to maternal diet led to differences in phenotype, has offered the best understanding of metastable epialleles [147, 151-153].

A series of Human studies using a cohort of individuals form a Gambian tribe showed, deprivation of nutrients during seasonal changes have also provided evidence for the development of metastable epialleles as a concept [154, 155]. However, due to the nature of metastable epialleles being established during so early in development, it is very hard to pinpoint the precise time that the genomes of developing offspring are most sensitive to environmental exposure. To further understand metastable epialleles and their role in disease phenotypes, future work investigating DNA methylation differences induced by environmental exposures over the whole of the embryogenesis period need to be examined.

1.2.6 Tobacco *in utero*

It is widely known that tobacco smoking adversely influences every organ in the body, causing the onset of disease that then reduces the health of a smoker substantially [156]. Maternal tobacco smoking, particularly during pregnancy, is considered to be the single largest modifiable lifestyle risk factor to adverse child development [157]. Cigarettes contain upwards of 600 ingredients, and when these are burned they contain over 7000 chemicals. Some of these 7000 chemicals can pass through the placenta [158], and there is an association between miscarriage rate and women who smoke tobacco during pregnancy [159]. Pregnancies are also more likely to have complications such as preterm delivery, lower birth weight, lung problems, and sudden infant death syndrome [160], all of which lead to perinatal compromise, or poor infant health [161]. Later-life outcomes of children whose mothers smoked tobacco during their pregnancy have shown associations with behavioural disorders such as autism,

attention deficit hyperactivity disorder (ADHD) and oppositional defiant disorder (ODD) [162], suggesting a link between *in utero* tobacco exposure and behavioural problems, which are collectively termed *conduct problem* (CP) phenotypes.

Consequences of *in utero* tobacco exposure can still occur postnatally; mothers who consumed tobacco during their pregnancy will continue to expose the new-born to tobacco, with adverse effects on their health [161]. In mouse models, environmental tobacco smoke exposure during critical periods of brain development showed pathogenesis of regions of the brain involved in sudden infant death and susceptibility to addiction [163], again suggesting a link between tobacco use during pregnancy, perinatal compromise and adverse health in later life. While the complex nature of these disorders means that it is almost impossible to identify a direct correlation between a handful of genes and the disease phenotype. However, given the impact of DNA methylation on brain plasticity and addiction, the role of DNA methylation in brain development, and the impact of tobacco on DNA methylation, we suggest that DNA methylation plays a crucial role in the link between maternal tobacco use during pregnancy and CP in exposed offspring.

1.3 The zebrafish

While DNA methylation changes are important, they can be considered a proxy – differential methylation can signal genomic regions that may be implicated in biologically interesting phenomena, but in order to prove that methylation changes have caused a measurable genomic and phenotypic change, it is imperative to link such methylation changes to changes in genome output (gene expression), and to correlate this with a phenotypic outcome. For instance, Genome wide association studies (GWAS) have identified chromosomal regions that appear to be involved in substance dependence including cannabis [164], but this information is not definitive. Answering questions such as these would help to emphatically link a particular environment to a phenotype, via epigenetic mechanisms. Thus, in this thesis we aim to establish a tractable model system in which to explore the interaction between the environment and the epigenome.

One of the most commonly used model organisms is the zebrafish, *Danio Rerio*. The zebrafish has become an increasingly popular model organism in molecular biology [165], to study the links between the environment and traits such as disease risk and behaviour [166]. Their short generation time, transparency and rapid development outside of the mother make them a tractable model system in which to explore the effect of the environment on the genome, and on phenotypes [167].

Further benefits of zebrafish as a model system that make them highly appropriate and relevant to this project are:

- zebrafish have similar DNA methylation machinery to humans and there is consistent distribution of 5-methylcytosine between zebrafish and mammals [168];
- numerous studies have explored cannabis and cannabinoid biology using zebrafish [169-171];
- zebrafish are frequently used in studies of environmental toxicology [169-171] [173, 174];
- cannabinoids induce behavioural effects in zebrafish that are comparable to some of those reported for mammals [169];
- there is widespread literature on behavioural assays in zebrafish that can test learning, memory and cognition [172] which have shown to be impaired in long term cannabis usage;
- many basic cellular and molecular pathways, regulated by different compounds, are similar between zebrafish and mammals [173, 174];
- their abundance of progeny produced (up to 50 embryos at one time) and their rapid time from fertilisation to completion of organogenesis (5 days post-fertilisation, dpf) means they are a time-efficient model [175]

The zebrafish genome sequencing project was initiated at the Wellcome Trust Sanger Institute and published in 2013 [176]. Subsequently it has become apparent how similar at a genetic level humans and zebrafish are, yet phenotypically very divergent from one another. Approximately 70% of all human genes have at least one functional homolog in zebrafish, providing evidence of more than 26,000 protein coding genes

that have the potential to be studied [176]. Zebrafish share genetic similarities with humans across many different organelles; the brain, digestive tract, musculature, vasculature, and innate immune system are all physiologically comparable. Due to this, diseases such as depression [177], autism [178], psychoses [179] and muscular dystrophies [180] can all be modelled in zebrafish [181].

Although other established model systems such as the fruitfly (*Drosophila melanogaster)* and the nematode worm (*Caenorhabditis elegans)* have some similar benefits to zebrafish (mass production and fast development) they lack the same 5-methylcytosine machinery exhibited by humans, which is conserved in zebrafish [182]. Additionally, there is a paucity of 5-methylcytosine in both fruitfly and nematodes [183]. Thus, given our focus on DNA methylation in this research, and coupled with our necessity to model the human condition, fruitfly, rodents and nematode systems are not suitable here with the research facilities available. As such, zebrafish were chosen to model the genomic and phenotypic consequences of environmentally-induced methylation changes in this research..

1.4 Summary

In the past decade, advancements within the field of epigenetics have unravelled a link between DNA methylation and human development and disease. As stated earlier, the epigenome is a complex and dynamic structure. Clearly, genomic variability and inheritance is not limited to genes alone, and our understanding of genomics is shifting - it is now commonly accepted that some phenotypic variation is environmentally induced, and that this 'missing heritability' (that which cannot be accounted for by DNA sequence alone) may be partly explained by epigenetics [184]. Epigenetic alterations such as DNA methylation are an important source of variation and regulation in the genome. Methylation is one of the most well studied epigenetic alterations, and it is dynamic, with the ability to impact gene expression. Nutrition, toxins, alcohol, and stress are just some of the various environmental factors that can cause DNA methylation changes and then also have an influence on gene expression. The evidence for the impact of epigenetic effects in health and disease is now unequivocal, but we do not understand the mechanisms underlying this effect. The work will directly

address these questions and will have broad applicability to our understanding of health, disease, wellbeing, and future health outcomes.

1.5 Statement of research

This research addresses the fundamental question of how the environment can alter DNA methylation.

Initial work will understand the impact of heavy cannabis use in the human genome. We will then use a targeted tool for establishing a pipeline for assessing regions of the genome for variants in DNA methylation. From there, we expand on our findings by using the model system, the zebrafish, to develop a tractable in-house model to link differential methylation with gene expression, facilitating the exploration of pathways involved with the biological response to cannabis that may be modified by epigenetic processes.

We then will assess the impact of maternal tobacco use during pregnancy on offspring DNA methylation, and its association with conduct problem, in both a targeted and genome-wide manner. Here, we will look for associations between induced methylation patterns and changes to behavioural output and social interaction.

Lastly, we will discuss the issue of sample type in studies of DNA methylation and whether associations between phenotypes and DNA methylation are consistent across different tissue types. It will be conducted as a meta-analysis using schizophrenia as a case study.

1.6 Research Design (Objectives)

The overall aim of this study is to further our understanding of the extent to which DNA methylation may change when exposed to specific environmental factors. To achieve this, the following aims will be carried out:

Chapter 2: Assess genome-wide DNA methylation alterations in response to heavy cannabis exposure, using the Illumina EPIC array system, and a cohort of individuals from the Christchurch Health and Development Study (CHDS);

Chapter 3: Validate differential DNA methylation observed via EPIC array, using a targeted bisulfite-based amplicon sequencing (BSAS) approach;

Chapter 4: Develop the zebrafish as a model for assessing the impact of THC and CBD on DNA methylation

Chapter 5: Using individuals from the CHDS cohort, quantify differential DNA methylation in individuals who were exposed to tobacco smoke during development (*in utero*). Analyse whether there is an association between maternal tobacco use during pregnancy and the development of conduct problem (CP) in offspring, at genes associated with neurodevelopment and CP phenotypes, using BSAS;

Chapter 6: Quantify genome-wide differential DNA methylation in response to maternal tobacco use during pregnancy, and probe the interaction between tobacco exposure during development and the onset of CP in offspring;

Chapter 7: Analyse whether choice of tissue is a limiting factor in detecting biologically relevant DNA methylation differences, by using publicly available data and assessing DNA methylation differences in individuals with schizophrenia.

Chapter 8: General discussion of the significance of the findings contained within this thesis, and suggestions for future research.

1.7 List of attributions of collaborative contributions to work in this thesis

Chapters 2, 3, 5 and 7 have all been submitted for publication and so there is some repetition of background in some cases.

Chapter 2

Blood samples for DNA extraction were provided by the Christchurch Health and Development Study. Sample extraction and quantification of DNA was under taken by Dr. Amy Osborne. Australian Genomics Research Facility (AGRF, Melbourne, VIC, Australia) processed the Infinium® Methylation EPIC BeadChip (Illumina, San Diego, CA USA).The candidate carried out all bioinformatics analysis with guidance from A/P John Pearson (University of Otago). Critical discussion was undertaken by Prof Martin Kennedy, Dr Miles Benton, Dr Donia Macartney-Coxson and Prof Neil Gemmell.

The data analysis in this chapter contributed to: *Osborne and Pearson, et al, (2020) Genome-wide DNA methylation analysis of heavy cannabis exposure in a New Zealand longitudinal cohort, Translational Psychiatry.*

Chapter 3

Blood samples for DNA extraction were provided by the CHDS. All laboratory and bioinformatics work was carried out by the candidate and Dr. Amy Osborne provided input into primer design and critical analysis of this work. Sequence libraries were prepared using the Illumina MiSeq™ 500 cycle Kit V2, and sequenced on an Illumina MiSeq™ system at Massey Genome Services (Palmerston North, New Zealand). Further bioinformatics guidance was provided by A/P John Pearson. Critical discussions around the research subject were undertaken with Prof Martin Kennedy and Prof Neil Gemmel.

The data in this chapter contributed to: *Noble et al, (2021) A validation of Illumina EPIC array system with bisulfite-based amplicon sequencing, Peer J.*

Chapter 4

Embryos were provided by the Otago Zebrafish Facility (Dunedin, New Zealand). All laboratory work and bioinformatics was carried out by the candidate. With critical analysis of this work from Dr. Amy Osborne, A/P John Pearson, and Prof Martin Kennedy.

Chapter 5

Blood samples for DNA extraction were provided by the CHDS. All laboratory work and bioinformatics was carried out by the candidate. Sequence libraries were prepared using the Illumina MiSeq™ 500 cycle Kit V2, and sequenced on an Illumina MiSeq™ system at Massey Genome Services (Palmerston North, New Zealand). Further bioinformatics guidance was provided by A/P John Pearson. Critical analysis provided by Dr Amy Osborne and Prof Martin Kennedy.

Chapter 6

Blood samples for DNA extraction were provided by the CHDS. All lab work and bioinformatics was carried by the candidate. Australian Genomics Research Facility (AGRF, Melbourne, VIC, Australia) processed the Infinium® Methylation EPIC BeadChip (Illumina, San Diego, CA USA).The candidate carried out all bioinformatics analysis with guidance from A/P John Pearson (University of Otago). Dr Amy Osborne and Martin Kennedy provided critical analysis into this work.

Chapter 7

All bioinformatics was carried out by candidate with support and critique by Dr Amy Osborne.

The data from this chapter is under revision at Frontiers Genetics.

1.8 References

1. Waddington, C.H., *Canalization of development and the inheritance of acquired characters.* Nature, 1942. **150**: p. 563.
2. Waddington, c.h., *genetic assimilation of an acquired character.* Evolution, 1953. **7**(2): p. 118-126.
3. Wong, C.C., J. Mill, and C. Fernandes, *Drugs and addiction: an introduction to epigenetics.* Addiction, 2011. **106**(3): p. 480-489.
4. Waddington, C.H., *The Strategy of the Genes; a Discussion of Some Aspects of Theoretical Biology.* Allen & Unwin, London, 1957.
5. Goldberg, A.D., C.D. Allis, and E. Bernstein, *Epigenetics: A Landscape Takes Shape.* Cell, 2007. **128**(4): p. 635-638.
6. Reik, W., *Stability and flexibility of epigenetic gene regulation in mammalian development.* Nature, 2007. **447**: p. 425.
7. Fusco, G., R. Carrer, and E. Serrelli, *The Landscape Metaphor in Development*. 2014. p. 114-128.
8. Hackett, J.A. and M.A. Surani, *DNA methylation dynamics during the mammalian life cycle.* Philosophical Transactions of the Royal Society B: Biological Sciences, 2013. **368**(1609): p. 20110328.
9. Su, Z., L. Han, and Z. Zhao, *Conservation and divergence of DNA methylation in eukaryotes: new insights from single base-resolution DNA methylomes.* Epigenetics, 2011. **6**(2): p. 134-40.
10. Goll, M.G., et al., *Methylation of tRNAAsp by the DNA methyltransferase homolog Dnmt2.* Science, 2006. **311**(5759): p. 395-8.
11. Seisenberger, S., et al., *Reprogramming DNA methylation in the mammalian life cycle: building and breaking epigenetic barriers.* Philos Trans R Soc Lond B Biol Sci, 2013. **368**(1609): p. 20110330.
12. Mayer, W., et al., *Demethylation of the zygotic paternal genome.* Nature, 2000. **403**(6769): p. 501-502.
13. Seisenberger, S., et al., *Reprogramming DNA methylation in the mammalian life cycle: building and breaking epigenetic barriers.* Philosophical transactions of the Royal Society of London. Series B, Biological sciences, 2013. **368**(1609): p. 20110330-20110330.
14. Jin, B., Y. Li, and K.D. Robertson, *DNA Methylation: Superior or Subordinate in the Epigenetic Hierarchy?* Genes & Cancer, 2011. **2**(6): p. 607-617.
15. Li, E., T.H. Bestor, and R. Jaenisch, *Targeted mutation of the DNA methyltransferase gene results in embryonic lethality.* Cell, 1992. **69**(6): p. 915-26.
16. Mitchell, C., L.M. Schneper, and D.A. Notterman, *DNA methylation, early life environment, and health outcomes.* Pediatric research, 2016. **79**(1-2): p. 212-219.
17. Fraga, M.F., et al., *Epigenetic differences arise during the lifetime of monozygotic twins.* Proc Natl Acad Sci U S A, 2005. **102**(30): p. 10604-9.
18. Wong, C.C.Y., et al., *A longitudinal study of epigenetic variation in twins.* Epigenetics, 2010. **5**(6): p. 516-526.
19. Bocklandt, S., et al., *Epigenetic Predictor of Age.* PLOS ONE, 2011. **6**(6): p. e14821.
20. Hannum, G., et al., *Genome-wide Methylation Profiles Reveal Quantitative Views of Human Aging Rates.* Molecular Cell, 2013. **49**(2): p. 359-367.
21. Edwards, J.R., et al., *DNA methylation and DNA methyltransferases.* Epigenetics & Chromatin, 2017. **10**(1): p. 23.
22. Goll, M.G. and T.H. Bestor, *Eukaryotic cytosine methyltransferases.* (0066-4154 (Print)).
23. Jeziorska, D.M., et al., *DNA methylation of intragenic CpG islands depends on their transcriptional activity during differentiation and disease.* Proceedings of the National Academy of Sciences, 2017. **114**(36): p. E7526-E7535.
24. Takai, D. and P.A. Jones, *Comprehensive analysis of CpG islands in human chromosomes 21 and 22.* Proc Natl Acad Sci U S A, 2002. **99**(6): p. 3740-5.
25. Mariño-Ramírez, L., et al., *Promoter Analysis: Gene Regulatory Motif Identification with A-GLAM.* Methods in molecular biology (Clifton, N.J.), 2009. **537**: p. 263-276.
26. Zhang, M.Q., *Computational analyses of eukaryotic promoters.* BMC Bioinformatics, 2007. **8**(Suppl 6): p. S3-S3.
27. Portela, A. and M. Esteller, *Epigenetic modifications and human disease.* Nature Biotechnology, 2010. **28**: p. 1057.

28. Deaton, A.M. and A. Bird, *CpG islands and the regulation of transcription.* Genes & development, 2011. **25**(10): p. 1010-1022.
29. Bird, A., *DNA methylation patterns and epigenetic memory.* Genes Dev, 2002. **16**(1): p. 6-21.
30. Weinberg, M.S. and K.V. Morris, *Transcriptional gene silencing in humans.* Nucleic Acids Research, 2016. **44**(14): p. 6505-6517.
31. Edith Heard, a. Philippe Clerc, and P. Avner, *X-CHROMOSOME INACTIVATION IN MAMMALS.* Annual Review of Genetics, 1997. **31**(1): p. 571-610.
32. Agrelo, R. and A. Wutz, *ConteXt of change--X inactivation and disease.* EMBO Mol Med, 2010. **2**(1): p. 6-15.
33. Struhl, K., *Is DNA methylation of tumour suppressor genes epigenetic?* eLife, 2014. **3**: p. e02475-e02475.
34. Robertson, K.D., *DNA methylation and human disease.* Nature Reviews Genetics, 2005. **6**(8): p. 597-610.
35. Denis, H., M.N. Ndlovu, and F. Fuks, *Regulation of mammalian DNA methyltransferases: a route to new mechanisms.* EMBO reports, 2011. **12**(7): p. 647-656.
36. Klose, R.J. and A.P. Bird, *Genomic DNA methylation: the mark and its mediators.* Trends in Biochemical Sciences, 2006. **31**(2): p. 89-97.
37. Robertson, K.D., et al., *The human DNA methyltransferases (DNMTs) 1, 3a and 3b: coordinate mRNA expression in normal tissues and overexpression in tumors.* (0305-1048 (Print)).
38. Bestor, T.H., *The DNA methyltransferases of mammals.* Human molecular genetics, 2000. **9**(16): p. 2395-2402.
39. Du, Q., Z. Wang, and V.L. Schramm, *Human DNMT1 transition state structure.* Proceedings of the National Academy of Sciences, 2016. **113**(11): p. 2916-2921.
40. Tuorto, F., et al., *The tRNA methyltransferase Dnmt2 is required for accurate polypeptide synthesis during haematopoiesis.* (1460-2075 (Electronic)).
41. Ashapkin, V., L. Kutueva, and B. Vanyushin, *Dnmt2 is the most evolutionary conserved and enigmatic cytosine DNA methyltransferase in eukaryotes.* Russian journal of genetics, 2016. **52**(3): p. 237-248.
42. Okano, M., et al., *DNA methyltransferases Dnmt3a and Dnmt3b are essential for de novo methylation and mammalian development.* (0092-8674 (Print)).
43. Okano, M., et al., *DNA Methyltransferases Dnmt3a and Dnmt3b Are Essential for De Novo Methylation and Mammalian Development.* Cell, 1999. **99**(3): p. 247-257.
44. Bourc'his, D., et al., *Dnmt3L and the establishment of maternal genomic imprints.* Science, 2001. **294**(5551): p. 2536-2539.
45. Li, E., T.H. Bestor, and R. Jaenisch, *Targeted mutation of the DNA methyltransferase gene results in embryonic lethality.* Cell, 1992. **69**(6): p. 915-926.
46. Susan, J.C., et al., *High sensitivity mapping of methylated cytosines.* Nucleic Acids Research, 1994. **22**(15): p. 2990-2997.
47. Booth, M.J., et al., *Oxidative bisulfite sequencing of 5-methylcytosine and 5-hydroxymethylcytosine.* Nature Protocols, 2013. **8**: p. 1841.
48. Masser, D.R., D.R. Stanford, and W.M. Freeman, *Targeted DNA Methylation Analysis by Next-generation Sequencing.* Journal of Visualized Experiments : JoVE, 2015(96): p. 52488.
49. Cline, J., J.C. Braman, and H.H. Hogrefe, *PCR Fidelity of Pfu DNA Polymerase and Other Thermostable DNA Polymerases.* Nucleic Acids Research, 1996. **24**(18): p. 3546-3551.
50. Ronaghi, M., *Pyrosequencing Sheds Light on DNA Sequencing.* Genome Research, 2001. **11**(1): p. 3-11.
51. Siqueira, J.F., Jr., A.F. Fouad, and I.N. Rôças, *Pyrosequencing as a tool for better understanding of human microbiomes.* Journal of oral microbiology, 2012. **4**: p. 10.3402/jom.v4i0.10743.
52. França, L.T., E. Carrilho, and T.B. Kist, *A review of DNA sequencing techniques.* Quarterly reviews of biophysics, 2002. **35**(2): p. 169-200.
53. Zhou, W., P.W. Laird, and H. Shen, *Comprehensive characterization, annotation and innovative use of Infinium DNA methylation BeadChip probes.* Nucleic acids research, 2017. **45**(4): p. e22-e22.
54. Staunstrup, N.H., et al., *Genome-wide DNA methylation profiling with MeDIP-seq using archived dried blood spots.* Clinical epigenetics, 2016. **8**: p. 81-81.
55. Taiwo, O., et al., *Methylome analysis using MeDIP-seq with low DNA concentrations.* Nature Protocols, 2012. **7**: p. 617.

56. Nair, S.S., 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): p. 34-44.
57. Clark, C., et al., *A comparison of the whole genome approach of MeDIP-seq to the targeted approach of the Infinium HumanMethylation450 BeadChip(®) for methylome profiling.* PloS one, 2012. **7**(11): p. e50233-e50233.
58. Vidal, E., et al., *A DNA methylation map of human cancer at single base-pair resolution.* Oncogene, 2017. **36**(40): p. 5648-5657.
59. Brinkman, A.B., et al., *Partially methylated domains are hypervariable in breast cancer and fuel widespread CpG island hypermethylation.* Nature Communications, 2019. **10**(1): p. 1749.
60. Olova, N., et al., *Comparison of whole-genome bisulfite sequencing library preparation strategies identifies sources of biases affecting DNA methylation data.* Genome Biology, 2018. **19**(1): p. 33.
61. Ji, L., et al., *Methylated DNA is over-represented in whole-genome bisulfite sequencing data.* Frontiers in genetics, 2014. **5**: p. 341-341.
62. Gu, H., et al., *Preparation of reduced representation bisulfite sequencing libraries for genome-scale DNA methylation profiling.* Nature Protocols, 2011. **6**: p. 468.
63. Gu, H., et al., *Genome-scale DNA methylation mapping of clinical samples at single-nucleotide resolution.* Nature methods, 2010. **7**(2): p. 133-136.
64. Lu, H., F. Giordano, and Z. Ning, *Oxford Nanopore MinION Sequencing and Genome Assembly.* Genomics, Proteomics & Bioinformatics, 2016. **14**(5): p. 265-279.
65. Laszlo, A.H., et al., *Detection and mapping of 5-methylcytosine and 5-hydroxymethylcytosine with nanopore MspA.* Proceedings of the National Academy of Sciences, 2013. **110**(47): p. 18904-18909.
66. Wescoe, Z.L., J. Schreiber, and M. Akeson, *Nanopores Discriminate among Five C5-Cytosine Variants in DNA.* Journal of the American Chemical Society, 2014. **136**(47): p. 16582-16587.
67. Liu, Y., et al., *Accurate targeted long-read DNA methylation and hydroxymethylation sequencing with TAPS.* Genome Biology, 2020. **21**(1): p. 54.
68. Kurdyukov, S. and M. Bullock, *DNA Methylation Analysis: Choosing the Right Method.* Biology, 2016. **5**(1): p. 3.
69. Armstrong, K.M., et al., *Global DNA methylation measurement by HPLC using low amounts of DNA.* Biotechnol J, 2011. **6**(1): p. 113-7.
70. Skene, N.G., et al., *Genetic identification of brain cell types underlying schizophrenia.* Nature Genetics, 2018. **50**(6): p. 825-833.
71. Pidsley, R., et al., *Critical evaluation of the Illumina MethylationEPIC BeadChip microarray for whole-genome DNA methylation profiling.* Genome biology, 2016. **17**(1): p. 208-208.
72. Tost, J. and I.G. Gut, *DNA methylation analysis by pyrosequencing.* Nature Protocols, 2007. **2**: p. 2265.
73. Delaney, C., S.K. Garg, and R. Yung, *Analysis of DNA Methylation by Pyrosequencing.* Methods in molecular biology (Clifton, N.J.), 2015. **1343**: p. 249-264.
74. Yang, A.S., et al., *A simple method for estimating global DNA methylation using bisulfite PCR of repetitive DNA elements.* Nucleic acids research, 2004. **32**(3): p. e38-e38.
75. McGregor, K., et al., *An evaluation of methods correcting for cell-type heterogeneity in DNA methylation studies.* Genome biology, 2016. **17**: p. 84-84.
76. Fasanelli, F., et al., *Hypomethylation of smoking-related genes is associated with future lung cancer in four prospective cohorts.* Nature communications, 2015. **6**: p. 10192-10192.
77. Reynolds, L.M., et al., *DNA Methylation of the Aryl Hydrocarbon Receptor Repressor Associations With Cigarette Smoking and Subclinical Atherosclerosis.* Circ Cardiovasc Genet, 2015. **8**(5): p. 707-16.
78. Reynolds Lindsay, M., et al., *DNA Methylation of the Aryl Hydrocarbon Receptor Repressor Associations With Cigarette Smoking and Subclinical Atherosclerosis.* Circulation: Cardiovascular Genetics, 2015. **8**(5): p. 707-716.
79. Zeilinger, S., et al., *Tobacco Smoking Leads to Extensive Genome-Wide Changes in DNA Methylation.* PLoS ONE, 2013. **8**(5): p. e63812.
80. Philibert, R.A., et al., *Changes in DNA methylation at the aryl hydrocarbon receptor repressor may be a new biomarker for smoking.* Clinical Epigenetics, 2013. **5**(1): p. 19.
81. Monick, M.M., et al., *Coordinated changes in AHRR methylation in lymphoblasts and pulmonary macrophages from smokers.* American Journal of Medical Genetics Part B: Neuropsychiatric Genetics, 2012. **159B**(2): p. 141-151.

82. Monick, M.M., et al., *Coordinated changes in AHRR methylation in lymphoblasts and pulmonary macrophages from smokers.* American journal of medical genetics part B: neuropsychiatric genetics, 2012. **159**(2): p. 141-151.
83. Kanherkar, R.R., N. Bhatia-Dey, and A.B. Csoka, *Epigenetics across the human lifespan.* Frontiers in cell and developmental biology, 2014. **2**: p. 49.
84. Peaston, A.E. and E. Whitelaw, *Epigenetics and phenotypic variation in mammals.* Mammalian genome : official journal of the International Mammalian Genome Society, 2006. **17**(5): p. 365-374.
85. Ragsdale, A.K., et al., *Epigenetic response to environmental change: DNA methylation varies with invasion status.* Environmental Epigenetics, 2016. **2**(2).
86. Bollati, V. and A. Baccarelli, *Environmental Epigenetics.* Heredity, 2010. **105**(1): p. 105-112.
87. Anderson, O.S., K.E. Sant, and D.C. Dolinoy, *Nutrition and epigenetics: an interplay of dietary methyl donors, one-carbon metabolism and DNA methylation.* The Journal of nutritional biochemistry, 2012. **23**(8): p. 853-859.
88. Ruiz-Hernandez, A., et al., *Environmental chemicals and DNA methylation in adults: a systematic review of the epidemiologic evidence.* Clinical epigenetics, 2015. **7**(1): p. 55-55.
89. Hing, B., C. Gardner, and J.B. Potash, *Effects of negative stressors on DNA methylation in the brain: implications for mood and anxiety disorders.* American journal of medical genetics. Part B, Neuropsychiatric genetics : the official publication of the International Society of Psychiatric Genetics, 2014. **165B**(7): p. 541-554.
90. Romens, S.E., et al., *Associations between early life stress and gene methylation in children.* Child development, 2015. **86**(1): p. 303-309.
91. Unternaehrer, E., et al., *Dynamic changes in DNA methylation of stress-associated genes (OXTR, BDNF) after acute psychosocial stress.* Translational Psychiatry, 2012. **2**: p. e150.
92. Liu, C., et al., *A DNA methylation biomarker of alcohol consumption.* (1476-5578 (Electronic)).
93. Breitling, L.P., et al., *Tobacco-smoking-related differential DNA methylation: 27K discovery and replication.* (1537-6605 (Electronic)).
94. Nielsen, D.A., et al., *Epigenetics of drug abuse: predisposition or response.* Pharmacogenomics, 2012. **13**(10): p. 1149-1160.
95. Levenson, J.M. and J.D. Sweatt, *Epigenetic mechanisms in memory formation.* (1471-003X (Print)).
96. Fuchs, E. and G. Flügge, *Adult neuroplasticity: more than 40 years of research.* Neural plasticity, 2014. **2014**: p. 541870-541870.
97. Bayraktar, G. and M.R. Kreutz, *Neuronal DNA Methyltransferases: Epigenetic Mediators between Synaptic Activity and Gene Expression?* The Neuroscientist, 2018. **24**(2): p. 171-185.
98. Jung, Y., et al., *An epigenetic mechanism mediates developmental nicotine effects on neuronal structure and behavior.* Nature Neuroscience, 2016. **19**: p. 905.
99. Kendler, K.S., et al., *The structure of genetic and environmental risk factors for common psychiatric and substance use disorders in men and women.* Arch Gen Psychiatry, 2003. **60**(9): p. 929-37.
100. Kendler, K.S., et al., *Genetic and Environmental Influences on Alcohol, Caffeine, Cannabis, and Nicotine Use From Early Adolescence to Middle Adulthood.* Archives of general psychiatry, 2008. **65**(6): p. 674-682.
101. Breitling, Lutz P., et al., *Tobacco-Smoking-Related Differential DNA Methylation: 27K Discovery and Replication.* The American Journal of Human Genetics, 2011. **88**(4): p. 450-457.
102. Bridgeman, M.B. and D.T. Abazia, *Medicinal Cannabis: History, Pharmacology, And Implications for the Acute Care Setting.* Pharmacy and Therapeutics, 2017. **42**(3): p. 180.
103. Hall, W. and L. Degenhardt, *Cannabis use and the risk of developing a psychotic disorder.* World psychiatry : official journal of the World Psychiatric Association (WPA), 2008. **7**(2): p. 68-71.
104. Fergusson, D. and J. Boden, *Cannabis use in adolescence.* Improving the Transition, 2011. **257**: p. 239-253.
105. Fergusson, D.M. and L.J. Horwood, *Cannabis use and dependence in a New Zealand birth cohort.* N Z Med J, 2000. **113**(1109): p. 156-8.
106. Fergusson D, B.J., *Cannabis Use in Adolescence, Improving the Transition: Reducing Social and Psychological Morbidity During Adolescence.*, O.o.t.P.M.s.S.A. Committee, Editor. 2011. p. 257-271.
107. Fergusson, D.M., L.J. Horwood, and N.R. Swain-Campbell, *Cannabis dependence and psychotic symptoms in young people.* Psychological Medicine, 2003. **33**(1): p. 15-21.

108. Kalant, H., *Adverse effects of cannabis on health: an update of the literature since 1996.* Progress in neuro-psychopharmacology and biological psychiatry, 2004. **28**(5): p. 849-863.
109. Hayatbakhsh, M.R., et al., *Cannabis and anxiety and depression in young adults: a large prospective study.* Journal of the American Academy of Child & Adolescent Psychiatry, 2007. **46**(3): p. 408-417.
110. Patton, G.C., et al., *Cannabis use and mental health in young people: cohort study.* Bmj, 2002. **325**(7374): p. 1195-1198.
111. Vassoler, F.M., N.L. Johnson, and E.M. Byrnes, *Female adolescent exposure to cannabinoids causes transgenerational effects on morphine sensitization in female offspring in the absence of in utero exposure.* Journal of psychopharmacology, 2013. **27**(11): p. 1015-1022.
112. Watson, C.T., et al., *Genome-wide DNA methylation profiling reveals epigenetic changes in the rat nucleus accumbens associated with cross-generational effects of adolescent THC exposure.* Neuropsychopharmacology, 2015. **40**(13): p. 2993.
113. Szutorisz, H. and Y.L. Hurd, *Epigenetic Effects of Cannabis Exposure.* Biological psychiatry, 2016. **79**(7): p. 586-594.
114. ORGAINZATION, W.H., 2016.
115. Institute of Medicine (US); Joy JE, W.S.J., Benson JA Jr., editors, *Marijuana and Medicine: Assessing the Science Base.* National Academies Press (US), 1991. **1**.
116. Baggio, S., et al., *Routes of administration of cannabis used for nonmedical purposes and associations with patterns of drug use.* J Adolesc Health, 2014. **54**(2): p. 235-40.
117. Loflin, M. and M. Earleywine, *A new method of cannabis ingestion: The dangers of dabs?* Addictive Behaviors, 2014. **39**(10): p. 1430-1433.
118. Jacobus, J., et al., *Functional consequences of marijuana use in adolescents.* Pharmacology, biochemistry, and behavior, 2009. **92**(4): p. 559-565.
119. Jarvis, S., S. Rassmussen, and B. Winters, *Role of the Endocannabinoid System and Medical Cannabis.* The Journal for Nurse Practitioners, 2017. **13**(8): p. 525-531.
120. Herkenham, M., et al., *Characterization and localization of cannabinoid receptors in rat brain: a quantitative in vitro autoradiographic study.* J Neurosci, 1991. **11**(2): p. 563-83.
121. Marsicano, G. and R. Kuner, *Anatomical Distribution of Receptors, Ligands and Enzymes in the Brain and in the Spinal Cord: Circuitries and Neurochemistry*, in *Cannabinoids and the Brain*, A. Köfalvi, Editor. 2008, Springer US: Boston, MA. p. 161-201.
122. Tibiriça, E., *The multiple functions of the endocannabinoid system: a focus on the regulation of food intake.* Diabetology & metabolic syndrome, 2010. **2**: p. 5-5.
123. Bermudez-Silva, F.J., P. Cardinal, and D. Cota, *The role of the endocannabinoid system in the neuroendocrine regulation of energy balance.* Journal of Psychopharmacology, 2011. **26**(1): p. 114-124.
124. du Plessis, S.S., A. Agarwal, and A. Syriac, *Marijuana, phytocannabinoids, the endocannabinoid system, and male fertility.* Journal of assisted reproduction and genetics, 2015. **32**(11): p. 1575-1588.
125. Elphick, M.R., *The evolution and comparative neurobiology of endocannabinoid signalling.* Philosophical transactions of the Royal Society of London. Series B, Biological sciences, 2012. **367**(1607): p. 3201-3215.
126. Barbara Bosiera, G.G.M., EmmanuelHerman, sc1Didier, M.Lamberta1, *Functionally selective cannabinoid receptor signalling: Therapeutic implications and opportunities.* Biochemical Pharmacology, 2010. **80**(1): p. 1-12.
127. Köfalvi, A., et al., *Involvement of Cannabinoid Receptors in the Regulation of Neurotransmitter Release in the Rodent Striatum: A Combined Immunochemical and Pharmacological Analysis.* The Journal of Neuroscience, 2005. **25**(11): p. 2874-2884.
128. Wilson, R.I. and R.A. Nicoll, *Endogenous cannabinoids mediate retrograde signalling at hippocampal synapses.* Nature, 2001. **410**(6828): p. 588-92.
129. Ohno-Shosaku, T., T. Maejima, and M. Kano, *Endogenous cannabinoids mediate retrograde signals from depolarized postsynaptic neurons to presynaptic terminals.* Neuron, 2001. **29**(3): p. 729-38.
130. Alger, B.E., *Getting high on the endocannabinoid system.* Cerebrum : the Dana forum on brain science, 2013. **2013**: p. 14-14.
131. Fride, E., et al., *Chapter 6 The Endocannabinoid System During Development: Emphasis on Perinatal Events and Delayed Effects*, in *Vitamins & Hormones*. 2009, Academic Press. p. 139-158.
132. Aguado, T., et al., *The endocannabinoid system drives neural progenitor proliferation.* The FASEB Journal, 2005. **19**(12): p. 1704-1706.

133. Malfait, A.M., et al., *The nonpsychoactive cannabis constituent cannabidiol is an oral anti-arthritic therapeutic in murine collagen-induced arthritis.* Proceedings of the National Academy of Sciences, 2000. **97**(17): p. 9561-9566.
134. Liu, Y., et al., *Inhibition of hepatocarcinoma and tumor metastasis to liver by gene therapy with recombinant CBD-HepII polypeptide of fibronectin.* International Journal of Cancer, 2007. **121**(1): p. 184-192.
135. Watt, G. and T. Karl, *In vivo Evidence for Therapeutic Properties of Cannabidiol (CBD) for Alzheimer's Disease.* Frontiers in Pharmacology, 2017. **8**(20).
136. Kolikonda, M.K., et al., *Medical Marijuana for Epilepsy?* Innovations in clinical neuroscience, 2016. **13**(3-4): p. 23-26.
137. Johnson, J.R., et al., *Multicenter, double-blind, randomized, placebo-controlled, parallel-group study of the efficacy, safety, and tolerability of THC:CBD extract and THC extract in patients with intractable cancer-related pain.* J Pain Symptom Manage, 2010. **39**(2): p. 167-79.
138. Murray, R.M. and M. Di Forti, *Cannabis and Psychosis: What Degree of Proof Do We Require?* Biological Psychiatry, 2016. **79**(7): p. 514-515.
139. Russo, E.B., *Taming THC: potential cannabis synergy and phytocannabinoid-terpenoid entourage effects.* British journal of pharmacology, 2011. **163**(7): p. 1344-1364.
140. Di Forti, M., et al., *Proportion of patients in south London with first-episode psychosis attributable to use of high potency cannabis: a case-control study.* The Lancet Psychiatry, 2015. **2**(3): p. 233-238.
141. Gluckman, P.D., et al., *Effect of in utero and early-life conditions on adult health and disease.* The New England journal of medicine, 2008. **359**(1): p. 61-73.
142. Joubert Bonnie, R., et al., *450K Epigenome-Wide Scan Identifies Differential DNA Methylation in Newborns Related to Maternal Smoking during Pregnancy.* Environmental Health Perspectives, 2012. **120**(10): p. 1425-1431.
143. Hernandez-Vargas, H., et al., *Exposure to aflatoxin B1 in utero is associated with DNA methylation in white blood cells of infants in The Gambia.* International Journal of Epidemiology, 2015. **44**(4): p. 1238-1248.
144. Cardenas, A., et al., *Differential DNA methylation in umbilical cord blood of infants exposed to mercury and arsenic in utero.* Epigenetics, 2015. **10**(6): p. 508-515.
145. Szyf, M., *The genome- and system-wide response of DNA methylation to early life adversity and its implication on mental health.* Can J Psychiatry, 2013. **58**(12): p. 697-704.
146. Szyf, M., *Nongenetic inheritance and transgenerational epigenetics.* Trends Mol Med, 2015. **21**(2): p. 134-44.
147. Dolinoy, D.C., et al., *Metastable Epialleles, Imprinting, and the Fetal Origins of Adult Diseases.* Pediatric Research, 2007. **61**(7): p. 30-37.
148. Slieker, R.C., et al., *DNA Methylation Landscapes of Human Fetal Development.* PLOS Genetics, 2015. **11**(10): p. e1005583.
149. Nazor, Kristopher L., et al., *Recurrent Variations in DNA Methylation in Human Pluripotent Stem Cells and Their Differentiated Derivatives.* Cell Stem Cell, 2012. **10**(5): p. 620-634.
150. Rakyan, V.K., et al., *Metastable epialleles in mammals.* Trends in Genetics, 2002. **18**(7): p. 348-351.
151. Dolinoy, D.C., et al., *Variable histone modifications at the Avy metastable epiallele.* Epigenetics, 2010. **5**(7): p. 637-644.
152. Dolinoy, D.C., et al., *Maternal genistein alters coat color and protects Avy mouse offspring from obesity by modifying the fetal epigenome.* Environmental health perspectives, 2006. **114**(4): p. 567-572.
153. Dolinoy, D.C., *The agouti mouse model: an epigenetic biosensor for nutritional and environmental alterations on the fetal epigenome.* Nutrition Reviews, 2008. **66**(suppl_1): p. S7-S11.
154. Waterland, R.A., et al., *Season of Conception in Rural Gambia Affects DNA Methylation at Putative Human Metastable Epialleles.* PLOS Genetics, 2010. **6**(12): p. e1001252.
155. Dominguez-Salas, P., et al., *Maternal nutrition at conception modulates DNA methylation of human metastable epialleles.* Nature communications, 2014. **5**: p. 3746.
156. Maritz, G.S. and M. Mutemwa, *Tobacco smoking: patterns, health consequences for adults, and the long-term health of the offspring.* Global journal of health science, 2012. **4**(4): p. 62-75.
157. Baler, R.D., et al., *Is fetal brain monoamine oxidase inhibition the missing link between maternal smoking and conduct disorders?* J Psychiatry Neurosci, 2008. **33**(3): p. 187-95.
158. Wang, N., et al., *The effect of maternal prenatal smoking and alcohol consumption on the placenta-to-birth weight ratio.* Placenta, 2014. **35**(7): p. 437-441.

159. Economides, D. and J. Braithwaite, *Smoking, pregnancy and the fetus.* J R Soc Health, 1994. **114**(4): p. 198-201.
160. Office of the Surgeon, G., S. Office on, and Health, *Reports of the Surgeon General*, in *The Health Consequences of Smoking: A Report of the Surgeon General*. 2004, Centers for Disease Control and Prevention (US): Atlanta (GA).
161. Wehby, G.L., et al., *The Impact of Maternal Smoking during Pregnancy on Early Child Neurodevelopment.* Journal of human capital, 2011. **5**(2): p. 207-254.
162. Nomura, Y., D.J. Marks, and J.M. Halperin, *Prenatal exposure to maternal and paternal smoking on attention deficit hyperactivity disorders symptoms and diagnosis in offspring.* The Journal of nervous and mental disease, 2010. **198**(9): p. 672-678.
163. Torres, L.H., et al., *Exposure to tobacco smoke during the early postnatal period modifies receptors and enzymes of the endocannabinoid system in the brainstem and striatum in mice.* Toxicology Letters, 2018.
164. Hopfer, C.J., et al., *A genome-wide scan for loci influencing adolescent cannabis dependence symptoms: evidence for linkage on chromosomes 3 and 9.* Drug & Alcohol Dependence, 2007. **89**(1): p. 34-41.
165. Dooley, K. and L.I. Zon, *Zebrafish: a model system for the study of human disease.* Current opinion in genetics & development, 2000. **10**(3): p. 252-256.
166. Dai, Y.J., et al., *Zebrafish as a model system to study toxicology.* Environmental toxicology and chemistry, 2014. **33**(1): p. 11-17.
167. Nusslein-Volhard, C. and R. Dahm, *Zebrafish*. 2002: Oxford University Press.
168. Kamstra, J.H., et al., *Zebrafish as a model to study the role of DNA methylation in environmental toxicology.* Environmental Science and Pollution Research, 2015. **22**(21): p. 16262-16276.
169. Akhtar, M.T., et al., *Developmental effects of cannabinoids on zebrafish larvae.* Zebrafish, 2013. **10**(3): p. 283-293.
170. Akhtar, M.T., et al., *Metabolic effects of cannabinoids in zebrafish (Danio rerio) embryos determined by 1 H NMR metabolomics.* Metabolomics, 2016. **12**(3): p. 44.
171. Thomas, R.J., *The toxicologic and teratologic effects of Δ9-tetrahydrocannabinol in the Zebrafish embryo.* Toxicology and applied pharmacology, 1975. **32**(1): p. 184-190.
172. Levin ED, C.D., *Behavioral Neuroscience of Zebrafish.* 2 ed. Methods of Behavior Analysis in Neuroscience. 2009: CRC Press/Taylor & Francis.
173. Voelker, D., et al., *Differential gene expression as a toxicant-sensitive endpoint in zebrafish embryos and larvae.* Aquatic toxicology, 2007. **81**(4): p. 355-364.
174. Schaaf, M.J., et al., *Discovery of a functional glucocorticoid receptor β-isoform in zebrafish.* Endocrinology, 2007. **149**(4): p. 1591-1599.
175. Rubinstein, A.L., *Zebrafish: from disease modeling to drug discovery.* (1367-6733 (Print)).
176. Howe, K., et al., *The zebrafish reference genome sequence and its relationship to the human genome.* Nature, 2013. **496**(7446): p. 498.
177. Ziv, L., et al., *An affective disorder in zebrafish with mutation of the glucocorticoid receptor.* Mol Psychiatry, 2013. **18**(6): p. 681-91.
178. Stewart, A.M., et al., *Developing zebrafish models of autism spectrum disorder (ASD).* Prog Neuropsychopharmacol Biol Psychiatry, 2014. **50**: p. 27-36.
179. Burgess, H.A. and M. Granato, *Sensorimotor Gating in Larval Zebrafish.* The Journal of Neuroscience, 2007. **27**(18): p. 4984-4994.
180. Parsons, M.J., et al., *Removal of dystroglycan causes severe muscular dystrophy in zebrafish embryos.* Development, 2002. **129**(14): p. 3505-12.
181. Zhao, S., J. Huang, and J. Ye, *A fresh look at zebrafish from the perspective of cancer research.* Journal of Experimental & Clinical Cancer Research, 2015. **34**(1): p. 80.
182. Dunwell, T.L. and G.P. Pfeifer, *Drosophila genomic methylation: new evidence and new questions.* Epigenomics, 2014. **6**(5): p. 459-461.
183. Hu, C.W., et al., *Trace analysis of methylated and hydroxymethylated cytosines in DNA by isotope-dilution LC-MS/MS: first evidence of DNA methylation in Caenorhabditis elegans.* Biochem J, 2015. **465**(1): p. 39-47.
184. Trerotola, M., et al., *Epigenetic inheritance and the missing heritability.* Human Genomics, 2015. **9**(1): p. 17.

1.9 Packages used throughout this thesis (in order of appearance)

Minfi- A Bioconductor tool to analyse and visualise Illumina Infinium methylation arrays [1]

SWAN- Subset- quantile within array normalisation. This Normalisation package is intended to remove sources of technical variation between measurements via randomly selecting a subset of probes defined to be biologically similar based on CpG content. [2]

Funnorm- Functional normalisation package for Illumina Infinium methylation arrays. This package uses 848 control probes as well as out-of-band probes into 42 summary measurements. [3]

Noob- Normal-exponential out-of-band (noob) is a background correction method with dye-bias normalization for Illumina Infinium methylation arrays.[4]

Flow.sorted.blood- Raw data objects for the Illumina 450k DNA methylation microarrays, and an object depicting which CpGs on the array are associated with cell type.[5]

Limma- Data analysis, linear models and differential expression for microarray data.[6]

Bacon- Bacon can be used to remove inflation and bias often observed in epigenome- and transcriptome-wide association studies. To this end bacon constructs an empirical null distribution using a Gibbs Sampling algorithm by fitting a three-component normal mixture on z-scores. [7]

Granges- The ability to efficiently represent and manipulate genomic annotations and alignments is playing a central role when it comes to analysing high-throughput sequencing data (a.k.a. NGS data). The GenomicRanges package defines general purpose containers for storing and manipulating genomic intervals and variables defined along a genome. More specialized containers for representing and manipulating short alignments against a reference genome, or a matrix-like summarization of an experiment, are defined in the GenomicAlignments and SummarizedExperiment packages, respectively. Both packages build on top of the GenomicRanges infrastructure. [8]

EnrichR/FishenrichR- Enrichment analysis is a popular method for analysing gene sets generated by genome-wide experiments. [9]

Ggplot2- A system for 'declaratively' creating graphics, based on "The Grammar of Graphics". You provide the data, tell 'ggplot2' how to map variables to aesthetics, what graphical primitives to use, and it takes care of the details. [10]

Bisearch- A Web server (http://bisearch.enzim.hu), a primer design software created for designing primers to amplify such target sequences [11]

SolexaQA++- SolexaQA calculates sequence quality statistics and creates visual representations of data quality for second-generation sequencing data. Originally developed for the Illumina system. *[12]*

Bowtie2- bowtie2 is an ultrafast and memory-efficient tool for aligning sequencing reads to long reference sequences. It is particularly good at aligning reads of about 50 up to 100s or 1,000s of characters, and particularly good at aligning to relatively long (e.g. mammalian) genomes [13].

Bismark- Bismark is a program to map bisulfite treated sequencing reads to a genome of interest and perform methylation calls in a single step [14].

edgeR- edgeR performs differential abundance analysis for pre-defined genomic features [15].

Survival - Contains the core survival analysis routines, including definition of Surv objects, Kaplan-Meier and Aalen-Johansen (multi-state) curves, Cox models, and parametric accelerated failure time models.[16]

FastQC- FastQC aims to provide a simple way to do some quality control checks on raw sequence data coming from high throughput sequencing pipelines. It provides a modular set of analyses which you can use to give a quick impression of whether your data has any problems of which you should be aware before doing any further analysis.

Trim Galore- Trim Galore! is a wrapper script to automate quality and adapter trimming as well as quality control, with some added functionality to remove biased methylation positions for RRBS sequence files (for directional, non-directional (or paired-end) sequencing).

UpsetR- Creates visualizations of intersecting sets using a novel matrix design, along with visualizations of several common set, element and attribute related tasks [17]

References

1. Aryee, M.J., et al., *Minfi: a flexible and comprehensive Bioconductor package for the analysis of Infinium DNA methylation microarrays.* Bioinformatics, 2014. **30**(10): p. 1363-1369.
2. Maksimovic, J., L. Gordon, and A. Oshlack, *SWAN: Subset-quantile Within Array Normalization for Illumina Infinium HumanMethylation450 BeadChips.* Genome Biology, 2012. **13**(6): p. R44.
3. Fortin, J.-P., et al., *Functional normalization of 450k methylation array data improves replication in large cancer studies.* Genome Biology, 2014. **15**(11): p. 503.
4. Triche, T.J., Jr., et al., *Low-level processing of Illumina Infinium DNA Methylation BeadArrays.* Nucleic Acids Research, 2013. **41**(7): p. e90-e90.
5. Jaffe, A.E., *FlowSorted.Blood.450k: Illumina HumanMethylation data on sorted blood cell populations.* 2019.
6. Ritchie, M.E., et al., *limma powers differential expression analyses for RNA-sequencing and microarray studies.* Nucleic Acids Research, 2015. **43**(7): p. e47-e47.
7. van Iterson, M., et al., *Controlling bias and inflation in epigenome- and transcriptome-wide association studies using the empirical null distribution.* Genome Biology, 2017. **18**(1): p. 19.
8. Lawrence, M., et al., *Software for Computing and Annotating Genomic Ranges.* PLOS Computational Biology, 2013. **9**(8): p. e1003118.
9. Chen, E.Y., et al., *Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool.* BMC Bioinformatics, 2013. **14**(1): p. 128.
10. Wickham, H., *ggplot2: Elegant Graphics for Data Analysis*. 2016: Springer-Verlag New York.
11. Arányi, T. and G.E. Tusnády, *BiSearch: ePCR tool for native or bisulfite-treated genomic template.* Methods Mol Biol, 2007. **402**: p. 385-402.
12. Cox, M.P., D.A. Peterson, and P.J. Biggs, *SolexaQA: At-a-glance quality assessment of Illumina second-generation sequencing data.* BMC Bioinformatics, 2010. **11**(1): p. 485.
13. Langmead, B. and S.L. Salzberg, *Fast gapped-read alignment with Bowtie 2.* Nature methods, 2012. **9**(4): p. 357-359.
14. Krueger, F. and S.R. Andrews, *Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications.* Bioinformatics, 2011. **27**(11): p. 1571-2.
15. Chen, Y., et al., *Differential methylation analysis of reduced representation bisulfite sequencing experiments using edgeR [version 2; peer review: 2 approved, 1 approved with reservations].* F1000Research, 2018. **6**(2055).
16. Fox, J. and M.S. Carvalho, *The RcmdrPlugin.survival Package: Extending the R Commander Interface to Survival Analysis.* 2012, 2012. **49**(7): p. 32.
17. Conway, J.R., A. Lex, and N. Gehlenborg, *UpSetR: an R package for the visualization of intersecting sets and their properties.* Bioinformatics, 2017. **33**(18): p. 2938-2940.

Chapter 2

2. The impact of heavy cannabis use on DNA methylation in the human genome

2.1 Introduction

2.1.1 Cannabis use and implications

Cannabis is one of the most widely used recreational drugs in the world [1], and its use is increasing in frequency [2, 3]. It is a widely debated topic due to the psychoactive component THC, and implications on adolescence and later life outcomes [4]. Although the full effects of its use are largely still unknown, legalisation of cannabis for recreational use has occurred in some jurisdictions around the globe. More so, harnessing cannabis for medicinal purposes has increased [5], due to another other active cannabinoid, CBD, which is not psychoactive [6]. With this in mind, it is particularly important to understand this drug's impact on the genome, particularly for heavy users.

2.1.2 Risks associated with cannabis use

While health risks associated with cannabis use, in the general population, are low [7], there is growing awareness about the spectrum of behavioural and neurological dysfunctions associated with cannabis use [8, 9]. Currently, a small number of cannabis users suffer neurological and behavioural effects due to the use of cannabis [10], and the acute effects of cannabis on cognitive function are well documented, including impaired working memory [11], increased risking taking, and deficiencies in planning and decision-making [12]. Further, while the long-term effects of cannabis use are still controversial and less well defined, we already understand that cannabis use in adolescence is associated with a 1.5 to 2.5 times higher risk of developing mental health conditions [13, 14] such as psychotic disorders like schizophrenia [15].

2.1.3 How drugs affect the genome

Regardless if cannabis is legal or not, people will still consume it, thus, research needs to investigate its effects on those exposed to cannabis during development. Understanding the true effects of cannabis is imperative for the New Zealand population, in particular our most vulnerable groups (youth, Māori). Tobacco use, for example, is currently reducing in NZ, yet rates remain high within Māori and Pasifika groups [16]. Cannabis use is more commonly seen in males and amongst Māori [17] and thus could be a driver in disparities between ethnicities. DNA methylation (a type of 'epigenetic' modification) is a mechanism that cells use to control gene expression. It is a chemical modification to the DNA strand that can be altered by the environment, and can determine whether or not a gene is expressed [18], and this can directly influence health outcomes [19].

If there are observed associations between cannabis use, health, and genomic impacts, it is vital that we seek to fully understand the biological effects of cannabis on the human body. In order to begin to address this, in this Chapter, we explore data from the DNA of heavy cannabis users, and assess levels of DNA methylation, compared to both controls (who have never used cannabis) and individuals who use both cannabis and tobacco. Both are important comparisons which will allow us to directly quantify the effect of cannabis, in isolation, on the DNA of users.

2.1.4 The Christchurch Health and Development Study

The Christchurch Health and Development Study (CHDS), is a longitudinal study of a birth cohort of 1265 children, all born in the Christchurch region in 1977 [20]. The cohort has been intensively studied from birth to 40 years thus far, and data obtained during this time have addressed numerous issues relating to health, development, and social wellbeing [20]. Importantly, the CHDS assessed cannabis use via self report rating (at ages 12,16 and 18) using frequenting items ranging from 'never' to 'daily', meaning there is a particular emphasis on usage during mid-adolescence and adulthood. Further, participant retention rate has remained high; at age 35, 962 respondents were studied, representing 79% of the original 1977 cohort.

Through their work, the CHDS has shown that cannabis use in late adolescence and early adulthood is associated with a range of adverse outcomes in later life [4], such as increased rates of psychotic symptoms [21]. Just like other substances, high use of cannabis can lead to dependency, and it has been estimated that 8-9% of cannabis users will become addicted to the drug [22, 23]. However, in the CHDS, 12.5% of the cohort met the Diagnostic and Statistical Manual of Mental Disorder (DSM-IV) criteria for dependence on cannabis by the age of 25 [24], a rate which is 3.5% higher than the global population rate of dependency. Thus, showing the particular importance of carrying out this study from a New Zealand context, as what is seen globally may not reflect the reality in New Zealand. To the best of our knowledge, the CHDS is the only cohort that contains participants where DNA has been extracted who have been diagnosed as heavy cannabis users, but who have never used tobacco .This creates the opportunity to investigate the genome for DNA methylation changes that are specific to cannabis. Our hypothesis is that the chemical composition differences between cannabis and tobacco has very different biological impacts [25]. Therefore, given the potential health implications it is important that we rigorously test the effect of cannabis on the methylome, using the best available tools and pipelines that ensure accuracy of result.

2.1.5 DNA methylation arrays

The Illumina EPIC array (and their 450k array predecessors) are a hybridising array system, and have enabled DNA methylation studies at the genome-wide level. Consequently, the scientific literature has seen an exponential increase of studies quantifying differential DNA methylation via 450k and EPIC array. The benefit of these arrays is that they are highly reproducible and consistent at analysing many methylation sites across datasets, meaning that it is possible to combine and analyses multiple datasets together (meta-analyses). However, one of the major challenges with array technology is the bioinformatics pipelines that are available for analysis of array data. As the study of DNA methylation is a fast-growing field, a diverse range of pipelines have been developed to analyse DNA methylation data. However, having a range of analytical options requires decisions about which pipeline is best for a given set of data. Therefore, the aim of this chapter is look at the impact of different analysis

techniques, and quantify the impact that this can have on the integrity and results of methylome data analysis.

2.1.5 The importance of normalisation

Normalisation is the process of adjusting for effects detected in biological datasets which arise due to the variation of the technique itself, rather than the biological variance between samples [26]. Normalisation is particularly important for EPIC array data because each EPIC array allows methylome analysis of eight distinct samples. Without normalisation, data analysis can be confounded by 'batch effects', where different batches of arrays as well as batches of the eight samples can give different biological results. Further, as previously mentioned, it is becoming increasingly common to combine multiple datasets into meta-analyses, meaning accurate normalisation across datasets, to remove any batch effect, is crucial.

Currently there is not a standardised 'best practise' normalisation pipeline for assessing EPIC array data. There are a variety of packages available for the platform, with each controlling for bias that may arise between arrays, such as background fluorescence corrections and colour dye adjustments. For example, Illumina's genome studio, SWAN, Funnorm and Noob are all pre-processing methods which are available under the 'minfi' Bioconductor package [27] which supports 27k, 450k and EPIC array platforms. Selecting the right tool is undertaken manually, through trial and error, and must be tailored to the unique design of each study. This is because different pre-processing pipelines can result in differences in the identified biological variation, because each normalisation method transforms the data in slightly different ways. Therefore, it is important to pick the best fit for the data, not the best result.

Here we assessed the impact of different normalisation methods on the reduction of batch effects across EPIC arrays that were sampled over two consecutive years (Table 2.2). Data from 48 EPIC arrays were collected in two separate batches, the first in 2016, and the second in 2017. We then proposed the question, what is the best normalisation tool for our study design? Finally, after choosing the normalisation method that best fits our data, we quantify the specific impact of heavy cannabis use on DNA methylation in the human genome.

Tobacco is one of the most researched lifestyle factors to be associated with genome wide differential DNA methylation [28].This provided an internal reference control, for comparison with individuals who use both cannabis and tobacco. However, it is important to specifically isolate the difference between tobacco and cannabis smoking.

2.2 Methods

2.2.1 Cohort and study design

CHDS participants between the ages of 28 to 30 were approached to provide a peripheral blood sample for DNA analysis. A subset of the >800 participants who consented and provided a blood sample was used in the present study, comprising a total of 96 participants. Cases (regular cannabis users, N= 48) were matched with controls (n=48) for sex, ethnicity and family of origin socioeconomic status (Table 2.1). Case participants were partitioned into two subsets: one that contained cannabis-only users (who had never consumed tobacco, N= 24), and one that contained cannabis users who also consumed tobacco (N= 24). Cases were a group of long term regular (>weekly) cannabis users, selected on the basis that they either met DSM-IV [29] diagnostic criteria for cannabis dependence or had reported using cannabis on a daily basis for a minimum of three years prior to age 28. The median duration of regular use for selected cases was 9 years (range 3-14 years). Control participants had never used cannabis or tobacco. Mode of cannabis consumption was via smoking, for all participants. All aspects of the study were approved by the regional Health and Disability Ethics Committee.

Table 2.1 Christchurch Health and Development Study (CHDS) participants selected for EPIC arrays. Cases and controls were matched as closely as possible by the following: sex, ethnicity and parental socioeconomic status/occupation.

		Cases	Controls
Sex	Male	37	37
	Female	11	11
Ethnicity	European	35	45
	Other	13	3
Socioeconomic status	Professional/managerial	6	6
	Clerical/technical/skilled	21	21
	Semi-skilled/unskilled	21	21

2.2.2 EPIC array methods

DNA was extracted from whole blood using the KingFisher Flex System (Thermo Scientific, Waltham, MA USA), as per the published protocols. DNA was quantified via NanoDrop™ (Thermo Scientific, Waltham, MA USA) and standardised to 100ng/μl. Equimolar amounts were shipped to the Australian Genomics Research Facility (AGRF, Melbourne, VIC, Australia) for processing via the Infinium® Methylation EPIC BeadChip (Illumina, San Diego, CA USA).The 2016 samples were prepared by Dr Amy Osborne. The DNA samples were sent in two different batches as shown in Table 2.2. Half the samples (N= 48) were measured in 2016 followed by the second round in 2017.

Table 2.2 Time frame of sampling

Batch/year	Cannabis only users	Cannabis + Tobacco users	Controls
2016	24		24
2017		24	24

2.2.3 Data processing

Analysis was carried out using R statistical software (Version 3.5.2), quality control was firstly performed on the raw data. Sex chromosomes and a total of 150 failed probes (detection *P* value < 0.01 in at least 50% of samples) were excluded from analysis. Furthermore, potentially problematic CpGs with adjacent SNVs, or that did not map to a unique location in the genome [30] were also excluded, leaving 700,296 CpG sites for further analysis. The raw data were then normalised using four different pipelines.

2.2.4 Selecting a normalisation tool

The raw data were normalised with Illumina, SWAN, Funnorm and Noob pre-processing tools in the minfi package [27]. Our decision around the most appropriate tool for our dataset was based on the following steps: i) normalisation was checked by visual inspection of intensity densities and the first two components from beta density distribution plots and Multi-Dimensional Scaling (MDS) of the 5000 most variable CpG sites, and; ii) Quantile-Quantile (QQ) plots were used to assess the distribution of residuals, with lambda values generated to compare normalisation tools.

2.2.5 Statistical analysis post-processing

After selection of the best-performing normalisation method, the proportions of cell types (CD4+, CD8+ T cells, natural killer, B cells, monocytes and granulocytes) in each sample were estimated with the Flow.Sorted.Blood package [34]. Linear models were fitted to the methylated/unmethylated or M ratios using limma [35]. Separate models were fitted for cannabis-only vs. controls, and cannabis with tobacco users vs. controls. Both models contained covariates for sex (bivariate), socioeconomic status (three levels), batch (bivariate), population stratification (four principal components from 5000 most variable SNPs) and cell type (five continuous).

The data were analysed in two ways: i) cannabis-only users, compared to controls, and ii) tobacco + cannabis users, compared to controls. β values were calculated as the ratio of the methylated probe intensity (M) / the sum of the overall intensity of both the unmethylated probe (U) + methylated probe (M). *β* values were calculated, defined

as the ratio of the methylated probe intensity (M)/the sum of the overall intensity of both the unmethylated probe (U) + methylated probe (M). *P* values were adjusted for multiple testing with the Benjamini and Hochberg method and assessed for genomic inflation with bacon [36].

Differentially methylated CpG sites that were intergenic were matched to the nearest neighbouring genes in Hg19 using Granges default settings [37], and the official gene symbols of all significantly differentially methylated CpG sites (nominal $P < 0.001$) in cannabis-only users were tested for enrichment in KEGG 2019 human pathways with EnrichR [38]. and ggplot was used to construct Manhattan plots [39].

2.3 Results

2.3.1 Raw data

Illumina EPIC array raw data was plotted based on beta density distribution giving an overall illustration of the distribution of methylated counts and unmethylated counts. Figure 2.1 shows plots of beta value density for each array, arranged by year of analysis. Density plots of the beta distribution have two peaks, the first at around 0.0-0.1 which indicates the number of unmethylated CpG sites, and the second peak at about 0.6-1.0 which indicates the methylated sites. The difference between these peaks indicates discrepancies between the samples measured in the different years, the aim of the section if to correct for this.

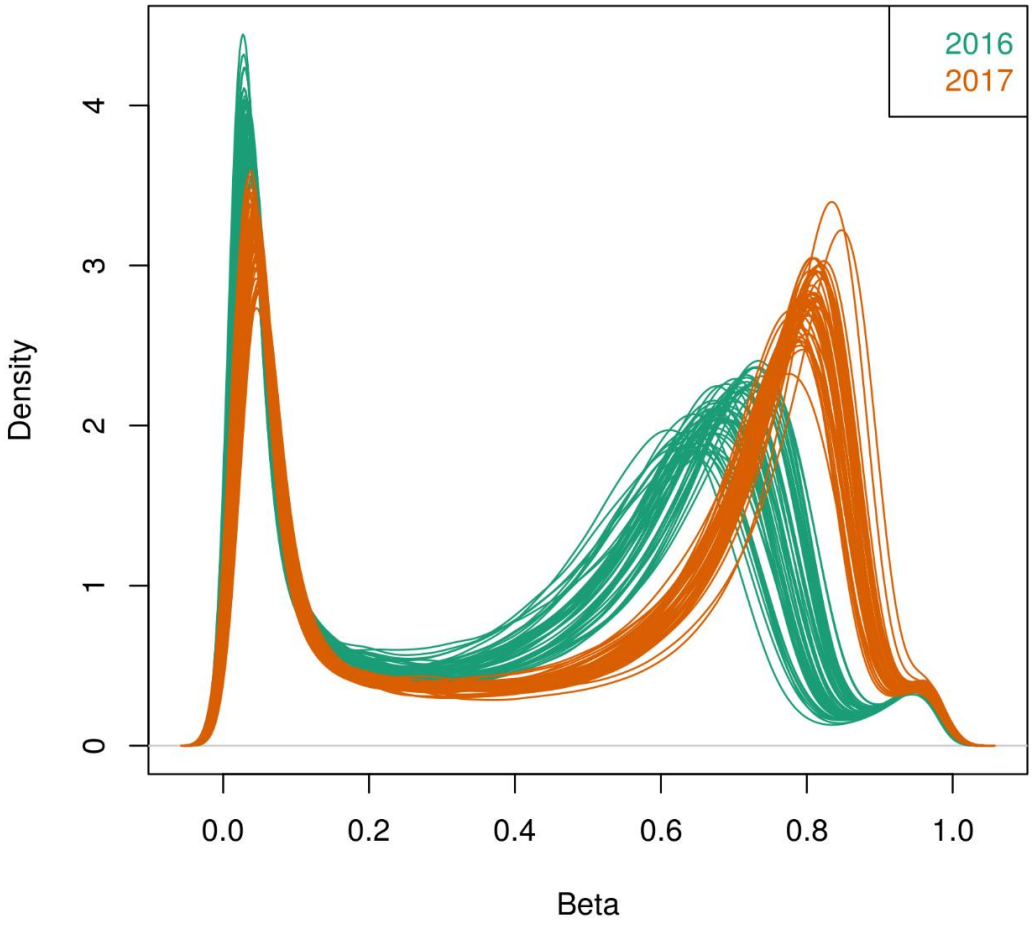


Figure 2.1 The raw density of the beta values across all samples analysed using the Illumina EPIC array system.The 2016 samples are indicated in green and the 2017 samples are indicated in orange.

2.3.2 Beta density profiles of raw data, compared to Illumina, SWAN and Noob normalisation methods

Four different normalisation tools were assessed for their fit to our data design (Table 2.3). The normalisation tool Funnorm showed no improvements of beta density distribution compared to the raw data, therefore was discontinued for all further analysis. The remaining three methods were compared to the raw EPIC data (Figure 2, A) and data processed with Illumina (Figure 2, B), SWAN (Figure 2, C), and Noob (Figure 2, D) normalisation methods were plotted as beta density plots, colour coded by analysis batch (year of EPIC array analysis).

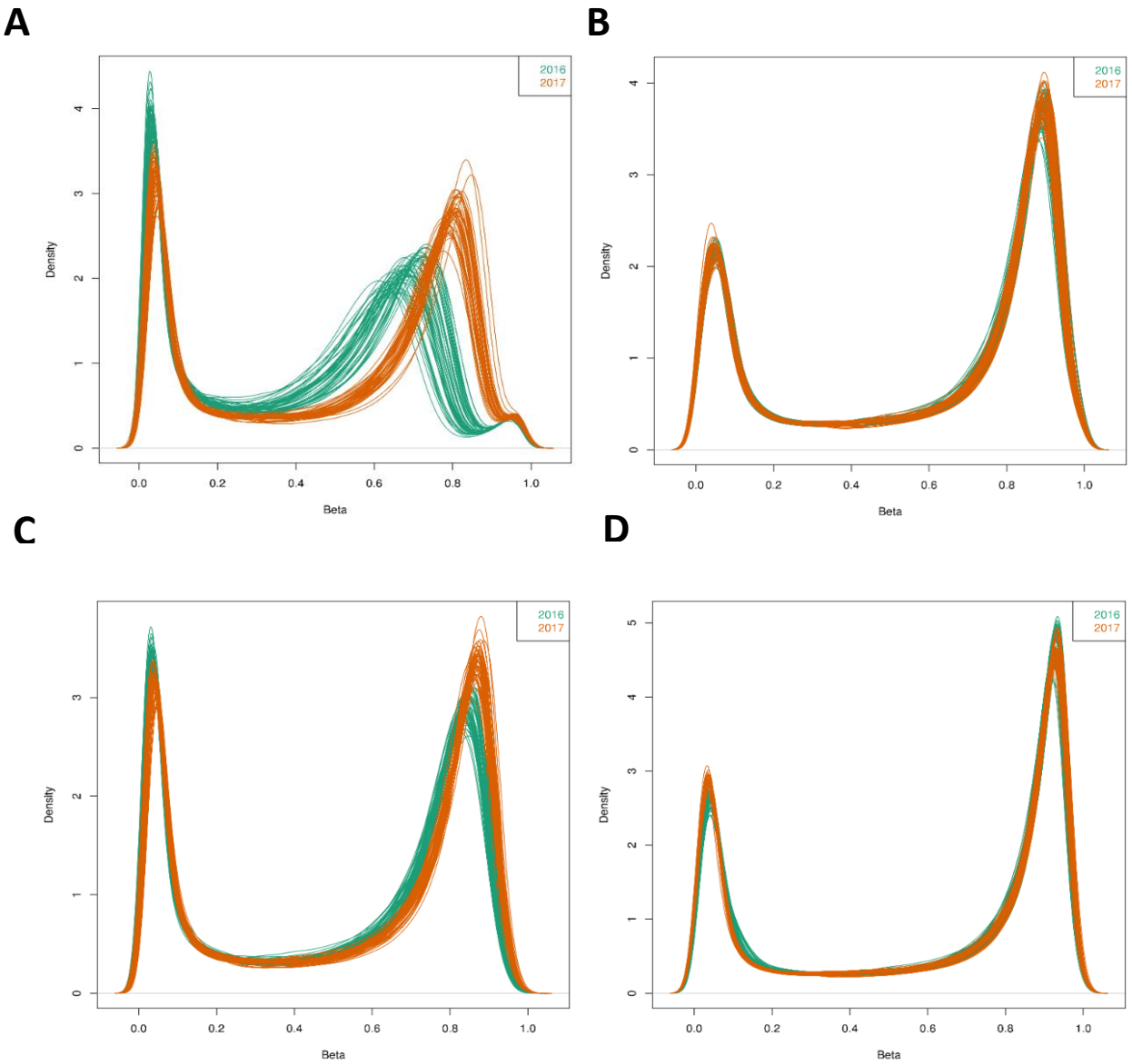


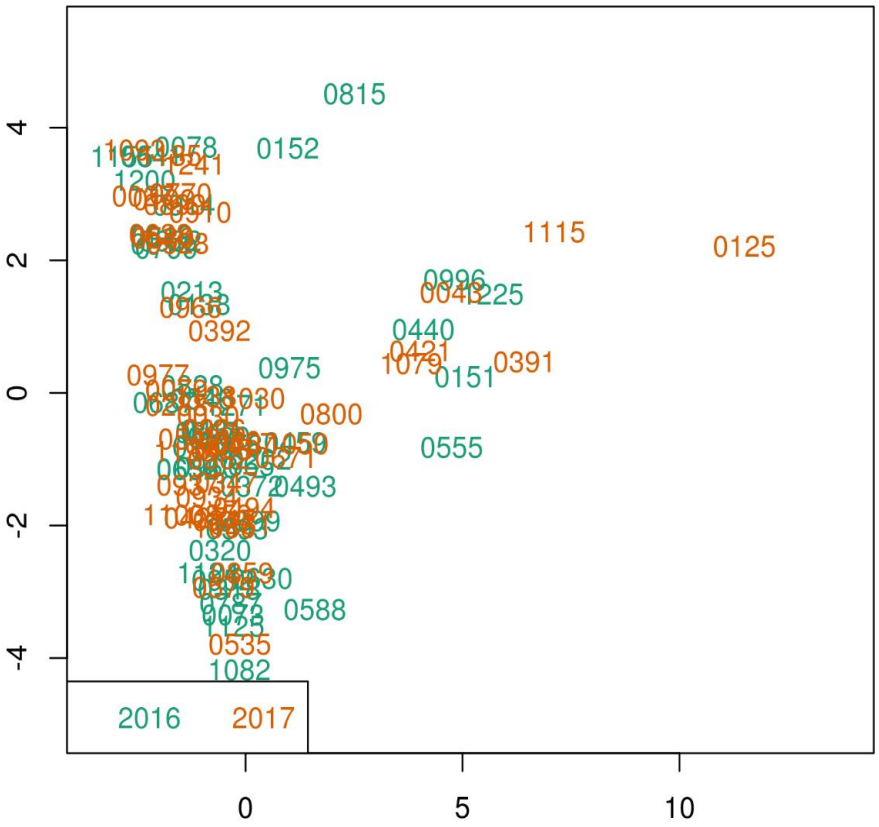
Figure 2.2 Density plots of the raw EPIC data compared after application of different normalisation tools. A) raw data, B) Illumina normalisation, C) SWAN normalisation, and D) Noob normalisation.

All beta density distribution plots generated by the normalisation methods showed an improved density distribution compared to the raw data, confirming that there was indeed a batch effect caused by the experiments being performed in two separate batches.

2.3.3 Multidimensional scaling plots using Illumina, SWAN and Noob normalisation methods

To further assess the best normalisation method for our data set, individual samples were displayed as a multidimensional scaling (MDS) plots for each of the normalisation method assessed. We can use this as a way of visually interpreting whether any individuals across batches reside closely to one another – this would indicate that the batch effect had not been normalised. Data from individual samples were each plotted, using 5000 of the most variable probes, with three normalisation tools: Illumina, SWAN and Noob (Figure 2.3). Illumina normalisation showed a random distribution of data points across the two years, indicating the batch effect was corrected (Figure 2.3A). The SWAN algorithm (Figure 2.3B), however, did not appear to effectively normalise the data, as data from each batch remained in discrete rather than overlapping clusters. Noob pre-processing of data (Figure 2.3C), showed similar results to Illumina normalisation. Here, no clustering based on array year was observed, indicating a correction of the batch effect.

A



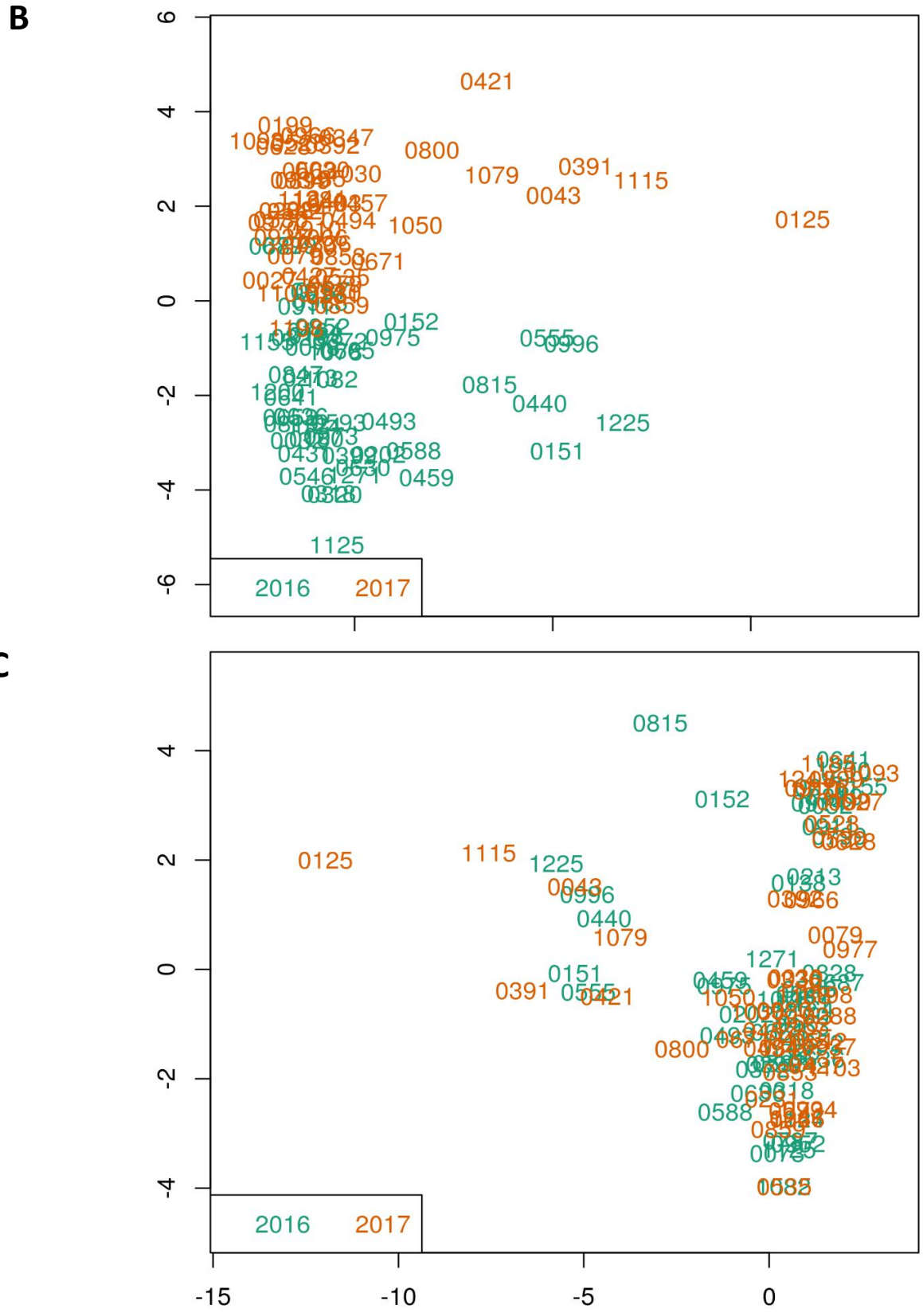


Figure 2.3 Multidimensional plots displaying the individuals of the study using the 5000 most variable positions post normalisation. A) Illumina B) SWAN and C) Noob. Individuals are grouped in colour by the year in which the samples were analysed – 2016 (green), 2017 (orange).

2.3.4 Genomic inflation - Quantile-Quantile plots for SWAN and Noob normalisation methods

Because post normalisation, statistical analysis were carried out to assess for differential DNA methylation between cannabis-only users versus controls (Figure 2.4 A and Figure 2.5 A), and cannabis with tobacco users versus controls (Figure 2.4 B and Figure 2.5 B), it was important to account for covariates that could lead to a bias in results. To determine the appropriate number of covariates to add to our model to prevent inflation of the test statistic, here we include data for ethnicity, sex, and social economic status, cell composition, and four principal components. To assess differences between residuals using SWAN and Noob, Quantile-Quantile plots were constructed, generating a lambda value which gives an indicator of the genomic inflation for both normalisation tools.

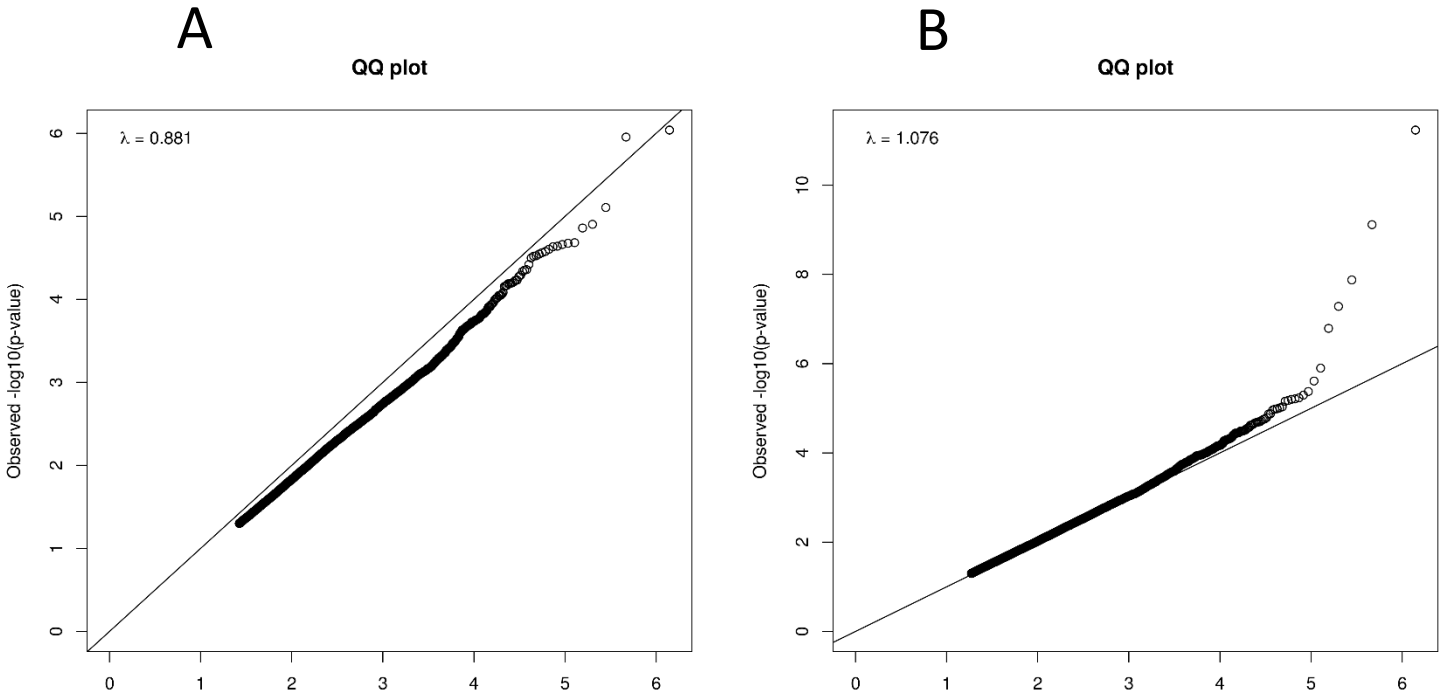


Figure 2.4 Post pre-processing using SWAN Quantile- quantile plots. Quantile plots were used to assess for overfitting of models. A) Cannabis only users vs controls B) Cannabis and tobacco users vs controls. Each dot displays the expected – log10 (p-values) under the model.

SWAN normalisation (Figure 2.4) shows the residuals plotted with a lambda estimate also displayed. The data generated using the model for cannabis-only users compared to controls resulted in λ= 0.881 (Figure 2.4A), and for cannabis with tobacco smokers compared to controls gave λ= 1.076 (Figure 2.4B). Residuals (CpG sites) are plotted, where the majority of the sites appear to follow the null hypothesis and show a normal distribution. Sites that appear outside of this normal distribution show significance in response to the variable of interest. In this instance (Figure 2.5A) all CpG sites analysed in response to cannabis only smoking compared to controls show a normal distribution.

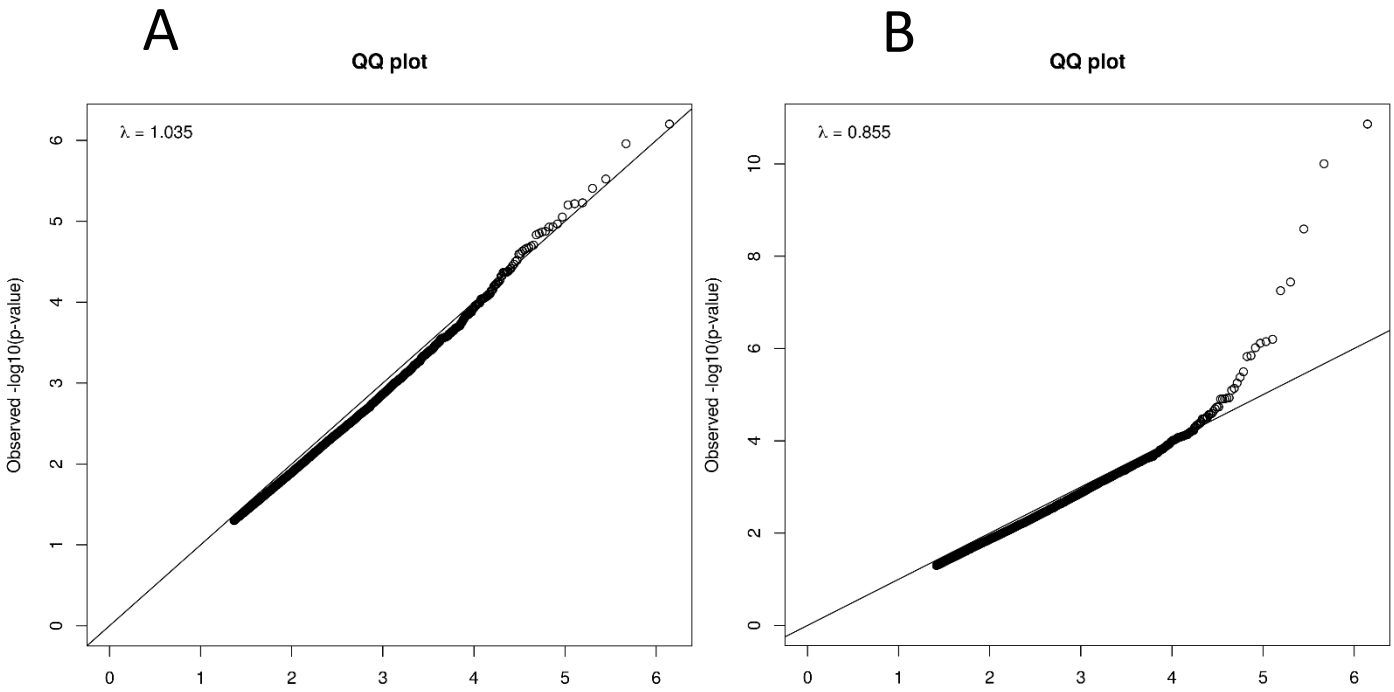


Figure 2.5 Post pre-processing using Noob Quantile- quantile plots. Quantile plots were used to assess for overfitting of models. A) Cannabis only users vs controls B) Cannabis and tobacco users vs controls. Each dot displays the expected –log10 (P values) under the model.

Similarly, with Noob normalisation, residuals using the model for cannabis-only users showed all residuals normally distributed (Figure 2.5A). Again, some residuals are seen to reach significance when assessing differences in cannabis and tobacco users versus controls (Figure 2.5B). Genomic inflation values have been improved to approach closer to 1. Cannabis-only users compared to controls model with λ= 1.035 and cannabis with tobacco users compared to controls generate λ= 0.855.

Under both SWAN and Noob almost all CpG sites follow a normal distribution indicating little variation between cannabis-only users and cannabis with tobacco users. Following the outcomes of the beta density plots (Figure 2.2), the multidimensional scaling plots (Figure 2.3), and Q-Q plots (Figure 2.4 and Figure 2.5) it was decided that Noob performed the best at normalising the batch effect. Therefore, the remainder of our analyses are performed on data normalised using Noob.

2.3.5 Differential DNA methylation in cannabis-only users, compared to controls.

Following selection of Noob as the sole processing method, further data analysis was carried out using the full data set. Table 2.4 displays the top 10 most highly differentially methylated CpG sites in cannabis-only users, compared to controls. Of the top CpG sites, none remain significant post multiple comparison adjustment. A total of six of the top 10 nominally significantly differentially methylated CpG reside within known genes, with *MYO1G* gene displaying two differentially methylated CpG sites. Most of the CpG sites that were found to be nominally significant reside within the gene body, as opposed to e.g. promoter regions or 5' untranslated regions. Four of the top 10 CpG sites were found to reside on chromosome 19. The beta values of the differences between cannabis-only users and controls vary amongst each of the CpG sites, and range from 1.1% differential DNA methylation to 9%. The greatest magnitude of change in differential DNA methylation is not associated with greatest nominal P value.

A genome-wide plot of the CpG sites measured using the Illumina EPIC array in cannabis-only users compared to controls is displayed in Figure 2.6. Labelled CpG sites have a –log10 P value of greater than 4.5. At multiple sites, CpG sites are close to adjusted P value significance.

Table 2.3 Top 10 CpG sites differentially methylated in response to cannabis-only users compared to controls. Beta values with P values, nominal and adjusted by the Benjamini and Hochberg method. Locations are relative to hg19 with gene names for overlapping genes or nearest 5' gene with distance to the 5' end shown. Missing UCSC locations are from new probes on the EPIC array, which have not yet been included in the UCSC annotation tracks..

Illumina ID	Gene	Chr	Location	Position in genome	Cannabis	Control	β difference	*Log FC*	*P value*	*Adjusted P value*
cg02234936		19	42420037		0.143	0.132	0.011	0.500	7.48E-07	0.269
cg12803068	*MYO1G*	7	45002919	Body	0.804	0.708	0.095	1.150	7.69E-07	0.269
cg01695406	*TMEM190*	19	55889276	Body	0.818	0.769	0.048	0.637	3.30E-06	0.700
cg24875484	*DPCR1*	6	30910583	Body	0.101	0.091	0.009	0.253	4.41E-06	0.700
cg05009104	*MYO1G*	7	45002980	Body	0.791	0.741	0.050	0.600	6.96E-06	0.700
cg00470351	*CDC20*	1	43825296	Exon	0.401	0.377	0.023	0.212	7.25E-06	0.700
cg24060040		19	5802267		0.108	0.078	0.029	0.798	7.45E-06	0.700
cg12322720		15	60447342		0.579	0.523	0.056	0.430	9.87E-06	0.700
cg06693983	*TMEM190*	19	55889216	Body	0.836	0.757	0.078	1.102	1.13E-05	0.700
cg06955687		11	125803030		0.739	0.702	0.036	0.366	1.21E-05	0.700

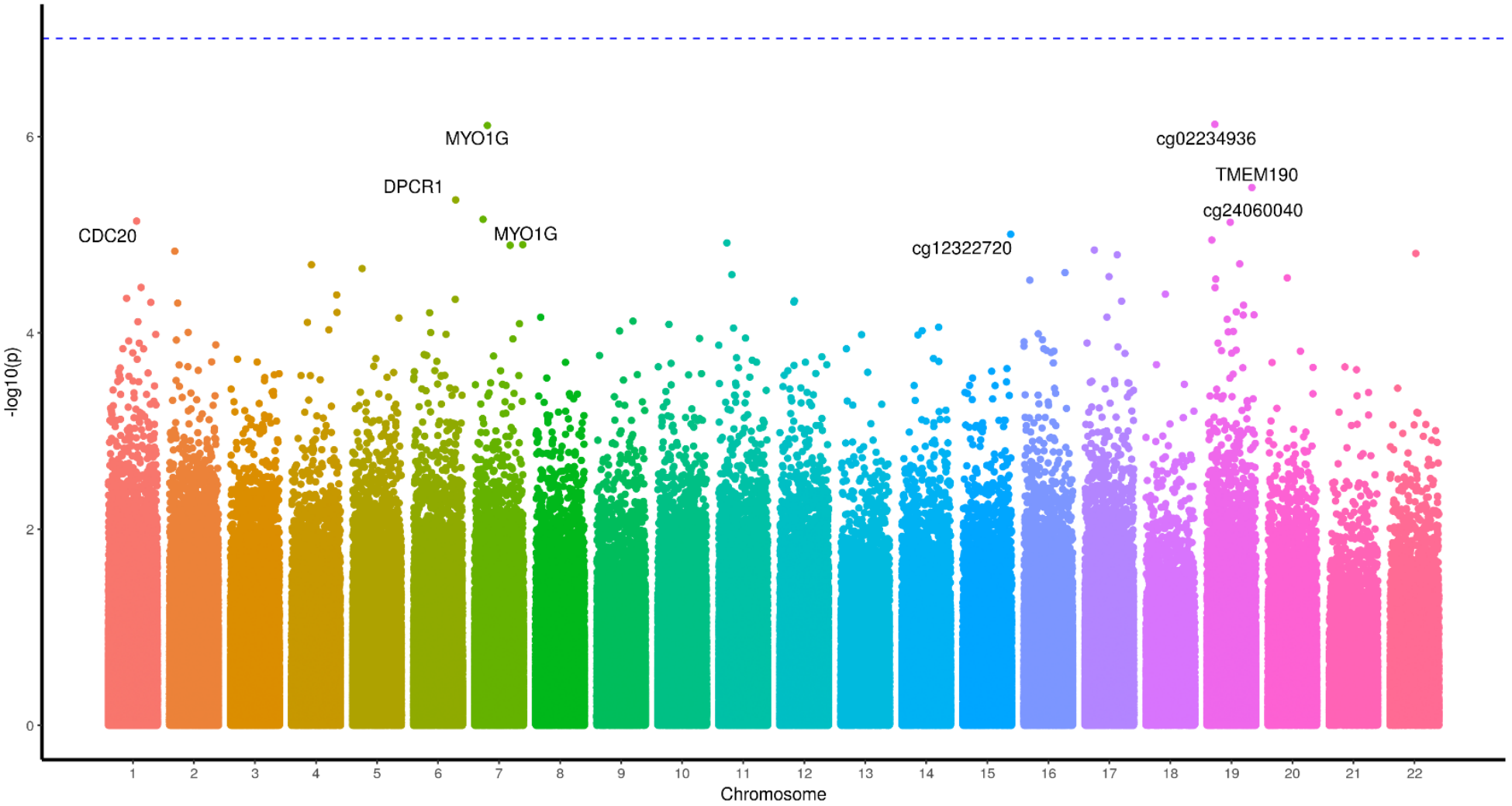


Figure 2.6 Manhattan plot of the genome-wide differential DNA methylation changes in response to cannabis only users compared to non-smoking controls. Each chromosome is listed along the X-axis, displaying the genome-wide differential DNA methylation changes found at each given CpG site. The dotted line represents the genome wide significance level, any adjusted P value significance observed at CpG sites would appear above this

2.3.6 Differential DNA methylation in response to cannabis with tobacco users

Cannabis with tobacco users were then compared to controls to assess for differential DNA methylation. A total of six CpG sites were found to be significant following Benjamini and Hochberg method (i.e. at the genome-wide level). Table 2.5 displays the top 10 most differentially methylated CpG sites ranked in order of P Value significance. Of the six CpG sites that were significantly differentially methylated at the genome-wide level, four were located in known genes *AHHR*, *RARA*, *F2RL3* and *PRSS23*. Of the top 10 CpG sites, three (*AHHR*, cg07219494 and cg12828729) reside on chromosome five.

Figure 2.7 displays the Manhattan plot of the genome-wide CpG sites differentially methylated between cannabis with tobacco users compared to controls. Note that the scaling is different to that used for Figure 2.7; in cannabis-only users the –log10(p) scale is scaled by two-fold change, compared to a three-fold change in cannabis with tobacco users. A total of five CpGs - *AHHR*, *RARA*, *F2RL3*, cg21566642 and cg01940273 - have –log10(p) values of greater than seven.

Table 2.5 Top differentially methylated CpG sites in cannabis and tobacco users compared to controls. Beta values with P values, nominal and adjusted by the Benjamini and Hochberg method. Locations are relative to hg19 with gene names for overlapping genes or nearest 5' gene with distance to the 5' end shown. Missing UCSC locations are from new probes on the EPIC array, which have not yet been included in the UCSC annotation tracks..

Illumina ID	Gene	Chr	Location	Position in genome	Cannabis + tobacco	Control	β difference	*Log FC*	*P value*	*Adjusted P value*
cg05575921	*AHRR*	5	373378	Body	0.661	0.895	-0.233	-2.071	5.33E-12	3.74E-06
cg21566642		2	233284661		0.445	0.619	-0.174	-0.990	7.24E-11	2.53E-05
cg01940273		2	233284934		0.533	0.628	-0.094	-0.557	9.29E-09	0.001
cg03636183	*F2RL3*	19	17000585	Body	0.590	0.682	-0.091	-0.527	1.04E-08	0.001
cg17739917	*RARA*	17	38477572	5'UTR	0.370	0.471	-0.100	-0.645	1.39E-08	0.001
cg14391737	*PRSS23*	11	86513429	5'UTR	0.362	0.421	-0.059	-0.467	3.71E-07	0.043
cg01541424		12	127874654		0.167	0.132	0.0349	0.605	1.33E-06	0.132
cg07219494		5	166408484		0.700	0.747	-0.047	-0.650	1.54E-06	0.134
cg12828729		5	134823969		0.561	0.504	0.057	0.372	2.06E-06	0.160
cg15651928	*PXMP4*	20	32290811	3'UTR	0.798	0.770	0.028	0.313	4.14E-06	0.290

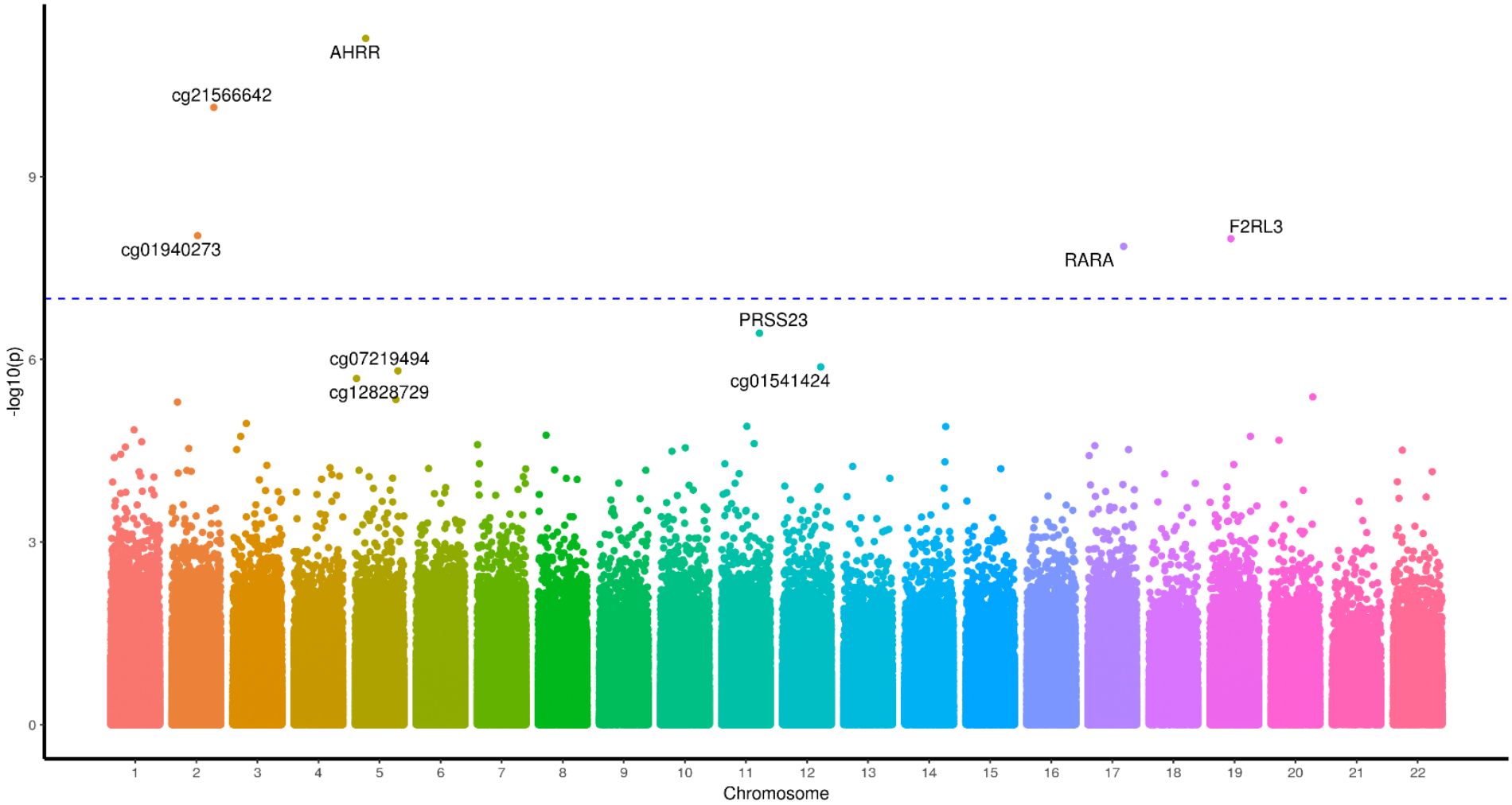


Figure 2.7 Manhattan plot of the genome-wide differential DNA methylation changes in response to cannabis and tobacco smoking users compared to controls. Each chromosome is listed along the X-axis, displaying the genome-wide differential DNA methylation changes found at each given CpG site. The dotted line represents the genome wide significance level, any adjusted P value significance observed at CpG sites would appear above this.

2.3.7 Functional gene annotation clustering (KEGG pathway analysis)

Functional gene annotation clustering was performed using Enrichr to annotate which KEGG pathways were most represented in the list of nominally significant differentially methylated CpG sites in the cannabis-only data. Specifically, the genes (or nearest genes) represented by the top 1000 nominally significant CpG sites were subjected to KEGG pathway analysis. All pathways that were found to have a significant adjusted P Value are included in the below tables.

Table 2.6 Pathway analysis from the top CpG sites and their associated genes in cannabis-only users compared to controls.

Pathway	*P value*	*Adjusted P value*	Odds Ratio	Combined Score
Cholinergic synapse	0.00004	0.013	3.15	31.61
Glutamatergic synapse	0.0001	0.020	2.90	24.81
Insulin secretion	0.0004	0.021	3.08	23.51
Long-term potentiation	0.0008	0.028	3.29	23.40
Circadian entrainment	0.0004	0.026	2.96	22.96
Aldosterone synthesis and secretion	0.0004	0.024	2.93	22.43
cAMP signalling pathway	0.00009	0.015	2.39	22.09
Dopaminergic synapse	0.0002	0.022	2.70	21.97
Arrhythmogenic right ventricular cardiomyopathy (ARVC)	0.001	0.034	3.07	20.04

Nine pathways were significantly enriched in the differential DNA methylation dataset. The top pathway was determined as cholinergic synapse (adjusted P= 0.01366), followed by glutamatergic synapse (adjusted P= 0.02005). Both of these pathways are involved in neurotransmission.

Table 2.7 Pathway analysis of the top CpG sites and their associated gene in response to cannabis and tobacco use.

Pathway	*P value*	*Adjusted P value*	Odds Ratio	Combined score
Gastric cancer	2.75E-06	0.000424	3.17	40.64
Pathways in cancer	3.07E-08	9.46E-06	2.25	38.96
Cushing syndrome	1.81E-05	0.001	2.91	31.74
Parathyroid hormone synthesis, secretion and action	6.79E-05	0.005	3.19	30.59
Basal cell carcinoma	0.0004	0.014	3.58	27.74
Cholinergic synapse	0.0001	0.006	3.02	27.03
Phospholipase D signalling pathway	0.0001	0.006	2.74	25.18
Signalling pathways regulating pluripotency of stem cells	0.0001	0.006	2.75	24.31
Renal cell carcinoma	0.0008	0.022	3.26	22.92
Breast cancer	0.0002	0.011	2.60	21.22
Melanoma	0.001	0.025	3.13	20.91
Cortisol synthesis and secretion	0.002	0.034	3.12	19.07
Circadian entrainment	0.001	0.026	2.79	18.72
Cellular senescence	0.0007	0.023	2.39	17.14
Hippo signalling pathway	0.0007	0.021	2.39	17.14
Fc gamma R-mediated phagocytosis	0.002	0.033	2.72	16.54
Hepatocellular carcinoma	0.001	0.025	2.28	15.08
Wnt signalling pathway	0.001	0.031	2.28	14.40
Proteoglycans in cancer	0.001	0.029	2.13	13.69

The gene or nearest gene represented by the top 1000 CpG sites identified in the cannabis with tobacco dataset were also investigated, to determine which KEGG pathways were significantly enriched in these data. A total of 19 pathways displayed significant enrichment after adjustment for multiple testing. Of the 19 pathways, seven are involved in cancer (gastric cancer and the more general pathways in cancer, adjusted P =0.000424 and 9.4 x 10^{-6} respectively). The top pathway in response to cannabis-only users, cholinergic synapse, is found to also be significant in cannabis with tobacco users.

2.4 Discussion

2.4.1 The Illumina EPIC array

High-throughput array technology has facilitated the next step in assessing associations between DNA methylation and response to a known phenotype at a genome wide level. The Illumina Infinium EPIC array (as well as the 27k and 450k) is one such platform that allows for the isolation of these DNA methylation changes. Selecting a pre-processing method is pivotal for the integrity of the data that is produced. The four pre-processing methods assessed in this chapter all performed differently on our data set. The raw density data (Figure 2.1) indicated that there were discrepancies between the two batches of samples which were measured in different years. Before any further analysis could begin these batch differences needed to be adjusted. Not addressing this issue could lead to bias and also misleading results, whereby the differential DNA methylation found is actually due to human/machine variation and not actually due to the variability seen from to the phenotype.

Variation can arise in data through numerous ways. For instance, only eight individual DNA samples can fit onto a slide to be measured. Each slide can be different, and each batch of slides can be different again. Variation also arises through operational processes and the use of different equipment. These can all result in subtle variations which can equate to a point of difference between samples which researchers cannot be aware of until quality checking of data is performed. The task then becomes to account for these sources of variation and take addition steps in bioinformatics pipelines to counteract these. The problem then arises, what is batch effect and what is biological variation?

The second problem with not having a uniform pipeline of analysis is the issue with validity and cross-comparison of other EPIC array experiments. Meta-analyses are a useful way of generating greater power to strengthen smaller analyses by combining datasets together. It is widely acknowledged that using the same technology is essential for meta-analyses, but further issues arise when different pre-processing methods have been applied to the different datasets. Thus, to ensure results are not biased by non-biological variability, all datasets should be processed in the same

manner. However, there is not yet a consensus on processing. Importantly, DNA methylation analyses, particularly via array, is a burgeoning field, and the more it grows, the more crucial it is that we have the methodology in place to be able to accurately combine data to increase our statistical power and determine the biological relevance of our results – often the most significant results come from those which combine multiple studies. Further, a consensus normalisation pipeline will future-proof research and yield cost savings - once array data has been generated for an individual DNA sample, the data can be applied to other hypotheses, enabling the investigation of epigenome-wide association analyses (EWAS). In most lab groups, sample size is the most common limiting factor for statistical power when detecting differential methylation in response to a stimulus. Thus, combining studies is the best way to combat this problem, however, batch effects need to be accounted for.

2.4.2 Comparison of four different normalisation methods

Overall, assessment of the "best normalisation tool" was decided empirically based on the many ways raw data can be assessed visually. Beta density distribution, multidimensional scaling plots and Q-Q plots all provided important visual evidence for determining which was the best. Furthermore, highlighting the need for effective data visualisation, rather than simply using tabulated numerical data.

All pre-processing tools were plotted to assess their adjustment and beta density distribution (Figure 2.2). Displaying this visually was crucial for understanding the true effects of the pre-processing normalisation methods. All three tools which could display beta density distributions showed a degree of correction for the batch effect compared to the raw unprocessed data. Funnorm showed no improvement of beta density distribution compared to the raw unprocessed data, therefore was discontinued. SWAN showed some improvements compared to the raw data however, discrepancies could still be seen. Illumina normalisation method and Noob both resulted in density plots which indicated that they had successfully corrected for the batch effect between the years that the samples were measured.

Further assessment of Illumina, SWAN and Noob was carried using the 5000 most variable CpG sites for each of the individuals in the study. These were plotted as

multidimensional scaling plots (Figure 2.3). With Illumina and Noob methods, a random distribution of individuals is seen (Figure 2.3 A and Figure 2.3 C), again indicating that the batch effect had been successfully corrected. However, the same cannot be said using SWAN where individuals cluster based upon the year of sampling (Figure 2.3 B).

Lambda values, as generated via Q-Q plot, are a quantitative measure of genome-wide distribution of the test statistic with the expected genomic inflation. A Lambda value of 1 would indicate that no inflation is present. In our analyses, the observed SWAN and Noob lambda values only showed marginal differences between both of our models. Specifically, using SWAN, the cannabis-only model genomic inflation was λ =0.881, and cannabis with tobacco users was λ =1.076. Using Noob, genomic inflation of our cannabis-only model was λ =1.035 and cannabis with tobacco was λ =0.855. In both of these instances, the values of both models appear to be either side of 1, by roughly a similar amount. A potential reason as to why results appear to be very similar here is that year of sampling was also included within the model for both Q-Q plot analyses. As this was included results were adjusted accordingly and therefore residual results appear to be very similar. Normalisation via the Illumina tool performed well in comparison to the other methods. However, its use was discontinued on the grounds of being outdated and as new innovations in the normalisation field have provided more robust tools [27]. Therefore, our results demonstrate that without visual interpretation of pre-processing batch normalisation, the underlying inaccuracies that were displayed by SWAN would not have been detected, and year of sampling could not therefore have been discounted as biasing our results.

Finally, while the end residuals results appear similar from both SWAN and Noob output, the discrepancies between batches seen using SWAN cannot be ignored, therefore the pre-processing method Noob was seen as the best fit for our study design.

2.4.3 Differential DNA methylation between cannabis only users and controls

Having successfully normalised the data, differential DNA methylation between cannabis-only users and controls was calculated. While we detected a large amount of differential DNA methylation between cannabis-only users and controls no individual CpG sites were found to reach adjusted P value significance (Table 2.4 and Figure 2.6). Within the top 10 most nominally significant CpG sites there are two CpG sites that reside within the same gene, *MYO1G.* The gene plays a role within the immune system as it is expressed specifically by haematopoietic tissue and cells [40]. Knockdowns of the gene show a decrease in cell elasticity [40].

Online tools such as EnrichR and KEGG (Kyoto Encyclopaedia of Genes and Genomes) provide further levels of understanding of the interaction of different genes in a pathway. DNA methylation sites within genes can then be compared and viewed for more functional roles. In Table 2.6, there are nine pathways that were found to contain genes with internal differentially methylated CpG sites. Interestingly, these pathways were related primarily to brain and cardiac function. Cholinergic synapse (adjusted P = 0.01366), glutamatergic synapse (adjusted P = 0.02005), long-term potentiation (adjusted P = 0.02816), dopaminergic synapse (adjusted P = 0.02230), and arrhythmogenic right ventricular cardiomyopathy (ARVC) (adjusted P = 0.03440). Both brain and cardiac alterations are consistent with the literature on the phenotypic impacts of cannabis use [41-44], supporting the biological relevance of our findings.

2.4.4 Differential DNA methylation between cannabis with tobacco users

When the data was partitioned to assess DNA methylation between cannabis with tobacco users, six CpG sites passed the Benjamini and Hochberg adjustment method. The top CpG site, *AHRR,* is the most well-known differentially methylated site resulting from tobacco exposure [45-48]. Validating this site with our cohort reiterates both the importance of that one CpG site but also the validity of our data and the methodology we applied to our analysis. These finding also gives us confidence in our cannabis-only data (for which there is no literature to compare our findings to). Thus in this instance, our detection of *AHRR* serves as a positive control.

KEGG pathway analysis for cannabis with tobacco users (Table 2.7) clearly indicates that KEGG pathways associated with cancer are a more dominant theme, rather than the brain or cardiac function which is seen in cannabis-only users. A total of 19 pathways had adjusted P value significance. Again, our data indicate biological relevance, as we know that tobacco smoking increases the risk of at least 17 classes of human cancers [49, 50], and induces DNA damage that can lead to an increase of somatic mutations and elevates the chance of acquiring driver mutations in cancer related genes [51].

2.4.5 Limitations

As previously discussed, our cannabis-only results are limited to nominal genome-wide significance which is to be partially expected, as our sample size (dictated by financial constraints) is a limiting factor. Expanding the number of cannabis-only users would aid in confirming truly positive sites of differential methylation. Also, if our study design was conducted in a way where not all the cannabis only individuals were sampled in 2016 and the cannabis with tobacco individuals sampled in 2017 we would have maybe been able to differentiate between biological variance and batch effect better.

Variance between individuals within the study could ultimately lead to bias in results, therefore it is very important that this is taken into account where possible. The statistical models which we used to compare cases to controls (cannabis-only and cannabis with tobacco) do take into account many forms of variance, as displayed by the residual plots in Figure 2.5. However, while this is necessary, it can also be a limiting factor - accounting for “too many” variables in a model can also mask true biological variance, due to the creation of an overly-stringent of the model. There is a fine line of inclusion/exclusion of co-variables, particularly in small studies. In our case, it is likely that we have over-compensated with covariates, if we were to remove some of these from our model we would expect to see some CpG sites reach the genome wide significance level. However, it is important to have a robust and replicable analysis to maintain the integrity of data, even if it does come at the expense of only nominal significant results. It is particularly important for genetic studies to have

individual variance within the population, and this must be accounted for wherever possible. We are fortunate that the CHDS records a tremendous amount of data which spans from birth to the present time, and key variables, such as, socioeconomic status is available to us to include in our analyses.

Thus, while our cannabis-only data is nominal, the apparent biological relevance of the findings demonstrate that these nominal results, in general, should be seen as interesting observations that require further follow up. The analysis illustrates the potential for DNA methylation to play a role in the human response to cannabis. The differences seen between the cannabis-only data, and the cannabis with tobacco data, highlights the unique mode of action of cannabis compared to tobacco, and stresses the importance of researching the biological effects of cannabis in isolation. By extension, however, our data also highlight the value of performing the same analysis on individuals who use both cannabis and tobacco; the large majority of cannabis users are also tobacco users, therefore joint repercussions on the genome may play a role in the development of a range of diseases.

2.5 Chapter summary

- Four normalisation tools were tested, and Noob was judged most effective at adjusting variance between batches of samples processed in different years.
- Differential DNA methylation was assessed between cannabis-only users and controls, as well as cannabis with tobacco users, versus controls.
- Nominal significance was found between cannabis-only users across CpG sites in the human genome, while six CpG were found to be significant post adjustment in the cannabis with tobacco users, compared to controls.
- Pathway analysis was carried out on the genes (or nearest genes) that housed the top 1000 differentially methylated CpG sites.
- Pathways differed between cannabis-only users, where the most significantly enriched KEGG pathways were involved in brain and cardiac functional. This is in contrast to the cannabis with tobacco users, where the most significantly enriched KEGG pathways were involved in cancer.
- Despite the limitation of small sample size the nominal results provide biologically relevant observations that should be expanded on.

2.6 References

1. Boden, J.M., D.M. Fergusson, and L.J. Horwood, *Illicit drug use and dependence in a New Zealand birth cohort.* Aust N Z J Psychiatry, 2006. **40**(2): p. 156-63.
2. Sevigny, E.L., R.L. Pacula, and P. Heaton, *The effects of medical marijuana laws on potency.* The International journal on drug policy, 2014. **25**(2): p. 308-319.
3. Organization, W.H., *The Health and Social Effects of Nonmedical Cannabis Use*. 2016.
4. Fergusson, D.M. and J.M. Boden, *Cannabis use and later life outcomes.* Addiction, 2008. **103**(6): p. 969-976.
5. Bridgeman, M.B. and D.T. Abazia, *Medicinal Cannabis: History, Pharmacology, And Implications for the Acute Care Setting.* P & T : a peer-reviewed journal for formulary management, 2017. **42**(3): p. 180-188.
6. Pellati, F., et al., *Cannabis sativa L. and Nonpsychoactive Cannabinoids: Their Chemistry and Role against Oxidative Stress, Inflammation, and Cancer.* BioMed research international, 2018. **2018**: p. 1691428-1691428.
7. Poulton, R., et al., *Patterns of recreational cannabis use in Aotearoa-New Zealand and their consequences: evidence to inform voters in the 2020 referendum.* Journal of the Royal Society of New Zealand, 2020. **50**(2): p. 348-365.
8. Crean, R.D., N.A. Crane, and B.J. Mason, *An evidence based review of acute and long-term effects of cannabis use on executive cognitive functions.* J Addict Med, 2011. **5**(1): p. 1-8.
9. Alegria, A.A., et al., *Comorbidity of generalized anxiety disorder and substance use disorders: results from the National Epidemiologic Survey on Alcohol and Related Conditions.* J Clin Psychiatry, 2010. **71**(9): p. 1187-95; quiz 1252-3.
10. Caspi, A., et al., *Moderation of the effect of adolescent-onset cannabis use on adult psychosis by a functional polymorphism in the catechol-O-methyltransferase gene: longitudinal evidence of a gene X environment interaction.* Biol Psychiatry, 2005. **57**(10): p. 1117-27.
11. Hart, C.L., et al., *Effects of acute smoked marijuana on complex cognitive performance.* Neuropsychopharmacology, 2001. **25**(5): p. 757-65.
12. Ramaekers, J.G., et al., *High-potency marijuana impairs executive function and inhibitory motor control.* Neuropsychopharmacology, 2006. **31**(10): p. 2296-303.
13. Hall, W. and L. Degenhardt, *Cannabis use and the risk of developing a psychotic disorder.* World Psychiatry, 2008. **7**(2): p. 68-71.
14. Barbee, J.G., et al., *Alcohol and substance abuse among schizophrenic patients presenting to an emergency psychiatric service.* J Nerv Ment Dis, 1989. **177**(7): p. 400-7.
15. Fergusson D, B.J., *Cannabis Use in Adolescence, Improving the Transition: Reducing Social and Psychological Morbidity During Adolescence.*, O.o.t.P.M.s.S.A. Committee, Editor. 2011. p. 257-271.
16. Barnett, R., J. Pearce, and G. Moon, *Community inequality and smoking cessation in New Zealand, 1981–2006.* Social Science & Medicine, 2009. **68**(5): p. 876-884.
17. Fergusson, D. and J. Boden, *Cannabis use in adolescence.* Improving the Transition, 2011. **257**: p. 239-253.
18. Weber, M., et al., *Distribution, silencing potential and evolutionary impact of promoter DNA methylation in the human genome.* Nature Genetics, 2007. **39**(4): p. 457-466.
19. Jirtle, R.L. and M.K. Skinner, *Environmental epigenomics and disease susceptibility.* Nature Reviews Genetics, 2007. **8**(4): p. 253-262.
20. Fergusson, D.M. and J.L. Horwood, *The Christchurch Health and Development Study: Review of Findings on Child and Adolescent Mental Health.* Australian & New Zealand Journal of Psychiatry, 2001. **35**(3): p. 287-296.
21. Fergusson, D.M., L.J. Horwood, and N.R. Swain-Campbell, *Cannabis dependence and psychotic symptoms in young people.* Psychological Medicine, 2003. **33**(1): p. 15-21.
22. Lopez-Quintero, C., et al., *Probability and predictors of transition from first use to dependence on nicotine, alcohol, cannabis, and cocaine: results of the National Epidemiologic Survey on Alcohol and Related Conditions (NESARC).* Drug and alcohol dependence, 2011. **115**(1-2): p. 120-130.
23. Flórez-Salamanca, L., et al., *Probability and predictors of transition from abuse to dependence on alcohol, cannabis, and cocaine: results from the National Epidemiologic Survey on Alcohol and Related Conditions.* The American journal of drug and alcohol abuse, 2013. **39**(3): p. 168-179.
24. Joseph M. Boden, D.M.F., L, John Horwood, *Illicit Drug use and Dependence in a New ZEaland Birth Cohort.* Australian & New Zealand JOurnal of Psychiatry, 2006. **40**(2): p. 156-163.

25. Melamede, R., *Cannabis and tobacco smoke are not equally carcinogenic.* Harm reduction journal, 2005. **2**: p. 21-21.
26. Smyth, G.K. and T. Speed, *Normalization of cDNA microarray data.* Methods, 2003. **31**(4): p. 265-273.
27. Aryee, M.J., et al., *Minfi: a flexible and comprehensive Bioconductor package for the analysis of Infinium DNA methylation microarrays.* Bioinformatics (Oxford, England), 2014. **30**(10): p. 1363-1369.
28. Alegría-Torres, J.A., A. Baccarelli, and V. Bollati, *Epigenetics and lifestyle.* Epigenomics, 2011. **3**(3): p. 267-277.
29. Association, A.P., *Diagnostic and Statistical Manual of Mental Disorders Fourth Edition*. Fourth ed. 1994, Washington, DC: American Psychiatric Association.
30. Pidsley, R., et al., *Critical evaluation of the Illumina MethylationEPIC BeadChip microarray for whole-genome DNA methylation profiling.* Genome Biology, 2016. **17**(1): p. 208.
31. Maksimovic, J., L. Gordon, and A. Oshlack, *SWAN: Subset-quantile within array normalization for illumina infinium HumanMethylation450 BeadChips.* Genome biology, 2012. **13**(6): p. R44.
32. Triche, T.J., Jr, et al., *Low-level processing of Illumina Infinium DNA Methylation BeadArrays.* Nucleic Acids Research, 2013. **41**(7): p. e90-e90.
33. Fortin, J.-P., et al., *Functional normalization of 450k methylation array data improves replication in large cancer studies.* Genome Biology, 2014. **15**(11): p. 503.
34. Jaffe, A.E., *FlowSorted.Blood.450k: Illumina HumanMethylation data on sorted blood cell populations.* 2019.
35. Ritchie, M.E., et al., *limma powers differential expression analyses for RNA-sequencing and microarray studies.* Nucleic acids research, 2015. **43**(7): p. e47-e47.
36. van Iterson, M., et al., *Controlling bias and inflation in epigenome-and transcriptome-wide association studies using the empirical null distribution.* Genome biology, 2017. **18**(1): p. 19.
37. Lawrence, M., et al., *Software for computing and annotating genomic ranges.* PLoS computational biology, 2013. **9**(8).
38. Kuleshov, M.V., et al., *Enrichr: a comprehensive gene set enrichment analysis web server 2016 update.* Nucleic acids research, 2016. **44**(W1): p. W90-W97.
39. Wickham, H., *ggplot2: Elegant Graphics for Data Analysis*. 2016: Springer-Verlag New York.
40. Olety, B., et al., *Myosin 1G (Myo1G) is a haematopoietic specific myosin that localises to the plasma membrane and regulates cell elasticity.* FEBS Lett, 2010. **584**(3): p. 493-9.
41. Singh, A., et al., *Cardiovascular Complications of Marijuana and Related Substances: A Review.* Cardiology and Therapy, 2018. **7**(1): p. 45-59.
42. Rezkalla, S. and R.A. Kloner, *Cardiovascular effects of marijuana.* Trends in cardiovascular medicine, 2019. **29**(7): p. 403-407.
43. Jones, R.T., *Cardiovascular system effects of marijuana.* The Journal of Clinical Pharmacology, 2002. **42**(S1): p. 58S-63S.
44. Goyal, H., H.H. Awad, and J.K. Ghali, *Role of cannabis in cardiovascular disorders.* Journal of thoracic disease, 2017. **9**(7): p. 2079.
45. Bojesen, S.E., et al., *AHRR (cg05575921) hypomethylation marks smoking behaviour, morbidity and mortality.* Thorax, 2017. **72**(7): p. 646-653.
46. Tantoh, D.M., et al., *Methylation at cg05575921 of a smoking-related gene (AHRR) in non-smoking Taiwanese adults residing in areas with different PM2.5 concentrations.* Clinical Epigenetics, 2019. **11**(1): p. 69.
47. Philibert, R., et al., *AHRR methylation predicts smoking status and smoking intensity in both saliva and blood DNA.* Am J Med Genet B Neuropsychiatr Genet, 2020. **183**(1): p. 51-60.
48. Prince, C., et al., *Investigating the impact of cigarette smoking behaviours on DNA methylation patterns in adolescence.* Human Molecular Genetics, 2018. **28**(1): p. 155-165.
49. Alexandrov, L.B., et al., *Mutational signatures associated with tobacco smoking in human cancer.* Science, 2016. **354**(6312): p. 618-622.
50. Phillips, D.H., *Smoking-related DNA and protein adducts in human tissues.* Carcinogenesis, 2002. **23**(12): p. 1979-2004.
51. Alexandrov, L.B., et al., *Mutational signatures associated with tobacco smoking in human cancer.* Science (New York, N.Y.), 2016. **354**(6312): p. 618-622.

Chapter 3

3. Validating DNA methylation using bisulfite-based amplicon sequencing (BSAS)

3.1 Introduction

Epigenetic modifications, such as DNA methylation, play a vital role in regulating gene expression [1] and have the potential to induce phenotypic changes [2-6]. DNA methylation occurs when a methyl group is covalently transferred to the C5 position of the cytosine ring of a DNA molecule by a methyltransferase enzyme, with the resulting modified cytosine then termed 5-methylcytosine (5mC) [7]. In mammals, most DNA methylation occurs at CpG dinucleotides. CpG sites themselves can be defined as a singular modified cytosine residue which are found throughout the genome, but are particularly dense in promoter regions [8].

DNA methylation is heavily influenced by the surrounding environment; factors such as tobacco smoking [9-12], alcohol [13, 14], nutrition [15, 16], stress [17] and aging [18, 19] can all impact on DNA methylation at CpG sites. Alterations to DNA methylation are associated with changes in phenotype and also, in some instances, methylation changes contribute to disease pathology [20-23].

3.1.1 Gold standard for DNA methylation analysis

As a result of these relatively recent observations, the assessment of differential DNA methylation in humans, and in particular, epigenome-wide association studies (EWAS), is a burgeoning field. High-throughput array technologies are a popular choice for EWAS, due to their robustness and accuracy [24]. The Illumina Infinium® MethylationEPIC array (hereafter 'EPIC array') quantifies methylation at 850,000 different CpG sites [25], and although this is still a small proportion of the total number of CpG sites in the genome (~28 million [26]) it represents a broad distribution of sites that give a specific and robust measurement of methylation at those sites.

3.1.2 Targeted techniques for the detection of differential DNA methylation

Further, the goal of many whole-genome studies of DNA methylation is often a pilot or scoping study to capture a range of targets that may be associating with, e.g., a particular environmental exposure. As such, once the genome has been investigated in a number of samples, a whole-genome approach is not always necessary if the user simply requires follow up and/or validation of identified loci in a larger cohort. To undertake further analyses and to validate methylation array-based experiments, several different methods exist that rely on bisulfite treatment of DNA: bisulfite-based amplicon sequencing (BSAS), bisulfite pyrosequencing and methylation-specific PCR (MS-PCR) are methods which can specifically target a predetermined area of interest in the genome at a low cost and higher sample throughput, compared to arrays. An informative study conducted by the BLUEPRINT consortium evaluated 27 predefined genomic regions, across 32 reference samples amongst 18 laboratories using six assays [27]. Good agreement was observed across methods, with amplicon bisulfite sequencing, and bisulfite pyrosequencing showing the best concordance [27]. A similar study also assessed bisulfite pyrosequencing, observing congruence to EPIC array analysis [28]. However, pyrosequencing is known to have quantitative flaws due to the output of sequences generated through fluorescence methods [29]. MS-PCR is a method often used in clinical settings [30], however it has a high false positive rate [31]. By contrast, BSAS detects cytosine methylation to base-pair scale resolution without reliance on light detection methods for sequencing [32]. BSAS is a multiplex procedure that can quantitatively assess each CpG site within numerous target regions at the same time [33].

Thus, given the limitations of pyrosequencing and MS-PCR, here we examine whether BSAS can also accurately validate EPIC array data, and be used as a replication, and/or expansion tool for targeted DNA methylation analyses, similar to what has been shown using pyrosequencing. Further, we wish to assess the multiple other CpG sites residing within the targeted amplicon region, to investigate differential methylated regions, which would not be able to be explored via EPIC array.

3.1.3 Study design

We used EPIC array data generated in Chapter 2 using the CHDS which evaluated differential DNA methylation in response to regular cannabis use (Chapter 2) [12].

For validation analysis we selected new individuals (N= 82), to serve as a validation and expansion cohort for the differential DNA methylation identified via EPIC array [12]. Specifically, we asked whether BSAS, after determination of the most appropriate normalisation method, produced the same average methylation values as EPIC arrays, when comparing case data to control data.

While both EPIC array and BSAS are readily used as standalone experiments, they would provide robust evidence if carried out together. Establishing a better understanding of how differential DNA methylation differs between regions within the genome, such as evaluating concordance between methods and then further assessing resultant CpG sites within a designated region, is valuable to the scientific community.

3.2 Methods

3.2.1 Illumina EPIC array samples and statistical analysis

Illumina EPIC array methods are described in Chapter 2.

3.2.2 Cohort selection and DNA extraction – BSAS experiments

BSAS analysis was carried out on two groups: cannabis plus tobacco users (N= 44) and controls (N= 38), who had never used cannabis. In contrast to the EPIC array analysis, no cannabis-only participants were used in BSAS; this is a consequence of the small number of individuals who use cannabis but who do not also use tobacco. Cannabis users were all selected on the basis that they either met DSM-IV diagnostic criteria [34] for cannabis dependence or had reported using cannabis consumption on a daily basis for a minimum of three years prior to age 28. Participants were matched as closely as possible for the following variables, sex, ethnicity, and parental socioeconomic status (Table 3.1). All participants were collected across a four month period so they are all of a similar age. Collection and analysis of DNA in the Christchurch Health and Development Study was approved by Southern Health and Disability Ethics Committee (CTB/04/11/234/AM10). DNA extraction protocols are previously described in [35]. Specifically, DNA was extracted from whole blood samples using a Kingfisher Flex System (Thermo Scientific, Waltham, MA USA) and quantified via nanodrop (Thermo Scientific, Waltham, MA USA). DNA was bisulfite treated using the EZ DNA Methylation-Gold kit (Zymo Research, USA) as per the manufacturer's instructions.

Table 3.1 The Christchurch Health and Developmental Study cohort selected for analysis by BSAS. Cases: cannabis and tobacco users; Controls: never cannabis users.

	Cases	Controls
Individuals	N= 44	N= 38
Gender		
Male	84%	76%
Female	16%	24%
Ethnicity		
Maori	20%	8%
Pacific Island	7%	3%
Asian	0%	0%
European	73%	89%
Socioeconomic status		
Professional/managerial	20%	37%
Clerical/technical	41%	39.%
Semi-skilled/unskilled	39%	24%
Tobacco smoking status		
Never	9%	92%
Occasional	4%	3%
Regular	87%	5%

3.2.3 CpG site selection, primer design and amplification – BSAS

A total of 15 CpG sites, representing 15 individual probes from the Illumina EPIC array were chosen based on their differential methylation status in cannabis plus tobacco users compared to controls (Table 3.2). A range of probes at differing levels of significance (not significant, nominally significant, and significant after P value adjustment) were chosen to reflect the range of data provided by the EPIC arrays. Primers to amplify bisulfite-treated DNA were designed using the online tool BiSearch [36] to amplify a ~250 base pair region which spanned the CpG site (Table 3.2). At the 5' end of each primer sequence, an Illumina overhang (33 base pair sequence) was included to ensure the ability to pool the amplicons and barcode them for high-throughput sequencing. All product lengths were all between 226 and 340 base pairs. To ensure primer specificity, Delta G's were designed to be no lower than -9 kcal/mol for efficiently, using the tool OligoAnalyzer (IDT®). A total of 30 primer pairs were initially designed for this experiment, and 15 of these are discussed here, as these were the primer pairs which performed efficiently at first usage.

Table 3.2 Forward and reverse primers used to target validation sites using bisulfite amplicon sequencing CpG sites including an Illumina overhand sequence.

Primer name	Illumina probe ID	Bisulfite converted primer (including Illuminia overhang sequence)
SLC17A7_F	Cg 02624701	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGGTATTAGAAGATTTYGAAGTTGTTT
SLC17A7_R	Cg 02624701	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGAAAATAAACCTATTCTCTCC
AHRR_F	Cg 05575921	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGGTTTTTTTGGTGTGGTTTTA
AHRR_R	Cg 05575921	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAG ACCACCATCTTATCTTATTT
ITPR1_F	Cg 08987995	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAG GATGGAATTTATTAGTGTTT
ITPR1_R	Cg 08987995	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGTAAAACACAACCCATTATCT
MAGI2_F	Cg 21121803	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGAGTTTAATTGAGTGTTTTTGAGG
MAGI2_R	Cg 21121803	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAG ACCCATTTTATTTATACCTTT
EHMT2_F	Cg 07829740	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAG AGAGGGGTTTAAATTTAAGTTTG
EHMT2_R	Cg 07829740	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAG CTAATAAATCACATATCTCC
PPM1L_F	Cg 26406186	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAG AATGTTAGTTGAATAAGTGG
PPM1L_R	Cg 26406186	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGCCACAAAAATACTCTAAAAAC
DPP10_F	Cg 05868547	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAG TTAAGGGAAGAAAGAAATGT
DPP10_R	Cg 05868547	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGCCTCTATAACAACATTTACTCAA
NIPAL4_F	Cg 17695979	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGGTTGGGAGAATTTATTTTTAGAG
NIPAL4_R	Cg 17695979	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGATATACCTATCACCAACTTC
CHD7_F	Cg 19926587	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAG TTTTAAAAGGATTTAAGGTAATG
CHD7_R	Cg 19926587	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGACTAACACAAAACAACCCAAT
PRDM5_F	Cg 01118724	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGATTTAAAAATGGTTGTGGTGAAG
PRDM5_R	Cg 01118724	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGTCCACTCATTACTCATATACTA
Cg11977356_F	Cg 11977356	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGGGAGGTGAGATGTTTTAATAATT
Cg11977356_R	Cg 11977356	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGAATAAACTATAATCATACCCCTC
Cg09078959_F	Cg 09078959	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGGTTTGAAAAGGGGAAATTTA
Cg09078959_R	Cg 09078959	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGACACTTAATAAAACACCAATC
Cg00571101_F	Cg 00571101	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGTGATAGGATATAAGAAGAAAGTA
Cg00571101_R	Cg 00571101	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGTCTACTTCAACCTAAAACAA
Cg11293828_F	Cg 11293828	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGTAGGGGGTTAGAGTATTTATTTT
Cg11293828_R	Cg 11293828	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGTTTTACTTTACTTAACTTCTCCC
Cg01614625_F	Cg 01614625	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGATGGAATTAGAAATTTTGGG
Cg01614625_R	Cg 0161462	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAG CCTCTCCATTTTATTTCTTTAA

Bisulfite-converted DNA was amplified via PCR, using KAPA Taq HotStart DNA Polymerase (Sigma, Aldrich) under the following conditions: 95 °C for 10 min, followed by 40 cycles of 95 °C for 30 sec, 59 °C for 20 sec, 72 °C for 7 min, and finally held at 4 C° using the Mastercycler Nexus (Eppendorf, Australia). PCR products were then purified with the Zymo DNA Clean & Concentrator Kit™ (Zymo Research, USA).
Following the PCR, DNA was cleaned up with Agencourt® AMPure® XP beads (Beckman Coulter) and washed with 80% ethanol and allowed to air-dry. DNA was then eluted with 52.5 µl of 10 mM Tris pH 8.5 before being placed back into the magnetic stand. Once the supernatant had cleared, 50 µl of supernatant was taken up and aliquotted into a fresh 96-well plate. DNA samples were quantified using the Quant-iT™ PicoGreen™ dsDNA Assay kit (Thermo Fisher) using the FLUROstar® Omega (BMG Labtech). Sequence libraries were prepared using the Illumina MiSeq™ 500 cycle Kit V2, and sequenced on an Illumina MiSeq™ system at Massey Genome Services (Palmerston North, New Zealand).

3.1.4 Bioinformatic and statistical analysis – BSAS data

Illumina MiSeq™ sequences were trimmed using SolexaQA++ software and aligned to FASTA bisulfite converted reference sequences using the package Bowtie2 (version 2.3.4.3). Each individual read was then aligned to all reference sequences (GRCh37/hg19) using the methylation-specific package Bismark [37]. Bismark produced aligned mapped reads with counts for methylated and unmethylated cytosines at each CpG site, thus BSAS returns additional CpG sites to the intended validation target, as each sequencing read contains multiple CpG sites. Cytosine proportion is calculated based upon the number of cytosines divided by the number of cytosines with the additions of the number of thymines present:

$$(C/(C_1) + T).$$

Where:

C Average cytosine methylation
C₁: Methylated cytosines
T: Number of thymines present/ Unmethylated cytosines

This gave the average methylation β values for each individual at each given CpG site. These β values could be anywhere between 0 - 1, with a β equal to 1 indicating 100% methylation at that CpG site across all sequencing reads. These data were imported into R Studio (RStudio Version 3.3.0) and the edgeR package [38] was used to determine differential DNA methylation between cannabis users and controls; coverage level was set to greater or equal to "8" across unmethylated and methylated counts under the recommendations of [38] whereby the conservative rule of thumb is total count (both methylated and unmethylated is at least "8" in every sample. Within the data set 96.5% of the reads were above a methylation coverage of 50. A negative binomial generalised model was used to fit the counts (methylated and unmethylated reads) in regards to the two variable groups, using the below model:

$$Y \sim Cannabis + e$$

Where:

Y = methylation M ratio

Cannabis = A cannabis user

e ~ N(0,s)

Summary tables compiled of the CpG sites of interest with nominal P value significance and post multiple testing using false discovery rate (FDR) of less than 0.05 were considered to be statistically significant. A scatter plot including a linear regression line with adjusted R^2 values was generated in R to quantify the correlation between β values produced with EPIC array and BSAS. Adjusted R^2 values were calculated for: i) BSAS cases versus EPIC cases, and; ii) BSAS controls versus EPIC controls. A Bland Altman analysis [39] was used to compare the agreement of the two techniques. Means were log transformed and lower and upper levels of agreement with 95% confidence intervals were calculated. Welch two sample t-tests were carried out on each of the loci (cases and controls separately) to assess differences between the two methods. All graphs were constructed using the package ggplot2 (version 3.3.2) [40].

3.2 Results

3.2.1 Validation and replication of EPIC array data using BSAS:

The differences between average methylation (β values) of cannabis plus tobacco users (cases) and controls were calculated for each method (EPIC array and BSAS, Table 3.3).

When comparing case vs control data from EPIC and BSAS individually, no significant difference in average methylation between case and control was observed for either detection method, with the exception of cg05575921 in *AHRR* and cg09078959. *AHRR* was significant in both BSAS and EPIC (P= 0.006, P= 5.33×10^{-12}), and cg05575921 was found to only be significant under BSAS (P= 0.001, P= 0.665).

Table 3.3 CpG site differences from EPIC array and the BSAS methods at the 15 loci of differing levels of significance (not significant, nominally significant and after P value adjustment). *When a cg number is listed, then there is no known gene associated with that CpG site. GB-Gene Body.

				Illumina EPIC array			BSAS			Difference between methods
	Cg/Gene	Position in genome	Illumina ID	β difference	*P value*	*FDR Adjusted P value*	β difference	*P value*	*FDR Adjusted P value*	β difference
1	*AHRR*	Chr5, GB	cg05575921	-0.233	5.33E-12	3.7E-06	-0.041	0.006*	0.245	-0.192
2	cg11977356*	Chr19	cg11977356	-0.040	0.474	0.999	-0.004	0.406	0.959	-0.036
3	*ITPR1*	Chr3, GB	cg08987995	-0.001	0.572	0.999	0.005	0.820	0.822	-0.006
4	*MAGI*	Chr7, GB	cg21121803	-0.008	0.572	0.999	-0.007	0.809	0.959	-0.0004
5	*EHMT2*	Chr6, GB	cg07829740	0.005	0.037	0.999	-0.015	0.071	0.579	0.020
6	*PPM1L*	Chr3, GB	cg26406186	-0.006	0.818	0.999	0.011	0.904	0.963	-0.017
7	cg00571101*	Chr12	cg00571101	0.004	0.368	0.999	-0.004	0.813	0.952	0.008
8	cg09078959*	Chr5	cg09078959	-0.001	0.893	0.999	-0.005	0.001*	0.245	0.004
9	cg01614625*	Chr7	cg01614625	-0.009	0.370	0.999	-0.006	0.569	0.952	-0.004
10	*DP10*	Chr2, GB	cg05868547	0.006	0.077	0.999	-0.003	0.713	0.952	0.009
11	cg11293828*	Chr12	cg11293828	-0.014	0.665	0.999	0.032	0.735	0.952	-0.045
12	*CHD7*	Chr5, 5'UTR	cg19926587	-0.007	0.960	0.999	-0.006	0.429	0.959	-0.001
13	*NIPAL4*	Chr5, TSS1500	cg17695979	-0.007	0.714	0.999	-0.003	0.106	0.713	-0.004
14	*PRDM5*	Chr4, GB	cg01118724	-0.004	0.734	0.999	0.005	0.116	0.713	-0.009
15	*SLC17A7*	Chr19, GB	cg02624701	-0.043	0.312	0.999	0.018	0.646	0.952	-0.061

3.2.2 Linear regression between BSAS and EPIC

Correlations between BSAS and EPIC were plotted individually for cases and controls. BSAS versus EPIC cases resulted in an adjusted R^2 of 0.8878 and BSAS versus EPIC controls gave an adjusted R^2 of 0.8683 (Figure 3.1).

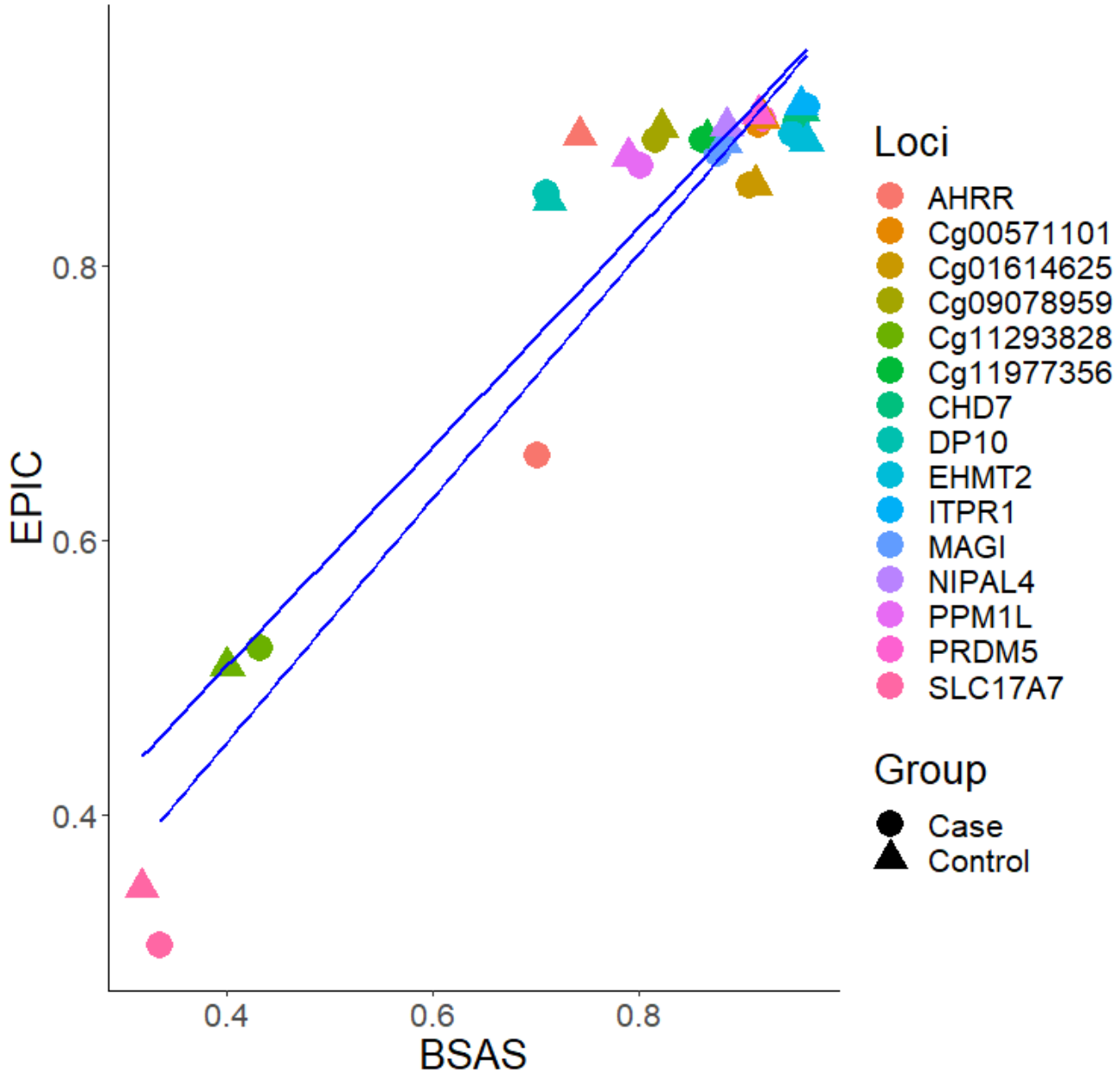


Figure 3.1 Scatter plots with linear regression of the β values at each loci for BSAS and EPIC array plotted against each other. Colours represent the loci of interest, with the shapes representing the case and controls. There are two regression lines: Correlation between cases with an adjusted R^2 =0.8878 and controls with R^2 = 0.8683.

3.2.3 Bland Altman correlations

A Bland Altman analysis was carried out on the loci investigated by BSAS and compared to data for the same loci produced using the Illumina EPIC array. Figure 3.2 A shows cannabis users (cases) measured using BSAS and the EPIC array on the X axis, while the Y axis represents the log differences between the measurements. The observed differences between loci in cannabis cases (EPIC and BSAS) fall within the lines of agreement. Figure 3.2 B shows the control group differences plotted for the same loci for BSAS and the EPIC array methods. Similar to above, all data points fall within the lower and upper lines of agreement.

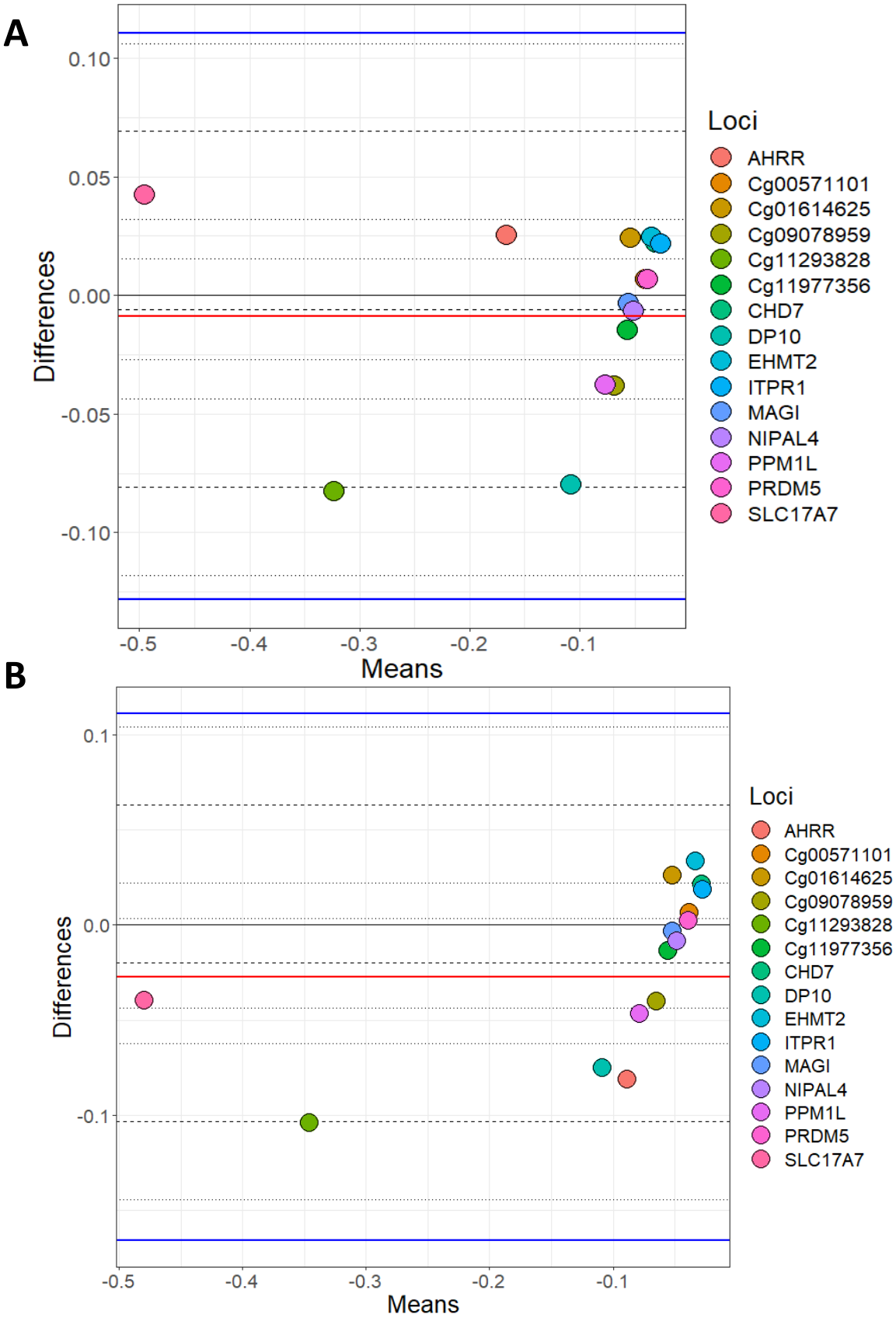
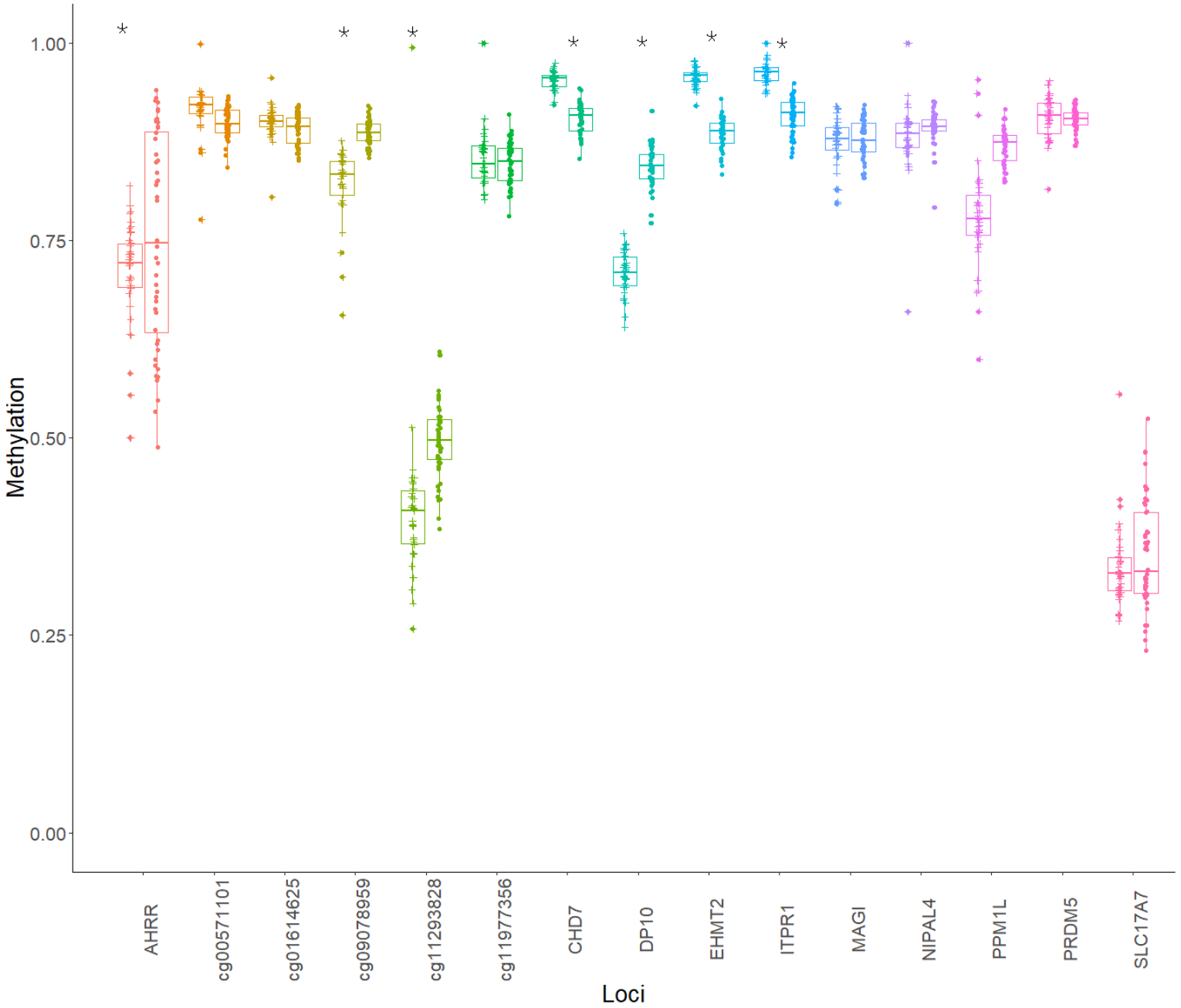


Figure 3.2 Bland Altman plots showing the log means differences between DNA methylation as measured by EPIC array vs. the same CpG sites measured using BSAS. A) Data from cannabis users, gathered using BSAS and the EPIC array (Cases) B) the control subjects used in BSAS and EPIC array. Each of the 15 points represent the CpG sites investigated.

3.2.4 Individual methylation across all 15 loci assessed for BSAS and EPIC

Mean methylation values for each individual were plotted for each of the 15 loci, and these were then compared between BSAS and EPIC, for cases (Figure 3.3 A) and controls (Figure 3.3 B). Loci displaying a significant shift in average methylation between the methods of detection are indicated with an * when using a Welsh two sample comparison. The following loci were found to display differences between BSAS and EPIC array: cases; *AHRR*, cg09078959, cg11293828, *CHD7*, *DP10*, *EHMT2* and *ITPR1*, and controls; *AHRR*, cg09078959, cg11293828, *CHD7*, *DP10*, *EHMT2*, *ITPR1, NIPAL4* and *PPM1L*.

A


Methylation
1.00
0.75
0.50
0.25
0.00
AHRR
cg00571101
cg01614625
cg09078959
cg11293828
cg11977356
CHD7
DP10
EHMT2
ITPR1
MAGI
NIPAL4
PPM1L
PRDM5
SLC17A7
Loci

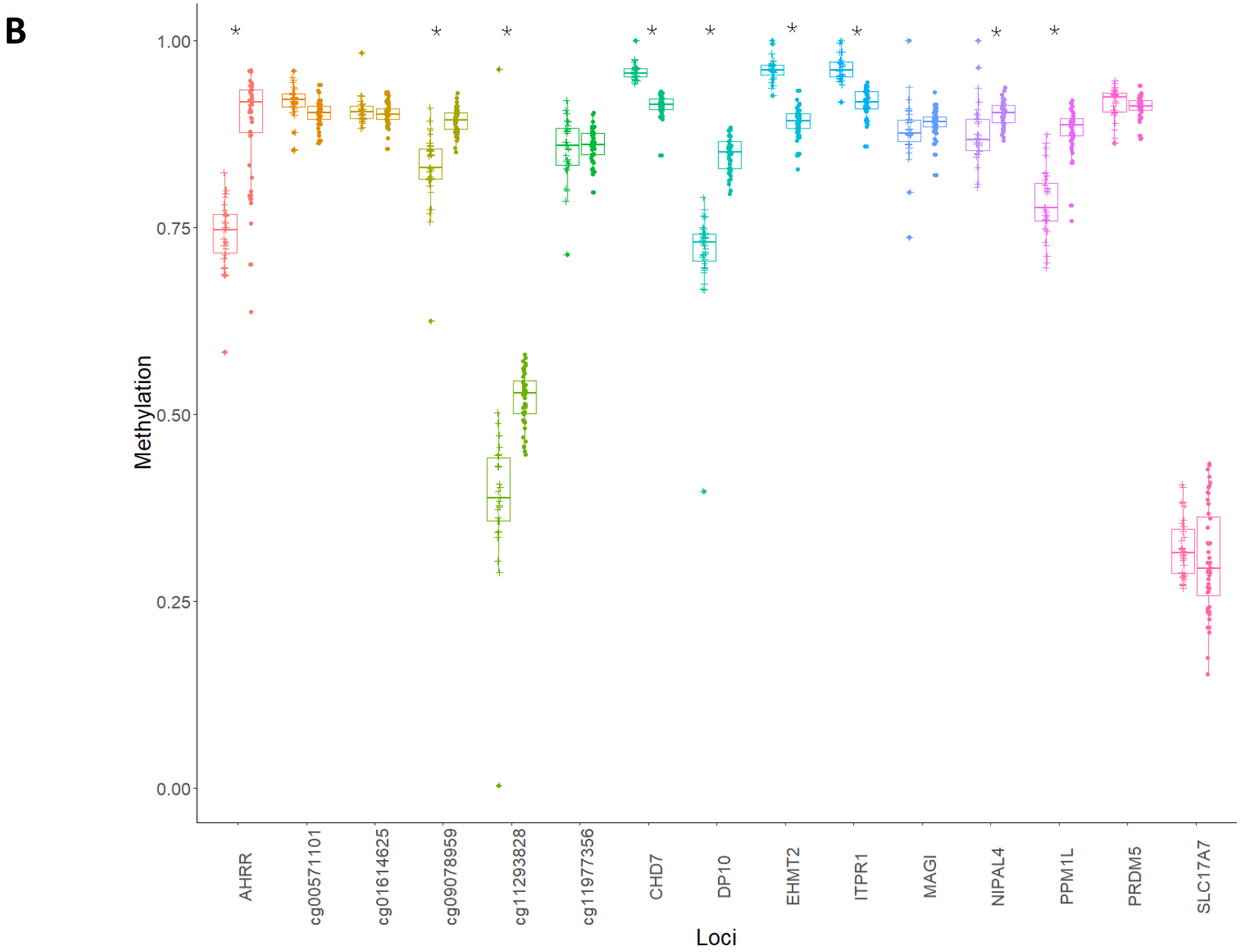


Figure 3.3 Average methylation for cases individuals across the 15 loci assessed using EPIC and BSAS . * represent those loci with significant differences in average methylation between EPIC and BSAS. A) case individuals B) control individuals for each of the studies.

3.2.5 Assessing amplicon regions

Multiple CpG sites residing within an amplicon can be sequenced using BSAS, providing information about a larger region of interest, rather than just a single CpG site. Figure 3.4 displays the multiple CpG sites found across each of the 15 amplicons in this study. A total of 9 of the 15 amplicons contained more than one CpG site. All CpG sites within the amplicons remained non-differentially methylated between cases and controls, except one site in *AHRR*. The amplicon from *SLC17A7* sequenced here contained a total of 15 CpG sites with in the 250 base pairs.

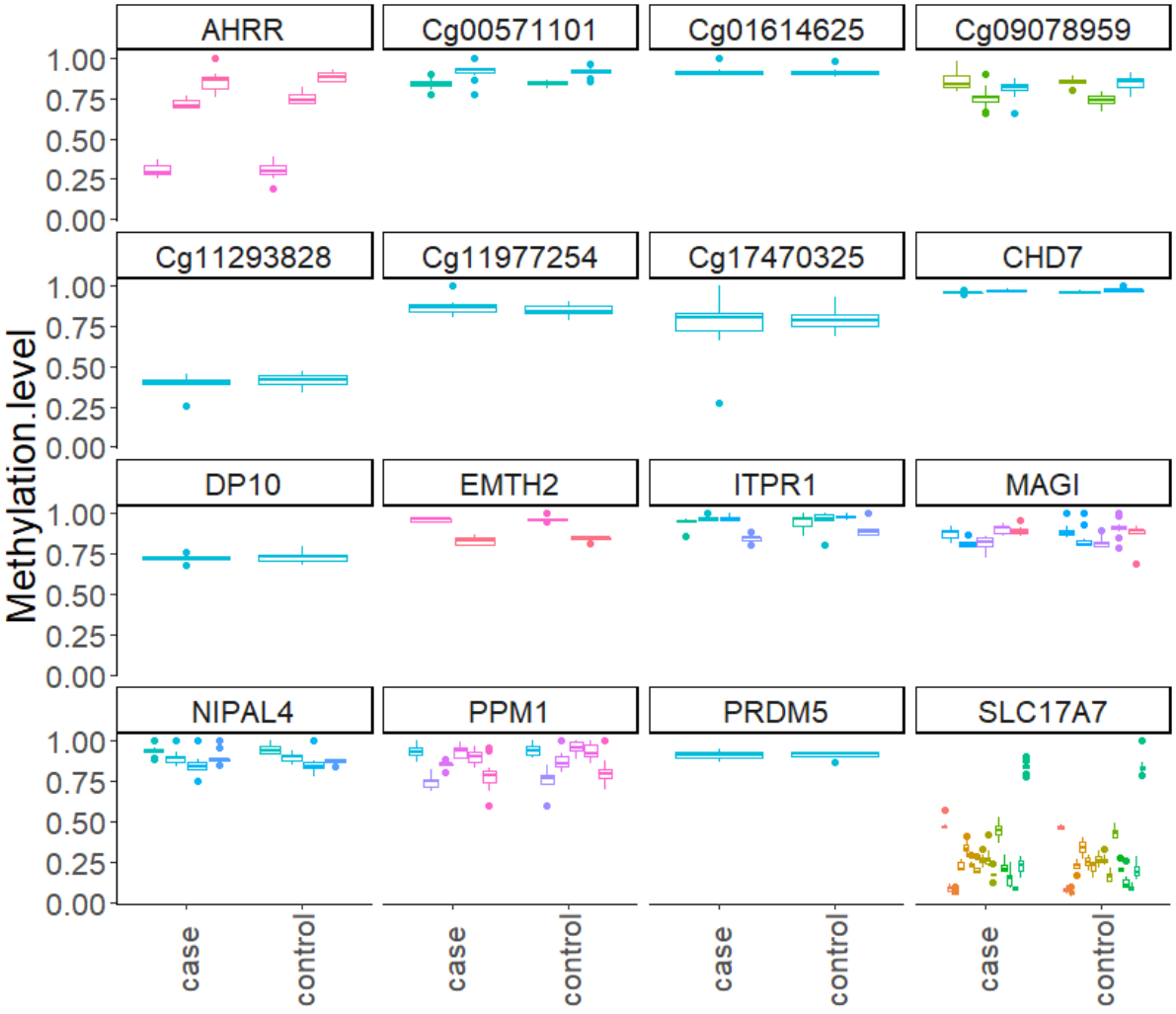


Figure 3.4 Average DNA methylation between cannabis users compared to controls across all CpGs that were investigated. A differing number of CpG sites are found within each amplicon.

3.3 Discussion

High throughput array technologies have facilitated the next step in assessing associations between DNA methylation changes in response to a known environmental exposure at a genome-wide level. The EPIC array (as well as the predecessor 27k and 450k arrays) is one such platform that allows for the characterisation of these DNA methylation changes. Through these approaches, various studies have furthered our understanding of how DNA methylation can play a role in response to different environmental exposures.

3.3.1 Validation of EPIC using BSAS

We selected the orthogonal method BSAS to determine its applicability as a validation, replication and/or expansion tool for EPIC array. BSAS is often used as a standalone method for assessing differential DNA methylation at specific CpG sites, usually because it is more cost-effective than EPIC arrays, and allows analysis of many samples at once, in multiplex. It returns data for all CpGs within a targeted region of interest (~250 base pairs) with results providing base pair-level specificity [32]. Overall, when considering average methylation between cases and controls as determined via BSAS or EPIC individually, we did not detect significant differences in average methylation for each detection method; the biological results are discussed in Chapter 2 [12] , however, it was expected that the smaller sample set used here would not have the statistical power to detect effects found in the larger cohort. The intent of this study was to compare average methylation as determined via BSAS, to that determined by EPIC array. We show here that the estimation of differential DNA methylation observed using BSAS correlated with differential methylation determined via EPIC array. However, although the data correlates between the methods (adjusted R^2 cases, 0.8878 and adjusted R^2 controls, 0.8683), we urge caution when interpreting this correlation as proof that BSAS will be a suitable independent validation of EPIC array data in every experiment. It is because while the data presented here correlated between BSAS and EPIC array as a whole dataset, some sites showed larger differences between average methylation estimated using BSAS vs. EPIC array. Most notably, where the differential methylation on EPIC array was greater than 5% between cases and controls, BSAS was unable to detect this differential DNA

methylation to the same magnitude as EPIC array. Further, a total of 9/15 loci had observed P value significance when carrying out a Welch two sample t-test on control data, and 7/15 on case data, implying there were differences between the methylation values for the methods. For instance, *AHRR* exhibited a 4% difference in methylation between cases and controls when assessed using BSAS (the highest value detected in using BSAS in this study), compared to 23% using EPIC array. Thus, while a strong correlation between EPIC array data and BSAS data was found across the 15 CpG sites investigated, which itself implies an association between the average methylation at each CpG for the two techniques, each locus must be validated on a case by case basis before being taken forward into high-throughput or large scale screening, to ensure it produces results that are equivalent to EPIC. In addition, further work on CpG sites with higher magnitude changes is needed to determine whether BSAS is limited by the magnitude of differential methylation it is able to detect. However, it is worth noting that most studies of differential methylation report modest (<5%) significant differential methylation observations, suggesting that BSAS may prove useful, given inclusion of rigorous controls of known differential methylation to ensure accuracy of results.

3.3.2 Advantages to using BSAS as a DNA methylation tool

Due to the sequence-based nature of BSAS data (compared to the probe-based nature of EPIC arrays) BSAS, as a standalone method, offers some advantages that are not applicable to EPIC arrays. For instance, BSAS has the potential to determine novel differentially methylated CpGs which may be near (in the same targeted region) but not the initial pre-determined CpG site of interest. This is possible because all CpGs within an, e.g., 500bp region are returned using BSAS data, only one of which may be on an EPIC array. Further, via this targeted sequencing process, BSAS may reveal novel differentially methylated regions (DMRs). DMRs are described as areas which exhibit multiple successive methylated CpG sites which may have biological impact within the genome [41]. Therefore targeting more than a single CpG site may provide further insight into genes and regions of interest. Consequently, while here we have used BSAS technology to replicate/validate differential methylation identified via EPIC array, given that BSAS outputs can correlate with EPIC data, equally, BSAS could be used as a “discovery-based tool”; if significantly differentially methylated

CpGs are identified via BSAS, this would serve to justify further investigation using a robust and more expensive high throughout method. The EPIC array still remains the most reproducible way to measure DNA methylation [42]. Largely, this is because the probe-based nature of the method frequently produces comparable results across research groups and arrays. For example, detection of differential methylation using the EPIC array found a difference of 23% in cannabis plus tobacco users, compared to controls, at *AHRR* (cg05575921, Table 3.3), a result that is supported by other studies in tobacco smokers using EPIC array [9, 43-46]. *AHRR* has an important role in controlling a range of different physiological functions; it contributes to regulation of cell growth, regulation of apoptosis and contributes to vascular and immune responses [47-50].

3.3.3 Methods of detection differences

BSAS and EPIC array rely upon different chemistries and methods to detect DNA methylation and this may account for the majority of the variation found between the two methods. BSAS relies upon PCR amplification of DNA that is treated with sodium bisulfite. When DNA is treated, unmethylated cytosine residues are converted into uracil via hydrolytic deamination. Amplification of uracil nucleotides during this process are replaced by thymine during replication and the 5-methylcytosines are left unreactive throughout the deamination process and then are amplified as cytosines. It then becomes possible to 'read' values of methylation for each cytosine in an amplicon via DNA sequencing [51]. The ability to treat DNA with sodium bisulfite has led to the expansion of research undertaken within this field [52]. However, it is important that we ensure the validity of the results are not limited by the manner in which the data was produced. Ensuring that we limit these discrepancies between technologies will allow for better validation of data. There is potential for errors to occur at this step, because incomplete bisulfite conversion cannot be distinguished from 5-methylcytosine, this can possibly introduce false positive methylation calls at this point [53] [54]. Although both techniques rely upon bisulfite treatment, it is this source of error followed by the PCR amplification that might explain the differences in results we have observed. Refining these sources of error may provide much more comparable results between the two methods.

3.4 Chapter summary

- We chose to validate EPIC array data by using the alternative method, BSAS, to detect differential methylation at CpG sites.
- While BSAS validated EPIC array data at some loci, and correlated across all loci as a whole, however some individual loci did not validate.
- BSAS was unable to reproduce the magnitude of changes that are shown in the EPIC array system, which may be a consequence of lack of specificity and addition error rate through PCR amplification.
- BSAS does offer some advantages such as being able to assess differentially methylated regions, rather than individual CpG site

3.5 References

1. Hackett, J.A. and M.A. Surani, *DNA methylation dynamics during the mammalian life cycle.* Philosophical Transactions of the Royal Society B: Biological Sciences, 2013. **368**(1609): p. 20110328.
2. Dolinoy, D.C., D. Huang, and R.L. Jirtle, *Maternal nutrient supplementation counteracts bisphenol A-induced DNA hypomethylation in early development.* Proceedings of the National Academy of Sciences, 2007. **104**(32): p. 13056-13061.
3. Sinclair, K.D., et al., *DNA methylation, insulin resistance, and blood pressure in offspring determined by maternal periconceptional B vitamin and methionine status.* Proceedings of the National Academy of Sciences, 2007. **104**(49): p. 19351-19356.
4. Kucharski, R., et al., *Nutritional control of reproductive status in honeybees via DNA methylation.* Science, 2008. **319**(5871): p. 1827-1830.
5. Gertz, J., et al., *Genistein and bisphenol A exposure cause estrogen receptor 1 to bind thousands of sites in a cell type-specific manner.* Genome research, 2012. **22**(11): p. 2153-2162.
6. Wang, H., et al., *Widespread plasticity in CTCF occupancy linked to DNA methylation.* Genome research, 2012. **22**(9): p. 1680-1688.
7. Mitchell, C., L.M. Schneper, and D.A. Notterman, *DNA methylation, early life environment, and health outcomes.* Pediatr Res, 2016. **79**(1-2): p. 212-9.
8. Takai, D. and P.A. Jones, *Comprehensive analysis of CpG islands in human chromosomes 21 and 22.* Proceedings of the National Academy of Sciences, 2002. **99**(6): p. 3740-3745.
9. Zeilinger, S., et al., *Tobacco Smoking Leads to Extensive Genome-Wide Changes in DNA Methylation.* PLOS ONE, 2013. **8**(5): p. e63812.
10. Breton, C.V., et al., *Prenatal tobacco smoke exposure affects global and gene-specific DNA methylation.* American journal of respiratory and critical care medicine, 2009. **180**(5): p. 462-467.
11. Ambatipudi, S., et al., *Tobacco smoking-associated genome-wide DNA methylation changes in the EPIC study.* Epigenomics, 2016. **8**(5): p. 599-618.
12. Osborne, A.J., et al., *Genome-wide DNA methylation analysis of heavy cannabis exposure in a New Zealand longitudinal cohort.* Translational Psychiatry, 2020. **10**(1): p. 114.
13. Philibert, R.A., et al., *The impact of recent alcohol use on genome wide DNA methylation signatures.* Frontiers in genetics, 2012. **3**: p. 54-54.
14. Liu, C., et al., *A DNA methylation biomarker of alcohol consumption.* Molecular Psychiatry, 2016. **23**: p. 422.
15. Delgado-Cruzata, L., et al., *Dietary modifications, weight loss, and changes in metabolic markers affect global DNA methylation in Hispanic, African American, and Afro-Caribbean breast cancer survivors.* J Nutr, 2015. **145**(4): p. 783-90.
16. Rampersaud, G.C., et al., *Genomic DNA methylation decreases in response to moderate folate depletion in elderly women.* The American Journal of Clinical Nutrition, 2000. **72**(4): p. 998-1003.
17. Murgatroyd, C., et al., *Dynamic DNA methylation programs persistent adverse effects of early-life stress.* Nature Neuroscience, 2009. **12**: p. 1559.
18. Horvath, S., et al., *Aging effects on DNA methylation modules in human brain and blood tissue.* Genome Biology, 2012. **13**(10): p. R97.
19. Marioni, R.E., et al., *DNA methylation age of blood predicts all-cause mortality in later life.* Genome Biology, 2015. **16**(1): p. 25.
20. Kim, M., et al., *DNA Methylation as a Biomarker for Cardiovascular Disease Risk.* PLOS ONE, 2010. **5**(3): p. e9692.
21. Mastroeni, D., et al., *Epigenetic changes in Alzheimer's disease: Decrements in DNA methylation.* Neurobiology of Aging, 2010. **31**(12): p. 2025-2037.
22. De Jager, P.L., et al., *Alzheimer's disease: early alterations in brain DNA methylation at ANK1, BIN1, RHBDF2 and other loci.* Nature Neuroscience, 2014. **17**(9): p. 1156-1163.
23. Rakyan, V.K., et al., *Identification of Type 1 Diabetes–Associated DNA Methylation Variable Positions That Precede Disease Diagnosis.* PLOS Genetics, 2011. **7**(9): p. e1002300.
24. Pidsley, R., et al., *Critical evaluation of the Illumina MethylationEPIC BeadChip microarray for whole-genome DNA methylation profiling.* Genome biology, 2016. **17**(1): p. 208-208.

25. Zhou, W., P.W. Laird, and H. Shen, *Comprehensive characterization, annotation and innovative use of Infinium DNA methylation BeadChip probes.* Nucleic acids research, 2017. **45**(4): p. e22-e22.
26. Lövkvist, C., et al., *DNA methylation in human epigenomes depends on local topology of CpG sites.* Nucleic Acids Research, 2016. **44**(11): p. 5123-5132.
27. Bock, C., et al., *Quantitative comparison of DNA methylation assays for biomarker development and clinical applications.* Nature Biotechnology, 2016. **34**(7): p. 726-737.
28. Roessler, J., et al., *Quantitative cross-validation and content analysis of the 450k DNA methylation array from Illumina, Inc.* BMC Research Notes, 2012. **5**(1): p. 210.
29. França, L.T., E. Carrilho, and T.B. Kist, *A review of DNA sequencing techniques.* Quarterly reviews of biophysics, 2002. **35**(2): p. 169-200.
30. Herman, J.G., et al., *Methylation-specific PCR: a novel PCR assay for methylation status of CpG islands.* Proc Natl Acad Sci U S A, 1996. **93**(18): p. 9821-6.
31. Claus, R., et al., *A systematic comparison of quantitative high-resolution DNA methylation analysis and methylation-specific PCR.* Epigenetics, 2012. **7**(7): p. 772-780.
32. Masser, D.R., D.R. Stanford, and W.M. Freeman, *Targeted DNA methylation analysis by next-generation sequencing.* Journal of visualized experiments : JoVE, 2015(96): p. 52488.
33. Masser, D.R., A.S. Berg, and W.M. Freeman, *Focused, high accuracy 5-methylcytosine quantitation with base resolution by benchtop next-generation sequencing.* Epigenetics & Chromatin, 2013. **6**(1): p. 33.
34. Association, A.P., *Diagnostic criteria from dsM-iV-tr*. 2000: American Psychiatric Pub.
35. Noble, A.J., et al., *Epigenetic signatures associated with the observed link between maternal tobacco use during pregnancy, and offspring conduct problems in childhood and adolescence.* bioRxiv, 2020: p. 2020.07.02.183285.
36. Arányi, T., et al., *The BiSearch web server.* BMC Bioinformatics, 2006. **7**(1): p. 431.
37. Krueger, F. and S.R. Andrews, *Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications.* Bioinformatics, 2011. **27**(11): p. 1571-1572.
38. Chen, Y., et al., *Differential methylation analysis of reduced representation bisulfite sequencing experiments using edgeR.* F1000Research, 2017. **6**: p. 2055-2055.
39. Martin Bland, J. and D. Altman, *STATISTICAL METHODS FOR ASSESSING AGREEMENT BETWEEN TWO METHODS OF CLINICAL MEASUREMENT.* The Lancet, 1986. **327**(8476): p. 307-310.
40. Wickham, H., *ggplot2: Elegant Graphics for Data Analysis*. 2016: Springer-Verlag New York.
41. Hotta, K., et al., *Identification of differentially methylated region (DMR) networks associated with progression of nonalcoholic fatty liver disease.* Scientific Reports, 2018. **8**(1): p. 13567.
42. Bibikova, M., et al., *Genome-wide DNA methylation profiling using Infinium® assay.* Epigenomics, 2009. **1**(1): p. 177-200.
43. Li, S., et al., *Causal effect of smoking on DNA methylation in peripheral blood: a twin and family study.* Clinical epigenetics, 2018. **10**: p. 18-18.
44. Ambatipudi, S., et al., *Tobacco smoking-associated genome-wide DNA methylation changes in the EPIC study.* Epigenomics, 2016. **8**(5): p. 599-618.
45. Su, D., et al., *Distinct Epigenetic Effects of Tobacco Smoking in Whole Blood and among Leukocyte Subtypes.* PLOS ONE, 2016. **11**(12): p. e0166486.
46. Guida, F., et al., *Dynamics of smoking-induced genome-wide methylation changes with time since smoking cessation.* Human Molecular Genetics, 2015. **24**(8): p. 2349-2359.
47. Lahvis, G.P., et al., *The aryl hydrocarbon receptor is required for developmental closure of the ductus venosus in the neonatal mouse.* Mol Pharmacol, 2005. **67**(3): p. 714-20.
48. Allan, L.L. and D.H. Sherr, *Constitutive activation and environmental chemical induction of the aryl hydrocarbon receptor/transcription factor in activated human B lymphocytes.* Mol Pharmacol, 2005. **67**(5): p. 1740-50.
49. Trombino, A.F., et al., *Expression of the aryl hydrocarbon receptor/transcription factor (AhR) and AhR-regulated CYP1 gene transcripts in a rat model of mammary tumorigenesis.* Breast Cancer Res Treat, 2000. **63**(2): p. 117-31.
50. Marlowe, J.L., et al., *The aryl hydrocarbon receptor binds to E2F1 and inhibits E2F1-induced apoptosis.* Mol Biol Cell, 2008. **19**(8): p. 3263-71.
51. Booth, M.J., et al., *Oxidative bisulfite sequencing of 5-methylcytosine and 5-hydroxymethylcytosine.* Nature Protocols, 2013. **8**: p. 1841.
52. Frommer, M., et al., *A genomic sequencing protocol that yields a positive display of 5-methylcytosine residues in individual DNA strands.* Proc Natl Acad Sci U S A, 1992. **89**(5): p. 1827-31.

53. Richards, R., et al., *Evaluation of massively parallel sequencing for forensic DNA methylation profiling.* Electrophoresis, 2018. **39**(21): p. 2798-2805.
54. Krueger, F., et al., *DNA methylome analysis using short bisulfite sequencing data.* Nature Methods, 2012. **9**: p. 145.

3.6 Supplementary Tables

Table 3.1 All CpG sites assessed for differential DNA methylation using BSAS

Gene	*Log FC*	*Log CPM*	*L R*	*P value*	*FDR*
AHRR	-0.112	11.080	7.216	0.007	0.268
SLC17A7	0.067	13.325	6.608	0.010	0.268
Cg09078959	-0.113	13.039	5.265	0.021	0.384
AHRR	-0.101	10.992	3.668	0.055	0.627
ITPR1	-0.068	14.057	3.274	0.070	0.627
EMTH2	-0.092	12.163	3.259	0.070	0.627
NIPAL4	0.069	13.011	2.613	0.105	0.773
PRDM5	-0.070	11.987	2.459	0.116	0.773
PPM1	-0.122	13.185	2.146	0.142	0.807
SLC17A7	0.076	13.104	2.047	0.152	0.807
SLC17A7	0.047	13.029	1.522	0.217	0.880
PPM1	-0.186	13.220	1.473	0.224	0.880
Cg17470325	-0.112	12.695	1.341	0.246	0.880
NIPAL4	0.050	12.909	1.238	0.265	0.880
NIPAL4	0.041	13.367	1.237	0.265	0.880
MAGI	-0.052	13.036	1.159	0.281	0.880
SLC17A7	-0.049	13.336	1.140	0.285	0.880
SLC17A7	0.027	13.549	1.079	0.298	0.880
SLC17A7	0.028	13.279	0.886	0.346	0.952
Cg09078959	-0.035	13.368	0.683	0.408	0.952
Cg11977356	-0.040	12.720	0.678	0.410	0.952
AHRR2	-0.033	11.343	0.640	0.423	0.952
CHD7	-0.031	13.351	0.625	0.429	0.952
SLC17A7	0.028	13.141	0.583	0.445	0.952
CHD7	-0.026	13.010	0.565	0.451	0.952
Cg01614625	-0.012	14.179	0.323	0.569	0.952
SLC17A7	0.015	13.579	0.319	0.571	0.952
SLC17A7	-0.018	13.086	0.280	0.596	0.952
SLC17A7	-0.014	12.674	0.273	0.600	0.952
MAGI	0.035	13.172	0.270	0.602	0.952
SLC17A7	0.013	13.394	0.211	0.645	0.952
ITPR1	0.027	13.919	0.197	0.654	0.952
MAGI	0.029	12.283	0.187	0.665	0.952
SLC17A7	0.004	12.982	0.177	0.673	0.952
MAGI	-0.019	13.162	0.176	0.674	0.952
PPM1	-0.020	13.684	0.156	0.692	0.952
DP10	-0.012	13.062	0.134	0.713	0.952
cg11293828	0.044	12.764	0.114	0.734	0.952

PPM1	-0.018	13.486	0.107	0.743	0.952
ITPR1	0.021	14.114	0.100	0.751	0.952
SLC17A7	-0.008	13.422	0.088	0.766	0.952
MAGI	-0.011	12.997	0.058	0.808	0.952
Cg00571101	-0.019	12.931	0.055	0.813	0.952
ITPR1	0.017	14.156	0.052	0.818	0.952
EMTH2	0.028	11.863	0.052	0.819	0.952
NIPAL4	-0.004	13.273	0.035	0.850	0.952
SLC17A7	-0.006	13.461	0.034	0.851	0.952
Cg09078959	0.009	13.468	0.029	0.863	0.952
Cg00571101	0.010	13.313	0.022	0.880	0.952
PPM1	0.006	13.005	0.014	0.904	0.958
SLC17A7	0.002	13.567	0.003	0.952	0.972
SLC17A7	-0.001	13.369	0.003	0.954	0.972
PPM1	-0.0006	13.425	0.0005	0.981	0.981

Chapter 4:

4. Development of the zebrafish *(Danio rerio)* as a model for assessing the impact of THC and CBD on DNA methylation

4.1 Introduction

So far, this thesis has worked with human cohorts to address the impact of an individual's environment (cannabis, tobacco exposures) on DNA methylation. Yet, assessing the impact of any one specific environmental exposure on DNA methylation is not without its challenges. The main challenge is that each individual's cumulative environmental exposures can vary greatly, and exposures also change throughout an individual's lifetime. Meaning that, in human cohorts, it can be challenging to definitively attribute differential DNA methylation to one cause or one exposure.

To counteract this diversity of exposures, here we utilise the model organism, *Danio rerio* (zebrafish), to assess the specific impacts of the most abundant cannabinoids within cannabis (THC and CBD) on DNA methylation, an experiment which is unable to be easily undertaken in humans. We then choose to address these two main hypotheses: i) THC and CBD exposure causes DNA methylation patterns compared to non-exposed in the zebrafish, and; ii) differential DNA methylation in response to THC and CBD will be identified within genes and pathways that are specific to the biological response to THC and CBD.

This research will therefore allow us to determine whether the zebrafish is an appropriate system for exploring the precise epigenetic effects of human cannabis exposure - while there is precedence for the use of zebrafish in cannabinoid research [15-17] [18, 19]. Their utility and applicability to probe the molecular basis of the biological response to cannabis has not yet been established. It will also provide novel insights into the specific genomic targets of THC and CBD, contributing to the scant knowledge in this area around the precise molecular effects of each component. Lastly, this research will allow us to better understand the health implications of cannabis use and will seed future research into the epigenetics of environmental exposures.

4.1.1 The zebrafish as a model organism

Model systems are an important component of research into the genetic bases of human diseases, and there are numerous well-established systems in which we can study human disease. There is no 'gold standard' model system for all research purposes [1]; each system is unique, and careful consideration is taken to weigh the positive and negative attributes of a model. Usually, the trade-off is between genetic similarity of models, and cost efficiency, because a higher degree of genetic similarity is associated with a higher research cost.

The zebrafish is an exceptional model for the study of embryonic development and has been utilised for genetic research for decades [2, 3]. Further, it has been vital in allowing the observation of developmental traits and the genetic basis of phenotypes like disease and behaviour. For example, studies of reward, learning, aggression and anxiety have all conserved regulatory processes in zebrafish and mammals [4], and thus, reward behaviour such as ethanol [5], nicotine [5] and opiates [6] have all been evaluated in the zebrafish. Further, zebrafish are a well-established model in which to study the epigenetic effects of environmental exposures such as nutrition and stress [7] and often, the zebrafish is the first port of call for toxicology research [8, 9]. More specifically, zebrafish have been used for analysis in response to environmental containments such as arsenic [10], bisphenol A [11] and benzopyrene [12], and importantly, studies such as the above are generally unable to be undertaken in human cohorts. Lastly, a vast network of data is curated in ZFIN (the Zebrafish Model Organism Database) which serves as a resource for genomic information and molecular tools for zebrafish research [13]. Thus, given its long history of use, its well-characterised genome, and the wealth of publicly available data on its genome, phenotypes and development, the zebrafish is a highly tractable model system to use to investigate the epigenetic effects of the environment.

4.1.2 Zebrafish and DNA methylation patterns

Zebrafish are an appropriate and relevant system in which to explore the effects of the environment on the epigenome because:

i) zebrafish have similar DNA methylation machinery to humans [14] and there is consistent distribution of 5-methylcytosine between zebrafish and mammals [14];

ii) numerous studies have explored cannabis and cannabinoid biology using zebrafish [15-17], particularly focussing on differences in gene expression [18, 19],

iii) cannabinoids induce behavioural effects in zebrafish that are comparable to some of those reported for mammals [20], with stimulation of locomotion at low concentration of cannabinoids, and suppression at higher concentrations [15],

iv) many basic cellular and molecular pathways, regulated by different compounds, are similar between zebrafish and mammals [15, 21, 22] and;

v) their applicability as a model of epigenetics in health and disease is becoming increasingly clear [23].

As such, they are an appropriate species in which to model the genomic and phenotypic consequences of environmentally induced methylation changes.

4.1.3 The endocannabinoid system in the zebrafish

Here we will investigate the epigenetic impact of (-)-trans-Δ^9-tetrahydrocannabinol (THC) and (-)-cannabidiol (CBD) on methylation in the zebrafish genome. THC binds to CB1 or CB2 receptors and is the psychoactive agent of cannabis [24] [25]. The CB1 receptor resides primarily within the central nervous system, mainly within key motor and behavioural centres, such as neocortex [26], olfactory system [26], hippocampus [26, 27] basal ganglia [28], cerebellum [28] and amygdala [26] and is the most abundant cannabinoid receptor [29, 30]. CBD is not thought to be psychoactive as it has a weaker affinity for the CB1 receptor [31]. CBD is recognised for its purported medicinal benefits, which are thought to be mediated via its binding to the CB2

receptor [32, 33]. The CB2 receptor is more generally distributed throughout the body, and less so in the central nervous system, in humans [34, 35]. The CB1 and CB2 receptor locations are conserved between humans and zebrafish [36], further supporting the use of zebrafish to quantify the impact of cannabinoid exposure [37].

4.1.5 DNA methylation, cannabinoid exposure and the zebrafish

Most recent research into zebrafish cannabinoid exposure has largely focussed on morphological and gene expression analysis, with little/no consideration of the epigenetic effects. Intriguingly, in zebrafish, CBD mirrors the developmental, morphological and behavioural impacts of THC, at much lower concentrations [38], however, it is unclear if the genomic basis of these similar phenotypic effects is shared between THC and CBD. Further, the effect of THC and CBD exposure on genome-wide DNA methylation patterns in zebrafish has not been established. Thus here, we use reduced representation bisulfite sequencing (RRBS) to quantify genome-wide DNA methylation patterns in response to THC and CBD exposure in zebrafish embryos. RRBS was chosen due to being a non-specific species method for DNA methylation detection as the Illumina EPIC array is specific to humans in its current form. We identify which CpG sites are differentially methylated in response to each ingredient, determine shared sites and conclude which genes and pathways are specifically targeted by each ingredient, paving the way for future research into the biological impacts of THC and CBD.

4.2 Methods

4.2.1 Zebrafish

Both male and female zebrafish (*Danio Rerio*, TB X pet shop), from the Otago Zebrafish facility Dunedin (New Zealand) were used for breeding. Zebrafish were kept in 45 L glass tanks containing ~35 fish per tank. The light cycle consisted of 14 h light and 10 h dark (lights on at 09.00). The temperature of the room was set at 28 ± 1 °C.

4.2.2 Breeding and embryo collection

The day before the morning of breeding, 1 female and 1 male (fish in a box) were set up in a 1.7 L beach breeding tank (Techniplast). The number of tanks set up would differ between experiments depending on the number of embryos needing to be produced. Males and females were separated by a divider overnight, which was removed with the onset of light. Pairs of zebrafish would then breed when the light cycle began at 09.00. Fish were then left for 1 h before they were put back into their designated tanks and their embryos collected. Embryos were then stored in 75 ml (cell culture containers) at N= 100 embryos per container and packaged with a heat pack and sent to the University of Canterbury, Christchurch, via overnight shipping. All embryos arrived within 24 hour post fertilisation (hpf). Embryos were placed in 28°C incubators if arrival was prior to 24 hpf.

4.2.3 Embryo treatment

Cannabidiol (CBD,1 mg/ml in EtOH 1 ml) was acquired from Echo pharmaceuticals (Leiden, Netherlands), and stored under a Ministry of Health Authority to Possess Medicines (Research/Study/Analysis) (Authority No: RI4570013-02). (-)-delta-9-tetrahydrocannabinol (THC, 1 mg/ml in EtOH 1 ml) was acquired from Echo pharmaceuticals (Leiden, Netherlands) and stored under a Ministry of Health Licence to Possess Controlled Drugs (Licence No: RI6910080-00).

Lethal Concentration50 (LC_{50}) experiments for treatments of both CBD and THC were carried out. Whereby, thec oncentration in the water was tested to determine the concentration which kills 50% of the zebrafish during the course of the observational period. Serial dilutions of concentrations for CBD and THC were identified from prior literature and used for initial calculations of LC_{50} (Table 4.1). Both CBD and THC are solubilised in ethanol, so an ethanol control was also added as an additional vehicle control to ensure that the effects seen from both CBD and THC were not due to ethanol. Thus, the 24 hpf embryos were exposed to either: i) e3 media (control); ii) e3 media with vehicle (ethanol); iii) e3 with CBD (0.6, 0.3, 0.15, 0.075 µg/ml), or; iv) e3 with THC (0.3, 0.6, 1.2, 2.5 µg/ml) for 96 hours (ceasing exposure at 120 hpf). Each concentration was set up in a petri dish with approx. 50 embryos per dish which each contained the desired concentration of CBD or THC in 30 ml of e3. Embryos were then left in an incubator at 28°C. The exposure medium was not replaced for the entirety of the time course. Probit and logit equations were used to assess the best concentrations for THC and CBD, which were taken forward into experiments to produce treatment embryos for DNA extraction, as per the methodology above.

Embryos were scored from 58 hours onwards at 30 min time intervals for individual hatchings from each of the treatment groups. Kaplan Meier [39] curves were fitted to the hatching data to describe survival probability with survival probability at time t, St, given by:

$$S_t = \frac{\begin{array}{c}Number\ of\ embryos\ at\ the\ begining\\ of\ treatment\ alive\ -\end{array}\ Number\ of\ embryos\ that\ hatched}{\begin{array}{c}Number\ of\ embryos\ at\ the\ begining\\ of\ treatment\ alive\end{array}}$$

The package survival [40] was used in R studio to return to Kaplan-Meier estimate and graphs were constructed using ggplot2.

4.2.4 DNA extraction

Embryos were removed from their treatment at 120 hpf and tissue was stored in TRIzol™ Reagent (Thermo Scientific, MA USA) in 1.5mL Eppendorf tubes at -20°C. For DNA extraction, liquid was then removed from the Eppendorf tubes and tissue was lysed in Solid Tissue Lysis buffer (Zymo Research, USA) and Proteinase K (Zymo Research, USA) for 3 h at 55°C. The kit *Quick*- DNA Miniprep Plus (Zymo Research, USA) was used to carry out the extraction as per the manufacturer's recommendations. DNA was assessed for quality using gel electrophoresis and Nanodrop™ (Thermo Scientific, Waltham, MA USA).

4.2.5 RRBS preparation

A total of 1 µg of DNA, from each of the treatment groups (control, vehicle ethanol, THC and CBD) were sent in duplicates to Custom Science (Auckland, New Zealand) for RRBS libraries to be prepared and sequencing to be carried out. Raw data was returned as FASTQ files and then processed in-house via the following pipeline (Figure 4.1):

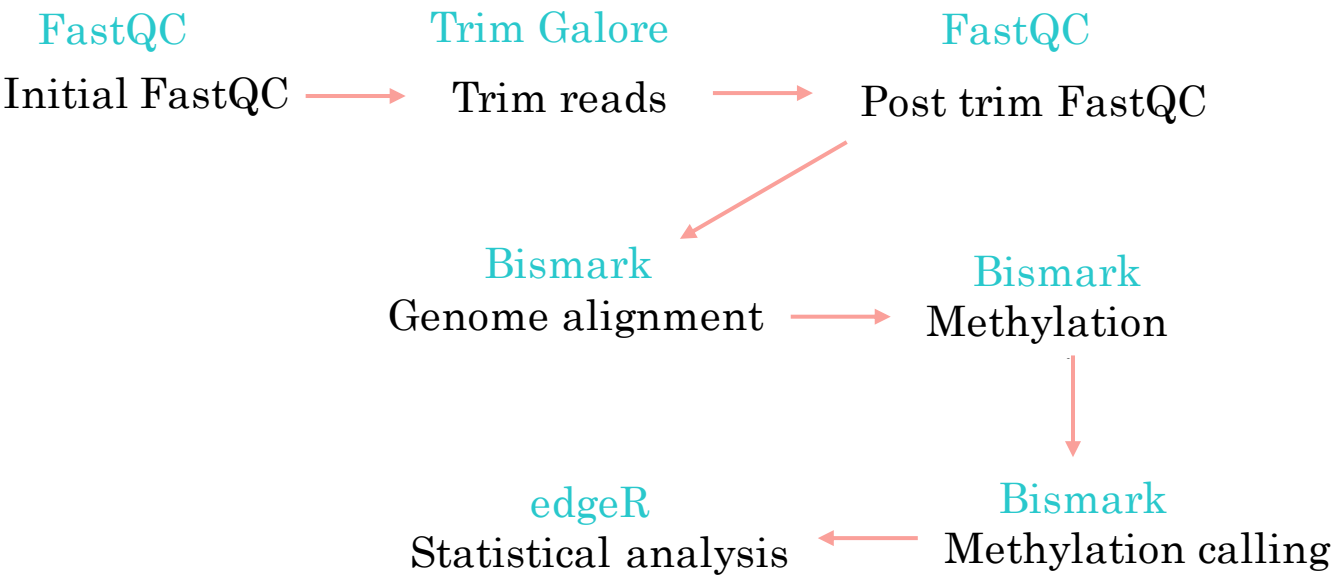


Figure 4.1 The pipeline used for quality control, pre-processing and methylation calling of RRBS data. Programs and methods are referenced in section 4.2.5.2.

4.2.6 Quality Control and alignment

Raw data was initially assessed for quality using the package FastQC. All sequences were then trimmed using the package Trim Galore (Version 0.6.5), and the 2 bp from the 3' end of the reads was trimmed to avoid the filled in cytosine position close to the second MspI site. A phred quality score cut-off was set at 20, and sequences of less than 20 bp were also removed. Trimmed reads for each of the treatments in duplicate were then aligned to the zebrafish reference genome (Version Zv9 –danRer7) using bowtie2 (Version 2.4.2). Genome indexing and methylation calling was carried out using Bismark (Version 0.22.3). Histograms were used to assess the frequency of percentage methylation these were plotted using ggplot2.

4.2.7 Methylation calling

Bismark coverage files were then loaded into R Studio (Version 3.3.0) and analysed using the package edgeR (Version 4.0). A minimum read coverage score of 10 was used to assess for differential CpG methylation. The two exposure duplicated were then pooled together for analysis. A linear model was applied to assess the difference between control, vehicle control, CBD and THC treatment groups.

4.2.8 Determining differential DNA methylation and gene regions

Two tables were generated with the number of significantly differentially methylated CpG sites for each treatment group (THC and CBD). The first contained the number differentially methylated CpG sites after FDR correction, and the second contained the number CpG sites which were differentially methylated with nominal P values ($P < 0.001$). Each table also showed the direction of methylation change (either hypermethylated or hypomethylated), this was displayed as MA and UpSetR plots (Version 1.4.0)[41] was used to assess for any overlapping CpG sites between the treatment groups. The top 50 most differentially methylated CpG sites for both CBD and THC were generated. We included the names of genes which housed the differentially methylated CpG sites were identified using the online tool annotation tool, GREAT [42]. GREAT assigns a gene to a CpG site if that CpG site is within 1000 kb

of the gene's regulatory domain and is particularly useful for genomes with missing annotation. CpG sites which were unable to assign to a gene were left blank.

4.2.9 Pathway analysis

To identify Molecular Function (MF) and Biological Process (BP) gene ontology pathways that are enriched in each of the treatment groups, Fish Enrichr [43, 44] was used. Fish Enrichr was supplied with a list of genes identified as housing CpG sites with nominal P values of $P < 0.001$. MF assess activities of molecules that perform actions and BP is a larger process of broader molecular functions. Fish Enrichr uses Fisher's exact test to assess the probability of a gene belonging to a set or given pathway. The pathways are then corrected for multiple testing via the Benhamini-Hochberg method. Rank or z-scores were also assigned which is a modification to Fisher's exact test for a deviation from the expected presence/absence of genes in the supplied lists. The combined enrichment score is thus a combination of the P value and the z-score.

4.3 Results

4.3.1 Calculating working solutions of cannabinoids

To calculate the working concentration for exposure of zebrafish embryos to CBD and THC, previous literature was consulted to establish a working range (Table 4.1). Our literature search showed that range of different exposure times were used, so that each may mimic a different stage in development, specific to the aim of each individual study. However, across all studies, all treatments were initiated between 5 hpf and 24 hpf.

Table 4.1 THC and CBD exposure concentration ranges and observed phenotypic differences identified from recent scientific literature and utilised here as a starting point for LC50 determination.

Treatment	Observations	Reference
CBD (1-4 mg/l) THC (2-10 mg/l)	Phenotype differences seen at 1 mg/l and above Phenotype differences seen at 2 mg/l and above	[45]
THC (1-10 mg/l)	Optimal THC concentration- 6 mg/l	[46]
CBD (0.07-1.25 mg/l) THC (0.3-5 mg/l)	CBD LC_{50} 0.53mg/l THC LC_{50} 3.65 mg/l LOAEL- pericardial edema THC 0.60, CBD 0.07 mg/l LOAEL jaw malformations THC- 5,CBD 0.3 mg/l LOAEL- axis curvature THC 2.5, CBD 0.6 mg/l LOAEL-trunk degradation THC 2.5, CBD 0.6 mg/l	[38]
THC (0.0.024- 0.6 mg/l exposing F0, F1 populations CBD (0.006-0.15 mg/l) exposing F0, F1 populations	Gene expression changes found to be different between both THC and CBD compared to controls in F0 population	[47]

4.3.2 Lethal concentrations of CBD and THC

Both the logit and probit values were generated from both cannabinoid LC_{50} experiments and were plotted against log concentrations (Supplementary Figure 4.1). THC concentrations (Supplementary Figure 4.1 A and B) were higher than that of CBD (Supplementary Figure 4.1 A and B). These values suggest that zebrafish embryos have a much higher mortality rate in CBD compared to THC at the same concentration.

Taking into account both the LC_{50} values for both CBD and THC and the previous literature (Table 4.1), treatment concentrations were calculated as 0.15 mg/l for CBD and 0.60 mg/l for THC.

4.3.3 Developmental and hatching differences between the different treatments

Embryos were exposed at 24 hpf to working concentrations of CBD (0.15 mg/l), THC (0.6 mg/l), vehicle ethanol (0.60 mg/l) and the non-exposed control group and left until 120 hpf at 28°C (96 hr exposure in total).

4.3.4 Hatching efficiency between treatments groups

From 57.5 hpf onwards, the number of embryos hatched was counted. Differences were seen between each of the treatment groups from the first initial recorded hatching (Figure 4.2). The rate at which hatching commences differs between each of the treatment groups. The control group is observed to have half the number of embryos hatched prior to 60 hpf. Both vehicle ethanol and THC have similar hatching efficiency rates with half the number of embryos hatched at 62 hpf. The CBD treatment group shows the greatest delay in hatching with half the number of embryos hatched at 65 hpf, with this being the most significant difference also (Table 4.2).

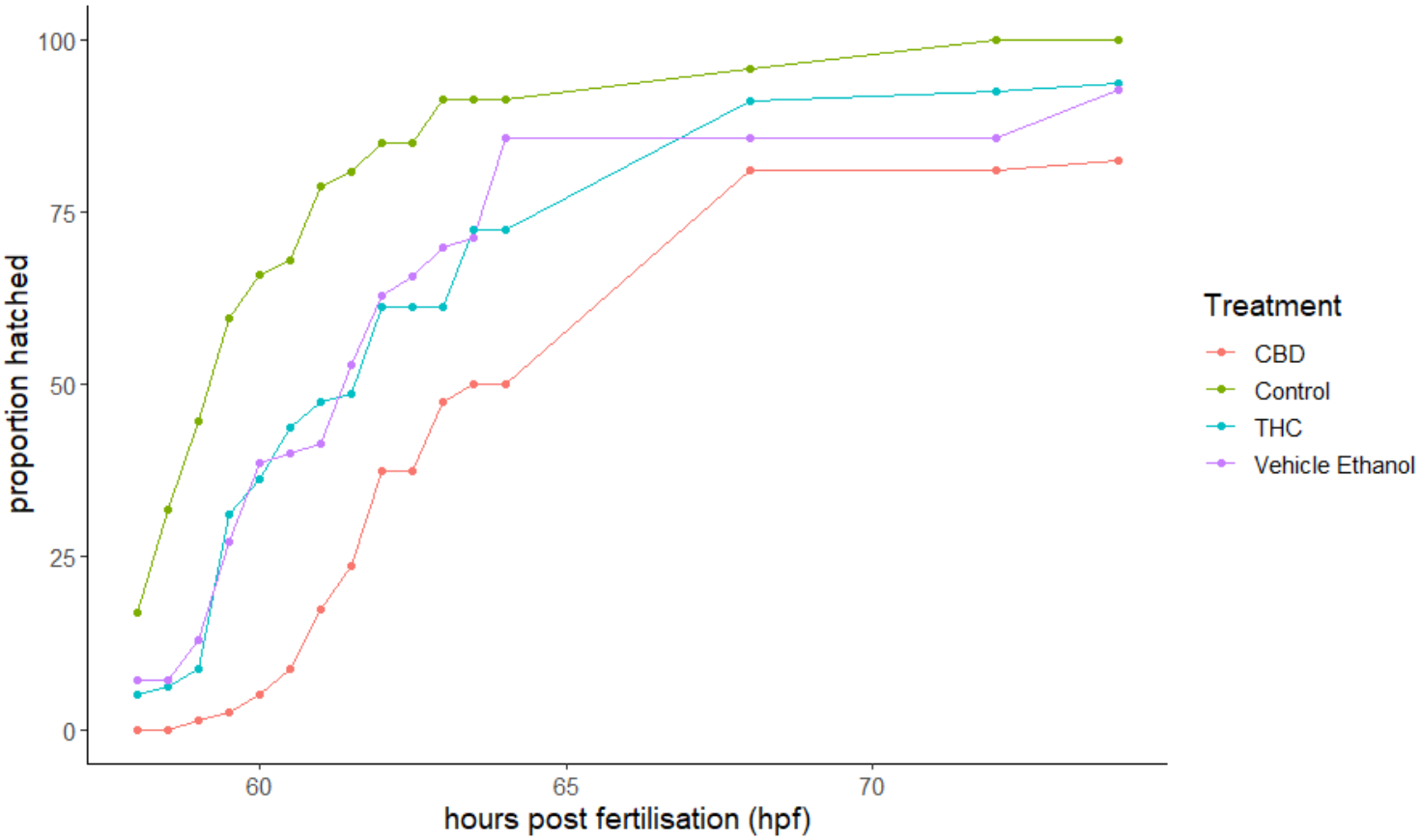


Figure 4.2 The proportion of embryos hatched at each of the time points for which data was collected, from 57.5 hpf. The number of embryos from each of the treatment groups were counted and plotted.

4.3.5 Survival probability

The hatching efficiency data was then assessed for quantitative differences between the control and each treatment group. The Kaplan-Meier method was fitted to estimate the statistical differences between each of the groups (Table 4.2 and Figure 4.2).

Table 4.2 The proportion of hatched embryos for each treatment groups compared to controls assessed for survival rate using the Kaplan-Meier method.

Treatment	Coefficient	Expected (coefficient)	Standard error (coefficient)	Z Pr(>|z|)
Vehicle Ethanol	-0.7778	0.4594	0.1933 -4.024	5.73×10^{-5} ***
CBD	-1.3694	0.2543	0.1945 -7.041	1.90×10^{-12} ***
THC	-0.8247	0.4384	0.1879 -4.389	1.14×10^{-5} ***

4.4 DNA methylation analysis

4.4.1 Genome alignment

To calculate differential DNA methylation between control and treatment groups, each of the eight samples (4 for each group, in duplicate) were mapped and aligned to the zebrafish reference genome (alignment statistics Table 4.3). Each sample differed slightly in mapped alignment ranging from 49.2% to 55.5%.

The coverage threshold was set to 10X; anything below this was disregarded from further analysis. Leaving between 2,273,488- 2,806,070 CpG sites per sample for analysis, the package edgeR was used for determining average methylation between samples.

Table 4.3 Genome alignment post processing information for the eight samples used for RRBS analysis. Sequence pair analysed- The total number of sequencing reads per sample. Number of reads <10X- The number of CpG sites which had greater than 10 reads.

Sample	% Aligned	Sequences analysed	Total no. CpG	Number of reads < 10X	Cytosine methylated in CpG context	Cytosine methylated in CHG context	Cytosine methylated in CHH context
Control-1	54.1	20606670	868622138	2699384	79.0	0.9	0.7
Control-2	55.5	39615887	1378858120	2806070	79.6	0.8	0.7
Vehicle ethanol-1	52.6	40424260	1366928024	2796302	78.1	0.8	0.7
Vehicle ethanol-2	50.6	32197049	1049997800	2518000	78.3	0.8	0.7
CBD-1	53.0	37913057	1269584406	2486994	78.8	0.8	0.7
CBD-2	49.2	25537266	793471211	2273488	76.5	0.8	0.7
THC-1	52.5	38781171	1283356654	2737199	78.6	0.8	0.7
THC-2	54.3	36925292	1246988505	2628683	81.4	0.8	0.7

4.4.2 Frequency of the percentage of methylation for the samples used for RRBS

The frequency of the number of reads based off the percentage methylated per sample were plotted as histograms (Figure 4.2 as exemplars of Control-1 and THC-1, remainder found in Supplementary Figure 4.2. The reads are bimodal with the highest counts observed at either 0% or above < 85% methylation in both examples.

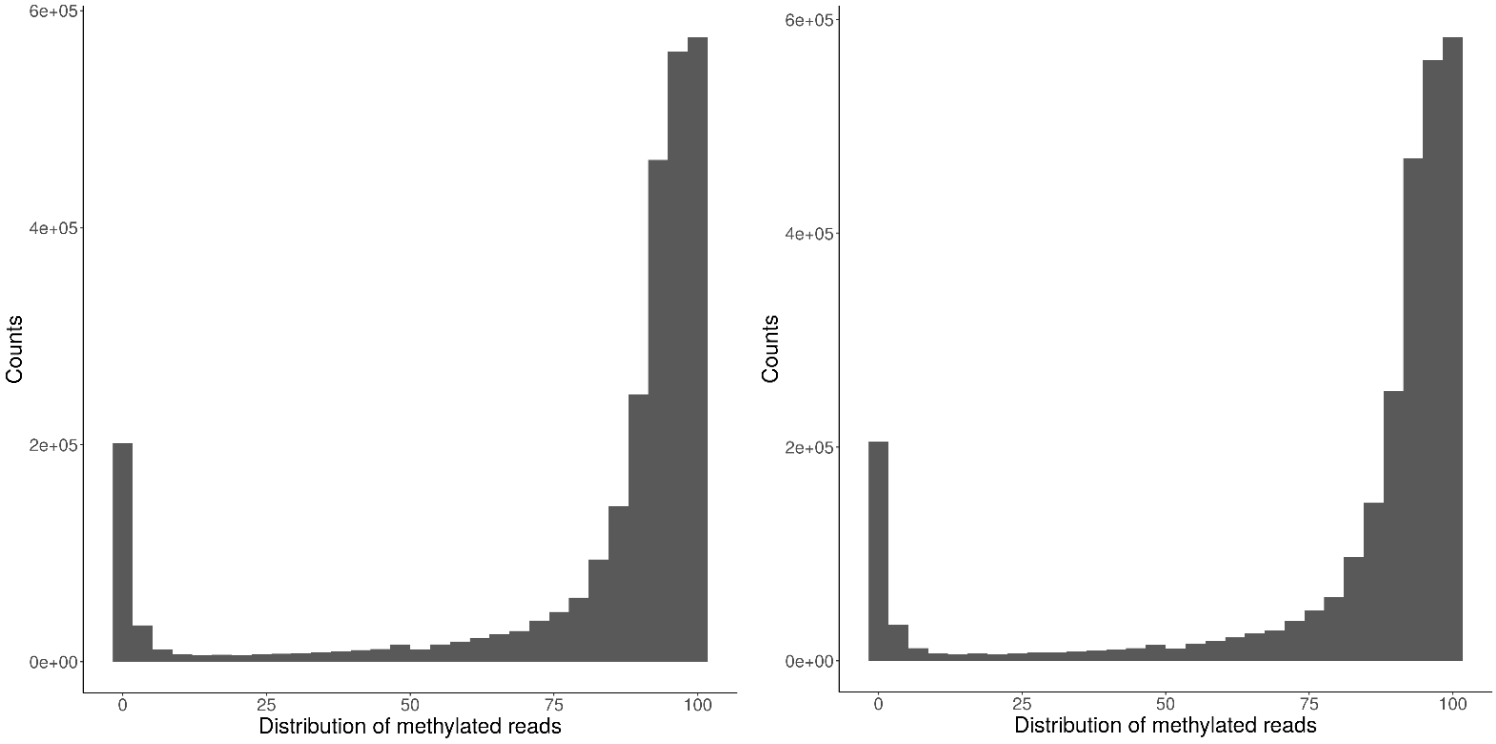


Figure 4.3 The distribution of reads measured by the distribution of methylated reads from the samples. Here are two examples of the distributions, with a control-1 sample of the left and THC-1 on the right.

4.4.3 Differential DNA methylation in each of the treatment groups

Each of the different exposures were compared to the control to asses for differential DNA methylation (Table 4.4 and Figure 4.4). The vehicle ethanol control was found to display a total of N= 662 CpG sites differentially methylated after FDR correction, compared to the control group. We identified N= 1939 sites significantly differentially methylated CpG sites in response to CBD treatment (Figure 4.4A), and N= 9 in response to THC (Figure 4.4B).

The differentially methylated CpG sites identified in the vehicle ethanol and CBD groups (Table 4.4 and Figure 4.4A) showed similar distributions of both hypermethylation and hypomethylation. Those differentially methylated CpG sites identified in response to THC showed a tendency towards hypomethylation (Table 4.4 and Figure 4.4B), however this may simply be an artefact of the small number of sites identified in response to THC. Any CpG sites in black that are more differentially methylated than the coloured dots are not significant due to standard error.

Table 4.4 Number of FDR adjusted significantly differentiated sites found with between the different treatment groups.

	Hypermethylated	Hypomethylated	Total
Vehicle ethanol	349	316	662
CBD	1005	934	1939
THC	2	7	9

Lists of significantly differentially methylated CpG sites in each treatment group were assessed to determine whether significant CpG sites overlapped between treatment groups (Figure 4.4c). No differentially methylated CpG sites were common to THC, CBD and the vehicle ethanol groups collectively. The vehicle ethanol and CBD groups shared the greatest number of differentially methylated CpG sites (N= 78) while CBD and THC shared N= 1. The one overlapping CpG site was the most significantly differentially methylated CpG in the THC treatment group, and 94th in the CBD treatment group. N=1860 CpG sites were therefore unique to CBD exposure, and 8 were unique to THC exposure.

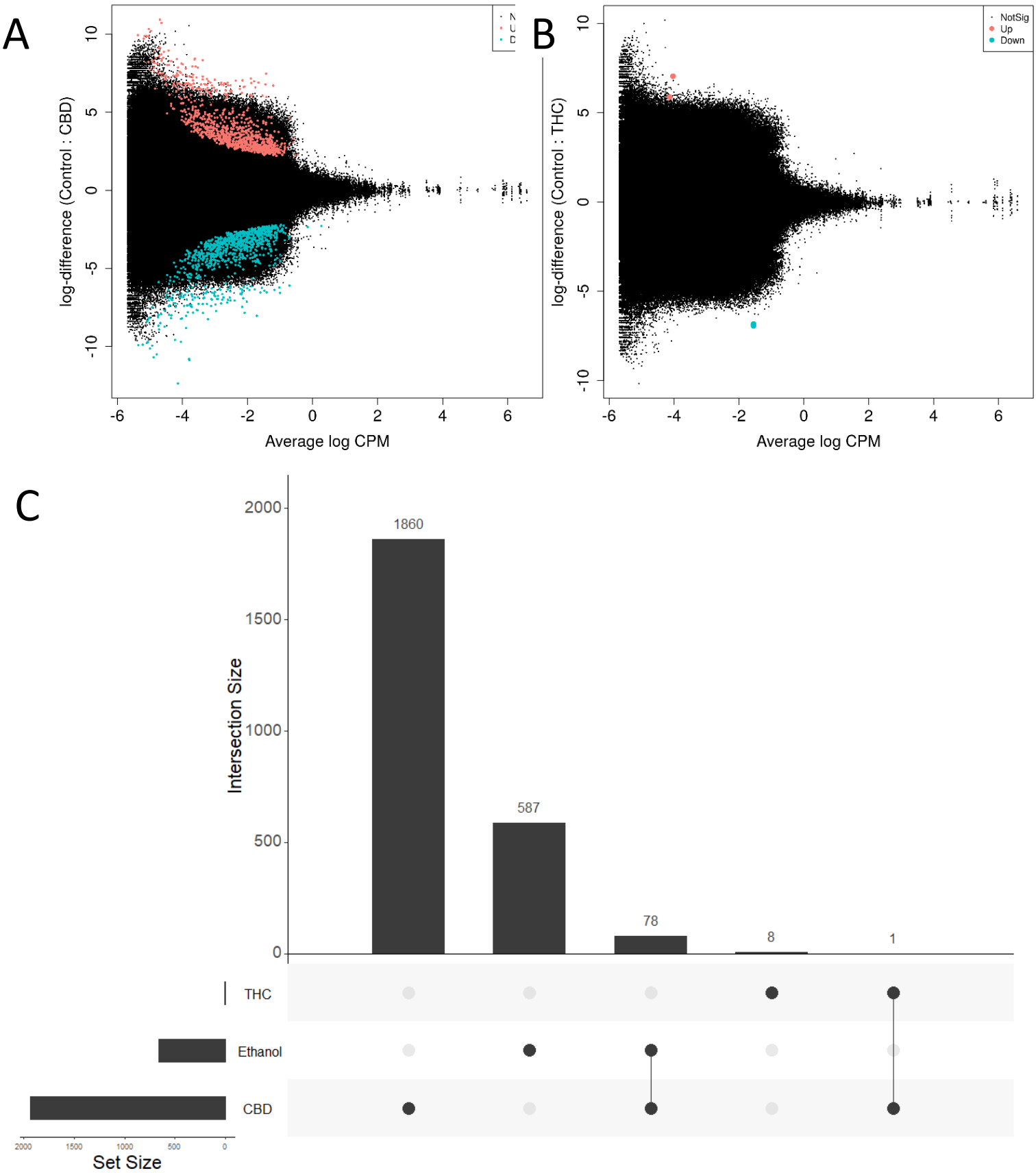


Figure 4.3 The top FDR corrected CpG sites found to be differentially methylated in response to CBD exposure (A), THC exposure (B). (C) – an upset plot to demonstrate shared or unique CpG sites between the treatment groups and the vehicle ethanol group.

4.4.4 Differential DNA methylation sites within each treatment group with nominal P value significance

Given the comparatively small number of differentially methylated CpG sites in response to THC that were significant after FDR correction, a less stringent threshold of significance was applied to allow an assessment of further overlap between the treatment groups. Consequently, differentially methylated CpG sites displaying a nominal P values of < 0.001 were counted (Table 4.5). Vehicle ethanol had a total of N= 7741 CpG sites that were differentially methylated at an unadjusted P < 0.001, N= 12148 were identified in response to CBD exposure, and THC exposure resulted in N= 3769 differentially methylated sites, when a less stringent significance threshold was used.

Table 4.5 The number of differentially methylated CpG sites with a nominal P value of < 0.001.

	Hypermethylated	Hypomethylated	Total
Vehicle ethanol	4715	3026	7741
CBD	6584	5564	12148
THC	1760	2009	3769

With the less stringent threshold, more overlap is observed between the different treatment groups (Figure 4.5c). Vehicle ethanol treatment shares a total of N= 34 CpG sites with both the CBD and THC exposure groups, N= 700 sites just with CBD exposure and N= 102 with just THC exposure.

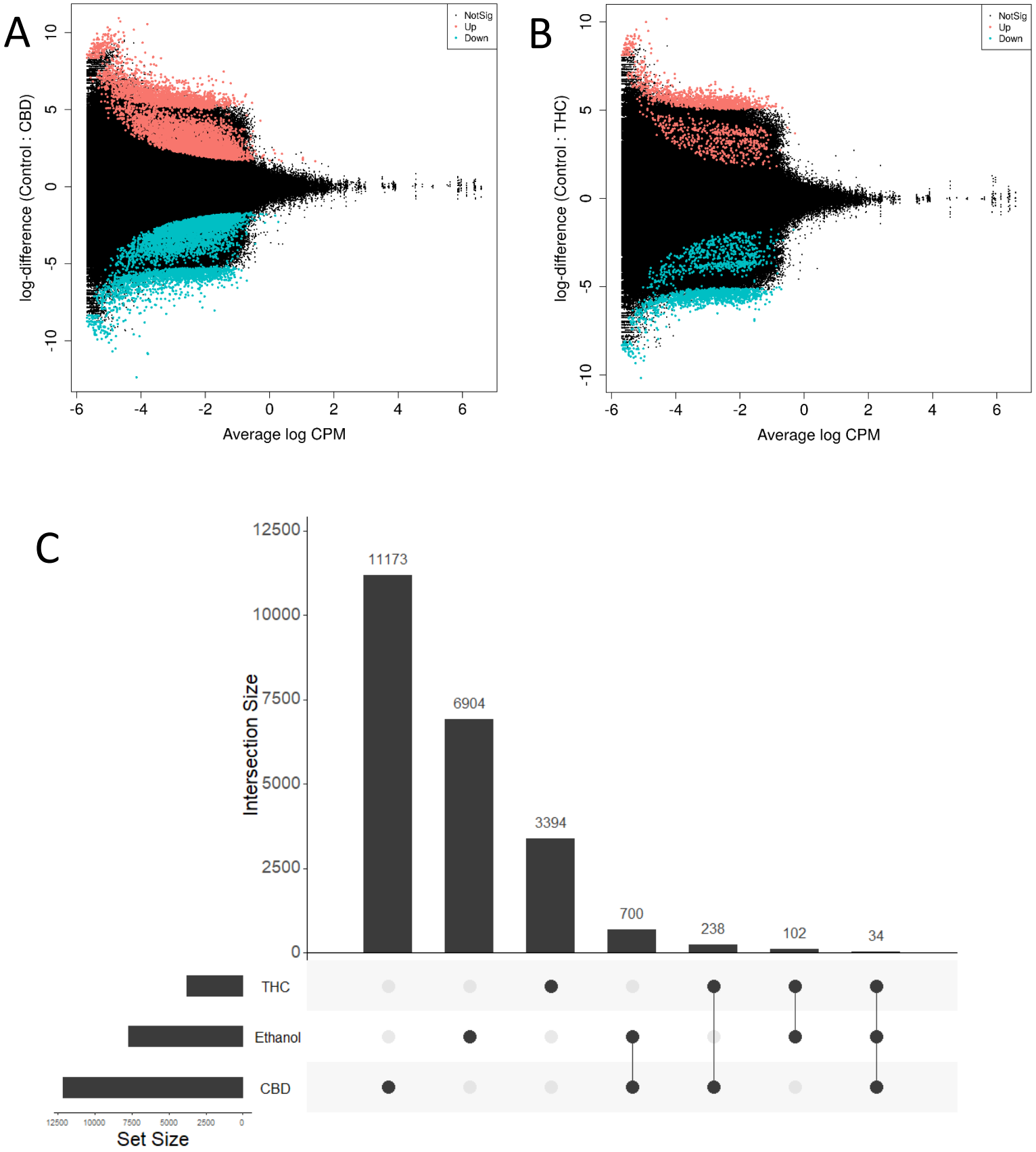


Figure 4.5 Top nominal (P < 0.001) CpG sites found to be differentially methylated in response to CBD (A) and THC (B) exposure. (C) - the overlap shared between the top sites with the vehicle ethanol group and within exposure groups.

4.4.5. Top 50 differentially methylated CpG sites in response to CBD treatment

Top tables were constructed to display the top 50 differentially methylated CpG sites between CBD treatment and control. Within the top 50 CpG sites, N= 26 CpG sites overlapped with CpG sites from the top 50 differentially methylated sites identified in the vehicle ethanol (nominal P < 0.001, Supplementary Table 4.1). These shared sites were disregarded from further analysis.

Table 4.6 The top 50 most significantly differentially methylated CpG sites in response to CBD treatment, compared to the untreated control and independent of vehicle ethanol control. The chromosome and location of the CpG site, and the CpG site's nearest gene, is included. Log FC- Log Fold Change and FDR- False Discovery Rate.

Chromosome	Location	Gene	*Log FC*	*P value*	*FDR*
NW_003040930.2	148543		-4.825	1.31E-15	4.57E-09
chr25	20353012	*kcna6*	-10.866	2.02E-14	2.71E-08
chr24	41876972	*scospondin*	8.384	2.32E-14	2.71E-08
chr3	58258458	*socs3a*	-10.807	6.69E-14	4.68E-08
chr3	57436826	*rnf213a*	-6.355	8.04E-14	4.69E-08
chr25	20353011	*kcna6*	-7.281	4.04E-13	2.02E-07
chr3	29330751	*cacna1i*	7.935	1.20E-12	5.22E-07
chr14	12445473	*hdac3*	-4.538	2.62E-12	8.31E-07
chr3	11274009	*zgc:165627*	-4.627	1.27E-11	3.16E-06
chr14	12445474	*hdac3*	-4.464	2.70E-11	5.55E-06
chr3	58317921	*cyth1a*	-9.709	3.34E-11	6.26E-06
chr3	59293008	*zgc:171489*	-6.832	3.40E-11	6.26E-06
chr3	58317922	*cyth1a*	-6.161	6.54E-11	1.09E-05
chr16	1893913	*sim1a*	8.077	1.28E-10	1.91E-05
chr10	5486494	*auh*	-8.253	1.31E-10	1.91E-05
chr3	58828626		-5.965	5.16E-10	5.58E-05
chr3	58258599	*socs3a*	-8.814	5.78E-10	5.80E-05
chr22	25789526	*si:ch211-226h8.14*	-5.318	7.29E-10	6.62E-05
chr3	58354011	*cyth1a*	-4.850	7.68E-10	6.71E-05
chr10	44002663	*fgfr1b*	-3.836	8.91E-10	7.17E-05
chr3	58349704	*cyth1a*	-7.407	9.02E-10	7.17E-05
chr10	44002528	*fgfr1b*	-3.767	1.34E-09	9.90E-05
chr16	1893844	*sim1a*	7.401	1.39E-09	9.91E-05
chr25	18423006	*zgc:103499*	-7.461	1.46E-09	0.0001
chr10	44224476	*zcchc9*	-4.049	1.61E-09	0.0001
chr3	60569101	*zgc:194562*	-5.018	1.85E-09	0.0001
chr3	57905454	*fscn2a*	-4.616	1.89E-09	0.0001
chr25	23673051	*mob2a*	6.087	1.90E-09	0.0001
chr10	43819553	*npy8ar*	-4.109	2.18E-09	0.0001
chr13	3088535	*park2*	3.753	2.22E-09	0.0001

Chr6	30338752	*zgc:171930*	-4.373	2.43E-09	0.0001
NW_003336528.1	95691		6.135	2.80E-09	0.0001
chr25	23673065	*mob2a*	5.489	2.83E-09	0.0001
chr13	38459928	*si:ch211-69e5.1*	-8.036	2.84E-09	0.0001
chr6	9442520	*acp5a*	4.658	2.97E-09	0.0001
chr25	23673067	*mob2a*	6.021	3.02E-09	0.0001
chr6	3841545	*slc25a12*	8.336	3.11E-09	0.0001
chr7	13641206	*zgc:158785*	6.209	3.20E-09	0.0001
chr10	37545264	*or104-2*	-3.442	3.26E-09	0.0001
chr1	37987032	*si:ch211-15e22.3*	-4.672	3.46E-09	0.0001
chr1	37987235	*si:ch211-15e22.3*	-4.672	3.46E-09	0.0001
chr5	1910297	*rcl1*	-4.673	3.63E-09	0.0006
chr3	57905609	*fscn2a*	-4.546	3.67E-09	0.0001
chr1	35174499	*gab1*	-4.157	3.71E-09	0.0001
chr23	2716155	*ncoa6*	3.623	3.97E-09	0.0001
chr3	25397589	*ddx5*	4.480	3.99E-09	0.0001
chr25	17774819	*e2f4*	-4.805	4.35E-09	0.0001
chr3	59679574	*luc7l3*	-4.440	4.97E-09	0.0002
chr12	9623582	*reep3*	4.126	5.24E-09	0.0002

4.4.6 Top 50 differentially methylated CpG sites found in response to THC treatment

Top tables were constructed to display the top 50 differentially methylated CpG sites between THC treatment and control. Within the top 50 CpG sites, a total of five CpG sites overlapped with vehicle ethanol (nominal $P < 0.001$, Supplementary Table 4.2). Table 4.7 lists the top sites which are independent of the sites found to be differentially methylated in response to the vehicle ethanol control.

Within the top 50 CpG sites displaying the most significant differential methylation between THC treatment and control (independent of ethanol), one gene reoccurs throughout the list. *si:dkey-85h7.1* is a zebrafish-specific gene which has no known human orthologue, and 12 CpG sites within this gene are differentially methylated in response to THC treatment (both with FDR adjusted and nominal significance). Another gene, *nlgn2a* contained the two other CpG sites found to have FDR adjusted significant P values. The top site in *nlgn2a*, is the one overlapping site shared between THC and CBD exposure after FDR correction.

Table 4.7 The top 50 most significantly differentially methylated CpG sites in response to THC treatment, compared to the untreated control and independent of vehicle ethanol control. The chromosome and location of the CpG site, and the CpG site's nearest gene, is included. Log FC- Log Fold Change and FDR- False Discovery Rate

Chromosome	Location	Gene	*Log FC*	*P value*	*FDR*
chr7	23338868	*nlgn2a*	7.045	2.01E-10	0.0007
chr6	20523802	*si:dkeyp-85h7.1*	-6.934	4.48E-08	0.034
chr6	20523860	*si:dkeyp-85h7.1*	-6.900	5.46E-08	0.034
chr6	20523902	*si:dkeyp-85h7.1*	-6.886	6.28E-08	0.034
chr6	20523745	*si:dkeyp-85h7.1*	-6.870	6.60E-08	0.034
chr7	23338729	*nlgn2a*	5.875	7.51E-08	0.034
chr6	20523883	*si:dkeyp-85h7.1*	-6.850	7.74E-08	0.034
chr6	20523755	*si:dkeyp-85h7.1*	-6.836	8.84E-08	0.034
chr6	20523757	*si:dkeyp-85h7.1*	-6.836	8.84E-08	0.034
chr17	9972703	*mgaa*	6.603	5.99E-07	0.209
chr6	20523743	*si:dkeyp-85h7.1*	-4.727	1.15E-06	0.364
chr17	9972789	*mgaa*	6.468	1.33E-06	0.386
chr7	51252313	*slc1a2a*	-10.171	1.55E-06	0.416
chr11	27554596	*map1lc3a*	7.117	1.96E-06	0.489
chr6	20523808	*si:dkeyp-85h7.1*	-4.648	2.11E-06	0.491
chr3	24452781	*prr15la*	-6.884	3.79E-06	0.767
chr3	24452797	*prr15la*	-6.884	3.79E-06	0.766
chr6	18869158	*zgc:174863*	-6.296	3.95E-06	0.766
chr6	19761188	*ppp1r27*	-6.510	4.22E-06	0.776
chr6	28542737	*tp63*	6.300	5.45E-06	0.787
chr14	28904090	*tsc22d3*	3.210	6.18E-06	0.787
chr23	23298992	*samd11*	-6.265	6.53E-06	0.787
chr4	29657600	*fnta*	6.366	7.07E-06	0.787
chr23	8906861	*sox18*	-7.962	7.12E-06	0.787
chr6	20788733	*znf644b*	-6.287	7.47E-06	0.787
chr7	23126983	*dock11*	6.246	8.19E-06	0.787
chr14	38793741	*atrx*	-6.228	8.20E-06	0.787
NW_001877452.3	719613		7.733	8.82E-06	0.787
chr7	20565248	*dock11*	-4.424	8.91E-06	0.787
chr15	43987300	*csf1ra*	-6.417	9.33E-06	0.787
chr6	20523868	*si:dkeyp-85h7.1*	-3.862	1.00E-05	0.787
chr20	27714926	*zbtb25*	8.488	1.01E-05	0.787
chr14	27510990	*smad5*	4.787	1.03E-05	0.787
chr18	41767582	*pvrl1b*	-6.149	1.12E-05	0.787
chr5	71262591	*zgc:175280*	6.078	1.21E-05	0.787
chr24	37667946	*pak1ip1*	-8.115	1.22E-05	0.787
chr14	7957224	*zgc:110843*	6.144	1.23E-05	0.787
chr18	26054544	*si:ch211-234p18.3*	6.429	1.24E-05	0.787

chr17	11610599	*efcab2*	5.717	1.25E-05	0.787
chr6	20523741	*si:dkeyp-85h7.1*	-3.824	1.35E-05	0.787
chr5	9454165	*atp5ib*	6.560	1.36E-05	0.787
chr6	20523722	*si:dkeyp-85h7.1*	-3.505	1.46E-05	0.787
chr9	57746726	*arsh*	6.802	1.46E-05	0.787
chr10	16251516	*slc12a2*	-6.225	1.49E-05	0.787
chr15	45699086	*igsf11*	-6.574	1.50E-05	0.787
chr23	24783311	*sult1st5*	-6.301	1.55E-05	0.787
chr23	4600674	*nup210*	6.142	1.62E-05	0.787
chr11	39055961	*etnk2*	6.119	1.71E-05	0.787

4.4.7 Pathway analysis for CBD CpG sites in genes

The less stringent P value cut off for significance ($P < 0.001$) was used to compile a list of the most significantly differentially methylated CpG sites within genes for biological pathway analysis. A total of 12,148 CpG sites had a $P < 0.001$, and within this, 11,745 CpG sites could be associated with a named zebrafish gene. The gene list was submitted to FishEnrichr to calculate pathway enrichment. Two different tables of results are displayed below: GO Molecular Function (Table 4.8) and GO Biological Process (Table 4.9).

A total of 13 molecular functions were nominally enriched in response to CBD treatment, and those that remained significant after BH adjustment were involved in membrane transport and the cellular response to stress. Seven biological functional pathways remained significantly enriched in response to CBD after BH correction, and include pathways involved in the maturation of sensory organs and those involved in the negative regulation of cell communication/signalling and the cellular response to stimuli.

Table 4.8 Molecular Function pathway analysis of the genes or nearest genes which house nominally significantly differentially methylated (P< 0.001) CpG sites in response to CBD exposure. P values were adjusted using Benjamini Hochberg.

Name	*P value*	*Adjusted P value*	*Z-score*	Combined score
transmembrane receptor protein tyrosine kinase activity (GO:0004714)	5.1E-06	0.003	-1.37	16.67
transmembrane receptor protein kinase activity (GO:0019199)	4.07E-05	0.008	-1.52	15.36
MAP kinase activity (GO:0004709)	2.43E-05	0.008	-1.30	13.78
mitogen-activated protein kinase binding (GO:0031434)	4.21E-05	0.008	-1.20	12.10
protein tyrosine kinase activity (GO:0004713)	7.46E-05	0.01	-1.19	11.29
acyl-CoA dehydrogenase activity (GO:0003995)	0.001	0.14	-4.40	29.72
transforming growth factor beta receptor binding (GO:0005160)	0.008	0.89	-1.54	7.39
heparan sulfate 6-O-sulfotransferase activity (GO:0017095)	0.034	1	-5.55	18.70
phosphatidylcholine transporter activity (GO:0008525)	0.024	1	-3.80	14.10
alpha2-adrenergic receptor activity (GO:0004938)	0.034	1	-4.14	13.93
NAD binding (GO:0051287)	0.030	1	-3.15	11.00
bioactive lipid receptor activity (GO:0045125)	0.044	1	-3.25	10.14

Table 4.9 Biological Process pathway analysis of the genes or nearest genes which house nominally significantly differentially methylated (P< 0.001) CpG sites in response to CBD exposure. P values were adjusted using Benjamini Hochberg.

Name	*P value*	*Adjusted P value*	*Z-score*	Combined score
sensory organ development (GO:0007423)	1.51E-06	0.003	-1.02	13.67
negative regulation of cell communication (GO:0010648)	2.97E-05	0.010	-2.00	20.81
negative regulation of signalling (GO:0023057)	2.97E-05	0.010	-1.99	20.77
negative regulation of response to stimulus (GO:0048585)	2.97E-05	0.010	-1.57	16.34
axon guidance (GO:0007411)	1.81E-05	0.010	-1.24	13.60
positive regulation of transferase activity (GO:0051347)	4.4E-05	0.015	-1.95	19.61
regulation of kinase activity (GO:0043549)	0.0001	0.033	-2.12	18.66
positive regulation of phosphorylation (GO:0042327)	0.0003	0.064	-1.64	13.03
central nervous system projection neuron axonogenesis (GO:0021952)	0.001	0.161	-2.64	17.50
semicircular canal development (GO:0060872)	0.001	0.188	-2.01	12.88
nucleus localization (GO:0051647)	0.002	0.229	-2.97	18.00
posterior lateral line neuromast hair cell development (GO:0035677)	0.003	0.280	-4.02	22.50
pattern specification involved in pronephros development (GO:0039017)	0.003	0.280	-3.29	18.42
posterior lateral line neuromast hair cell differentiation (GO:0048923)	0.004	0.313	-2.83	15.12
rRNA modification (GO:0000154)	0.006	0.385	-3.30	16.62
regulation of insulin secretion (GO:0050796)	0.006	0.385	-3.06	15.39
nuclear migration (GO:0007097)	0.008	0.453	-2.96	14.01
anterior/posterior pattern specification involved in pronephros development (GO:0034672)	0.011	0.483	-3.52	15.69
anterior/posterior pattern specification involved in kidney development (GO:0072098)	0.011	0.483	-3.21	14.35
rhombomere boundary formation (GO:0021654)	0.012	0.492	-3.10	13.60
rhombomere 4 morphogenesis (GO:0021661)	0.034	0.741	-6.61	22.28
adenylate cyclase-inhibiting adrenergic receptor signalling pathway (GO:0071881)	0.034	0.741	-5.46	18.38
epithelial cell fate commitment (GO:0072148)	0.034	0.741	-4.10	13.82
regulation of neutrophil differentiation (GO:0045658)	0.034	0.741	-4.07	13.70
negative regulation of myeloid cell differentiation (GO:0045638)	0.034	0.741	-4.00	13.46

4.4.8 THC Pathway analysis

The P value cut off of P < 0.001 was also used to assess the CpG sites within genes for pathway analysis of THC treatment (N= 3769). Of this, a total of N= 3620 CpG sites resided or could be assigned to a nearest gene and was used for biological pathway analysis as per section 4.4.5.3.

A total of eight molecular function pathways were nominally enriched in response to THC treatment (Table 4.10). The top two pathways are also the same top two pathways found in response to CBD exposure. The most significantly enriched pathway, transmembrane receptor protein tyrosine kinase activity, was the only molecular function to remain significant after P value adjustment. The biological pathway results (Table 4.11) show a bias towards brain related activity; axon guidance (also found in response to CBD Table 4.9), retinal ganglion cell axon guidance and neuron projection fasciculation are all significantly enriched and remain so after P value adjustment.

Table 4.10 Molecular Function pathway analysis of the genes or nearest genes which house nominally significantly differentially methylated (P< 0.001) CpG sites in response to THC exposure. P values were adjusted using Benjamini Hochberg.

Name	*P value*	*Adjusted P value*	*Z-score*	Combined score
transmembrane receptor protein tyrosine kinase activity (GO:0004714)	1.16E-05	0.007	-1.37	15.54
transmembrane receptor protein kinase activity (GO:0019199)	0.0001	0.058	-1.53	13.17
sodium:phosphate symporter activity (GO:0005436)	0.005	0.489	-2.68	14.10
sodium-dependent phosphate transmembrane transporter activity (GO:0015321)	0.007	0.565	-3.72	18.40
phosphate ion transmembrane transporter activity (GO:0015114)	0.012	0.753	-3.45	15.02
BMP receptor binding (GO:0070700)	0.020	0.753	-2.87	11.13
double-stranded DNA exodeoxyribonuclease activity (GO:0008309)	0.032	0.957	-4.29	14.75
transmembrane receptor protein serine/threonine kinase binding (GO:0070696)	0.030	0.957	-3.01	10.46

Table 4.11 Biological Process pathway analysis of the genes or nearest genes which house nominally significantly differentially methylated (P< 0.001) CpG sites in response to THC exposure. P values were adjusted using Benjamini Hochberg.

Name	*P value*	*Adjusted P value*	*Z-score*	Combined score
axon guidance (GO:0007411)	3.39E-10	7.12E-07	-1.25	27.20
retinal ganglion cell axon guidance (GO:0031290)	1.2E-05	0.007	-1.93	21.92
neuron projection fasciculation (GO:0106030)	5.54E-05	0.016	-3.04	29.80
axonal fasciculation (GO:0007413)	0.0002	0.044	-2.81	23.68
embryonic skeletal joint development (GO:0072498)	0.001	0.098	-4.33	28.26
negative regulation of hemopoiesis (GO:1903707)	0.003	0.133	-5.21	29.60
negative regulation of cellular response to transforming growth factor beta stimulus (GO:1903845)	0.003	0.133	-3.62	20.41
phosphate ion homeostasis (GO:0055062)	0.005	0.164	-4.23	22.24
cellular phosphate ion homeostasis (GO:0030643)	0.007	0.187	-4.16	20.57
retinoic acid metabolic process (GO:0042573)	0.032	0.419	-6.54	22.48

4.5 Discussion

The zebrafish offers many advantages as a model organism, in particular their rapid development from embryo to larvae stage, and so this model was utilised to investigate the epigenetic effects of environmental exposures. Here we assessed the impact of exposure to the two main active ingredients of cannabis, THC and CBD, on DNA methylation in zebrafish. The data shows that CBD drives a greater degree of differential methylation in the zebrafish genome compared to THC, and its effects are more broadly distributed across molecular functions and biological processes. In contrast, the impact of THC exposure on the zebrafish genome is less widespread and differential methylation is more localised to biological processes that function in the brain. Thus, our results highlight a role for DNA methylation in the biological response to cannabis. While provisional, given that here we detect differential DNA methylation at CpG sites within or near genes that contribute to molecular functions and biological pathways that have relevance to the biological mode of action of each cannabinoid, our findings demonstrate the potential for the broad use and applicability of the zebrafish as a model for probing the genomic effects of cannabinoids, and would benefit from further exploration.

4.5.1 Concentration of cannabinoids

LC_{50} experiments are a tool to determine the delicate balance between a concentration that is biologically relevant, and one in which either there is no biological effect or one which leads to major mortality. Initial dose concentrations of cannabinoids required an extensive literature review of previous zebrafish and cannabinoid research (Table 4.1). Previous LC_{50} experiments provided a range for initial testing. Our range-finding experiments yielded similar observations to that which had been previously described [38, 45-47], and the LC_{50} for CBD was calculated to be four times lower than that of THC (0.15 mg/l and 0.60 mg/l). The concentration for the vehicle ethanol treatment group was calculated based on the highest concentration that was used for either CBD or THC – as THC was calculated as a concentration of 0.60 mg/l (dissolved in e3 buffer) the vehicle ethanol was also calculated to the same concentration (0.60 mg/l in e3 buffer). Thus ensuring that we could accurately account for any methylation

changes that might be confounded by the presence of ethanol in the THC and CBD products.

4.5.2 Hatching efficiency and survival analysis

Alteration to hatching times differed between each of the treatment groups compared to the control, such that each treatment displayed statistical significance (Table 4.2). CBD exposure (0.15 mg/l) was lead to a greater reduction in hatching efficiency compared to the control as determined via Kaplan-Meier survival analysis ($P = 1.90 \times 10^{-12}$).

Generally there is considerable variation in the hatching rates of zebrafish larvae under normal conditions; usually hatching takes place between 48 and 72 hpf [48, 49]. There are a range of different factors that can influence this, such as temperature and light cycles [50]. However, our exposure experiments were conducted at the same time, under constant conditions; all of the embryos were housed in the same incubator which was kept at a constant temperature of 28.5 °C for the duration of the experiment. Allowing us to minimise any hatching variation that could be attributed to environmental conditions or by experiment time.

The prompting of a zebrafish embryo to hatch into a larvae requires the secretion of proteolytic enzymes to soften the outer shell of the embryo known as the chorion, this then allows the larvae's movements to break it open [51]. It has been suggested that alterations in developmental pathways are responsible for the disruption of this process [51]. Previous research has associated a range of chemical exposures with a delay in hatching efficiency, for example, exposure to butyl benzyl phthalate (BBP) [52], ionizing radiation [53], gamma radiation [54], tobacco condensate [51] and graphene oxide [55]. More so, findings from exposure to alcohol have suggested that late hatching larvae may model alcohol response later in life as a predisposition to alcohol tolerance and dependency [56]. Here, we demonstrate using the Kaplan Meier method (Table 4.2) that exposure to CBD ($P= 1.9E^{-12}$), THC ($P= 1.14E^{-05}$) and vehicle ethanol ($P= 5.73E^{-05}$) all resulted in a delay of hatching compared to the control. The delay was most pronounced in those embryos exposed to CBD, prompting us to hypothesise that the reduction in hatching efficiency detected here implies that THC

and CBD exposure may be altering developmental pathways in the zebrafish, and that CBD exposure may be having a more pronounced impact at the molecular level compared to any of the other exposure groups.

4.5.3 Overall differential DNA methylation found in the treatment groups

Each of the treatment groups were compared to the unexposed control group to assess for differential DNA methylation. Firstly, sites which reached an FDR cut off $P < 0.05$ were investigated (Table 4.4). The CBD treatment group showed the greatest amount of DNA methylation differences, with N= 1939 CpG sites identified as significant after FDR correction, followed by vehicle ethanol control (N= 662). The least amount of differential DNA methylation was seen in the THC treatment group, at N= 9 after FDR correction. Some cross over was seen between treatment groups, which is to be expected largely due to the inclusion of the vehicle ethanol group. CBD and vehicle ethanol shared N= 78 significantly differentially methylated CpG sites, and THC and CBD shared N= 1 CpG site. The single CpG site resides in the gene, Neuroligin (*NLGN2a)*, which will be discussed in detail below.

In order to investigate differential methylation in response to THC more fully, we increased the significance threshold to nominal $P < 0.001$ across all treatment groups. As expected, a greater level of differential DNA methylation was identified (Table 4.5). Again CBD had the greatest number of differentially methylated sites (N= 12148), followed by the vehicle ethanol group (N= 7741) and then by the THC treatment group (N= 3769). Accordingly, a greater level of overlap was seen at this less stringent significance level, with N= 700 CpG sites shared between CBD and vehicle ethanol (out of 12148 sites for CBD, or 5.8%), N= 102 CpG sites shared between THC and vehicle ethanol (2.7%). Thus, while there is a degree of overlap between the drug treatment groups and the ethanol control, the overlap is not so great as to impede further downstream analyses, and the effects of alcohol in the drug treatment groups were able to be taken into account in further analyses.

Again at this lower stringency, we identified N= 238 CpG sites that were shared between CBD and THC, and with a total of N= 34 CpG sites shared between all treatment groups (CBD, THC and vehicle ethanol). The results of this analysis implies

that the impact of CBD and THC on DNA methylation are dissimilar, and we hypothesise that the genes and pathways that house the THC- or CBD-specific CpG sites may highlight the precise biological pathways that are impacted by each cannabinoid.

4.5.4 Differentially methylated CpG sites in response to CBD exposure

To probe the biological relevance of the CBD-specific differential methylation, we further explored the differential CpG sites due to CBD exposure. Due to the larger number of CpG sites that remained significant after FDR correction (N= 1939) compared to THC, FDR-corrected data was used for generation of top tables and CpG sites with a P value < 0.001 were used for pathway analysis. In contrast with our approach to THC which used P values < 0.001 for both top tables and pathway analysis, which will be discussed in section 4.5.6. Significantly differentially methylated CpG sites showed an even distribution of hypomethylated and hypermethylated sites (Table 4.5), indicating that the response of DNA methylation to CBD is not biased towards hyper or hypomethylation. We then calculated the top 50 most significantly differentially methylated CpG sites using the FDR adjustment method (Table 4.6). The top 50 CpG sites identified in response to CBD exposure include a range of zebrafish-specific genes, as well as genes with human homolog. Specifically, we identified multiple significantly differentially methylated CpG sites within the top 50 (Table 4.6) in the genes Potassium Voltage-Gated Channel subfamily A Member 6 (*KCNA6*, associated with neurotransmitter release [57] and heart rate [58]), Cytohesin 1a (*CYTH1A*, immune defence [59]), MOB kinase activator 2a (*MOB2A,* neuron projection development [60]*)*, Histone Deacetylase 3 (*HDAC3,* white matter neurostructure in the brain [61, 62]) and consequences for behaviour [63] and Fibroblast growth factor receptor B (*FGFR1B,* associated with schizophrenia [64, 65]*).* Given that CBD is the non-psychoactive component of cannabis, the inclusion of many brain-related genes in the top 50 most significantly differentially methylated sites was unexpected and warrants further exploration.

Limited research has been conducted on assessing CBD exposure independently of THC and in non-disease systems. Often CBD exposure studies are in conjunction with

illness such as multiple sclerosis (MS) [66] and severe epilepsy [67, 68] and show promising health-related outcomes. However, we are still unsure of the full extent of the impact of CBD on the human body, and more so on DNA methylation, particularly in light of our findings above. Thus, while we have found differences in DNA methylation in response to CBD exposure, we are unable to comment on the clinical implications of these findings and we suggest that the impact of CBD on the brain should be explored more fully.

4.5.5 Pathway analysis of differentially methylated CpG sites in genes from CBD exposure

Differentially methylated CpG sites were annotated with their gene of residence or their nearest gene, and this list was then used for pathway analysis. Two types of pathway analysis was carried out, the first assessed Molecular Function which is described as the gene product ontologies (the role of the gene product), the second assessed Biological Process and is the based off the wider terminology of the process itself. From here we were able to assess sites showing differential methylation for gene pathways that may have been enriched due to CBD exposure (Table 4.8 and 4.9). Gene ontology (GO) molecular functions displayed three pathways specific to receptor protein tyrosine and kinase function (Table 4.8) both of which have broad roles in cell signalling. GO biological process pathway analysis (Table 4.9) showed a diverse enrichment of pathways with pathways relevant to sensory organ development, cell communication and axon guidance reaching significance (adjusted $P < 0.05$). The diversity of the molecular function and biological process pathways are indicative of the nature of the locations of CB2 receptors, which are found abundantly [69]. CBD is also often strongly associated with an effect on the immune system [19, 70], and although we identified the gene *CYTH1A* as differentially methylated, pathway analysis did not support this further.

4.5.6 Differentially methylated CpG sites in response to THC exposure

A total of nine CpG sites were found to be differentially methylated in response to THC exposure (FDR-corrected) compared to the unexposed control groups. Seven of the

sites displayed hypomethylation and two displayed hypermethylation. Within these nine CpG sites, seven of them are found to be within one region of the genome, with *SI:DKEYP-85H7.1-201* as the nearest gene. The gene is found in zebrafish and some birds species, however, no mammalian homologues are thought to exist. Due to this, there is limited information available about the functional implications of this gene, however it is predicted to be involved in signal transduction – it possesses a Rho GTP-ase-activating protein domain, and these domains have crucial roles in neuronal development and synaptic functions [71] which highlights the biological relevance of this gene. The remaining two CpG sites with FDR significance are both located within the gene *NLGN2a*, and one of these CpG sites was shared with the CBD exposure group. Differential methylation in *NLGN2a* has previously been identified in rodents, in a study assessing the cross-generational effects of THC exposure on offspring DNA methylation in the nucleus accumbens [72]. The gene presents an interesting finding as there has been a very recent surge in research assessing the association between paternal and maternal cannabis use and the development of autism in exposed offspring [73-78]. In both humans and mice, *NLGN2* variants have been associated with autism, intellectual disabilities, behavioural disorders and schizophrenia [79-82]. Implying that while our data show few differences in response to THC at an FDR-corrected level, identification of differential methylation at *NLGN2* is biologically relevant. It further suggests that differential methylation at this gene, in response to THC, is conserved across species, highlighting the value of the zebrafish as a model for human cannabis exposure.

In order to probe the impact of THC on DNA methylation more fully, we extended the significance threshold ($P < 0.001$) and identified N= 3769 CpG sites as differentially methylated. Of these, N= 2009 were hypomethylated and N= 1760 hypermethylated. Within the top most differentially methylated 50 CpG sites, five further CpG sites within *si:dkeyp-85h7.1-201* are observed. We suggest that this gene is targeted for further investigation in zebrafish and work should focus on identification of a human homologue as it may be important in the human response to THC.

4.5.7 Pathway analysis of differentially methylated CpG sites in genes from THC exposure

Similarly to the findings of CBD pathway analysis, molecular function pathway analysis of the genes which house nominally significant CpG sites in response to THC revealed that protein kinase activity was enriched (Table 4.10). Biological process enrichment analysis displayed enrichment for brain-related functions, for example, axon guidance, retinal ganglion cell axon guidance, and neuron projection fasciculation (Table 4.11), all of which reached an adjusted P value significance level.

4.5.8 How do these data relate to cannabis use in humans?

Experiments here were undertaken with pure THC and CBD, independently of each other. Although both THC and CBD are the most abundant cannabinoids in cannabis, they are still only two of approximately 100 potential cannabinoids that constitute cannabis. Importantly, CBD and THC have synergistic properties, for example, they have been described to be more effective in reducing symptoms of MS in combination rather than as independent chemicals [83-85]. Carrying out this same research with an additional CBD:THC treatment group is needed to further understand the genomic impact of these cannabinoids.

Contrary to the literature around THC, there is limited evidence to suggest that CBD is associated with detrimental health outcomes in humans. However, animal studies have previously reported unfavourable effects in response to CBD [86] such as a decrease in BDNF expression [87], decrease in circulating testosterone [88], reduced fertility [89, 90], hypertension and cardiac arrest [91]. Although we are unable to determine whether the DNA methylation differences we identify here in response to CBD are having a positive or negative phenotypic impact, they are associated with a decrease in hatching efficiency, compared to the control group. Meaning that given the increasing popularity of CBD as a therapeutic substance, the impact of CBD on the human genome needs to be explored more fully, particularly with regards to: i) the effect of CBD on neurodevelopment and neurotransmission, and; ii) developmental exposure, when the genome is more sensitive to environmental perturbation.

4.5.10 Limitations and considerations

Although in this study we identify differentially methylated CpG sites that reach genome-wide significance, we consider sample size to be a limitation, as this results in low statistical power. Increasing our sample size, as well as the number of replicates in each treatment group, would provide more robust evidence to support the biologically relevant results presented here.

Secondly, due to the constraints of working with controlled drugs and prescription medicines, licences and authorities for possession and use are required from the Ministry of Health, which is time consuming. We suggest that next steps in this research would be to validate differential methylation of CpG sites using a targeted approach (e.g. bisulfite-based amplicon sequencing or Sequenom MassARRAY EpiTYPER analysis) and probed for functional significance using complementary quantitative PCR analysis to interrogate gene expression changes in response to differential methylation. Thus would serve to validate our results and provide functional support for the role of methylation in the biological response to THC and CBD.

Lastly, although this study has highlighted the value of the zebrafish as a model for human THC and CBD exposure, there are still limitations to consider. Specifically, the THC and CBD used here have been solubilised in ethanol, and then given to the zebrafish via its environment. However, this is not the same mode of consumption as human cannabis use: the cannabis plant is composed of many different cannabinoids, and further, the combustion process which is involved in human cannabis consumption changes the molecular composition of the substance which is inhaled, and this cannot be replicated in the zebrafish model. However, while the precise effects of combustion of “smoked cannabis” cannot be replicated, our experimental design is more indicative of edible cannabis-based products, which are becoming increasingly available. Furthermore, we know that ethanol can also induce DNA methylation changes and can be associated with phenotype differences in humans, for example Fetal Alcohol Syndrome Disorder (FASD). Thus, it is important to further investigate what role ethanol may be causing in our result.

4.6 Chapter Summary

- Zebrafish embryos were exposed to two cannabinoids, THC and CBD, as well as a vehicle ethanol control, at 24 hpf.
- Hatching time discrepancies were observed in all treatment groups compared to the unexposed controls, with CBD displaying the greatest difference.
- Differential DNA methylation was identified via RRBS with genome-wide (FDR-corrected) significant differential DNA methylation observed in all treatment groups.
- The greatest number of differentially methylated CpG sites was seen in CBD (N= 1939), followed by vehicle ethanol (N= 662), and THC (N= 9).
- GO pathway analysis of CBD exposure showed enrichment for a diverse range of functional pathways, including cell communication and signalling.
- In response to CBD, multiple differentially methylated CpG sites were identified in genes that have roles in neurodevelopment, neurotransmission, behaviour and schizophrenia.
- GO pathway analysis for THC exposure was enriched for axon guidance, retinal ganglion and neuron projection fasciculation.
- Twelve differentially methylated CpG sites were identified in the zebrafish-specific gene *si:dkeyp-85h7.1-201,* which has predicted roles in neuronal development and synaptic function.
- We demonstrate that the zebrafish shows promise and value as a model in which to probe the genomic impacts of human cannabinoid exposure.

4.7 References

1. Council, N.R., *Cells and Surveys: Should Biological Measures Be Included in Social Science Research?*, ed. C.E. Finch, J.W. Vaupel, and K. Kinsella. 2001, Washington, DC: The National Academies Press. 388.
2. Ingham, P.W., *Zebrafish Genetics and Its Implications for Understanding Vertebrate Development.* Human Molecular Genetics, 1997. **6**(10): p. 1755-1760.
3. Veldman, M.B. and S. Lin, *Zebrafish as a Developmental Model Organism for Pediatric Research.* Pediatric Research, 2008. **64**(5): p. 470-476.
4. Norton, W. and L. Bally-Cuif, *Adult zebrafish as a model organism for behavioural genetics.* BMC Neuroscience, 2010. **11**(1): p. 90.
5. Kily, L.J.M., et al., *Gene expression changes in a zebrafish model of drug dependency suggest conservation of neuro-adaptation pathways.* Journal of Experimental Biology, 2008. **211**(10): p. 1623-1634.
6. Lau, B., et al., *Dissociation of food and opiate preference by a genetic mutation in zebrafish.* Genes, Brain and Behavior, 2006. **5**(7): p. 497-505.
7. Kamstra, J.H., et al., *Zebrafish as a model to study the role of DNA methylation in environmental toxicology.* Environ Sci Pollut Res Int, 2015. **22**(21): p. 16262-76.
8. Bambino, K. and J. Chu, *Chapter Nine - Zebrafish in Toxicology and Environmental Health*, in *Current Topics in Developmental Biology*, K.C. Sadler, Editor. 2017, Academic Press. p. 331-367.
9. Hill, A.J., et al., *Zebrafish as a Model Vertebrate for Investigating Chemical Toxicity.* Toxicological Sciences, 2005. **86**(1): p. 6-19.
10. Hallauer, J., et al., *The Effect of Chronic Arsenic Exposure in Zebrafish.* Zebrafish, 2016. **13**(5): p. 405-412.
11. Tse, W.K.F., et al., *Early embryogenesis in zebrafish is affected by bisphenol A exposure.* Biology Open, 2013. **2**(5): p. 466-471.
12. Knecht, A.L., et al., *Developmental benzo[a]pyrene (B[a]P) exposure impacts larval behavior and impairs adult learning in zebrafish.* Neurotoxicology and teratology, 2017. **59**: p. 27-34.
13. Howe, D.G., et al., *ZFIN, the Zebrafish Model Organism Database: increased support for mutants and transgenics.* Nucleic acids research, 2012. **41**(D1): p. D854-D860.
14. Kamstra, J.H., et al., *Zebrafish as a model to study the role of DNA methylation in environmental toxicology.* Environmental Science and Pollution Research, 2015. **22**(21): p. 16262-16276.
15. Akhtar, M.T., et al., *Developmental effects of cannabinoids on zebrafish larvae.* Zebrafish, 2013. **10**(3): p. 283-293.
16. Akhtar, M.T., et al., *Metabolic effects of cannabinoids in zebrafish (Danio rerio) embryos determined by 1H NMR metabolomics.* Metabolomics, 2016. **12**(3): p. 1-11.
17. Krug, R.G. and K.J. Clark, *Elucidating cannabinoid biology in zebrafish (Danio rerio).* Gene, 2015. **570**(2): p. 168-179.
18. Carty, D.R., et al., *Developmental Effects of Cannabidiol and Δ9-Tetrahydrocannabinol in Zebrafish.* Toxicological Sciences, 2018. **162**(1): p. 137-145.
19. Jensen, H.M., et al., *Cannabidiol effects on behaviour and immune gene expression in zebrafish (Danio rerio).* PLoS One, 2018. **13**(7): p. e0200016.
20. Achenbach, J.C., et al., *Analysis of the Uptake, Metabolism, and Behavioral Effects of Cannabinoids on Zebrafish Larvae.* Zebrafish, 2018. **15**(4): p. 349-360.
21. Voelker, D., et al., *Differential gene expression as a toxicant-sensitive endpoint in zebrafish embryos and larvae.* Aquatic toxicology, 2007. **81**(4): p. 355-364.
22. Schaaf, M.J., et al., *Discovery of a functional glucocorticoid receptor β-isoform in zebrafish.* Endocrinology, 2008. **149**(4): p. 1591-1599.
23. Aluru, N., *Epigenetic effects of environmental chemicals: insights from zebrafish.* Current Opinion in Toxicology, 2017.
24. Munro, S., K.L. Thomas, and M. Abu-Shaar, *Molecular characterization of a peripheral receptor for cannabinoids.* Nature, 1993. **365**(6441): p. 61-65.
25. Zou, S. and U. Kumar, *Cannabinoid Receptors and the Endocannabinoid System: Signaling and Function in the Central Nervous System.* International journal of molecular sciences, 2018. **19**(3): p. 833.
26. Marsicano, G. and B. Lutz, *Expression of the cannabinoid receptor CB1 in distinct neuronal subpopulations in the adult mouse forebrain.* European Journal of Neuroscience, 1999. **11**(12): p. 4213-4225.

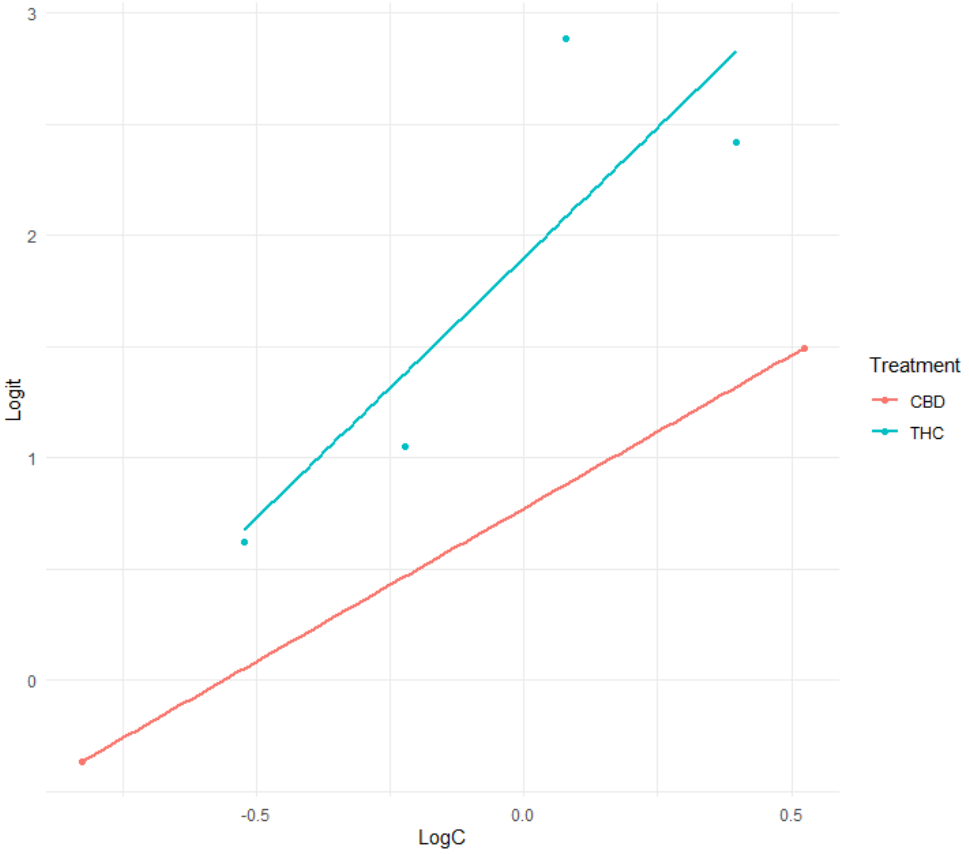
27. Katona, I., et al., *Presynaptically Located CB1 Cannabinoid Receptors Regulate GABA Release from Axon Terminals of Specific Hippocampal Interneurons.* The Journal of Neuroscience, 1999. **19**(11): p. 4544-4558.
28. Herkenham, M., et al., *Characterization and localization of cannabinoid receptors in rat brain: a quantitative in vitro autoradiographic study.* The Journal of Neuroscience, 1991. **11**(2): p. 563-583.
29. Ligresti, A., L.D. Petrocellis, and V.D. Marzo, *From Phytocannabinoids to Cannabinoid Receptors and Endocannabinoids: Pleiotropic Physiological and Pathological Roles Through Complex Pharmacology.* Physiological Reviews, 2016. **96**(4): p. 1593-1659.
30. Manzanares, J., M. Julian, and A. Carrascosa, *Role of the cannabinoid system in pain control and therapeutic implications for the management of acute and chronic pain episodes.* Current neuropharmacology, 2006. **4**(3): p. 239-257.
31. Laprairie, R.B., et al., *Cannabidiol is a negative allosteric modulator of the cannabinoid CB1 receptor.* British journal of pharmacology, 2015. **172**(20): p. 4790-4805.
32. Turcotte, C., et al., *The CB(2) receptor and its role as a regulator of inflammation.* Cellular and molecular life sciences : CMLS, 2016. **73**(23): p. 4449-4470.
33. Bie, B., et al., *An overview of the cannabinoid type 2 receptor system and its therapeutic potential.* Current opinion in anaesthesiology, 2018. **31**(4): p. 407-414.
34. Soderstrom, K., E. Soliman, and R. Van Dross, *Cannabinoids Modulate Neuronal Activity and Cancer by CB1 and CB2 Receptor-Independent Mechanisms.* Frontiers in Pharmacology, 2017. **8**(720).
35. Turcotte, C., et al., *The CB2 receptor and its role as a regulator of inflammation.* Cellular and Molecular Life Sciences, 2016. **73**(23): p. 4449-4470.
36. Sufian, M.S., et al., *CB₁ and CB₂ receptors play differential roles in early zebrafish locomotor development.* The Journal of Experimental Biology, 2019. **222**(16): p. jeb206680.
37. Ellis, L., *Zebrafish as a High-Throughput In Vivo Model for Testing the Bioactivity of Cannabinoids*, in *Recent Advances in Cannabinoid Research*. 2018, IntechOpen.
38. Carty, D.R., et al., *Developmental Effects of Cannabidiol and Δ9-Tetrahydrocannabinol in Zebrafish.* Toxicological Sciences, 2017. **162**(1): p. 137-145.
39. Kaplan, E.L. and P. Meier, *Nonparametric Estimation from Incomplete Observations.* Journal of the American Statistical Association, 1958. **53**(282): p. 457-481.
40. Grambsch, T.M.T.a.P.M., *Modeling Survival Data: Extending the Cox Model}*. 2000, New York: Springer.
41. Lex, A., et al., *UpSet: Visualization of Intersecting Sets.* IEEE Trans Vis Comput Graph, 2014. **20**(12): p. 1983-92.
42. Hiller, M., et al., *Computational methods to detect conserved non-genic elements in phylogenetically isolated genomes: application to zebrafish.* Nucleic Acids Res, 2013. **41**(15): p. e151.
43. Chen, E.Y., et al., *Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool.* BMC Bioinformatics, 2013. **14**: p. 128.
44. Kuleshov, M.V., et al., *Enrichr: a comprehensive gene set enrichment analysis web server 2016 update.* Nucleic Acids Res, 2016. **44**(W1): p. W90-7.
45. Ahmed, K.T., et al., *Motor neuron development in zebrafish is altered by brief (5-hr) exposures to THC (∆(9)-tetrahydrocannabinol) or CBD (cannabidiol) during gastrulation.* Scientific reports, 2018. **8**(1): p. 10518-10518.
46. Amin, M.R., K.T. Ahmed, and D.W. Ali, *Early Exposure to THC Alters M-Cell Development in Zebrafish Embryos.* Biomedicines, 2020. **8**(1): p. 5.
47. Carty, D.R., et al., *Multigenerational consequences of early-life cannabinoid exposure in zebrafish.* Toxicology and Applied Pharmacology, 2019. **364**: p. 133-143.
48. Kimmel, C.B., et al., *Stages of embryonic development of the zebrafish.* Dev Dyn, 1995. **203**(3): p. 253-310.
49. Parichy, D.M., et al., *Normal table of postembryonic zebrafish development: staging by externally visible anatomy of the living fish.* Developmental dynamics : an official publication of the American Association of Anatomists, 2009. **238**(12): p. 2975-3015.
50. Villamizar, N., et al., *Impact of Daily Thermocycles on Hatching Rhythms, Larval Performance and Sex Differentiation of Zebrafish.* PLOS ONE, 2012. **7**(12): p. e52153.
51. Ellis, L.D., et al., *Use of the Zebrafish Larvae as a Model to Study Cigarette Smoke Condensate Toxicity.* PLOS ONE, 2014. **9**(12): p. e115305.

52. Sun, G. and K. Liu, *Developmental toxicity and cardiac effects of butyl benzyl phthalate in zebrafish embryos.* Aquatic Toxicology, 2017. **192**: p. 165-170.
53. Freeman, J.L., et al., *Embryonic ionizing radiation exposure results in expression alterations of genes associated with cardiovascular and neurological development, function, and disease and modified cardiovascular function in zebrafish.* Frontiers in Genetics, 2014. **5**(268).
54. Praveen Kumar, M.K., et al., *Effects of gamma radiation on the early developmental stages of Zebrafish (Danio rerio).* Ecotoxicology and Environmental Safety, 2017. **142**: p. 95-101.
55. Chen, Y., et al., *Specific nanotoxicity of graphene oxide during zebrafish embryogenesis.* Nanotoxicology, 2016. **10**(1): p. 42-52.
56. Leite-Ferreira, M.E., H. Araujo-Silva, and A.C. Luchiari, *Individual Differences in Hatching Time Predict Alcohol Response in Zebrafish.* Frontiers in Behavioral Neuroscience, 2019. **13**(166).
57. Wisniewska, M.B., et al., *Novel β-catenin target genes identified in thalamic neurons encode modulators of neuronal excitability.* BMC Genomics, 2012. **13**(1): p. 635.
58. Nerbonne Jeanne, M., et al., *Genetic Manipulation of Cardiac K+ Channel Function in Mice.* Circulation Research, 2001. **89**(11): p. 944-956.
59. Mohanan, V., et al., *C1orf106 is a colitis risk gene that regulates stability of epithelial adherens junctions.* Science, 2018: p. eaan0814.
60. Duhart, J.C. and L.A. Raftery, *Mob Family Proteins: Regulatory Partners in Hippo and Hippo-Like Intracellular Signaling Pathways.* Frontiers in Cell and Developmental Biology, 2020. **8**(161).
61. Kassis, H., et al., *Histone deacetylase expression in white matter oligodendrocytes after stroke.* Neurochem Int, 2014. **77**: p. 17-23.
62. Norwood, J., et al., *Histone deacetylase 3 is necessary for proper brain development.* The Journal of biological chemistry, 2014. **289**(50): p. 34569-34582.
63. Shang, A., S. Bylipudi, and K.M. Bieszczad, *Inhibition of histone deacetylase 3 via RGFP966 facilitates cortical plasticity underlying unusually accurate auditory associative cue memory for excitatory and inhibitory cue-reward associations.* Behav Brain Res, 2019. **356**: p. 453-469.
64. Terwisscha van Scheltinga, A.F., S.C. Bakker, and R.S. Kahn, *Fibroblast growth factors in schizophrenia.* Schizophrenia bulletin, 2010. **36**(6): p. 1157-1166.
65. Narla, S.T., et al., *Common developmental genome deprogramming in schizophrenia - Role of Integrative Nuclear FGFR1 Signaling (INFS).* Schizophrenia research, 2017. **185**: p. 17-32.
66. Mecha, M., et al., *Cannabidiol provides long-lasting protection against the deleterious effects of inflammation in a viral model of multiple sclerosis: a role for A2A receptors.* Neurobiology of disease, 2013. **59**: p. 141-150.
67. Tzadok, M., et al., *CBD-enriched medical cannabis for intractable pediatric epilepsy: The current Israeli experience.* Seizure, 2016. **35**: p. 41-44.
68. Szaflarski, J.P., et al., *Cannabidiol improves frequency and severity of seizures and reduces adverse events in an open-label add-on prospective study.* Epilepsy & Behavior, 2018. **87**: p. 131-136.
69. Atwood, B.K. and K. Mackie, *CB2: a cannabinoid receptor with an identity crisis.* British journal of pharmacology, 2010. **160**(3): p. 467-479.
70. Yang, X., et al., *Cannabidiol Regulates Gene Expression in Encephalitogenic T cells Using Histone Methylation and noncoding RNA during Experimental Autoimmune Encephalomyelitis.* Scientific Reports, 2019. **9**(1): p. 15780.
71. Moon, S.Y. and Y. Zheng, *Rho GTPase-activating proteins in cell regulation.* Trends Cell Biol, 2003. **13**(1): p. 13-22.
72. Watson, C.T., et al., *Genome-Wide DNA Methylation Profiling Reveals Epigenetic Changes in the Rat Nucleus Accumbens Associated With Cross-Generational Effects of Adolescent THC Exposure.* Neuropsychopharmacology : official publication of the American College of Neuropsychopharmacology, 2015. **40**(13): p. 2993-3005.
73. Schrott, R., et al., *Sperm DNA methylation altered by THC and nicotine: Vulnerability of neurodevelopmental genes with bivalent chromatin.* Scientific Reports, 2020. **10**(1): p. 16022.
74. Schrott, R., et al., *Cannabis use is associated with potentially heritable widespread changes in autism candidate gene DLGAP2 DNA methylation in sperm.* Epigenetics, 2020. **15**(1-2): p. 161-173.
75. Reece, A.S. and G.K. Hulse, *Impacts of cannabinoid epigenetics on human development: reflections on Murphy et. al. 'cannabinoid exposure and altered DNA methylation in rat and human sperm' epigenetics 2018; 13: 1208-1221.* Epigenetics, 2019. **14**(11): p. 1041-1056.
76. Holloway, Z.R., et al., *Paternal cannabis extract exposure in rats: Preconception timing effects on neurodevelopmental behavior in offspring.* NeuroToxicology, 2020. **81**: p. 180-188.

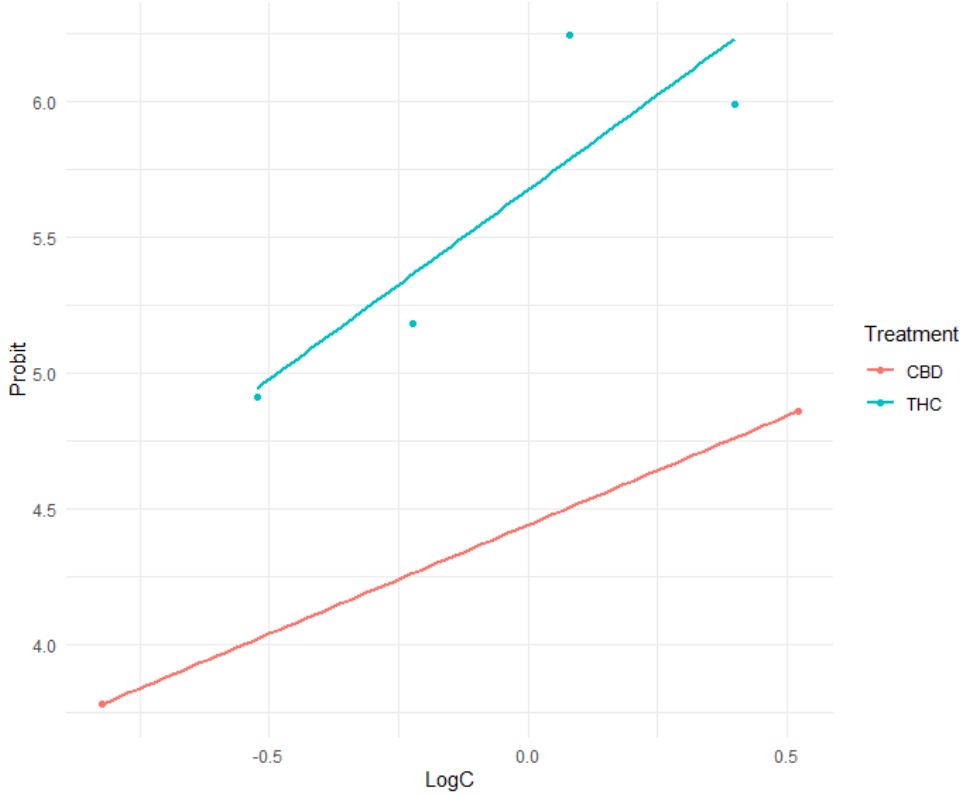
77. Slotkin, T.A., et al., *Paternal Δ9-Tetrahydrocannabinol Exposure Prior to Mating Elicits Deficits in Cholinergic Synaptic Function in the Offspring.* Toxicological Sciences, 2020. **174**(2): p. 210-217.
78. Corsi, D.J., et al., *Maternal cannabis use in pregnancy and child neurodevelopmental outcomes.* Nature Medicine, 2020. **26**(10): p. 1536-1540.
79. Parente, D.J., et al., *Neuroligin 2 nonsense variant associated with anxiety, autism, intellectual disability, hyperphagia, and obesity.* Am J Med Genet A, 2017. **173**(1): p. 213-216.
80. Wöhr, M., et al., *Developmental delays and reduced pup ultrasonic vocalizations but normal sociability in mice lacking the postsynaptic cell adhesion protein neuroligin2.* Behavioural Brain Research, 2013. **251**: p. 50-64.
81. Sun, C., et al., *Identification and functional characterization of rare mutations of the neuroligin-2 gene (NLGN2) associated with schizophrenia.* Human molecular genetics, 2011. **20**(15): p. 3042-3051.
82. Study, T.D.D.D., et al., *Large-scale discovery of novel genetic causes of developmental disorders.* Nature, 2015. **519**(7542): p. 223-228.
83. Langford, R.M., et al., *A double-blind, randomized, placebo-controlled, parallel-group study of THC/CBD oromucosal spray in combination with the existing treatment regimen, in the relief of central neuropathic pain in patients with multiple sclerosis.* Journal of Neurology, 2013. **260**(4): p. 984-997.
84. Sastre-Garriga, J., et al., *THC and CBD oromucosal spray (Sativex®) in the management of spasticity associated with multiple sclerosis.* Expert Review of Neurotherapeutics, 2011. **11**(5): p. 627-637.
85. Celius, E.G. and C. Vila, *The influence of THC: CBD oromucosal spray on driving ability in patients with multiple sclerosis-related spasticity.* Brain and behavior, 2018. **8**(5): p. e00962.
86. Huestis, M.A., et al., *Cannabidiol Adverse Effects and Toxicity.* Current neuropharmacology, 2019. **17**(10): p. 974-989.
87. ElBatsh, M.M., et al., *Anxiogenic-like effects of chronic cannabidiol administration in rats.* Psychopharmacology, 2012. **221**(2): p. 239-247.
88. Carvalho, R.K., et al., *Chronic exposure to cannabidiol induces reproductive toxicity in male Swiss mice.* Journal of Applied Toxicology, 2018. **38**(9): p. 1215-1223.
89. Schuel, H., et al., *Cannabinoids reduce fertility of sea urchin sperm.* Biochem Cell Biol, 1987. **65**(2): p. 130-6.
90. Rosenkrantz, H., R.W. Fleischman, and R.J. Grant, *Toxicity of short-term administration of cannabinoids to rhesus monkeys.* Toxicology and applied pharmacology, 1981. **58**(1): p. 118-131.
91. Garberg, H.T., et al., *High-Dose Cannabidiol Induced Hypotension after Global Hypoxia-Ischemia in Piglets.* Neonatology, 2017. **112**(2): p. 143-149.

4.7 Supplementary Figures and tables

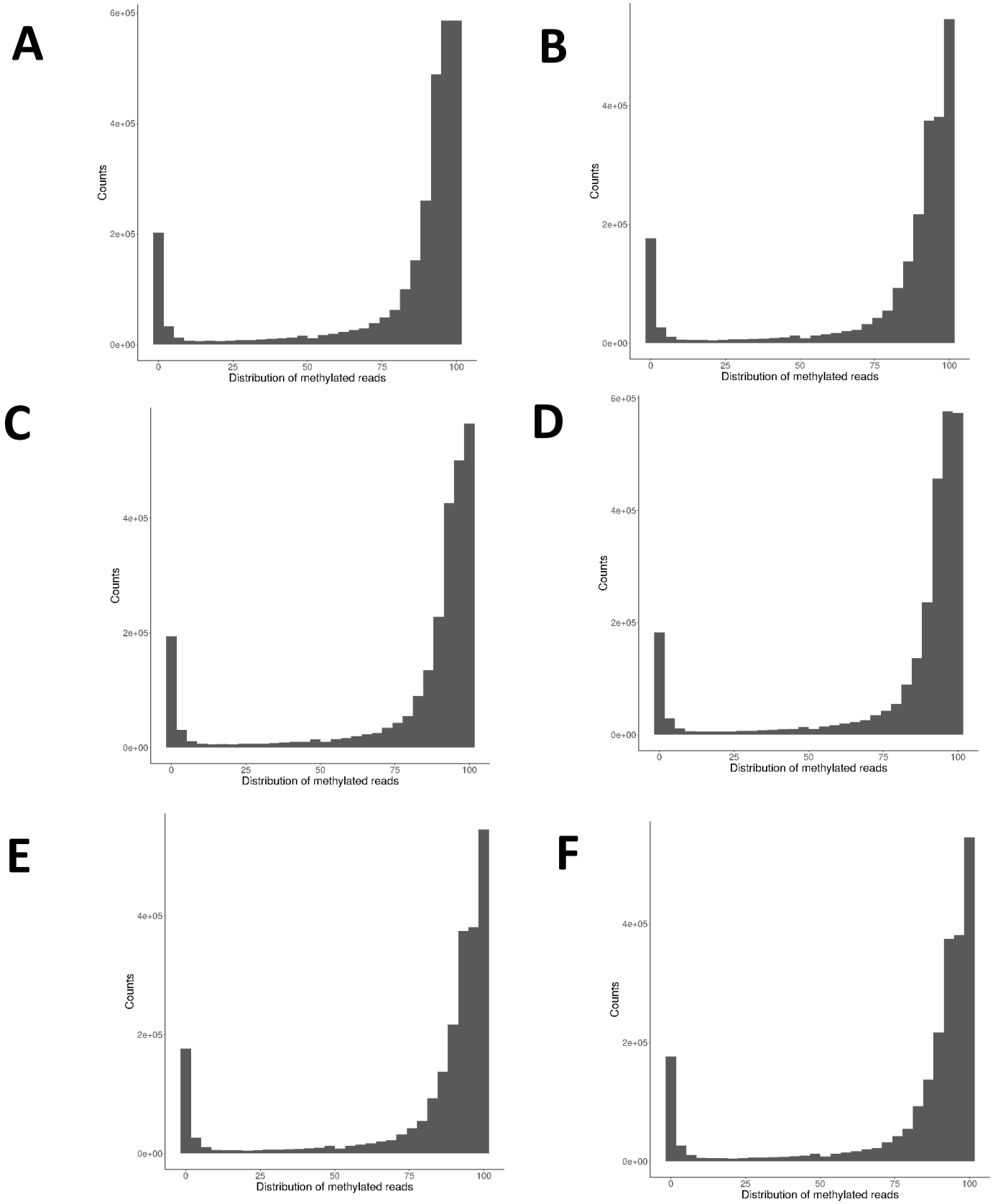
A



B



Supplementary Figure 4.1 Lethal concentration of THC and CBD experiment, the Log concentration plotted against Logit (A) and probit (B).



Supplementary 4.2 Frequency of the percentage of methylated reads for the remaining samples assessed for RRBS. A) Control 2, B) vehicle ethanol 1, C) Vehicle ethanol 2, D) THC 2, E) CBD 1, F) CBD 2.

Supplementary Table 4.1 CpG sites found to be differentially methylated in response to THC treatment that were also nominally significant (P value < 0.001) in response to the vehicle ethanol treatment.

Chromosome	Location	Gene	*LogFC*	*P value*	*FDR*
chr3	34907133	*cdk5r1a*	6.221	7.83E-06	0.787
chr14	10899857	*atrx*	6.192	9.29E-06	0.787
NW_003337037.1	71666		3.870	1.01E-05	0.787
chr1	34805751	*gab1*	-4.971	1.10E-05	0.787
chr6	21955808	*acox1*	6.088	1.12E-05	0.787
chr23	18588976	*hsd17b10*	-7.002	1.24E-05	0.787

Supplementary Table 4.2 CpG sites found to be differentially methylated in response to CBD treatment that were also nominally significant (P value < 0.001) in response to the vehicle ethanol treatment.

Chromosome	Location	Gene	*LogFC*	*P Value*	*FDR*
chr19	46554422	*ptpro*	-5.648	6.26E-14	4.68E-08
NW_003337101.1	6968		-7.284	1.70E-12	5.94E-07
NW_003337101.1	6927		7.284	1.70E-12	5.94E-07
NW_003040930.2	148542		-4.131	3.95E-12	1.15E-06
NW_003337101.1	6952		7.104	5.39E-12	1.45E-06
chr13	45886895	*si:ch211-168h21.3*	7.245	2.06E-11	4.75E-06
chr1	24511269	*plrg1*	-4.943	2.17E-11	4.75E-06
chr1	22503413	*slit2*	-4.361	6.25E-11	1.09E-05
chr19	43384306	*tinagl1*	-4.245	1.09E-10	1.74E-05
chr1	29281079	*tmem41ab*	-4.232	1.54E-10	2.15E-05
chr5	12159513	*zgc:112294*	4.455	1.65E-10	2.22E-05
chr5	1968493	*rcl1*	-5.642	2.40E-10	3.11E-05
chr19	43384367	*tinagl1*	4.159	2.81E-10	3.47E-05
chr5	7278569	*ostf1*	-5.516	2.88E-10	3.47E-05
NW_001877452.3	55072		4.872	3.09E-10	3.49E-05
chr5	6012759	*zgc:73226*	-4.444	3.10E-10	3.49E-05
chr18	25702611	*sema4ba*	-12.37	5.27E-10	5.58E-05
chr5	7278568	*ostf1*	-4.280	5.81E-10	5.80E-05
chr3	58674022	*stra13*	-6.146	6.20E-10	6.02E-05
chr1	31036330	*slc2a15b*	-4.299	6.47E-10	6.11E-05
chr5	5784586	*rabl6*	-4.343	7.38E-10	6.62E-05
chr14	8009994	*zgc:92242*	-4.540	8.05E-10	6.78E-05
chr1	22503410	*slit2*	-4.030	8.15E-10	6.78E-05
chr21	28252720	*cxxc5a*	-7.844	9.49E-10	7.37E-05
NW_001877452.3	7572		-5.083	1.08E-09	8.21E-05
chr5	3496936	*ywhag1*	5.239	1.36E-09	9.90E-05

Chapter 5

5. Epigenetic signatures associated with the observed interaction between maternal tobacco use during pregnancy, and offspring conduct problems in childhood and adolescence

5.1 Introduction

5.1.1 Maternal tobacco use during pregnancy

The use of tobacco during pregnancy is one of the leading causes of perinatal compromise for developing offspring, and one of the most preventable [1]. For example, low birth weight [2], congenital heart anomalies [3], asthma/respiratory illness [4, 5], and sudden infant death syndrome (SIDS)[6] are all associated with maternal tobacco use during pregnancy, the rate of which remains relatively high in New Zealand (18.4% [7]), despite declining tobacco use rates overall [8].

While immediate perinatal compromise in infants due to maternal smoking is well documented, the long term effects into later childhood, adolescence and adulthood are not understood. There is increasing evidence of linkages between maternal tobacco use in pregnancy and later risks of mental health and related adjustment problems in childhood and adolescence. In particular, there is evidence that maternal smoking during pregnancy is associated with increased risks of conduct disorders and antisocial behaviours in offspring [9] [10-12]. This association is not explained by post-natal environment [13]. Further associations have been identified between maternal tobacco use during pregnancy and the increased risk of cardiometabolic disease [14], and the development of attention-deficit hyperactivity disorder (ADHD) [15]. Also affected are offspring neurodevelopment and behaviour, suggesting that poor behavioural adjustment (often termed ‘conduct problems’, CP) can be considered a consequence of maternal smoking during pregnancy [9]. While these traits in themselves can be linked to other societal risk factors such as low socioeconomic status and early-life adversity [16], their association with maternal tobacco use during pregnancy is intriguing. Understanding the link between exposures such as tobacco use during pregnancy and the association with CP is crucial to further our understanding the paradigm of the developmental origins of human health and disease (DOHaD) [17].

5.1.2 Effect of prenatal tobacco exposure on DNA methylation

Recent research has demonstrated links between prenatal tobacco exposure and specific DNA methylation patterns of newborn offspring [18-21]. Tobacco-induced DNA methylation changes can persist into adolescence [22] [21, 23] with potential for these unexplained marks to be inherited by future generation of offspring of exposed individuals [24]. Further, meta-analyses of multiple CpG sites in the gene, *GFI1* (Growth Factor Independent one transcriptional repressor) were found to be differentially methylated in adult offspring in response to being exposed to tobacco *in utero*, at multiple sites within the gene [25]. However, these studies are limited in their scope - they provide evidence for differential DNA methylation induced in both children and adults by tobacco exposure *in utero*, but do not relate these DNA methylation changes to a phenotype that is associated with *in utero* tobacco exposure. Thus, while limited preliminary work has been carried out, in which three loci which indicated modest DNA methylation changes in response to maternal smoking during pregnancy and CP phenotypes [26], the etiology of this link has not been fully explored. One potential mechanism is that differential DNA methylation caused during the *in utero* time period is playing a role later in life of the affected offspring via the *in utero* generation of metastable epialleles (MEs). Evidence at this stage has largely come from animal studies, where *in utero* exposures cause the development of MEs [27-29]. Potentially these *in utero* exposures can generate permanent epigenetic changes to the genome [30] that may contribute to an individual's phenotype later in life [29-32]

5.1.3 Chapter scope, aims and hypotheses

Thus, given: i) the fact that maternal tobacco smoking during pregnancy is linked to offspring CP during early childhood and adolescence, and; ii) that maternal tobacco use during pregnancy can affect DNA methylation of offspring through to adolescence and adulthood, and; iii) that *in utero* exposures can create permanent epigenetic changes that can affect health in later life, here we hypothesise that DNA methylation is altered at genes involved in *in utero* brain development, and in those that associate with CP phenotypes, in the adult offspring of individuals who were exposed to tobacco *in utero*.

To test this hypothesis, we quantified DNA methylation at a suite of genes with known roles in *in utero* neurodevelopment and CP phenotypes, to assess whether DNA methylation may be implicated in the interaction between maternal tobacco use during pregnancy and the development of CP in offspring. We applied a targeted approach via bisulfite-based amplicon sequencing (BSAS) of each gene in our panel, to interrogate differential methylation in the DNA of participants from the Christchurch Health and Development Study (CHDS) whose mothers consumed tobacco during pregnancy.

5.2 Methods

5.2.1 Sample

A sub-group of individuals from the CHDS were selected for this study (Table 5.1). The longitudinal study originally included 97% of all the children (N = 1265) born in the Christchurch, New Zealand urban region during a three-month period in mid-1977 and has been studied at 24 time points from birth to age 40 (n = 987 at age 30). All participants were aged between 28-30 when blood samples for DNA were drawn.

For the subsets studied in this report, CHDS participants were chosen based on their *in utero* tobacco exposure status, their adult smoking status, and their CP scores (Table 5.1). Group 1 consisted of individuals who were exposed *in utero* to tobacco smoke, and never smokers at the time blood samples were taken (N= 32). Group 2 consisted of individuals who were exposed *in utero* to tobacco smoke and were themselves regular smokers at the time the blood was taken (N =32). Group 3 consisted of individuals who were not exposed to tobacco *in utero*, and never smokers at the time blood was taken (N =32). *In utero* tobacco exposure was defined as 10+ cigarettes per day throughout pregnancy. Within each group, 16 individuals were selected with a 'high' score on a measure of childhood CP at age 7-9 years and 16 with a 'low' score. Severity of childhood CP was assessed using an instrument that combined selected items from the Rutter and Conners child behaviour checklists [33-36] as completed by parents and teachers at annual intervals from 7-9 years. Parental and teacher reports were summed and averaged over the three years [37] to derive a robust scale measure of the extent to which the child exhibited conduct disordered/oppositional behaviours (mean (SD)=50.1(7.9) ; range 41-97). For the purposes of this report a 'high' score was defined as falling into the top quartile of the score distribution (scores> 53) and a 'low' score was defined as scores< 46.

A further control group consisting of non-exposed *in utero* who are adult smokers would have been beneficial for statistical analysis for this study. However, this group of individuals were unable to be sourced for this study.

Table 5.1 CHDS subsets selected for analysis *in utero* maternal tobacco exposure and the interaction of CP. The range of CP scores in each category is indicated in brackets. A score of 53 or more is the top quartile for CP, with a score of 60 or higher indicating the top decile for CP.

	Group 1 Exposed *in utero* and a never smoker	Group 2 Exposed *in utero* and a regular smoker	Group 3 Not *exposed in utero* and a never smoker
	N= 32	N= 32	N= 32
Sex			
Male	69%	72%	60%
Female	31%	28%	40%
Tobacco smoking status at the time of blood collection			
Never	100%	0%	100%
Occasional	0%	0%	0%
Regular	0%	100%	0%
Conduct problem Score (CPS)			
Below 46	N= 16 (42-46)	N= 16 (42-46)	N= 16 (41-43)
Above 53	N= 16 (53-75)	N= 16 (60-85)	N= 16 (53-68)

5.2.2 Bisulfite-based amplicon sequencing

Bisulfite-based amplicon sequencing (BSAS) and genome alignment was carried out as described in 3.3.1 [38].

Genes for sequencing (Table 5.2) were picked based upon several criteria: i) previously published differential DNA methylation in response to *in utero* tobacco smoking in human studies; ii) known associations with *in utero* brain development, and; iii) known associations with CP phenotypes.

Table 5.2 Genes selected to investigate the link between *in utero* tobacco exposure and CP.

Gene	Function	Significance
AHRR [43-47]	Mediates toxicity of dioxin (found in cigarette smoke)	Hypomethylated in tobacco smokers and their offspring
ASH2L [48]	Histone lysine methyltransferase	Associated with schizophrenia
BDNF [49, 50]	Nerve growth factor	Promotes neuronal survival. Implicated in neurodegenerative disease
CNTNAP2 [44, 51, 52]	Neurexin family – functions in vertebrate nervous system	Implicated in schizophrenia, autism, ADHD, intellectual disability. Hypomethylated in offspring of maternal smoking
CYP1A1 [43-47, 53]	Monooxygenase – expression is induced by hydrocarbons found in cigarette smoke	Hypomethylated in offspring of maternal smoking
DUSP6 [54]	Protein phosphatase, cellular proliferation and differentiation	Regulates neurotransmitter homeostasis
GFI1 [43, 46, 47]	Zinc finger protein - transcriptional repressor	Part of a complex that controls histone modifications and gene silencing. Hypermethylated in offspring of maternal smoking
GRIN2B [55]	Glutamate receptor – expressed early in the brain and is required for normal brain development	Mutations associated with autism, ADHD, schizophrenia
MEF2C [54]	MEF2C is associated with hippocampal-dependent learning and memory	MEF2C is crucial for normal neuronal development. Associated with ADHD
PRDM8 [51]	Histone methyltransferase - Controls expression of genes involved in neural development and neuronal differentiation	Hypomethylated in offspring of maternal smoking

Primers were then designed (Table 5.3) to flank the CpG sites of interest, ~350 base pairs (bp) in total, or to amplify ~350bp of the promoter region of the gene if a specific CpG site was not known. Multiple pairs of primers were designed to amplify larger regions.

Table 5.3 Forward and reverse primers (5' – 3') used to target potential candidates of *in utero* tobacco exposure and the interaction of CP. Primers for CpG sites of interest include the Illumina overhang sequence at the 5' end.

Primer name	Illumina Probe ID	Bisulfite converted primer (including the Illumina overhang sequence)
AHRR_F	Cg05575921	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGGTTTTTTTGGTGTGGTTTTA
AHRR_R	Cg05575921	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAG ACCACCATCTTATCTTATTT
CNTNAP2_F	Cg2594950	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGTTTGTTTTGGAGTAGTTTTA
CNTNAP2_R	Cg2594950	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGCATAATCTTCACTTTTCATTCAC
CYP1A1_F	Cg05549655	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGTATAGTAGTTGTTTGGTAAA
CYP1A1_R	Cg05549655	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGRATACAAAAAATCTAAATCTAC
GFI1_F	Cg09935388	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGGGGGGGAAGGAATGAGTAGAT
GFI1_R	Cg09935388	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGACTAAAACTAATAACCCCAA
GFI1_F	Cg09662411	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGTATAGTAGTTTYGATTTTATTTTGA
GFI1_R	Cg09662411	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGAACCCTTCCCCCTACCTTTC
DUSP6_F	Promoter region	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGAGTAAATAGAGTTGGGTTTT
DUSP6_R	Promoter region	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGTACAAACAAACTACAACAAC
BDNFpro1_F	Promoter region 1	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGAAAAGGGAAAGTTGTTGGGTT
BDNFpro1_R	Promoter region 1	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGCTAAAAAACTTATTACTTATC
BDNFpro2_F	Promoter region 2	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGTTTTTTATTTTTTTTTTGTT
BDNFpro2_R	Promoter region 2	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGATTTCCTAAAACTACCTTCTAAC
BDNFpro3_F	Promoter region 3	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGTTTTTATTTTTTTTGGGAAT
BDNFpro3_R	Promoter region 3	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGRTCTCCCCAACAAATACTAAA
PRDM8pro1_F	Promoter region 1	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGTGGTTGAAGTAGTTGTTTT
PRDM8pro1_R	Promoter region 1	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGAAATATATAAAAATCATAAC
PRDM8pro2_F	Promoter region 2	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGTATTTTTTTATATTATTTTTTTT
PRDM8pro2_R	Promoter region 2	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGAAAACTATAAAACTCCTTCC
MEF2Cpro1_F	Promoter region	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGATGGAAAGATTGATTTATTAAG
MEF2Cpro1_R	Promoter region	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGTTTTATCCTTACCTTTACTT
ASH2L_F	Promoter region	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGYGGGTAGGGAGTGTTAGATTTTA
ASH2L_R	Promoter region	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGCTAAAAAAACATAATTCCAC
SLC6A1pro2_F	Promoter region	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGATGTTTTAAGTGAATTTTATTG
SLC6A1pro2_R	Promoter region	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGRATCTTATTATTCCAAATAA
GRIN2Bpro2_F	Promoter region	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGCTGTGGGAAATGCGGGGTTT
GRIN2Bpro2_R	Promoter region	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGCAAAGGTAATTCAGGGTATG

5.2.3 Statistical analysis

Differential DNA methylation was assessed using the package edgeR [41]. MA plots were carried out for clustering based on group and for the top differentially methylated sites via edgeR. The following models were fitted to the data:

Univariate regression:

Model 1 - effect of in utero tobacco exposure on DNA methylation (Table 5.5 and Table 5.6)

$$Y \sim U + e$$

Model 2 - effect of CP on DNA methylation (Table 5.7)

$$Y \sim C + e$$

Model 3 - effect of adult smoking on DNA methylation (fitted on Exposed participants only, Table 5.8)

$$Y \sim AS + e$$

Multiple Regression:

Model 4 - effect of in utero tobacco exposure and CP on DNA methylation (Table 5.9)

$$Y \sim U + C + U{:}C + e$$

Where:

Y = methylation M ratio

U = Exposed/Unexposed *in utero* to maternal smoking

C = Conduct problem/Non-conduct problem

$e \sim N(0,s)$

AS = Adult smoking/Non-adult smoking

U:C is interaction term between U and C

Models 1,2 and 3 all assessed differential DNA methylation from the one variable of interest.

Model 4 took into account *in utero* exposure and CP score into the interaction between the two variables. It was fitted with both ANOVA parameters and with contrasts between *in utero* exposure groups (exposed – non-exposed) within CP score levels.

Top tables were constructed using the topTags function in edgeR, Log fold change, average log counts per million, and in some cases F statistic and were calculated and nominal significance was given for $P < 0.05$, these were then corrected using FDR. The F value takes into account the standard error for each of the data sets being assessed. Co-variates such as ethnicity and sex were not corrected for. Box plots were constructed from log transformed normalized methylated and unmethylated counts. A statistical package called Predict Means [42] was used to assess the overall methylation differentiation between the various interactions as a whole data set. Similarly, general linear model with a binomial distribution was used for this analysis and a Bonferroni correction method was applied. The P value significance threshold based off the total number of different tests conducted.

$$\alpha \text{ altered} = \frac{\alpha \text{ original } 0.05}{\text{number of tests}}$$

5.3 Results

5.3.1 Assessing *AHRR* methylation differences in smokers versus non smokers-model 3

To assess the validity of this study, we compared differential DNA methylation between the CHDS subset used in these analyses, against that observed from the subset of the CHDS cohort used in Chapter 3, at one CpG site within the gene *AHRR (*Illumina ID cg05575921); this amplicon was used in Chapter 3 and so is included here as a control. The magnitude of difference between the individuals in this study who smoked tobacco (N = 32) compared to non-smokers (N= 64) was compared to Chapter 3 cannabis with tobacco smokers (N = 48), compared to non-smoking controls (N = 38).

Our previous data from Chapter 3 demonstrated an average β difference between cases and controls of 4.1% (Table 5.4). The methylation difference here was found to be conservative, however statistically relevant between smokers and non-smokers (as well as cannabis smokers). In this new analyses, we detect a methylation difference of 3.1%. The direction of change was the same, showing hypomethylated in cases vs. controls.

Table 5.4 β differences in the gene *AHRR* between BSAS in Chapter 3 using tobacco and cannabis users and here in this new cohort which has sub-selected the adult smoker for this comparison.

	Cannabis and tobacco users in Chapter 3	Controls in Chapter 3	Methylation difference	Smokers in the *in utero* study	Controls in the *in utero* study	Methylation difference
AHRR cg05575921	0.701	0.742	-0.04	0.716	0.748	-0.031

5.3.2 Validating previously reported CpG sites in response to *in utero* exposure to tobacco

Initially, we attempted to validate in our cohort (age ~28-30 years) 5 CpG sites which have been previously reported to be differentially methylated in the DNA of cord blood from newborns, and whole blood from children and adolescents (ages newborn to 17) in response to *in utero* tobacco exposure (Table 5.5). Data were partitioned into those individuals exposed *in utero*, and those who were not, and corrected for CP score (Model 1, Methods).

Table 5.5 Previously reported CpG sites showing differential DNA methylation in response to *in utero* tobacco exposure, and their average methylation values in individuals from this cohort (Model 1).

Gene	Illumina ID	Exposed *in utero* methylation	Non-exposed *in utero* methylation	β difference	*P value (nominal)*
AHRR	cg05575921	72.287	75.448	-3.161	0.022
CNTNAP2	cg2594950	3.8457	3.8600	-0.014	0.991
CYP1A1	cg05549655	26.894	21.699	5.195	0.425
GFI1	cg09935388	75.151	75.330	-0.582	0.055
GFI1	cg09662411	95.837	97.400	-1.583	0.274

AHRR (cg05575921) displayed a 3.1% decrease in DNA methylation between exposed and non-exposed individuals, at a nominal P value of 0.02. This site has been previously identified as hypomethylated in adults, as well as in postnatal cord blood samples between *in utero* tobacco-exposed and non-exposed individuals. The probe cg05549655 in the gene *CYP1A1* displayed a 5.19% increase in DNA methylation in the *in utero*-exposed group, however, this site this site did not reach nominal statistical significance in our cohort. Cg09935388 and cg09662411 in *GFI1* were unable to be replicated as differentially methylated between the exposed and the non-exposed groups (no significant change in methylation). Both CpG sites did show

hypomethylation, supporting previous observations of differential methylation within this gene. *CNTNAP2* (cg2594950) was similarly unable to be validated in our cohort.

5.3.3 Differentially methylated CpGs by *in utero* tobacco exposure status

Data were partitioned according to *in utero* exposure status only (exposed vs. unexposed) using Model 1 (Methods). Of the 10 genes (encompassing a total of 280 CpG sites) selected for BSAS, 6 genes showed nominally significant differential methylation between *in utero*-exposed and non-exposed controls, across 22 different CpG sites that resided in those regions: *AHRR2*, *GRIN2b*, *GFI1*, *BDNF*, *ASH2L* and *DUSP6* (Table 5.6). The remaining genes, *CNTNAP2*, *MEF2C*, *SLC9A9* and *CYP1A1*, showed no differential methylation across the region in response to *in utero* tobacco exposure alone.

The top log fold changes (2.1 and 1.78) in differential methylation between *in utero* exposed individuals verses non-exposed individuals both come from CpG sites in *GRIN2b* (Chr12: 14133243 and Chr12: 14133359), followed by two further larger log fold changes in two CpG sites in *BDNF* (Chr11:, 27743857 and Chr11:, 27743730).

Table 5.6 Top CpG sites found to be nominally significantly differentially methylated (unadjusted P < 0.05) in the *in utero* tobacco exposed group (Model 1). Asterisk, *, indicates CpG sites in genes identified as differentially methylated in response to adult smoking status (Table 5.8) Abbreviations: FC, fold change; CPM, counts per million; FDR, FDR-corrected P value.

Gene	Illumina ID, CpG site location	*Log FC*	*Average Log CPM*	*P value*	*FDR*
**AHRR*	Chr5, 373398	-0.369	12.699	0.0009	0.187
**GFI1*	Chr1, 92946546	-0.588	12.284	0.002	0.192
**BDNF*	Chr11, 27743856	-1.323	10.237	0.004	0.192
**GRIN2b*	Chr12, 14133243	2.100	10.113	0.004	0.192
**GFI1*	Chr1, 92947559	-0.507	9.068	0.005	0.192
**GFI1*	Chr1, 92947752	-0.433	9.844	0.006	0.192
GRIN2b	Chr12, 14133359	1.789	10.523	0.007	0.192
**GFI1*	Chr1, 92946452	-0.374	12.211	0.008	0.192
**GIF1*	Chr1, 92946429	-0.558	12.163	0.009	0.192
BDNF	Chr11, 27743594	-0.773	11.078	0.010	0.192
GFI1	Chr1, 92946514	-0.477	10.053	0.011	0.200
**BDNF*	Chr11, 27743729	-1.266	8.550	0.016	0.262
GFI1	Chr1, 92946568	-0.339	12.218	0.019	0.284
**AHRR*	cg05575921	-0.270	12.687	0.022	0.291
AHRR	Chr5, 373355	-0.228	12.749	0.022	0.291
**GIF1*	Chr1, 92946418	-0.512	12.160	0.030	0.365
DUSP6	Chr12, 89746641	-0.635	10.060	0.033	0.371
GFI1	Chr1, 92946434	-0.314	12.193	0.035	0.371
GFI1	Chr1, 92946340	-0.368	12.360	0.047	0.413
**GFI1*	Chr1, 92946132	-0.420	12.295	0.048	0.413
DUSP6	Chr12, 89746479	0.813	10.285	0.049	0.413
ASH2L	Chr8, 37962720	0.692	11.626	0.049	0.413

A MA plot of the log average difference between individuals exposed *in utero*, and non-exposed individuals (Figure 5.1, Table 5.6) indicates those sites with the highest log fold changes, and demonstrates the direction of change in methylation of the 22 nominally significantly differentially methylated CpGs ($P < 0.05$); 4 are hypermethylated (pink) and 18 are hypomethylated (cyan).

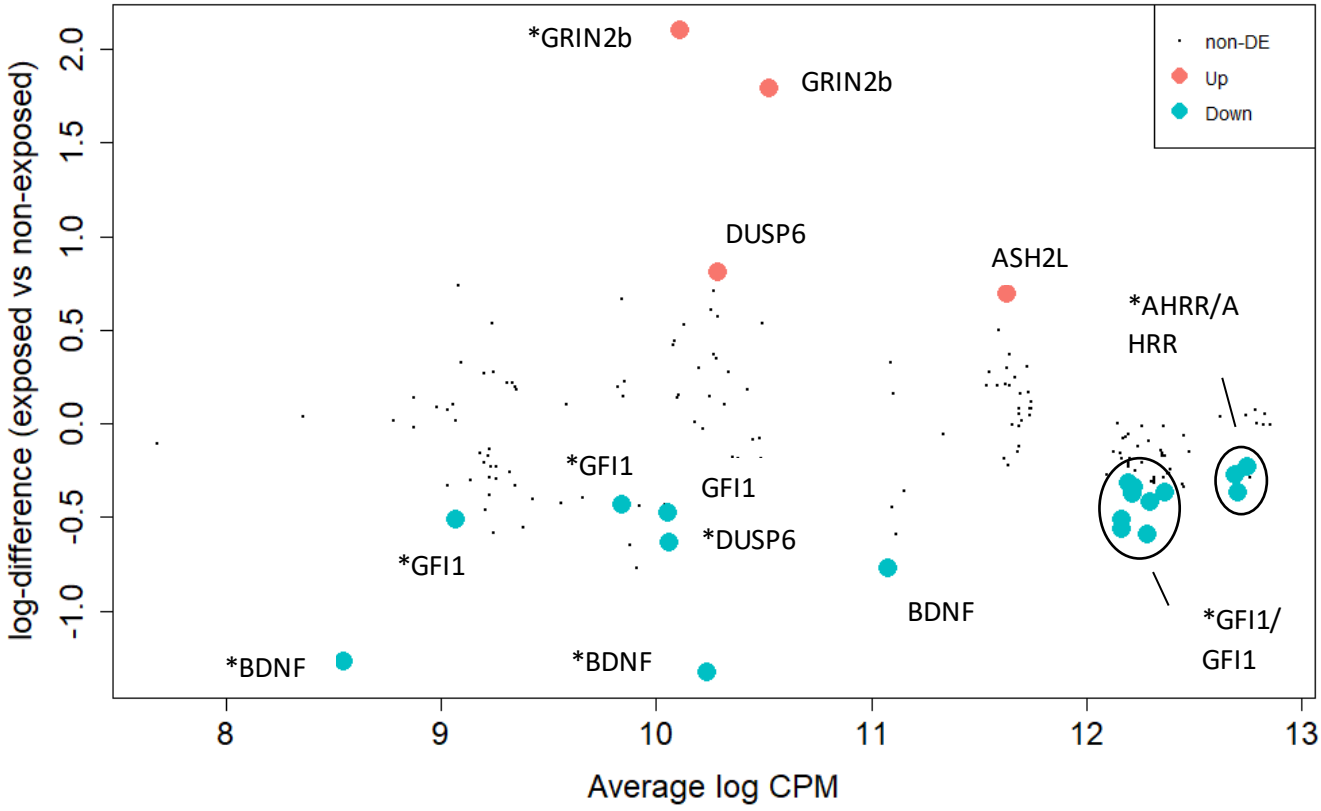


Figure 5.1 Differential DNA methylation of individuals exposed to tobacco *in utero* vs non-exposed *in utero* individuals, across 280 CpG sites within 10 genes. Dots that are displayed in colour represent those that are differentially methylated: cyan, hypomethylation; pink, hypermethytion; black, non-differentially methylated sites. *previously shown to be differentially methylated in response to adult smoking status

5.3.4 Differentially methylated CpG sites in response to CP

Data were then partitioned based upon CP and non-CP status (Model 2). A total of nine CpG sites were found to be differentially methylated (Table 5.7). Four CpG sites were independent of *in utero* exposure, while the remaining five were also identified as differentially methylated in the *in utero* exposed group (Table 5.6).

Table 5.7 Top CpG sites found to be nominally significant differentially methylated (unadjusted P < 0.05) in response to CP.

Gene	Illumina ID, CpG site location	*log FC*	*Average log CPM*	*P value*	*FDR*
DUSP6	Chr12, 89746479	-1.042	10.285	0.004	0.776
GIF1	Chr1, 92946568	0.335	12.218	0.013	0.776
CYP1A1	Chr15, 75019185	-0.943	9.079	0.014	0.776
GIF1	Chr1, 92946472	0.317	12.125	0.018	0.776
BDNF	Chr11, 27743694	1.036	10.338	0.020	0.776
CNTNAP2	Chr7, 145814223	0.345	12.826	0.024	0.776
GIF1	Chr1, 92946132	0.419	12.295	0.033	0.817
DUSP6	Chr12, 89746470	0.873	9.202	0.040	0.817
GIF1	Chr1, 92946421	0.243	12.160	0.046	0.817

CpG sites of nominal significance in response to CP were plotted in Figure 5.2. Compared to Figure 5.1 (*in utero* exposure vs. non-*in utero* exposure), CpG sites for this analysis (CP vs. non-CPs) showed seven CpG sites hypermethylated and two sites that are hypomethylated. Four of the CpG sites with nominal significance display a log fold change difference ~1 fold, there is a secondary cluster of five CpG sites of ~0.5 fold change difference.

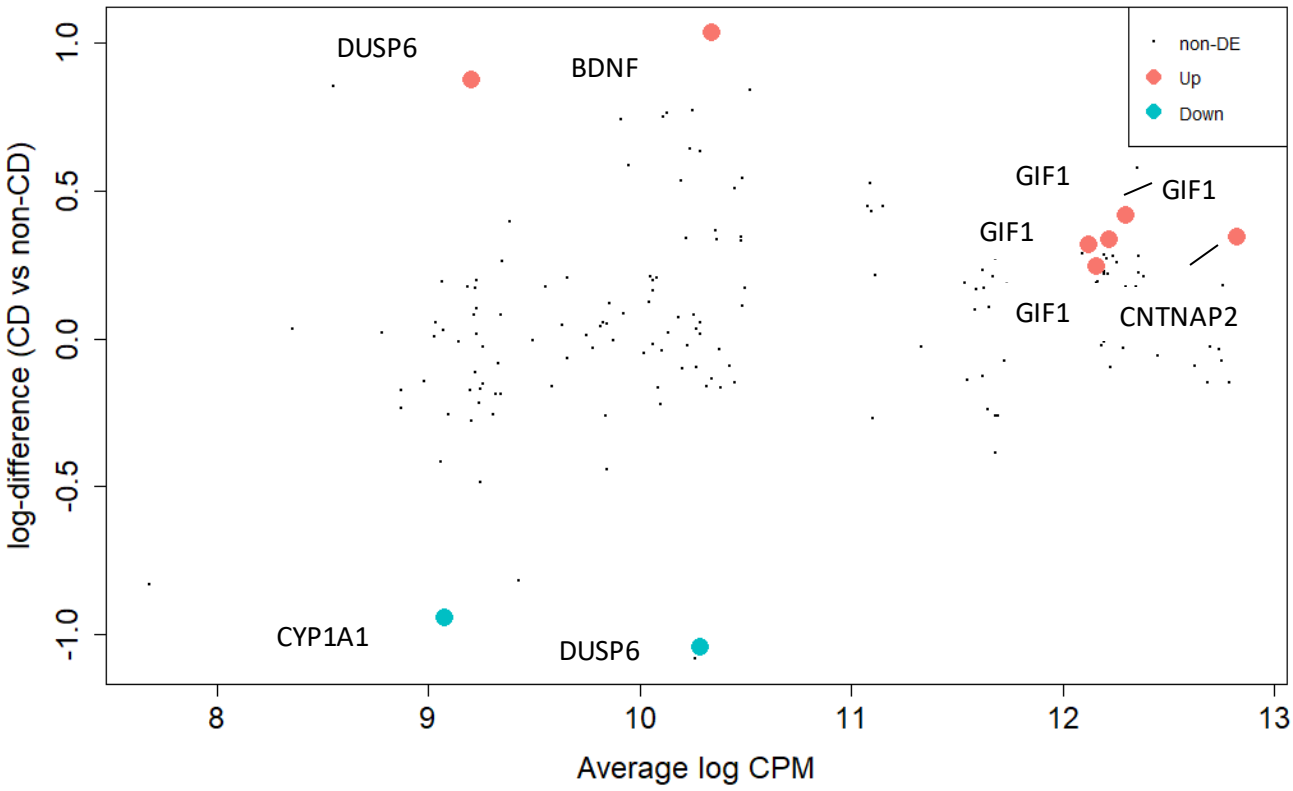


Figure 5.2 Differentially methylated sites in high CP individuals verse people with low CP scores. Dots that are displayed in colour represent those that are differentially methylated: cyan, hypomethylation; pink, hypermethylation; black, non-differentially methylated sites.

5.3.5 Differential methylation in response to adult smoking status

Smoking in adulthood was assessed for its confounding effect on DNA methylation across the amplicons of genes of interest. The data was partitioned into those individuals who were tobacco smokers in adulthood, and those who were never smokers. When differential methylation was calculated in smokers vs. never smokers, 26 out of 280 CpG sites in total were identified as significantly differentially methylated (nominal $P < 0.05$, Table 5.8). These loci were in general hypomethylated, consistent with the literature for the same or near sites with the only hypermethylated site located in the GRIN2b promoter. There were a total of 12 CpG sites that were also found to be differentially methylated in response to both of the univariate analyses of adult smoking status and *in utero* exposure (indicated by * in Table 5.6). 14 CpG sites were found solely to be differentially methylated in response to adult smoking status and 10 CpG sites differentially methylated only in response to *in utero* exposure.

Table 5.8 Top CpG sites found to be nominally significantly differentially methylated (unadjusted P < 0.05) in response to adult smoking status. Abbreviations: Log FC, Log fold change, Log CPM, Log counts per million.

Gene	CpG site location	*Log FC*	*Average Log CPM*	*P value*	*FDR*
AHHR	Chr5, 373398	-0.343	12.699	0.002	0.273
GFI1	cg09662411	-0.444	12.314	0.005	0.273
GFI1	Chr1, 92946923	-0.372	12.378	0.007	0.273
GFI1	Chr1, 92946222	-0.492	12.299	0.007	0.273
GFI1	cg09935388	-0.458	9.2268	0.008	0.273
GFI1	Chr1, 92946429	-0.560	12.163	0.008	0.273
ASH2L	Chr8, 37962657	-0.129	11.333	0.010	0.273
GFI1	Chr1, 92947752	-0.422	12.093	0.012	0.273
GFI1	Chr1, 92947586	-0.445	9.229	0.013	0.273
GRIN2b	Chr12, 14133243	2.388	10.113	0.015	0.273
GFI1	Chr1, 92946270	-0.315	12.363	0.018	0.273
ASH2L	Chr8, 37962793	-0.674	11.685	0.021	0.273
GFI1	Chr1, 92946452	-0.336	12.125	0.022	0.273
GFI1	Chr1, 92 947581	-0.332	9.2268	0.022	0.273
GFI1	Ch1, 92946415	-0.303	12.160	0.022	0.273
GFI1	Chr1, 92946620	-0.263	12.195	0.022	0.273
BDNF	Chr11, 27743452	-0.674	10.381	0.026	0.286
GFI1	cg06338710	-0.402	12.198	0.029	0.286
GFI1	Chr1, 92946434	-0.327	12.193	0.029	0.286
BDNF	Chr11, 27743729	-1.214	8.550	0.030	0.286
GFI1	Chr1, 92946418	-0.500	12.160	0.031	0.286
GFI1	Chr1, 92946235	-0.428	12.311	0.034	0.295
GFI1	Chr1, 92947559	-0.336	12.356	0.041	0.337
GFI1	Chr1, 92946132	-0.436	12.295	0.043	0.337
GFI1	Chr1, 92946452	-0.287	12.211	0.045	0.337
AHRR	cg05575921	-0.233	12.687	0.048	0.337

5.3.6 Differentially methylated CpGs dependent on both *in utero* tobacco exposure and CP

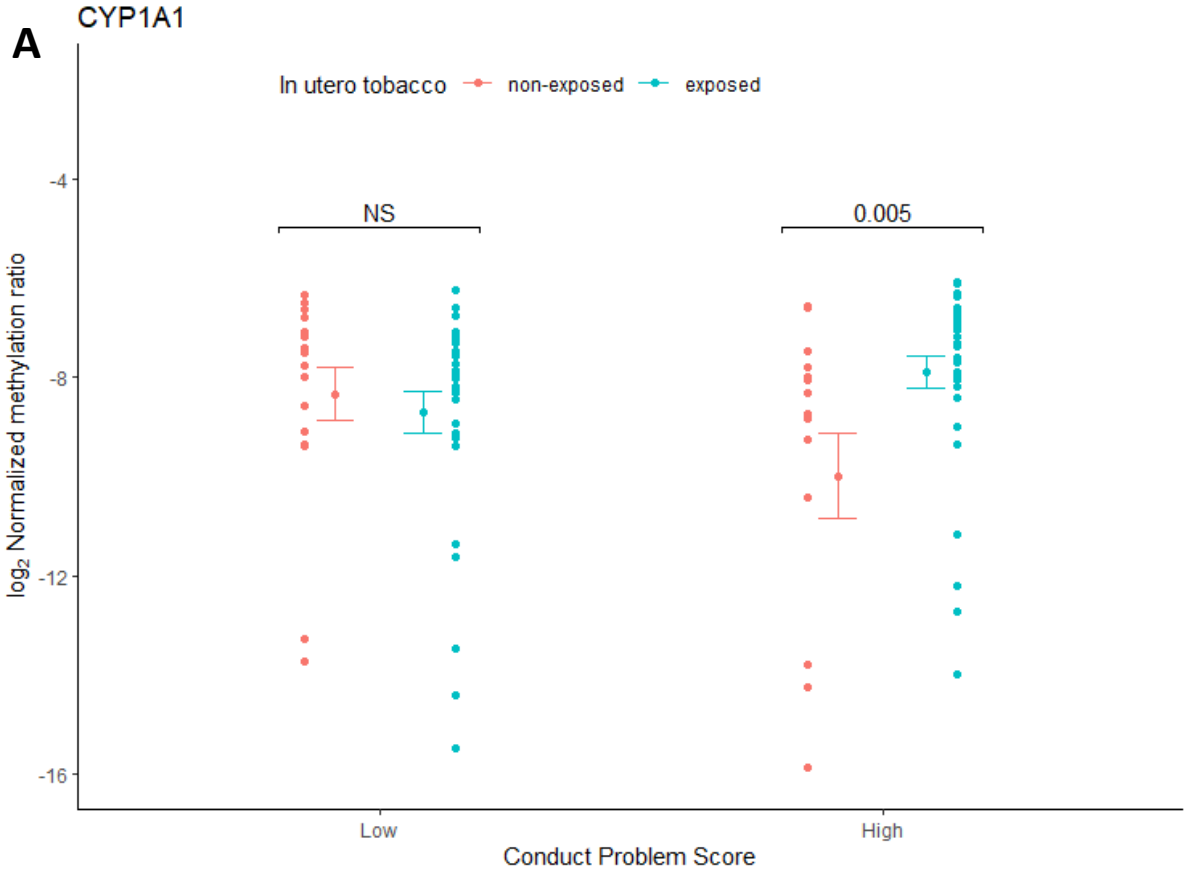
Differential methylation dependent on both *in utero* exposure and CP score was found at 10 loci in six genes at nominal significance level, however none were significant after correcting for false discovery rate (Table 5.9).

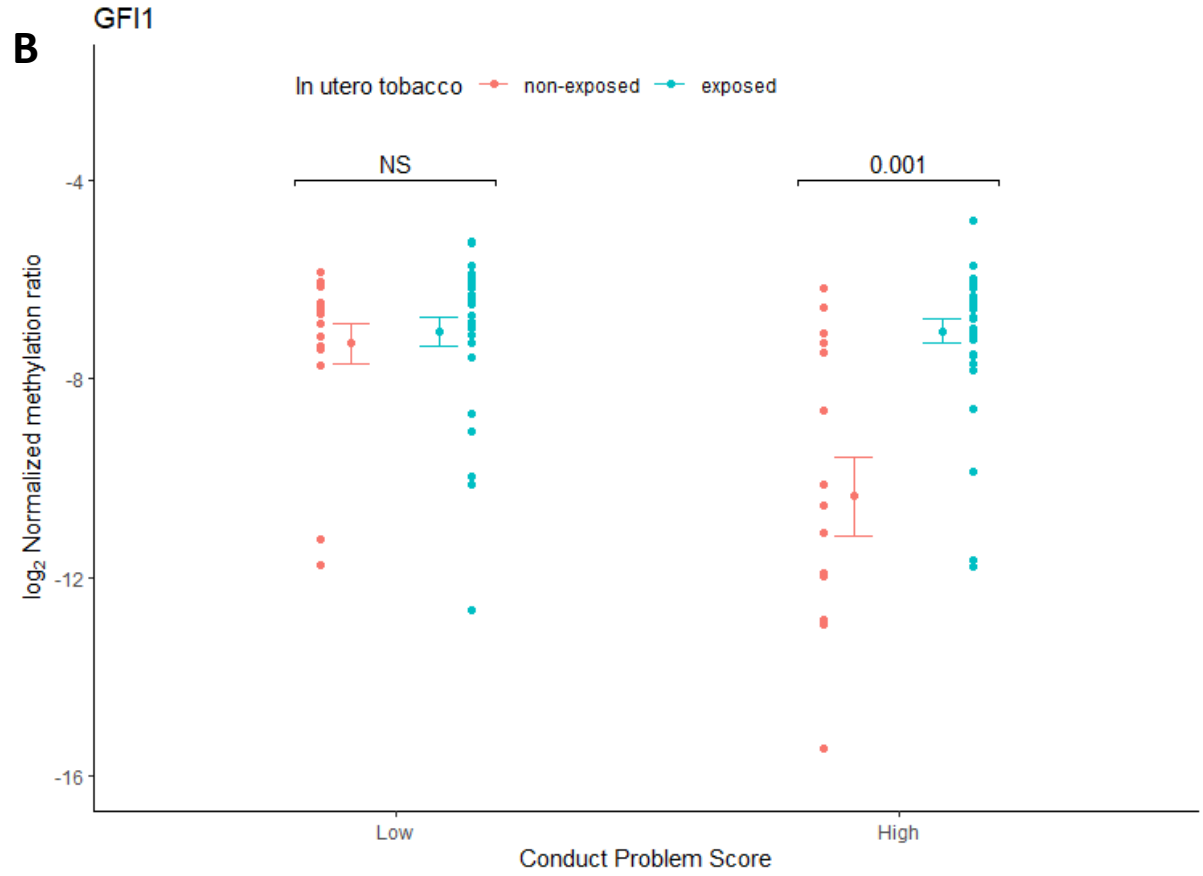
Nine out of the 10 sites (all except *DUSP6*) displayed a greater level of differential methylation between *in utero* exposure states for high conduct scores, with 5/10 nominally significant, compared to low conduct scores (no nominal significance). The CpG sites which were nominally significantly differentially methylated ($P < 0.05$) in the DNA of *in utero*-exposed individuals with high CP score were sites within *CYP1A1*, *GFI1*, ASH2L, *and GRIN2b* (Model 4, Table 5.9).

Table 5.9 CpG sites where differential methylation between conduct problem scores differs with *in utero* exposure at $P< 0.05$. Log Fold Change (FC) and P values (unadjusted) from log ratio tests for the effect on normalized methylation ratios of: (1) P value of differential methylation for the interaction between *in utero* exposure and Conduct Problem score. Then to determine whether this P value was driven by low CP score or high CP score we assessed (2) *In utero* exposed versus non-exposed in the Low CP group and (3) within High CP participants. Loci with nominally significant ($P<0.05$) interaction shown, all FDR P values > 0.05.

Gene	CpG location	Interaction(1)		Low CP(2)		High CP(3)	
		Log FC	*P value*	*Log FC*	*P value*	*Log FC*	*P value*
CYP1A1	Chr15, 75019290	-2.013	0.010	0.344	0.493	-1.669	0.005
GFI1	Chr1, 92947705	-0.957	0.011	0.002	0.992	-0.955	0.001
ASH2L	Chr8, 37962878	1.257	0.024	-0.447	0.253	0.811	0.042
MEF2C	Chr5, 88179596	-1.679	0.040	0.678	0.174	-1.000	0.122
DUSP6	Chr12, 89746588	-1.444	0.041	0.864	0.107	-0.580	0.204
ASH2L	Chr8, 37962657	-0.199	0.042	0.052	0.455	-0.147	0.033
CYP1A1	Chr15, 75019127	-1.221	0.045	0.403	0.319	-0.819	0.072
ASH2L	Chr8, 37962901	1.250	0.046	-0.561	0.205	0.688	0.121
GRIN2b	Chr12, 14133359	2.711	0.048	0.121	0.903	2.832	0.004
MEF2C	Chr5, 88179541	-1.336	0.050	0.615	0.139	-0.720	0.190

Negative log fold change values for the significantly differentially methylated sites within the high CP score group correspond to hypomethylation within the exposed group, whereas positive log fold changes correspond to hypermethylation in the *in utero* exposed group as the log normalized ratios are negative, three examples are shown in Figure 5.3. These associations were not detected when data was partitioned and analysed to assess the impact of CP only on DNA methylation (Model 2,Table 5.7).


A
CYP1A1
In utero tobacco
non-exposed
exposed
NS
0.005
-4
-8
-12
-16
log2 Normalized methylation ratio
Low
High
Conduct Problem Score


B
GFI1
In utero tobacco
non-exposed
exposed
NS
0.001
-4
-8
-12
-16
log2 Normalized methylation ratio
Low
High
Conduct Problem Score

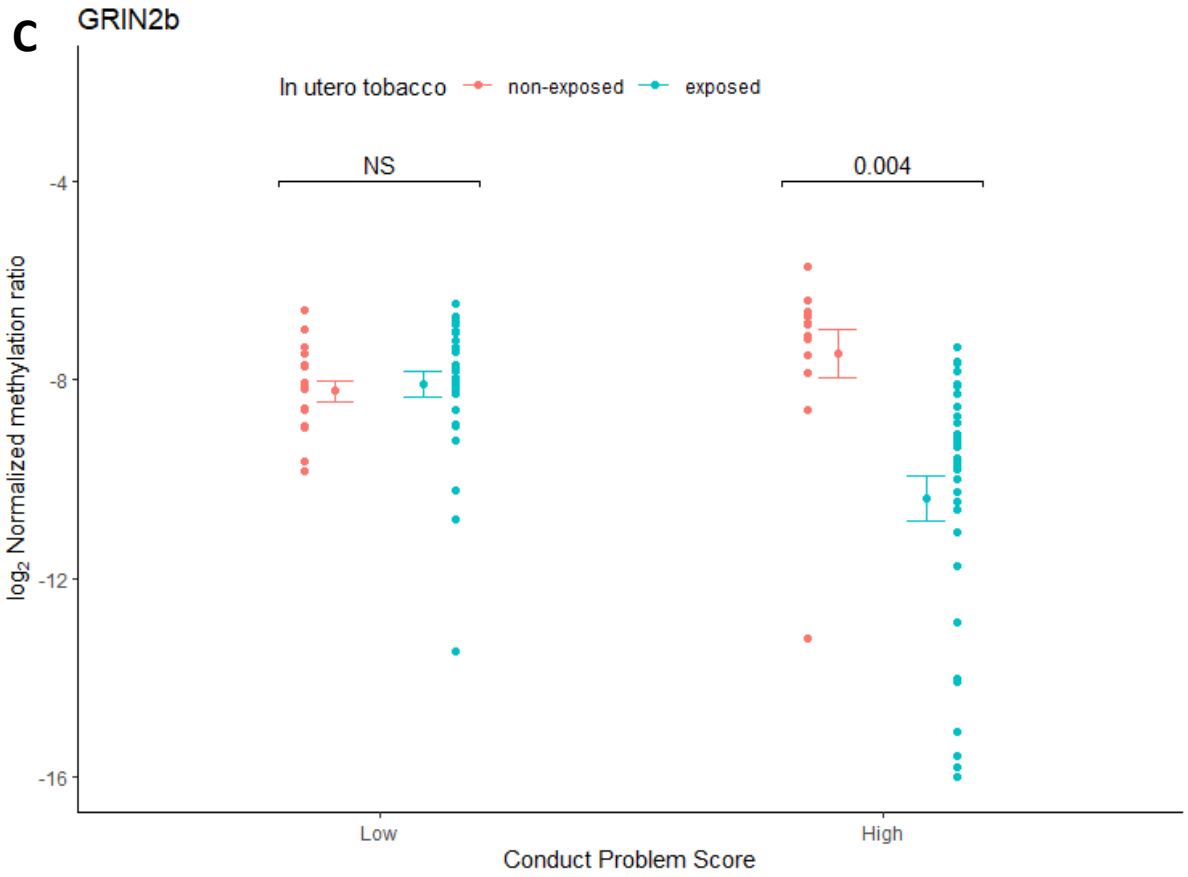


Figure 5.3 Differential methylation found *in utero* tobacco exposed for individuals with high conduct problem score that is not observed in individuals with low conduct problem score. A- CYP1A1 (Chr15, 75019290), B- GFI1 (Chr1, 92947705) and C- , GRIN2b (Chr12, 14133359).

5.3.7 Overall methylation levels across all amplicon regions

Overall methylation was assessed across all 10 gene regions (280) by each of the interactions to identify any overall patterns of differential methylation, which may give an indication of what would be represented across the whole epigenome.

Each interaction is was assessed individually between either *in utero* exposed and not exposed, CP vs non-CP and then smoking versus non-smoking. Interactions were then combined, with *in utero* exposed versus non-in utero exposed vs CP vs non-CP. Then lastly, CP vs non-CP vs smoker versus non-smoking.

Table 5.10 Overall DNA methylation differences found compared to control groups from the 280 CpG sites assessed under the different variable assessed in this study,

Variable	*Df*	*Sum Sq*	*Mean Sq*	*P value*
In utero maternal tobacco vs non-exposed	1	30908	30908	$5.15x10^{-6}$
Low CP vs high CP	1	1249	1249	0.358
Adult Smoking status vs non-smokers	1	7727	7727	0.022
In utero maternal tobacco exposure and the interaction of CP	1	850	850	0.449
CP and the interaction of adult smoking status	1	12212	12212	0.004
Residuals	4059	6020897	1483	

Overall methylation across the 10 gene regions showed one interaction in particular as being significantly statistically different: *in utero* exposed individuals have differential methylation overall, compared to non-exposed individuals. If a Bonferroni P value correction method was applied to this interaction, a P value of less than 0.01 would pass the significance threshold. *In utero* exposed versus non exposed *in utero* has a P= $5.15x10^{-6}$ which gives an indication that there is a significant difference between these two groups. The other two significant interactions were between: i) smoking and non- smoking (P= 0.02252), and; ii) CP vs non-CP with the addition of adult smoking status (P= 0.00414). Only CP with the addition adult smoking status remains significant post Bonferroni correction method.

5.4 Discussion

In utero tobacco exposure is known to alter DNA methylation at the genome-wide level in offspring [18, 19] [20, 21]. The later-life implications of these tobacco-induced DNA methylation changes are unclear, however, an association between *in utero* tobacco exposure and CP has previously been observed [26]. Given the complex etiology of CP phenotypes [56-58] and the vast array of socioeconomic variables associated with tobacco use [59], proving a causal link between maternal smoking and offspring CP is inherently challenging. However, here, we provide initial observations within our *in utero* tobacco exposed cohort that may show DNA methylation changes that are associated with CP phenotypes in offspring. These methylation changes are within a panel of genes that have known roles in *in utero* brain development and CP phenotypes.

5.4.1 Study design limitations

All individuals in this study came from the CHDS longitudinal study cohort. The study commenced in 1977, at which time the effects of maternal *in utero* smoking had not been clearly defined. In more recent times, numerous studies have found associations between maternal smoking and adverse health outcomes and long term effects of offspring [60-62].

There are many other co-variables that have the potential to confound the effects of *in utero* smoking on DNA methylation. These are variables that also have a high chance of co-occurring with maternal tobacco smoking. For example, if a women is likely to smoke throughout her pregnancy this is also a high chance she will continue to smoke throughout the upbringing of that child. Differentiating between *in utero* exposure and second-hand smoke exposure is not possible with this study design or with this cohort of individuals. The problem is also the same for parental smoking as if a male partner also is a smoker then there is a higher proportionate chance that the female partner also smokes. Minimal research has been carried out on the effect of second-hand smoking, however this is very much a limitation for a various study who investigate the

effects of *in utero* exposure. Similarly, to second-hand smoking, alcohol consumption is another co variable which we are unable to account for.

This data set was not corrected for on ethnicity or sex, which is another confounding issue. However, due to the small sample size, correcting for these factors would have been detrimental to detecting differential DNA methylation. To validate and confirm our findings from this work, a larger sample size should be used. At that point, it would then be appropriate to correct for these two variables, similar to that which was carried out in Chapter 2, where the data was also corrected for five cell types, batch effects, four principal components and parent socioeconomic status, as well as ethnicity and sex. However, in this Chapter, in order to fully explore our hypotheses on a limited dataset with a small number of loci, we have not taken these variables into account.

5.4.2 Validation of previously identified differentially methylated CpG from *in utero* tobacco exposure

First, we asked whether differentially methylated CpGs that have been previously associated with *in utero* tobacco exposure were supported by this cohort. Here, we present validation of differential methylation of a CpG site within the gene *AHRR* (cg05575921). *AHRR* is a well-defined tobacco smoking gene, which is consistently represented in tobacco methylation data. *AHRR* has previously been found to be differentially methylated in response to *in utero* tobacco exposure [22, 45, 63]. We find that this particular CpG within *AHRR* remains differentially methylated in response to *in utero* tobacco exposure in our adult cohort at age ~28-30 (Table 5.1). However, in this study, differential methylation at this CpG site was also explained by adult smoking status (Table 5.8). The four CpG sites (*AHRR, CYP1A1, CNTNAP2* and GFI1) investigated here due to previous association with *in utero* tobacco exposure were not differentially methylated in our data. However, the direction of methylation change was supported at all five sites investigated [47, 64, 65]. We suggest that further investigation in a larger cohort may lead to nominal significance at the sites in *CYP1A1*, *CNTNAP2,* and *GFI1*.

5.4.3 Identification of *in utero* exposure-related differentially methylated CpGs

Next, we compared all individuals exposed to tobacco *in utero*, to individuals not exposed to tobacco *in utero,* and we identified a large number of differentially methylated CpG sites (22, Table 5.6). Of these, 20 represent novel sites, which are not target CpG sites in the Illumina EPIC or 450K array systems (the most commonly used methylation arrays for which published data is available). Thus, these sites were unable to be previously identified as differentially methylated in response to *in utero* tobacco exposure. This highlights the benefits of the BSAS method, which enables estimates of differential methylation of all CpGs within a particular amplicon [38]. Further, the novel CpG sites we identify here are all in relatively close proximity to one another, suggesting that these sites may represent differentially methylation regions. Differentially methylation regions have important roles in regulating gene expression, thus potentially leading to changes in phenotype that could influence health outcomes [66]. None of the 22 CpG sites identified as being differentially methylated in response to *in utero* tobacco exposure remained significantly differentially methylated after FDR correction, which was expected because of small sample size. However, while our data are nominally significant, it does suggest that *in utero* tobacco exposure may be affecting DNA methylation at CpG sites within genes that had no overlap with adult smoking status in this study.

5.4.4 Some changes in response to adult smoking status and *in utero* exposure unable to be differentiated

We assessed what effect adult smoking status was having on differential methylation within well studied genes, in order to determine differential methylation patterns specifically impacted by *in utero* tobacco exposure. The premise here was that CpG sites which were not identified in response to adult smoking status would indicate that the differential methylation we identify was much more likely to be induced during development, and not a by-product of adult smoking status. When the data were partitioned based on adult smoking status (Model 3), we identified 26 differentially methylated CpGs (Table 5.8). Of these, 12 CpG sites overlapped with the CpG sites found to be differentially methylated when the data was partitioned based upon *in utero*

tobacco exposure status (Table 5.6, Model 1). This indicates that differential DNA methylation identified in genes which overlap between Models 1 and 3 may be explained by adult smoking status, or *in utero* exposure. However, the remaining ten CpG sites observed in our panel of genes are not explained by adult smoking status. This implies that differential methylation at these CpG sites is explained more fully by *in utero* tobacco exposure, and provides confidence that the differential methylation we observe within these genes is more likely due to the *in utero* environment, than to adult smoking. We cannot ignore the fact that adult tobacco smoking may still be playing a role in differential DNA methylation at these sites, but it does not appear to explain the variation in methylation we observe at the sites investigated in this study as fully as the *in utero* environment.

Differential methylation within *AHRR* (cg05575921) was explained by adult smoking status in this study (Table 5.8). This was an expected result as this site is one of the most pronounced and associated sites found to be differentially methylated in tobacco smoking [67, 68]. The site, however, also showed nominal significance in response to *in utero* maternal tobacco exposure. The reason for this may be due to the study design; this study was limited by sample size, and as such, distinguishing between adult smoking status and *in utero* tobacco exposure is difficult; CpG sites which could show differences in response to both variables may have skewed the results when independently assessing them within this relatively small sample.

Tobacco smoking is known to greatly affect DNA methylation, and because the DNA samples used in this study are from individuals who were between 28 and 30 years old, adult smoking is closer temporally than *in utero* exposure. Thus we hypothesise that the data used in the *in utero* exposure model could be expected to be confounded to some extent by adult smoking status, meaning that, in these data, differential methylation at certain sites can be explained independently by both *in utero* tobacco exposure and adult smoking status. Further investigations in larger cohorts, preferably at the genome-wide level, are required. To further rule out adulthood smoking status as an explanatory factor in the differential methylation we observe within our panel of brain development and CP genes, this study should be expanded to include an additional group of individuals that were not exposed to tobacco *in utero*, but are smokers as adults.

5.4.5 Identification of *in utero* exposure-related differentially methylated CpGs that are specific to individuals with CP

An overwhelming amount of epidemiological data has shown an increased association between *in utero* tobacco exposure and behavioural disorder in children and adolescents [69, 70]. Thus, here, we investigated DNA methylation changes induced by *in utero* tobacco exposure as a potential molecular mechanism of dysfunction that could link the phenotypic trait of CP to maternal tobacco use during pregnancy. We therefore analysed DNA methylation patterns within our gene panel in response to *in utero* tobacco exposure and its interaction with CP status. A total of 10 CpG sites in six genes were found to display nominal significance in DNA methylation in response to *in utero* tobacco exposure and CP in this cohort (Table 5.9, Model 4). Differential methylation at none of these CpG sites could not be explained by adult smoking status.

The candidate genes explored here have been shown to be differentially methylated in response to both adult smoking and *in utero* smoking. We observed that when *in utero* smoking and CP score were considered together, differential methylation attributed to *in utero* exposure was significantly different in those with high CP scores than in those with low CP scores. In the 10 loci we identified with interactive differential methylation, all but the loci in *DUSP6* showed greater magnitude differential methylation in high CP scores (exposed *in utero* vs. non-exposed with high CPS), with reduced, reversed or no evidence of differential methylation at the same sites with low CP score. While we cannot assert causality, our results are consistent with *in utero* tobacco exposure altering methylation at loci associated with neural phenotypes which persist into adulthood and are then associated with increased risk of CP.

Our results indicate that *in utero* tobacco exposure is associated with a greater level of *MEF2C* hypomethylation in participants who were exposed to tobacco *in utero* with CP in this cohort, although not at the FDR significance level. We identified differential methylation at two CpG sites that are located next to each other within the gene *MEF2C* (chr5, 88179596 and 88179541). *MEF2C* (Myocyte enhancer factor 2C) is a transcription factor which regulates gene expression for development and maintenance in a variety of tissues [71]. It has been shown to play an important role in the brain [72-76], particularly, in neuronal migration and neuronal differentiation [77-79]. More so, *MEF2C* in plays a role in neural crest formation during development,

where tissue-specific inactivation of the gene results in embryonic lethality [80]. Further, *MEF2* interacts with oxytocin, which is affiliated with prosocial behaviours [81, 82]. Alterations to oxytocin have been shown to change the morphology of neurons via *MEF2A* [83, 84]. Functional roles of the gene in relation to early neuronal development still remain unclear, however it is thought to play a crucial role [85].

Three CpG sites from the gene *ASH2L* (ASH2 like histone lysine methyltransferase complex subunit) were also found to display differential methylation in response to *in utero* tobacco exposure and CP. *ASH2L* has been found to interact with *MEF2C* to mediate changes in histone 3 lysine 4 trimethylation (H3K4me3 [86]). Recent research in animal models suggests that nicotine-dependent induction of the *ASH2L* and *MEF2C* complex during development induces alterations that could lead to fundamental changes in the brain. These consist of dendritic branching and hypersensitive passive avoidance behaviour which is a consequence of developmental nicotine exposure [86]. Our findings support this hypothesis by providing molecular evidence of CpG site alterations in these genes via *in utero* tobacco exposure in individuals with high CP score.

However, these sites were not differentially methylated in response to CP vs non-CP alone (Model 2, Table 4.7), suggesting that DNA methylation changes in developmental genes are both induced by maternal tobacco use during pregnancy, and involved in pathways in development of CP phenotypes. Further, the persistence of specific *in utero* related DNA methylation changes into adulthood, as identified here, indicates that methylation differences at these genes may be induced during development and stable over the life course, potentially indicating the presence of metastable epialleles within these genes.

Although adult smoking status was the only other variable able to control for in this study we cannot account for many other confounding variables when assessing *in utero* effects. Other genetic factors such as sex and ethnicity, as well as social interactions of economic status are all confounding variables. Ideally, this study should be repeated in a larger cohort to further for assess these confounding variables on *in utero* tobacco exposure.

5.4.6 Overall hypomethylation found

When average DNA methylation across all CpG sites investigated (280) differences were found between the different variables. The most significant differences was in the *in utero* exposed group with a Bonferroni corrected value of P= 5.15E-06. Hypomethylation was displayed within this group. These finding lead to the question, if we were to conduct a genome wide assessment of *in utero* exposure would we see this same result as seen in the preselected 10 genes? To further this observation, a genome-wide approach should be used to investigate the overall differences of methylation, as this may provide additional useful information at the genome-wide level, both supporting and expanding on the findings reported here. We also detected overall methylation differences in response to adult smoking status (P= 0.02252), however this did not pass Bonferroni correction. This implies that the DNA methylation we observe within this panel of genes is more likely to be driven by the *in utero* environment, rather than adult smoking status, highlighting the importance of developmentally-induced DNA methylation changes to offspring phenotypes.

5.4.7 Significance

It is widely known that tobacco smoking has significant genome-wide effects on DNA methylation. Thus here, we asked whether tobacco smoking during pregnancy affected offspring DNA methylation in the CHDS cohort. As chemical compounds in tobacco are so harmful to an adult [87], we hypothesised that these same chemicals would also be having an impact on offspring, if smoking continued throughout pregnancy. What we were unsure of, was if these effects *in utero* could still be detected in individuals as adults. Some CpG sites showing differential DNA methylation effects due to tobacco smoke have been found to be reversible over time, or somewhat reformed [88], however, here we present preliminary data which suggests that developmentally-induced DNA methylation changes can persist into adulthood.

While adult smoking status was able to be controlled for in this study, we were not able to control for many other confounding variables. This can be problematic, as other genetic factors such as sex and ethnicity, as well as social interactions of economic status can all play a role in variability of results. Secondhand smoke throughout one's

lifetime can also cause changes to DNA methylation. Whether these differences are in fact integrated into the genome *in utero*, or acquired during development in childhood or adolescence, is unknown. However, this is something which we are unable to address in our cohort, as the cohort only consists of adult DNA samples. Several studies, however, are now addressing this and taking samples of newborn cord blood or placenta tissues [20, 89-91] and will take further samples throughout childhood to account for secondhand smoke as a variable. Ideally, taking multiple blood samples beginning at birth, throughout childhood and into adulthood will provide the best study design for eliminating secondhand smoke, and other childhood exposures, as a variables.

The frequency of maternal smoking during pregnancy in New Zealand, as of 2010, was estimated to be 18.4% [7]. This is a substantial proportion of pregnancies, and this research will serve to increase our knowledge base around the risk of such activities, most of which are preventable risks [1]. Given the prevalence of maternal tobacco use, it is a very important health issue in today's society. Providing a molecular link between maternal tobacco use and an adverse phenotypic outcome is therefore highly valuable, as this research will directly contribute to prevention methods, via early identification of at-risk individuals, and timely behavioural interventions.

While this study focuses on maternal tobacco use, it is also a model for a variety of other exposures. For instance, maternal tobacco smoking is an extreme exposure, but more importantly there are various other maternal lifestyle factors that can cause differences in the epigenome of offspring [92, 93]. For example, nutrition [92], sugar [94], caffeine [95], alcohol [96] and cannabis [97] intake may all affect the epigenome of the developing offspring [98]. Thus, although these findings provide interesting observations, it further re-iterates the complexity of environmental exposures on DNA methylation, particularly as many of these environmental exposures will co-occur. Since there is not one distinct gene or CpG site showing a highly significant difference between exposed and control groups, it is still very difficult to understand precisely how DNA methylation may be contributing to the development of CP. However, by investigating specific genes, or regions within genes, we are able to offer support for the role of DNA methylation in the observed link between maternal tobacco use during pregnancy and development of CP in exposed offspring.

5.5 Chapter summary

- Maternal tobacco smoking during pregnancy is still prevalent within the New Zealand population, and it has been associated with in an increased risk of adverse outcomes for exposed offspring.
- Nominal significance was found in response to in utero exposed vs non-exposed, Low CP vs high CP and adult smoking vs non-adult smokers.
- Our preliminary data suggests that there may be an association between maternal tobacco use during pregnancy and the development of CP in children and adolescents.
- We acknowledge the limitations of this study and the data presented here are suggestive of a role for DNA methylation in the link between *in utero* tobacco exposure and offspring CP.
- Our findings should stimulate further study using a larger sample size, preferably with analysis at the genome-wide level.

5.6 References

1. Castles, A., et al., *Effects of smoking during pregnancy. Five meta-analyses.* Am J Prev Med, 1999. **16**(3): p. 208-15.
2. Blatt, K., et al., *Association of reported trimester-specific smoking cessation with fetal growth restriction.* Obstet Gynecol, 2015. **125**(6): p. 1452-9.
3. Baardman, M.E., et al., *Combined adverse effects of maternal smoking and high body mass index on heart development in offspring: evidence for interaction?* Heart, 2012. **98**(6): p. 474-479.
4. Hayatbakhsh, M.R., et al., *Maternal smoking during and after pregnancy and lung function in early adulthood: a prospective study.* Thorax, 2009. **64**(9): p. 810-814.
5. Beyer, D., H. Mitfessel, and A. Gillissen, *Maternal smoking promotes chronic obstructive lung disease in the offspring as adults.* European Journal of Medical Research, 2009. **14**(4): p. 27.
6. Kandall, S.R. and J. Gaines, *Maternal substance use and subsequent Sudden Infant Death Syndrome (SIDS) in offspring.* Neurotoxicology and Teratology, 1991. **13**(2): p. 235-240.
7. Andrews, A., et al., *Smoking prevalence trends: An analysis of smoking at pregnancy registration and at discharge from a midwife Lead Maternity Carer, 2008 to 2010.* New Zealand College of Midwives Journal, 2014. **49**.
8. Health, M.o., *Annual Update of Key Results 2017/18*, N.Z.H. Survey, Editor. 2018, In, Ministry of Health: Wellington, Ministry of Health.
9. Fergusson, D.M., L.J. Woodward, and L.J. Horwood, *Maternal Smoking During Pregnancy and Psychiatric Adjustment in Late Adolescence.* Archives of General Psychiatry, 1998. **55**(8): p. 721-727.
10. Pratt, T.C., J.M. McGloin, and N.E. Fearn, *Maternal Cigarette Smoking During Pregnancy and Criminal/Deviant Behavior: A Meta-Analysis.* International Journal of Offender Therapy and Comparative Criminology, 2006. **50**(6): p. 672-690.
11. Brennan, P.A., E.R. Grekin, and S.A. Mednick, *Maternal smoking during pregnancy and adult male criminal outcomes.* Arch Gen Psychiatry, 1999. **56**(3): p. 215-9.
12. Roza, S.J., et al., *Maternal smoking during pregnancy and child behaviour problems: the Generation R Study.* International Journal of Epidemiology, 2008. **38**(3): p. 680-689.
13. Gaysina, D., et al., *Maternal smoking during pregnancy and offspring conduct problems: evidence from 3 independent genetically sensitive research designs.* JAMA psychiatry, 2013. **70**(9): p. 956-963.
14. Rauschert, S., et al., *Maternal smoking during pregnancy induces persistent epigenetic changes into adolescence, independent of postnatal smoke exposure and is associated with cardiometabolic risk.* Frontiers in Genetics, 2019. **10**(770).
15. Langley, K., et al., *Maternal smoking during pregnancy as an environmental risk factor for attention deficit hyperactivity disorder behaviour. A review.* Minerva Pediatr, 2005. **57**(6): p. 359-71.
16. Kazdin, A.E., *Practitioner review: Psychosocial treatments for conduct disorder in children.* Child Psychology & Psychiatry & Allied Disciplines, 1997. **38**(2): p. 161-178.
17. Gluckman, P.D., M.A. Hanson, and T. Buklijas, *A conceptual framework for the developmental origins of health and disease.* Journal of Developmental Origins of Health and Disease, 2009. **1**(1): p. 6-18.
18. Joubert, B.R., et al., *450K epigenome-wide scan identifies differential DNA methylation in newborns related to maternal smoking during pregnancy.* Environ Health Perspect, 2012. **120**(10): p. 1425-31.
19. Joubert, B.R., et al., *DNA Methylation in Newborns and Maternal Smoking in Pregnancy: Genome-wide Consortium Meta-analysis.* Am J Hum Genet, 2016. **98**(4): p. 680-96.
20. Joubert Bonnie, R., et al., *450K Epigenome-Wide Scan Identifies Differential DNA Methylation in Newborns Related to Maternal Smoking during Pregnancy.* Environmental Health Perspectives, 2012. **120**(10): p. 1425-1431.
21. Richmond, R.C., et al., *Prenatal exposure to maternal smoking and offspring DNA methylation across the lifecourse: findings from the Avon Longitudinal Study of Parents and Children (ALSPAC).* Human Molecular Genetics, 2014. **24**(8): p. 2201-2217.
22. Richmond, R.C., et al., *Prenatal exposure to maternal smoking and offspring DNA methylation across the lifecourse: findings from the Avon Longitudinal Study of Parents and Children (ALSPAC).* Human molecular genetics, 2015. **24**(8): p. 2201-2217.

23. Lee, K.W., Richmond, R., Hu, P., French, L., Shin, J., Bourdon, C., … Pausova, Z. (2015). Prenatal exposure to maternal cigarette smoking and DNA methylation: epigenome-wide association in a discovery sample of adolescents and replication in an independent cohort at birth through 17 years of age. Environmental health perspectives, 123(2), 193–199. doi:10.1289/ehp.1408614.
24. Osborne, A.J., et al., *Chapter 4 - Epigenetics and the Maternal Germline*, in *Transgenerational Epigenetics*, T. Tollefsbol, Editor. 2014, Academic Press: Oxford. p. 27-41.
25. Parmar, P., et al., *Association of maternal prenatal smoking GFI1-locus and cardio-metabolic phenotypes in 18,212 adults.* EBioMedicine, 2018. **38**: p. 206-216.
26. Sengupta, S.M., et al., *Locus-specific DNA methylation changes and phenotypic variability in children with attention-deficit hyperactivity disorder.* Psychiatry Research, 2017. **256**: p. 298-304.
27. Dolinoy, D.C., D. Huang, and R.L. Jirtle, *Maternal nutrient supplementation counteracts bisphenol A-induced DNA hypomethylation in early development.* 2007. **104**(32): p. 13056-13061.
28. Waterland, R.A., et al., *Maternal methyl supplements increase offspring DNA methylation at Axin fused.* 2006. **44**(9): p. 401-406.
29. Waterland, R.A. and R.L. Jirtle, *Transposable Elements: Targets for Early Nutritional Effects on Epigenetic Gene Regulation.* 2003. **23**(15): p. 5293-5300.
30. Waterland, R.A., et al., *Season of conception in rural gambia affects DNA methylation at putative human metastable epialleles.* PLoS genetics, 2010. **6**(12): p. e1001252-e1001252.
31. Harris, R.A., D. Nagy-Szakal, and R. Kellermayer, *Human metastable epiallele candidates link to common disorders.* Epigenetics, 2013. **8**(2): p. 157-163.
32. Kühnen, P., et al., *Interindividual Variation in DNA Methylation at a Putative POMC Metastable Epiallele Is Associated with Obesity.* Cell Metabolism, 2016. **24**(3): p. 502-509.
33. Rutter M, T.J., Whitmore K, *Education, Health and Behaviour.* London: Longmans, 1970.
34. Conners, C.K., *Symptom patterns in hyperkinetic, neurotic, and normal children.* Child Development, 1970. **41**(3): p. 667-682.
35. Conners, C.K., *A teacher rating scale for use in drug studies with children.* The American Journal of Psychiatry, 1969. **126**(6): p. 884-888.
36. Fergusson, D.M., L.J. Horwood, and M. Lloyd, *Confirmatory factor models of attention deficit and conduct disorder.* J Child Psychol Psychiatry, 1991. **32**(2): p. 257-74.
37. Fergusson, D.M., L.J. Horwood, and E.M. Ridder, *Show me the child at seven: the consequences of conduct problems in childhood for psychosocial functioning in adulthood.* J Child Psychol Psychiatry, 2005. **46**(8): p. 837-49.
38. Noble, A., et al., *A validation of Illumina EPIC array system with bisulfite-based amplicon sequencing.* bioRxiv, 2020: p. 2020.05.25.115428.
39. Cox, M.P., D.A. Peterson, and P.J. Biggs, *SolexaQA: At-a-glance quality assessment of Illumina second-generation sequencing data.* BMC Bioinformatics, 2010. **11**(1): p. 485.
40. Krueger, F. and S.R. Andrews, *Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications.* Bioinformatics, 2011. **27**(11): p. 1571-1572.
41. Chen, Y., et al., *Differential methylation analysis of reduced representation bisulfite sequencing experiments using edgeR.* F1000Research, 2017. **6**: p. 2055-2055.
42. Dongwen Luo, S.G.a.J.K., *Calculate Predicted Means for Linear Models*. 2020-04-14.
43. Joubert, B.R., et al., *450K epigenome-wide scan identifies differential DNA methylation in newborns related to maternal smoking during pregnancy.* Environmental health perspectives, 2012. **120**(10): p. 1425.
44. Richmond, R.C., et al., *Prenatal exposure to maternal smoking and offspring DNA methylation across the lifecourse: findings from the Avon Longitudinal Study of Parents and Children (ALSPAC).* Human Molecular Genetics, 2015. **24**(8): p. 2201-2217.
45. de Vocht, F., et al., *Assessment of offspring DNA methylation across the lifecourse associated with prenatal maternal smoking using Bayesian Mixture Modelling.* International journal of environmental research and public health, 2015. **12**(11): p. 14461-14476.
46. van Otterdijk, S.D., A.M. Binder, and K.B. Michels, *Locus-specific DNA methylation in the placenta is associated with levels of pro-inflammatory proteins in cord blood and they are both independently affected by maternal smoking during pregnancy.* Epigenetics, 2017. **12**(10): p. 875-885.
47. Rotroff, D.M., et al., *Maternal smoking impacts key biological pathways in newborns through epigenetic modification in Utero.* BMC genomics, 2016. **17**(1): p. 976.
48. Li, L., et al.

49. Skogstrand, K., et al., *Reduced neonatal brain-derived neurotrophic factor is associated with autism spectrum disorders.*
50. Jiao, S.S., et al., *Brain-derived neurotrophic factor protects against tau-related neurodegeneration of Alzheimer's disease.*
51. Joubert, B.R., et al., *DNA methylation in newborns and maternal smoking in pregnancy: genome-wide consortium meta-analysis.* The American Journal of Human Genetics, 2016. **98**(4): p. 680-696.
52. Rzehak, P., et al., *Maternal Smoking during Pregnancy and DNA-Methylation in Children at Age 5.5 Years: Epigenome-Wide-Analysis in the European Childhood Obesity Project (CHOP)-Study.* PLOS ONE, 2016. **11**(5): p. e0155554.
53. Suter, M., et al., *In utero tobacco exposure epigenetically modifies placental CYP1A1 expression.* Metabolism-Clinical and Experimental, 2010. **59**(10): p. 1481-1490.
54. Demontis, D., et al., *Discovery of the first genome-wide significant risk loci for ADHD.* BioRxiv, 2017: p. 145581.
55. Riva, V., et al., *GRIN2B predicts attention problems among disadvantaged children.* European child & adolescent psychiatry, 2015. **24**(7): p. 827-836.
56. Acosta, M.T., M. Arcos-Burgos, and M. Muenke, *Attention deficit/hyperactivity disorder (ADHD): Complex phenotype, simple genotype?* Genetics in Medicine, 2004. **6**(1): p. 1-15.
57. Beaver, K.M., et al., *A gene × gene interaction between DRD2 and DRD4 is associated with conduct disorder and antisocial behavior in males.* Behavioral and Brain Functions, 2007. **3**(1): p. 30.
58. Salvatore, J.E. and D.M. Dick, *Genetic influences on conduct disorder.* Neuroscience & Biobehavioral Reviews, 2018. **91**: p. 91-101.
59. Lantz, P.M., et al., *Socioeconomic Factors, Health Behaviors, and MortalityResults From a Nationally Representative Prospective Study of US Adults.* JAMA, 1998. **279**(21): p. 1703-1708.
60. Cornelius, M.D., L. Goldschmidt, and N.L. Day.
61. Robinson, M., et al., *Smoking cessation in pregnancy and the risk of child behavioural problems: a longitudinal prospective cohort study.* Journal of Epidemiology and Community Health (1979-), 2010. **64**(7): p. 622-629.
62. Chatterton, Z., et al., *In utero exposure to maternal smoking is associated with DNA methylation alterations and reduced neuronal content in the developing fetal brain.* Epigenetics & chromatin, 2017. **10**: p. 4-4.
63. Joubert, B.R., et al., *450K epigenome-wide scan identifies differential DNA methylation in newborns related to maternal smoking during pregnancy.* Environ Health Perspect, 2012. **120**(10): p. 1425-31.
64. Tehranifar, P., et al., *Maternal cigarette smoking during pregnancy and offspring DNA methylation in midlife.* Epigenetics, 2018. **13**(2): p. 129-134.
65. Rauschert, S., et al., *Maternal Smoking During Pregnancy Induces Persistent Epigenetic Changes Into Adolescence, Independent of Postnatal Smoke Exposure and Is Associated With Cardiometabolic Risk*. Frontiers in genetics, 2019. **10**: p. 770-770.
66. Tobi, E.W., et al., *DNA methylation signatures link prenatal famine exposure to growth and metabolism.* Nature Communications, 2014. **5**(1): p. 5592.
67. Philibert, R.A., et al., *Changes in DNA methylation at the aryl hydrocarbon receptor repressor may be a new biomarker for smoking.* Clinical Epigenetics, 2013. **5**(1): p. 19.
68. Lee, D.-H., et al., *Performance of urine cotinine and hypomethylation of AHRR and F2RL3 as biomarkers for smoking exposure in a population-based cohort.* PLOS ONE, 2017. **12**(4): p. e0176783.
69. Carter, S., et al., *Maternal smoking during pregnancy and behaviour problems in a birth cohort of 2-year-old Pacific children in New Zealand.* Early Human Development, 2008. **84**(1): p. 59-66.
70. Mick, E., et al., *Case-Control Study of Attention-Deficit Hyperactivity Disorder and Maternal Smoking, Alcohol Use, and Drug Use During Pregnancy.* Journal of the American Academy of Child & Adolescent Psychiatry, 2002. **41**(4): p. 378-385.
71. Pon, J.R. and M.A. Marra, *MEF2 transcription factors: developmental regulators and emerging cancer genes.* Oncotarget, 2016. **7**(3): p. 2297-2312.
72. Li, H., et al., *Transcription factor MEF2C influences neural stem/progenitor cell differentiation and maturation in vivo.* Proceedings of the National Academy of Sciences, 2008. **105**(27): p. 9397-9402.

73. Li, Z., et al., *Myocyte Enhancer Factor 2C as a Neurogenic and Antiapoptotic Transcription Factor in Murine Embryonic Stem Cells.* The Journal of Neuroscience, 2008. **28**(26): p. 6557-6568.
74. Chen, S.X., et al., *The transcription factor MEF2 directs developmental visually driven functional and structural metaplasticity.* Cell, 2012. **151**(1): p. 41-55.
75. Brusco, J. and K. Haas, *Interactions between mitochondria and the transcription factor myocyte enhancer factor 2 (MEF2) regulate neuronal structural and functional plasticity and metaplasticity.* The Journal of Physiology, 2015. **593**(16): p. 3471-3481.
76. Shalizi, A., et al., *A Calcium-Regulated MEF2 Sumoylation Switch Controls Postsynaptic Differentiation.* Science, 2006. **311**(5763): p. 1012-1017.
77. Gong, X., et al., *Cdk5-Mediated Inhibition of the Protective Effects of Transcription Factor MEF2 in Neurotoxicity-Induced Apoptosis.* Neuron, 2003. **38**(1): p. 33-46.
78. Mao, Z., et al., *Neuronal Activity-Dependent Cell Survival Mediated by Transcription Factor MEF2.* Science, 1999. **286**(5440): p. 785-790.
79. Okamoto, S.-i., et al., *Dominant-interfering forms of MEF2 generated by caspase cleavage contribute to NMDA-induced neuronal apoptosis.* Proceedings of the National Academy of Sciences, 2002. **99**(6): p. 3974-3979.
80. Verzi, M.P., et al., *The Transcription Factor MEF2C Is Required for Craniofacial Development.* Developmental Cell, 2007. **12**(4): p. 645-652.
81. Kosfeld, M., et al., *Oxytocin increases trust in humans.* Nature, 2005. **435**(7042): p. 673-676.
82. Zak, P.J., A.A. Stanton, and S. Ahmadi, *Oxytocin increases generosity in humans.* PloS one, 2007. **2**(11): p. e1128-e1128.
83. Meyer, M., et al., *Oxytocin alters the morphology of hypothalamic neurons via the transcription factor myocyte enhancer factor 2A (MEF-2A).* Mol Cell Endocrinol, 2018. **477**: p. 156-162.
84. Meyer, M., et al., *Myocyte Enhancer Factor 2A (MEF2A) Defines Oxytocin-Induced Morphological Effects and Regulates Mitochondrial Function in Neurons.* International Journal of Molecular Sciences, 2020. **21**(6): p. 2200.
85. Harrington, A.J., et al., *MEF2C regulates cortical inhibitory and excitatory synapses and behaviors relevant to neurodevelopmental disorders.* eLife, 2016. **5**: p. e20059.
86. Jung, Y., et al., *An epigenetic mechanism mediates developmental nicotine effects on neuronal structure and behavior.*
87. Talhout, R., et al., *Hazardous Compounds in Tobacco Smoke.* International Journal of Environmental Research and Public Health, 2011. **8**(2): p. 613-628.
88. Tsaprouni, L.G., et al., *Cigarette smoking reduces DNA methylation levels at multiple genomic loci but the effect is partially reversible upon cessation.* Epigenetics, 2014. **9**(10): p. 1382-1396.
89. Joubert, B.R., et al., *Maternal Smoking and DNA Methylation in Newborns: In Utero Effect or Epigenetic Inheritance?* Cancer Epidemiology Biomarkers & Prevention, 2014. **23**(6): p. 1007-1017.
90. Guerrero-Preston, R., et al., *Global DNA hypomethylation is associated with in utero exposure to cotinine and perfluorinated alkyl compounds.* Epigenetics, 2010. **5**(6): p. 539-546.
91. Wilhelm-Benartzi Charlotte, S., et al., *In Utero Exposures, Infant Growth, and DNA Methylation of Repetitive Elements and Developmentally Related Genes in Human Placenta.* Environmental Health Perspectives, 2012. **120**(2): p. 296-302.
92. Li, Y., *Epigenetic Mechanisms Link Maternal Diets and Gut Microbiome to Obesity in the Offspring.* Frontiers in genetics, 2018. **9**: p. 342-342.
93. Girchenko, P., et al., *Associations between maternal risk factors of adverse pregnancy and birth outcomes and the offspring epigenetic clock of gestational age at birth.* Clinical Epigenetics, 2017. **9**(1): p. 49.
94. Bédard, A., et al., *Maternal intake of sugar during pregnancy and childhood respiratory and atopic outcomes.* The European respiratory journal, 2017. **50**(1): p. 1700073.
95. Fang, X., et al., *In Utero Caffeine Exposure Induces Transgenerational Effects on the Adult Heart.* Scientific Reports, 2016. **6**(1): p. 34106.
96. Hutchinson, D., et al., *Prenatal alcohol exposure and infant gross motor development: a prospective cohort study.* BMC Pediatrics, 2019. **19**(1): p. 149.
97. Morris, C.V., et al., *Molecular mechanisms of maternal cannabis and cigarette use on human neurodevelopment.* The European journal of neuroscience, 2011. **34**(10): p. 1574-1583.
98. Banik, A., et al., *Maternal Factors that Induce Epigenetic Changes Contribute to Neurological Disorders in Offspring.* Genes, 2017. **8**(6): p. 150.

Chapter 6

6. Genome wide methylation analysis of *in utero* tobacco exposure and risk of conduct disorder in adolescence

6.1 Introduction

In Chapter 5, we used a targeted approach to quantify DNA methylation (bisulfite-based amplicon sequencing, BSAS), and we demonstrated differential DNA methylation that was specific to the interaction between individuals exposed to tobacco *in utero,* and the risk of conduct problem (CP) in childhood and adolescence. Specifically, a total of seven CpG sites were found to be nominally significantly differentially methylated in individuals who were exposed to tobacco *in utero,* and who had high CP scores. These CpG sites resided in six different genes, which all have roles in neurodevelopment and CP phenotypes. The *in utero* effects of tobacco exposure and the manifestation of CP later in childhood and adolescence have been described previously [1, 2]. However, a molecular mechanism between the two has not been established. Our pilot data suggested that DNA methylation could play a role in the association between maternal tobacco use during pregnancy and the development of CP. However, this highly targeted study displayed nominal significance at a handful of pre-selected genes. Here, we further investigate this association by employing a genome-wide approach (the Illumina EPIC array), applied to a new subset of the Christchurch Health and Development (CHDS) cohort, to probe genome-wide DNA methylation changes that are specific to individuals exposed to tobacco *in utero* with high CP scores.

The EPIC array tool, while expensive, does have its advantages – data obtained via EPIC arrays is highly reproducible [8]., meaning that raw data obtained in previous studies can be included in new analyses, allowing array data to be used to answer a further hypothesis. Although, this is also highly beneficial because increasing sample size will also increase statistical power to detected genome-wide associations. Thus, individuals that were used for analyses in Chapter 2, are combined with new array data in analyses here, along with their *in utero* tobacco exposure and CP score phenotypes.

Analysis at the genome-wide level allows the further exploration of DNA methylation differences due to tobacco exposure *in utero*. The additional sub-grouping of people with low and high CP scores allows us to build on our previous analysis into the interaction between *in utero* tobacco exposure and high CP score. Although further investigation of our pilot data using a similar sample size will still be bound by the same caveats as stated in Chapter 5 (e.g whole blood sampling, and limited covariate adjustments), here, we now ask if *in utero* tobacco exposure changes DNA methylation at the genome-wide level, where they are in the genome, and whether these changes associate with CP score. To further interrogate DNA methylation at the whole genome level and investigate whether we can detect differentially methylated *regions* rather than sole CpG sites. These findings are an important advance on the previous chapter, because differentially methylated regions are possible drivers of further downstream molecular changes, such as genes expression changes, histone modifications and chromatin confirmation [3, 4], all of which can lead to adverse health outcomes [5-7] . Therefore, identification of differentially methylated regions in response to *in utero* tobacco exposure will allow novel exploration of how exposure may be leading to disease (CP) in later life.

6.2 Methods

6.2.1 Study design

To utilise the maximum number of individuals for this analysis, here we utilise raw data from individuals from Chapter 2 (2016/2017 data, Table 6.1) if the maternal tobacco status during pregnancy and the individual CP score was known), along with newly acquired EPIC array data specific to this Chapter (2020, Table 6.1).

Of the cohort of individuals from Chapter 2 a total of N= 19 were exposed to tobacco *in utero* to tobacco and are non-smokers, N= 22 were exposed to tobacco *in utero* and are adult smokers, and a further N= 42 were not exposed to tobacco *in utero* and were non-smokers. These individuals were then combined with a further N=18 individuals who were sampled in 2020, to assess *in utero* tobacco exposure and its risk with conduct problems at the genome-wide level.

Table 6.1 EPIC array samples used in this study based upon year of measurement each were placed into the following sub groups, *in utero* exposed non-smokers, *in utero* exposed smokers, non-exposed *in utero* non-smokers and non-exposed *in utero* smokers.

	in utero exposed non-smoker	*in utero* exposed smoker	*in utero* non-exposed non-smoker	*in utero* non-exposed smoker
2016	12	6	28	-
2017	7	16	14	8
2020	7	1	10	-
Total	26	23	52	8

The final cohort for analysis (Table 6.2) comprises N= 109 individuals, N= 49 of which were exposed to tobacco *in utero*, and N= 60 are non-exposed controls. A subset of individuals who were exposed to tobacco *in utero,* who are also tobacco smokers themselves (N= 23) were included to control for lifetime tobacco exposure. A total of eight individuals who were not exposed in tobacco *in utero* who are adult smokers were also included in this study, which was a subgroup unable to be included in Chapter 5, which was a major limitation that we are able to address here. Subgroups

of CP low and CP high scored individuals were included for each of the *in utero* status groups. Diagnosis of CP is described in detail in section 5.1.1.

Table 6.2 Cohort characteristics of the *in utero* maternal tobacco exposed group and their matched controls.

	in utero maternal tobacco exposed N= 49	*in utero* non-exposed N= 60
Sex		
Male	36	46
Female	13	14
Paternal socioeconomic status		
1	3	13
2	21	30
3	25	17
Adult tobacco smoking status		
Never smoker	26	52
Regular smoker	23	8
Adult cannabis use status		
Never user	18	39
Regular user	31	21
Conduct problem score (CP)		
Low CP (< 46)	26	41
High CP (> 53)	23	19

6.2.2 DNA samples

All samples from Table 6.2 were prepared as per the DNA extraction protocol in section 2.2.3. Briefly, the 2020 samples were taken from whole blood samples and DNA extractions were conducted using the Kingfisher Flex System (Thermo Scientific, Waltham, MA USA), as per the published protocols. DNA was quantified via NanoDrop™ (Thermo Scientific, Waltham, MA USA) and standardised to 100ng/μl. Equimolar amounts were shipped to the Australian Genomics Research Facility (AGRF, Melbourne, VIC, Australia) for processing via the Infinium® Methylation EPIC

BeadChip (Illumina, San Diego, CA USA). With 8 samples being organised onto one chip.

6.2.3 Data processing

Analysis was carried out in R statistical software (Version 3.5.2). Quality control and data was processed via the protocols established in 2.2.4. Sex chromosomes and a total of 90 failed probes (detection P value of 0.01 in at least 50% of samples) were excluded from the analysis. CpG sites known to be problematic with adjacent SNVs or which did not map to a location in the genome were also excluded [8]. Leaving a total of 699,916 CpG sites for further analysis. Pre-processing was also performed using the noob, swan and Illumina normalisation methods. Normalisation was then visually inspected for performance using beta density distribution plots and Multi-dimensional scaling of the 5,000 most variable CpG sites.

6.2.4 Statistical analysis

Hierarchical regression was used to investigate the best linear model to be fitted to the methylated/unmethylated or M ratios (Table 6.3). Baseline models (Model 1, 4, 8, 10) were corrected for the following variables: i) year of sampling (3 levels), and; ii) population stratification (four principal components from 5000 most variable SNPs). Further models included combinations of the variables tobacco status (bivariate), sex (bivariate), socioeconomic status (three levels) (Model 2,5,7), cannabis status (bivariate), conduct problem (bivariate) and *in utero* tobacco smoking status (bivariate) (Model 3,6,9). Q-Q plots of the residuals were also used to compare lambda values for over-inflation.

Table 6.3 Hierarchical regression models which were used to investigate differences between each of the variables assessed in this study. CP- Conduct Problems, IU- *In utero* exposed to tobacco, PC- Principal Components, SES- Socioeconomic status.

Variable	Model	Multiple regression equation
CP	Model 1	Y ~ CP + Year + PC + *e*
CP	Model 2	Y ~ CP + Year + PC + SES + Smoking + Sex + *e*
CP	Model 3	Y ~ CP + Year + PC + SES + Smoking + Sex + IU + cannabis+ *e*
IU	Model 4	Y ~ IU + Year + PC + *e*
IU	Model 5	Y ~ IU + Year + PC + SES + Smoking + Sex + *e*
IU	Model 6	Y ~ IU + Year + PC + SES + Smoking + Sex + CP + Cannabis + *e*
IU:CP	Model 7	Y ~ IU + CP + Year + PC + SES+ + Smoking + Sex + IU:CP + *e*
IU:CP	Model 8	Y ~ IU + CP + Year + PC + IU:CP + *e*
IU:CP	Model 9	Y ~ IU + CP + Year + PC + SES + Smoking + Sex + Cannabis IU:CP + *e*
Adult tobacco smoking status	Model 10	Y ~ Smoker + Year + PC + *e*

Linear regression models used to generate the top tables of differentially methylated CpG sites were correct for multiple testing using Benjamini-Hochberg (BH). Differentially methylated CpG sites that were intergenic were matched to the nearest neighbouring genes within the Hg19 using Granges default settings [9]. The official gene symbols of all significantly differentially methylated CpG sites (nominal $P < 0.001$) for 1) *in utero* tobacco exposure, 2) CP and 3) *in utero* tobacco exposure with the interaction of CP were tested for enrichment in KEGG 2019 human pathways with EnrichR [10] and ggplot package (Version 3.3.2) was used to construct all dotplot graphs [11].

6.3 Results

6.3.1 Data pre-processing

Following the critical analysis of individual normalisation methods in Chapter 2, the same evaluation of pre-processing methods was conducted for this analysis. Specifically, the pre-processing methods noob, swan and Illumina were all fitted to the data. The data in its raw form is displayed as a beta density plot (Figure 6.1), which confirms that a batch effect is present amongst array batches in our data. The 2020 samples are more congruent with the 2017 samples than the 2016 measurements (Figure 6.1).

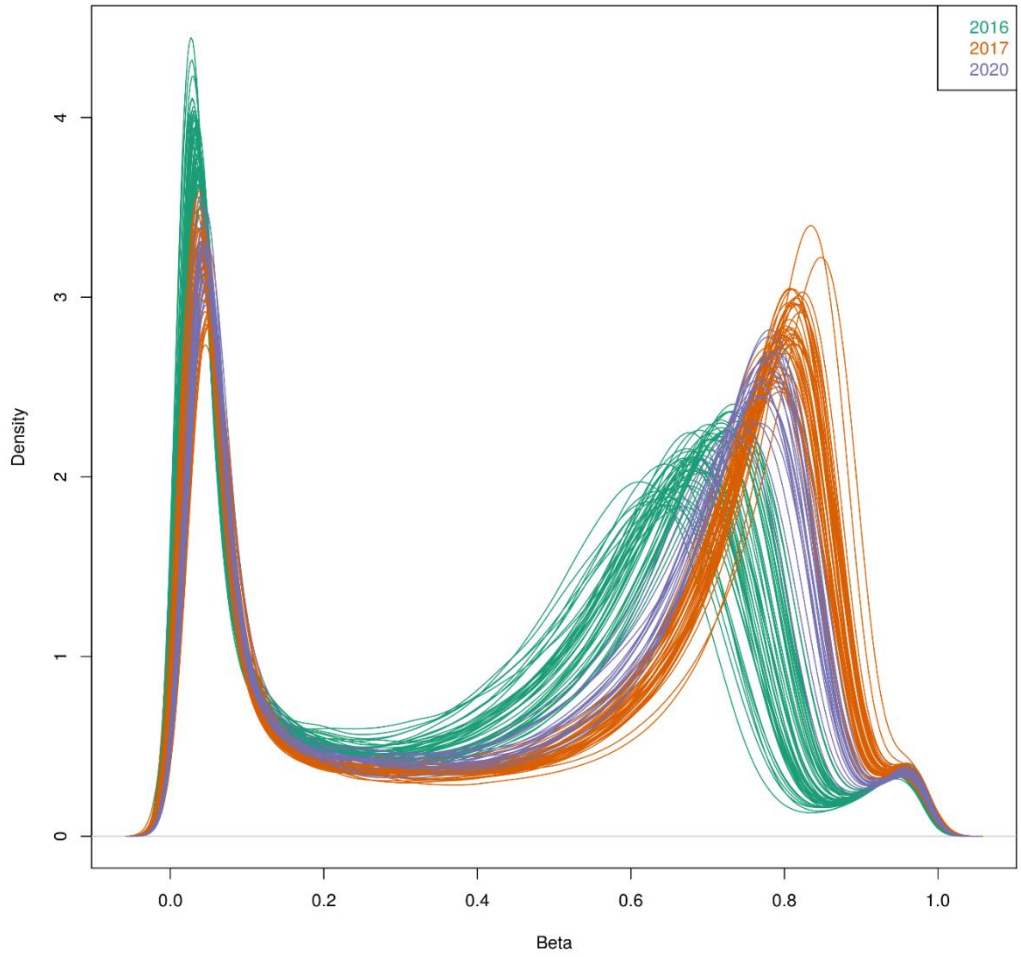


Figure 6.1 The raw density distributions plotted by year of Illumina EPIC array measurement. 2020 samples are shown in purple, and are distributed between the 2016 (green) and 2017 (orange) samples previously collected and analysed in Chapter 2.

Normalisation was then trialled with a range of different pre-processing tools. In support of our findings from Chapter 2, the density plot produced using the pre-processing tool noob indicates that it has successfully normalised the data to correct for the batch effect (Figure 6.2).

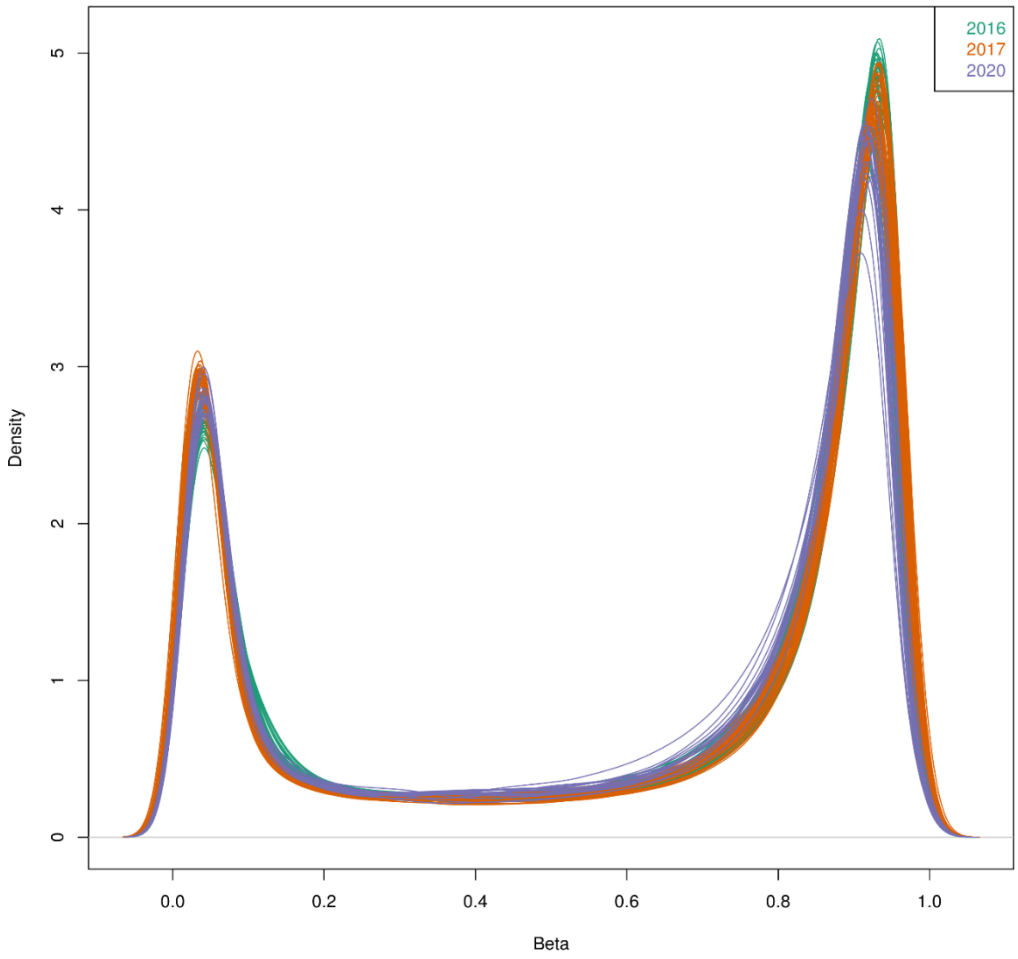
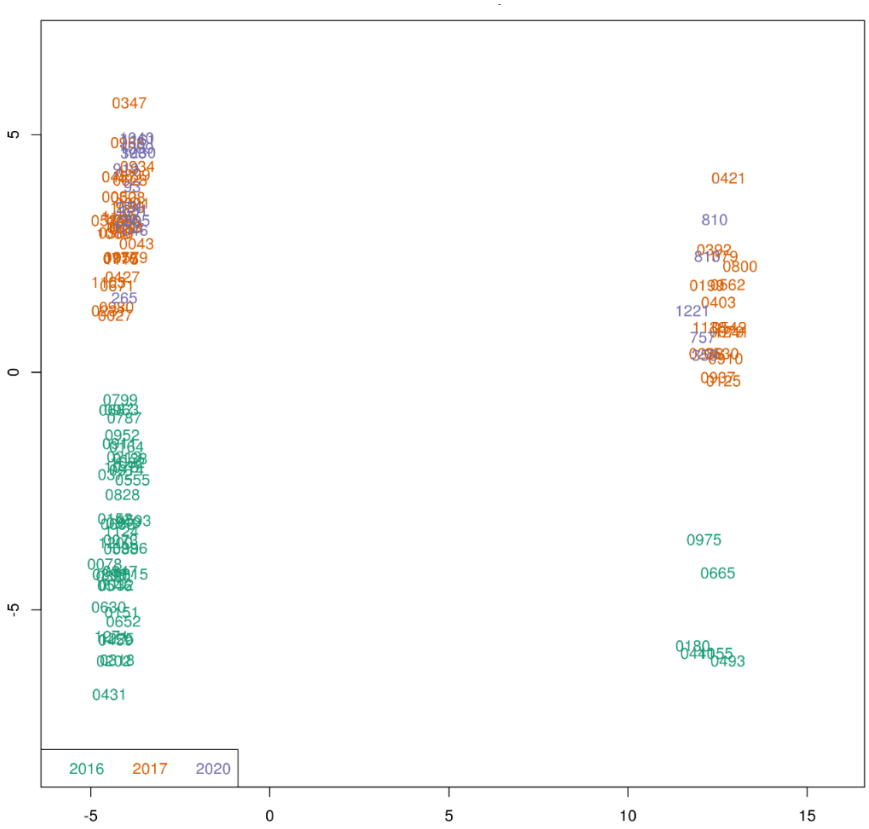


Figure 6.2 Beta density distributions by year of the Illumina EPIC array samples measured by year after using the pre-processing method of noob normalisation.

Multidimensional scaling plots were then produced to assess the 5000 most variable probes, and was plotted for each of the individuals in the study. Figure 6.3 A) shows the raw data, B) is the most variable plots following pre-processing using noob.

A



B

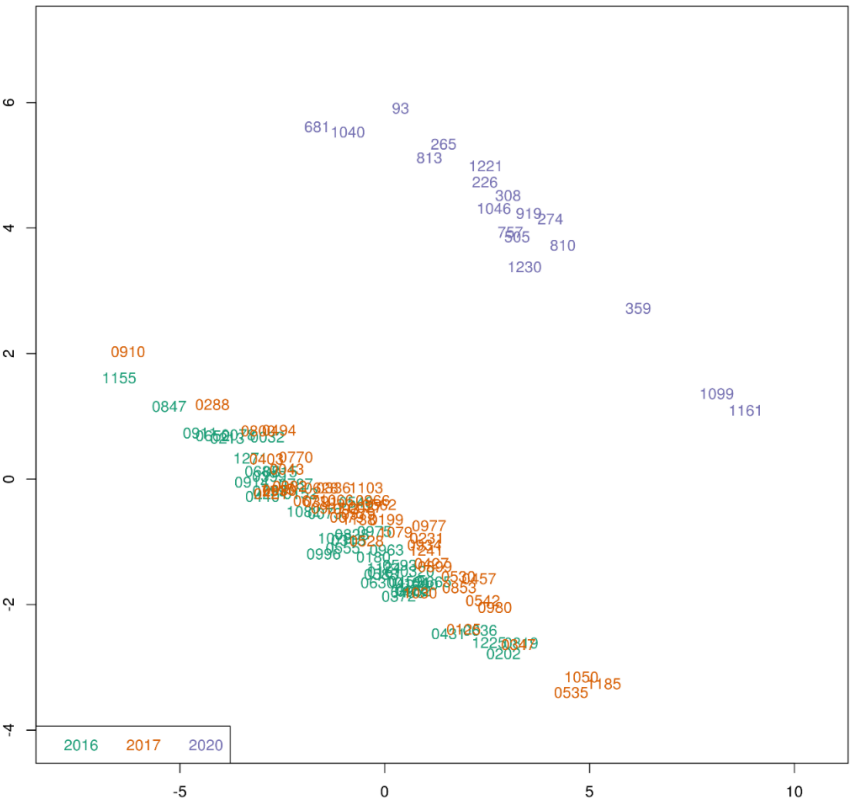


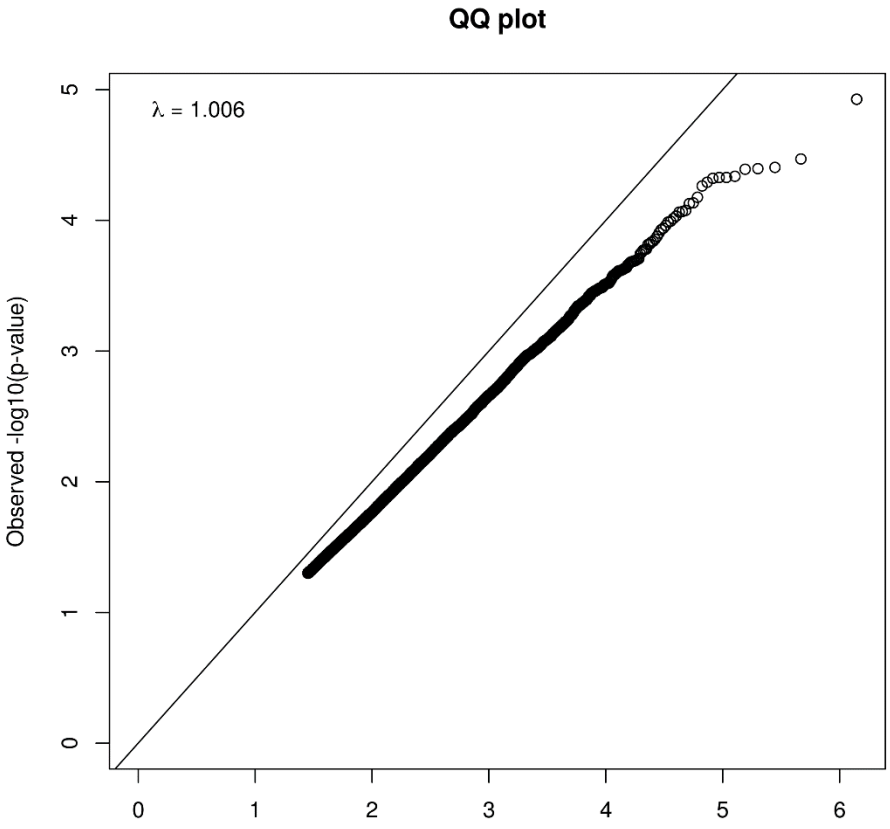
Figure 6.3: Multidimensional scaling plots of the 5000 most variable CpG positions analysed A) the raw data non-normalised and B) post normalisation-using noob. Each of the plots display the samples, which were analysed by year.

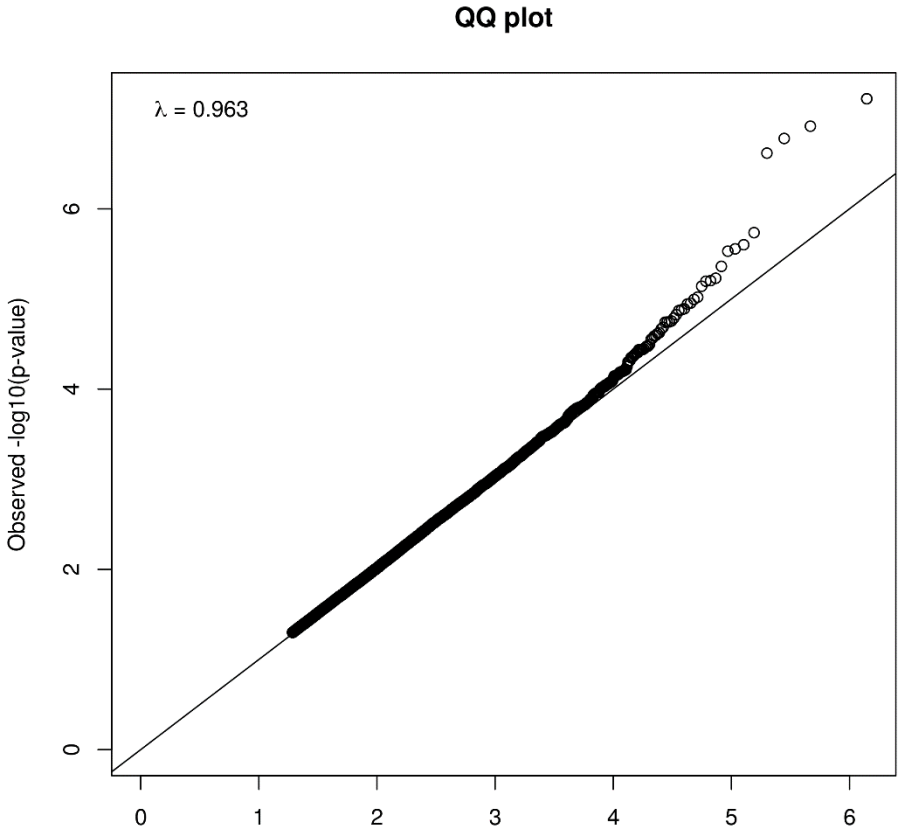
The two plots are grouped by colour depending on the year the samples were analysed. In A) the prenormalisation data displayed MDS plot, displays a similarity between the 2020 and the 2017 samples with the 2016 samples less aligned.

Following noob normalisation, the batch effect between the 2016 and the 2017 samples is corrected, however the same was not seen for the 2020 samples. The same difference between the 2016, 2017 groupings and 2020 samples was found when using both Illumina and swan pre-processing methods (data not shown). It was concluded that the batch effect between these samples was unable to be adjusted for with any of the pre-processing methods available. Thus, it was decided to include year of sampling in the model as the best way to adequately adjust for this. Therefore, the following analyses are based on noob normalisation, but with the addition of the year of sampling variable in all models, to ensure adjustment for a batch effect.

6.3.2 Hierarchical clustering

A number of models were fitted to each of the variables of interest with differing levels of covariates (Table 6.3): baseline models (1,4,7,10) included year samples were analysed to normalise batch effect and four principal components, models 2,5 and 8 included these covariates with addition of adult smoking status, and models 3, 6 and 9 included all of the above covariates with the addition of sex and adult cannabis use status. Each of models we applied were visualised as a Q-Q plots, with chosen models presented in Figure 6.4 and non-used models in Supplementary Figure 6.1.


QQ plot
λ = 1.006
Observed -log10(p-value)
0
1
2
3
4
5
6


QQ plot
λ = 0.963
Observed -log10(p-value)
0
2
4
6
0
1
2
3
4
5
6

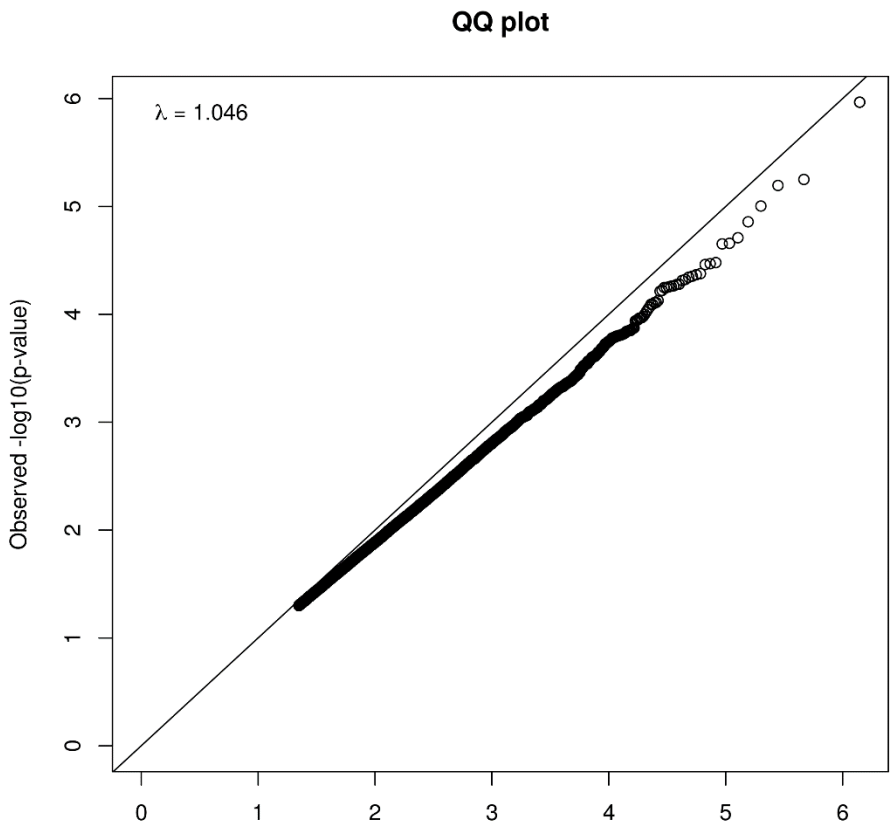


Figure 6.4 Q-Q plots of each of the chosen models which will be discussed in depth throughout the chapter. A) model 1- CP low vs CP high, B) model 4- maternal tobacco exposure vs controls, C) model-7 maternal tobacco exposure and the interaction of CP.

The models used to generate Q-Q plots were calculated with lambda values to infer how much inflation each model was producing (Table 6.4). These were then compared to the tobacco smoking (model 10), λ=0.818, for validation purposes. Based on Q-Q value, it was determined that the baseline models 1, 4 and 7 were the best to carry forward with in our analysis.

Table 6.4 All models fitted to the data set based upon the variable or variables of interest and their associated lamba values. n.b bolded are the models which were used to generate the results for the rest of this chapter.

Variable	Model (Table 6.3)	Lamba (Q-Q value)	How many CpG significant post *P* value adjustments?
CP	**Model 1**	**1.006**	**0**
CP	Model 2	0.936	0
CP	Model 3	0.953	0
IU	**Model 4**	**0.963**	**4**
IU	Model 5	0.957	0
IU	Model 6	0.956	0
IU:CP	**Model 7**	**1.046**	**0**
IU:CP	Model 8	1.031	0
IU:CP	Model 9	1.170	0
Adult tobacco smoking status	Model 10	0.818	2

Following the fitting of the above models (Table 6.4), we assessed the robustness of our estimates by comparing data of differential DNA methylation to our previous EPIC array study. To do this we generated top differentially methylated CpG sites in response to adult tobacco smoking status (model 10) (Supplementary Table 6.1) and then directly compared those tables to tobacco cannabis top tables (Table 2.5 in Chapter 2) . The top two differentially methylated sites in Supplementary Table 6.1 were *AHRR* (cg05575921) and *F2RL3* (cg03636183), which both remained significant at the genome-wide level after P value adjustment. These two sites were ranked in Chapter 2 (Table 2.5) as the 1st and 4th most significantly differentially methylated CpG sites in that analysis. The agreement between analyses provides evidence that the addition of year into the model has combatted the batch effect displayed in Figure 6.3.

6.3.3 Genome wide alterations from *in utero* tobacco exposure on offspring

Following hierarchical clustering, model 4 (Table 6.3) was chosen to be fitted to the data. Results of this analysis identified significant differential DNA methylation between individuals exposed to tobacco *in utero* compared to the non-exposed control group.

Top tables of the most significant CpG sites were then constructed; the top 10 CpG sites are displayed in Table 6.5. Within this group, the top four CpG sites were found to display *P* value significance following BH adjustment. These four sites resided in three genes, *MYOG1* (7.4%) two sites in *FRMD4A* (5.1 and 4.6%) and *RTN1* (3.1%). At all four of these CpG sites methylation differences decreased, showing hyopmethylation in response to *in utero* tobacco exposure. The most differentially methylated site in *MYOG1* (cg04180046) has a methylation difference of 7.4% in the *in utero* exposed group compared to the non-exposed controls (Table 6.5 and Figure 6.5). The same trend is observed in the other remaining top CpG sites with a tendency towards hypomethylation in the exposed group. The observation is further illustrated by scatter plots of the top four most significant CpG sites in Figure 6.4. These three of these four sites have all previously been shown to be hypermethylated in response to *in utero* tobacco exposure.

Pathway analysis was then conducted on nominal *P* value significant CpG sites of less than 0.01 within genes. Table 6.6 displays the pathways that were found to be significant after BH adjustment, with a total of 39 pathways shown to be enriched within the differentially methylated CpG sites. Analysis of these pathway showed that a total of 14 of the 39 pathways played specific roles in signalling function. Other pathway implications include hormone related functions, cancer and immune response.

Table 6.5 Top 10 most differentially methylated CpG sites in response to *in utero* maternal tobacco exposure in offspring, Beta values with P values, nominal and adjusted by the Benjamini and Hochberg method. Locations are relative to hg19 with gene names for overlapping genes or nearest 5' gene with distance to the 5' end shown. Missing UCSC locations are from new probes on the EPIC array, which have not yet been included in the UCSC annotation tracks..

Illumina ID	Gene	Chr	Position in gene	UCSC Location	*in utero* exposed	Non-exposed	β difference	*Log FC*	*Nominal P value*	*Adjusted P value*
cg04180046	*MYO1G*	7	Body	chr7:45002111-45002845	0.587	0.512	0.074	0.072	6.01E-08	0.038
cg15507334	*FRMD4A*	10	TSS200	chr10:14372914 -14372914	0.648	0.597	0.051	0.047	1.21E-07	0.038
cg01604380	*RTN1*	14	Body	chr14:60336951-60337461	0.249	0.217	0.031	0.030	1.66E-07	0.038
cg25464840	*FRMD4A*	10	TSS200	chr10:14372911-14372911	0.755	0.708	0.046	0.043	2.41E-07	0.042
cg06284231	*CLEC14A*	14	1st Exon	chr14:38724254-38725537	0.183	0.160	0.022	0.021	1.84E-06	0.257
cg05009104	*MYO1G*	7	Body	chr7:45002111-45002845	0.798	0.740	0.057	0.054	2.50E-06	0.258
cg15433297	*PKHD1L1*	8	1st Exon	chr8:110374552-110374793	0.345	0.322	0.023	0.023	2.78E-06	0.258
cg12282552	*LTBP3*	11	Body	chr11:65321225-65321823	0.553	0.514	0.039	0.041	2.95E-06	0.258
cg06671242	*PRSS23*	11	Body		0.836	0.804	0.031	0.031	4.35E-06	0.338
cg11866719	*RTN1*	14	Body	chr14:60336951-60337461	0.528	0.481	0.046	0.04	5.89E-06	0.370

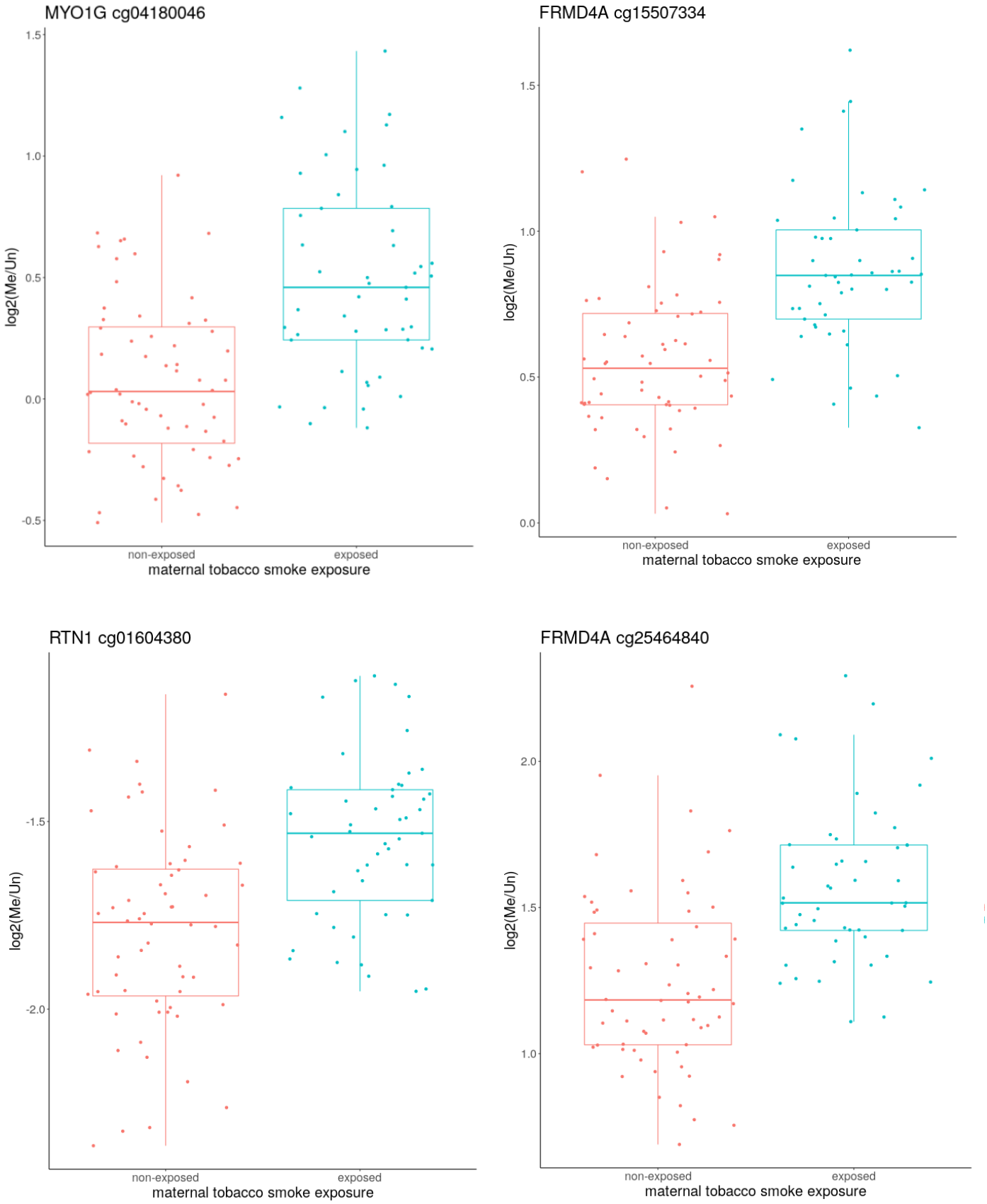


Figure 6.5 The top four CpG sites differentially methylated due to *in utero* maternal tobacco exposure, these sites resided in genes *MYO1G, RTN1* and two sites in *FRMD4A*.

Table 6.6 KEGG pathway analysis on nominally significant (P < 0.01) CpG sites differentially methylated between individuals exposed to tobacco *in utero* and non-exposed individuals. All KEGG pathways included here have significant adjusted P values (adjusted P < 0.05).

Name	*P value*	*Adjusted P value*	Odds Ratio	Combined score
Focal adhesion	4.30E-07	0.0001	1.81	26.57
ErbB signalling pathway	5.13E-06	0.0005	2.15	26.23
Glutamatergic synapse	3.92E-06	0.0006	2.00	24.84
Phospholipase D signalling pathway	7.09E-06	0.0005	1.84	21.78
Axon guidance	7.59E-06	0.0004	1.75	20.60
Vibrio cholerae infection	0.0002	0.0057	2.22	18.65
B cell receptor signalling pathway	0.0001	0.0048	2.03	17.81
Insulin secretion	0.0001	0.0048	1.94	17.17
Rap1 signalling pathway	3.09E-05	0.0015	1.64	17.06
Calcium signalling pathway	5.34E-05	0.0023	1.65	16.26
T cell receptor signalling pathway	0.000276	0.0060	1.81	14.85
MAPK signalling pathway	6.59E-05	0.0025	1.5	14.49
Wnt signalling pathway	0.0002	0.0059	1.65	13.96
Neurotrophin signalling pathway	0.0003	0.0071	1.72	13.62
Cushing syndrome	0.0002	0.0061	1.65	13.59
Fc gamma R-mediated phagocytosis	0.0009	0.0160	1.77	12.24
Choline metabolism in cancer	0.0009	0.0160	1.74	12.12
Cortisol synthesis and secretion	0.0016	0.0193	1.88	12.05
Ras signalling pathway	0.0003	0.0073	1.51	11.96
Oxytocin signalling pathway	0.0007	0.0132	1.00	11.52
GnRH signalling pathway	0.0014	0.0185	1.73	11.31
Prostate cancer	0.0013	0.0187	1.72	11.28
Gastric acid secretion	0.0024	0.0258	1.78	10.68
Adherens junction	0.0030	0.0312	1.77	10.27
Type II diabetes mellitus	0.0049	0.0414	1.93	10.23
Thyroid hormone signalling pathway	0.0019	0.0222	1.63	10.14
Insulin resistance	0.0021	0.0237	1.64	10.09
Regulation of actin cytoskeleton	0.0010	0.0168	1.48	10.08
Proteoglycans in cancer	0.0011	0.0172	1.49	10.05
Chemokine signalling pathway	0.0016	0.0199	1.49	9.57
Aldosterone synthesis and secretion	0.0034	0.0343	1.64	9.31
Colorectal cancer	0.0039	0.0382	1.68	9.27
Cholinergic synapse	0.0040	0.0375	1.59	8.74
HIF-1 signalling pathway	0.0047	0.0418	1.61	8.61
Inflammatory mediator regulation of TRP channels	0.0047	0.0406	1.61	8.61
Amoebiasis	0.0050	0.0406	1.62	8.57
Pathways in cancer	0.0013	0.0187	1.29	8.51
Gap junction	0.0055	0.0417	1.64	8.51
Arrhythmogenic right ventricular cardiomyopathy (ARVC)	0.0066	0.0457	1.70	8.49

6.3.4 DNA methylation analysis of low CP compared to high CP scored individuals

Model 1 was fitted to the normalised array data to investigate DNA methylation analysis on low CP compared to high CP scored individuals. Here, no adjusted P value significance was seen at any of the CpG sites (Table 6.7), however nominal significance was observed. Top tables were generated for the 10 most nominally significant CpG sites (Table 6.7). Two CpG sites in the top 10 most nominally significant sites resided within the same gene, *PDE9A.* There are two CpG sites that have no known gene association (cg06632577 and cg17695791) and we present these as novel findings. The top three nominally significant loci display hypomethylation in the high CP group compared to the low CP group (Table 6.7 and Figure 6.6). In comparison to the tobacco exposed *in utero* top 10 CpG sites (Table 6.5), the methylation differences were smaller, with all CpG sites in the top 10 displaying less than 1.7% differential DNA methylation.

The top four most nominally significant CpG sites are plotted in Figure 6.6. One CpG site, cg20474266, which resides in the gene *STEAP1B,* displayed an increase in methylation in the high CP group. The top four observed CpG sites show a consistent pattern of variability in the range of methylation values for each of the individuals.

KEGG pathway analysis was carried out on CpG sites within genes (or annotated to the nearest gene), which displayed nominal P value significance of less than 0.01 to assess for KEGG pathway enrichment. A total of 48 pathways were identified as enriched in the comparison of high CP vs. low CP (Table 6.8). Within the pathways that showed adjusted P value significance (N= 10), five have specific roles in the brain (cholinergic synapse, axon guidance, cGMP-PKG signalling pathway, GABAergic synapse and glutamatergic synapse). One pathway, calcium signalling pathway, plays multiple roles in muscle contraction, neuron signalling and fertilisation. The remaining three pathways (adrenergic signalling in cardiomyocytes, insulin secretion and type II diabetes mellitus) all rely upon calcium signalling processes for proper function.

Table 6.7 Top 10 differentially methylated CpG sites in response to low CP compared to high CP. Beta values with P values, nominal and adjusted by the Benjamini and Hochberg method. Locations are relative to hg19 with gene names for overlapping genes or nearest 5' gene with distance to the 5' end shown. Missing locations are from new probes, which have not been properly annotated.

Iluminia ID	Gene	Chr	Position in gene	Location	Low CP	High CP	β difference	*Log FC*	*Nominal P value*	*Adjusted P value*
cg20218460	*LRRFIP1*	2	Body	chr2:238583504-238583504	0.039	0.038	-0.001	-0.006	1.18E-05	0.992
cg05064509	*EYA2*	20	Body		0.947	0.935	-0.011	-0.008	3.39E-05	0.992
cg06632577		10			0.948	0.936	-0.012	-0.008	3.93E-05	0.992
cg20474266	*STEAP1B*	7	Body		0.055	0.069	0.014	0.014	4.02E-05	0.992
cg11570752	*PDE9A*	21	Body		0.099	0.151	0.051	0.077	4.06E-05	0.992
cg15495039	*BAG4*	8	TSS1500	chr8:38033408-38034643	0.037	0.043	0.005	0.008	4.60E-05	0.992
cg01036746	*C14orf37*	14	TSS1500	chr14:58618291-58619220	0.062	0.073	0.011	0.009	4.69E-05	0.992
cg04916741	*PDE9A*	21	5'UTR		0.941	0.926	-0.014	-0.009	4.70E-05	0.992
cg17695791		11		chr11:91913958-91913958	0.862	0.844	-0.017	-0.024	4.77E-05	0.992
cg03935183	*GATA3*	10	Body	chr10:8100384-8100768	0.945	0.932	-0.012	-0.010	5.11E-05	0.992

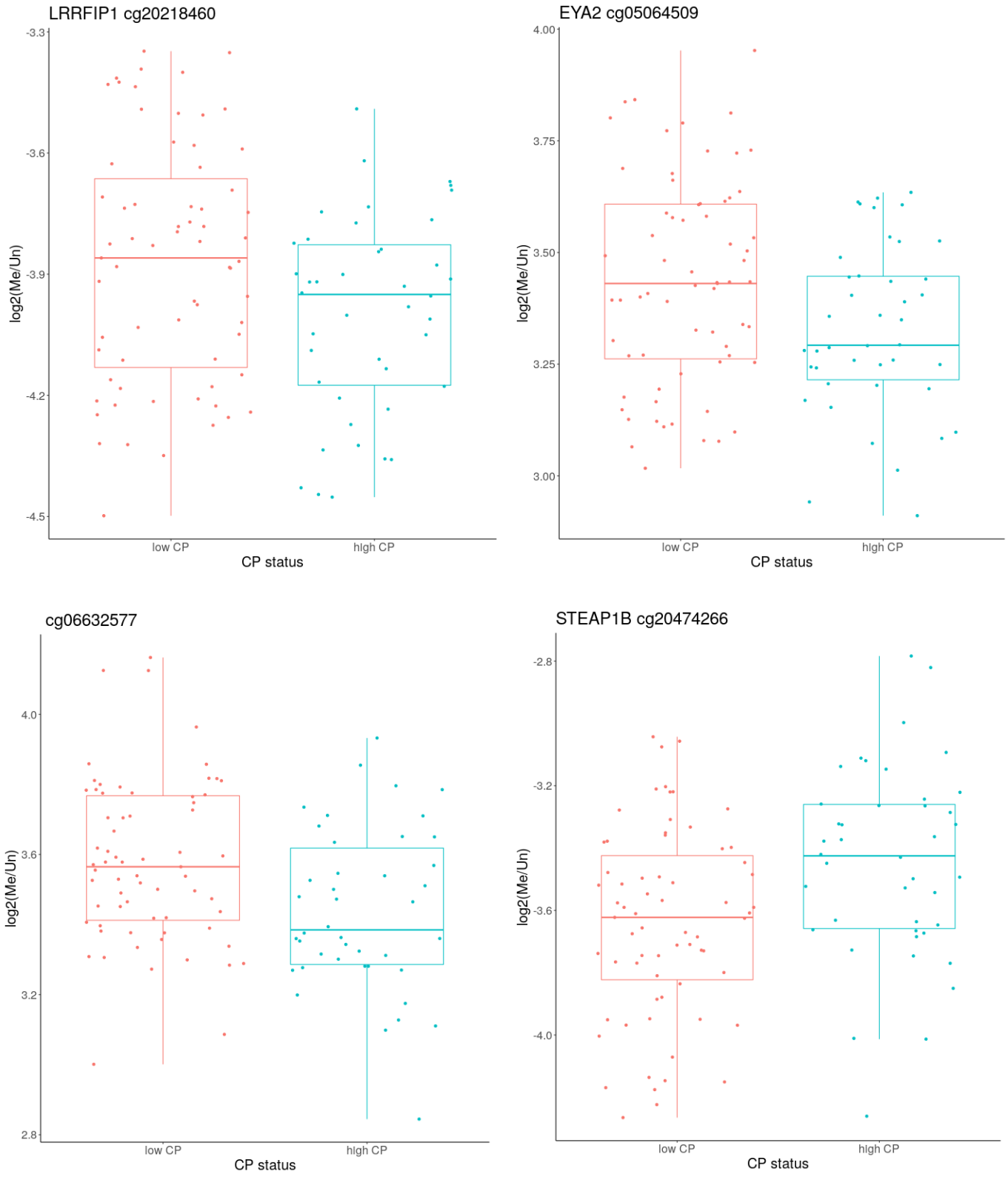


Figure 6.6 The top four CpG sites differentially methylated between high CP and low CP scored individuals. CpG sites resided in genes *LRRFIP1, EYA2* and *STEAP1B*, and one site, cg06632577 had no known gene association.

Table 6.8 KEGG pathway analysis of CpG sites differentially methylated (nominal P < 0.01) in low CP vs high CP.

Name	*P value*	*Adjusted P value*	Odds Ratio	Combined score
Cholinergic synapse	3.23E-06	0.0004	2.36	29.86
Axon guidance	1.65E-06	0.0005	2.07	27.62
Type II diabetes mellitus	0.0001	0.009	2.78	24.47
cGMP-PKG signalling pathway	1.48E-05	0.001	2.00	22.3
Calcium signalling pathway	2.6E-05	0.001	1.91	20.13
GABAergic synapse	0.0001	0.009	2.21	19.01
Insulin secretion	0.0007	0.029	2.08	14.95
Adrenergic signalling in cardiomyocytes	0.0006	0.027	1.82	13.46
Gastric acid secretion	0.002	0.060	2.05	12.56
Glutamatergic synapse	0.001	0.042	1.87	12.33
Parathyroid hormone synthesis, secretion and action	0.002	0.057	1.85	11.15
Platelet activation	0.002	0.056	1.79	10.95
Thyroid hormone synthesis	0.004	0.088	1.96	10.55
Morphine addiction	0.003	0.079	1.88	10.42
PI3K-Akt signalling pathway	0.001	0.043	1.47	9.79
Progesterone-mediated oocyte maturation	0.004	0.087	1.81	9.65
Long-term depression	0.008	0.122	1.99	9.52
Hippo signalling pathway	0.003	0.071	1.65	9.46
AMPK signalling pathway	0.005	0.101	1.71	8.74
beta-Alanine metabolism	0.023	0.194	2.20	8.27
Aldosterone synthesis and secretion	0.009	0.127	1.74	8.18
Aldosterone-regulated sodium reabsorption	0.024	0.196	2.08	7.72
Dopaminergic synapse	0.009	0.124	1.63	7.61
Ras signalling pathway	0.007	0.121	1.47	7.12
TGF-beta signalling pathway	0.015	0.173	1.71	7.08
Oxytocin signalling pathway	0.010	0.135	1.56	7.05
MAPK signalling pathway	0.007	0.120	1.42	6.95
Circadian entrainment	0.016	0.170	1.67	6.85
Viral carcinogenesis	0.010	0.136	1.49	6.75
C-type lectin receptor signalling pathway	0.017	0.171	1.64	6.66
Insulin signalling pathway	0.016	0.171	1.56	6.43
Choline metabolism in cancer	0.020	0.195	1.64	6.38
Phosphatidylinositol signalling system	0.020	0.189	1.64	6.38
cAMP signalling pathway	0.014	0.167	1.45	6.17
Focal adhesion	0.015	0.175	1.46	6.09
Serotonergic synapse	0.021	0.194	1.59	6.09
Cysteine and methionine metabolism	0.042	0.254	1.82	5.71
Phospholipase D signalling pathway	0.022	0.198	1.50	5.68
Retrograde endocannabinoid signalling	0.022	0.193	1.50	5.68
Thyroid hormone signalling pathway	0.027	0.215	1.54	5.53
TNF signalling pathway	0.030	0.220	1.55	5.43
Adherens junction	0.038	0.245	1.66	5.41
Relaxin signalling pathway	0.028	0.213	1.51	5.38
Fc gamma R-mediated phagocytosis	0.034	0.234	1.59	5.38
Pancreatic secretion	0.034	0.234	1.57	5.26
Th1 and Th2 cell differentiation	0.037	0.246	1.58	5.17
Oocyte meiosis	0.033	0.237	1.50	5.11
Glycerolipid metabolism	0.048	0.272	1.68	5.08

6.3.5 *In utero* tobacco exposure and the interaction with CP

Model 7 was used to investigate genome-wide differential methylation specific to the interaction between *in utero* tobacco exposure and CP score (high CP and low CP). Top tables of the most significantly differentially methylated CpG sites were constructed, with the top 10 CpG sites displayed in Table 6.9. No CpG sites were found to be significant after adjustment for multiple testing. Within the top 10 CpG sites there is a clear bias towards hypomethylation, with nine CpG sites displaying a decrease in methylation. The differences between the b0 and the b1 variables under the interaction are all minor, with less than 1.2%. By comparison, differential methylation between individuals exposed to tobacco *in utero*, and unexposed controls (Table 6.5) were all greater than 2.2%.

The top four most differentially methylated sites under this interaction are plotted in Figure 6.7. Scatter plots, overlaid with box are fitted to the four sub categories, non-exposed low CP, non-exposed high CP, *in utero* exposed low CP and *in utero* exposed high CP. Each of the four sub categories here has a range of methylation values for the individuals within the group. The main difference is that of the non-exposed high CP vs the exposed high CP, which are shown to be differentially methylated in each of the four CpG sites displayed.

KEGG pathway analysis was conducted on the nominally significant CpG sites ($P < 0.01$) within gene or near to genes (Table 6.10). One KEGG pathway, small lung cancer, reached an adjusted P value of significance (adjusted $P = 0.0307$). There were 19 other pathways that were nominally significantly enriched under the interaction model.

Table 6.9 Differentially methylation CpG sites between *in utero* maternal tobacco exposure and the interaction with CP Beta values with P values, nominal and adjusted by the Benjamini and Hochberg method. Locations are relative to hg19 with gene names for overlapping genes or nearest 5' gene with distance to the 5' end shown. Missing locations are from new probes, which have not been properly annotated.

Illumina ID	Gene	Chr	Position in gene	Location	*In utero* maternal tobacco exposed low CP	*In utero* maternal tobacco exposed high CP	β difference	*Log FC*	*Nominal P value*	*Adjusted P value*
cg13339919	*SLC10A7*	4	Body		0.947	0.943	-0.003	-0.014	1.08E-06	0.755
cg01394525	*LAMC3*	9	Body	chr9:133901745-133901956	0.912	0.912	-0.0004	-0.021	5.62E-06	0.974
cg13787134	*PHF2*	9	Body	chr9:96362103-96362103	0.941	0.935	-0.005	-0.017	6.39E-06	0.974
cg12163448	*FASTKD1*	2	TSS200	chr2:170430473-170430473	0.136	0.166	0.030	0.0758	9.91E-06	0.974
cg17343033		5			0.886	0.872	-0.014	0.031	1.39E-05	0.974
cg09125477	*C16orf91*	16	Body	chr16:1470502-1471164	0.106	0.105	-0.0003	-0.034	1.95E-05	0.974
cg24835473	*CHCHD6*	3	Body		0.933	0.923	-0.009	-0.015	2.20E-05	0.974
cg25849390	*CCT6A*	7	Body	chr7:56131778-56132226	0.887	0.874	-0.012	-0.033	2.22E-05	0.974
cg02809796	*PCGF3*	4	Body	chr4:752029-753788	0.920	0.915	-0.005	-0.020	3.31E-05	0.974
cg13353442	*CDC27*	17	Body		0.920	0.909	-0.011	-0.022	3.39E-05	0.974

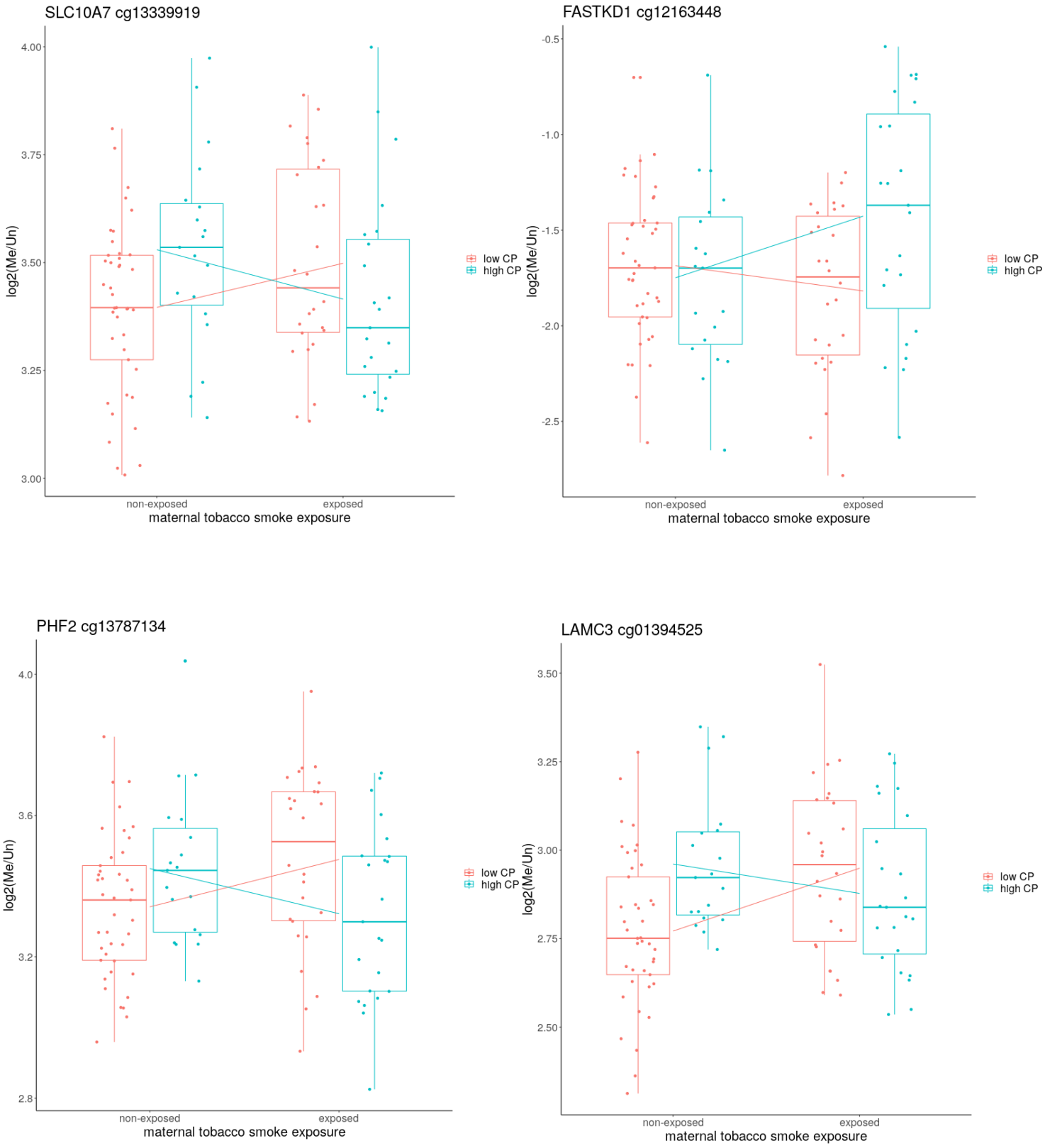


Figure 6.7 The top four most significantly differentially methylated (nominal P < 0.01) CpG sites when *in utero* tobacco exposure was assessed with the interaction CP. Non-exposed individuals are plotted on the left of each plot, colour coded for either low CP (salmon) or high CP (cyan), with exposed individuals on the right. Lines from the non-exposed group to the exposed group represent the median methylation between non-exposed and exposed with (salmon) and without (cyan) CP.

Table 6.10 KEGG pathway analysis on CpG sites that are nominally significantly differentially methylated in the interaction between individuals exposed to tobacco *in utero* tobacco and high CP.

Name	*P value*	*Adjusted P value*	Odds Ratio	Combined score
Small cell lung cancer	0.0001	0.037	1.97	17.76
Morphine addiction	0.001	0.114	1.81	12.31
Arrhythmogenic right ventricular cardiomyopathy (ARVC)	0.002	0.125	1.86	11.55
Longevity regulating pathway	0.001	0.119	1.73	11.21
Regulation of actin cytoskeleton	0.001	0.161	1.51	10.37
Parathyroid hormone synthesis, secretion and action	0.002	0.128	1.67	9.74
Chronic myeloid leukaemia	0.004	0.131	1.77	9.63
Dilated cardiomyopathy (DCM)	0.005	0.113	1.68	8.82
Fructose and mannose metabolism	0.013	0.167	2.03	8.82
cAMP signalling pathway	0.002	0.128	1.47	8.79
Glycosaminoglycan biosynthesis	0.008	0.137	1.84	8.68
Autophagy	0.004	0.11	1.57	8.53
Focal adhesion	0.003	0.119	1.47	8.50
Prostate cancer	0.006	0.120	1.63	8.30
AMPK signalling pathway	0.005	0.112	1.58	8.20
Proteoglycans in cancer	0.003	0.130	1.46	8.11
GABAergic synapse	0.007	0.132	1.64	7.99
Mannose type O-glycan biosynthesis	0.025	0.234	2.12	7.82
Endocytosis	0.004	0.111	1.40	7.50
MAPK signalling pathway	0.004	0.122	1.36	7.41

6.3.6 Overall CpG differential methylation between exposure models

Globally the total number of differentially methylated sites with a nominal *P* value of less than 0.01 were scored from each of the model variations assessed (Model 1, 4 7 and 10, Table 6.11). Model 4 (exposure to tobacco *in utero* vs. non-exposed controls) produced 7228 differentially methylated CpG sites, compared to models 4, 7 and 10, which all detected fewer differentially methylated CpG sites.

Table 6.11 Overall genome-wide nominally significantly differentially methylated CpG sites for each of the variables assessed. Adult smoking status (model 10), tobacco exposure *in utero* (model 4), Conduct problems (model 1) and *in utero* maternal tobacco exposure and the interaction of CP (model 7).

	Adult smoking status (model 10)	*in utero* maternal tobacco exposed (model 4)	Conduct problems (model 1)	*in utero* maternal tobacco exposure and the interaction with CP (model 7)
Number of differentially methylated CpG sites (*P* <0.01)	4745	7228	3858	5226

6.3.7 Assessing differential DNA methylated regions within genes in individuals exposed to tobacco *in utero,* compared to non-exposed controls

Data from models, 1, 4 and 7 were then used to assess differentially methylated regions. We defined a differentially methylated region to have the following, a gene containing nominally significant P values of less than 0.01 at greater that had five CpG sites. There was a threshold cut-off for region size based on the frequency of genes that had more than five differentially methylated CpG sites.

Model 4, tobacco exposure *in utero* displayed a large number of differentially methylated CpG sites. Due to this, differentially methylated regions in genes containing seven or more CpG sites are displayed in Table 6.12. Genes containing either five or six CpG sites are displayed in appendices (Supplementary Table 6.3).

Across these regions, a trending pattern of hypermethylation is seen, with eight of the genes displayed in Table 6.12 all showing hypermethylation at each cg site in the differentially methylated region. Functional analysis of the genes for which we detect differentially methylated regions show a vast majority of brain development genes or known brain related diseases.

Table 6.12 Genes that contain seven or more differentially methylated CpG sites, here defined as differentially methylated regions, found between individuals exposed to *in utero* maternal tobacco compared to non-exposed individuals.

Gene	Illumina ID	Correlation	Location	CpG island	Functional association
BCL11B	cg23479730	hyper	chr14:99681758-99681758	Body	Neurodevelopment [12, 13]
	cg13987489	hyper	chr14:99664107-99664107	Body	
	cg08129129	hyper	chr14:99711839-99713431	Body	
	cg04162647	hyper	chr14:99736040-99737584	Body	
	cg15530474	hyper		Body	
	cg12737475	hyper		Body	
	cg03205581	hyper		Body	
	cg26791805	hyper		Body	
CAMTA1	cg09068636	hypo	chr1:7764593-7765856	Body	Neurobehavioral phenotypes [14]
	cg08647349	hyper	chr1:7439692-7,439692	Body	
	cg00452133	hyper	chr1:7308117-7308117	Body	
	cg10349142	hypo		Body	
	cg09914736	hypo		Body	
	cg15063687	hyper	chr1:7,359,621-7,359,621	Body	
	cg12710648	hyper	chr1:7,308092-7308092	Body	
	cg26669159	hyper	chr1:6,967037-6967037	Body	
	cg11294564	hypo		Body	
	cg05233894	hypo		Body	
CASZ1	cg02396224	hyper	chr1:10698299-10698910	3'UTR	Cardiac development [15]
	cg22849913	hypo	chr1:10702136-10702340	Body	
	cg24661860	hyper	chr1:10,784363-10784363	5'UTR	
	cg22513691	hyper	chr1:10738,664-10738664	Body	
	cg00787856	hyper	chr1:10853894-10856964	TSS1500	
	cg13553158	hyper	chr1:10853894-10856964	5'UTR	
	cg16436377	hyper	chr1:10725187-10725617	Body	
FAM84B	cg08568155	hyper	chr8:127568676-127570873	Body	Prostate cancer [16, 17]
	cg16566518	hyper	chr8:127568676-127570873	Body	
	cg21390512	hyper	chr8:127568676-127570873	Body	
	cg13636698	hyper	chr8:127568676-127570873	Body	
	cg06230848	hyper	chr8:127568676-127570873	Body	
	cg06532751	hyper	chr8:127568676-127570873	TSS1500	
	cg02925049	hyper	chr8:127568676-127570873	Body	
	cg03374695	hyper	chr8:127568676-127570873	3'UTR	
FOXP1	cg10715905	hyper	chr3:71542871-71542871	5'UTR	Autism spectrum disorder [18] Intellectual disability syndrome [19]
	cg07278181	hypo	chr3:71293767-71293767	5'UTR	
	cg11670533	hyper		5'UTR	
	cg07324822	hyper		5'UTR	
	cg14398973	hyper		5'UTR	
	cg21993077	hyper		TSS200	
	cg12423097	hypo		Body	
	cg27419618	hyper		TSS1500	
	cg20642055	hyper		Body	
	cg21458836	hyper		Body	
FRMD4A	cg15507334	hyper	chr10:14372914-14372914	TSS200	Alzheimer's disease [20]
	cg25464840	hyper	chr10:14372911-14372911	TSS200	
	cg11813497	hyper	chr10:14372879-14372879	TSS200	
	cg14630801	hyper		5'UTR	
	cg20344448	hyper	chr10:14372432-14372432	5'UTR	
	cg17538881	hyper		5'UTR	
	cg17808360	hypo		Body	
	cg20643833	hypo		Body	
LTBP3	cg12282552	hyper	chr11:65321225-65321823	Body	Amylogenesis and skeletal development [21]
	cg20588859	hyper	chr11:65321225-65321823	Body	
	cg21547324	hyper	chr11:65321225-65321823	Body	
	cg19332572	hyper	chr11:65321225-65321823	Body	
	cg24716530	hyper	chr11:65321225-65321823	Body	
	cg23743554	hyper	chr11:65321225-65321823	Body	

	cg23272978	hyper	chr11:65314912-65315476	Body	
PCDHGA4	cg04637478	hyper	chr5:140753654-140753952	Body	Brain regulation [22]
	cg21908557	hyper	chr5:140762401-140762768	Body	
	cg07231479	hyper	chr5:140794358-140795045	Body	
	cg10917547	hyper	chr5:140794358-140795045	Body	
	cg07017875	hyper	chr5:140789094-140789762	Body	
	cg22737624	hyper	chr5:140750050-140750264	Body	
	cg12145907	hyper	chr5:140864527-140864748	Body	
PRDM16	cg18240463	hyper	chr1:3163969-3164643	Body	Angiogenesis [23]
	cg09990962	hyper	chr1:3163969-3164643	Body	neural stem and
	cg12648819	hyper		Body	neuronal cell
	cg08262220	hyper	chr1:2997272-2997473	Body	maintenance [24,
	cg12701603	hyper		Body	25]
	cg19317333	hyper	chr1:3321269-3322310	Body	
	cg15156029	hyper		Body	
PRRT1	cg11617964	hyper	chr6:32118101-32118544	Body	Synapse function
	cg15194163	hyper	chr6:32118101-32118544	Body	and development
	cg12320039	hyper	chr6:32118101-32118544	Body	[26, 27]
	cg05764839	hyper	chr6:32118101-32118544	Body	
	cg21398794	hyper	chr6:32118101-32118544	Body	
	cg25845985	hyper	chr6:32118101-32118544	Body	
	cg19227031	hyper	chr6:32118101-32118544	Body	
	cg22268510	hyper	chr6:32118101-32118544	Body	
SH2B2	cg17190891	hyper	chr7:101961741-101962226	3'UTR	Insulin signalling
	cg07512361	hyper	chr7:101943785-101944557	Body	and glucose
	cg24707573	hyper	chr7:101961741-101962226	Body	metabolism [28]
	cg01723606	hyper	chr7:101943785-101944557	Body	
	cg15355015	hyper	chr7:101943785-101944557	Body	
	cg06785147	hyper	chr7:101943785-101944557	Body	
	cg05302531	hyper	chr7:101936317-101936548	Body	
VENTX	cg12554483	hyper	chr10:135048797-135052077	TSS200	Regulation of
	cg22165685	hyper	chr10:135048797-135052077	TSS200	dendritic cells [29]
	cg04665423	hyper	chr10:135048797-135052077	TSS200	acute myeloid
	cg23845574	hyper	chr10:135048797-135052077	TSS200	leukaemia [30, 31]
	cg14539179	hyper	chr10:135048797-135052077	TSS1500	
	cg18727936	hyper	chr10:135048797-135052077	TSS1500	
	cg02645368	hyper	chr10:135048797-135052077	TSS200	
	cg04347264	hyper	chr10:135048797-135052077	TSS200	
	cg06500714	hyper	chr10:135048797-135052077	TSS1500	
	cg12666165	hyper	chr10:135048797-135052077	1s tExon	
	cg07370771	hyper	chr10:135048797-135052077	TSS1500	

6.3.8 Detecting differential DNA methylated regions in individuals with high CP scores compared to low CP scores.

We detected five genes which displayed multiple differentially methylated CpG sites in response to low vs high CP. Four of these gene regions are displayed below in Table 6.13 along with the methylation direction change, location and their functional annotation. The fifth gene region, not displayed in Table 6.13 *PCDHGA4* will be discussed further in section 6.3.9. Functional analysis reveals that two of the four genes with differentially methylated regions are associated with brain pathologies that are relevant to the CP phenotype.

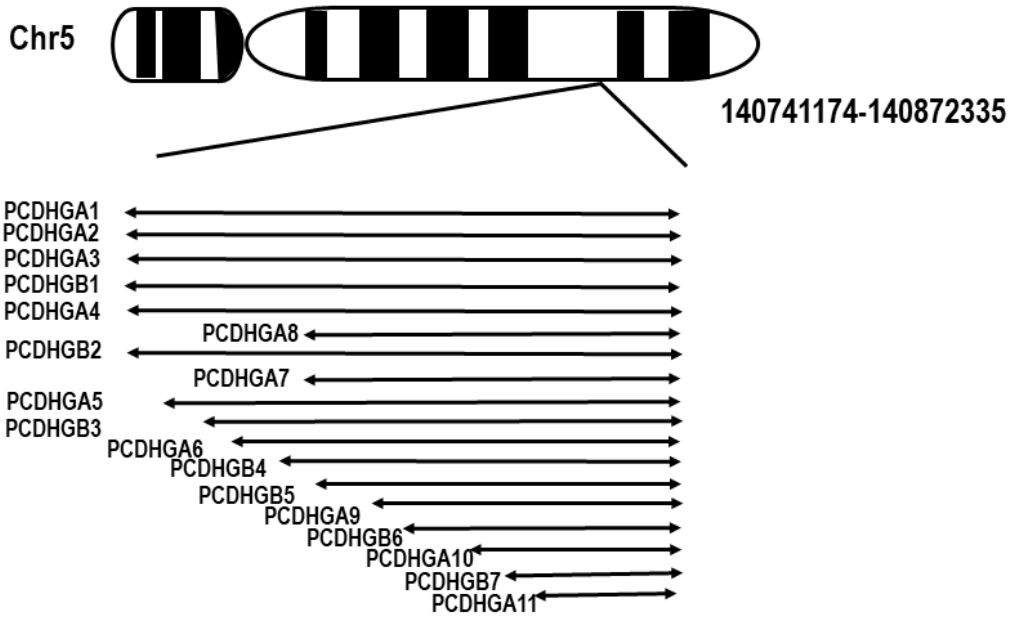
Table 6.13 Genes for which differentially methylated regions were detected between individuals with a low conduct problem score and those with high conduct problem scores.

Gene	Illumina ID	Correlation	Location	CpG island	Functional association
CCKBR	cg25740457	hyper	chr11:6291338-6291558	3'UTR	Anxiety related behaviours [32, 33]
	cg21112490	hyper	chr11:6291338-6291558	Body	
	cg13580265	hyper	chr11:6292256-6292693	Body	
	cg08101193	hyper	chr11:6291338-6291558	Body	
	cg19364351	hyper	chr11:6292256-6292693	Body	
CPT1B	cg08260245	hyper	chr22:51016253-51017020	5'UTR	Post traumatic stress disorder [34]
	cg24363820	hyper	chr22:51016253-51017020	5'UTR	
	cg06530441	hyper	chr22:51016253-51017020	TSS200	
	cg16386697	hyper	chr22:51016253-51017020	TSS200	
	cg00983520	hyper	chr22:51016253-51017020	1st Exon	
	cg27502912	hyper	chr22:51016253-51017020	1st Exon	
	cg00270625	hyper	chr22:51016253-51017020	1st Exon	
	cg17952465	hyper	chr22:51016253-51017020	5'UTR	
DIP2C	cg20684696	hypo	chr10:518192-518471	Body	Breast and lung cancers [35, 36]
	cg10809719	hypo		Body	
	cg16942135	hyper	chr10:711896-712395	Body	
	cg10441401	hyper	chr10:575898-576131	Body	
	cg15030662	hypo	chr10:575898-576131	Body	
PGAM2	cg07075347	hyper	chr7:44104746-44105116	1st Exon	Associated with glycogen storage disease type X [37]
	cg16627090	hyper	chr7:44104746-44105116	1st Exon	
	cg14219560	hyper	chr7:44104746-44105116	Body	
	cg23616741	hyper	chr7:44104746-44105116	TSS1500	
	cg17459793	hyper	chr7:44104746-44105116	Body	
	cg03470754	hyper	chr7:44104746-44105116	1st Exon	

6.3.9 Protocadherin Gamma Subfamily differential methylation between low CP and high CP scored individuals

A large region of the genome was found to be differentially methylated between low CP and high CP scored individuals. The region spans approx. 130,000bp of chromosome 5 (Figure 6.8 A, chr5:140741174-140872335), which is predicted to contain any of the following genes: *PCDHGA4, PCDHGA1, PCDHGA6, PCDHGA5, PCDHGB1, PCDHGA3, PCDHGA2, PCDHGA6, PCDHGB2, PCDHGB3.* The region is poorly annotated and highly repetitive, so a specific gene name is unable to be assigned to this observed region. In this region, 30 CpG sites were differentially methylated (Figure 6.8 B). All sites were hypermethylated and had nominal P values of less than 0.01.

A



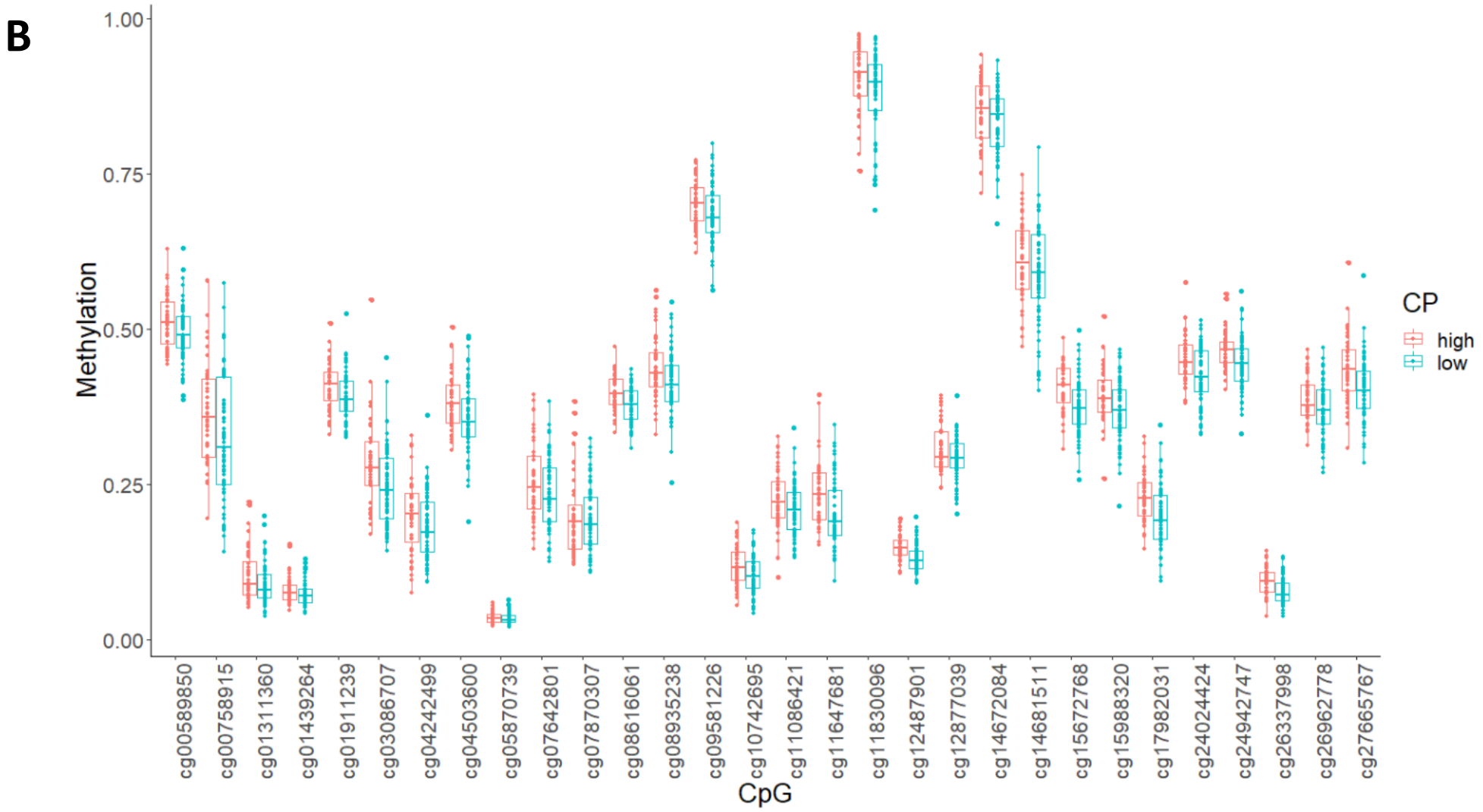


Figure 6.8 Chromosome 5:140741174-140872335, located within the gene, protocadherin gamma displayed consistent DNA methylation differences between low and high CP individuals.

6.3.9 Differentially methylated regions under the interaction of *in utero* tobacco exposure and CP scores

Multiple CpG sites were differentially methylated from the IU:CP interaction in the following six genes: *CUX1, DIP2C, INPP5A, MAD1L1, PDE4B* and *PTPRN2* (Table 6.14), all of which had five or more differentially methylated CpG sites. Five of the six genes had roles in brain development and brain pathologies that are directly relevant to the CP phenotype. In contrast to the differentially methylated regions in Table 6.12 and 6.13 hypomethylation is predominantly observed at differentially methylated regions detected for this this interaction.

Table 6.14 Genes for which differentially methylated regions were detected via the interaction of *in utero* maternal exposed and CP scored individuals.

Gene	Illumina ID	Correlation	Location	CpG island	Functional association
CUX1	cg02169185	hypo	chr7:101,899,429-101,899,429	Body	Global developmental delay [38]
	cg22202558	hyper	chr7:101,500,303-101,500,303	Body	
	cg10141789	hypo	chr7:101,807,411-101,807,411	Body	
	cg07266412	hypo	chr7:101,723,522-101,723,522	Body	
	cg24420432	hypo	chr7:101,518,619-101,518,619	Body	
DIP2C	cg22869706	hypo	chr10:357,285-357,285	Body	Breast and lung cancers [35, 36]
	cg05764011	hypo	chr10:409201-409523	Body	
	cg25488288	hypo	chr10:734707-735606	1st Exon	
	cg04854162	hypo		Body	
	cg12724894	hyper	chr10:711896-712395	Body	
INPP5A	cg21730012	hypo	chr10:134477298-134477515	Body	Cerebellar degeneration [39] Spinocerebellar [40]
	cg05174943	hypo	chr10:134,513,773-134,513,773	Body	
	cg04391569	hypo	chr10:134595293-134595694	Body	
	cg11740348	hypo	chr10:134,420,209-134,420,209	Body	
	cg08195412	hypo	chr10:134,372,696-134,372,696	Body	
MAD1L1	cg17309904	hyper	chr7:1,960,073-1,960,073	Body	Schizophrenia [41, 42] Bipolar [43, 44]
	cg17712928	hypo	chr7:2,124,974-2,124,974	Body	
	cg11994639	hypo	chr7:1,997,029-1,997,029	Body	
	cg23001930	hypo	chr7:2,094,478-2,094,478	Body	
	cg01886004	hypo	chr7:1,923,613-1,923,613	Body	
	cg11824316	hypo	chr7:1950279-1950482	Body	
PDE4B	cg02077669	hypo	chr1:66,730,524-66,730,524	Body	Neurological disorders [45, 46]
	cg17726558	hypo		Body	
	cg15393946	hypo		Body	
	cg11741987	hypo		TSS1500	
	cg02288344	hypo		Body	
PTPRN2	cg09193477	hypo	chr7:158,188,683-158,188,683	Body	Frontal temporal dementia [47] ADHD [48] Cocaine dependence and Major depressive episode [49]
	cg05869732	hypo	chr7:158318970-158319172	Body	
	cg23053506	hypo	chr7:158,075,705-158,075,705	Body	
	cg06094238	hypo	chr7:157568163-157568404	Body	
	cg03916382	hypo	chr7:157937614-157937841	Body	
	cg16492510	hypo	chr7:157560974-157561195	Body	
	cg13486056	hypo		Body	

6.4 Discussion

6.4.1 Overall analysis

Within this chapter, we aimed to assess genome-wide differences between *in utero* tobacco exposure vs non-exposed, low CP vs high CP and the interaction between *in utero* tobacco exposure and CP. Here, we are presenting work based off three models, the first showed adjusted significance for genome wide differential DNA methylation and nominal significance was observed for the other two. More so, differential methylation was detected in CpG sites located in biological relevant genes, particularly in the *in utero* tobacco exposure and CP interaction model. Which leads us to support our hypothesis that DNA methylation may be involved in the development of CP in individuals exposed to tobacco *in utero*.

6.4.2 Sample size and batch effects

Our pilot study in Chapter 5 was limited by sample size. A benefit of expanding our previous study into a genome-wide approach (using the Illumina EPIC array) we were able to increase our sample size (from 96 in Chapter 5 to 109 here). As we included array data from Chapter 2, in combination with new 2020 samples, to maximise our sample size, providing more statistical power. In saying this, the sample size was still moderately small, hence why nominal significance, rather than genome-wide significance, was the best outcome for some of the models presented here. However, this is a continuing challenge faced throughout this thesis, and one which we attempt to address by validating findings of previously published studies in our dataset. We show when differential methylation is assessed based on adult smoking status, our top differentially methylated sites replicate those identified in Chapter 2, that were significant at the genome-wide level. Therefore, this gives us confidence that the models which reach nominal significance are likely to be biologically relevant, and would benefit from an increase in sample size to probe the statistical associations further.

Similar to that which we have previously undertaken in Chapter 2, a selection of normalisation techniques were trialled to adjust the profound batch effect which we

know is a problem in metanalyses of array data. Unlike our post-normalisation results in Chapter 2, we were unable to correct for the additional samples added as the 2020 samples via any of the published methods trialled. Our analyses in this chapter show that the tool noob corrected for differences in the 2016 and 2017 samples, which were the same samples in Chapter 2, but was unable to adjust for the additional 2020 samples, which were the new additional samples.

Other techniques were also trialled here, but none successfully corrected for this batch effect. We decided to persevere with the analysis and mitigate the batch effect problem by including year of sampling as a confounding variable in the model. It is noteworthy that our analysis of our combined dataset to explore differential methylation in response to adult smoking status (essentially our control analysis) rendered very similar results to the data output from Chapter 2. The similarity observed gives us confidence that the inclusion of year of sampling in our model is correcting for this batch effect and that are results may be biologically meaningful. If we had not observed concordance with Chapter 2 (Table 2.5) then we would have sought a new pipeline for analysis. However, our results here did reflect our previous analyses (Supplementary Table 6.1) with the top two CpG sites (*AHRR* and *F2RL3)* displaying adjusted *P* values of significance, and both ranked as 1st and 4th most differentially methylated CpG sites in Table 2.5. We conclude here that the inclusion of year into each of the baseline models has corrected batch effect, and we suggest that this process is as a valid way of normalising data, without having to develop a new pipeline.

6.4.3 Hierarchical model selection

Multiple models were fitted to the data to get a clear understanding of how many covariables could be added prior to genomic inflation appearing in the results. Often, there is a fine balance between adjusting for as many confounders and overfitting, which can result in loss of relevant biological data.

Previous research on exposure to tobacco (adult smoking status) and its effect on genome-wide methylation has identified reproducible differential DNA methylation at specific loci [50, 51]. Here we demonstrated validation of these previously identified loci (Supplementary Table 6.1), meaning that our less stringent models with fewer

confounding variables were producing robust data. Thus, we chose to use the least stringent models, correcting for year of sampling and four principal components, to give us the greatest amount of power, reduce genomic inflation, and provide the most biologically relevant data possible with our sample size.

Other models explored but not used to generate the results of this chapter are also found in Table 6.3. These models featured additional covariates, such as adult tobacco smoking status, cannabis smoking status and sex. Although lambda values of these other models did not give an indication of over fitting, they were deemed as secondary cofounders which will not bias the downstream analyses. For example, while we exclude the X and Y chromosomes probes from analysis as there is the potential for them to skew the analysis, since our cohort here is matched for sex, ethnicity and socioeconomic status, this means that we also have both female and male participants for each of the cases and controls, essentially self-correcting for this confounder. The same rationale applies to adult tobacco smoking status and cannabis use status. Thus, although these are not fully corrected for in the models these variables are included in both the control and exposure groups.

6.4.4 DNA methylation differences from individuals exposed to tobacco *in utero* vs non-exposed controls

When we assessed the effect of exposure to tobacco *in utero* compared to non-exposed controls, four CpG sites reached an adjusted P value significance (Table 6.5) in the genes *MYO1G, RTN1* and two sites in *FRMD4A*. CpG sites in genes *MYO1G* and *FRMD4A* have been previously found to be differentially methylated due to maternal smoking during pregnancy [52-54]. However, no known literature has reported on the CpG site cg01604380 within the gene *RTN1*.

Reticulon 1 (*RTN1)* is a part of the RTN protein family, which resides in the endoplasmic reticulum. The proteins are predominantly involved in trafficking and axonal regeneration [55] and *RTN1* has been linked to Alzheimer's disease [56-58]. Within the top 10 most differentially methylated CpG sites (Table 6.5) there are two sites that resided within *RTN1* (cg01604380 and cg11866719). These two sites are 141bp away from one another (chr14:60,336,293 and chr14:60,336,434) therefore

may be having an additive effect on gene transcription. Myosin 1 G *(MYO1G)* is highly expressed in T cells and is responsible for innate and adaptive immune related functions [59]. The gene also plays a role in cell elasticity [60] with a loss of *MYO1G* causing a decrease in cell tension affecting migration and both endocytosis and phagocytosis [61]. Ferm domain containing 4A *(FRMD4A*), in which two CpG sites were differentially methylated, has strong relevance to the brain and is important for neuronal development and synaptic processes; mutations in this gene are associated with intellectual disability, microcephaly and are potentially a risk marker for Alzheimer's disease [20, 62].

MYO1G and *FRMD4A* are more established biomarkers for tobacco *in utero* exposure [52-54]. However, these prior results explore DNA methylation in childhood. Our results expand on this current knowledge into DNA methylation stability, as here we show that these well-established biomarkers are detectable in exposed individuals through to adulthood (age ~28 years). Thus, it is clear that *MYO1G* and *FRMD4A* are specifically differentially methylated in response to maternal tobacco use during pregnancy, and that these methylation changes induced in utero appear to be stable into adult. However, the effect of methylation changes in these genes on the development of CP is unknown. We suggest further research to quantify DNA methylation changes over the life course in a cohort where matched samples from childhood and adulthood are available, and explore the association with development of CP phenotypes.

KEGG pathway analysis of CpG sites within or near genes that are differentially methylated in response to maternal tobacco use during pregnancy demonstrate enrichment for pathways that have functional relevance to *MYO1G* and *FRMD4A.* Specifically, the top KEGG pathways (Table 6.6) included focal adhesion, ErbB signalling pathways and both B and T cell reception signalling, all of which have strong relevance to *MYO1G.* Also within this list are brain specific KEGG pathways such as glutamatergic synapse and axon guidance, which have relevance to the biological functions of *FRMD4A*. Further findings from the pathway analysis showed an overwhelming amount of signalling pathways affected, with 39 pathways in total displaying an adjusted *P* value of less than 0.05.

Interestingly, the top 10 differentially methylated CpG sites in response to maternal tobacco use (Table 6.5) all displayed hypermethylation in response to the exposure. The same observation were present in the differentially methylated regions of genes displaying greater than five CpG sites (Table 6.12). Of the exposures we assessed, *in utero* tobacco exposure showed the largest proportion of differentially methylated regions (DMRs). When we assessed the functional associations between DMRs and phenotypes from previously published literature of the genes we identify in this analysis, the genes which the DMRs resided in were associated a range of diseases; total of nine out of the 12 DMRs (Table 6.12) had direct brain developmental phenotypes and the others were associated with cardiac function, cancer and metabolism. Thus while CP status was not included in this model, these data further support the role of DNA methylation in the link between *in utero* tobacco exposure and the development of CP.

6.4.5 Differences in methylation in low CP compared to high CP scored individuals

When we assessed the differences in DNA methylation between low CP and high CP scored individuals, we found nominal significance throughout the genome. While this was the least profound effect measured, with the observed adjusted *P* values all being greater than (P<0.992), our results produced biologically relevant findings, which we will discuss below.

Two CpG sites in the gene Phosphodiesterase 9A (*PDE9A)* were identified in the top 10 most differentially methylated sites between high and low CP scored individuals. The two CpG sites in our analysis have display differential methylation in opposite directions; cg11570752 is hypermethylated and cg04916741 is hypomethylated. *PDE9A* gene is predominantly expressed in neurons in the brain and disruption in this gene has been associated with several neurological deficits [63, 64]. *PDE9A* is also targeted previously as a therapeutic to treat cognitive disorders [37, 65, 66]. Therefore, its association here with CP score is intriguing and would benefit from further investigation.

In contrast to the observations from exposure to tobacco *in utero*, high CP scored individuals displayed greater levels of hypomethylation compared to low CP.

Differential methylation changes were also much smaller, with very few CpG sites displaying greater than 3% differential methylation. KEGG pathway analysis identified nine pathways that were significantly enriched after adjustment, within which brain related pathways predominated (e.g Cholinergic synapse, axon guidance and GABAergic synpase).

Protocadherin gamma family (*PCDHG*) displayed a long region of differential methylation in response to both *in utero* tobacco exposure (seven CpG sites), adult tobacco smoking status (11 CpG sites, Supplementary Table 6.2) but 30 CpG sites differentially methylated in low CP vs high CP, all of which were observed to be hypermethylated. In fact, this was the largest differentially methylated region observed in response to any of the exposures assessed in this chapter. When we investigated the specific location of this differentially methylated region it was challenging to distinguish between the many transcripts associated with that location (Figure 6.8a). Thus, we believe this area of the genome is still poorly annotated, and we are unable to be more specific about the exact gene within this family that our data relates to. However, the *PCDHG* gene family is of important relevance to fetal brain development, implying that this differentially methylated region is biologically relevant and should be explored further.

Further supporting its role in development of CP, differential DNA methylation within this same region (Chr5:140,750,000-140,850,000) has been found to be altered in numerous disorders, for example, Down syndrome [67], dyslexia [68], cancer (colorectal cancer and gastric cancer primarily) [69], fetal alcohol syndrome [70] and Williams syndrome [71]. In all of these disorders, including our observations of high CP scored individuals, all sites are hypermethylated. The common theme across these disorders is brain development; Down syndrome, fetal alcohol syndrome, dyslexia, Williams's syndrome and high CP are likely to have brain related dysfunction too. Highlighting the importance of this genome region in brain development and suggests that altered methylation at this region may be contributing to the development of CP phenotypes.

6.4.6 *In utero* exposure with the interaction of CP

The last model assessed here was the interaction between *in utero* exposure and CP. No CpG sites displaying genome-wide significance were detected under this interaction model, however, nominal significance was observed. Beta differences within the top 10 most nominally ($P<0.01$) significant CpG sites remained low, with the largest being 1.2%. The top four most differentially methylated CpG sites resided in the following genes: *SLC10A7*, *LAMC3*, *PHF2* and *FASTKD1*. At CpG sites within these genes, specific differences in methylation were found between individuals with high CP scores who were exposed to tobacco *in utero*, versus those that were not. There was no difference in methylation between individuals exposed *in utero* versus those who were not, who had low CP scores. Implying that, at these loci, differential methylation is specifically detected in exposed individuals with high CP scores, suggesting that these differences specifically associate with the development of CP in exposed individuals. These findings should be explored in a larger cohort to fully investigate this association. What this finding further suggests is that the mechanism of CP development in non-exposed individuals with high CP is likely to be different to the individuals who were exposed to maternal tobacco smoke with the same phenotype. This is not surprising – CP is a highly complex phenomenon and encompasses a range of phenotypes [72], which will have a range of aetiologies. What this study does show, however, is that DNA methylation at specific CpG sites within the genome associate with the link between maternal tobacco use during pregnancy and high CP score in exposed offspring, suggesting that tobacco-induced DNA methylation changes may be playing a role in the development of CP in exposed individuals. Thus, although the results in this chapter are, in the main, only nominally significant, they do offer some insight into the contributing genes, and how they function, could be playing a role in the phenotype of high CP.

Of the four genes specific to the high CP/exposed group under the interaction model, the most significantly differentially methyaled CpG site was found within the gene, Solute Carrier Family 10 Member 7 *(SLC10A7)*. The gene does not have a clear biological link to our investigated phenotype but mutations in this gene show dysfunction in skeletal development [73, 74]. Its role is essential for the biosynthesis and trafficking of glycoproteins for the functioning of the extracellular matrix [74].

The other three CpG sites within the top four all share common functional biological roles which exhibit relevance to the CP phenotype. Firstly, Laminin gamma 3 (*LAMC3)*, has diverse roles in cell migration, apoptosis and adhesion. Mutations within this gene have been found to contribute to cortical malformations [75-77]. More so, it has been reported that individuals with this mutation also have specific behavioural outcomes, e.g., impairments in endogenous attentional processes [76]. FAST Kinase domain 1 (*FASTKD1)* plays a role in the regulation of mitochondrial RNA [78, 79]. Single nucleotide variants within this gene have been associated with glaucoma [80] . PHD Finger protein 2 (*PHF2)* is a key regulator in neural stem cell proliferation [81]. The gene was first described as mutation known as hereditary sensory neuropathy type 1, which is a disorder of the sensory neurons [82]. Mutations that arise have been linked to genome instability [81] and also been found in high CP related phenotypes such as Autism Spectrum Disorder [83]. Although these three genes all share similar visual impairment phenotypes along with their roles in brain development, this was not reflected in the pathway analysis results (Table 6.10).

6.4.7 Overall genome-wide significance

Overall, the model though which we detected the highest number of genome-wide methylation changes *in utero* tobacco exposure versus non-exposed (Table 6.11) with a total of 7,228 CpG sites differentially methylated (nominal *P* <0.01). Interestingly, there were ~2,000 more differentially methylated sites here compared to our model which explored methylation in response to adult tobacco smoking status, which displayed 4,747 differentially methylated CpG sites (nominal *P* value <0.01).

Tobacco smoking is one of the most potent exposures to DNA methylation patterns in the genome, however, in our analysis there were far less changes in response to adult tobacco smoking status compared to *in utero* tobacco exposure. Eluding further to the key role DNA methylation plays *in utero* development and disruption during these vulnerable times can lead to long lasting alterations.

6.5 Chapter Summary

- Quantification of genome-wide differential DNA methylation in response to *in utero* tobacco exposure found adjusted P values of significance at four CpG sites: *MYO1G*, two sites in *FRMD4A* and *RTN1*.
- Detection of differential DNA methylation between individuals with low CP and high CP scores was observed only nominal significance.
- We sought to determine if there was an interaction between *in utero* tobacco exposure and high CP scores, nominal significance was observed here. We did demonstrate specific CpG methylation differences were seen between the unexposed group with high CP and the exposed group with high CP.
- Our findings support the hypothesis that DNA methylation is inovlved in the link between *in utero* tobacco exposure and CP development. They also highlight the detrimental effect that *in utero* tobacco exposure has on the genome, and suggest that DNA methylation may have implications for development of disease later on in life.

6.6 References

1. Cornelius, M.D., et al., *Effects of prenatal cigarette smoke exposure on neurobehavioral outcomes in 10-year-old children of adolescent mothers.* Neurotoxicology and teratology, 2011. **33**(1): p. 137-144.
2. Robinson, M., et al., *Smoking cessation in pregnancy and the risk of child behavioural problems: a longitudinal prospective cohort study.* J Epidemiol Community Health, 2010. **64**(7): p. 622-9.
3. Flam, E.L., et al., *Differentially Methylated Super-Enhancers Regulate Target Gene Expression in Human Cancer.* Scientific reports, 2019. **9**(1): p. 15034-15034.
4. Cao, W., et al., *Multi-faceted epigenetic dysregulation of gene expression promotes esophageal squamous cell carcinoma.* Nature Communications, 2020. **11**(1): p. 3675.
5. Nativio, R., et al., *An integrated multi-omics approach identifies epigenetic alterations associated with Alzheimer's disease.* Nature Genetics, 2020. **52**(10): p. 1024-1035.
6. Gusev, F.E., et al., *Chromatin profiling of cortical neurons identifies individual epigenetic signatures in schizophrenia.* Translational Psychiatry, 2019. **9**(1): p. 256.
7. Cheung, P., et al., *Single-Cell Chromatin Modification Profiling Reveals Increased Epigenetic Variations with Aging.* Cell, 2018. **173**(6): p. 1385-1397.e14.
8. Pidsley, R., et al., *Critical evaluation of the Illumina MethylationEPIC BeadChip microarray for whole-genome DNA methylation profiling.* Genome Biology, 2016. **17**(1): p. 208.
9. Lawrence, M., et al., *Software for computing and annotating genomic ranges.* PLoS computational biology, 2013. **9**(8).
10. Kuleshov, M.V., et al., *Enrichr: a comprehensive gene set enrichment analysis web server 2016 update.* Nucleic acids research, 2016. **44**(W1): p. W90-W97.
11. Wickham, H., *ggplot2: Elegant Graphics for Data Analysis*. 2016: Springer-Verlag New York.
12. Lessel, D., et al., *BCL11B mutations in patients affected by a neurodevelopmental disorder with reduced type 2 innate lymphoid cells.* Brain, 2018. **141**(8): p. 2299-2311.
13. Simon, R., C. Wiegreffe, and S. Britsch, *Bcl11 Transcription Factors Regulate Cortical Development and Function.* Frontiers in molecular neuroscience, 2020. **13**: p. 51-51.
14. Shinawi, M., et al., *Intragenic CAMTA1 deletions are associated with a spectrum of neurobehavioral phenotypes.* Clin Genet, 2015. **87**(5): p. 478-82.
15. Liu, Z., et al., *Essential role of the zinc finger transcription factor Casz1 for mammalian cardiac morphogenesis and development.* The Journal of biological chemistry, 2014. **289**(43): p. 29801-29816.
16. Wong, N., et al., *Upregulation of FAM84B during prostate cancer progression.* Oncotarget, 2017. **8**(12): p. 19218-19235.
17. Jiang, Y., et al., *FAM84B promotes prostate tumorigenesis through a network alteration.* Therapeutic Advances in Medical Oncology, 2019. **11**: p. 1758835919846372.
18. Siper, P.M., et al., *Prospective investigation of FOXP1 syndrome.* Molecular Autism, 2017. **8**(1): p. 57.
19. Meerschaut, I., et al., *FOXP1-related intellectual disability syndrome: a recognisable entity.* Journal of Medical Genetics, 2017. **54**(9): p. 613.
20. Yan, X., et al., *FRMD4A–cytohesin signaling modulates the cellular release of tau.* Journal of Cell Science, 2016. **129**(10): p. 2003.
21. Huckert, M., et al., *Mutations in the latent TGF-beta binding protein 3 (LTBP3) gene cause brachyolmia with amelogenesis imperfecta.* Human molecular genetics, 2015. **24**(11): p. 3038-3049.
22. Fukuda, E., et al., *Down-regulation of protocadherin-α A isoforms in mice changes contextual fear conditioning and spatial working memory.* European Journal of Neuroscience, 2008. **28**(7): p. 1362-1376.
23. Su, L., et al., *PRDM16 orchestrates angiogenesis via neural differentiation in the developing brain.* Cell Death & Differentiation, 2020. **27**(8): p. 2313-2329.
24. Shimada, I.S., et al., *Prdm16 is required for the maintenance of neural stem cells in the postnatal forebrain and their differentiation into ependymal cells.* Genes & Development, 2017. **31**(11): p. 1134-1146.
25. Baizabal, J.-M., et al., *The Epigenetic State of PRDM16-Regulated Enhancers in Radial Glia Controls Cortical Neuron Position.* Neuron, 2018. **98**(5): p. 945-962.e8.
26. Matt, L., et al., *SynDIG4/Prrt1 Is Required for Excitatory Synapse Development and Plasticity Underlying Cognitive Function.* Cell reports, 2018. **22**(9): p. 2246-2253.

27. Kirk, L.M., et al., *Distribution of the SynDIG4/proline-rich transmembrane protein 1 in rat brain.* The Journal of comparative neurology, 2016. **524**(11): p. 2266-2280.
28. Minami, A., et al., *Increased insulin sensitivity and hypoinsulinemia in APS knockout mice.* Diabetes, 2003. **52**(11): p. 2657-2665.
29. Wu, X., et al., *Homeobox transcription factor VentX regulates differentiation and maturation of human dendritic cells.* J Biol Chem, 2014. **289**(21): p. 14633-43.
30. Gentner, E., et al., *VENTX induces expansion of primitive erythroid cells and contributes to the development of acute myeloid leukemia in mice.* Oncotarget, 2016. **7**(52): p. 86889-86901.
31. Rawat, V.P., et al., *The vent-like homeobox gene VENTX promotes human myeloid differentiation and is highly expressed in acute myeloid leukemia.* Proc Natl Acad Sci U S A, 2010. **107**(39): p. 16946-51.
32. Chen, Q., et al., *Elevated cholecystokininergic tone constitutes an important molecular/neuronal mechanism for the expression of anxiety in the mouse.* Proceedings of the National Academy of Sciences, 2006. **103**(10): p. 3881-3886.
33. Horinouchi, Y., et al., *Reduced anxious behavior in mice lacking the CCK2 receptor gene.* European neuropsychopharmacology, 2004. **14**(2): p. 157-161.
34. Zhang, L., et al., *Mitochondria-focused gene expression profile reveals common pathways and CPT1B dysregulation in both rodent stress model and human subjects with PTSD.* Translational Psychiatry, 2015. **5**(6): p. e580-e580.
35. Larsson, C., et al., *Loss of DIP2C in RKO cells stimulates changes in DNA methylation and epithelial-mesenchymal transition.* BMC Cancer, 2017. **17**(1): p. 487.
36. Li, J., et al., *DIP2C expression in breast cancer and its clinical significance.* Pathology - Research and Practice, 2017. **213**(11): p. 1394-1399.
37. Sidhu, M., et al., *Novel heterozygous mutations in the PGAM2 gene with negative exercise testing.* Mol Genet Metab Rep, 2018. **17**: p. 53-55.
38. Platzer, K., et al., *Haploinsufficiency of CUX1 Causes Nonsyndromic Global Developmental Delay With Possible Catch-up Development.* Annals of Neurology, 2018. **84**(2): p. 200-207.
39. Yang, A.W., A.J. Sachs, and A.M. Nystuen, *Deletion of Inpp5a causes ataxia and cerebellar degeneration in mice.* neurogenetics, 2015. **16**(4): p. 277-285.
40. Liu, Q., et al., *Cerebellum-enriched protein INPP5A contributes to selective neuropathology in mouse model of spinocerebellar ataxias type 17.* Nature Communications, 2020. **11**(1): p. 1101.
41. Su, L., et al., *Genetic association of GWAS-supported MAD1L1 gene polymorphism rs12666575 with schizophrenia susceptibility in a Chinese population.* Neurosci Lett, 2016. **610**: p. 98-103.
42. Chang, S., et al., *Network-Based Analysis of Schizophrenia Genome-Wide Association Data to Detect the Joint Functional Association Signals.* PLOS ONE, 2015. **10**(7): p. e0133404.
43. Trost, S., et al., *Investigating the Impact of a Genome-Wide Supported Bipolar Risk Variant of MAD1L1 on the Human Reward System.* Neuropsychopharmacology, 2016. **41**(11): p. 2679-87.
44. Zhao, L., et al., *Replicated associations of FADS1, MAD1L1, and a rare variant at 10q26.13 with bipolar disorder in Chinese population.* Translational Psychiatry, 2018. **8**(1): p. 270.
45. Siuciak, J.A., et al., *Behavioral and neurochemical characterization of mice deficient in the phosphodiesterase-4B (PDE4B) enzyme.* Psychopharmacology, 2008. **197**(1): p. 115-126.
46. Siuciak, J.A., et al., *Antipsychotic profile of rolipram: efficacy in rats and reduced sensitivity in mice deficient in the phosphodiesterase-4B (PDE4B) enzyme.* Psychopharmacology, 2007. **192**(3): p. 415-424.
47. van der Ende, E.L., et al., *Novel CSF biomarkers in genetic frontotemporal dementia identified by proteomics.* Annals of Clinical and Translational Neurology, 2019. **6**(4): p. 698-707.
48. Lionel, A.C., et al., *Rare Copy Number Variation Discovery and Cross-Disorder Comparisons Identify Risk Genes for ADHD.* Science Translational Medicine, 2011. **3**(95): p. 95ra75-95ra75.
49. Yang, B.-Z., et al., *A Genomewide Linkage Scan of Cocaine Dependence and Major Depressive Episode in Two Populations.* Neuropsychopharmacology, 2011. **36**(12): p. 2422-2430.
50. Prince, C., et al., *Investigating the impact of cigarette smoking behaviours on DNA methylation patterns in adolescence.* Human Molecular Genetics, 2019. **28**(1): p. 155-165.
51. Elliott, H.R., et al., *Differences in smoking associated DNA methylation patterns in South Asians and Europeans.* Clinical Epigenetics, 2014. **6**(1): p. 4.
52. Küpers, L.K., et al., *DNA methylation mediates the effect of maternal smoking during pregnancy on birthweight of the offspring.* International Journal of Epidemiology, 2015. **44**(4): p. 1224-1237.

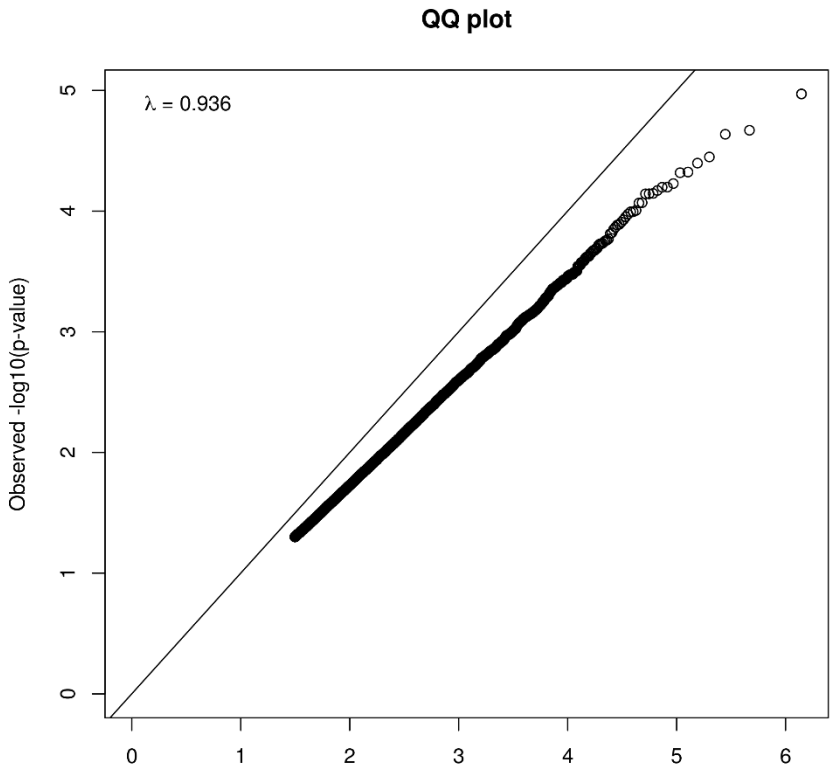
53. Wiklund, P., et al., *DNA methylation links prenatal smoking exposure to later life health outcomes in offspring.* Clinical Epigenetics, 2019. **11**(1): p. 97.
54. Vives-Usano, M., et al., *In utero and childhood exposure to tobacco smoke and multi-layer molecular signatures in children.* BMC Medicine, 2020. **18**(1): p. 243.
55. van de Velde, H.J., et al., *NSP-encoded reticulons, neuroendocrine proteins of a novel gene family associated with membranes of the endoplasmic reticulum.* Journal of Cell Science, 1994. **107**(9): p. 2403-2416.
56. He, W., et al., *Reticulon family members modulate BACE1 activity and amyloid-β peptide generation.* Nature Medicine, 2004. **10**(9): p. 959-965.
57. Heath, J.E., et al., *Widespread distribution of reticulon-3 in various neurodegenerative diseases.* Neuropathology, 2010. **30**(6): p. 574-579.
58. Masliah, E., et al., *Genetic deletion of Nogo/Rtn4 ameliorates behavioral and neuropathological outcomes in amyloid precursor protein transgenic mice.* Neuroscience, 2010. **169**(1): p. 488-494.
59. Olety, B., et al., *Myosin 1G (Myo1G) is a haematopoietic specific myosin that localises to the plasma membrane and regulates cell elasticity.* FEBS Letters, 2010. **584**(3): p. 493-499.
60. Kuleshov, M.V., et al., *Enrichr: a comprehensive gene set enrichment analysis web server 2016 update.* Nucleic Acids Res, 2016. **44**(W1): p. W90-7.
61. López-Ortega, O., et al., *Myo1g is an active player in maintaining cell stiffness in B-lymphocytes.* Cytoskeleton, 2016. **73**(5): p. 258-268.
62. Fine, D., et al., *A syndrome of congenital microcephaly, intellectual disability and dysmorphism with a homozygous mutation in FRMD4A.* European journal of human genetics : EJHG, 2015. **23**(12): p. 1729-1734.
63. van der Staay, F.J., et al., *The novel selective PDE9 inhibitor BAY 73-6691 improves learning and memory in rodents.* Neuropharmacology, 2008. **55**(5): p. 908-918.
64. Vardigan, J.D., et al., *The selective phosphodiesterase 9 (PDE9) inhibitor PF-04447943 attenuates a scopolamine-induced deficit in a novel rodent attention task.* Journal of neurogenetics, 2011. **25**(4): p. 120-126.
65. West, A.R. and A.A. Grace, *The nitric oxide-guanylyl cyclase signaling pathway modulates membrane activity states and electrophysiological properties of striatal medium spiny neurons recorded in vivo.* Journal of Neuroscience, 2004. **24**(8): p. 1924-1935.
66. Boland, K., et al., *A phase I, randomized, proof-of-clinical-mechanism study assessing the pharmacokinetics and pharmacodynamics of the oral PDE9A inhibitor BI 409306 in healthy male volunteers.* Human Psychopharmacology: Clinical and Experimental, 2017. **32**(1): p. e2569.
67. El Hajj, N., et al., *Epigenetic dysregulation in the developing Down syndrome cortex.* Epigenetics, 2016. **11**(8): p. 563-578.
68. Naskar, T., et al., *Ancestral Variations of the PCDHG Gene Cluster Predispose to Dyslexia in a Multiplex Family.* EBioMedicine, 2018. **28**: p. 168-179.
69. Vega-Benedetti, A.F., et al., *Clustered protocadherins methylation alterations in cancer.* Clinical Epigenetics, 2019. **11**(1): p. 100.
70. Laufer, B.I., et al., *Associative DNA methylation changes in children with prenatal alcohol exposure.* Epigenomics, 2015. **7**(8): p. 1259-1274.
71. Strong, E., et al., *Symmetrical Dose-Dependent DNA-Methylation Profiles in Children with Deletion or Duplication of 7q11.23.* The American Journal of Human Genetics, 2015. **97**(2): p. 216-227.
72. Kazdin, A.E., *Conduct Disorders in Childhood and Adolescence*. 1995: SAGE Publications.
73. Dubail, J., et al., *SLC10A7 mutations cause a skeletal dysplasia with amelogenesis imperfecta mediated by GAG biosynthesis defects.* Nat Commun, 2018. **9**(1): p. 3087.
74. Ashikov, A., et al., *Integrating glycomics and genomics uncovers SLC10A7 as essential factor for bone mineralization by regulating post-Golgi protein transport and glycosylation.* Hum Mol Genet, 2018. **27**(17): p. 3029-3045.
75. Barak, T., et al., *Recessive LAMC3 mutations cause malformations of occipital cortical development.* Nature genetics, 2011. **43**(6): p. 590-594.
76. Urgen, B.M., et al., *Homozygous LAMC3 mutation links to structural and functional changes in visual attention networks.* NeuroImage, 2019. **190**: p. 242-253.
77. Barak, T., et al., *Recessive LAMC3 mutations cause malformations of occipital cortical development.* Nature Genetics, 2011. **43**(6): p. 590-594.

78. Simarro, M., et al., *Fast kinase domain-containing protein 3 is a mitochondrial protein essential for cellular respiration.* Biochemical and biophysical research communications, 2010. **401**(3): p. 440-446.
79. Boehm, E., et al., *FASTKD1 and FASTKD4 have opposite effects on expression of specific mitochondrial RNAs, depending upon their endonuclease-like RAP domain.* Nucleic acids research, 2017. **45**(10): p. 6135-6146.
80. Information, N.C.f.B. *NM_024622.6(FASTKD1):c.2230T>A (p.Tyr744Asn)*. [VCV000548952.1], July,17th, 2018 24 Nov 2020].
81. Pappa, S., et al., *PHF2 histone demethylase prevents DNA damage and genome instability by controlling cell cycle progression of neural progenitors.* Proceedings of the National Academy of Sciences of the United States of America, 2019. **116**(39): p. 19464-19473.
82. Nicholson, G.A., et al., *The gene for hereditary sensory neuropathy type I (HSN–I) maps to chromosome 9q22.1–q22.3.* Nature Genetics, 1996. **13**(1): p. 101-104.
83. Iossifov, I., et al., *De Novo Gene Disruptions in Children on the Autistic Spectrum.* Neuron, 2012. **74**(2): p. 285-299.
84. Zhang, H., et al., *Bach2 Deficiency Leads to Spontaneous Expansion of IL-4-Producing T Follicular Helper Cells and Autoimmunity.* Frontiers in Immunology, 2019. **10**(2050).
85. Miotto, E., et al., *Frequent aberrant methylation of the CDH4 gene promoter in human colorectal and gastric cancer.* Cancer research, 2004. **64**(22): p. 8156-8159.
86. Chen, Y.-A., I.-L. Lu, and J.-W. Tsai, *Contactin-1/F3 Regulates Neuronal Migration and Morphogenesis Through Modulating RhoA Activity.* Frontiers in Molecular Neuroscience, 2018. **11**(422).
87. Canali, G., et al., *Genetic variants in autism-related CNTNAP2 impair axonal growth of cortical neurons.* Human Molecular Genetics, 2018. **27**(11): p. 1941-1954.
88. McAllister, J.M., et al., *Overexpression of a DENND1A isoform produces a polycystic ovary syndrome theca phenotype.* Proceedings of the National Academy of Sciences, 2014. **111**(15): p. E1519-E1527.
89. Bodmer, D., et al., *Disruption of a novel gene, DIRC3, and expression of DIRC3-HSPBAP1 fusion transcripts in a case of familial renal cell cancer and t(2;3)(q35;q21).* Genes, Chromosomes and Cancer, 2003. **38**(2): p. 107-116.
90. Fan, Z., et al., *DLGAP1 and NMDA receptor-associated postsynaptic density protein genes influence executive function in attention deficit hyperactivity disorder.* Brain and Behavior, 2018. **8**(2): p. e00914.
91. Kim, G.-C., et al., *ETS1 Suppresses Tumorigenesis of Human Breast Cancer via Trans-Activation of Canonical Tumor Suppressor Genes.* Frontiers in Oncology, 2020. **10**(642).
92. Verschoor, M.L., C.P. Verschoor, and G. Singh, *Ets-1 global gene expression profile reveals associations with metabolism and oxidative stress in ovarian and breast cancers.* Cancer & Metabolism, 2013. **1**(1): p. 17.
93. Zhong, J., et al., *Integration of GWAS and brain eQTL identifies FLOT1 as a risk gene for major depressive disorder.* Neuropsychopharmacology, 2019. **44**(9): p. 1542-1551.
94. Sakaguchi, M., et al., *FoxK1 and FoxK2 in insulin regulation of cellular and mitochondrial metabolism.* Nature Communications, 2019. **10**(1): p. 1582.
95. Zai, G., et al., *Possible association between the gamma-aminobutyric acid type B receptor 1 (GABBR1) gene and schizophrenia.* European Neuropsychopharmacology, 2005. **15**(3): p. 347-352.
96. Sando, R., 3rd, et al., *HDAC4 governs a transcriptional program essential for synaptic plasticity and memory.* Cell, 2012. **151**(4): p. 821-834.
97. Perttilä, J., et al., *OSBPL10, a novel candidate gene for high triglyceride trait in dyslipidemic Finnish subjects, regulates cellular lipid metabolism.* Journal of molecular medicine (Berlin, Germany), 2009. **87**(8): p. 825-835.
98. Sisodiya, S.M., et al., *PAX6 haploinsufficiency causes cerebral malformation and olfactory dysfunction in humans.* Nat Genet, 2001. **28**(3): p. 214-6.
99. Klann, M. and E.C. Seaver, *Functional role of pax6 during eye and nervous system development in the annelid Capitella teleta.* Developmental Biology, 2019. **456**(1): p. 86-103.
100. Lee, S., *The association of genetically controlled CpG methylation (cg158269415) of protein tyrosine phosphatase, receptor type N2 (PTPRN2) with childhood obesity.* Scientific Reports, 2019. **9**(1): p. 4855.
101. Sarachana, T. and V.W. Hu, *Genome-wide identification of transcriptional targets of RORA reveals direct regulation of multiple genes associated with autism spectrum disorder.* Molecular Autism, 2013. **4**(1): p. 14.

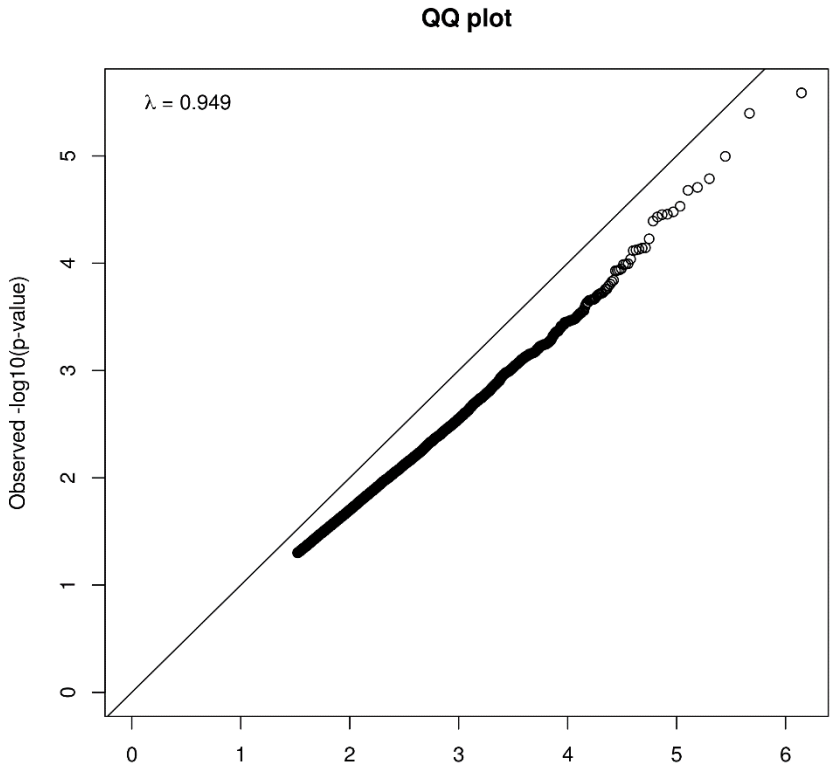
102. Lee, H.S., et al., *Circulating methylated septin 9 nucleic acid in the plasma of patients with gastrointestinal cancer in the stomach and colon.* Translational oncology, 2013. **6**(3): p. 290-IN4.
103. Kuo, I.-Y., et al., *Prognostic CpG methylation biomarkers identified by methylation array in esophageal squamous cell carcinoma patients.* International journal of medical sciences, 2014. **11**(8): p. 779.
104. Yang, I.V., et al., *Epigenetic marks of in utero exposure to gestational diabetes and childhood adiposity outcomes: the EPOCH study.* Diabetic Medicine, 2018. **35**(5): p. 612-620.
105. Leloup, N., L.M.P. Chataigner, and B.J.C. Janssen, *Structural insights into SorCS2–Nerve Growth Factor complex formation.* Nature Communications, 2018. **9**(1): p. 2979.
106. Glerup, S., et al., *SorCS2 is required for BDNF-dependent plasticity in the hippocampus.* Molecular Psychiatry, 2016. **21**(12): p. 1740-1751.
107. Marangi, G., et al., *TRAPPC9-related autosomal recessive intellectual disability: report of a new mutation and clinical phenotype.* European Journal of Human Genetics, 2013. **21**(2): p. 229-232.
108. Hnoonual, A., et al., *Novel Compound Heterozygous Mutations in the TRAPPC9 Gene in Two Siblings With Autism and Intellectual Disability.* Frontiers in Genetics, 2019. **10**(61).
109. Zhou, J., et al., *Ubiquitin-specific protease-44 inhibits the proliferation and migration of cells via inhibition of JNK pathway in clear cell renal cell carcinoma.* BMC Cancer, 2020. **20**(1): p. 214.
110. Tarcic, O., et al., *RNF20 and histone H2B ubiquitylation exert opposing effects in Basal-Like versus luminal breast cancer.* Cell Death & Differentiation, 2017. **24**(4): p. 694-704.

6.7 Supplementary Figures and Table

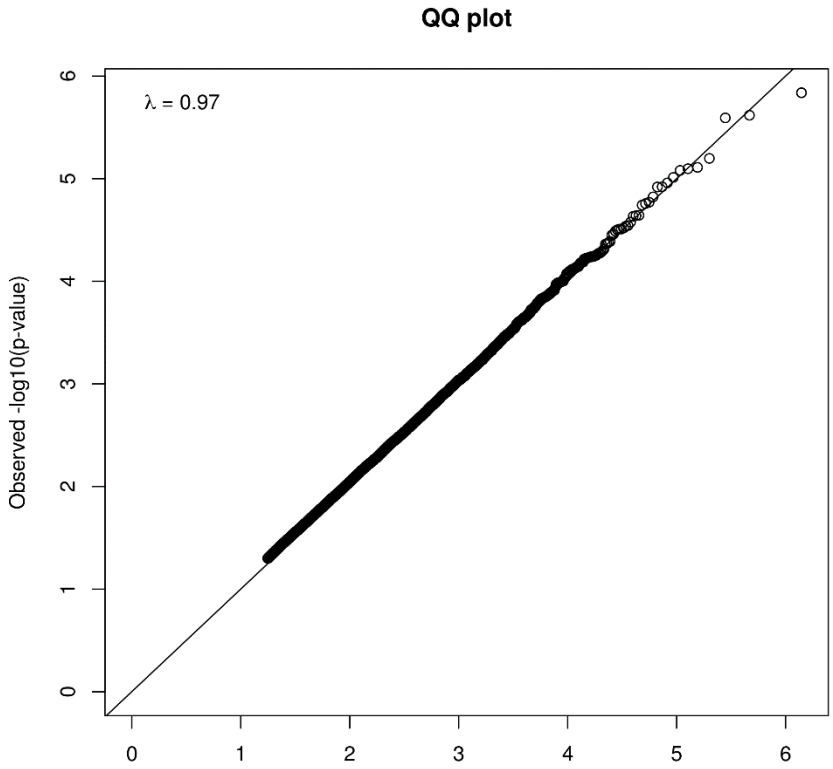
A



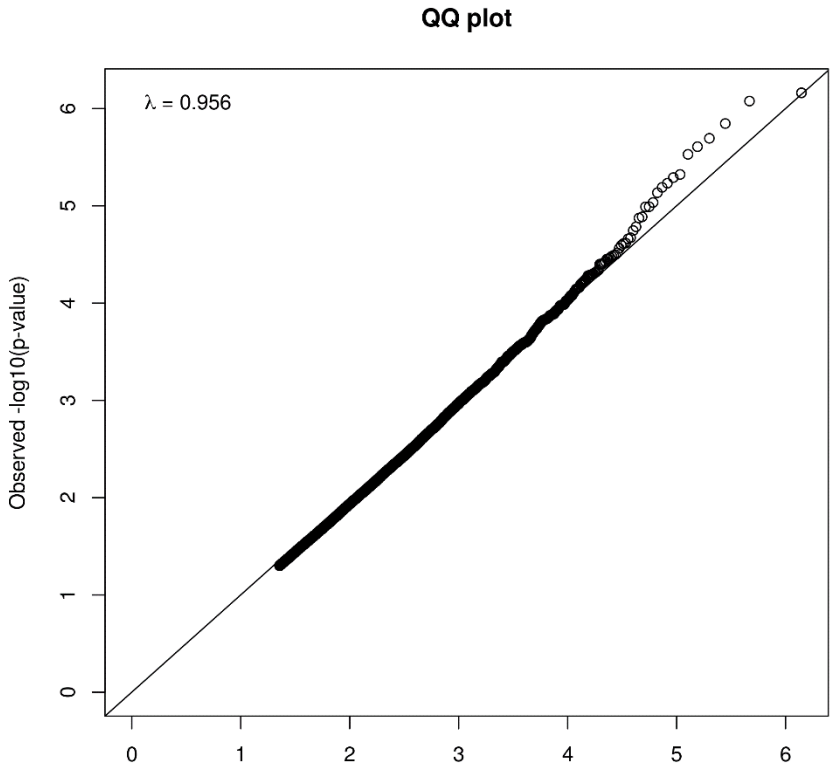
B



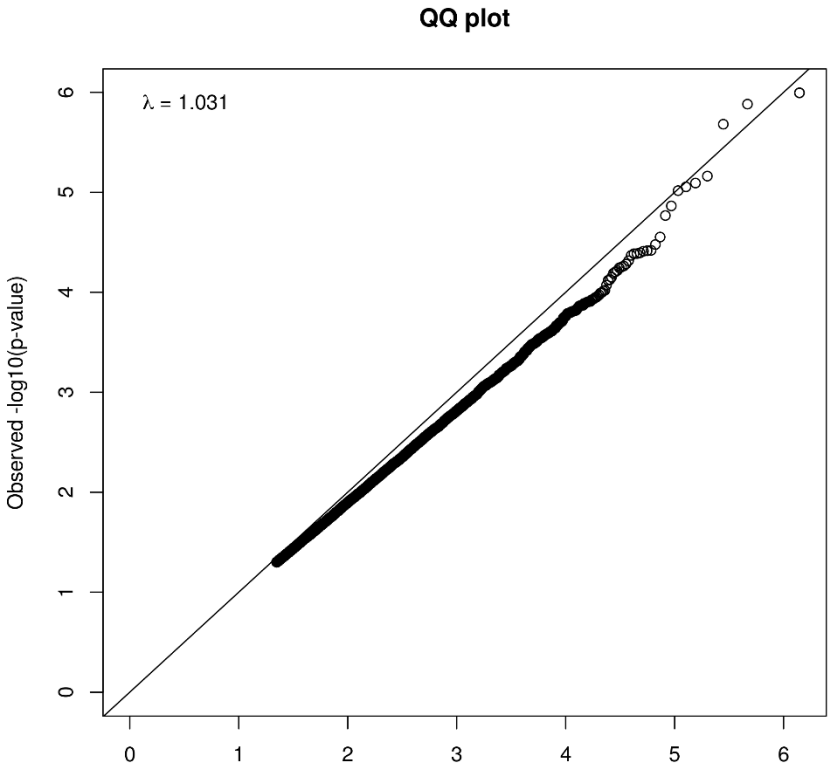
C


QQ plot
λ = 0.97
Observed -log10(p-value)
0
1
2
3
4
5
6

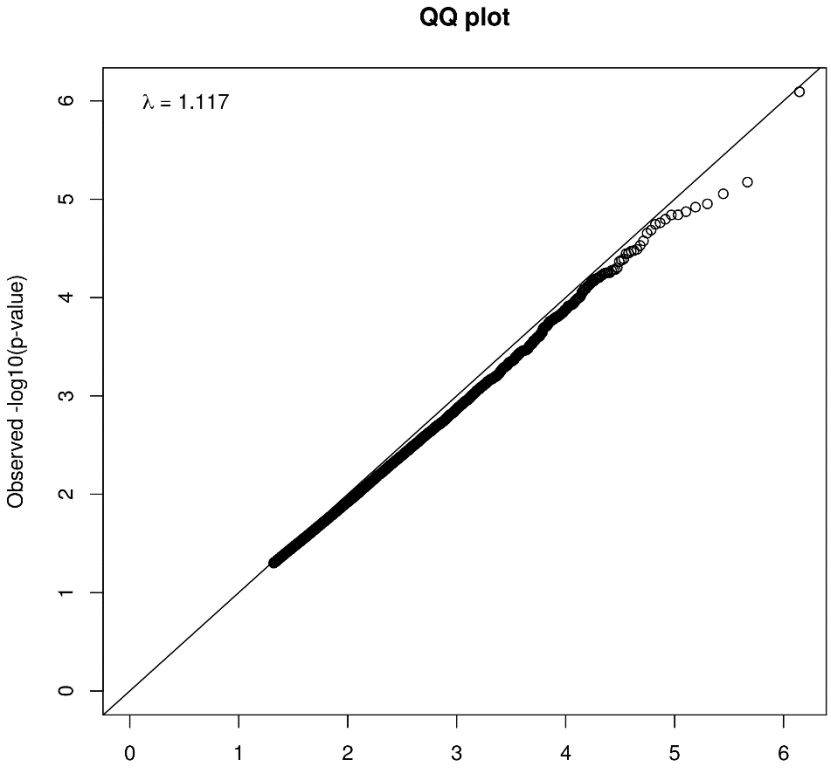
D


QQ plot
λ = 0.956
Observed -log10(p-value)
0
1
2
3
4
5
6

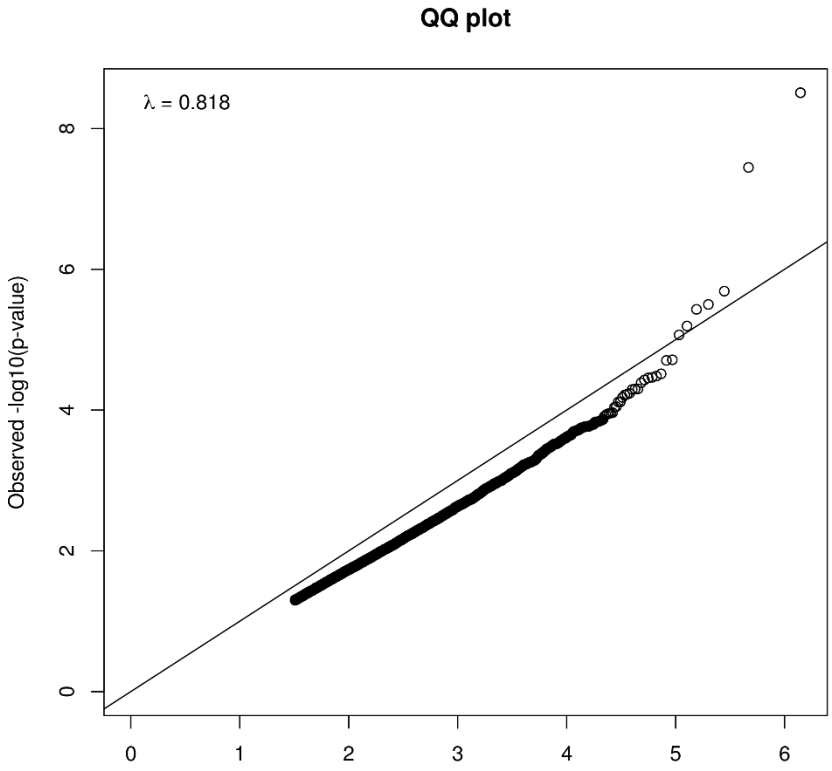
E


QQ plot
λ = 1.031
Observed -log10(p-value)
0
1
2
3
4
5
6

F


QQ plot
λ = 1.117
Observed -log10(p-value)
0
1
2
3
4
5
6

G



Supplementary Figure 6.1 Q-Q plots of the models that were fitted, Part A-F model results were not used to generate the results of this chapter. Part G was used for the tobacco smoking data. A) Model 2, B) Model 3, C) Model 5, D) Model 6, E) Model 8, F) Model 9, G) Model 10.

Supplementary Table 6.1 Differentially methylation CpG sites between adult tobacco smoking vs non-smoking controls (model 10).

Rank	Illumina ID	Gene	Chr	Position in genome	Location	Control	Adult tobacco smokers	β difference	Log FC	*P* Value	Adjusted *P* value
1	cg05575921	*AHRR*	5	Body	chr5:373842-374426	0.880	0.717	-0.162	-0.150	2.90E-09	0.0020
2	cg03636183	*F2RL3*	19	Body	chr19:17000627-17001398	0.674	0.611	-0.062	-0.061	9.50E-09	0.0033
3	cg03329539		2		chr2:233283397-233285959	0.412	0.370	-0.042	-0.042	7.59E-07	0.175
4	cg05767409	*NDUFS7*	19	TSS1500	chr19:1383437-1384251	0.053	0.058	0.004	0.0112	1.01E-06	0.175
5	cg21161138	*AHRR*	5	Body		0.746	0.701	-0.045	-0.048	1.28E-06	0.178
6	cg21566642		2		chr2:233283397-233285959	0.589	0.500	-0.089	-0.079	2.27E-06	0.264
7	cg21911711	*F2RL3*	19	TSS1500	chr19:17000627-17001398	0.846	0.815	-0.031	-0.031	4.50E-06	0.450
8	cg10870815	*CACNA1C*	12	Body		0.589	0.656	0.067	0.071	5.17E-06	0.452
9	cg23327011	*MBP*	18	5'UTR		0.929	0.923	-0.006	-0.012	6.97E-06	0.542
10	cg01940273		2		chr2:233283397-233285959	0.622	0.563	-0.058	-0.055	8.87E-06	0.594

Supplementary Table 6.2 Differentially methylated regions in genes (greater than 5 nominal P <0.01 CpG sites) in comparison to the same genes from the other exposure models. * genome poorly annotated in this area so unable to distinguish gene name

Gene	Exposed to tobacco *in utero*	CP	Exposed to tobacco *in utero* : CP	Adult tobacco smoking status
AHRR	3	2	1	7
ATP11A	5	2	1	3
BACH2	6	0	0	2
BCL11B	8	2	0	0
CAMTA1	10	0	1	1
CCKBR	0	5	0	0
CASZ1	7	0	2	1
CDH4	5	2	2	1
CNTN1	5	1	0	0
CNTNAP2	5	4	0	1
CUX1	0	0	5	3
CPT1B	0	8	0	0
DENND1A	5	0	1	2
DIP2C	7	5	5	5
DIRC3	5	1	1	0
DLGAP1	6	2	0	1
ETS1	5	0	0	0
FAM84B	8	2	0	0
FLOT1	5	0	1	0
FOXK1	5	0	2	1
FOXP1	11	2	0	1
FRMD4A	8	2	0	1
GABBR1	6	2	0	1
HDAC4	5	1	4	2
INPP5A	1	4	5	2
LTBP3	7	0	0	3
MAD1L1	4	1	6	4
MTHFR	0	3	0	0
MYO1G	5	0	0	2
OSBPL10	5	0	0	0
PAX6	5	1	1	2
*PCDHGA4**	7	30	4	11
PDE4B	4	0	5	0
PGAM2	0	6	0	0
PRDM16	7	2	0	7
PRRT1	8	1	0	0
PTPRN2	6	2	7	3
RORA	6	0	0	4
RPTOR	3	2	3	5
SH2B2	7	0	0	0
SH3PXD2A	5	0	3	0
Sept09	5	0	1	0
SORCS2	6	3	2	6
TRAPPC9	5	1	3	0
USP44	5	0	0	0
VAV2	5	0	1	3
VENTX	11	0	0	0

Supplementary Table 6.3 Differentially methylated regions for tobacco exposure *in utero* within genes displaying 5<6 CpG sites

Gene	Illumina ID	Correlation	Location	CpG island	Functional association
BACH2	cg20886265	hypo		3'UTR	Autoimmune
	cg01462343	hypo		5'UTR	disease[84]
	cg19039673	hypo		5'UTR	
	cg03035849	hyper	chr6:91004795-91006944	5'UTR	
	cg26961240	hypo		5'UTR	
	cg17486314	hyper		5'UTR	
CDH4	cg21538645	hyper		Body	Colorectal and
	cg03215161	hyper		Body	gastric cancer
	cg09051966	hyper	chr20:60470080-60470335	Body	[85]
	cg13021439	hyper		Body	
	cg17036908	hyper		1stExon	
CNTN1	cg15087347	hyper	chr12:41086522-41087102	TSS200	Formation of
	cg11743675	hyper		5'UTR	cortical neurons
	cg01443755	hyper		5'UTR	[86]
	cg10514886	hyper		TSS200	
	cg02234487	hyper		TSS200	
CNTNAP2	cg25949550	hypo	chr7:145813030-145814084	Body	Autism [87]
	cg15932065	hypo		Body	
	cg11207515	hypo		Body	
	cg05640346	hyper	chr7:148036494-148036848	Body	
	cg08374341	hyper		Body	
DENND1A	cg17501842	hyper		Body	Polycystic ovary
	cg00619207	hyper	chr1:111746337-111747303	TSS200	syndrome [88]
	cg15591803	hypo		Body	
	cg19269039	hyper	chr1:111746337-111747303	1stExon	
	cg20317872	hyper	chr1:111746337-111747303	1stExon	
	cg23184711	hyper	chr1:111746337-111747303	TSS200	
DIP2C	cg00332951	hypo	chr10:518192-518471	Body	Breast and lung
	cg19140503	hyper		Body	cancers [35, 36]
	cg13471336	hypo	chr10:669070-669336	Body	
	cg10591926	hypo	chr10:652259-652528	Body	
	cg03287763	hypo	chr10:465928-466396	Body	
	cg07102380	hypo		Body	
	cg24723457	hypo	chr10:396943-397228	Body	
DIRC3	cg00991467	hypo		Body	Renal cancer
	cg15335768	hypo		Body	[89]
	cg12596243	hypo		Body	
	cg14216322	hypo		Body	
	cg19918866	hypo		Body	
DLGAP1	cg16128363	hypo	chr18:3879202-3880087	TSS1500	Obsessive
	cg15874411	hypo		TSS1500	compulsive
	cg04214965	hypo	chr18:3879202-3880087	TSS1500	disorder [89]
	cg06291743	hypo	chr18:3879202-3880087	Body	ADHD [90]
	cg11159132	hypo		Body	
	cg05111420	hypo		Body	
ETS1	cg18898103	hypo		5'UTR	Cancer [91, 92]
	cg15555017	hyper		5'UTR	
	cg23514374	hyper		Body	
	cg23531640	hyper		Body	
	cg23800023	hyper		Body	
FLOT1	cg10513302	hyper		Body	Major depressive
	cg16646298	hyper		Body	disorder [93]
	cg02684104	hyper		Body	
	cg17988780	hyper		Body	
	cg09284772	hypo	chr6:30710307-30712440	Body	
FOXK1	cg03077364			Body	Insulin regulation
	cg22581896			Body	[94]
	cg01974478			Body	

	cg00208274			Body	
	cg05066096		chr7:4784820-4785058	Body	
GABBR1	cg08862148	hyper	chr6:29595298-29595795	Body	Schizophrenia
	cg21100518	hyper	chr6:29595298-29595795	Body	[95]
	cg00594408	hyper	chr6:29595298-29595795	Body	
	cg25642476	hyper	chr6:29595298-29595795	Body	
	cg02014853	hyper	chr6:29595298-29595795	Body	
	cg25729445	hyper	chr6:29595298-29595795	Body	
HDAC4	cg03281426	hypo	chr2:240111314-240111577	Body	Synaptic
	cg23367987	hypo		Body	plasticityand
	cg24634565	hypo	chr2:240101503-240101764	Body	memory[96]
	cg02812817	hypo		5'UTR	
	cg22296756	hypo		5'UTR	
MYO1G	cg04180046	hyper	chr7:45002111-45002845	Body	Maternal tobacco
	cg05009104	hyper	chr7:45002111-45002845	Body	smoke during
	cg12803068	hyper	chr7:45002111-45002845	Body	pregnancy[52]
	cg19089201	hyper	chr7:45002111-45002845	3'UTR	
	cg21188037	hyper		5'UTR	
OSBPL10	cg02057211	hypo		Body	Regulation of
	cg08541624	hypo		Body	lipid metabolism
	cg23774003	hypo		Body	[97]
	cg08947058	hypo		Body	
	cg24458780	hypo		Body	
PAX6	cg09041678	hyper	chr11:31841315-31842003	TSS1500	Cerebral and
	cg15301794	hyper	chr11:31820060-31821416	Body	olfactory
	cg01587682	hyper	chr11:31820060-31821416	Body	dysfunction [98,
	cg18082638	hyper	chr11:31827696-31827921	Body	99]
	cg12798259	hyper	chr11:31820060-31821416	Body	
PTPRN2	cg03983213	hyper	chr7:157476886-157486719	Body	Metabolic
	cg07176561	hyper		Body	disease [100]
	cg15080590	hyper	chr7:157494510-157494739	Body	
	cg08242024	hyper	chr7:157550547-157551025	Body	
	cg25906770	hyper	chr7:157568163-157568404	Body	
	cg16747052	hyper		Body	
RORA	cg27167601	hyper	chr15:61519621-61520031	TSS1500	Autism [101]
	cg16261097	hypo		Body	
	cg21241560	hyper		Body	
	cg12340454	hypo	chr15:61519621-61520031	Body	
	cg09782034	hyper	chr15:61520423-61521716	TSS1500	
	cg24053032	hyper		Body	
Sept09	cg21579666	hypo		5'UTR	Cancer [102,
	cg19277969	hypo		Body	103]
	cg05783080	hypo		Bod	
	cg01320579	hyper		Body	
	cg16293484	hyper		Body	
SH3PXD2A	cg13289509	hyper		Body	Gestational
	cg14467781	hyper	chr10:105614511-105615456	Body	diabetes [104]
	cg12975399	hyper	chr10:105614511-105615456	Body	
	cg17687265	hyper		Body	
	cg12312107	hyper	chr10:105452338-105453230	Body	
SORCS2	cg21445325	hyper		Body	Neuronal
	cg08268947	hyper		Body	plasticity[105,
	cg16356712	hypo	chr4:7593369-7593586	Body	106]
	cg11450537	hypo		Body	
	cg08145989	hyper		Body	
	cg17574602	hypo		Body	
TRAPPC9	cg14745383	hypo		Body	Intellectual
	cg14689150	hypo	chr8:141359155-141359621	Body	disability[107,
	cg23671279	hypo	chr8:140971270-140971524	Body	108]
	cg06924606	hypo		Body	
	cg24617008	hypo	chr8:141467218-141467927	TSS1500	
USP44	cg14565151	hyper	chr12:95941906-95942979	TSS200	Cancer hallmark
	cg08948170	hyper	chr12:95941906-95942979	TSS200	[109, 110]
	cg06476970	hyper	chr12:95941906-95942979	TSS200	
	cg04488758	hyper	chr12:95941906-95942979	TSS200	
	cg27100916	hyper	chr12:95941906-95942979	TSS200	

Chapter 7:

7. Is tissue really an issue? DNA methylation differences between whole blood and brain tissue in schizophrenia: a meta-analysis

7.1 Introduction

7.1.1 DNA methylation and whole blood

To understand how epigenetic modifications can impact on disease processes, these modifications need to be quantified from a tissue sample. Despite the fact that DNA methylation patterns are highly tissue-specific [1], one of the most common forms of tissue used for DNA methylation analysis, regardless of the environmental exposure or disease under investigation, is whole blood. There are two main reasons for this: i) samples can be obtained via a simple blood test which is easily accessible [2, 3], and; ii) blood samples are often collected routinely in clinical trials and biomedical studies, which means studies of DNA methylation can be applied retrospectively to complement ongoing or past research questions.

Throughout the contents of this thesis we have been assessing the proxy sample of whole blood for measuring diseases associated within the brain. Within a whole blood sample there are three major components; plasma, white blood cells and platelets, and red blood cells [4]. Thus, within this one sample there is a heterogeneous population of different cell types, each with its own unique epigenetic identity [5]. Because the different blood cell types are present in different proportions at different times (e.g. during infection), determination of methylation at CpG sites within a whole blood sample, if measured using bisulfite sequencing this is based upon an overall average of all sequencing reads at that site [6], can be confounded by cell counts within whole blood [7]. To further confound estimates of average DNA methylation, we know that there is a strong association between DNA methylation patterns and an individual's age [8, 9]. Thus, failing to account for cellular heterogeneity can cause differences in average methylation calculations, potentially leading to a bias in results [10] and false positives [11]. In fact, it is highly likely that tissue heterogeneity is one of the main causes of lack of reproducibility of methylome studies [12]. Adjustment

tools which are specific to and can correct for the individual composition of the principle immune cells present in whole blood (B cells, granulocytes, monocytes, natural killer cells, and T cells [13]) have been developed and can be fitted when using the Illumina EPIC array system. However, this is specific to the array system only, and analysis tools for other forms of DNA methylation quantification cannot adjust for heterogeneous cell populations Despite this, whole blood is still the most commonly used tissue for studies of DNA methylation.

7.1.2 Is whole blood a good measure of overall DNA methylation?

Given that DNA methylation patterns are highly cell type-specific, it is important to question whether DNA methylation patterns in whole blood are indicative of methylation patterns in other tissues. There is evidence to support whole blood as a good overall predictor of methylation status for other tissues in some instances [8, 14]. For example, several studies have asked whether whole blood can be used as a surrogate for brain tissue [15, 16], and have found it to be concordant [17]. However, others have suggested that whole blood is not the most reliable proxy for other tissues [18, 19] and that it should be used with a side of caution [18]. Therefore, the literature around this issue presents contrasting results. The contradiction could be a consequence of the heterogeneous nature of whole blood, or may be a consequence of the heterogeneous nature of brain pathologies; diseases such as schizophrenia have multiple causes and routes to disease progression [20], which may not have a shared genomic or epigenetic basis.

7.1.3 Our investigation of differential methylation between tissue types

Thus, to investigate whether whole blood can be a good predictor of the DNA methylation status of other tissues, here, publicly available DNA methylation array data is used to ask whether brain tissue (prefrontal cortex, PFC) samples show the same significant differential DNA methylation signatures as whole blood samples, from individuals with schizophrenia. Allowing us to determine how reliable whole blood is as a proxy tissue for assessing DNA methylation that might be relevant to phenotypes

that manifest in the brain. Although using samples with CP would have been more fitting, we instead picked schizophrenia as a more defined disease. CP covers many disoders such as Autism, ADD and ADHD to name a few and the broadness of the term CP might induce a bias into our investigation due to diagnosis.

Answering this question is important because the health of individuals with schizophrenia can benefit from early identification and intervention [21] . Therefore, development of biomarkers to diagnose or identify at-risk individuals that is non-invasive, such as a blood sample, will go a long way to aiding in early diagnosis and therapy. Further, this work will contribute to a better understanding of the value of proxy tissues in DNA methylation-disease associations.

7.2 Methods

7.2.1 Acquiring data

All data was acquired through the National Genomics Data Centre (NGDC), https://bigd.big.ac.cn/ewas/index. The database consists of epigenome-wide association studies, and processed data is available for open access [22]. The database is made up of Illumina 27k, 450k and EPIC array where a total of 3,087 (as of May 2020) cohort studies have been collected. Through this database one may search a specific CpG site, a trait of interest, a cohort, tissue or cell type, or a specific study or a publication. The keyword used to search for the data used in this chapter was “schizophrenia”. Studies associated with this trait were then listed with further details and the data was selected for inclusion based on a strict criterion for omission.

7.2.2 Inclusion/exclusion criteria

All data needed to be generated from either the Illumina 27k, 450k or EPIC array system. Although there are other ways of assessing the genome for DNA methylation changes, here we specifically target studies which utilised the array system to assess accordance. All cohort data in response to the searched trait needed to be peer reviewed and published prior to data being acquired.

All available data for a certain tissue sample was used. Studies were dismissed if the tissue definition was not clear or whole blood samples were sub selected for either buffy coat or leukocytes.

The number of individuals in the study and the ethnicity of participants varied between studies. There was no minimum or maximum number of individuals within a study classified in our criteria for inclusion. Ethnicity was not stated in every study, however we included all available data even where ethnicity was not clear. Pre-processing and model design of data was also varied between studies. No exclusion was made on the basis of the pre-processing or statistical design, as studies have been peer reviewed prior to publication and this was seen as stringent enough.

All data derived for this analysis (Table 7.1) came from individuals who had been diagnosed via the Diagnostic and Statistical Manual of Mental Disorders (*DSM-IV)*. The matched controls were free from psychiatric and or neurological diagnoses and substance abuse according to *DSM-IV*. Studies using whole blood samples had a minimum age criteria of 18 years old. The same criteria was not applied to samples of brain tissue due to the difficulty of acquiring any type of brain tissue.

Table 7.1 Previously published studies used in the analysis of DNA methylation changes in individuals with schizophrenia from PFC and whole blood samples.

Study ID	Tissue Type	Publication/Reference
ES00498	Prefrontal cortex (PFC)	[23]
ES00841	Whole blood (WB)	[24]

7.2.3 Methodology between studies for assessing top CpG sites

Each of the top hits/statistically significant CpG sites from the studies in Table 7.1 were downloaded as CSV files. Statistically significant CpG sites/top hits were determined slightly differently between each of the studies. Across all studies X and Y chromosomes were excluded and SNPs were removed. Pre-processing methods were varied, as well as multiple comparison testing methods.

More specifically, for the schizophrenia study, the PFC study [23] determined significance of differential methylation via a linear model, and CpG sites were adjusted based upon the conservative method of Bonferroni correction. However, this contrasts with the whole blood study [24], which was adjusted for sex, age, race/ethnicity, smoking status, estimates cell proportions of six cell types and the first two principal components. An FDR cut off of less than 0.2 was used for significance for whole blood.

For each significant CpG site, data files included the Illumina ID, the region of each CpG site in the genome, associated genes, associated trait, and associated Genome wide association studies (GWAS). Additional columns for tissue of origin and study ID were also included.

7.2.4 Assessing CpG locations

Data was then imported into R statistical software (Version 3.5.2). Unique statistically significant CpG sites/probes were counted in response to each tissue type from each of the studies. CpG sites were then assigned to their known associated gene. Genes were searched in genecards [25] to look for mRNA expression in relevant tissue. Differentially methylated CpG sites that were intergenic were matched to the nearest neighbouring genes within Hg19 using Granges default settings [26], and the official gene symbols of all significantly differentially methylated CpG sites were obtained. Pathway analysis was carried out using KEGG 2019 human pathways with EnrichR [27], and P values were adjusted using FDR. Tables and Venn diagrams were constructed in R studio.

7.3 Results

7.3.1 Is whole blood telling the story of the brain?

The two separate studies used for this analysis was found using the EWAS Atlas database. A total of 2357 individuals (PFC, N= 526, whole blood, N= 1831) were used for investigating differential DNA methylation. Ethnicity varied between samples obtained from whole blood and PFC, however, both sample types contained European individuals. Whole blood samples which also contained individuals of African American or Afro Caribbean origin, reported all ethnicities of their cohort, whereas the other ethnicities of the PFC samples were not reported.

Table 7.2 Cohort characteristics in the studies analysed based off whole blood and prefrontal cortex- WB whole Blood, PFC prefrontal cortex.

	WB	PFC
The number of individuals in each of the cohort studies	1831	526
Ethnicity	African American Afro Caribbean European	European Not reported

7.3.2 Differentially methylated CpG sites

Significantly differentially methylated probes were counted and recorded between the two sample types (Table 7.3). There were more differentially methylated probes identified in prefrontal cortex (N= 1772) compared to that of the whole blood (N= 95) of individuals with schizophrenia. No overlap in significant CpG sites was observed between whole blood and PFC (Figure 7.1).

Table 7.3 Differentially methylated CpG sites found within whole blood (WB) and prefrontal cortex (PFC).

	WB	PFC
The number of unique probes (excluding duplicate probes/CpG's from the individual cohorts)	95	1772
The number of CpG sites that were found to be statistically significant that were associated with a named gene	77	1567

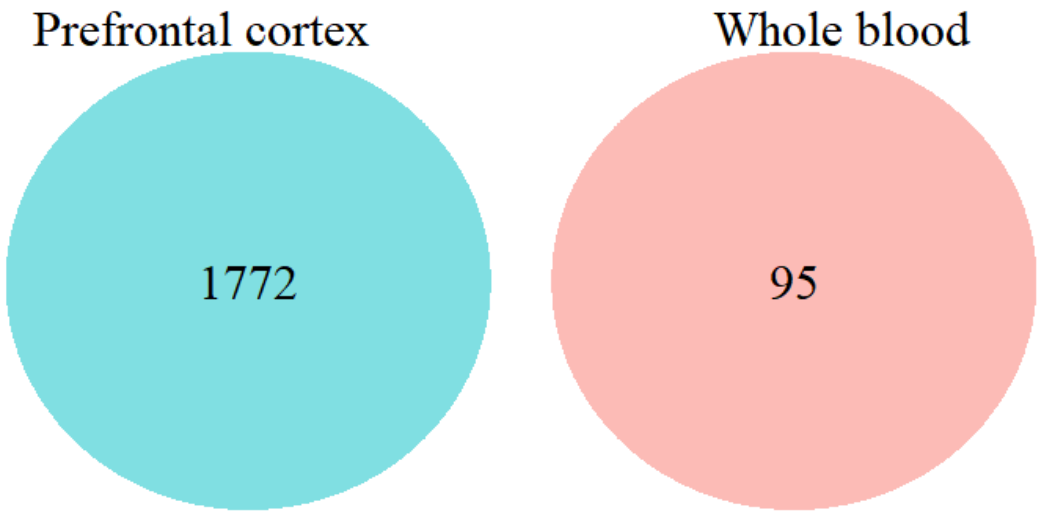


Figure 7.1- The number of CpG sites that were statistically significant within the prefrontal cortex (salmon) and whole blood sample (blue). Within these tissue types there was no overlap between CpG sites found.

Of the total number of CpG sites that have a known gene associated with them, N= 8 genes were identified in both tissue types (Table 7.4). A total of N= 1559 stayed unique to prefrontal cortex samples and N= 69 to whole blood samples. Seven of the eight genes that were shared between tissues (with the exception of *CTD-2175A23*) are expressed in both whole blood and brain tissues. The direction of methylation change was consistent at CpG sites within half of the overlapping genes, *RPTOR* (hypermethylated in both tissues*), CTD-2175A23.1*, *GFI1* and *KIFC3* (all hypomethylated in both tissues). The remaining four genes showed contrasting direction of methylation change between whole blood and prefrontal cortex. A literature search of the shared genes indicated that two have roles in schizophrenia, and a further five have roles in brain and neurological impairment/disease.

Table 7.4 CpG site locations within common genes of both whole blood and prefrontal cortex tissue. Functional associations determined via genome wide association studies are cited. Hyper= hypermethylation, Hypo = hypomethylation.

Gene	Tissue type	Cg Identifier	Correlation	Location	CpG island	Functional association
RPTOR	Whole blood	cg13549638	Hyper	chr17: 78860076	Shelf	Schizophrenia [28]
	Whole blood	cg16660971	Hyper	chr17: 78860029	Shelf	
	Whole blood	cg27457201	Hyper	chr17: 78854232	Shelf	
	Prefrontal cortex	cg22882460	Hyper	chr17: 78654648	Other	
SEC14L1	Whole blood	cg20610950	Hyper	chr17: 75096202	Other	Cognitive performance [29]
	Whole blood	cg11597902	Hyper	chr17: 75096239	Other	
	Whole blood	cg11186858	Hyper	chr17: 75096382	Other	
	Prefrontal cortex	cg26547236	Hypo	chr17: 75136326	Island	
FBXO46	Whole blood	cg09277709	Hyper	chr19: 46224285	Shelf	Alzheimer's disease [30]
	Prefrontal cortex	cg26562171	Hypo	chr19: 46220049	Shelf	
CTD-2175A23.1	Whole blood	cg05036937	Hypo	chr5: 52283760	Shore	
	Prefrontal cortex	cg15676241	Hypo	chr5: 52285231	Island	
GMDS	Whole blood	cg06315217	Hyper	chr6: 1629850	Other	Grey matter volume [31], Depression in smokers [32], PHF-tau measurement [33]
	Prefrontal cortex	cg08932320	Hypo	chr6: 2246077	Island	
GFI1	Whole blood	cg04535902	Hypo	chr1: 92947332	Island	Multiple sclerosis [34, 35]
	Whole blood	cg04777348	Hypo	chr1: 92952897	Shore	
	Whole blood	cg24517501	Hypo	chr1: 92952702	Shore	
	Prefrontal cortex	cg14475915	Hypo	chr1: 92952268	Island	
KIFC3	Whole blood	cg01115923	Hypo	chr16: 57793728	Shore	Alzheimer's disease [36]
	Prefrontal cortex	cg07685869	Hypo	chr16: 57836706	Island	
MAD1L1	Whole blood	cg25323444	Hyper	chr7: 2111060	Shelf	Schizophrenia [37]
	Prefrontal cortex	cg20935553	Hypo	chr7: 2272059	Island	

7.3.3 Pathway analysis of prefrontal cortex tissue and whole blood

Genes in which significantly differentially methylated CpG sites resided were then analysed via Enrichr to determine which KEGG pathways were enriched in whole blood (Table 7.5) and prefrontal cortex (Table 7.6). For whole blood, six pathways were found to be nominally enriched. These pathways included the long-term potentiation, mTOR signalling, and the mRNA surveillance pathways.

Table 7.5 List of KEGG pathways calculated from gene lists containing statistically significant CpG sites found between whole blood and individuals with schizophrenia.

Name	*P value*	*Adjusted P value*	Odds Ratio	Combined score
Glycosaminoglycan biosynthesis	0.001	0.331	14.90	101.78
Insulin resistance	0.008	1.000	7.31	35.21
Long-term potentiation	0.026	1.000	7.86	28.41
Oocyte meiosis	0.012	1.000	6.32	27.92
mTOR signalling pathway	0.020	1.000	5.19	20.27
mRNA surveillance pathway	0.046	1.000	5.78	17.69

When genes that housed significantly differentially methylated CpG sites in prefrontal cortex tissue were analysed, 23 KEGG pathways were identified as nominally significantly enriched. The top four KEGG pathways were cancer-related pathways. There were two pathways which overlapped between the tissues; the mTOR signalling pathway, and the mRNA surveillance pathway.

Table 7.6 List of KEGG pathways calculated from gene lists containing statistically significant CpG sites, found between whole blood and individuals with schizophrenia.

Name	*P Value*	*Adjusted P Value*	Odds Ratio	Combined score
Basal cell carcinoma	0.0004	0.141	2.72	20.90
Hepatocellular carcinoma	0.001	0.180	1.89	12.79
Thyroid cancer	0.008	0.302	2.65	12.51
Breast cancer	0.001	0.193	1.92	12.01
Propanoate metabolism	0.013	0.289	2.68	11.60
Gastric cancer	0.002	0.173	1.89	11.51
Nucleotide excision repair	0.012	0.300	2.34	10.24
Cushing syndrome	0.003	0.231	1.82	10.14
mRNA surveillance pathway	0.006	0.295	2.02	10.1
Wnt signalling pathway	0.004	0.247	1.78	9.52
Melanogenesis	0.007	0.295	1.94	9.44
N-Glycan biosynthesis	0.018	0.320	2.20	8.76
Lysosome	0.011	0.340	1.79	8.07
Cell cycle	0.011	0.335	1.78	7.86
mTOR signalling pathway	0.012	0.311	1.69	7.46
Colorectal cancer	0.022	0.339	1.85	7.06
AMPK signalling pathway	0.017	0.345	1.73	6.97
RNA transport	0.015	0.316	1.63	6.81
Signalling pathways regulating pluripotency of stem cells	0.018	0.331	1.67	6.69
Hippo signalling pathway	0.020	0.336	1.61	6.23
Acute myeloid leukemia	0.040	0.546	1.85	5.93
Pancreatic cancer	0.040	0.568	1.80	5.75
Lysine degradation	0.048	0.628	1.87	5.63

7.4 Discussion

Selecting the most appropriate tissue type for investigating differential DNA methylation is an important aspect of any epigenetic study. However, this is sometimes a non-negotiable aspect of a study design and usually analysis is undertaken of the tissue type that is most readily available, even if the sample is not the most appropriate for answering the hypothesis. This is because, often the "best" (most disease-specific) tissue is also the most invasive, therefore collecting those samples is not suitable. To advance our understanding of the role of DNA methylation in disease, it is important to gauge whether 'proxy' tissues (e.g. whole blood) reflect the same disease-associated DNA methylation differences as disease-specific tissue (e.g. brain tissue).

7.4.1 Schizophrenia cohort characteristics

Thus, to address the value of whole blood as a proxy tissue in examining the role of DNA methylation in the development of schizophrenia, we utilised publicly available DNA methylation array data and analysed the top significantly differentially methylated CpG sites from individuals with schizophrenia in two different tissue types: prefrontal cortex, which represents the specific site in the body affected by the disease, and whole blood, which is, as mentioned previously, the most common proxy tissue used in methylation analyses.

There was very little available DNA methylation data from previously published schizophrenia studies via the EWAS database. We believe this is because brain tissue can only be sampled postmortem. Nevertheless, three studies were identified, but only two of them passed our inclusion criteria, due to the source of sample.

7.4.2 Schizophrenia differential DNA methylation between prefrontal cortex and whole blood

A total of N= 1772 significantly differentially methylated CpG sites were observed in the prefrontal cortex group (N= 526 individuals). In contrast, a total of N= 95 statistically significant CpG sites found from the whole blood samples (N= 526 individuals). We

hypothesise that this is due to the cellular composition of the prefrontal cortex, this will be further discussed in section 7.4.4. Further, different correction methods were used to determine statistically significant CpG sites: prefrontal cortex samples were adjusted using Bonferroni correction of $P > 0.05$, which is a conservative method of testing, whereas whole blood significance was determined using a threshold of $P > 0.2$ FDR threshold, which essentially allows 20% of 'significant' CpG sites to be a false positive. Without access to the raw data, we are unable to correct for this discrepancy here, however we are aware that this, along with the heterogeneous nature of whole blood, is likely to be driving the difference in the number of statistically significant CpG sites identified in each study (but not the assignment of the 'top hits' themselves).

We did not detect an overlap between CpG sites found to be statistically significant between the prefrontal cortex tissue cohort and the whole blood cohort. However, this is not unexpected; we know that schizophrenia, like many other complex diseases, is highly heterogeneous and thus, lack of concordance between statistically significant CpG sites identified between whole blood and prefrontal cortex might be an offset characteristic of the disease itself. One may expect some overlap between truly significant CpG sites if, on an individual level, they were affecting gene transcription to such an extent that they were playing a major role in disease, however, like SNPs in genome wide association studies, finding one particular nucleotide (or in this case, CpG site) that associates strongly with a heterogeneous disease, across populations, is rare, with phenotypes being a product of multiple loci, each with a small individual effect [38].

In saying this, a further reason for this lack of concordance may be a consequence of the environmentally-induced nature of DNA methylation and differing cell type lifespans. Specifically, whole blood cells have a shorter life span compared to cells in the brain. A monocyte, one type of white blood cell in whole blood, has a life span of 24 hours [39]. When this is compared to a neuron, a key cell in the prefrontal cortex [40, 41], which is debated as to whether or not it undergoes replicative aging at all [42]. Implying that a neuron has a maximum lifespan is similar to that of the individual [43]. Therefore, given DNA methylation can be induced by the environment [44] this would potentially allow prefrontal cortex cells to accumulate many more DNA methylation differences over its lifespan than a white blood cell [45], meaning that, when compared to DNA methylation patterns in white blood cells, any pathological DNA methylation

differences might be lost due to 'noise' in the prefrontal cortex cells. This is supported by our results that suggest differential methylation is more widespread in prefrontal cortex samples than whole blood. This would also make estimates of concordance of individual CpG sites between tissues more challenging, since whole blood samples may not show long term effects of DNA methylation that have occurred in the prefrontal cortex samples due to their much higher turnover rate.

7.4.3 The overlapping genes found to contain differentially methylated CpG sites

Interestingly, when we annotated the CpG sites to their gene (or nearest gene), a total of eight genes, *RPTOR, SEC14L1, FBXO46, CTD-2175A23.1, GMDS, GFI1, KIFC3* and *MAD1L1,* overlapped between prefrontal cortex and whole blood (Table 7.4). When we further investigated what tissues expressed these genes, all apart from one, *CTD-2175A23.1* (gene expression not known), were found to be ubiquitously expressed in circulating blood and the brain. Of these eight genes, seven have been identified via genome wide association studies to be implicated in schizophrenia and other related brain pathologies. For example, SNPs in *RPTOR* and *MAD1L1* have both been associated with schizophrenia [28, 37]. The loss of grey matter phenotype associated with SNPs in *GMDS* is a pathology common to schizophrenia patients [46]. The genes *FBXO46* and *KIFC3* are associated with dementia; meta-analyses suggest that individuals with schizophrenia have an increased risk of dementia [47] and that the shared psychiatric symptoms in schizophrenia and Alzheimer's disease, along with the effects on the dopaminergic/cholinergic axis common in both diseases, suggest similarities in the pattern of regional brain dysfunction [48]. From this, we suggest that, while individual CpG sites are not conserved between tissue types, the genes that are specifically differentially methylated and associate with the phenotype under investigation are conserved between tissue types, marking molecular pathways that are relevant to the phenotype. Thus we can conclude from this that, in this study, the two tissue types are consistent in their identification of differential methylation at phenotypically relevant genes, and we stress that concordance between tissue types across studies should focus on the genes in which differential methylation is detected, rather than individual CpG sites.

In addition to this, and as indicated in Table 7.4, only half of the overlapping genes showed the same observed direction of differential methylation at CpG sites. For example, within *GMDS*, whole blood identified a CpG site that was hypermethylated, while a hypomethylated CpG was identified in prefrontal cortex samples. We do not consider this evidence of discordance between tissue samples. This is because: i) the CpG sites are at different locations in the gene, DNA methylation is dynamic and known to vary between different genomic locations within the same gene, thus we have no reason to expect the direction of change to match, and; ii) we cannot assume that hypo- vs. hypermethylation indicates either positive or negative effect on resulting gene expression, when evidence suggests that either could be the result [49]. Thus, concordance between tissues cannot be rejected on the bases of the direction of differential methylation at CpG sites within shared genes.

7.4.4 Pathway analysis of genes found to be associated in prefrontal cortex and whole blood

Pathway analysis was then carried out on these same gene lists. Table 7.5 displays the CpG sites in genes from the prefrontal cortex pathway analysis. There was a total of 23 pathways which had nominal significance, the majority of which were cancer related. We hypothesise that the abundance of cancer-related pathways identified may be a potential consequence of the low turnover of cells in the prefrontal cortex compared to whole blood; given the known accumulation of cancer-causing mutations over the lifespan, the low cell turnover may be allowing the accumulation in these cells of differential methylation in genes that could play a role in cancer. Therefore, the identification of numerous cancer-related pathways could be interpreted as biological noise in response to our trait of interest, schizophrenia, or also a consequence of the inherent knowledge bias towards cancer that exists in such databases.

Whole blood pathway ontology (Table 7.6) showed that the CpG sites within genes were enriched in just six KEGG pathways. However, two of these pathways, mTOR signalling pathway, and mRNA surveillance were found in both prefrontal cortex tissue and whole blood. The mechanistic Target of Rapamycin (mTOR), is an important pathway during neurodegeneration [50-52]. It has been hypothesised that the pathway prevents apoptotic cell death in the nervous system [53], and loss of mTOR leads to

apoptosis of neuronal cells [54]. There are several hypothesis as to how mTOR dysfunction is linked to schizophrenia [55-60], one of which is that disruption in the pathway when influenced by several extracellular and environmental factors could have implications for the onset of schizophrenia [61]. Since we know that DNA methylation is heavily influenced by the environment, it is feasible that environmental factors may be influencing methylation at CpG sites of relevance to the mTOR pathway, which could be contributing to the pathology of schizophrenia.

The second pathway, mRNA surveillance, is also associated with neurodegenerative conditions via its role in the prevention of the production of potentially toxic proteins in protein aggregation. Loss of mRNA surveillance has been shown to lead to an increase in protein aggregation in the brain [62]. More so, work in human sibling pairs indicated that, compared to human reference sequence, brothers affected with childhood onset schizophrenia and autism spectrum disorders were found to have a mutation in the gene, *UPF3B* [63]. The gene encodes for a complex involved in mRNA surveillance and has been found to regulate expression and degradation of various mRNA present at the synapse [64].

Thus, the pathway analysis carried out for schizophrenia provided interesting findings based upon the genes containing differential DNA methylation at CpG sites. Specifically, while there was little overlap in individual CpG sites, studies looking at the same phenotype in different tissue types showed enrichment for the same KEGG pathways, both of which had biological relevance to the disease in question. We hypothesis that although specific CpG sites are not differentially methylated in response to tissue type, the genes that are potentially causing dysregulation maybe the same in a wider network.

A heterogeneous disease such as schizophrenia does not have one clear disease-causing mutation. Therefore, it would also be highly unlikely that the same exact CpG sites in an individual are causing disease. The results of this study supports the heterogeneous nature of schizophrenia development, and supports the assertion that dysregulation within gene networks are more important to disease development, rather than DNA methylation at any one CpG site in particular. Thus, while different tissues display different differentially methylated CpG sites, this work demonstrates that genes and functional pathways of relevance to the disease can be shared across different

tissues. Further, accepting that more definitive research needs to be done in this area, e.g. comparing prefrontal cortex and whole blood from the same patient, we would suggest that whole blood is not an inferior tissue for studies of organ-specific diseases, and that tissue-specific sampling may not always necessary, particularly when that sampling is highly invasive.

7.4.4 Limitations of this analysis

Meta analyses usually have a stringent inclusion and exclusion criteria. In this analysis, data was firstly limited and therefore inclusion rules were somewhat less stringent. Although this objective was to try and combat a major limitation that we face when investigating DNA methylation, it also brings up its own limitations. When analysing metadata or taking top hits from selected cohorts you are rarely working with a uniformed pipeline of analysis. For instance, one cohort analysed may have used algometric tools which were needed for the integrity of their study design, but then not applicable for other study designs. Every study design is unique, therefore the way in which the analyses is carried out is different. Therefore, we entrust that data which is made publicly available is to the highest degree of integrity and has been pre-processed adequately. However, it would be naive to say data processing methods could be ignored when interpreting the findings of our results. Unified pipelines would be the best way to ensure there is as little variation between each individual analysis. However, implementing this would be very difficult, as gaining access to raw data is challenging. For the most part, we had to trust the fact that all studies presented where data was analysed in response to the variable “schizophrenia” were accurate. Here we entrust the stringent peer review process that would have queried misinformation or biases in the data sets we chose to examine.

It is important to note here, that although we found genes that displayed commonality between the different tissue groups, the actual CpG site of differential DNA methylation differed in all instances. Within a gene, there can be a multitude of CpG sites, especially in promoter regions of genes. So, conclusions on the relevance of any one CpG site to the phenotype should be made cautiously. When pathway analysis was performed, we found nominal significance in multiple pathways. Although this is a consequence of sample size, further work here needs to be carried out to be able to address this limitation.

One last issue that needs to be addressed is that of individual sample variance. Each of these tissue types shows a unique pattern, meaning that there will be a large amount of variation present. The variation can be caused by range of factors (underlying permanent epigenetic variation, environmental variation) and this variation can add up. It is normal to see some variation. However; it becomes a major problem when the variation fluctuates across the individual cohorts, particularly when we want to look at datasets as a collective. We often refer to this as heteroscedasticity, so for instance, the presence of variability differences between two groups e.g. whole blood vs prefrontal cortex. Heteroscedasticity could potentially be confounding these results, because more variability will be present between samples of whole blood; the multiple cell types present in whole blood means that they are more likely to show more variation across individuals. In contrast to the prefrontal cortex, where variation across the majority of sites analysed will be more uniform, such that heteroscedasticity could be considered to be "non-applicable" to prefrontal cortex samples. Therefore, heteroscedasticity may be a reason why we see more significant CpG sites associated with the prefrontal cortex, when compared to whole blood. However, this is something that in this study we are unable to investigate as we are only taking "top hits". However, conducting analysis using log fold form data is one way to overcome this problem.

The observations in this chapter have provided support for the use of whole blood as a proxy tissue for brain pathologies. Given the biological and phenotypic relevance of the genes and pathways we identify between tissues, we suggest that, in order to make the best use of proxy tissues for heterogeneous diseases such as schizophrenia, studies should focus on the genes and the pathways that house differentially methylated CpG sites, rather than individual CpG sites themselves.

7.5 Chapter summary

- We found no overlap between prefrontal cortex tissue and whole blood at specific differentially methylated CpG sites in individuals with schizophrenia in the two studies used in this analysis.
- KEGG pathway analysis of the genes that housed significantly differentially methylated CpG sites in each tissue revealed that two major pathways, the mTOR signalling pathway and the mRNA surveillance pathway, which both play a role in neurodegeneration, were enriched in both tissue types.
- We hypothesise that in complex diseases such as schizophrenia, differential methylation at individual CpG sites is less informative than the identification of the precise genes which house those CpG sites.
- This is potentially due to individual variation in DNA methylation which leads to different CpG sites being differentially methylated in different individuals, especially when assessing whole blood samples.

7.6 References

1. Lokk, K., et al., *DNA methylome profiling of human tissues identifies global and tissue-specific methylation patterns.* Genome biology, 2014. **15**(4): p. r54-r54.
2. Jiang, R., et al., *Discordance of DNA Methylation Variance Between two Accessible Human Tissues.* Scientific Reports, 2015. **5**(1): p. 8257.
3. Mill, J. and B.T. Heijmans, *From promises to practical strategies in epigenetic epidemiology.* Nature Reviews Genetics, 2013. **14**(8): p. 585-594.
4. L, D., *Blood Groups and Red Cell Antigens [Internet]*. 2005: Bethesda (MD): National Center for Biotechnology Information (US).
5. Buenrostro, J.D., et al., *Single-cell chromatin accessibility reveals principles of regulatory variation.* Nature, 2015. **523**(7561): p. 486-490.
6. Frommer, M., et al., *A genomic sequencing protocol that yields a positive display of 5-methylcytosine residues in individual DNA strands.* Proceedings of the National Academy of Sciences of the United States of America, 1992. **89**(5): p. 1827-1831.
7. Tollefsbol, T., *Personalized Epigenetics*. 2015: Elsevier Science.
8. Horvath, S., et al., *Aging effects on DNA methylation modules in human brain and blood tissue.* Genome Biology, 2012. **13**(10): p. R97.
9. Martin, G.M., *Epigenetic drift in aging identical twins.* Proceedings of the National Academy of Sciences of the United States of America, 2005. **102**(30): p. 10413-10414.
10. Altschuler, S.J. and L.F. Wu, *Cellular heterogeneity: do differences make a difference?* Cell, 2010. **141**(4): p. 559-563.
11. Jaffe, A.E. and R.A. Irizarry, *Accounting for cellular heterogeneity is critical in epigenome-wide association studies.* Genome Biology, 2014. **15**(2): p. R31.
12. Marusyk, A. and K. Polyak, *Tumor heterogeneity: causes and consequences.* Biochimica et biophysica acta, 2010. **1805**(1): p. 105-117.
13. Houseman, E.A., et al., *DNA methylation arrays as surrogate measures of cell mixture distribution.* BMC bioinformatics, 2012. **13**: p. 86-86.
14. Li, X., W. Li, and Y. Xu, *Human Age Prediction Based on DNA Methylation Using a Gradient Boosting Regressor.* Genes, 2018. **9**(9): p. 424.
15. Walton, E., et al., *Correspondence of DNA Methylation Between Blood and Brain Tissue and Its Application to Schizophrenia Research.* Schizophrenia Bulletin, 2015. **42**(2): p. 406-414.
16. Davies, M.N., et al., *Functional annotation of the human brain methylome identifies tissue-specific epigenetic variation across brain and blood.* Genome Biology, 2012. **13**(6): p. R43.
17. Masliah, E., et al., *Distinctive patterns of DNA methylation associated with Parkinson disease: identification of concordant epigenetic changes in brain and peripheral blood leukocytes.* Epigenetics, 2013. **8**(10): p. 1030-1038.
18. Reinius, L.E., et al., *Differential DNA methylation in purified human blood cells: implications for cell lineage and studies on disease susceptibility.* PloS one, 2012. **7**(7).
19. Münzel, M., et al., *Quantification of the Sixth DNA Base Hydroxymethylcytosine in the Brain.* Angewandte Chemie International Edition, 2010. **49**(31): p. 5375-5377.
20. Jablensky, A., *The diagnostic concept of schizophrenia: its history, evolution, and future prospects.* Dialogues in clinical neuroscience, 2010. **12**(3): p. 271-287.
21. Kulhara, P., A. Banerjee, and A. Dutt, *Early intervention in schizophrenia.* Indian journal of psychiatry, 2008. **50**(2): p. 128-134.
22. Li, M., et al., *EWAS Atlas: a curated knowledgebase of epigenome-wide association studies.* Nucleic Acids Res, 2019. **47**(D1): p. D983-d988.
23. Jaffe, A.E., et al., *Mapping DNA methylation across development, genotype and schizophrenia in the human frontal cortex.* Nature Neuroscience, 2016. **19**(1): p. 40-47.
24. Montano, C., et al., *Association of DNA Methylation Differences With Schizophrenia in an Epigenome-Wide Association Study.* JAMA Psychiatry, 2016. **73**(5): p. 506-14.
25. Stelzer, G., et al., *The GeneCards Suite: From Gene Data Mining to Disease Genome Sequence Analyses.* Current Protocols in Bioinformatics, 2016. **54**(1): p. 1.30.1-1.30.33.
26. Lawrence, M., et al., *Software for computing and annotating genomic ranges.* PLoS computational biology, 2013. **9**(8).
27. Kuleshov, M.V., et al., *Enrichr: a comprehensive gene set enrichment analysis web server 2016 update.* Nucleic acids research, 2016. **44**(W1): p. W90-W97.

28. Goes, F.S., et al., *Genome-wide association study of schizophrenia in Ashkenazi Jews.* American Journal of Medical Genetics Part B: Neuropsychiatric Genetics, 2015. **168**(8): p. 649-659.
29. Need, A.C., et al., *A genome-wide study of common SNPs and CNVs in cognitive performance in the CANTAB.* Human Molecular Genetics, 2009. **18**(23): p. 4650-4661.
30. Kwok, M.K., S.L. Lin, and C.M. Schooling, *Re-thinking Alzheimer's disease therapeutic targets using gene-based tests.* EBioMedicine, 2018. **37**: p. 461-470.
31. Alliey-Rodriguez, N., et al., *NRXN1 is associated with enlargement of the temporal horns of the lateral ventricles in psychosis.* Transl Psychiatry, 2019. **9**(1): p. 230.
32. Heinzman, J.T., et al., *GWAS and systems biology analysis of depressive symptoms among smokers from the COPDGene cohort.* J Affect Disord, 2019. **243**: p. 16-22.
33. Wang, H., et al., *Genome-wide interaction analysis of pathological hallmarks in Alzheimer's disease.* Neurobiology of Aging, 2020. **93**: p. 61-68.
34. Andlauer, T.F., et al., *Novel multiple sclerosis susceptibility loci implicated in epigenetic regulation.* Sci Adv, 2016. **2**(6): p. e1501678.
35. *Multiple sclerosis genomic map implicates peripheral immune cells and microglia in susceptibility.* Science, 2019. **365**(6460).
36. Mez, J., et al., *Two novel loci, COBL and SLC10A2, for Alzheimer's disease in African Americans.* Alzheimers Dement, 2017. **13**(2): p. 119-129.
37. Bergen, S.E., et al., *Genome-wide association study in a Swedish population yields support for greater CNV and MHC involvement in schizophrenia compared with bipolar disorder.* Mol Psychiatry, 2012. **17**(9): p. 880-6.
38. Eichler, E.E., et al., *Missing heritability and strategies for finding the underlying causes of complex disease.* Nature Reviews Genetics, 2010. **11**(6): p. 446-450.
39. Monie, T.P., *Section 1 - A Snapshot of the Innate Immune System*, in *The Innate Immune System*, T.P. Monie, Editor. 2017, Academic Press. p. 1-40.
40. Puig, M.V. and A.T. Gulledge, *Serotonin and prefrontal cortex function: neurons, networks, and circuits.* Molecular neurobiology, 2011. **44**(3): p. 449-464.
41. Gabi, M., et al., *No relative expansion of the number of prefrontal neurons in primate and human evolution.* Proceedings of the National Academy of Sciences, 2016. **113**(34): p. 9617-9622.
42. Magrassi, L., K. Leto, and F. Rossi, *Lifespan of neurons is uncoupled from organismal lifespan.* Proceedings of the National Academy of Sciences, 2013. **110**(11): p. 4374-4379.
43. Eriksson, P.S., et al., *Neurogenesis in the adult human hippocampus.* Nature Medicine, 1998. **4**(11): p. 1313-1317.
44. Flores, K.B., F. Wolschin, and G.V. Amdam, *The Role of Methylation of DNA in Environmental Adaptation.* Integrative and Comparative Biology, 2013. **53**(2): p. 359-372.
45. Réu, P., et al., *The Lifespan and Turnover of Microglia in the Human Brain.* Cell Rep, 2017. **20**(4): p. 779-784.
46. Vita, A., et al., *Progressive loss of cortical gray matter in schizophrenia: a meta-analysis and meta-regression of longitudinal MRI studies.* Transl Psychiatry, 2012. **2**(11): p. e190.
47. Cai, L. and J. Huang, *Schizophrenia and risk of dementia: a meta-analysis study.* Neuropsychiatr Dis Treat, 2018. **14**: p. 2047-2055.
48. White, K.E. and J.L. Cummings, *Schizophrenia and Alzheimer's disease: Clinical and pathophysiologic analogies.* Comprehensive Psychiatry, 1996. **37**(3): p. 188-195.
49. Wan, J., et al., *Characterization of tissue-specific differential DNA methylation suggests distinct modes of positive and negative gene expression regulation.* BMC Genomics, 2015. **16**(1): p. 49.
50. Maiese, K., *Stem cell guidance through the mechanistic target of rapamycin.* World journal of stem cells, 2015. **7**(7): p. 999-1009.
51. Johnson, S.C., et al., *Modulating mTOR in aging and health.* Interdiscip Top Gerontol, 2015. **40**: p. 107-27.
52. Maiese, K., *The mechanistic target of rapamycin (mTOR) and the silent mating-type information regulation 2 homolog 1 (SIRT1): oversight for neurodegenerative disorders.* Biochemical Society Transactions, 2018. **46**(2): p. 351-360.
53. Chong, Z.Z., et al., *Shedding new light on neurodegenerative diseases through the mammalian target of rapamycin.* Progress in Neurobiology, 2012. **99**(2): p. 128-148.
54. Chen, L., et al., *Hydrogen peroxide inhibits mTOR signaling by activation of AMPKα leading to apoptosis of neuronal cells.* Laboratory Investigation, 2010. **90**(5): p. 762-773.

55. Howell, K.R. and A.J. Law, *Neurodevelopmental concepts of schizophrenia in the genome-wide association era: AKT/mTOR signaling as a pathological mediator of genetic and environmental programming during development.* Schizophrenia Research, 2020. **217**: p. 95-104.
56. Pham, X., et al., *The DPYSL2 gene connects mTOR and schizophrenia.* Translational Psychiatry, 2016. **6**(11): p. e933-e933.
57. Meffre, J., et al., *5-HT6 receptor recruitment of mTOR as a mechanism for perturbed cognition in schizophrenia.* EMBO Molecular Medicine, 2012. **4**(10): p. 1043-1056.
58. Liu, Y., et al., *Functional Variants in DPYSL2 Sequence Increase Risk of Schizophrenia and Suggest a Link to mTOR Signaling.* G3: Genes|Genomes|Genetics, 2015. **5**(1): p. 61-72.
59. Swiech, L., et al., *Role of mTOR in physiology and pathology of the nervous system.* Biochimica et Biophysica Acta (BBA) - Proteins and Proteomics, 2008. **1784**(1): p. 116-132.
60. Ryskalin, L., et al., *mTOR-Related Brain Dysfunctions in Neuropsychiatric Disorders.* International Journal of Molecular Sciences, 2018. **19**(8): p. 2226.
61. Gururajan, A. and M. van den Buuse, *Is the mTOR-signalling cascade disrupted in Schizophrenia?* J Neurochem, 2014. **129**(3): p. 377-87.
62. Jamar, N.H., P. Kritsiligkou, and C.M. Grant, *Loss of mRNA surveillance pathways results in widespread protein aggregation.* Scientific Reports, 2018. **8**(1): p. 3894.
63. Addington, A.M., et al., *A novel frameshift mutation in UPF3B identified in brothers affected with childhood onset schizophrenia and autism spectrum disorders.* Molecular Psychiatry, 2011. **16**(3): p. 238-239.
64. Laumonnier, F., et al., *Mutations of the UPF3B gene, which encodes a protein widely expressed in neurons, are associated with nonspecific mental retardation with or without autism.* Molecular Psychiatry, 2010. **15**(7): p. 767-776.

8. Discussion

8.1 General findings of the Chapters in this thesis

The body of work presented within this thesis assessed the impact of two main environmental exposures, cannabis, and *in utero* tobacco exposure, on DNA methylation in the human and zebrafish genomes. We have presented evidence of differential DNA methylation changes in response to these exposures using both genome-wide and targeted (amplicon-based) techniques. Each of these chapters offers further insight into our understanding of the impact of environmental exposures on DNA methylation, and how this might relate to adverse phenotypic effects, but they also highlight further challenges which need to be addressed.

Firstly, assessment of the response of DNA methylation in the human genome to heavy cannabis exposure was conducted using the Illumina EPIC array. We demonstrated that combining array data from different years led to a batch effect between sampling years, and that choice of normalisation methods for sample pre-processing led to significant discrepancies in ability to correct for the batch effect between sample batches. If sampling could be carried out again, cannabis only samples and cannabis with tobacco samples would be sent at the same time. However, we determined that the tool noob was able to adequately correct for the batch effects, and did so more successfully than other methods. Differential DNA methylation was observed at a nominal level in cannabis-only users compared to controls, while in the cannabis with tobacco group, FDR adjustment levels were met at several CpG sites indicating CpG sites that were differentially methylated at the genome-wide level. KEGG pathway analysis was carried out on the genes (or nearest genes) which housed the top differentially methylated CpG sites in the cannabis-only group, and the cannabis with tobacco group. The cannabis-only genes displayed enrichment for genes involved in brain and cardiac function, whereas the cannabis with tobacco genes enriched for pathways involved in cancer. Given the observed phenotypic effects of long-term cannabis exposure in humans, these results, while nominal, are biologically meaningful and highlight a role for DNA methylation in the biological response to cannabis, and should be explored further.

Following on from these findings, we established a cost-efficient pipeline for the replication and validation of differential DNA methylation identified via EPIC array using the tool BSAS, in a targeted and high-throughput manner. The aim was to determine whether BSAS is an accurate tool for further exploration of differentially methylated CpG sites of interest. CpG sites from the cannabis with tobacco data from Chapter 2 were picked on the basis of their statistical significance (statistically significant, nominally significant and no observed difference between cannabis with tobacco users vs control). Here we found that BSAS was able to validate some CpG sites from the EPIC array but we caution that each locus should be explored individually on a small scale before being chosen for large-scale use. While BSAS was unable to reproduce the magnitude of differential methylation change shown in the EPIC array, BSAS did display some distinct advantages; it can be used to assess multiple CpG sites within a region in a gene, and therefore could be used as a tool for investigating specific differentially methylated gene regions efficiently and thoroughly.

In Chapter 4 we demonstrated that the zebrafish was a tractable model system in which to assess the impact of cannabinoid exposure on DNA methylation. Specifically, we show that THC and CBD exposure reduces zebrafish embryo hatching efficiency compared to controls, however CBD exposure shows the greatest effect. DNA methylation differences were investigated using RRBS, and we detected differential DNA methylation in response to all treatment groups, at an FDR corrected adjustment, indicting significant results at the genome-wide level. CBD exposure resulted in N= 1939 significantly differentially methylated CpG sites, and THC exposure displayed N= 9 significantly differentially methylated individual CpG sites. Intriguingly, biological pathway analysis of the genes which housed significantly differentially methylated CpG sites in response to CBD showed that these data were enriched for genes involved in cell communication and axon guidance, which was unexpected due to the non-psychoactive nature of CBD.

The impact of *in utero* tobacco exposure on DNA methylation, and the interaction between exposure and CP was quantified using BSAS. In this pilot study we identified 10 genes from the literature known to play a role in neurodevelopment to investigate this. We identified nominally significant differential DNA methylation at specific CpG sites in individuals with CP who were exposed to tobacco *in utero*. These findings

highlighted the potential role for DNA methylation in the association between *in utero* tobacco exposure and CP, and therefore we investigated this association further, at the genome-wide level, in Chapter 6.

In Chapter 6 we presented Illumina EPIC array data which assessed differential methylation under three different models: i) maternal tobacco use during pregnancy (*in utero* exposure) vs. non-exposed; ii) low CP scores vs high CP scores, and; iii) interaction between *in utero* exposure and CP score. We detected significant genome-wide DNA methylation differences between individuals exposed to tobacco *in utero* and those that were not, and this remained significant after adjustment for multiple testing. In addition, nominal significance was observed across the genome when comparing high vs. low CP scores, and when modelling the interaction between *in utero* exposure and CP score (interaction model). The top CpG sites identified under this interaction model all have functional relevance to visual impairment and brain function, suggesting that visual impairment may be an additional phenotypic response to *in utero* tobacco exposure.

Lastly, explored the use of whole blood samples for DNA methylation analysis and how indicative these marks were at predicting DNA methylation changes in brain cells, using DNA methylation in blood and brain in individuals with schizophrenia as a model. Here, we found very little overlap between differentially methylated CpG sites between whole blood and brain samples. However, KEGG pathway analysis of the genes containing the top differentially methylated CpGs from each tissue identified an overlap between whole blood and prefrontal cortex in the mTOR signalling and mRNA surveillance pathways, both of which have roles in schizophrenia, highlighting the value of whole blood as a proxy tissue for organ-specific diseases.

8.2 Contributions to the field

8.2.1 Cannabinoid exposure

Here, we have provided evidence of differential DNA methylation in response to heavy cannabis exposure in humans. Currently, there is little research that has investigated the response of DNA methylation to cannabis, which is largely due to the fact that cannabis is most often consumed in combination with tobacco. Here, we have had the unique opportunity to specifically investigate the effect of cannabis, in isolation from tobacco, on DNA methylation in the human genome. These findings has provided new insights into how cannabis alone is interacting with DNA methylation in the human genome. For example, we now know that DNA methylation is altered at genes with roles in brain and cardiac function, indicating that DNA methylation may play a role in the biological effects of cannabis. These observations can contribute to the debate around the safety and efficacy of cannabis and its constituents as a therapeutic agent, as well as contribute to the ongoing debates around decriminalisation and legalisation.

Of importance to the above, the growing popularity of medicinal cannabis, and CBD-based therapeutic products, highlights the need for investigation into the precise modes of action of each of the main ingredients of cannabis. However, this is a question that could not be readily answered in humans, therefore we sought to begin to explore this in the zebrafish. Our data showed surprising results, namely that the impact of CBD on DNA methylation was widespread and included significant differential DNA methylation at genes and pathways that function in the brain. Therefore, our RRBS results seem contrary to what we would expect, given that THC is the main psychoactive component of cannabis.

Although we cannot use these data to assign positive or negative phenotypic impacts for the individual, this evidence justifies the need for further research into the precise biological impact of CBD exposure. We see this as particularly relevance given the popularity of CBD for medicinal purposes. Specifically, pilot data from other groups suggests that CBD could be beneficial in the treatment of multiple sclerosis and severe epilepsy [1]. Thus, while there may be evidence for the use of cannabinoids as a therapy for multiple sclerosis and epilepsy, the unexpected nature of the identified differential DNA methylation in response to CBD implies that there is much we do not

now about the impact of CBD on the genome, and what this might mean for health. While we were not able to validate the impact of CBD and THC on gene expression within the scope of this thesis, this serves as a valuable observations around the use of cannabinoids and justifies further investigation.

8.2.2 *In utero* tobacco exposure

Maternal tobacco use during pregnancy is common, and has been associated with perinatal compromise and CP. However, CP is an umbrella term that encompasses a number of different disorders, each of which may be influenced by numerous genetic, environmental, or socioeconomic factors. Further, diagnosis of CP is via a numerical scale (i.e. not binary). Thus, proving that CP is definitely linked to *in utero* tobacco exposure is challenging. While nominal, our initial findings at both the amplicon and the whole genome level identified differential methylation at CpG sites that were specific to the interaction between *in utero* tobacco exposure and high CP score, supporting the role of DNA methylation in the association between *in utero* tobacco exposure and the development of CP. This provides evidence to further support the risks associated with maternal tobacco use during pregnancy, and further research into this association will support policy and education around maternal tobacco use.

Our data also provided evidence to suggest that developmentally-derived DNA methylation may be maintained into adulthood. Specifically, here we identify four differentially methylated CpG sites in the DNA of adults that remain significant after FDR correction, in response to *in utero* tobacco exposure, that are independent of adult smoking status. Three of these CpG sites have been identified as differentially methylated in response to maternal tobacco use during pregnancy in newborns and young children [2-6]. Implying that some *in utero* tobacco-induced DNA methylation changes may be stable through the life course. These sites should be further investigated, as this observation may have further implications for human health.

8.3 Avenues for further research

8.3.1 Sample size and genome-wide significance

A limitation in many human studies, including this one, is sample size and access to human DNA. The number of individuals in each of the studies assessed here are of modest size. Chapter 5 was the largest working cohort assessed in this thesis (N= 109). While our nominally significant data were biologically relevant, each hypothesis that we test is subjected to correction for multiple testing for ~850,000 tests (the number of CpG sites on the EPIC array). Due to this, it impedes our power to detect CpG sites that are significant at the genome-wide level. We hypothesise that further investigation with a sample size of at least N= 500 individuals may provide enough power to reach genome wide significance, providing further support for our conclusions. However, achieving this target number, particularly in retrospective studies such as this one, is a challenge.

An alternative approach would be to combine data from other comparable cohorts in order to validate our results. Again this approach has challenges, in particular, the majority of published data which has been discussed throughout this thesis was generated using the EPIC array predecessor, the 450K array, therefore it would remain impossible to validate half of our data points using this approach. Therefore the most tractable way forward would be to continue development of the zebrafish as a model for cannabinoid exposure and take this work forward into larger human cohorts.

8.3.2 Functional relevance of KEGG pathway analysis

Pathway analysis can support observations of differential DNA methylation by highlighting biological pathways that may be over-represented in differential methylation data. It is important because it may indicate functional relevance of a dataset, as pathway analysis tools group genes which have similar functional annotations. However, while KEGG pathway analysis can indicate the functional relevance of observed methylation changes, it remains important to link these methylation changes to a genomic output. These data serve as justification for the addition of gene expression data to complement our methylation findings. We further suggest that a useful addition to KEGG analyses here would be a functional

investigation of the impact of cannabis on the genome that specifically asks how CBD and THC affect chromatin structure and the 3-dimensional organisation of the genome. Asking how environmental exposures change genomic interactions would improve our understanding of how substances such as CBD and THC affect genome regulation, and this will increase our understanding of the mechanistic link between DNA methylation and phenotypes related to cannabis exposure.

8.3.3 Validating zebrafish data

Ideally further validation would have been carried out on the DNA methylation analysis of cannabinoid exposure. Carrying out bisulfite based amplicon sequencing to validate regions of interest would offer more robust results. Further assessing RNA-seq data would give more insight into the role these differential DNA methylation marks play in gene expression. One further point, we would also like to explore is that using a different vehicle. Ethanol is known to have profound effects on the genome and thus, could also be playing more of a role in the differences seen from each exposure group. Investigating the use of another control would offer further insight into the true effects cannabinoids are playing on the zebrafish genome.

8.4 Overall relevance

The findings we present here show how differential DNA methylation marks can be shaped by the surrounding environment. They further highlight the role of DNA methylation in the biological response to cannabis and tobacco. We have demonstrated that the environment can affect the genome during early development (*in utero*) and in adulthood. Our results also indicate that developmentally-induced changes can persist into adulthood. Further investigation is required to understanding the mechanism by which DNA methylation is contributing to disease. However, the results of the thesis contribute useful observations for future research in this area.

8.5 References

1. Stockings, E., et al., *Evidence for cannabis and cannabinoids for epilepsy: a systematic review of controlled and observational evidence.* Journal of Neurology, Neurosurgery & Psychiatry, 2018. **89**(7): p. 741-753.
2. Chatterton, Z., et al., *In utero exposure to maternal smoking is associated with DNA methylation alterations and reduced neuronal content in the developing fetal brain.* Epigenetics & chromatin, 2017. **10**: p. 4-4.
3. Chhabra, D., et al., *Fetal lung and placental methylation is associated with in utero nicotine exposure.* Epigenetics, 2014. **9**(11): p. 1473-84.
4. Rotroff, D.M., et al., *Maternal smoking impacts key biological pathways in newborns through epigenetic modification in Utero.* BMC Genomics, 2016. **17**(1): p. 976.
5. Vives-Usano, M., et al., *In utero and childhood exposure to tobacco smoke and multi-layer molecular signatures in children.* BMC Medicine, 2020. **18**(1): p. 243.
6. Joubert, B.R., et al., *Maternal Smoking and DNA Methylation in Newborns: In Utero Effect or Epigenetic Inheritance?* Cancer Epidemiology Biomarkers & Prevention, 2014. **23**(6): p. 1007-1017.