Using functional genetics and epigenetics to dissect the molecular architecture of schizophrenia.

Submitted by

Sam James Washer

to the University of Exeter

as a thesis for the degree of

Doctor of Philosophy in Medical Studies

In December 2019

I certify that all material in this thesis which is not my own work has been identified and that no material has previously been submitted and approved for the award of a degree by this or any other University.

Signature: ……..

Abstract

Schizophrenia is a severe neuropsychiatric disorder which is within the top ten causes of disability in the developed world. While we currently do not fully understand the aetiology of schizophrenia, a wealth of genetic, epigenetic, and epidemiological evidence suggests a neurodevelopmental origin, and that dysregulation of the immune system and infection can play a role in aetiology. This has been supported throughout the last decade and the “big data” revolution. Genetic and epigenetic data gathered through genome and epigenome wide association studies, have identified over 100 genetic risk loci for schizophrenia. Further to this there have been a number of epidemiological studies examining the effect of environmental risk factors to schizophrenia. However, many of these genetic, epigenetic, and environmental risk factors have no clear mechanism by which they cause disease. Many of these studies published present evidence for the “risk factor” involvement but fail to validate their findings in model systems, to establish how these risk factors cause disease.

The main aim of this thesis was to develop functional assays and methods to validate genetic, epigenetic, and environmental risk factors in human cellular based models to elucidate how these risks can contribute to disease. From this work I have developed a protocol for knocking out risk genes, and have identified a role for the schizophrenia risk gene *AS3MT* in neuronal development. I have also developed several pipelines to validate epigenome wide association studies, and link how changes in DNA methylation can affect gene expression, at both the regional level and at individual genomic loci. And finally, I have examined what effect, if any, the active component of cannabis, which is an established environmental risk factor for schizophrenia, has on the DNA methylome, to understand how environmental risk can cause disease.

In summary, the work presented in this thesis represents the movement from large scale (epi)genetic analysis and statistics to functional lab validation. This thesis contains several methods for the functional validation of identified (epi)genetic risk loci and risk factors for schizophrenia in a cellular model. From this we can now begin to explore how these changes in genetics, epigenetics, and environmental risk factors biologically link to the development and progression of schizophrenia and other neuropsychiatric disorders.

Acknowledgements

“I see no future for Sam in science” is a quote from my KS2 science teacher to my parents at parents evening. I hope this PhD thesis proves how wrong they were!

One of the key elements of a good PhD is a great supervisory team, and I would argue that I had one of the best. Therefore, I would first like to start by thanking my first supervisor, Dr Emma Dempster for all of her continued help, support, and guidance throughout my PhD. Whenever there has been an up or a down she has always been there to celebrate or comfort, she has been an inspiration and I could not have asked for a better supervisor. I would also like to thank my second supervisor, Dr Aaron Jeffries for being a fountain of knowledge and support on everything lab and computer based (despite his lack of hard copy protocols!). And finally my third supervisor, Professor Jonathan Mill, for his endless support, jokes, and for constantly mentioning my non-existent iPSC work at every conference. This supervisory team have made the last three years incredible, and I have learnt so much about how to be an outstanding researcher, for which I am forever grateful.

I would like to thank the entire Complex Disease Epigenetics Group, I have never met a bunch of people who made coming into work so much fun. Jenny, you have been so supportive of everything both in and out of work and I don’t know how I would have got through the PhD without you, no doubts that we will remain close for a very long time! Janou (Jaynow), despite being out of the country for half the time you’ve still managed to put a smile on my face with those crazy dutch “would you rathers” and jokes, you’ll always have a place to visit in the UK! Greg, for your constant banter at me for supporting the best football team in the North West, I will miss the abuse, trips to the football, obscure Simpson’s references, and to clarify, the saxophone is a woodwind instrument. Josh for your musical recommendations, I hope you consider the submission of this thesis as the passing of the “lab bitch” baton onto you, treasure it well. Thea for the support, post-it reviews of my baking, and eye contact during meetings when Jon said something stupid, no meeting will be the same again! Bex, for being the post-doc of support, I will miss popping into your office for a quick five minute chat about life, drag race, games, or what has annoyed me during the day! Barry, for being one of the only people in the lab who cares about football (although can Arsenal be called a real team?), I will miss having our Monday morning catch ups

and weekly rantings. Eilis for her continued support with the statistical aspects of the thesis. Joe for his support and guidance within the lab and for the assistance with FACS and general lab issues. I would also like to thank Rob Flynn and Michael Schrauben for their continued help with some of the projects, particularly at the end of the PhD when everything was getting very stressful, you are some of the best students I've ever had the pleasure of working with.

My sincere thanks go to Dr Asami Oguro-Ando and Dr Rosie Bamford and the Oguro-Ando Lab, for their help and guidance with the cellular assays. Despite their busy schedules, they always found time to help and assist when needed, and I will miss the CRISPR related troubleshooting, journal clubs, and student carnage!

To my parents and sister, thank you so much for your endless support for not only the last three years but for the last 25. You have supported me no matter what and have always been there for more me, despite not knowing what I talk about half the time. You have sacrificed so much to get me to where I am today, and I am forever grateful for the opportunities you have given me. But most importantly, thank you for being the most amazing and bonkers family in the world!

Finally, I would like to dedicate this work to my late grandfather Brendon, who knew I was accepted into Exeter for my undergraduate all those years ago, but never saw me start. He was the wisest, kindest, and loving man I have ever known, and his words of wisdom will stay with me forever.

List of Contents

List of Figures

List of Tables

Publications arising from this thesis

Chapter 5 (published manuscript presented in Appendix B)

Stefania Policicchio, **Sam Washer**, Joana Viana, Artemis Iatrou, Joe Burrage, Eilis Hannon, Gustavo Turecki, Zachary Kaminsky, Jonathan Mill, Emma L Dempster, Therese M Murphy (2020). Genome-Wide DNA Methylation Meta-analysis in the Brains of Suicide Completers. Translational Psychiatry, 10:69, pp1-13.

Declarations

All laboratory work and associated analysis were carried out by myself with the following exceptions:

- **Chapter 3. Characterisation and functional analysis of the schizophrenia associated gene *AS3MT*.** Neurite tracing in imageJ was performed by Josan Gandawijaya, a PhD student within the Oguro-Ando Laboratory, under my guidance.
- **Chapter 4. Transcriptomic profiling of the *AS3MT* knockout cell lines.** Library preparation and RNA sequencing was performed by Audrey Farbos and Dr Karen Moore at the University of Exeter Sequencing Service
- **Chapter 6. Functional epigenetic editing using CRISPR/Cas9.** Pyrosequencing was carried out by Michael Schrauben, a PhD student within the lab.
- **Chapter 7. Examining the effect of ∆9-tetrahydrocannabinol on the methylome.** Bisulphite conversion and Illumina EPIC arrays were run by Dr Joe Burrage, laboratory manager within the lab.

All bioinformatics analysis were carried out by myself with a few exceptions:

- De-multiplexing of raw sequence files, contamination check and spike-in internal control checks for **Chapter 4. Transcriptomic profiling of the *AS3MT* knockout cell lines** were performed by Dr Paul O'Neill at the University of Exeter Sequencing Service.

Abbreviations

Abbreviation	Term
5mC	5-methylcytosine
5hmC	5-hydroxymethylcytosine
5fC	5-formylcytosine
5caC	5-carboxylcytosine
ANOVA	Analysis Of Variance
ATP	Adenosine Triphosphate
BC	Bisulphite Converted
BDNF	Brain Derived Neurotrophic Factor
BER	Base Excision Repair
Bp	Base Pair
BSA	Bovine Serum Albumin
Btn	Biotinylated
C	Cytosine
Cas	CRISPR Associated
Cas9	CRISPR Associated Protein 9
CBD	Cannabidiol
cDNA	Complementary DNA
CGI	CpG Island
Chr	Chromosome
CI	Confidence Interval
CNV	Copy Number Variation
CpG	Cytosine-Guanine Dinucleotide
CRISPR	Clustered Regularly Interspaced Short Palindromic Repeats
CRISPRa	CRISPR Activation
CRISPRi	CRISPR Interference
CTCF	CCCTC-Binding Factor
dCas9	Dead Cas9
ddH_2O	Double Distilled Water

DEG	Differentially Expressed Gene
DMEM	Dulbecco's Modified Eagle's Medium
DMR	Differentially Methylated Region
DMSO	Dimethyl Sulfoxide
DNA	Deoxyribonucleic Acid
DNMT	DNA Methyltransferase
dNTPs	Deoxyribonucleotide Mix
DSB	Double Strand Break
dsDNA	Double Stranded DNA
DSM	Diagnostic and Statistical Manual of Mental Disorders
DTT	Dithiothreitol
E	Empty Backbone Vector
ESC	Embryonic Stem Cell
EGFP	Enhanced Green Fluorescent Protein
eQTL	Expression Quantitative Trait Loci
EWAS	Epigenome Wide Association Study
Ext	External
FACS	Fluorescent Activated Cell Sorting
FBS	Foetal Bovine Serum
FDR	False Discovery Rate
GFP	Green Fluorescent Protein
GO	Gene Ontology
gRNA	Guide RNA
GTF	General Transfer Format
GWAS	Genome Wide Association Study
HDR	Homology Directed Repair
Het	Heterozygote
HIV	Human Immunodeficiency Virus
Hom	Homozygote
IGV	Integrative Genomic Viewer

Indel	Insertion or Deletion
Int	Internal
iPSC	Induced Pluripotent Stem Cell
IQR	Interquartile Range
Kb	Kilobase
KRAB	Krüppel-associated Box
L	Ladder
LB	Lysogeny Broth
LD	Linkage Disequilibrium
M	Methylated
MAM	Methylazoxymethanol Acetate
MHC	Major Histocompatability Complex
Min	Minute
miRNA	Micro RNA
mQTL	Methylation Quantitative Trait Loci
mRNA	Messenger RNA
n	Sample Size
ncRNA	Non-coding RNA
NGS	Next Generation Sequencing
NHEJ	Non Homologous End Joining
P	Passage
PAM	Protospacer Adjacent Motif
PAGE	Polyacrylamide Gel Electrophoresis
PBS	Phosphate-Buffered Saline
PCA	Principle Component Analysis
PCR	Polymerase Chain Reaction
PDL	Poly-d-lysine
Pen/Strep	Penicillin Streptomycin
PFA	Paraformaldehyde
PGC	Psychiatric Genomics Consortium

piRNA	Piwi-interacting RNA
PLB	Passive Lysis Buffer
PNK	Polynucleotide 5’-hydroxyl-kinase
pQTL	Protein Quantitative Trait Loci
Pre-crRNA	Precursor CRISPR RNA
PRS	Polygenic Risk Score
qPCR	Quantitative Polymerase Chain Reaction
QQ	Quantile-Quantile
QC	Quality Control
RA	Retinoic Acid
RFU	Relative Fluorescent Unit
RIN	RNA Integrity Number
RISC	RNA Induced Splicing Complex
RNA	Ribonucleic Acid
rRNA	Ribosomal RNA
RNAseq	RNA Sequencing
SAM	S-adenosylmethionine
SD	Standard Deviation
SE	Standard Error
SDS	Sodium Dodecyl Sulphate
Sec	Second
siRNA	Small Interfering RNA
SMR	Summary-based Mendelian Randomisation
snRNA	Small Nuclear RNA
snoRNA	Small Nucleolar RNA
SNP	Single Nucleotide Polymorphism
S.O.C	Super Optimal broth with Catabolite repression
ssDNA	Single Stranded DNA
STAR	Splicer Transcripts Alignment to Reference
stRNA	Small Temportal RNA

T	Thymine
TALE	Transcription Activator-Like Effector
TALEN	Transcription Activator-Like Effector Nuclease
TBS	Tris Buffered Saline
TBE	Tris-borate EDTA
TBST	Tris Buffered Saline Tween
TDG	Thymidine DNA glycosylase
TE	Tris-EDTA
TET	Ten-eleven dioxygenase
THC	Δ^9-tetrahydrocannabinol
TG	Tris Glycine
TGS	Tris Glycine SDS
tracRNA	Trans-encoded Small RNA
TWAS	Transcriptome Wide Association Study
U	Unmethylated
UTR	Untranslated Region
UV	Ultra-violet
WT	Wild Type
ZFN	Zinc Finger Nuclease
-ve	Negative Control
+ve	Positive Control

Chapter 1. General Introduction

1.1 Schizophrenia

1.1.1 Clinical presentation

Schizophrenia is a neuropsychiatric disorder with a prevalence of 1%, it affects 21 million people worldwide, and is within the top 10 causes of disability in developed countries (Whiteford *et al.*, 2013; World Health Organization, 2014). Schizophrenia onset usually occurs during early adulthood with earlier prevalence in males compared to females (Häfner *et al.*, 1994). Furthermore to an earlier risk, there is an increased relative risk of approximately 1.4 of being diagnosed with schizophrenia if you are male compared to female, resulting in a higher incidence rate in males (McGrath, 2005). The disorder is primarily characterised by two sets of symptoms the “positive” and “negative” symptoms. The “positive” symptoms include hallucinations, both auditory and visual, delusions and thought process interference, whereas the “negative” symptoms are effects on social tendencies, such as social isolation, lack of apathy and drive (Burmeister, Mcinnis and Zöllner, 2008). This diverse set of symptoms makes the disease difficult to define as a single entity, as two patients with schizophrenia may present with two completely different sets of symptoms and could respond to different treatments, highlighting the heterogeneity in schizophrenia presentation (Kendell, 1987). In order to address the heterogeneity in symptoms, the Diagnostic and Statistical Manual of Mental Disorders (DSM) version 5 provides an up-to-date diagnostic criteria that now encompasses all of the classic subtypes of schizophrenia such as; schizophreniform disorder, schizoaffective disorder and delusional disorder (Tandon *et al.*, 2013).

1.1.2 Neuropathology

As schizophrenia is a disorder of the brain there is mounting evidence for specific regions of the brain that are affected by the disorder (**Error! Reference source not found.**). The prefrontal cortex is located at the anterior part of the frontal lobe and is involved in attention, planning, communication with other brain regions, learning and memory (Tamminga and Buchsbaum, 2004). Neuroimaging studies have shown reduced white matter in the frontal lobes through diffusion tensor imaging (Ellison-Wright and Bullmore, 2009), however magnetic resonance imaging have had conflicting results, although this could be due to the analysis pipelines adopted due to the complexity of the region (Shenton *et al.*, 2001). Nevertheless the prefrontal cortex has been identified as one of the prime areas affected in schizophrenia patients due

to its role in the connectivity to other regions of the brain (Zhou *et al.*, 2015).The striatum is the hub of connectivity in the brain, and is an integral part of the corticobasal ganglia circuitry involved in facilitating voluntary movement (McCutcheon, Abi-Dargham and Howes, 2019). This region is home to the limbic pathway, where it has been shown that lesions of the limbic system in humans produced the symptoms seen in schizophrenia, therefore it is of significance in the aetiology of the disease (Meltzer and Stahl, 1976). Another region of the brain that is part of the limbic system is the hippocampus, this region is involved in long term potentiation of memories from short term to long term (Lynch, 2004). Hippocampal volume differences have been observed in schizophrenic patients however reproducibility is limited, potentially due to methodological differences (Haukvik *et al.*, 2018). Further, memory impairments have been observed in schizophrenic patients, for a comprehensive review see (Guo, Ragland and Carter, 2019), further enhancing the role of the hippocampus in schizophrenia aetiology. Finally the cerebellum is thought to have a role in cognition and emotional processing, as lower activity has been observed in schizophrenia patients when exposed to emotional response tests (Mothersill, Knee-Zaska and Donohoe, 2016). It is clear that there is not one part of the brain solely responsible for the symptoms seen in schizophrenia, therefore there must be an underlying environmental or genetic component which gives rise to the phenotype.

This image has been removed by the author of this thesis for copyright reasons.

1.1.3 Environmental influencers of schizophrenia

1.1.3.1 Pregnancy and birth complications

There have been numerous studies examining the concordance of pregnancy and schizophrenia. An individual patient meta-analysis of 12 studies containing 700 schizophrenia patients and 835 controls identified shorter gestational age, premature rupture of membranes, and use of resuscitation or incubators increased schizophrenia risk (Geddes *et al.*, 1999). This was followed up by a larger meta-analysis of population based studies, and found three groups of obstetric complications associated with schizophrenia: pregnancy complications, abnormal foetal growth and development, complications during delivery (Cannon, Jones and Murray, 2002). Furthermore, there is an increased risk of developing schizophrenia if born between winter and spring, although this is only observed in the northern hemisphere (Torrey *et al.*, 1997). It has been suggested that this is due to increased maternal infection during these time points, particularly influenza (Brown and Patterson, 2011). There are not only maternal exposures which lead to an increased risk, increased paternal age was thought to be strongly associated with schizophrenia from a large cohort study in Jerusalem (Malaspina *et al.*, 2001). The increased risk due to paternal age can be explained by increased chromosomal abnormalities and mutations which arise due to the aging germline (Torrey *et al.*, 2009). However, when paternal age was meta-analysed the large effect is reduced, so overall paternal age provides a relatively small risk for schizophrenia (Torrey *et al.*, 2009).

1.1.3.2 Geographical risk

There have been many epidemiological studies trying to identify if there is a geographical risk for schizophrenia. The first study of geographical risk in 1939 identified there were differences in the prevalence of schizophrenia across the city of Chicago (Faris and Dunham, 1939). This first suggested there could be a geographical link to schizophrenia, however more recent studies find that this is not consistent across urban centres and that geographical location can be confounded by many different factors (Kirkbride *et al.*, 2006). Migration has also shown to increase the risk of developing schizophrenia. A meta-analysis found a 2.7 relative risk of developing schizophrenia in first-generation migrants and 4.5 relative risk in second-generation migrants perhaps due to being in an ethnic minority (Cantor-Graae and Selten, 2005). This minority hypothesis has been further supported in a study in the UK which

identified that African-Caribbeans and Black Africans are at increased risk of developing schizophrenia (Fearon *et al.*, 2006). Current hypotheses believe the increased risk in migratory and ethnic minority populations is due to social stressors and other social risk factors rather than an underlying genetic component, however this is still disputed (Sharpley *et al.*, 2001).

1.1.3.3 Childhood adversity

Parental divorce or separation has been suggested to increase the development of psychosis. A large study carried out using 390 cases of psychiatric illness in the UK and matched controls identified a two to three fold increase in the risk of developing psychosis. However this study did not discriminate between different psychotic disorders therefore it cannot be linked to schizophrenia (Morgan *et al.*, 2007). In examining schizophrenia only, a study in 1999 identified a 3.8 fold increased risk of developing schizophrenia if one suffered from early parental loss, characterised by divorce or death (Agid *et al.*, 1999). However this study was relatively small comprising of 76 matched pairs. Of these pairs very few of the schizophrenia and control cases reported early parental loss (15 vs four). Furthermore there are confounding factors, one large confounder is that within the divorce cases there is no reported of level of contact following separation. These studies are also confounded further as there could be other confounders at play leading up to death and/or separation. Further research has focused on abuse during childhood. There have been numerous studies examining the effects on abuse and psychosis however these are, very often, small studies, and the findings have not been consistent (Morgan and Fisher, 2007).

1.1.3.4 Cannabis exposure

Epidemiological studies have shown that cannabis usage is associated with psychotic experiences and disorders including schizophrenia, however the causal link between cannabis and psychosis is difficult to ascertain from population based studies using observational approaches (Gage, Hickman and Zammit, 2016). The first longitudinal study into cannabis exposure and schizophrenia risk was carried out in Sweden in the 1980's. This study identified a six-fold relative risk of developing schizophrenia if you are a high user of cannabis (defined as patients with over 50 exposures to cannabis) (Andreasson *et al.*, 1987). Since then other studies have been published examining a causal link between cannabis usage and schizophrenia. A meta-analysis of cannabis use and psychosis from a multitude of different studies found the overall risk of

developing schizophrenia was two-fold higher with cannabis usage (Arseneault *et al.*, 2004). More recently, a larger meta-analysis over multiple centres examined the effect of Δ^9-tetrahydrocannabinol (THC) in cannabis on psychosis, this study showed a three-fold increase in psychosis in low potency cannabis (<10% THC content) but this rose to five-fold with high potency cannabis (>10% THC) (Di Forti *et al.*, 2019). This study is of importance as it has been shown that the potency of cannabis has been increasing over the last decade with an increase in THC content compared to cannabidiol (CBD) (Freeman *et al.*, 2019). While THC usage has thought to be an entirely environmental risk, it has been shown that there is a genetic link between cannabis usage and schizophrenia. In a recent Genome Wide Association Study (GWAS) on cannabis usage it was shown that single nucleotide polymorphism (SNP) signatures for cannabis usage overlap with the schizophrenia GWAS (1.1.4.3 The polygenicity of schizophrenia) (Pasman *et al.*, 2018). So it could be that schizophrenics are genetically predisposed to taking cannabis, whether this is causal or as a result of the disease is still unknown. In Chapter 7. I discuss the effect of THC on deoxyribonucleic acid (DNA) methylation, and how environmental risk factors for schizophrenia can be studied within the laboratory.

1.1.4 Genetics of schizophrenia

1.1.4.1 Twin studies

Familial studies have shown an approximate 85% heritability of schizophrenia indicating there is a credible genetic risk of one developing schizophrenia if one or both parents have the disease (Craddock, O'Donovan and Owen, 2005). The genetic component of schizophrenia has also been confirmed in twin studies, where the concordance rate is much higher in monozygotic twins compared to dizygotic twins (50% vs 17% respectively) (Cardno and Gottesman, 2000; Sullivan, Kendler and Neale, 2003). The heritability estimates are also consistent as study sizes increase and the data is meta-analysed, this provides strong evidence that genetics are accountable for part of the risk of developing schizophrenia but are not solely responsible. Considerable efforts have been made over the last two decades to identify schizophrenia causing genetic mutations, while a few rare highly penetrant genetic regions have been identified it is generally now accepted that most cases of schizophrenia are complex and polygenic, where there are multiple genes with low penetrance which are common in the population.

1.1.4.2 Rare and de novo mutations

There have been a few rare and highly penetrant genomic changes which that have been identified through genetic pedigrees as susceptibility risk loci for schizophrenia (Craddock, O'Donovan and Owen, 2005). One of the first characterised genes involved in schizophrenia is *DISC1* which was identified in a large Scottish family susceptible to psychiatric disorders (Millar *et al.*, 2000). Further to this, other pedigree studies have identified various other mutations in single genes which increase schizophrenia susceptibility including: *NRG1* (Stefansson *et al.*, 2002)*, DTNBP1* (Van Den Bogaert *et al.*, 2003), *DAOA* (Chumakov *et al.*, 2002)*,* and *RGS4* (Chowdari, 2003). However these only account for a very small proportion of risk for schizophrenia. Therefore if schizophrenia is not a monogenic disorder a polygenic hypothesis was then proposed where multiple genes are affected. This is supported by copy number variations (CNV), where genes or parts of whole chromosomes are duplicated or deleted (McCarroll and Altshuler, 2007). CNVs have been associated with many diseases including schizophrenia. One of the most commonly associated CNVs in schizophrenia is 22q11.2 deletion, where deletion of 22q11.2 results in a 20 fold increased risk of developing schizophrenia (Bassett *et al.*, 2017). Moreover duplication of 22q11.2 has been shown to become protective of schizophrenia (Rees *et al.*, 2014). While 22q11.2 is one of the most well characterised CNVs associated with schizophrenia a recent analysis of over 21,000 schizophrenia cases and 20,000 controls identified a number of significant CNV associations including; 1q21.1, 2p16.3, 3q29, 7q11.2, 15q13.3, 16p11.2 and 22q11.2 (Marshall *et al.*, 2017). However the relative load of CNVs on global schizophrenia cases is low, for example 22q11.2 accounts for approximately only 2% of cases of schizophrenia despite it being the most significantly associated CNV (Bassett, Scherer and Brzustowicz, 2010). There are also a number of new *de novo* mutations in schizophrenia which are being reported, this could explain how the rates of schizophrenia have remained stable despite reduced fecundity (Rees *et al.*, 2012; Fromer *et al.*, 2014) and links back to environmental risk factors of paternal age (see section 1.1.3).

1.1.4.3 The polygenicity of schizophrenia

While there are a handful of highly penetrant genetic alterations implicated in schizophrenia these are rare occurrences and do not explain the relatively high prevalence of the disease. This has led to the realisation that schizophrenia is a

polygenic disease, where there are multiple genes with low penetrance which are common in the population. In the last decade improvements in sequencing technology, open collaborations, and large patient/control cohorts has led to a genetic revolution. GWAS have identified genomic regions which harbour multiple genes and SNPs associated with complex diseases (including schizophrenia) (Visscher *et al.*, 2017). These datasets require large numbers to uncover the common but low penetrant variants associated with diseases. To combat this, in respect to schizophrenia and other psychiatric disorders, the Psychiatric Genetic Consortium (PGC) was established to gather enough data to have the power to detect common variants associated with schizophrenia (Sullivan, 2010). One of the largest published schizophrenia GWAS by the PGC used 36,989 patients with schizophrenia and 113,075 controls identified 108 genomic *loci* associated with developing schizophrenia (Schizophrenia Working Group of the Psychiatric Genomics Consortium, 2014). A more recent GWAS by (Pardiñas *et al.*, 2018) which used 11,260 schizophrenia cases and 24,542 controls was then meta-analysed with the PCG study and increased the 108 risk *loci* to 145 (Pardiñas *et al.*, 2018) (**Figure 1.2**). The top 20 most significant risk loci for schizophrenia are shown in **Table 1.1**. From the table you can see that multiple genes are present at single risk loci making functional analysis complex as we do not know which gene (or genes) is affected by the causal variant. This is made even more complex as many SNPs are within linkage disequilibrium (LD) of each other, therefore identifying the causal locus in these risk *loci* is difficult (Slatkin, 2008). However advances in genome editing technologies (discussed later in Chapter 3.) has made it easier to characterise the risk genes independent of the causal risk variant. One schizophrenia risk gene in which I characterised in this thesis is *AS3MT*, this is discussed in more detail in Chapter 3. and Chapter 4. So while we have made breakthroughs in identifying the risk genes and SNPs of schizophrenia, we are no closer to knowing how these risk variants have a causal effect, however delving deeper into the risk genes identified they support some of the common hypothesis of schizophrenia (Visscher *et al.*, 2017).

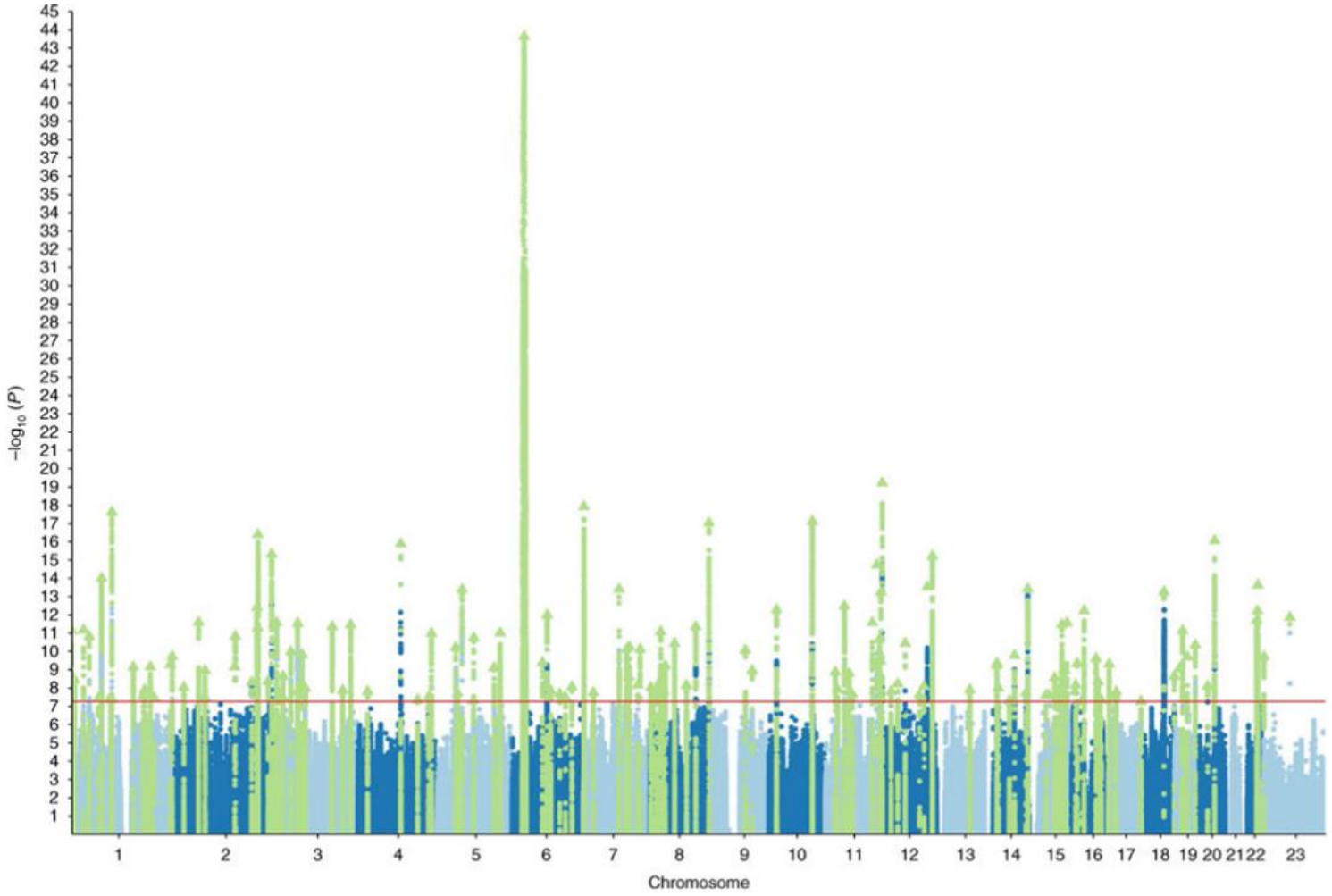


Figure 1.1: A Manhattan plot showing the 145 identified genetic risk loci for schizophrenia from (Pardinas et al 2018). *Chromosomes are along the X axis with adjusted p-values on the Y axis. The red line shows the genome-wide significant level ($5x10^{-8}$). Each dot represents a single SNP, green dots are the 145 identified loci which reach genome-wide significance.*

***Table 1.1 The top 20 schizophrenia risk loci identified by the (Pardinas et al 2018) meta-analysis**. Shown are the schizophrenia risk loci ranked by genome-wide significant p-value. The gene AS3MT has been extensively characterised in this thesis using state of the art molecular tools (Chapters 3 and 4).*

Chromosome	Start Location	Length (Kb)	Number of SNPs	p-value	Genes
6	24,988,105	8854.772	12997	2.12E-44	*xMHC*
12	2,321,868	201.904	80	5.63E-20	*CACNA1C, CACNA1C-AS4, CACNA1C-IT3*
7	1,877,502	312.598	216	1.10E-18	*MAD1L1, MIR4655*
1	98,341,152	217.941	189	2.18E-18	*DPYD, MIR137, MIR137HG, MIR2682*
10	104,570,118	489.778	386	7.09E-18	*AS3MT, C10orf32, C10orf32-ASMT, CNNM2, CYP17A1, INA, NT5C2, PCGF6, RPEL1, WBP1L*
8	143,276,678	81.638	33	8.65E-18	*LINC00051, MIR4472-1, TSNARE1*
2	200,547,937	761.610	194	3.69E-17	*C2orf47, C2orf69, FTCDNL1, TYW5, SPATS2L*
20	37,361,504	124.482	53	7.91E-17	*ACTR5, PPP1R16B, SLC32A1*
4	103,001,649	196.433	7	1.19E-16	*BANK1, SLC39A8*
2	233,559,312	231.337	64	4.33E-16	*C2orf82, EFHD1, GIGYF2, KCNJ13, NGEF*
12	123,447,928	454.433	361	5.55E-16	*ABCB9, ARL6IP4, C12orf65, CDK2AP1, LOC100507091, MIR4304, MIR8072, MPHOSPH9, OGFOD2, PITPNM2, RILPL2, SBNO1, SETD8*
11	124,610,011	10.136	11	1.70E-15	*ESAM, MSANTD2, NRGN, VSIG2*
1	73,736,562	255.089	204	8.94E-15	*LOC101927295*
22	42,315,790	373.580	119	2.15E-14	*CENPM, LINC00634, MIR33A, SEPT3, SHISA8, SREBF2, TNFRSF13C, CYP2D6, CYP2D7P, FAM109B, LOC388906, LOC101929829, NAGA, NDUFA6, NDUFA6-AS1, SEPT3, SMDT1, TCF20, WBP2NL*
12	110,723,245	0.000	2	2.68E-14	*ATP2A2*
14	104,188,920	348.760	188	3.44E-14	*LINC00637, PPP1R13B, XRCC3, ZFYVE21, ASPG, C14orf2, RD3L, TDRD9*

7	86,403,263	544.810	214	3.57E-14	*DMTF1, KIAA1324L, TMEM243, TP53TG1, GRM3*
5	60,563,907	279.799	41	3.72E-14	*ZSWIM6*
18	52,747,689	602.184	164	5.03E-14	*TCF4, LOC101927273, LOC100505474*
11	133,822,133	30.875	36	5.32E-14	*IGSF9B*

1.1.4.4 Neurodevelopmental hypothesis

The neurodevelopmental hypothesis has been established as one of the core hypotheses of schizophrenia. This hypothesis states that schizophrenia results from deficits which arise during neurodevelopment (Owen *et al.*, 2011). There is mounting evidence which supports this hypothesis, from neuroimaging studies, genetic association, and epidemiological stances (for a comprehensive review see (Fatemi and Folsom, 2009). Neuroimaging studies using a range of different techniques, regions, and study groups have brought up consistent findings of reduced connectivity in chronic schizophrenia patients, however these findings are less consistent in first episode schizophrenia patients, possibly due to more subtle changes early on in the development of the disease (for a review see (Wheeler and Voineskos, 2014)). One observation noted is that patients who have a direct relation with schizophrenia show similar structural changes to their relative compared to non-relatives, this links into the recent GWAS findings (Wheeler and Voineskos, 2014). Several of the most robust signals from GWAS have been found to overlap with genes involved in neurodevelopment and are in genes which are expressed during brain development (Duarte *et al.*, 2016; Ma *et al.*, 2018). A recent transcriptome wide association study (TWAS) also identified how schizophrenia genetics can alter transcripts involved in neuronal development (Gusev *et al.*, 2018). One of the associations identified was in *MAPK3,* which they chose to characterise in zebrafish. *MAPK3* interacts with *KCTD13* and *MVP* in a co-regulation network, with increased expression of *KCTD13* resulting in microcephaly and reduced neuronal proliferation. The TWAS result suggested a transcriptional interation between *MAPK3* and *KCTD13* and that suppression of *MAPK3* would resuce the microencephal and neuronal loss. When *mapk3* was knocked down in *KCTD13* overexpressed zebrafish using morpholinos, neuronal proliferation and head size were rescued back to wild type levels (Gusev *et al.*, 2018). This links schizophrenia genetic risk to transcripts involved in neurodevelopment. This is further supported by several prenatal environmental factors that are thought to contribute to schizophrenia including maternal stressors (Khashan *et al.*, 2008), maternal infection (Brown and Patterson, 2011), maternal famine (Wang and Zhang, 2017). Adolescent cannabis usage has also been associated with developing schizophrenia as the brain is undergoing extensive developmental changes (Arseneault *et al.*, 2004; Radhakrishnan, Wilkinson and D'Souza, 2014). For example, Catechol-O-methyl transferase (*COMT*) is a risk gene for developing psychosis

following THC exposure, it is also within the 22q11 locus associated with schizophrenia (Williams, Owen and O'Donovan, 2007; Costas *et al.*, 2011). These all provide compelling evidence for the neurodevelopmental hypothesis of schizophrenia.

1.1.4.5 Inflammation and immunity

The PGC and Pardinas GWAS studies on schizophrenia identified genetic variants within the major histocompatibility complex (MHC) *locus* on chromosome 6 as the most significant and strongly associated region with schizophrenia (Schizophrenia Working Group of the Psychiatric Genomics Consortium, 2014). The MHC region contains 18 highly polymorphic human leukocyte antigen genes which are involved in antigen presentation during the immune response to pathogens. As such the MHC region is very gene rich and incredibly polymorphic, making it very difficult to fine map and identify risk genes and variants (Benacerraf, 1981). The role of the immune system in schizophrenia has previously been implicated before the GWAS study of 2014 (for a comprehensive review see (Strous and Shoenfeld, 2006). For instance, increased maternal interleukin-8 during pregnancy has been associated with an increased risk of developing schizophrenia in their offspring (Brown *et al.*, 2004). For a comprehensive review of the maternal infection and schizophrenia see (Brown and Derkits, 2010). There is evidence of the classical complement cascade, an innate immune response to pathogens and cellular debris being involved in schizophrenia. The complement component 4 (C4) genes (*C4A* and *C4B*) are located within the MHC complex are of interest in schizophrenia as the C4 protein localises to neuronal synapses, dendrites and axons (Sekar *et al.*, 2016). Increased expression of C4A has been identified in post mortem brain tissue from schizophrenic patients versus controls, this can be explained due to increased CNV in *C4A* which correlates to the increase in C4 expression (Sekar *et* al., 2016). In mice is was identified that C4 is involved in synaptic pruning which could explain the synaptic loss observed in schizophrenia patients (Sekar *et al.*, 2016). Because of the highly variable nature of the MHC complex there are multiple haplotypes of C4 genes with varying frequency within the MHC locus which makes disentangling the causal haplotype difficult.

1.2 Gene expression and regulation

Despite recent advances in genetics and risk loci identification we are still are no closer to understanding the role of genetics in disease pathology. This is further hindered as many of the risk variants associated with schizophrenia are in non-coding and

regulatory regions. It is clear then that we first need to understand the role of transcriptional regulation and how these SNPs can exert an effect on gene expression. In the human body the DNA sequence is generally identical within every cell, although recent single cell sequencing experiments have identified somatic mutations occurring at very low frequency in individual cells (Lodato *et al.*, 2015). DNA holds the instructions required for synthesising everything which forms a functional organism (The ENCODE Project Consortium, 2007). As of the latest build on Ensembl, humans possess approximately 20,000 protein-coding genes, 23,000 non coding genes, and 15,000 pseudogenes resulting in a transcriptome pool of approximately 230,000 difference transcripts (Cunningham *et al.*, 2019).

The central dogma of molecular biology, first proposed by Francis Crick in the 1960's states that during transcription a coding portion of the DNA is used as a template by a ribonucleic acid (RNA) polymerase to produce a messenger RNA (mRNA) molecule. This mRNA molecule then undergoes splicing, which removes any transcribed non-coding repetitive sequences known as introns, leaving behind the coding exons, to produce a mature mRNA. This mature mRNA then leaves the nucleus where it itself is then used as a template to synthesise a protein by ribosomes, in a process known as translation (Crick, 1970) (**Error! Reference source not found.**). The fine control of the central dogma and gene transcription is regulated by the presence of transcription factors, co-factors, and epigenetic markers. These epigenetic markers include DNA modifications, alterations to chromatin states, and non-coding RNA (ncRNA) (Allis and Jenuwein, 2016).

This image has been removed by the author of this thesis for copyright reasons.

1.2.1 Epigenetics

Epigenetics was introduced in 1942 by Conrad Waddington to bridge the gap of study between genes and their products (Waddington, 2012). Waddington then went on to propose in 1957 “an epigenetic landscape” representing how gene regulation modulates cell fates throughout development (Waddington, 1957). The figure (shown in **Figure 1.4**) shows a stem cell represented at the top of the landscape, where it progresses through development it can traverse down different courses to ultimately end at one cell fate, this is mediated by epigenetic mechanisms. The modern epigenetic term refers to the study of mitotically heritable (but reversible) changes in gene expression which occur independently from the DNA sequence (Henikoff and Matzke, 1997; Jaenisch and Bird, 2003). Therefore epigenetics is thought of as factors “above the genome”, acting to regulate gene expression.

Epigenetic processes are critical for development, differentiation, and regulation of gene function. Epigenetics have key roles in several aspects of development, including genomic imprinting (Morison, Ramsay and Spencer, 2005) and X-chromosome inactivation (Gartler and Riggs, 1983). Epigenetic mechanisms are highly dynamic throughout development (Zhang *et al.*, 2019) and show tissue specific patterns, this explains how the same DNA gives rise to multiple different cell types (Rakyan *et al.*, 2008). Epigenetic markers are inherited through mitotic divisions, however like somatic mutations in DNA, there can be environmental influences or stochastic changes which can arise across one’s life which can alter gene expression and result in disease (Egger *et al.*, 2004). Alterations in epigenetic markers have been shown to associate with a number of diseases, including schizophrenia (Shorter and Miller, 2015; Hannon, Dempster, *et al.*, 2016; Viana *et al.*, 2017), depression and suicide (Murphy *et al.*, 2017), Alzheimer’s (Lunnon and Mill, 2013; Jager *et al.*, 2015), and cancer (Sharma, Kelly and Jones, 2010). Epigenetic research has the potential to help develop treatments for diseases as many epigenetic processes can be altered by environmental influencers such as therapeutics or drugs, therefore it makes it paramount that we understand epigenetic processes and their interplay in disease. As mentioned previously, there are several mechanisms of epigenetic processes, including DNA methylation, histone modifications, and ncRNAs, their roles in epigenetic control are explained in the following sections.

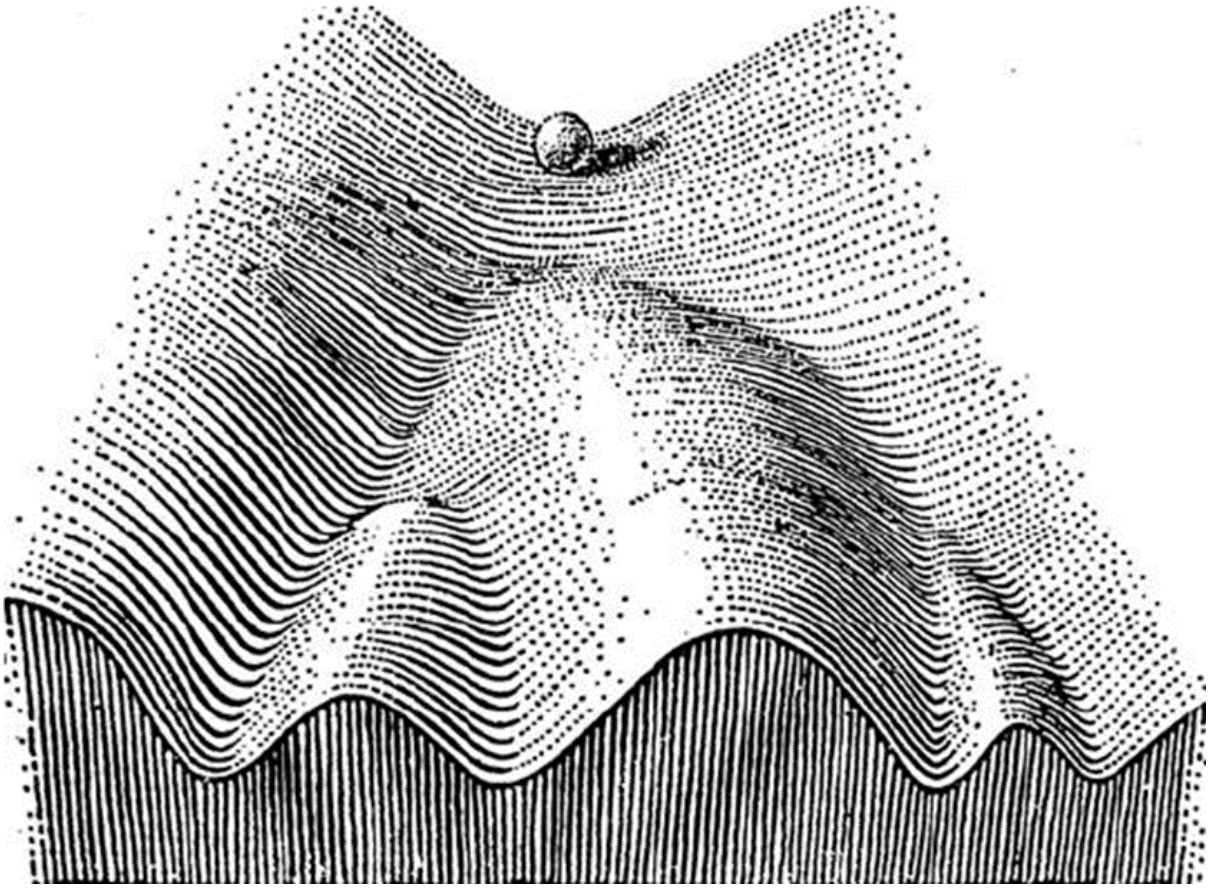


***Figure 1.2: Waddington's epigenetic landscape**. Figure taken from (Waddington 1957). The figure represents a stem cell as a ball at the top of the landscape. Through development there are many different trajectories it can take, represented as valleys. The pathway to its final cell state is determined by genetic regulation through epigenetic means.*

1.2.1.1 DNA methylation

DNA methylation is one of the most well characterised and stable epigenetic mechanisms and plays a key role in transcriptional regulation (Jirtle and Skinner, 2007). DNA methylation in eukaryotes involves the transfer of a methyl group to the 5th position of the cytosine ring at a CpG dinucleotide to form 5-methylcytosine (5mC). The reaction uses S-adenosylmethionine (SAM) as a methyl donor and is catalysed by a group of enzymes known as DNA methyltransferases (DNMT) (Klose and Bird, 2006). There are several known DNMTs, DNMT1 is the maintenance methyltransferase which maintains methylation patterns during cell division, whereas DNMT3a and DNMT3b are *de novo* methyltransferases which methylate previously unmethylated cytosines (Lyko, 2018). Methylation primarily occurs within CpG islands (CGIs), these dense regions of CpG sites are usually located within 5' promoters of genes. Promoters are near transcription start sites and are involved in transcriptional regulation of gene expression. Methylation therefore disrupts gene expression by blocking transcription factor binding and initiates chromatin contraction (Bird, 1986; Klose and Bird, 2006).

Although it is accepted that methylation can be actively and passively removed from 5mC, the mechanisms have only recently been discovered (Tahiliani *et al.*, 2009). It has been shown that methyl groups can be removed in a sequence of oxidation reactions catalysed by the ten-eleven translocation (TET) dioxygenases to form three intermediate modifications; 5-hydroxymethylcytosine (5hmC), 5-formylcytosine (5fC) and 5-carboxylcytosine (5caC) (He *et al.*, 2011; Ito *et al.*, 2011). 5caC can then be converted back into cytosine through a combination approach of base excision repair and thymine DNA glycosylase activity (He *et al.*, 2011; Weber *et al.*, 2016). The cycle of methylation and de-methylation is shown in **Figure 1.5** which is adapted from a comprehensive review by (Wu and Zhang, 2017).

The relationship between 5mC and transcription is not as linear as first thought, exciting new developments have identified that 5mC within gene bodies acts to regulate alternative splicing (Flores *et al.*, 2012; Maunakea *et al.*, 2013) and methylation at non CpG sites has also been identified to activate gene expression (Lister *et al.*, 2009; Aran *et al.*, 2011). Further research has now identified that other cytosine modifications are considered to be epigenetic markers of their own rather than just by-products of 5mC demethylation. 5hmC for example, has been shown to

be altered in Alzheimer's disease (Smith *et al.*, 2019), is highly dynamic through human development (Spiers *et al.*, 2017), decreases through embryonic stem cell differentiation (Pfaffeneder *et al.*, 2011), is enriched in purkinje neurones, a GABAnergic neuron located in the cerebellum (Kriaucionis and Heintz, 2009), and is enriched at exon-intron boundaries of constitutive exons (Khare *et al.*, 2012). Indicating a potential role for other cytosine modifications in disease.

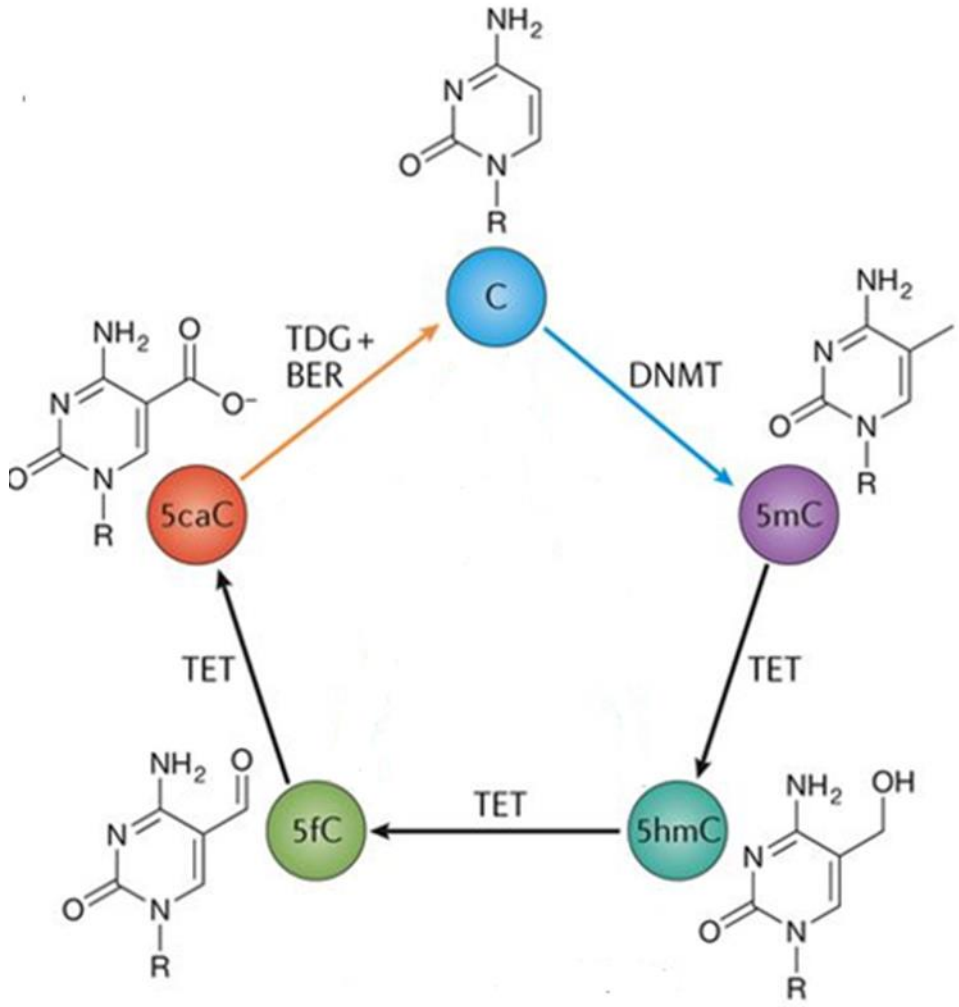


Figure 1.3: Cytosine modifications at stages of methylation and de-methylation. *Cytosine is methylated to 5mC by DNA methyltransferase (DNMT), 5mC can then be oxidised to 5hmC, 5fC and 5caC by ten-eleven translocation (TET) dioxgenase. 5fC can be returned to cytosine through thymidine DNA glycosylase (TDG) and base excision repair (BER). Adapted from (Wu and Zhang 2017).*

1.2.1.2 Histone modifications

While cytosine modifications are the most well characterised modification they do not regulate expression alone, there is a complex interplay with how DNA is packaged in 3D space. Eukaryotic DNA is packaged into chromatin by structural subunits called nucleosomes, the DNA wraps around nucleosomes to form a spool like structure which enables the 10,000-20,000 fold level of compaction to package the DNA into a nucleus (Zentner and Henikoff, 2013) (**Figure 1.6**). These nucleosomes in turn are made up of eight different histone proteins arranged in an octamer core. The octamer is made up of four different subunits of histone protein, known as H2A, H2B, H3, and H4. It is the N-terminal tails of these histone proteins, which extend out from the nucleosome, which are subject to post-translational modifications (**Figure 1.5**) (Spencer and Davie, 1999).

Common histone modifications include acetylation, adenosine diphosphate-ribosylation, citrullination, methylation, phosphorylation, sumonylation, and ubiquitination (Sadakierska-Chudy and Filip, 2015). These histone modifications result in changes to the conformation of chromatin, either changing it from a relaxed form of euchromatin to a compact heterochromatin preventing gene expression. The histone modification can also result in the opening of the heterochromatin to euchromatin which allows access to transcription start sites and promoter regions within the DNA for RNA polymerase to bind and initiate the transcription of genes (Bannister and Kouzarides, 2011).

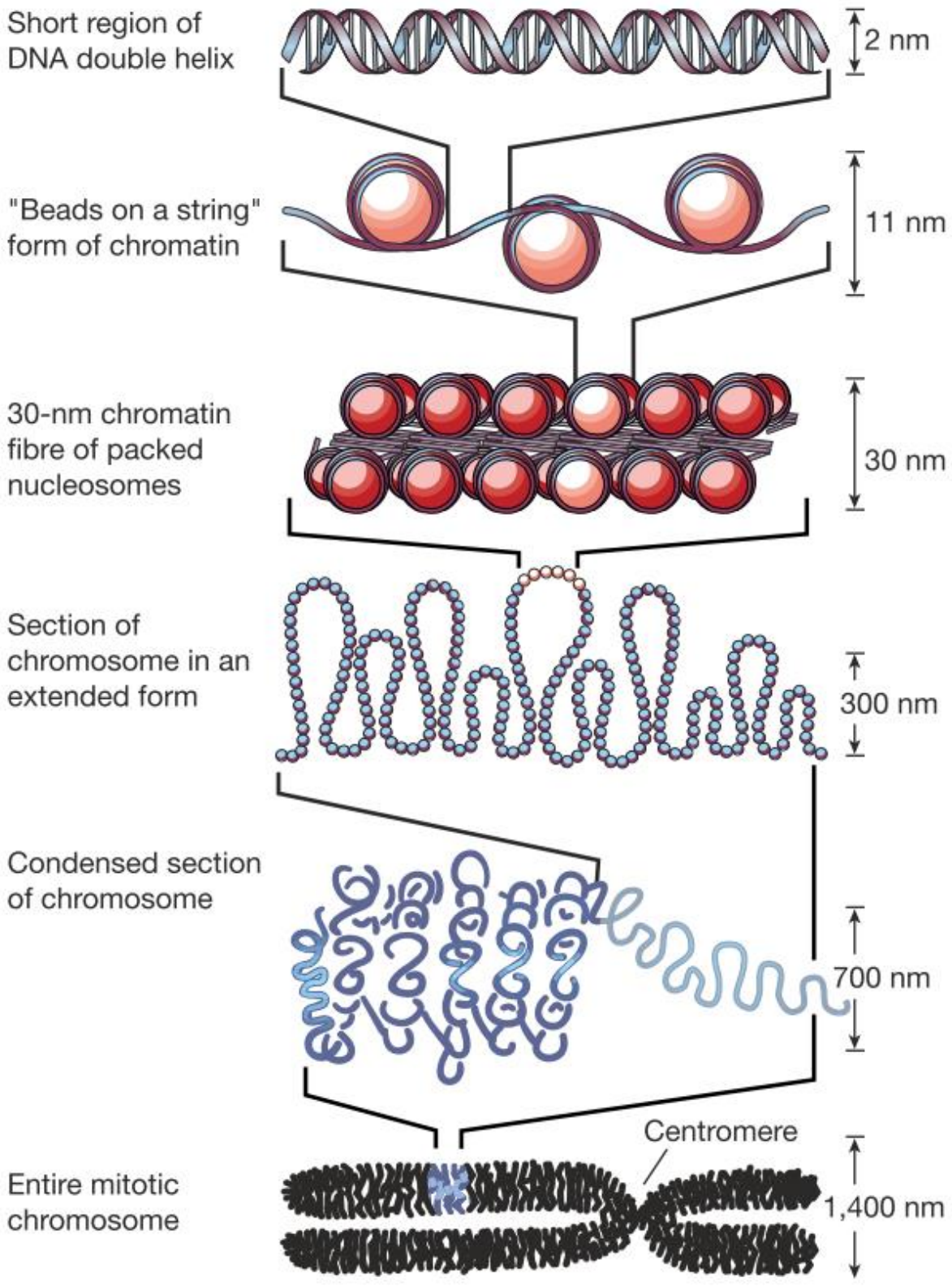


Figure 1.4: Diagram showing how chromatin helps condense the DNA sequence. *The nucleosomes are made of eight histone subunits, these nucleosomes are then tightly packed to form chromatin. This chromatin then forms higher structures to eventually form a chromosome. Image adapted from Fesenfeld and Groudine (2003).*

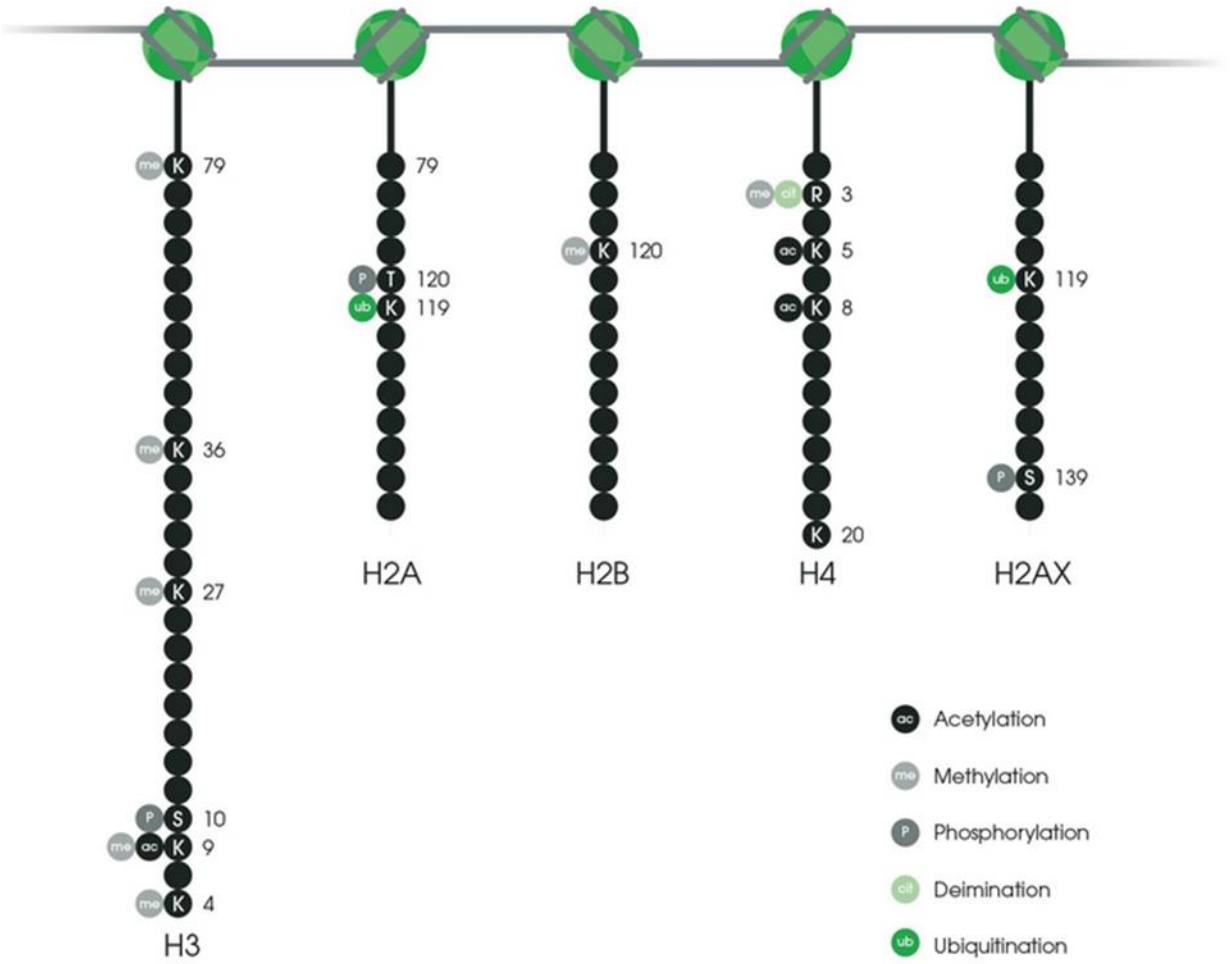
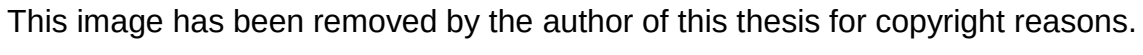


Figure 1.5: Common histone tail modifications which have been well characterised for their role in gene expression and regulation*. Image from Abcam (https://www.abcam.com/epigenetics/histone-modifications).*

1.2.1.3 Non-coding RNA

Non-coding RNA (ncRNA) are another method of epigenetic control. There are several forms of ncRNA; small nuclear RNAs (snRNA), small nucleolar RNAs (snoRNA), piwi-interacting RNAs (piRNA), small interfering RNAs (siRNA), small temporal RNAs (stRNA), and micro RNAs (miRNA) (Mattick and Makunin, 2006)(**Error! Reference source not found.**). Each of these ncRNA have a distinct role in regulation of gene expression. snRNA are involved in the creation of small nuclear riboproteins which are crucial to alternative splicing of differential mRNA isoforms (Lerner *et al.*, 1980). snoRNA were first discovered in nucleus and are involved in ribosomal RNA modifications and processing although they might be closer related to miRNA function (Matera, Terns and Terns, 2007; Scott and Ono, 2011). siRNAs are involved in RNA interference by binding to an RNA induced silencing complex (RISC), this then recognises complementary mRNA strands and cleaves the mRNA before translation to prevent gene expression (Dana *et al.*, 2017). piRNA are again involved in splice mediation like siRNA and form a RISC known as piRISC, these specifically target and silence transposon elements (Iwasaki, Siomi and Siomi, 2015). stRNA were identified in *C. elegans* development, these repress mRNA expression by binding to complementary 3' untranslated region (UTR) of their target genes mRNA preventing translation (Banerjee and Slack, 2002). Finally miRNA are small sequences of nucleotides which regulate gene expression post-transcription and inhibit translation (Bartel, 2004). Therefore it is important to recognise the roles that the many ncRNA play in mediating gene expression.


This image has been removed by the author of this thesis for copyright reasons.

1.2.1.4 How genetic variation can influence transcriptional variation

Traditionally it has been thought that genetic variation within non-coding regions alters gene expression independent of epigenetic mechanisms, namely through modification of transcription-factor binding motifs. SNPs within these regions could either remove or create a motif leading to a decrease or increase in gene expression respectively (The ENCODE Project Consortium., 2007). However now it has been shown that genetic variation also influences transcriptional regulation through epigenetic modifications: including chromatin (Kasowski *et al.*, 2013), histones (Mcvicker *et al.*, 2013) and DNA methylation (Bell *et al.*, 2011). Furthermore it has been shown that genetic variation on DNA methylation works in a tissue dependent manner (Gutierrez-Arcelus *et al.*, 2015). The way in which genetics can influence epigenetics is twofold; first alterations in the genomic sequence through a SNP may influence epigenetic marks, for example a SNP located in a guanine or cytosine at a CpG site may introduce or remove a CpG thus altering methylation, secondly SNPs may alter the epigenetic machinery indirectly. SNPs which have an effect on a measurable phenotype, such as height or weight, are known as quantitative trait *loci* (Members of the Complex Trait Consortium, 2003). Over the years these have been further refined as the field has progressed, SNPs shown to alter DNA methylation are known as methylation quantitative trait *loci* (mQTLs), and SNPs shown to alter gene expression are termed expression quantitative trait *loci* (eQTLs) (**Figure 1.9**). mQTLs and eQTLs have been shown to overlap and affect gene expression of genes in cis of their location. It is therefore possible that the genome is altering the methylome to bring about gene expression changes which, as a result, cause complex phenotypes (Wagner *et al.*, 2014). There have been many publications which have studied these eQTLs (Schizophrenia Working Group of the Psychiatric Genomics Consortium, 2014; Dobbyn *et al.*, 2018) and mQTLs (Hannon, Dempster, *et al.*, 2016; Hannon, Spiers, *et al.*, 2016) and their interactions (Hannon *et al.*, 2018). This provides support to how these common variants associated with schizophrenia may function through altering gene regulation.

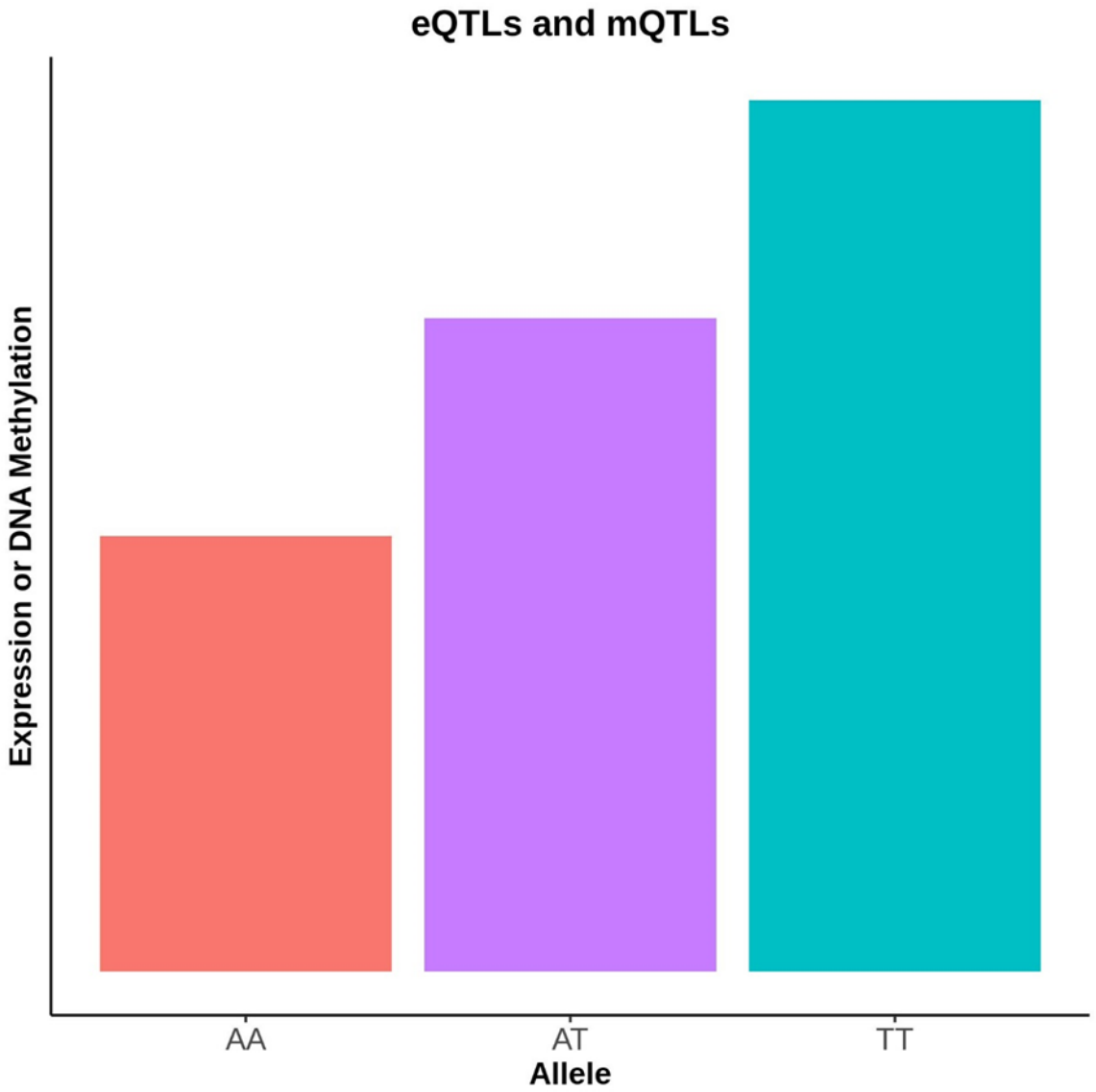


Figure 1.6: eQTLs and mQTLs share an underlying principle, the alleles inherited have an effect on either gene expression (eQTL) or DNA methylation (mQTL). *In this hypothetical example, being homozygous for T results in increased gene expression or DNA methylation compared to if you are homozygous for A.*

1.3 DNA methylation studies in schizophrenia

Some of the earliest DNA methylation studies in schizophrenia were targeted approaches to hypothesised schizophrenia risk genes such as *RELN* (Abdolmaleky *et al.*, 2005; Grayson *et al.*, 2005), *SOX10* (Iwamoto *et al.*, 2005), and *COMT* (Abdolmaleky *et al.*, 2009). The problem with these studies are the small sample sizes, due to post mortem brain samples being difficult to obtain as currently these studies cannot be completed *in vivo.* Further to this, targeted approaches are limited to hypothesised risk genes which limits the discovery of novel genes and loci, and you have to be confident that the risk genes targeted are directly involved in schizophrenia. However as technology advanced towards microarrays, genome wide changes in DNA methylation could then be observed. The first genome-wide methylation study of schizophrenia was carried out by (Mill *et al.*, 2008) and discovered genome-wide methylation changes in frontal cortex of 35 schizophrenia samples compared to controls. Since this landmark, further genome-wide studies have further validated genome-wide methylation changes in schizophrenia patients in different regions of the brain using increasing sample sizes.

Two early studies using the predecessor to the HumanMethylationBeadChipEPIC the Illumina 450k array (Illumina, San Diego, CA, USA) started to unravel the complex nature of DNA methylation in respect to schizophrenia. The use of these array based approaches allows for genome-wide methylation profiles to be carried out with relative ease, resulting in high throughput data. One of the first studies using the 450k approach used frontal cortex from 24 schizophrenic patients and controls, and identified genome-wide methylation changes in genes previously associated with schizophrenia (Wockner *et al.*, 2014). In the same year (Pidsley *et al.*, 2014) published a study using prefrontal cortex and cerebellum from schizophrenic patients and controls and identified hypo-methylation of *NRN1* along with an enrichment of changes in neurodevelopmental genes and pathways. However these studies were still carried out on relatively small samples, and many changes in DNA methylation could have been lost through confounding factors such as cell type, post-mortem interval, storage time, pH, and cellular heterogeneity (Pidsley and Mill, 2011). One of the first larger cohort studies used prefrontal cortex from 191 schizophrenic patients and 335 controls and again discovered methylation enrichment in genes involved with neuronal development and differentiation (Jaffe *et al.*, 2016). This was then further validated by

(Viana *et al.*, 2017) using the most comprehensive cohort of schizophrenia samples including multiple brain regions (prefrontal cortex, striatum, hippocampus, and cerebellum). DNA methylation studies using enriched bisulphite conversion sequencing and array based methods have been shown to be highly reproducible in identifying robust alterations in DNA methylation profiles (Chan *et al.*, 2019).

Further analysis has integrated GWAS data from the schizophrenia GWAS to identify mQTLs and how SNPs can alter methylation profiles through development. Several studies have been published which show the top differentially methylated regions in schizophrenia are enriched in regions which show methylation changes throughout development (Pidsley *et al.*, 2014; Spiers *et al.*, 2015). As previously mentioned, a recent study in our lab identified an enrichment of schizophrenia GWAS variants in foetal brain mQTLS (Hannon, Spiers, *et al.*, 2016). In this study 166 foetal brains from the first and second trimester had both DNA methylation and SNP profiled to generate a database of mQTLs, where a SNP is associated with DNA methylation at a given CpG. From this 16,809 mQTL associations were discovered in foetal brain with 96% of the associated SNP and DNA methylation site occurring on the same chromosome. Further to this, these sites are co-localised to areas of gene regulation, specifically with repressive histone markers H3K9me3 and H3K27me3 and CTCF transcription binding motifs. When the overlap between these mQTL sites and the PGC schizophrenia GWAS was examined (Schizophrenia Working Group of the Psychiatric Genomics Consortium, 2014) there was a strong enrichment of these GWAS hits within the foetal mQTL dataset (relative enrichment = 4.11, p-value=3.0×10^{-6}). This study has identified how these schizophrenia risk loci could have an effect on DNA methylation, thus gene regulation during foetal development which would support the neurodevelopmental hypothesis of schizophrenia. However it is crucial to note, that this approach cannot determine when there are two causal SNPs in complete LD with one another, one responsible for DNA methylation and one responsible for schizophrenia. Disentangling how SNPs effect gene expression through DNA methylation was the focus of Chapter 6. and is discussed in more detail later on.

1.4 Transcriptome studies in schizophrenia

Early transcription studies were targeted to risk genes of schizophrenia. The dopamine D4 receptor was reportedly increased in schizophrenia patients' post-mortem however

mRNA expression in post-mortem brains failed to support this increase (Roberts *et al*., 1996). Serotonin receptors were the next target, mRNA expression and autoradiography analysis on post-mortem brain identified imbalances within 5-HT_{1A} and 5-HT_{2A} receptor ratio which could contribute to impairment of corticocortical association pathways (Burnet *et al*., 1996). Finally GABAergic genes were examined, with particular focus on *RELN* and *GAD67,* where expression was found to be decreased in schizophrenia patients (Impagnatiello *et al.*, 1998; Guidotti *et al.*, 2000). Following on, the advent of microarrays increased the throughput of transcription studies to examine the entire transcriptome. Examining the entire transcriptome allows for a more complex landscape to be uncovered, including gene pathways and networks which have been disrupted in schizophrenia. Studies using microarrays in schizophrenia implicated pathways involved with immune and inflammatory responses, neurodevelopment, and neurotransmission (for a comprehensive review see (Kumarasinghe, Tooney and Schall, 2012)).

The advent of short read RNA sequencing (RNA-seq) in the last decade has accelerated the field in transcriptional studies into schizophrenia and other diseases (Ozsolak and Milos, 2011). One of the largest brain tissue transcriptional studies into schizophrenia used dorsolateral cortex of 258 schizophrenic patients and 279 control subjects, 693 differentially expressed genes were identified that were involved with neuronal development and synaptic processing (Fromer, Roussos, Solveig K. Sieberts, *et al.*, 2016). Further analysis identified that up to 20% of GWAS associated schizophrenia variants could result in these identified transcriptional changes (Fromer, Roussos, Solveig K. Sieberts, *et al.*, 2016). This integration of GWAS and expression datasets can provide extra dimensions to studies and infer how GWAS SNPs alter gene expression. A large TWAS in schizophrenia was undertaken to identify how genetic variants affect transcriptional profiles by looking at SNP – expression correlation in adipose, blood, and brain tissue of non-psychiatric controls (Gusev *et al.*, 2018). These were then cross-referenced to SNP – schizophrenia correlation from the previous schizophrenia GWAS (Schizophrenia Working Group of the Psychiatric Genomics Consortium, 2014). This cross-referencing identified 157 schizophrenia-related SNP/transcriptome associations in the non-psychiatric controls with 42 of these related to chromatin features, identifying the role of gene regulation in schizophrenia (Gusev *et al.*, 2018). Another computational integration approach further supports

schizophrenia GWAS loci affecting transcriptional changes across multiple different brain regions (Huckins *et al.*, 2019). However these still need to be modelled in biological systems to validate the findings.

1.5 Current models of schizophrenia

Most of the work discussed thus far has been carried out using post mortem brain tissue. Post mortem brain tissue has many advantages, including quick and efficient access to relevant brain regions with comprehensive medical records allowing for a multitude of confounders to be corrected for based on environmental influencers. There are a large number of caveats of using this tissue, primarily which these tissues are heterogeneous, and contain many different cell types of which the proportions can change dependent on the sample examined. This makes identifying the causal cell population difficult, as complex deconvolution of resultant data needs to be undertaken. Furthermore, these tissues only collect a snapshot of the disease at a given time which is usually at the later stages when it has highly progressed, so results uncovered could be resultant of the disease instead of a cause. Finally post mortem tissues could have been handled in different ways dependent on the brain bank they have been acquired from, the lack of standardisation could result in different post mortem intervals, mistaken diagnosis, or incorrect brain region identification, all impacting on the reliability of the results. Despite these caveats post-mortem brain tissue has been widely used to study schizophrenia as there is no current *in vitro* or *vivo* model of the disease. However there are a range of models organisms which are increasingly being used to study schizophrenia, including rodent and neuronal cell models. The following section will go into more detail on the advantages and caveats of each of these models, in relation to modelling schizophrenia.

1.5.1 Rodent models

To fully characterise a disease extensive modelling must take place to uncover the mechanisms of disease aetiology. One of the most common models for disease are rodent models. These models are beneficial as you can remove many confounders which can influence a study such as, environmental factors, genetic differences, and reverse causation. With rodent models you also have ease of access to different tissues and specific regions of the brain, and can examine the brain at different points of development. As rodents are a whole system organism there is also the ability to establish system wide networks of changes, compared to singular pathways in a cell

model. There are also extensive tests which can be can be used which measure cognition in the rodents which can be translated to human schizophrenic phenotypes (Young *et al.*, 2009).

Rodent models of schizophrenia can be split into distinct categories, dependent on the researcher's needs and question being addressed. The most common type of rodents models for schizophrenia are either; neurodevelopmental, pharmacological, or genetic. Neurodevelopmental rodent models are based around the theory of maternal stressors during development contributing to the disease. There are two well-known models; the methylazoxymethanol acetate (MAM) model, where administration of MAM at gestation day 17 results in rats with a "schizophrenic like" phenotype (for a review see (Lodge and Grace, 2009)), and post weaning social isolation which causes behavioural, neurochemical and morphological changes which resemble schizophrenia (Fone and Porkess, 2008). The limitations of these models are that they are not definitive to one disorder, but could correspond to other psychiatric disorders such as anxiety or depression.

The pharmacological models are the most frequently used models to study schizophrenia, partially due to their ease of usage. Common pharmacological interventions include dopamine enhancers, such as amphetamines (Sams-Dodd, 1998), and non-competitive N-methyl-D-aspartate antagonists such as ketamine (Jeevakumar *et al.*, 2015) and phencyclidine (Janhunen *et al.*, 2015). The biggest caveat of pharmacological interventions is the targeting of single pathways or neurotransmitters as there could be a multitude of different systems working in symphony to develop the schizophrenic phenotype which we have not yet characterised.

Finally genetic models have been established using traditional genetic engineering techniques to remove genes or groups of genes associated with schizophrenia. These range from single gene knockout from associated schizophrenia risk genes, such as *DISC1* (Clapcote *et al.*, 2007), *NRG1* (Mei and Xiong, 2008), Dysbindin (Cox *et al.*, 2009), and Reelin (Fatemi, 2001), to chromosomal deletions such as 22q11.2 (Bassett *et al.*, 2005; Bassett and Chow, 2008; Ellegood *et al.*, 2014). Although the limiting factors of genetic manipulation is that, as previously discussed, there is no one gene which causes schizophrenia, instead it is a combination of genes having small

changes with additive effect. A summary of the current approaches and their advantages and disadvantages are shown in **Table 1.2** which comes from a comprehensive review on animal models in psychiatric research (Nestler and Hyman, 2010).

One of the defining features of schizophrenia are hallucinations and delusions, these are very human characteristics and are not easily translated into a mouse model (Canetta and Kellendonk, 2018). The heterogeneity within schizophrenia makes it difficult to establish, as the DSM-V criteria for diagnosing schizophrenia makes it difficult to have a mouse model which fills all of these criteria, or even just some. The divergence of rodents from humans over 60 million years ago has led to differences in brain development and structure between humans and rodents (Clowry, Molnár and Rakic, 2010). Therefore rodent models are not optimal for studying a complex heterogeneous psychiatric disorder such as schizophrenia.

Table 1.2: Current approaches to animal models in schizophrenia and their strengths and weaknesses. *Adapted from (Nestler and Hyman 2010)*

General Approach	Specific Method	Strengths	Weaknesses
Genetics	Selective breeding	Focus on phenotypes of interest	May produce a phenocopy of human disorder
	Random mutation and screening	Focus on phenotypes of interest	May produce a phenocopy of human disorder
	Transgenic animals (for example, knockouts, knockins, overexpression)	Recapitulates genomic abnormality in human disorder; focus on gene of interest	Variable penetrance of genetic abnormality in rodents Human relevance of phenotype may be difficult to establish
	Virally mediate gene delivery to brain	Spatial and temporal control over genetic change; focus on gene of interest	Does not recapitulate genetic cause of human disorder
Pharmacological	Administration of neurotransmitter agonist or antagonist	Temporal and some spatial (with intracranial delivery) control; focus on neurotransmitter system of interest	Lack of evidence that common mental disorders involve selective lesions of a single neurotransmitter system
Environmental	Chronic social stress (adult or during development)	May recapitulate risk factors in humans	Lack of specificity for given human disorder
	Chronic physical stress	Easy to administer	Lack of construct validity for most human disorders

1.5.2 Cell models

While animal models as complex organisms are useful to examine system wide changes they have their limitations and some findings are untranslatable into humans (for review see (Denayer, Stöhrn and Van Roy, 2014)). Because of this, researchers regularly use human cell lines to model disease, these can be either immortalised cell lines or induced pluripotent stem cell models (iPSC) taken from schizophrenia patients or carriers of risk mutations.

The common cell lines used for research have been derived from cancerous tumours as these will readily expand and continue to proliferate in culture. The current most common used human neural cell line is the SH-SY5Y cell line derived from a metastatic neuroblastoma (Biedler *et al.*, 1978). (More detailed description of the SH-SY5Y cell line in **2.1.1.1.**). These cell lines have the advantage of being highly reproducible between experimental repeats and laboratories as they are all genetic clones, however genetic drift can result in divergent genotypes with increased time in culture so cell lines with a higher passage number are generally avoided. Cancer derived cell lines are incredibly easy to culture and require far less complex media than some neuronal cell lines, therefore ideal for most experiments as this reduces batch effects through creation of complex media. There are some key limitations of cancer derived cell lines, as they are cancerous they can show complex karyotypes which alters gene expression. Furthermore you are limited to the cell types you can derive from certain cancers and these derived cell types will only be a pseudo-type of the cell they are trying to represent.

One new avenue for examining the complex genetics and epigenetics of schizophrenia is through using stem cells. Embryonic stem cells (ESCs) derived from the blastocyst of an embryo, have the unique ability to differentiate into any cell of the body (Kaufman and Evans, 1981). These provide a better model of complex disease as you can have ESC with different polygenic risk score (PRS) to complex diseases, including regions which are in strong LD of each other. This can then allow fine mapping of the regions of LD to identify the causal variant using newer technologies (discussed later in this thesis) by engineer changes in individual SNPs to elucidate its mechanism of action in different cell types. However, acquiring ESCs is difficult due to ethical concerns therefore these are not routinely used, recent techniques have been developed to to convert adult human fibroblasts from patients with complex diseases back into stem

cells, these are known as iPSCs (Takahashi *et al.*, 2007) (**Error! Reference source not found.**). These iPSCs have been shown to have similar morphology, proliferation, surface antigens, gene expression and epigenetic profiles of ESCs and can differentiate into any cell type (Takahashi *et al.*, 2007). Cross-referencing RNAseq data from early to late embryogenesis and iPSC derived neurones has identified shared transcriptional profiles which mimic neuronal development (Nicholas *et al.*, 2013; Stein *et al.*, 2014). Therefore, the ability to generate iPSC from human fibroblasts makes them a powerful tool for studying complex disease, including schizophrenia. iPSC lines have been derived from schizophrenic patients with mutations in *DISC1* (Wen *et al.*, 2014), 22q11.2 deletions (Lin *et al.*, 2016), or with increased polygenic risk (Brennand *et al.*, 2011). Neurones from patients with schizophrenia derived from iPSCs have shown diminished neural connectivity, neurite number, and glutamate receptor expression (Brennand *et al.*, 2011). Further analysis identified that these neurones also responded to loxapine, an antipsychotic currently used in schizophrenia treatment (Brennand *et al.*, 2011).

Nevertheless iPSC models of schizophrenia have numerous disadvantages for research. Firstly the reprogramming and differentiation of fibroblasts to neurones results in heterogeneous populations of different cell types, which can vary between experimental repeats due to genetic effects between different iPSC donors (Rouhani *et al.*, 2014). In order to combat this, many different cell lines should be established from multiple donors, however current studies are only using tissue from minimal donors. The cost of generating iPSC models is exponentially more expensive than non iPSC cell culture, along with this most differentiation protocols take many weeks of continuous culturing to achieve the desired population which cannot be possible in some laboratories. Furthermore, genetic drift can also result with consistent culturing as previously discussed with cancer derived cell lines. Currently we do not know the specific cell types primarily responsible for schizophrenia, this then limits which cells we should be examining using iPSC, it could be that a multitude of different cell types are interacting to produce the complex schizophrenia phenotype.

This image has been removed by the author of this thesis for copyright reasons.

1.6 Functional validation using genetic engineering

Recent advances in genetic engineering have made it relatively straightforward to edit the genomes of animals and cells to functionally correlate genetic risk to disease. One of the largest issues in complex disease genetics is understanding the pathways which are implicated in disease. From the schizophrenia GWAS there are over 145 loci and over 300 individual genes related to the disease (Pardiñas *et al.*, 2018). Older technologies such as zinc fingers transcription activator-like effector nucleases (ZNF) resulted in single knockout of risk genes in models and then characterising the observed phenotype and correlating this to disease (Kim, Cha and Chandrasegaran, 1996). These techniques were long, laborious, singleplex, and required highly skilled teams of technicians to achieve genetic knockout and are discussed in more detail below. However with the advent of new technologies in the last decade we have now moved into a new foray of functionally validating not only GWAS studies but also epigenome-wide association studies (EWAS) using cutting edge molecular biology techniques. In this section I will discuss the advances made in genetic engineering and how these can be used to create new models of complex disease. A summary of the techniques is shown in **Error! Reference source not found.** and **Table 1.3**.

1.6.1 Methods of editing the genome and epigenome

1.6.1.1 Zinc finger nucleases (ZFN)

One of the first mechanisms of genome engineering were the ZFN, these have been used to alter the DNA of *Drosophila* (Bibikova *et al.*, 2002)*, Xenopus* (Bibikova *et al.*, 2001)*, Mus* (Meyer *et al.*, 2010), and human cells (Porteus and Baltimore, 2003). ZFN work by recognising a specific trinucleotide sequence, and when linked together, multiple ZFN can recognise larger sequences of up to nine base pair (bp) improving targeting specificity. The ZFN is fused to the endonuclease *FokI* which, when the ZFN is bound to DNA, can cleave the DNA and result in a double strand break (DSB) (Kim, Cha and Chandrasegaran, 1996) (**Error! Reference source not found.**). Following creation of a DSB the DNA attempts to repair itself by one of two mechanisms: non-homologous end joining (NHEJ), which results in small insertions or deletions (indel), or homology directed repair (HDR), which results in identical replacement from the homologous chromosome (**Error! Reference source not found.**). (Symington and Gautier, 2011; Chapman, Taylor and Boulton, 2012). NHEJ results in indels which shift the open reading frame of the protein and results in the knockout of a functional

protein. Whereas HDR utilizes a template strand of DNA, usually the intact homologous chromosome, to repair the gap with an exact copy, however this pathway can be utilised for knock-in of reporter genes (Moehle *et al.*, 2007). One limiting factor of ZFN is that *FokI* works as a dimer, therefore two ZFN need to be designed for the sense and antisense strand of the targeting region for cleavage to occur (Bibikova *et al.*, 2001). Because of the need for two constructs there is considerable more expense and time to result in a functional knockout, therefore these are no longer the preferred method for genetic engineering.

This image has been removed by the author of this thesis for copyright reasons.

1.6.1.2 Transcription activator-like effector nucleases

Transcription activator-like effector nucleases (TALENs) were the next generation of genetic engineering tools and were easier to construct compared to ZFNs. TALENs are very similar to ZFN in that they use a *FokI* endonuclease to induce DSB at a sequence specific site. Rather than using zinc fingers to target the *FokI* TALENs use a DNA binding domain composed of highly conserved sequences from transcription activator-like effectors (TALE) (**Error! Reference source not found.**). These TALEs are proteins secreted by *Xanthomonas* bacteria to alter gene transcription by binding to DNA using highly conserved 30-35 amino acid repeats (Boch and Bonas, 2010). The individual TALE repeats in each array is specific to a single base of DNA by utilizing two hypervariable residues at amino acid 12 and 13 (Boch *et al.*, 2009). Therefore, by arranging these TALE repeats in a specific order you can direct the fused *FokI* to your target region of DNA resulting in DSB and either NHEJ or HDR. These are advantageous over ZFN as they can recognise a single base pair making it easier to target specific regions of the genome. Again like ZFN, TALENs have been utilised for the editing of human cells (Miller *et al.*, 2011), *Drosophila* (Liu *et al.*, 2012), zebrafish (Sander *et al.*, 2011), and mice (Haute *et al.*, 2013). Much like ZFN, the use of *FokI* endonuclease means two constructs are required for dimerization and cleavage of target DNA, increasing cloning times and cost. Another disadvantage of TALENs is that they are sensitive to 5mC which results in reduced efficiency (Valton *et al.*, 2012).

1.6.1.3 Clustered regularly interspaced short palindromic repeats

1.6.1.3.1 Genetic

Clustered regularly interspaced short palindromic repeats (CRISPR) loci were first identified in prokaryotes in 1987, these are short, partially palindromic DNA repeats which occur at regular intervals to form loci which alternate variable sequences and repeated elements (Ishino *et al.*, 1987). These loci are usually surrounded by CRISPR-associated (*cas)* genes, of which the corresponding proteins show DNA binding ability and exonuclease activity (Jansen *et al.*, 2002). Further research identified that the spacers located within CRISPR regions have homology to phage genes and other extrachromosomal elements (Bolotin *et al.*, 2005). This led to a hypothesis that the CRISPR-Cas system has evolved as a form of bacterial immunity to foreign DNA such as phage or viruses (Barrangou and Marraffini, 2014). The CRISPR-Cas mediated

defence can be divided into several stages: step one is the insertion of new spacers from foreign elements into the CRISPR locus, step two is transcription of the *cas* genes and the long precursor CRISPR RNA (pre-crRNA), which is then processed into mature crRNA by accessory factors. The third and final step is interference where the invading nucleic acid is targeted and destroyed by a crRNA and Cas protein complex (Rath *et al.*, 2015) (**Error! Reference source not found.**).

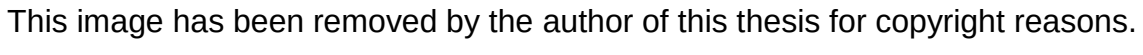
Further characterisation has identified multiple *cas* genes of different properties and mechanisms of action which has resulted in the formation of different classes of CRISPR-Cas systems (Makarova *et al.*, 2011). Where the most common used is the type II system which includes Cas9, identified in *S. thermophilus* as an endonuclease which cleaves DNA (Garneau *et al.*, 2010). Further research identified that Cas9 requires t*rans*-encoded small RNA (tracrRNA) as well as pre-crRNA to cause DSB at specific DNA sites (Deltcheva *et al.*, 2011). Once this was discovered, an engineered fusion of the pre-crRNA and tracrRNA, henceforth known as a guide RNA (gRNA), was found to direct Cas9 to a specific sequence of 20 DNA nucleotides to cause a DSB (Jinek *et al.*, 2012) (**Error! Reference source not found.**). One final requirement for Cas9 activity is a protospacer adjacent motif (PAM) directly next to the target site to allow cleavage of DNA, in the *S. pyogenes* Cas9 system the PAM is an NGG sequence (Shah *et al.*, 2013).

Following the characterisation of the type II CRISPR-Cas9 system in bacteria, researchers began to adapt the system for use in mammalian cells. One of the first instances of use was in HEK cells (Hsu *et al.*, 2013) but the system has also been used for editing human iPSC (Srikanth *et al.*, 2015), mice (Wang *et al.*, 2013), and zebrafish (Hwang *et al.*, 2013). The advantage of the CRISPR-Cas9 system for genetic engineering is the ease of targeting as only a single gRNA is required for Cas9 activity. The RNA-DNA interaction in CRISPR-Cas9 versus Protein-DNA interaction in ZFN and TALENs allows for easier characterisation of on target and off target effects. Finally the system can be easily modified for multiplexing or screening, resulting in ease of knockout of multiple genes and screening for phenotypes of interest (Sakuma *et al.*, 2014; Datlinger *et al.*, 2017). Further to this the CRISPR-Cas9 system can be modified by creating a catalytically dead version of Cas9 and fusing it to other enzymes, including epigenetic modifiers (Doudna and Charpentier, 2014; Tadi *et al.*, 2016; Vojta *et al.*, 2016).

Table 1.3: A summary of current genetic engineering tools.

Feature	Zinc Finger	TALENs	CRISPR/Cas9
Identified From	Eukaryotic gene expression regulators	*Xanthomonas* plant pathogens	Adaptive immune system of bacteria and archaea
Target Sequence	9-18bp	30-40bp	20bp + PAM
Mechanism of recognition	DNA-Protein	DNA-Protein	DNA-RNA
Targeting Predictability	Low	Low	High
Nuclease	FokI	FokI	Cas9
Design	Two complementary motifs around target to be disrupted	Two complementary motifs around target to be disrupted	Singular gRNA complementary to target sequence
Cost	High	Affordable	Cheap
Time	High	High	Quick
Multiplexing	No	No	Yes
Off Targets	Unknown to determine due to unknown protein-DNA interactions that can occur	Unknown to determine due to unknown protein-DNA interactions that can occur	Easier to predict based on Watson and Crick base pairing rules
Size	Smallest	Intermediate	Largest
Ease of Delivery	Easiest	Intermediate	Hardest
Key Limitation	Difficulty in engineering the ZFN construct	Requirement of a T before the 5’ end of the targeting sequence	Location of PAM sequences in the human genome required for Cas9 activity

This image has been removed by the author of this thesis for copyright reasons.


This image has been removed by the author of this thesis for copyright reasons.

1.6.1.3.2 Epigenetic CRISPR tools

Until recently the CRISPR molecular toolbox available was limited to creating breaks in the targeted DNA, however recent developments utilising a catalytically dead Cas9 (dCas9) has expanded the CRISPR toolkit. This dCas9 no longer has the catalytic ability to cleave DNA but still targets and binds to DNA when complexed with gRNA (Doudna and Charpentier, 2014). Because of this, researchers have begun to fuse a variety of gene regulatory enzymes to the dCas9 to target these enzymes to specific parts of the genome, including enhancers, CCCTC-binding factor (CTCF) loops, and promoters. There are a variety of different epigenetic tools dependent on the research questions being asked, however the most common are CRISPR activators (CRISPRa), CRISPR interference (CRISPRi) which modify chromatin to allow activation or repression respectively, or DNA methylation editors which edit DNA methylation at specific CpG sites.

The first CRISPRa systems involved the fusion of dCas9 to either VP64 or p65 activation domain. These systems resulted in an increase of green fluorescent protein (GFP) expression by 25 fold and 12 fold respectively in a HEK293 reporter cell line by targeting an upstream activator sequence of GFP (Gilbert *et al.*, 2013). Newer techniques have now been developed to further increase the activation ability of these vectors including a hybrid fusion of dCas9 fused to VP64-p65-Rta. This hybrid showed 10 fold increase in *NGN2* mRNA expression and 18 fold increase in *NEUROD1* mRNA expression compared to VP64 in neuronal differentiated iPSC (Chavez *et al.*, 2015). Further work has looked at CRISPRa fused to the p300 core domain of human acetyltransferase which acetylates histone H3 lysine 27 resulting in transcriptional activation. These have shown promise by increasing transcriptional activation of *IL1RN*, *MYOD* and *OCT4* in HEK293T cells (Hilton *et al.*, 2015). Along with CRISPRa there are also CRISPRi which are not as plentiful as the CRISPRa vectors. The most common CRISPRi system available is dCas9 fused to Krüppel-associated box (KRAB) repressor which induces trimethylation at histone H3 lysine 9 resulting in decreased chromatin accessibility and gene repression in targeted genes. In a GFP reporter HEK293 cell line this system resulted in 15 fold repression of GFP (Gilbert *et al.*, 2013). This system has also been used for the reduction of gene expression of *HBE1* and *HBG1/2* in human K562 cells when the HS2 enhancer is targeted, helping us to

identify genes in *cis* and *trans* which are under the control of enhancers (Thakore *et al.*, 2015).

Along with the CRISPRa and CRISPRi systems to edit chromatin states there are also systems to edit DNA methylation at specific CpGs. The DNA methylation inducers are dCas9 fused to DNMT3a which introduce a methyl group to unmethylated CpG sites (Liu *et al.*, 2016; Vojta *et al.*, 2016). Targeting the promoter region of a GFP reporter mouse ESCs cell line using a lentivirus containing these editors showed an approximate 70% increase in methylation which resulted in 12% of cells no longer expressing GFP (Liu *et al.*, 2016). Further uses identified that targeted methylation of CTCF loops involved in chromatin organisation increased expression of targeted genes in mouse ESCs (Liu *et al.*, 2016). Another system using transient transfection rather than lentiviral has shown to engineer fold change reductions in *IL6ST* and *BACH2* of 0.4 and 0.55 in HEK293 cells, by increasing methylation up to 55% over a promoter region (Vojta *et al.*, 2016). As well as DNA methylation inducers there are also DNA methylation erasers which consist of dCas9 fused to Tet1CD which induces the conversion of 5mC to 5hmC to allow conversion back to unmethylated CpG sites (Liu *et al.*, 2016, 2018). This system was first used to turn on GFP expression in a GFP reporter mouse ESCs with a fully methylated promoter. By targeting the promoter a decrease of methylation from 80% to approximately 50% was observed over the promoter region which resulted in 26% of cells expressing GFP (Liu *et al.*, 2016). The system also has been shown to increase gene expression of *BDNF* six fold in mouse cortical neurones and induced reprogramming of fibroblasts to myoblasts by demethylating the *MyoD* distal enhancer (Liu *et al.*, 2016). The system can also be used *in vivo* on transgenic mice to examine the effect of specific methylation changes on a whole system. To do this, a transgenic mouse containing a GFP reporter with a hypermethylated promoter was targeted using direct lentiviral injection of dCas9 fused to Tet1CD. Following injection into the mouse brain and subsequent demethylation of the promoter, GFP fluorescence was reported, identifying how this system can then be utilised to examine specific methylation changes *in vivo* (Liu *et al.*, 2016). Furthermore, this system has also been used to rescue fragile X by restoring *FMR1* expression in iPSC derived from fragile X patients by targeting the aberrantly methylated CGG repeat in the 5’ UTR (Liu *et al.*, 2018). It is clear to see that there are a plethora of different tools to activate and repress gene expression using a variety of

different fusions and modifications to the CRISPR/Cas9 system. These tools will further help us to understand the role of genetic regulation and is causal link to disease. These tools were the focus of Chapter 6. to characterise and validate DNA methylation changes derived from computational EWAS studies in our lab.

1.6.2 Functional validation of genome wide association studies

As discussed previously there are limitations in interpreting GWAS study results, namely that it is very difficult to identify a causal variant within a genetic associated loci for a diseases as many SNPs are within LD of each other. In order to combat this there are numerous statistical approaches to elucidate the causal SNP within GWAS. QTL analysis has managed to refine the loci to SNPs of interest through examining overlapping datasets, either known expression (for eQTL), methylation datasets (mQTL), and proteins (pQTL). These studies are of interest because they can be used to refine and identify how the SNP can change biological function within the cell and open avenues for further validation. However these statistical approaches are limited mainly by how QTLs work in a tissue dependent manner; we are currently limited to datasets in specific cell types that may not be relevant to the disease being studied. Moving on from the statistical approaches, with the advent of CRISPR technologies we are now able to move into functionally validating GWAS signatures and findings in the laboratory with relative ease (**Figure 1.14**).

Genetic knockout approaches can be taken to validate GWAS findings using CRISPR/Cas9. These “dirty” methods can quickly examine and refine where the causal SNP is located within area of LD. This can be achieved by designing a tile of gRNA to knockout several individual SNPs at once over the region. Following generation of the cell lines, transcriptional, expression, or methylation analysis can then be undertaken to identify if any of the knocked out SNPs are eQTLs or mQTLs. This would be beneficial as these can be carried out in a cell line of interest to the disease being studied and will provide more relevance than publicly available datasets. This does have a caveat of knocking out multiple SNPs at once, this can help fine map where the causal variant is and what affect it is having on the cell but it will never identify the precise SNP, therefore it would be better to perform individual SNP analysis. Newer techniques have being generated to aid in this functional validation.

Until recently it was thought that you could only generate knockouts using CRISPR/Cas9, however as previously discussed the technique has been used to edit DNA methylation. Other systems have been developed to mediating changes of C to T using cytidine deaminase (Komor *et al.*, 2016) and A to G using adenine deaminases (Gaudelli *et al.*, 2017). These utilise the dCas9 and do not result in DSB at the target sites, making these techniques relatively easy to use to deliver the correct base edit. This therefore can allow for all four transition mutations (C to A, A to G, T to C and G to A) and can be a powerful tool for the validation of GWAS studies. Nonetheless, there is an inherent problem of off targets produced using these systems for editing single bases. Currently these systems have a window of editing where any cytosine within the target site could be edited, however these windows are decreasing in size with newer iterations of the systems (Rees and Liu, 2018). These editing windows makes the system more problematic when it comes to repetitive sequences in non-coding regions of the genome, where a majority of SNPs occur, and could result in pseudo associations if the wrong base is edited. One newer system which has been developed can prevent unwanted editing of the wrong base by utilising the HDR arm of the DSB repair system to insert corrected bases using a template with relative ease. This system uses the piggyBac transposon system which allows for easier selection and screening of edited cell lines. This system was first used with CRISPR/Cas9 to correct mutations in iPSC associated with β-thalassemia and examine how these mutations are functionally linked to disease (Xie *et al.*, 2014). This technique does have scope for use in validation of GWAS studies. Interestingly one recent study used CRISPR/Cas9 base editing to validate eQTL associations in HEK293T cells (Brandt *et al.*, 2019). To do this the authors designed gRNA to cut over an eQTL site and co-transfected a ssDNA template containing the alternative allele to repair the cut by HDR. However this technique has a relatively low success rate due to HDR repair pathways only being active in currently dividing cells. This low efficiency is shown as the authors wished to edit 33 eQTL sites but only 13 passed QC prior to data analysis. Although only 13 sites passed QC, ten of these eQTL associations examined have the same direction of effect on the predicted gene as predicted by the Genotype-Tissue Expression (GTEx) database (Brandt *et al.*, 2019). This shows how we are beginning to move forward into validating genomic changes in context of gene expression and hopefully how this causes disease.

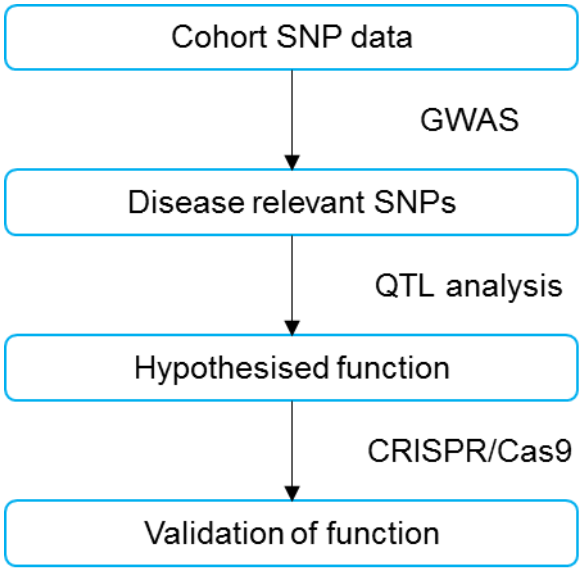


Figure 1.7: A hypothetical pipeline for Genome Wide Association Studies (GWAS) and functional validation. *Many GWAS studies published tend to identify disease relevant SNPs but do not follow up with functional validation. Some GWAS studies will go further and cross reference SNPs to other datasets and perform QTL analysis to try and elucidate if any have a functional role. Very few studies utilise CRISPR/Cas9 techniques to validate these QTL findings, this is potentially because CRISPR/Cas9 is relatively new to the field of genetics and could be further exacerbated by the lack of expertise within many statistical labs.*

1.7 General aims

Considering the literature discussed above there is a clear link between genetic and epigenetic variation and schizophrenia, however the causal links are difficult to disentangle due to current methodologies. Computational models provide us with lists of statistically relevant genes and CpG sites however there is need for further validation for how these sites are causal in the disease. The aim of this project was to develop functional methodologies to functionally validate computational studies into schizophrenia and elucidate how identified genes (Chapter 3. and Chapter 4.), differentially methylated regions (Chapter 5.) or CpGs (Chapter 6.), and environmental risks (Chapter 7.) can contribute to schizophrenia. An overview of each chapter is provided in the summary in **Figure 1.15**.

To do this I used a range of cutting edge molecular techniques which have been discussed previously to achieve the following objectives:

1. Chapter 3. examines how CRIPSR/Cas9 can be used to knock out and characterise the schizophrenia risk gene *AS3MT* in SH-SY5Y neuroblastomas cells. I developed a protocol for editing SH-SY5Y cells using CRISPR/Cas9 which is previously unreported in this cell line. I also identified morphological changes in differentiated SH-SY5Y cells following knockout of *AS3MT* which have been previously unknown.
2. Chapter 4. identifies the transcriptional response following knockout of the schizophrenia associated risk gene *AS3MT.* In this analysis, I identified a significant change in gene expression in genes involved in ribosomal function and neurodevelopment. I also identified an enrichment of previously associated schizophrenia risk genes as being differentially expressed in *AS3MT* knockout lines.
3. Chapter 5. examines the uses and limitations of luciferase based reporters to validate differentially methylated regions from EWAS studies. Utilising a DMR identified in schizophrenia and depression I identify that *in vitro* methylation of this sequence results in the quenching of luciferase expression. Validating the DMR identified through computational science is influenced by DNA methylation state.
4. Chapter 6. utilises a modified form of CRISPR/Cas9 designed to edit specific DNA methylation sites and how these can be used to functionally link DNA

methylation changes and the genes which they control expression of. I developed a modified form of CRISPR/Cas9 to show that DNA methylation can be edited specifically and does not necessarily lead to gene expression changes in predicted genes.

5. Finally Chapter 7. examines how a known schizophrenia risk inducing environmental factor (THC exposure) can impact on the DNA methylome using a cell model. This is the first study of its kind examining genome-wide DNA methylation, while we do not observe any significant findings the top differentially methylated probes are within genes of relevance to neurodevelopment and neuronal function.

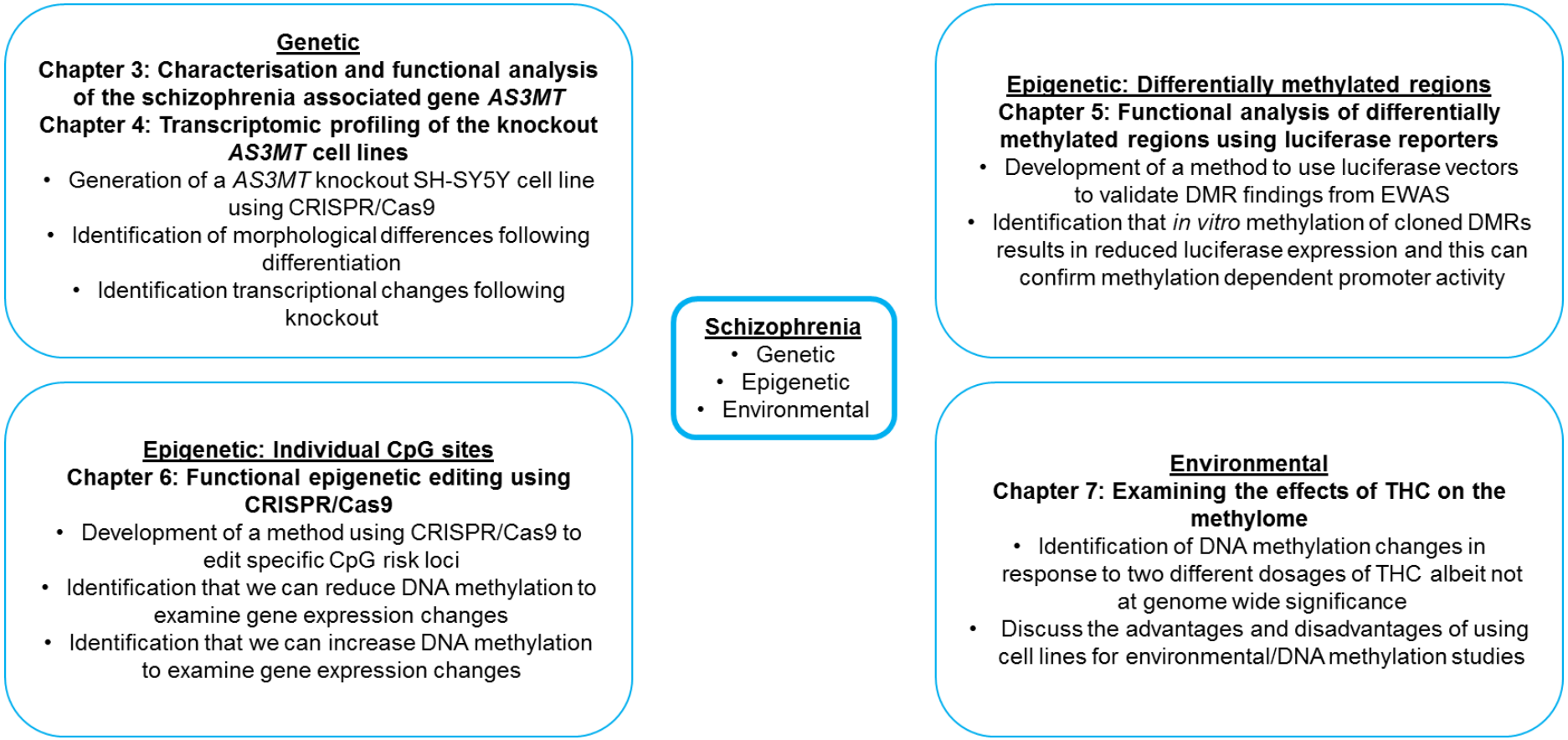


Figure 1.8: Levels of functional investigation studied in this thesis, including genetic, epigenetic, and environmental risks to schizophrenia

Chapter 2. General Methods

This chapter lists the general materials and methods used throughout this thesis in multiple chapters. Detailed descriptions of methods and analysis specific to each chapter are given in each empirical chapter.

2.1. Tissue culture

2.1.1. Cell lines

The three major cell lines used through this thesis are the human neuroblastoma cell line SH-SY5Y and the human embryonic kidney cell lines HEK293/HEK293T. All cell lines were acquired from the American Type Culture Collection (https://www.lgcstandards-atcc.org/?geo_country=gb).

2.1.1.1. SH-SY5Y

The SH-SY5Y neuroblastoma cell line is the most cited *in vitro* model in neuropsychiatric research, partly because of its ease of culture, cheap cost, reproducibility, and literature available. The SH-SY5Y cell line is a subline of the SK-N-SH cell line established in the 1970's from a bone marrow biopsy of a metastatic neuroblastoma of a four year old female which has undergone clonal selection (Biedler *et al.*, 1978). The cell line also has the ability to differentiate into a more neuronal phenotype, to help understand synapse formation, with the simple supplement of retinoic acid (RA) (Jahn, Blumer and Pathak, 2017). The karyotype of the SH-SY5Y cell line is diploid, in contrast many cell lines have triploid or tetraploid karyotypes, this means that gene expression is more similar to a normal human cell with limited genomic rearrangement. A trisomy of 1q is what results in the neuroblastoma and drives the uncontrolled division of the cell line (Spengler, Biedler and Ross, 2002). Compared to induced pluripotent stem cells (iPSC), which are up and coming in neuropsychiatric research as the go to cell model, SH-SY5Y cells are still preferred. iPSC are still incredibly expensive to culture, require complex growth media for correct differentiation, and have reproducibility issues, as a result the SH-SY5Y cell line has been used for multiple studies of neuropsychiatric disorders such as schizophrenia, Parkinson's Disease, Bipolar disorder and Alzheimer's (Saavedra *et al.*, 2014; Xia *et al.*, 2015; Trost *et al.*, 2016; Xicoy, Wieringa and Martens, 2017).

2.1.1.2. HEK293

The HEK293 cell line is one of the most cited *in vitro* cell model used in laboratories around the world (52,000 citations on Pubmed to date). The HEK293 cell line, much like the SH-SY5Y cell line, is simple to culture, cheap to maintain, highly reproducible,

easy to transfect, and is highly characterised (Thomas and Smart, 2005). The HEK293 cell line was created by transforming human embryonic kidney cells with sheared fragments of adenovirus type 5 DNA, these HEK293 cells then express a virus-specific tumour antigen allowing them to form tumours and continue to divide (Hill and Iaa, 1977). The HEK293 cell line is highly heterogeneous and comprises mainly of endothelial, epithelial or fibroblast cell types. Because of this the karyotype is complex, there are two or more copies of each chromosome with a modal chromosome number of 64, with three copies of the X chromosome and four copies of chromosome 17 and 22 (Stepanenko and Dmitrenko, 2015).

2.1.1.3. HEK293T

The HEK293T cell line, also known as 293tsA1609neo, is a highly transfectable derivative of the aforementioned HEK293 cell line. These cells contain the temperature sensitive mutant SV40 T-antigen tsA1609, which allows for replication of SV40-orgin vectors at 33°C compared to 40°C (Rio, Clark and Tjian, 1985; Dubridge *et al.*, 1987). The end result is much higher viral titres in cells transfected with lentiviral vectors containing the SV40 origin (Pear *et al.*, 1993). As a result, the HEK293T cell line is the cell line of choice for generation of high quality titred lentiviruses.

2.1.2. Maintenance

Cell lines were maintained in Nunc™ cell culture treated T75 flasks with filter caps (ThermoFisher Scientific, Massachusetts, USA) with Dulbecco's Modified Eagle's Medium (DMEM) F12 containing GlutaMAX™, and supplemented with filtered 10% foetal bovine serum (FBS). Cells were passaged at 80% confluency into two further T75 flasks at a ratio of 1:10 or cells counted using a haemocytometer and seeded at the required density in plates required. Media was changed every three days, or when required. Cells were maintained in an incubator set at 37°C, 5% CO_2 to mimic *in vivo* conditions.

2.1.3. Passaging

All media's were pre-warmed to 37°C in a water bath before being passaged. Cells were passaged once they reached 70-80% confluency to maintain the cell line. Briefly growth media was removed with a disposable 1ml pipette attached to an air aspirator. 3ml of StemPro Accutase (Gibco, ThermoFisher Scientific, Massachusetts, USA) was added to the cells to break down the extracellular matrix and allow for collection. The

flask was tilted until the base was coated with StemPro Accutase before being returned to the 37°C incubator for three minutes. Cells were collected into a sterile falcon tube before the flask was washed with 10ml DMEM F12/10% FBS to collect any remaining cells which could still be attached to the flask. The falcon tube containing cells was spun in a centrifuge at 1000rpm for five minutes to pellet the cells. After pelleting, the supernatant was removed using a disposable pipette tip attached to an aspirator. Cells were resuspended in fresh 10ml of DMEM/F-12 10% FBS media by pipetting up and down 10 times with a 10ml stripette to ensure single cell suspension. Cells were split at 1 in 10 where 1ml of resuspended cells were added to 10ml of DMEM/F-12 10% FBS in a T75 before being returned to the incubator (Splitting 1:10). If specific cell numbers were required cells were counted using a haemocytometer which is described below. The passage (P) number was then increased by one, giving a rough estimate to the age of the cell line.

2.1.4. Cell counting

Cells required at a certain density underwent counting with a haemocytometer before being diluted and added to culture flasks/dishes. Briefly, following resuspension during passaging, 10µl of resuspension is added to a haemocytometer (Hirschmann laboratories, Eberstadt, Germany). The haemocytometer is a set of gridded squares, where the area between the lines and the depth of the chamber is known, therefore the concentration of cells in a fluid can be measured. If the cells are too dense then the resuspension of cells was further diluted 1:10 in DMEM/F-12 10% FBS in an Eppendorf (resulting dilution factor of 10). The concentration of the cells in the original resuspension (cells/ml) was calculated using the following formula:

$$concentration\ of\ cells\ = (\frac{number\ of\ cells\ counted}{4})\ x\ 10^4 x\ dilution\ factor$$

A worked example of this would be that I counted a total of 120 cells over the four corner quadrants from a 1:10 stock dilution of a 10ml resuspension. Therefore the calculation would look like:

$$concentration\ of\ cells = \left(\frac{120}{4}\right)\ x\ 10^4 x\ 10$$

$$concentration\ of\ cells = 3\ x\ 10^6 cells/ml$$

Once the stock cell concentration has been calculated required number of cells can be calculated using the formula:

$$C1\ x\ V1 = C2\ x\ V2$$

Where C1 is the stock cell concentration and V1 is the required volume of stock cells to make a working concentration of C2 in a working volume of V2. The working volume was made up of V1 cells and the remainder with DMEM/F-12 10% FBS media.

2.1.5. Cryogenic preservation and recovery

Cell line stocks underwent long-term storage through cryopreservation under a liquid nitrogen store. Briefly, cells were collected following standard passaging protocol but were resuspended in DMEM/F-12 Glutamax 10% FBS supplemented with 10% Dimethyl sulfoxide (DMSO), the addition of DMSO prevents the formation of ice crystals within the cells preventing cytolysis. 1ml of resuspension was divided into cryogenic storage cryovials before being frozen at -80°C in a Mr Frosty™ (ThermoFisher Scientific, Massachusetts, USA) which cools the cells down at a rate of -1°C/minute to prevent cell lysis. The following day, cryovials are transported on dry ice to the liquid nitrogen store where they are stored in the vapour phase long term at -196°C. When cells are required from liquid nitrogen they are collected on dry ice before being thawed quickly in a water-bath at 37°C. The thawed cells are transferred to a falcon tube containing 10ml DMEM/F-12 Glutamax 10% FBS to dilute the DMSO and prevent cell death. Cells are then pelleted in a centrifuge at 1000rpm for 5 minutes. Diluted freezing media is then removed using a disposable pipette tip connected to an aspirator before being resuspended in 10ml DMEM/F-12 Glutamax 10% FBS. Resuspended cells are then transferred to a T75 flask and incubated at 37°C 5% CO_2 until confluent. Cells were passaged at least twice following thawing before being used in experiments.

2.2. Flow cytometry and fluorescent activated cell sorting

Flow cytometry allows for analysis of multiple individual cells from a heterogeneous population. In simple terms, thousands of cells from a population are passed one by one in single file through one or more lasers which measure the scattering of light and/or fluorescence emissions from antibody staining (Picot *et al.*, 2012). From this we can gather information regarding cell numbers and cell type within this heterogeneous population. We can also use antibody or fluorescent markers to isolate

a pure subpopulation of cells from the bulk sample, this is known as Fluorescent Activated Cell Sorting (FACS). The process is comprised of several steps which are described in more detail below.

Firstly the cell sample is pressurised and injected into the core of a sheath fluid before passing through a nozzle of 100µM width. This hydrodynamic focusing allows cells to align in a single file based on coaxial laminar flow properties to ensure cells are analysed on a one by one (Reynolds, 1883). This step is crucial, as too high pressure allows for more lateral movement of cells and decreases precision as more than one cell could be counted at once. Once the cells have been focused they are passed through the flow cell, this is where the lasers and cells interact. When cells pass through, the laser is refracted in all directions, light diffusion at small angles is known as the forward scatter and correlated to the size of the cell and light diffusion at larger angles (90°) is known as the side scatter, this corresponds to the granularity of the cell. Forward scatter and side scatter can be used in combination to identify different cell types within a population (Salzman *et al.*, 1975). Specificity can be improved in combination with conjugated primary antibodies against the cell type required or by transfection with a fluorescent marker, as the fluorescence can also be recorded and used as a distinguishing marker. Once the detector has identified the light source it is amplified and converted into an electronic signal. The computer then determines if the signal is positive or negative by looking at the pulse height, width and area. From this specific subsets of cells can then be sorted using electrostatic deflection (Chapman, 2000). Cells emerging from the flow cell are put through a vibrating nozzle, which breaks up the stream of individual cells into individual droplets, with one droplet containing one cell. These droplets are then assigned a positive or negative charge, based on the population to be isolated, and are passed through an electric field between two deflection plates. The cell droplets are then deflected into corresponding collection tubes, resulting in a pure subpopulation of desired cells. An overview of the process is shown in **Error! Reference source not found.**.

All FACS was carried out using the BD Biosciences FACSAria III instrument using the 100µM nozzle. The instrument was set up following manufacturer's instructions and the same parameters were kept for all experiments. Cells to be sorted were harvested using accutase as described in 2.1.3. , however cells were resuspended in 1ml of FACS sorting media (10% FBS, DMEM/F-12 w Glutamax No Phenol Red) with the

addition of 1x penicillin streptomycin (pen/strep) if cells were being cultured following sorting. The following laser settings were used for the different flurochromes used in this thesis:

- GFP/eGFP: FITC
- mCherry: PE-CF594
- BFP: DAPI

For each run, the negative control sample was sorted first, this allowed for the gating strategy to be identified. Single cells are identified by their low forward and side scatter and fluorescent positive cells are sorted from this sub-population. An example of a negative stained sample is shown on the left of **Figure 2.1** where anything within the gate (box) will be called as GFP positive. On the right of **Figure 2.1** is a GFP transfected sample, where the GFP positive cells have been called in the gate, these cells would then be sorted as GFP positive single cells.

Cells were sorted into empty Eppendorfs for DNA, Eppenforfs containing 500µl of TRIzol LS (ThermoFisher Scientific, Massachusetts, USA) for RNA, or into 96 well plates containing growth media (10% FBS, DMEM/F-12 w Glutamax, 1x Pen/Strep) if a monoclonal cell line was to be established.

This image has been removed by the author of this thesis for copyright reasons.

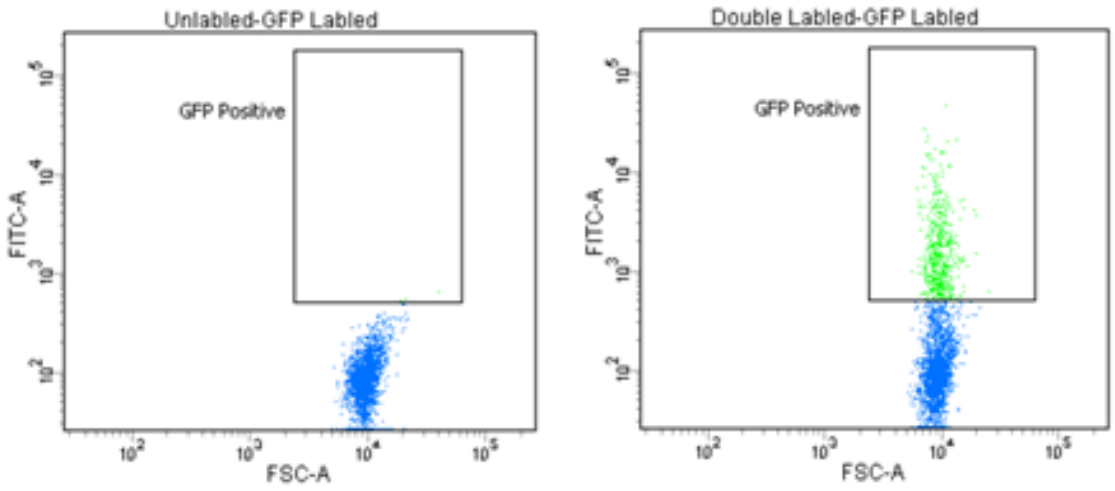


Figure 2.1: Example gating strategy for FACS. *Negative stained sample is on the left where all cells are GFP negative and single cells are identified by blue dots. A GFP transfected sample is shown on the right where GFP positive cells are shown as green dots within the box. X axis is the forward scatter, Y Axis is the FITC intensity.*

2.3. Nucleic acid extraction

This section outlines the methods used throughout this thesis to isolate nucleic acids, (both genomic DNA and total RNA) from cultured cells. All plasticware used during these procedures were sterile and RNase/DNase-free. Before beginning extractions all surfaces and pipettes were cleaned using ethanol and/or RNaseZap (ThermoFisher Scientific, Massachusetts, USA) to remove RNases present and avoid RNA degradation.

2.3.1. Extraction of DNA/RNA from cells

This section describes how cells were harvested for DNA/RNA isolation. Cells were grown to confluency then dissociated using StemPro Accutase (Gibco, ThermoFisher Scientific, Massachusetts, USA). The dissociated cells were then split equally between two sterile Eppendorf tubes. Samples were then spun at 13,000g for one minute in a table top centrifuge to pellet the cells. The supernatant was removed and for DNA the whole cell pellet was collected and stored frozen at -20°C until DNA isolation. For RNA samples the cell pellet was resuspended in 250µl TRIzol™ reagent (ThermoFisher Scientific, Massachusetts, USA) before being stored at -80°C until total RNA isolation.

2.3.2. Genomic DNA isolation

This section describes isolation of genomic DNA from cells. Cell pellets were collected from -20°C and 250µl Slagboom buffer (**Appendix A**) was added directly to the pellet. 0.5µl of RNase A (ThermoFisher Scientific, Massachusetts, USA) was added to each sample to prevent RNA carry over (Meulenbellt *et al.*, 1995). Samples were incubated at 37°C in a heat block for 45 minutes to activate the RNAse to remove residual RNA. After 45 minutes, 2.5µl of Proteinase K (20mg/ml) was added to each sample and mixed by inversion, samples were then incubated at 60°C for one hour in a water bath. Following incubation, samples were cooled to room temperature for five minutes before addition of 50µl majik mix (1:1 ratio yeast Reagent 3 (Autogen Bioclear, Caine, Wiltshire, UK) and 100% ethanol). Samples were vigorously inverted before centrifugation at 13,000 rpm for 10 minutes. The supernatant was collected into a fresh Eppendorf tube and another 50µl majik mix added followed by inversions and centrifugation as above.

The supernatant was collected into a fresh Eppendorf to which an equal volume of isopropanol (ThermoFisher Scientific, Massachusetts, USA) was added to each

sample to precipitate the DNA. Samples were inverted 10 times before being centrifuged at 13,0000rpm for 15 minutes at room temperature to pellet the precipitated DNA. The supernatant was discarded and 300µl of 80% ethanol was added to the samples to wash and remove residual salts from the DNA pellet. Samples were centrifuged at 13,0000rpm for five minutes at room temperature. The supernatant was removed and discarded and the Eppendorf's containing the DNA pellets were left with the lids off inverted on the bench to air dry for 30 minutes or until dry. The DNA was resuspended in 25µl of Tris-EDTA (TE) buffer (10mM Tris, 1mM EDTA) overnight on the bench to fully resuspend.

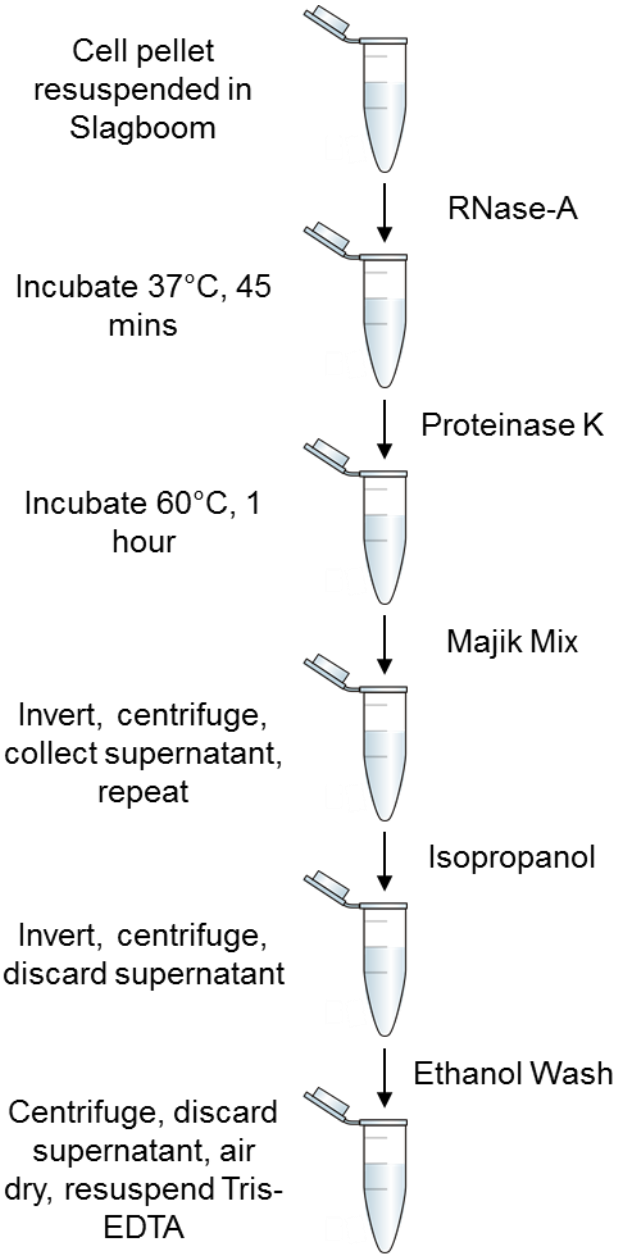


Figure 2.2: Overview of the genomic DNA extraction protocol.

2.3.3. Total RNA isolation

This section describes total isolation of RNA from samples extracted in **2.3.1.** . Total RNA was extracted using the Direct-zol™ RNA MiniPrep kit (Zymo Research, California, USA). All the reagents are supplied with the kit unless otherwise stated. Before starting, the Direct-zol™ RNA PreWash buffer was prepared by adding 10ml of 100% ethanol (ThermoFisher Scientific, Massachusetts, USA) and the RNA wash buffer was prepared by adding 48ml of 100% ethanol (ThermoFisher Scientific, Massachusetts, USA) as indicated in the manufacturer's instructions.

The RNA samples containing TRIzol™ from **2.3.1.** were collected from the -80°C freezer and allowed to thaw to room temperature. Cells were homogenised in TRIzol™ by pipetting up and down ~10 times. 250µl of 100% ethanol (ThermoFisher Scientific, Massachusetts, USA) was added to each RNA sample and mixed thoroughly. The samples were then transferred to a Zymo-Spin™ IIC Column in a collection tube and centrifuged at 12,000g for 30 seconds. The flow through was discarded and the column was transferred to a new collection tube. 400µl of RNA Wash Buffer was added to each column before being centrifuged again at 12,000g for 30 seconds, the supernatant was discarded. In an RNase-free tube 5µl DNase I (6 U/µl) and 75µl DNA digestion buffer were mixed per sample of RNA being isolated and added to the column, columns were then incubated at room temperature for 15 minutes to allow digestion of DNA. 400µl of Direct-zol™ RNA PreWash was added to each column, samples were then centrifuged at 12,000g for 30 second and the flow through discarded, and this step was then repeated. 700µl of RNA Wash Buffer was then added to each column, samples were then centrifuged at 12,000g for two minutes to ensure complete removal of the wash buffer. The column was then transferred to a clean RNase free Eppendorf for elution. For elution, 25µl of DNase/RNase-Free Water was added directly to the column and then centrifuged at 12,000g for 30 seconds. Eluted RNA was then stored at -80°C until used in downstream experiments.

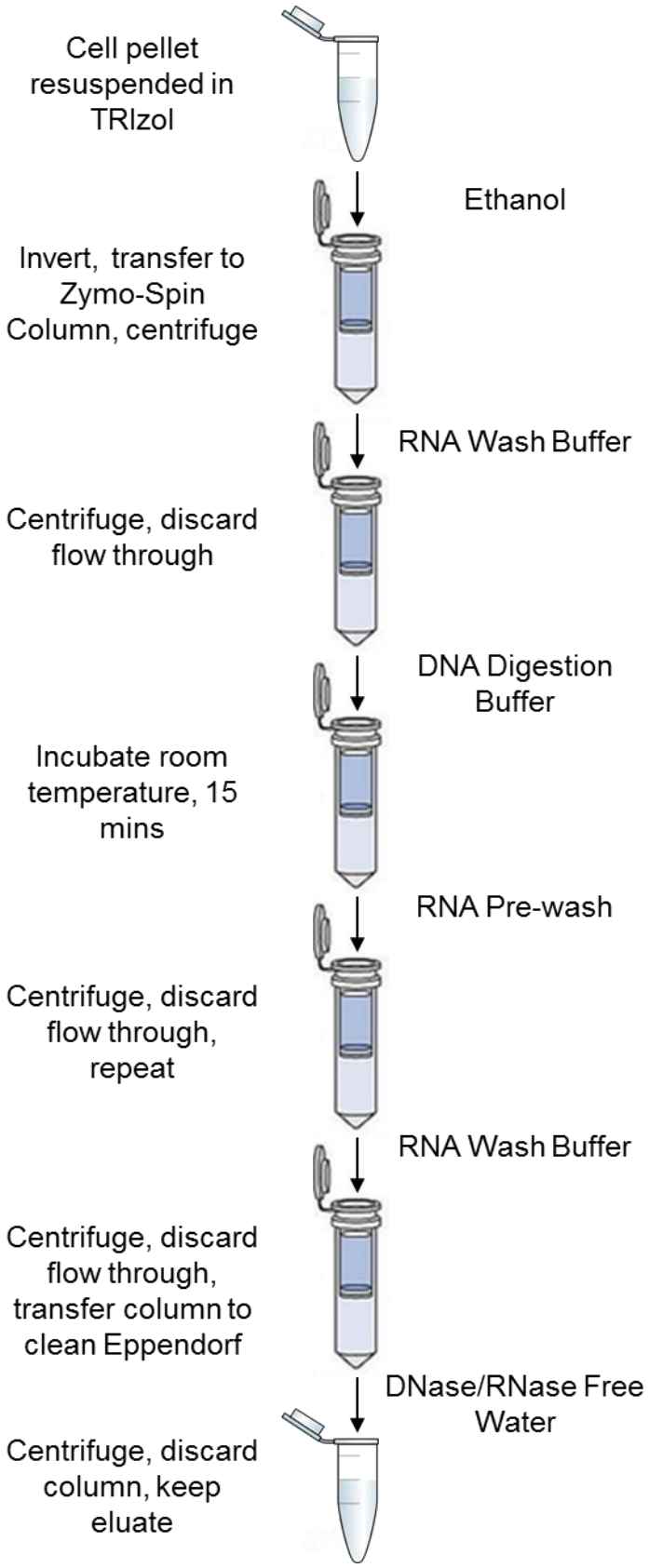


Figure 2.3: Overview of the RNA clean-up protocol.

2.3.4. Quantification/Quality

After extraction, DNA and RNA samples were quantified and checked for purity using the Nanodrop ND-8000 spectrophotometer (ThermoFisher Scientific, Massachusetts, USA), and the Agilent 2200 TapeStation (Agilent Technologies, California, USA) for RNA integrity. These methods are used in other sections of this thesis and their usage has been identified in subsequent chapters.

2.3.4.1. Nanodrop

The Nanodrop spectrophotometer utilises the wavelength absorbance of nucleic acids and other contaminates to calculate the quantity of nucleic as well as its purity. Nucleic acids absorb ultra-violet (UV) light at a wavelength of 260nm, however proteins absorb UV light at 280nm and other common contaminants such as EDTA, ethanol, and phenol absorb UV at approximately 230nm. Therefore the absorbency ratios of 260/280 and 260/230 indicate the presence of potential contaminants in the nucleic acid samples. A 260/280 ratio of ~1.8 and ~2.0 for DNA and RNA respectively and a 260/230 ratio between 1.8 and 2.2 is indicative of a high purity sample. The Nanodrop was always blanked before use using the resuspension/elution buffer corresponding to the samples being measured. We can use this to our advantage to decipher where contamination may have arose in the experiment and if samples can be used for downstream experiments. If Nanodrop results were inconclusive through samples being too dilute or concentrated an agarose gel was run to confirm if samples was present and is described below. **Figure 2.5** shows a typical absorbance reading for DNA.

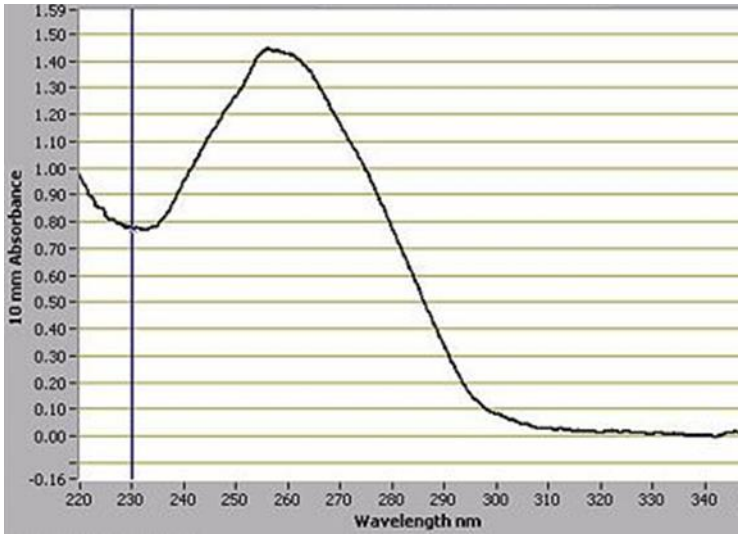


Figure 2.4: A typical DNA sample gives a characteristic profile on the Nanodrop.

2.3.4.2. Tapestation

RNA samples sent for sequencing underwent a further check using an Agilent 2200 TapeStation (Agilent Technologies, California, USA) in conjunction with RNA ScreenTape (Agilent Technologies, California, USA). The TapeStation uses a fluorescent dye which binds to RNA run on a gel electrophoresis chip, the software then analyses the resultant gel and calculates an RNA integrity number (RIN). This RIN can range between 1 and 10, with 1 suggesting very high degradation of RNA and 10 suggesting perfect integrity of RNA. These RINs are calculated by looking at peaks in the 18S and 28S subunits of the ribosomal RNA (Schroeder *et al.*, 2006). Samples with high RIN numbers were sent for sequencing and are outlined in future chapters in more detail.

2.4. Polymerase chain reaction

Polymerase chain reaction (PCR) is a method used to amplify a segment of DNA to generate thousands to millions of copies of the same segment. The components of the reaction are combined in a single reaction and are subjected to cycles of heating and cooling in a thermocycler. A standard PCR reaction mix set up is described in **Table 2.1**. The first step of PCR is to heat the mix to activate the heat sensitive taq polymerase. This is then followed in a cycle of three different temperatures:

1. Denaturing step: The mix is heated to 95°C to denature the double stranded DNA to be amplified.
2. Annealing step: The mix is then cooled to a primer specific temperature (usually between 50°C and 65°C). This allows the primers to bind specifically and prevent the formation of non-specific amplicons.
3. Elongation step: The mix is then heated back up to 72°C to allow the taq polymerase to synthesize the complementary strand of DNA using free deoxyribonucleotides (dNTPs) within the mixture.

These three steps are repeated for a number of cycles to allow for the synthesis of DNA amplicons, usually between 35 and 45 times dependent of application. The final number of amplicons will be 2^n where n represents the number of cycles. A final elongation step of 72°C is added at the end to allow for final extension. The cycling conditions, primer sequences, PCR reagents used in this thesis are outlined in individual chapters where required.

Table 2.1: Polymerase Chain Reaction (PCR) reagents and quantities used in standard PCR through this thesis.

Component	Function	PCR Reagent (Concentration)	Quantity (µl)
Genomic DNA	Single stranded DNA provides the template for the PCR Reaction	Genomic DNA (50-100ng/ul)	1
DNA Primers	Short, single stranded, DNA oligonucleotides which are complementary to the target sequence and allow *Taq* polymerase to bind.	Forward and reverse primer (10µM each)	0.5
DNA Nucleotides	Nucleotide bases required for synthesis of new DNA strands	Mix of dATP, dGTP, dTTP, dCTP (2.5mM)	0.4
Taq DNA polymerase	Heat-resistant enzyme which binds to primed DNA and synthesises new DNA strands complementary to the target sequencing using available dNTPs	Qiagen HotStar *Taq* Polymerase (5 units/µl) (Qiagen, Venlo, Holland)	0.1
PCR Buffer	Buffer which maintains the optimum pH and salt concentration for PCR reaction	10 x PCR Buffer (Qiagen, Venlo, Holland)	2
RNase/DNase Free Water	Ensures consistent reaction volume and salt concentration	Water to final volume of 20µl	16

2.4.1. Agarose gel electrophoresis

Agarose gel electrophoresis is a method for separating out different sized fragments of DNA by exploiting the negatively charged properties of DNA. This technique is commonly used for identifying if a PCR reaction or restriction digest has worked correctly. An agarose gel is a 3D matrix composed of pores, of which DNA can migrate through when a charge is applied. The gel is made by combining agarose powder with Tris-borate EDTA (TBE) buffer at different percentages, usually 0.5-2.5% weight/volume (w/v). The percentage of agarose depends on the size of fragments to be separated, with lower percentages having bigger pores for larger fragments and high percentage gels have smaller pores. After addition of DNA to the mix with a loading dye an electrical current is applied to the gel, causing the negatively charged DNA to migrate towards the cathode, at a speed determined by the size of the fragment. Smaller molecules can move faster and further through the agarose pores, with larger fragments remaining near the anode. The separated DNA is then viewed using a stain, the most commonly used stain is ethidium bromide, this intercalates to double stranded nucleic acids and is visualised under UV light. The UV light is absorbed by the ethidium bromide bound to nucleic acid and is re-emitted as visible light, allowing the fragments to be observed. In this thesis the percentage of agarose gels, running voltage and time are listed in subsequent chapters and methods.

2.4.2. PCR column clean up

Following PCR, amplicons may be used for downstream purposes such as restriction digestion or sequencing. As a result, the PCR reaction should be purified to remove any buffer, salts, unused dNTPs, and primers which may inhibit downstream applications. PCR reactions were cleaned up using the GeneJET PCR Purification Kit (ThermoFisher Scientific, Massachusetts, USA), which uses a column based approach to bind PCR products while washes remove any impurities. Before starting, the Wash Buffer was diluted using 45ml of 100% ethanol (ThermoFisher Scientific, Massachusetts, USA).

PCR products to be purified were combined in a 1:1 volume ratio with Binding Buffer and mixed thoroughly. If the amplicon was less than 500bp a 1:2 volume of 100% isopropanol (ThermoFisher Scientific, Massachusetts, USA) was added to the samples and mixed thoroughly. Samples were then transferred to a GeneJET

purification column and centrifuged at 12,000g for one minute at room temperature. The flow through was discarded and 700µl of Wash Buffer added to the column, the columns were then centrifuged at 12,000g for one minute at room temperature and the flow through discarded. The empty columns were then centrifuged for a further minute at 12,000g to remove any residual ethanol. The column was then transferred to a DNAse free sterile Eppendorf and 25-50µl of Elution Buffer was added directly to the centre membrane of the column, this was incubated at room temperature for one minute. Columns were then centrifuged for one minute at 12,000g to elute the PCR product. Samples were then quantified by Nanodrop and used in downstream processes or stored at -20°C. An overview of the protocol is outlined in **Figure 2.6**.

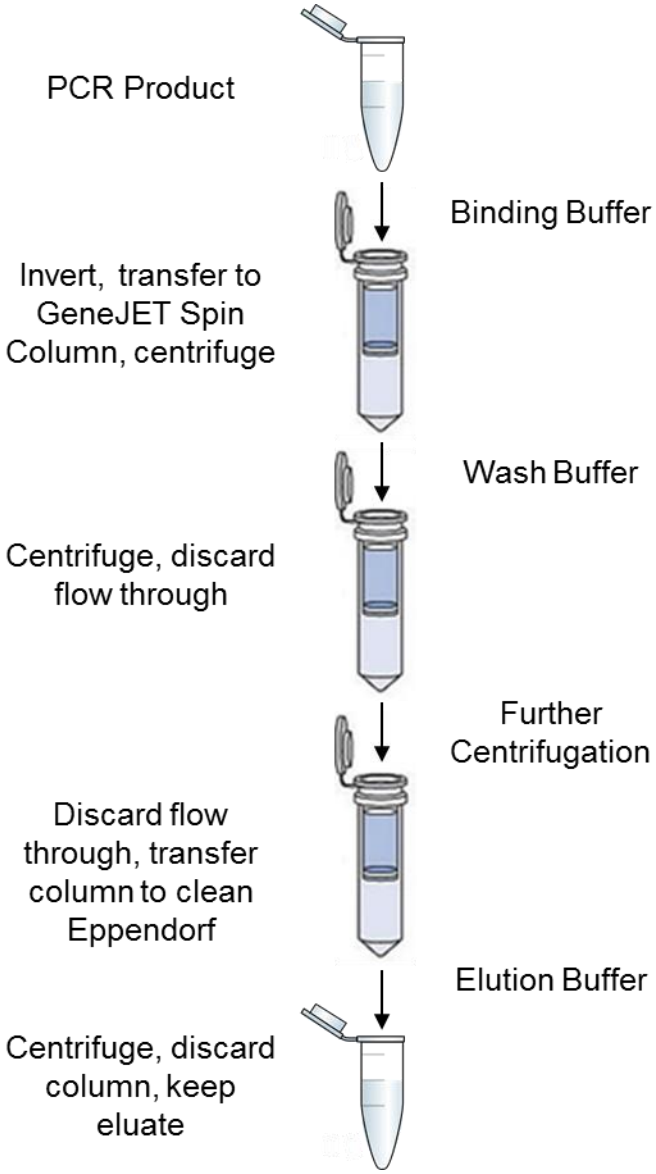


Figure 2.5: Overview of the PCR clean-up protocol.

2.4.3. Agarose gel DNA extraction

Where DNA was amplified through PCR and required for cloning, or plasmids cut with restriction enzymes resulted in multiple bands, a gel extraction was carried out. Agarose gels were moved to a UV Box and bands of interested were excised from the gel using a fresh clean scalpel blade and collected into a sterile Eppendorf. To reduce the risk of UV directed mutagenesis, agarose gels were placed into a pipette tip box lid before being placed onto the UV box, the gel was only exposed to UV light for several seconds to visualise the band to reduce the chance of mutation. Following excision of the band the DNA was extracted following the GeneJET Gel Extraction Kit Protocol (ThermoFisher Scientific, Massachusetts, USA). Before starting the Wash Buffer was diluted using 45ml of 100% ethanol (ThermoFisher Scientific, Massachusetts, USA).

The excised gel slice was weighed and the empty weight of the Eppendorf subtracted. Binding buffer was added in a 1:1 ratio to the gel slice (volume:weight) and the mixture was incubated at 60°C for 10 minutes or until the gel slice was dissolved. Samples were inverted every two minutes to aid in melting. If the gel fragment was less than 500bp in length one gel volume of 100% isopropanol was added, if larger than 10kb one gel volume of double distilled water (ddH_2O) was added. Samples were then transferred to a GeneJET purification column and centrifuged at 12,000g for one minute at room temperature. The flow through was discarded and 700µl of Wash Buffer added to the column, the columns were then centrifuged at 12,000g for one minute at room temperature and the flow through discarded. The empty columns were then centrifuged for a further minute at 12,000g to remove any residual ethanol. The column was then transferred to a DNAse free sterile Eppendorf and 25-50µl of Elution Buffer was added directly to the centre membrane of the column, this was incubated at room temperature for one minute. Columns were then centrifuged for one minute at 12,000g to elute the PCR product. Samples were then quantified by Nanodrop and used in downstream processes or stored at -20°C. An overview of the protocol is outlined in **Figure 2.7**.

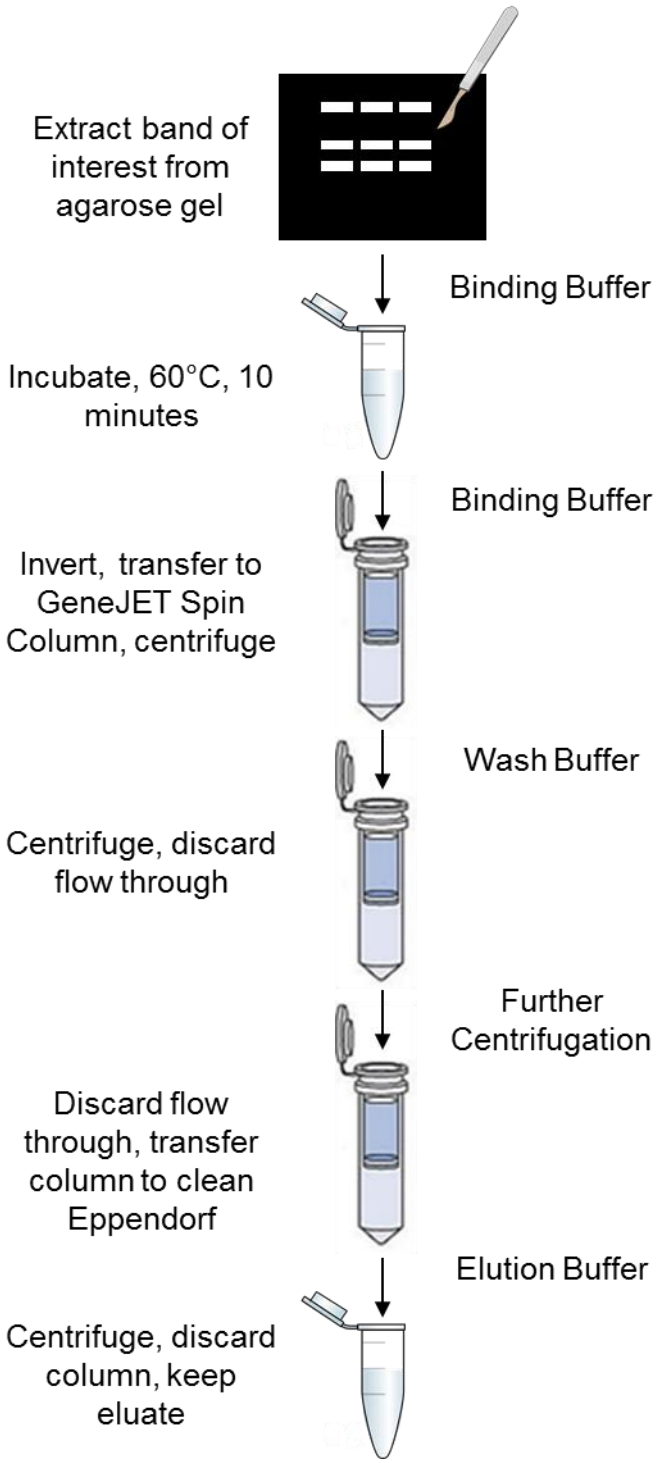


Figure 2.6: Overview of the DNA extraction from agarose gel protocol.

2.5. Quantitative PCR (qPCR)

Quantitative PCR (qPCR) is used to relatively quantify the levels of RNA transcripts within a complementary DNA (cDNA) sample in real time. As the number of PCR cycles increases as does the number of PCR amplicons relative to the number of starting template cDNA, with higher expressed transcripts reaching the plateau phase of PCR at earlier cycles than that of lower expressed transcripts. qPCR can use several methods to detect PCR amplicons in a PCR reaction in real time, for this thesis EVAGreen technology was used. EVAgreen is a fluorescent dye which intercalates into dsDNA, thus as the number of PCR amplicons increases as does the fluorescence. Every cycle the fluorescence of each qPCR reaction is recorded and once the fluorescence signal crosses a threshold above the background fluorescence a CT value is calculated. This CT value can then be compared across samples to work out a fold change in gene expression relative to a control group. Input cDNA is normalised through reverse transcription and through qPCR of known housekeeping genes such as *β-Actin* or *GAPDH*, which have stable expression within different cells types and conditions.

2.5.1. cDNA synthesis

For cDNA generation the SuperScript VILO cDNA Synthesis Kit (Invitrogen, Carlsbad, California, USA) was used. 100ng of RNA was converted into cDNA in a 20µl reaction set up listed in **Table 2.2** under thermocycling conditions listed in **Table 2.3**. A 100% conversion efficiency from RNA to cDNA was assumed to calculate the concentration of cDNA at 5ng/µl. cDNA was diluted to 1ng/µl for use in qPCR reactions.

2.5.2. qPCR set up

qPCR reactions were set up as listed in **Table 2.4** in white walled 96 well plates or 384 well plates. Each sample, including input and negative controls were run in triplicate for all primer sets. The qPCR primers for the housekeeping genes *β-Actin* and *GAPDH* are listed in **Table 2.5**. The qPCR primers used in subsequent chapters are listed in each chapter.

Table 2.2: Creation of cDNA using the SuperScript VILO cDNA synthesis kit.

Component	Function	VILO cDNA Synthesis Reagent (Concentration)	Quantity (µl)
RNA	Single strand RNA acts as a template for cDNA synthesis	RNA (100ng/µl)	1
Reaction Mix	Includes random primers, $MgCl_2$, and dNTPs required for successful cDNA synthesis	5x VILO Reaction Mix (Invitrogen, Carlsbad, California, USA)	4
Reverse Transcriptase	The enzyme mix which contains SuperScript III reverse transcriptase for conversion of RNA into cDNA, RNase OUT for removal of RNase, and a helper protein.	10x SuperScript Enzyme Mix (Invitrogen, Carlsbad, California, USA)	2
RNase/DNase Free Water	Ensures consistent reaction volume and salt concentration	Water to final volume of 20µl	Variable

Table 2.3: Thermocycling conditions for cDNA synthesis.

Step	Temperature (°C)	Time	Number of Cycles
Priming	25	10 min	1
Synthesis	42	60 min	2
Termination	85	5 min	1
Storage	4	∞	1

Table 2.4: Reaction set up for qPCR using EvaGreen.

Component	Function	qPCR Reagent (Concentration)	Quantity (µl)
cDNA	Single stranded cDNA provides the template for the PCR Reaction	cDNA (1ng/µl)	1
qPCR Primers	Short, single stranded, DNA oligonucleotides which are complementary to the target sequence and allow *Taq* polymerase to bind.	Forward and reverse primer (10µM each)	0.5
EvaGreen Polymerase Master Mix	Nucleotide bases required for synthesis of new DNA strands	Complete master mix containing Mg^{2+}, nucleotides and *Taq* polymerase. (Solis Biodyne, Tartu, Estonia)	4
RNase/DNase Free Water	Ensures consistent reaction volume and salt concentration	Water to final volume of 20µl	14

Table 2.5: qPCR primers for the housekeeping genes.

Gene	Primer Name	Sequence (5'–3')
β-Actin	ACTB F	CTGGGAGTGGGTGGAGGC
	ACTB R	TCAACTGGTCTCAAGTCAGTG
GAPDH	GAPDH F	GTGAAGGTCGGAGTCAACG
	GAPDH R	TGAGGTCAATGAAGGGGTC

2.5.3. QuantStudio and StepOnePlus set up

384 well qPCR reactions were carried out using the QuantStudio 6 Flex qPCR machine (Applied Biosystems, Foster City, California, USA) and 96 well qPCR reactions in the StepOnePlus Real-Time PCR System (Applied Biosystems, Foster City, California, USA). The qPCR program used is listed in **Table 2.6**, EvaGreen fluorescence was measured after the extension stage.

2.5.4. Data analysis

Raw CT values were collected from the software and the standard deviation (SD) of the triplicates of each primer set per sample were calculated. Any CT values +/- 1.5 SD outside of the mean were excluded and the average CT value was calculated. Differential expression was calculated using the $2^{-\Delta\Delta CT}$ method using the CT values for *β-Actin* and *GAPDH* as the normalising controls (Livak and Schmittgen, 2001).

Table 2.6: qPCR cycling conditions used for all qPCR reactions in this thesis.

Step	Temperature (°C)	Time	Number of Cycles
Hotstart	95	12 min	1
Denaturation	95	30 sec	
Annealing	60	30 sec	45
Extension	72	30 sec	

2.6. Molecular cloning

Multiple chapters within this thesis use plasmids generated through molecular cloning. Molecular cloning involves the creation of recombinant DNA to allow expression of certain genes within specific organisms. These recombinant DNA are replicated in bacterial cells. Once extracted, these can be later used in human cells for functional purposes such as gene expression.

The basics of molecular cloning are the digestion of a recipient plasmid (or backbone vector) using a restriction endonuclease to linearize the circular plasmid and to create single stranded DNA overhangs known as sticky ends. The section of DNA to be added to these vectors (Insert) can be either a PCR product whose ends have been digested with the same restriction endonuclease as the backbone vector, or individually ordered oligos containing sticky ends which have been annealed with its complementary sequence to give double stranded fragments. Inserts are then added to digested backbone vector at a molar ratio of 3:1 and the sticky ends annealed through a T4 ligase enzyme to form a recombinant plasmid. Recombinant plasmids are then transformed into a competent bacterial host, usually a strain of *Escherichia coli*, through heat shock treatment. These bacteria now contain host DNA and recombinant plasmids and are then plated on a selective agar where they undergo replication. During cell replication, both the host DNA and recombinant plasmid are copied resulting in a colony of clones which have identical host DNA and recombinant plasmid. These bacterial cells are then collected and the plasmid extracted for use in downstream experiments.

Detailed descriptions of the molecular cloning techniques used through this thesis are outlined below. **Figure 2.8** shows an overview of molecular cloning and the generation of recombinant plasmids using PCR.

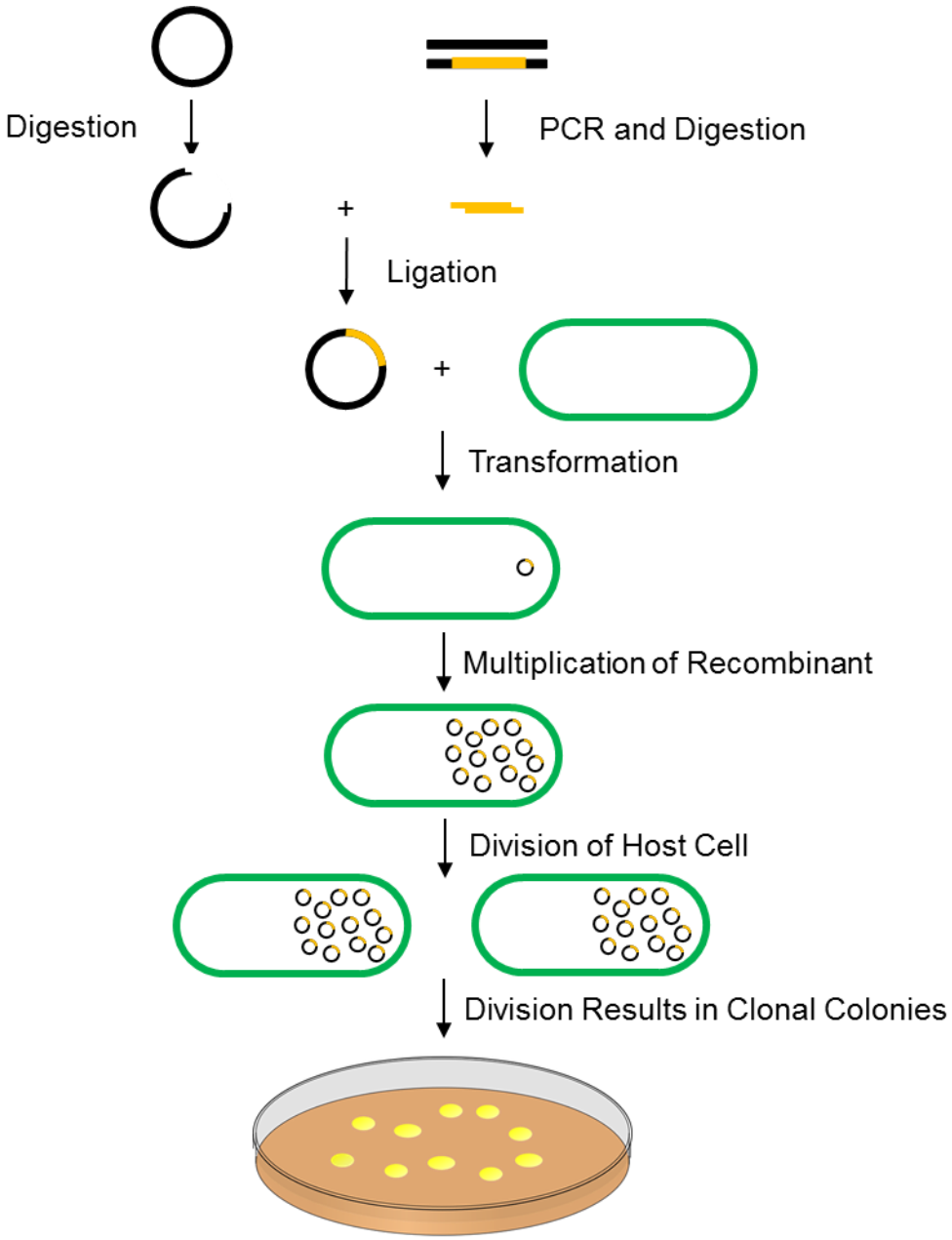


Figure 2.7: An overview of molecular cloning. *The region of interest (orange) is amplified from DNA using PCR. The destination vector (black) and PCR product are digested with the same restriction enzyme resulting in complementary sticky ends. These digested products are ligated together to give a recombinant plasmid. These recombinants are then transformed by heat-shock into a host cell E.coli (green). These host cells are then plated on selective agar. Recombinants then multiply using the host cell machinery and the host cells divide. Any host cells containing the recombinant survive on selective agar and form clonal colonies. These colonies are then screened for the insert of interest.*

2.6.1. Generation of insert through PCR.

One of the most common ways for generating an insert for cloning is to PCR the region of interest from human DNA. Primers are designed to amplify the region to be inserted, these primers then have the nucleotide sequence for the restriction enzyme(s) used to linearize the vector backbone with six additional random nucleotides added at the 5' end of the primer. The DNA is then amplified through PCR using a higher fidelity polymerase to reduce mistakes made through standard *taq* polymerase, where *taq* polymerase has an error rate of $2.3x10^{-5}$ errors per bp per PCR cycle (Bobilin and Ching, 2006). Phusion® High-Fidelity DNA polymerase (New England Biolabs, Ipswich, Massachusetts, USA) was used for these PCR reactions as it as 50 times higher fidelity than standard *taq* polymerase ($4.2x10^{-7}$ errors per bp per PCR cycle) and possesses 3' to 5' exonuclease proofreading activity to prevent mismatching through the PCR (Bobilin and Ching, 2006). A standard reaction set up is listed in **Table 2.7**. Following PCR, the resultant amplicon then contains at the 5' end six random bases, the restriction site, and the sequence of interest, the product is then run on an agarose gel to separate the amplicons and recovered through a gel extraction (see **2.4.3.**).

Following PCR amplification, the amplicon contains blunt ends comprising of six random bases, the restriction site and then the sequence of interest. These blunt ends are then digested by the restriction endonuclease to form sticky ends (**Error! Reference source not found.**). These reactions are cleaned through PCR clean up columns (as outlined in **2.4.2.**) to yield a pure insert to be used in ligation. Further information is given through this thesis regarding enzymes used, primers, and PCR conditions.

Table 2.7: High fidelity PCR reaction components used to generate inserts for molecular cloning.

Component.	Function	PCR Reagent (Concentration)	Quantity (µl)
Genomic/Plasmid DNA	Single stranded DNA provides the template for the PCR Reaction	Genomic/Plasmid DNA (50-100ng/ul)	1
DNA Primers	Short, single stranded, DNA oligonucleotides which are complementary to the target sequence and allow *Taq* polymerase to bind. These contain the restriction site for the enzyme used for cloning at the 5' end along with six extra random bases to aid with enzyme binding.	Forward and reverse primer (10µM each)	1
DNA Nucleotides	Nucleotide bases required for synthesis of new DNA strands	Mix of dATP, dGTP, dTTP, dCTP (2.5mM each)	0.4
High Fidelity DNA polymerase	Heat-resistant enzyme which binds to primed DNA and synthesises new DNA strands complementary to the target sequencing using available dNTPs. This polymerase contains 3'-5' proofreading ability and results in fewer errors than standard *Taq* polymerase	Phusion® High-Fidelity DNA polymerase Polymerase (5 units/µl) (New England Biolabs, Ipswich, Massachusetts, USA)	0.2
PCR Buffer	Buffer which maintains the optimum pH and salt concentration for PCR reaction	5 x Phusion HF Buffer (New England Biolabs, Ipswich, Massachusetts, USA)	4
RNase/DNase Free Water	Ensures consistent reaction volume and salt concentration	Water to final volume of 20µl	13.4

This image has been removed by the author of this thesis for copyright reasons.

2.6.2. Digestion of plasmid vector backbone

In order to clone the region of interest into a plasmid vector, the plasmid must first be cut with a restriction enzyme. Restriction enzymes are endonucleases present in bacterial DNA which are used in viral immunity (Roberts, Murray and Roberts, 1976). These endonucleases recognise specific sequences of the DNA, known as a restriction site, and cleave the DNA causing a double strand break. This cleavage can result in either a sticky end or a blunt end at the cleavage site. Sticky ends result in a 4-6bp overhang which can be exploited for cloning, whereas blunt ends result in no overhangs but can still be exploited for molecular cloning, albeit with reduced efficiency. The amount of available restriction sites are varied. For example, of the 3,000 restriction enzymes stocked by the company Promega, 230 different DNA sequences are recognised and cut.

As most of the recognition sites are palindromic the vector could spontaneously ligate back together during the ligation step, or the insert could be ligated in the incorrect orientation. To minimise this, two differing restriction enzymes are used for each end of the insert and for cutting the vector backbone. This then forces the insert into the vector in the correct orientation and prevents the spontaneous re-ligation of the plasmid backbone. Another way to minimise this is to use a restriction enzyme which contains N_n recognition sites, this means it can target one site but results in two incompatible overhangs, allowing for the usage of a single enzyme. It is beneficial to use a single enzyme as using multiple restriction enzymes which have unique buffers decreases the endonuclease activity in single reactions. To overcome this vectors may have to be digested sequentially with different enzymes.

All restriction enzymes used and mode of digestion are outlined in sequential chapters. A standard enzyme digestion of vectors and inserts is listed in **Table 2.8**.

Table 2.8. Standard restriction digestion reaction for digestion of vector backbones and inserts.

Component	Function	Quantity (µl)
Plasmid DNA/PCR Insert	Backbone for which insert is to be cloned into, contains genes involved in vector function, antibiotic resistance genes, and/or fluorescent markers. Insert to be added to plasmid to provide additional functions	500ng
Restriction Enzyme	Two restriction endonuclease enzymes required for cutting at specific sequence. Binds to recognition sequence and cleaves DNA, resulting in 4-6bp sticky end overhangs which are not compatible with each other	1 Each
Restriction Buffer	Optimal buffer to the restriction enzymes being used. Contains salts and cofactors used for successful digestion. Maintains pH required for reaction	2
RNase/DNase Free Water	Ensures consistent reaction volume and salt concentration	To 20µl

2.6.3. Ligation of plasmid and insert

Ligating the insert into the plasmid vector is possible due to the complementary overhangs created by digesting both the plasmid vector and insert with the same restriction enzyme. The 5’ overhangs on the insert match by Watson-Crick complementary base pairing to the 3’ overhangs present in the digested plasmid. This is facilitated by the T4 DNA ligase, which is a ligase originally derived from bacteriophage T4, to ligate the insert to the vector backbone. The typical molar ratio for ligation of insert:backbone was 3:1 and was calculated using the online calculator from New England Biolabs (https://nebiocalculator.neb.com/#!/ligation). A standard ligation was set up as listed in **Table 2.9** and incubated at 23°C for one hour then 16°C overnight in a thermocycler.

Table 2.9: Composition of a standard T4 ligation reaction.

Component	Function	Reagent	Quantity (µl)
Digested Vector Backbone	Digested vector backbone into which insert it to be inserted	Plasmid DNA (50-100ng)	Variable
Digested Insert	Digested PCR or plasmid insert to be inserted into digested vector backbone	Insert DNA (1:3 Molar Ratio over Vector)	Variable
T4 DNA Ligase	Ligase enzyme from T4 bacteriophage which ligates the phosphate backbone between the insert and backbone vector once paired by Watson-Crick complementary base pairing.	T4 DNA Ligase (ThermoFisher Scientific, Massachusetts, USA)	1
Buffer	Buffer which contains ATP, salts, and maintains the optimum pH and concentration for ligation reaction	5 x T4 Ligase Buffer (ThermoFisher Scientific, Massachusetts, USA)	4
RNase/DNase Free Water	Ensures consistent reaction volume and salt concentration	Water to final volume of 20µl	Variable

2.6.4. Transformation of recombinant plasmid into bacteria

The ligation reaction from **2.6.3.** will contain a mix of different plasmids, including recombinants and non-digested backbone vector which have been carried over. In order to overcome this the ligation reaction is mixed with a bacterial cell, *E. coli.* These bacteria are then placed under stress by heat shocking at 42°C to induce horizontal gene transfer of a single plasmid into an *E. coli*. The cells are then plated on a selective agar containing antibiotic, for which the plasmid contains resistance genes, thus allowing survival of bacteria which have undergone horizontal gene transfer. These bacteria then multiply and, in turn, copy the transformed plasmid. This results in clonal colonies formed on the agar plates, each with millions of copies of a single recombinant plasmid which can then be screened.

On the day of transformation, a vial of sub cloning efficiency DH5α strain of *E. coli* (Invitrogen, Carlsbad, California, USA) were thawed on ice for one hour. A bottle of nutrient rich Super Optimal broth with Catabolite repression (S.O.C) and pre-made agar plates containing selection antibiotic were removed from the cold store and placed in a 37°C incubator. After one hour, 50µl of DH5α *E. coli* were added to curved bottom propylene tubes along with 5µl of the ligation reaction from **2.6.3.** or 2.5µl of pUC19 transformation control vector. Tubes were then gently flicked to mix and incubated on ice for 30 minutes. After 30 minutes the tubes were submerged in a water bath for set at 42°C for 20 seconds to heat shock the DH5α *E. coli*, promoting horizontal gene transfer. After 20 seconds tubes were placed back on ice for two minutes. After two minutes 950µl of pre-warmed S.O.C media was added to each tube and these were then incubated at 37°C with shaking incubation at 225rpm for one hour. This step is to allow for recovery following heat shock and to allow transcription of antibiotic resistant genes. Following incubation, two different volumes of bacteria were spread plated on agar containing selection antibiotic, these were then inverted and incubated overnight at 37°C. One or both plates would contain individual colonies which could be counted in the morning.

An overview of the bacterial transformation is provided in **Figure 2.10**. Bacterial lines and antibiotics used are listed in subsequent chapters.

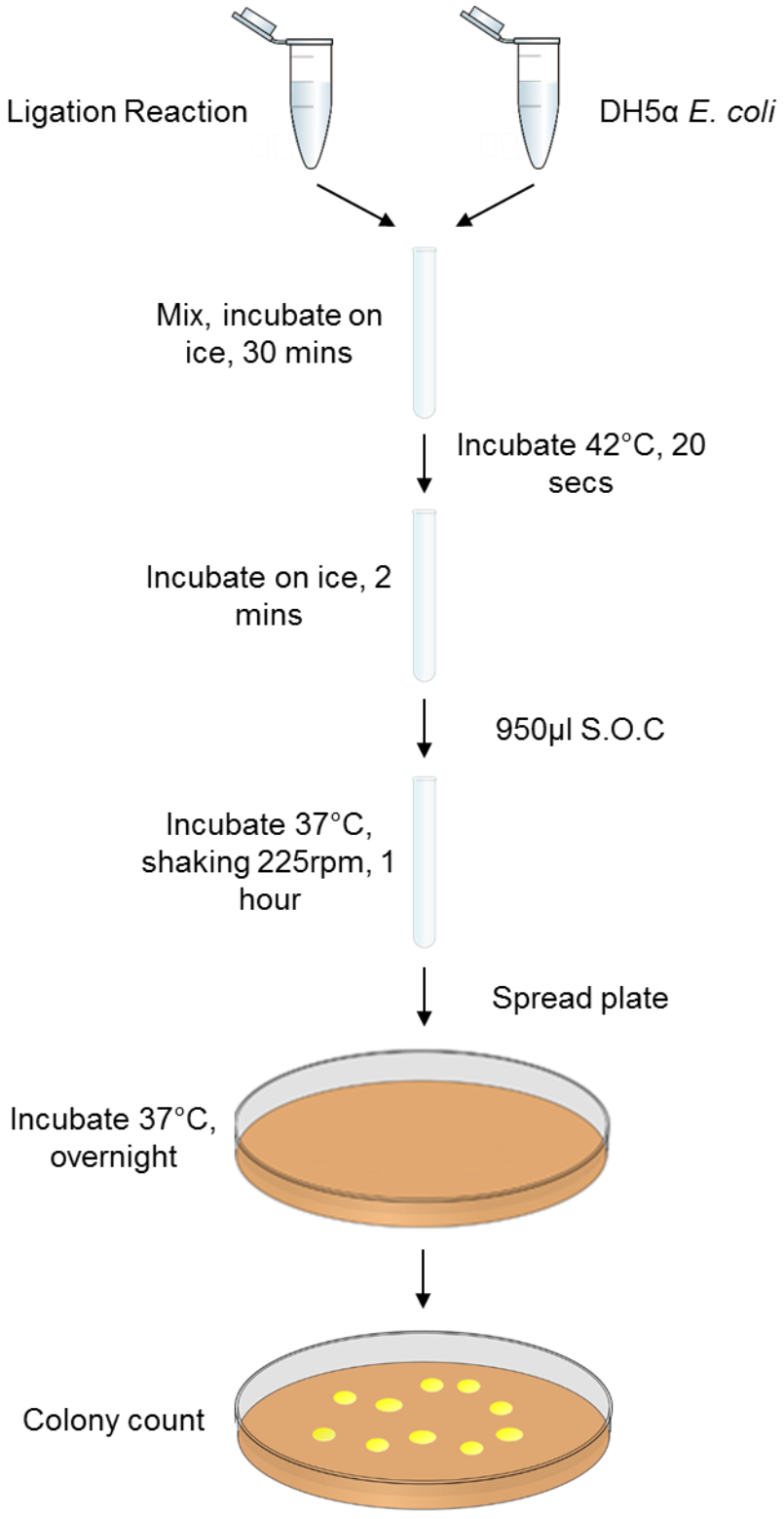


Figure 2.8: Overview of bacterial transformation of ligated plasmids into E.coli.

2.6.5. Colony PCR

Overnight incubation results in hundreds of clonal colonies present on the agar plate, with each colony containing a genetically different recombinant plasmid. Colony PCR is carried out to reduce workload on downstream steps as a screening method to identify colonies which have the recombinant of interest. Colony PCR utilises the sequencing primer of the recombinant plasmid backbone, along with a primer to the reverse complement of the insert. This results in a PCR product being observed in insert positive recombinants but not in insert negative, allowing for quick and robust screening of colonies. Eight colonies are selected at random from each agar plate and numbered, half of the colony are then scraped using a sterile pipette tip. The bacteria are then transferred into an eight strip PCR tube by rubbing the pipette tip against the bottom of the well. A colony PCR master mix, as listed in **Table 2.10**, is added directly to the eight strip tube and cycling carried out in a thermocycler with the cycles listed in **Table 2.11**. Agarose gel electrophoresis is carried out as listed in **2.4.1.** using 2µl of PCR product, on a 2.5% agarose gel, with running voltage of 120V for 30 minutes, an example result of colony PCR is shown in **Figure 2.11**.

Table 2.10: Components of the colony PCR reaction.

Component	Function	PCR Reagent (Concentration)	Quantity (µl)
Bacterial Colony	Bacterial cells which contain copies of the recombinant plasmid required for experiments	Plasmid DNA (Varies)	Variable
Plasmid Sequencing Primer	The sequencing primer is upstream of the cloning site, the primers used for sequencing are listed in subsequent chapters	Forward sequencing primer (5µM)	1
Insert Reverse Complementary Primer	The insert reverse complementary primer is unique to the insert being cloned. This allows for amplicons only in recombinants containing the insert. The primers are listed in subsequent chapters	Insert reverse complement primer (10µM)	0.5
DNA Nucleotides	Nucleotide bases required for synthesis of new DNA strands	Mix of dATP, dGTP, dTTP, dCTP (2.5mM)	0.4
Taq DNA polymerase	Heat-resistant enzyme which binds to primed DNA and synthesises new DNA strands complementary to the target sequencing using available dNTPs	Qiagen HotStar *Taq* Polymerase (5 units/µl) (Qiagen, Venlo, Holland)	0.1
PCR Buffer	Buffer which maintains the optimum pH and salt concentration for PCR reaction	10 x PCR Buffer (Qiagen, Venlo, Holland)	2
RNase/DNase Free Water	Ensures consistent reaction volume and salt concentration	Water to final volume of 20µl	16

Table 2.11: Colony PCR thermocycling conditions.

Step	Temperature (°C)	Time	Number of Cycles
Hotstart	95	5 min	1
Denaturation	94	30 sec	
Annealing	60	30 sec	35
Extension	72	1 min	
Final extension	72	5 min	1

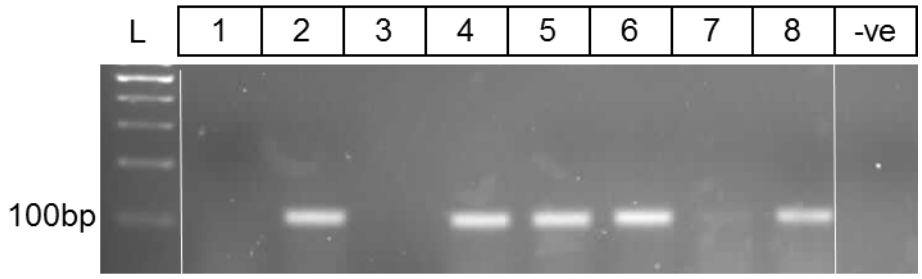


Figure 2.9: Example agarose gel electrophoresis result from a colony PCR reaction. *Each column is an individual colony from one agar plate. Colonies 2,4,5,6 and 8 contain the insert, colonies 1 and 3 do not, and colony 7 is inconclusive. The negative control (-ve) is master mix with no bacterial colony. L is DNA ladder (100bp DNA Ladder Ready to Load (Solis biodyne)).*

2.6.6. Overnight bacterial cultures

Colonies on agar plates are not large enough to extract sufficient plasmid DNA for downstream experiments. In order to get enough plasmid DNA, individual colonies are inoculated into Lysogeny Broth (LB) with selection antibiotic and are grown overnight. LB is a nutrient rich broth which allows for the support of larger volumes of bacteria, thus producing enough plasmid DNA for downstream experiments.

Two or more colonies, which are positive for the insert from the colony PCR, are taken into overnight inoculation with LB and appropriate selective antibiotic. The volumes of LB are dependent on the type of prep carried. For miniprep, 5ml LB containing antibiotic is aliquoted into curved bottom tubes, the remaining half of the colony (identified from **2.6.5.**) is collected using a sterile pipette tip, and this pipette tip is then dropped into the curved bottom tubes. For maxiprep, 200ml of LB containing antibiotic is aliquoted into autoclaved conical flasks, the remaining half of the colony (identified from **2.6.5.**) is collected using a sterile pipette tip, and this pipette tip is then placed into the flask and pipetted up and down several times before discarding.

Cultures are incubated at 37°C overnight, shaking at 225rpm. A negative control is including containing a pipette tip with no bacteria, to confirm aseptic techniques. Cultures are checked the following morning for growth before glycerol stocks are taken. For glycerol stocks 500µl of overnight culture are mixed by gentle pipetting with 500µl 50% Glycerol in a 2ml screw top tube, these are then stored at -80°C for long term storage. Plasmids are extracted from the remaining LB cultures.

2.6.7. Purification of plasmid DNA

Following overnight culturing, the recombinant plasmids need to be extracted from the bacterial culture. The bacterial culture will contain bacterial chromosomal DNA as well as bacterial proteins and nucleic acids, as well as salt from LB, and the antibiotic. To isolate pure plasmid DNA, the overnight cultures were processed using either the GeneJET Plasmid Miniprep Kit (ThermoFisher Scientific, Massachusetts, USA) or the EndoFree Plasmid Maxi Kit (Qiagen, Venlo, Holland).

2.6.7.1. Miniprep using GeneJET Plasmid Miniprep Kit

For the 5ml overnight cultures the GeneJET Plasmid Miniprep Kit was used (ThermoFisher Scientific, Massachusetts, USA). The kit lyses the bacterial cells to

release plasmids, the cell debris is then removed by centrifugation and the supernatant is cleaned using a column based approach. Before starting, the Wash Solution was diluted using 35ml of 100% ethanol (ThermoFisher Scientific, Massachusetts, USA) and the provided RNase A solution was added to the Resuspension Solution. An overview of the protocol is shown in **Figure 2.12**.

Overnight cultures were collected and glycerol stocks taken as described in **2.6.6.** . The pipette tips used for inoculation are removed and the tubes are centrifuged for 10 minutes at 4,000g in a bench top centrifuge to pellet the bacterial cells. Once complete, the supernatant is discarded and the bacterial pellets are resuspended in 250µl Resuspension Solution and transferred to Eppendorf tubes. 250µl of Lysis Solution is added and the Eppendorf inverted until the solution becomes slightly clear, the reaction is incubated for four minutes at room temperature. 350µl of Neutralisation Solution is then added and the Eppendorf's are mixed vigorously to stop bacterial lysis. Samples are spun at 12,000g for five minutes to pellet cell debris, once complete the supernatant is transferred to a GeneJET spin column and the cell debris discarded.

The spin columns are centrifuged at 12,000g for one minute and the flow through discarded, the plasmid DNA is bound to the column. 500µl of wash solution is added to the column and the column is centrifuged at 12,000g for one minute, the flow through is discarded and this step is repeated. A final centrifugation of 12,000g for one minute allows for removal of any residual ethanol carry over. The spin column is then transferred to a clean Eppendorf and 20µl of elution buffer is added directly onto the column membrane. The column is then incubated at room temperature for two minutes before being centrifuged at 12,000g for two minutes to elute plasmid DNA. Eluted plasmid DNA is then quantified using Nanodrop as described in **2.3.4.1.** . Plasmid is then stored at -20°C for long term storage.

2.6.7.2. Maxiprep of plasmid DNA

For the 200ml overnight cultures the EndoFree Plasmid Maxi Kit (Qiagen, Venlo, Holland) was used. The kit lyses the bacterial cells to release plasmids, the cell debris is then removed by filtration through a QIAfilter. Endotoxins are then removed by the addition of the endotoxin removal buffer. The plasmids are then bound to a column and cleaned through washes. The plasmids are then eluted and precipitated with isopropanol and resuspended in endotoxin-free resuspension buffer. Before starting,

RNase A solution is added to Buffer P1, 40ml of 100% ethanol is added to endotoxin-free water, and buffer P3 is chilled to 4°C. An overview of the protocol is shown in **Figure 2.13**.

Overnight cultures were collected and glycerol stocks taken as described in **2.6.6.** . The 200ml overnight culture was then split into 4x 50ml falcons and spun at 4,000g for 10 minutes to pellet the bacteria, the supernatant was discarded. One pellet was resuspended in 10ml buffer P1 before moving to another falcon and re-suspending the next pellet, this continued until all four pellets were resuspended. 10ml buffer P2 was then added and samples inverted vigorously, the suspension turns blue. Samples were incubated at room temperature for five minutes. 10ml of chilled buffer P3 was then added to each tube and mixed by vigorous inversions, samples then turn white to indicate complete mixing, the precipitate contains genomic DNA, proteins and cell debris. Samples are then transferred to a QIAfilter cartridge by pouring, these are incubated at room temperature for 10 minutes. After 10 minutes the plunger is inserted and samples filtered into a fresh 50ml falcon, approx. 25ml of lysate is recovered.

2.5ml of buffer ER is added to the filtered lysate, samples are mixed and incubated on ice for 30 minutes. During the 30 minutes a QIAGEN-tip 500 is equilibrated by adding 10ml buffer QRT, the column empties by gravity flow. After 30 minutes the lysate is added to the equilibrated QIAGEN-tip 500 and allowed to flow through by gravity. QIAGEN-tips are then washed twice with 30ml of Buffer QC. Plasmid DNA is eluted with 15ml buffer QN into a fresh 50ml falcon tube. 10.5ml room-temperature isopropanol is added to the eluate to precipitate the DNA, samples are mixed and centrifuged at 4,700g for one hour. The supernatant is poured off and the plasmid DNA pellet is washed with 5ml of 70% ethanol, the pellet is then allowed to air dry for 30 minutes. 500µl of endotoxin-free Buffer TE is added to each pellet and plasmid DNA is allowed to resuspend overnight on the bench. Samples are then transferred to 1.5ml tubes (eppendorfs) the following morning and quantified using Nanodrop as described in **2.3.4.1.** . Plasmid is then stored at -20°C for long term storage.

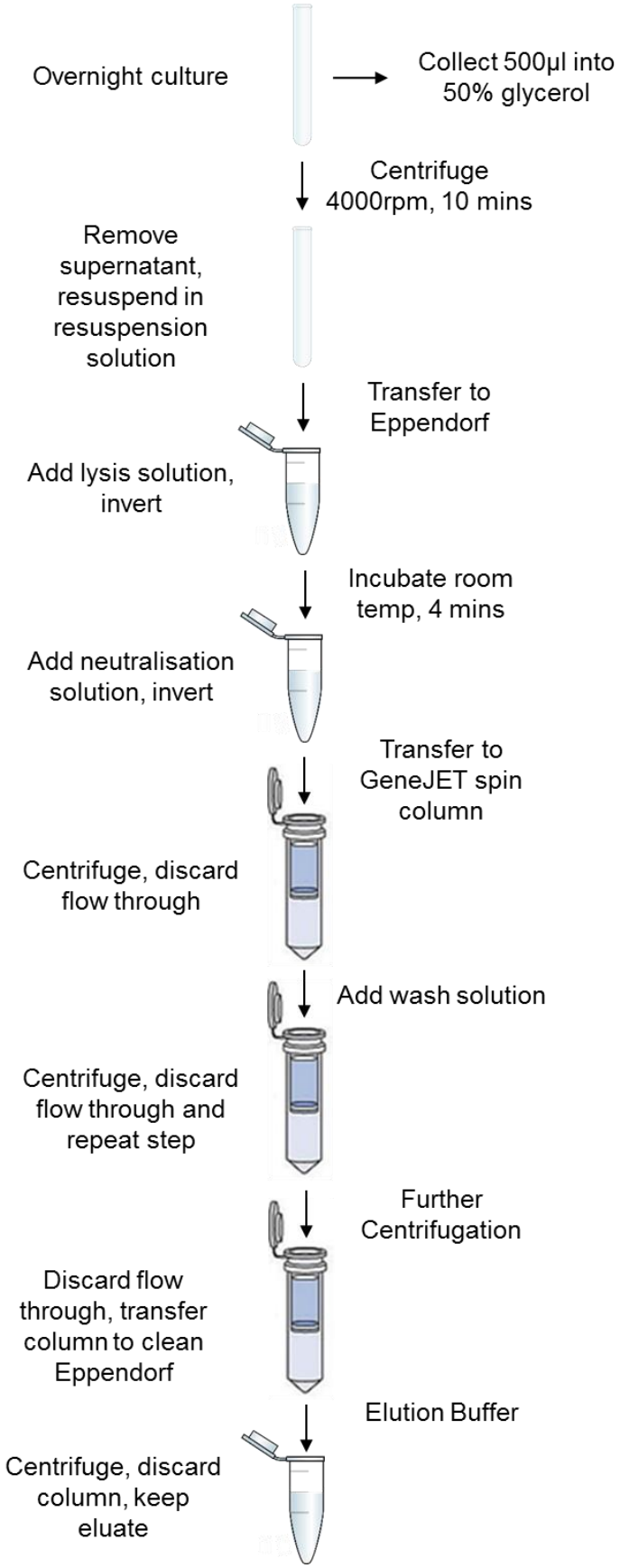


Figure 2.10: Overview of the GeneJET plasmid miniprep protocol.

Figure 2.11: Overview of the EndoFree plasmid maxi kit protocol. Image lifted from Qiagen supplied manual.

2.6.8. Diagnostic restriction digest

To confirm successful cloning of the insert into the backbone, mini/maxi prepped plasmids are digested by two or more restriction enzymes which give a distinct pattern to distinguish recombinants from background. A double digest restriction digestion is set up as listed in **Table 2.8** with purified plasmid and the starting backbone vector. Restriction enzymes used for diagnostic digests are described in subsequent chapters. Products are then visualised using agarose gel electrophoresis as outlined in **2.4.1.** . Gel percentages, voltages, and running times are described in subsequent chapters. An example of a diagnostic digest is provided in **Figure 2.14**.

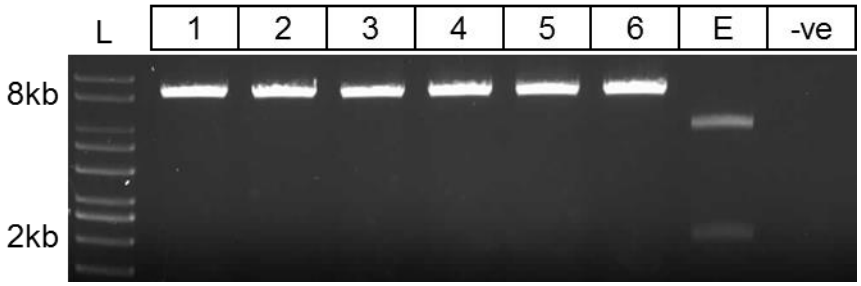


Figure 2.12: Example diagnostic digest results in distinct restriction patterns in correctly cloned plasmids compared to empty backbone vector. *Miniprepped plasmids 1-6 contain the insert as their restriction pattern is different to that of the empty backbone vector (E). The negative control (–ve) is water replacing the plasmid. L is DNA ladder (1kb DNA Ladder Ready to Load (Solis Biodyne)).*

2.6.9. Sanger sequencing of cloned plasmids

The final check of cloned plasmids is to sequencing the cloning site using Sanger sequencing. Sanger sequencing was first developed by Fredrick Sanger in 1977 and is based on chain-termination using fluorescently tagged dideoxynucleotides (Sanger and Nicklen, 1977). These dideoxynucleotides result in chain termination and prevent elongation, once integrated into the sugar phosphate backbone they release their fluorochrome which corresponds to the base added. Capillary electrophoresis is used to separate out the different sized products and the fluorescent readout is recorded in a chromatogram, resulting in the DNA sequence.

Plasmids are sent to an external company (Genewiz ltd, Leipzig, Germany) where Sanger sequencing is undertaken. As per their protocol, 800ng of plasmid DNA is sent with 25µM of sequencing primer in a 10µl of water in a PCR tube. Once the results are received, the sequencing is then searched for the insert of interest. Plasmids which do not contain the insert are destroyed. The sequencing primers used are listed in subsequent chapters of this thesis. An example chromatogram showing a gRNA insert is shown in **Figure 2.14**.

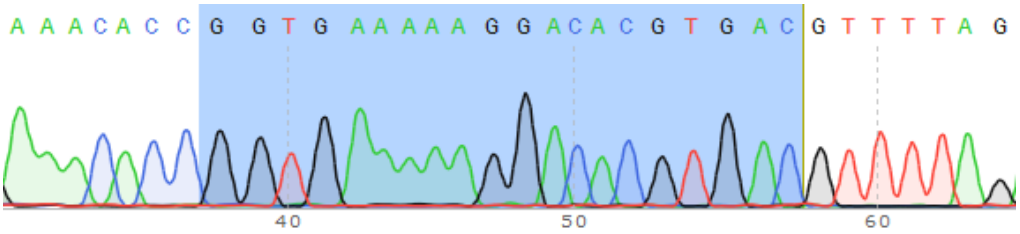


Figure 2.13: Chromatogram of a cloned region following Sanger sequencing. *Each colour peak corresponds to the incorporated dideoxynucleotide, with green representing adenine, blue representing cytosine, black representing guanine, and red representing thymine. The successfully cloned gRNA insert is highlighted in blue and is found by searching the sequence for the insert.*

2.7. DNA methylation analysis

2.7.1. DNA bisulphite conversion (BC)

DNA BC treatment is a method which provides base-pair level sensitivity to measure DNA methylation and was originally described by Frommer et al (Frommer *et al.*, 1992). In this process, un-modified cytosine residues are converted to uracil in the presence of sodium bisulphite. However modified cytosine residues, including 5mC, are unreactive and do not convert. This treated DNA sample can then be amplified by PCR, where the uracil residue is converted into a thymine, and the modified cytosine residues are converted to cytosine. This then allows us to compare sequences and identify which cytosine's are 5mC and which are unmodified. For this thesis, BC was carried out using the Zymo EZ-96 DNA Methylation-Gold™ Kit (Cambridge Bioscience, Cambridge, UK). This method has been used for treatment of DNA samples studies within our lab (Fisher *et al.*, 2015; Jager *et al.*, 2015; Hannon, Spiers, *et al.*, 2016; Murphy *et al.*, 2017; Hannon *et al.*, 2018). Before using the kit 144ml of ethanol is added to the M-Wash Buffer provided.

Briefly, 500ng (25ng/µl in 20µl) of DNA was added to each well of a 96 well PCR plate. 130µl of the CT conversion reagent (as prepared in **Table 2.12**) was added to each sample and mixed. The plate was then sealed and incubated in thermocycler for the cycles outlined in **Table 2.13**. 400µl of M-Binding buffer was added to the wells of a Silicon-A™ binding plate mounted on a collection plate. Following thermocycling, samples were transferred from the PCR plate to the corresponding wells of the Silicon-A™ binding plate and mixed by pipetting. The plate was centrifuged at 3,500g for five minutes then the flow through was discarded. 400µl of M-Wash buffer was added to each well and the plate centrifuged at 3,500g for five minutes, the flow through was discarded. Another 400µl of M-Wash buffer was added to each well and the plate centrifuged at 3,500g for 10 minutes. The Silicon-A™ binding plate was transferred onto an elution plate. 15µl of M-Elution buffer was added directly to each well of Silicon-A™ binding plate and the plate incubated for five minutes at room temperature. The plate was then centrifuged at 4,000g for 3 minutes to elute the DNA. Another 15µl of M-Elution buffer was added to the Silicon-A™ binding plate and the above steps repeated.

The eluted DNA was used for immediate analysis using the Infinium HumanMethylationEPIC BeadChip Array or was stored at -20°C for future use.

Table 2.12: CT Conversion Reagent required for BC using the Zymo EZ-96 DNA Methylation-Gold Kit. *The CT Conversion Reagent is supplied as a powder and needs to be dissolved prior to use. Solutions are mixed with frequent vortexing or shaking for 15 minutes.*

Name	Volume (ml)
M-Dissolving Buffer	0.5
M-Dilution Buffer	3
CT Conversion Reagent	Bottle Provided
Water	9

Table 2.13: Thermocycler conditions for BC.

Step	Temperature (°C)	Time	Cycles
Denaturation	98	10 min	1
Conversion	64	2.5 hrs	1
Storage	4	∞ (up to 20 hrs)	1

2.7.2. Infinium HumanMethylationEPIC BeadChip Array

The EPIC array combines BC DNA and whole-genome amplification with direct, assay-based capture to score the methylation state at individual CpG Loci (Pidsley *et al.*, 2016). The EPIC array allows for quantification of 866,836 CpG sites, these sites include over 90% of the sites on the previous HumanMethylation450K plus more than 350,000 CpGs at potential enhancers identified by the FANTOM5 and ENCODE projects (Siggens and Ekwall, 2014; Lizio *et al.*, 2015). The HumanMethylationEPIC array covers important regulatory regions such as CpG Islands, island shores and shelves, 5' and 3' UTRs, gene promoters and also gene bodies, which therefore provides genome wide DNA methylation status. Signal intensity is measured using the Illumina iScan system to generate beta values, a measure of methylation status, at each locus. One or two probes are used to infer the methylation state of a single CpG loci, dependent on the probe design for the particular CpG, the Infinium I design has two probes per site and Infinium II has one probe per site. Further to this, the 3' end of the probe is positioned directly across from the CpG site for Infinium I or immediately adjacent for Infinium II. The type I probe design assumes that the methylation state of the CpG of interest is directly correlated to the methylation state of CpG sites within 50bp of the CpG of interest. This therefore can lead to biases where the probes are not associated with nearby methylation status. However the type II probes use a single bead type with the methylation state determined by a single base extension step following hybridisation with two different coloured dyes representing methylation state (green for methylated and red for unmethylated). This works by the inclusion of a degenerate A/G base at underlying CpG sites and therefore is less influenced by nearby methylation patterns. The coloured signal is amplified and this improves the signal to noise ratio (Illumina, 2015).

A full description of the methodology can be found on the Illumina website (https://emea.illumina.com/products/by-type/microarray-kits/infinium-methylation-epic.html). Briefly, the BC DNA is denatured and neutralised to prepare for amplification. The denatured DNA is isothermally amplified in an overnight step, this allows for uniform whole genome amplification without significant amplification biases. A controlled enzymatic process fragments the amplified DNA, this uses endpoint fragmentation to prevent over fragmentation. The fragmented DNA is then precipitated using isopropanol and collected by centrifugation at 4°C. The precipitated DNA is then

resuspended in Hybridisation buffer. The resuspended DNA fragments are then dispensed onto BeadChips, these are incubated in the Illumina Hybridisation Oven to allow for hybridisation of DNA onto the BeadChips. 12 samples are then applied to each BeadChip, these samples are kept separate by the IntelliHyb seal, these are then incubated overnight. The amplified and fragmented DNA samples anneal to a locus-specific site on the BeadChip array during the hybridisation process. Following hybridisation unhybridised and nonspecific hybridised DNA is washed away and the BeadChip undergoes extension and staining in capillary flow-through chambers. Single-base extension of the oligos on the BeadChip using the captured DNA as a template incorporates the detectable labels (green for methylated, red for unmethylated) to determine methylation state of the CpG site. Finally the BeadChips are scanned on the Illumina iScan System which uses a laser to excite the fluorophore, the scanner then records a high resolution image of the light emitted. These were the imported into the statistical software R and the wateRmelon package was used to generate Beta values of methylation (Pidsley *et al.*, 2013).

2.7.2.1. Data normalisation and quality control

The DNA modification status at each CpG site was determined by calculating the ratio of the intensity of the fluorescent signal for M (methylated/green) and U (unmethylated/red). This results in a β value for each site ranging from 0 (i.e all cytosines unmethylated) to 1 (i.e all cytosines are methylated). The β value is calculated by the following equation:

$$\beta = \frac{Intensity\ M}{Intensity\ M + Intensity\ U + 100}$$

Data was normalised using the watermelon package in R, which has been developed by our group. This package offers a range of normalisation methods and tools to pre-process and normalise the DNA modification data presented in this thesis (Pidsley *et al.*, 2013).

The presence of SNP variation within close proximity of the CpG site is an issue with the methylation array, therefore the methylation value recorded could simply represent the polymorphism and not the methylation state (Chen *et al.*, 2013; Price *et al.*, 2013). Further to this, approximately 5% of probes on the array have been found to cross-hybridise with other genomic locations and do not accurately estimate the methylation levels (Pidsley *et al.*, 2016). As a result probes known to either cross-hybridise with

another site or contain SNP variation were excluded prior to further analysis. The array also contains 59 probes targeting SNPs which are used for genotyping and sample identification. These probes were removed before data analysis.

Chapter 3. Characterisation and functional analysis of the schizophrenia associated gene *AS3MT*

3.1. Introduction

Selecting candidate or causal genes for functional analysis in polygenic diseases is complex as many genes contribute to the disease. For schizophrenia over 108 risk loci have been identified through GWAS, with the total number of candidate genes exceeding 300, and this list is continually expanding (Schizophrenia Working Group of the Psychiatric Genomics Consortium, 2014; Pardiñas *et al.*, 2018). One of the added problems of GWAS studies is that many of the identified risk SNP variants are within LD of each other, making identification of the causal SNP incredibly complex. However we have identified that many of the SNPs identified through GWAS are correlated with foetal development providing evidence for the developmental hypothesis of schizophrenia (Duarte *et al.*, 2016; Ma *et al.*, 2018). Following on from this, a recent EWAS on foetal brain samples through different stages of development identified enrichment of mQTLs in 10 of the aforementioned 108 risk loci from the schizophrenia GWAS (Hannon, Spiers, *et al.*, 2016). Even this reduction still results in a variety of different genes, however this has refined the data down to genes which could be involved in brain development and schizophrenia. In this chapter I set out to identify the neuronal function of one of these genes, *AS3MT,* in neuronal cell culture by generating a knockout cell lines using CRISPR/Cas9.

Arsenite Methyltransferase (*AS3MT)* is located at 10q24 region on Chromosome 10, this is the most statistically significant GWAS locus outside the MHC region associated with schizophrenia and other neuropsychiatric disorders (The Schizophrenia Psychiatric Genome-Wide Association Study Consortium, 2012; Duarte *et al.*, 2016; The Autism Spectrum Disorders Working Group of The Psychiatric Genomics Consortium, 2017). *AS3MT* is a methyltransferase which primarily methylates and detoxifies arsenic using S-adenosyl-L-methionine (SAM) as a methyl donor to form intermediate compounds of monomethylated and dimethylated forms of arsenate and arsenite (Thomas *et al.*, 2007). *AS3MT* comprises of 11 exons in length, of which all encode for a full functional transcript, henceforth known as $AS3MT^{full}$, there is an alternatively spliced transcript of *AS3MT* with a deletion of exon 2/3 henceforth known as $AS3MT^{d2/3}$. The main protein domain, the methyltransferase, is located within exons 4 to 6 and is present in both $AS3MT^{full}$ and $AS3MT^{d2/3}$ (**Figure 3.2**). *AS3MT* is expressed in neuronal tissue as well as somatic tissue, however looking at the splice variants the $AS3MT^{full}$ transcript shows no difference in expression between somatic

and neuronal whereas the *AS3MT*$^{d2/3}$ is expressed 2 fold higher in brain tissue than somatic tissue (M. Li *et al.*, 2016). Interestingly the ratio of the two transcripts is highest in the hippocampus, the region which is implicated prominently in psychiatric diseases (Harrison, 2004). The isoforms also have differential expression throughout brain development, *AS3MT*^{full} expressed higher prenatally compared to postnatally, whereas the *AS3MT*$^{d2/3}$ expression does not change, further supporting that *AS3MT* splice variation has a role in neurodevelopment (M. Li *et al.*, 2016). Further evidence supports the role of *AS3MT* splice ratio in neuropsychiatric disorders, the GWAS schizophrenia risk SNP rs7085104 has been identified to increase the expression of the *AS3MT*$^{d2/3}$ isoform with no effect on *AS3MT*^{full} in multiple brain tissues (M. Li *et al.*, 2016).

The 10q24 region has been highly characterised with schizophrenia and is highly replicable with multiple GWAS studies (Ripke *et al.*, 2013; Schizophrenia Working Group of the Psychiatric Genomics Consortium, 2014; Pardiñas *et al.*, 2018). The schizophrenia risk SNPs rs11191419 shows increased expression of *AS3MT* and *BORCS7* in foetal and adult brain (Duarte *et al.*, 2016) and rs7085104 with increased differential isoform expression of *AS3MT* (M. Li *et al.*, 2016). There is strong evidence for co-localisation of schizophrenia GWAS and EWAS at *AS3MT* in adult brain and blood, where the risk SNP rs11191419 could be acting to change gene expression through methylation (**Figure 3.1**) (Hannon, Dempster, *et al.*, 2016). These methylation signals are also prevalent in the developing foetus, identifying a potential role of *AS3MT* in development (Hannon, Spiers, *et al.*, 2016). *AS3MT* has also appeared more recently in transcriptome wide association studies of schizophrenia (Gusev *et al.*, 2018), a GWAS of Alzheimer's (Demichele-sweet *et al.*, 2018) and shows altered expression in patients with depression (M. Li *et al.*, 2016). A recent meta-analysis looking at the 10q24.32-33 region (containing rs7085104) identified a strong association with major affective disorders, further strengthening the link between *AS3MT* isoforms and neuropsychiatry (L. Li *et al.*, 2016). Further functional work has recently shown that the schizophrenic risk SNP rs7085104 located within *AS3MT* has a significant impact on dopamine synthesis pathways identified through positron emission tomography scans (Ambrosio *et al.*, 2019).

While the role of *AS3MT* in arsenic metabolism has been extensively studied the individual contributions of the splice variants has been less studied. The *AS3MT*^{full}

transcript possesses arsenic methylation ability, as shown through previous work, however the *AS3MT*$^{d2/3}$ transcript does not methylate arsenic indicating an alternative function (M. Li *et al.*, 2016). This combined with increased expression in the brain, through development, and its upregulation through differentiation of human pluripotent stem cells towards dorsal telencephalic fates, identifies a potential role in neuronal processing and function. To elucidate the role of *AS3MT* in neuronal cell function a knockout cell line of *AS3MT* was generated in the dopaminergic SH-SY5Y cell line using CRISPR/Cas9. We then used this line to study morphological, molecular, and biological changes as a result of *AS3MT* knockout and how these link to neuronal development and schizophrenia.

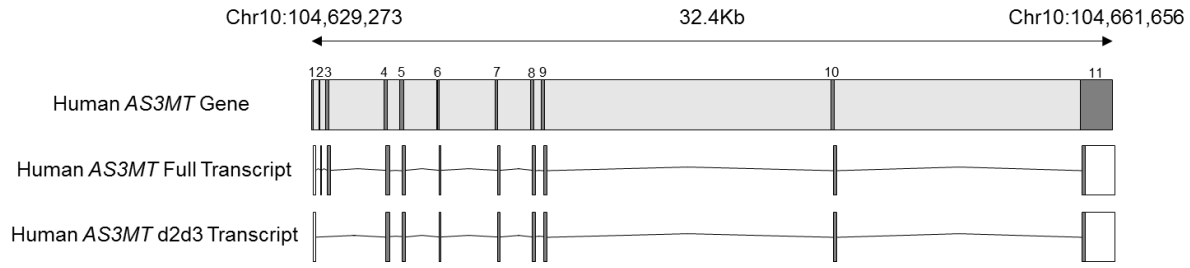


Figure 3.2: The human AS3MT gene and the alternatively spliced isoforms and their structure. *Introns are shown in light grey with exons shown in dark grey, the white boxes are untranslated regions.*

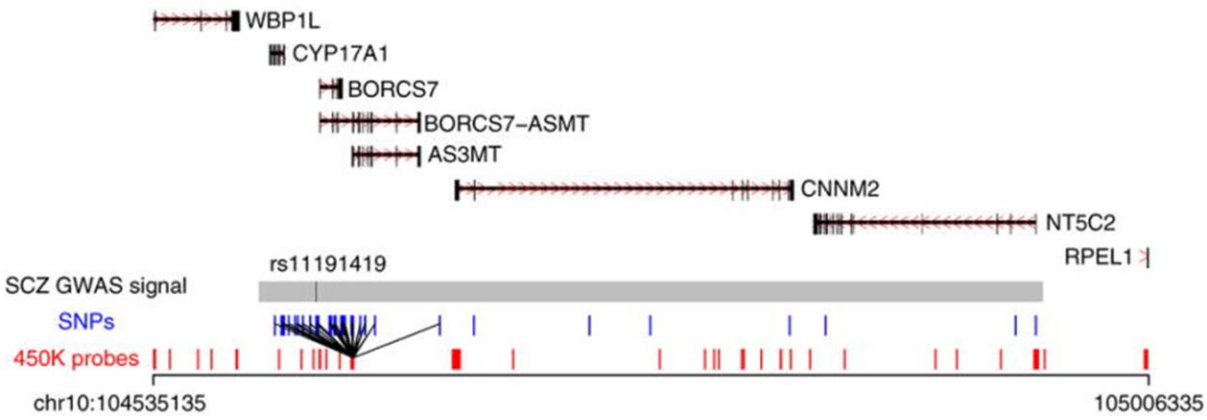


Figure 3.1: Identification of genetic variants associated with both schizophrenia and DNA methylation using brain mQTL data. *Known SNPs are shown in blue with 450K methylation probes shown in red. A line connecting the SNP to the 450K probes indicates an mQTL, where the SNP is having a direct effect on DNA methylation at that 450K site. In the above image, lots of SNPs are affecting several 450K probes, this is due to LD and our inability to identify the causal variant. One of these mQTL SNPs, rs11191419, has been identified as a risk SNP in the schizophrenia GWAS. The effect of genetics on DNA methylation is explained in more detail in Chapter 6. Image from (Hannon, Spiers et al 2016).*

3.2. Aims

The main aim of this chapter was to develop a pipeline to functionally validate neuropsychiatric associated genes in SH-SY5Y cells. I then set out to apply this to the schizophrenia candidate gene *AS3MT* and elucidate its role, if any, in neuronal development. The specific aims of this chapter were:

1. To generate a protocol to edit the genome using CRISPR/Cas9 in SH-SY5Y cells.
2. Confirm differential expression of *AS3MT* through differentiation of SH-SY5Y cells towards neuronal fates.
3. Knock out *AS3MT* within SH-SY5Y cells and generate homozygous cell lines for functional analysis.
4. Analyse changes in neuronal morphology of *AS3MT* knockout on differentiated SH-SY5Y neurons through MAP2 immunocytochemistry.

3.3. Methods

3.3.1. Gene selection

I selected a gene for functional validation based on the analysis from (Hannon, Spiers, *et al.*, 2016) which used DNA methylation data to identify genes associated with schizophrenia that are highly expressed in foetal brain development. I then further refined the gene list by cross-referencing the genes to another schizophrenia EWAS study in adult blood and brain samples which reduced the gene list from 10 loci to nine individual genes (Hannon, Dempster, *et al.*, 2016). In order to refine the nine genes to one for functional analysis the CommonMind Consortium was consulted for known eQTLs for these nine genes (Fromer, Roussos, Solveig K Sieberts, *et al.*, 2016). Genes with known eQTLs were considered for functional analysis. Gene ontology for each of the genes was then established from the panther database to identify gene function (Mi *et al.*, 2017). Gene expression was then explored using the EMBL-EBI Expression Atlas and genes with known neuronal and foetal expression were selected (Papatheodorou *et al.*, 2018). Phenotypes showing behavioural or neurological changes in humans and mice where the genes have been knocked out or knocked down were extracted from Online Mendelian inheritance in Man for disease association for humans, and MGI for mouse (Hamosh *et al.*, 2002; Eppig and Eppig, 2017). Along with this known orthologue identification in zebrafish was included to allow ease of follow up functional work in an animal model routinely used for development (Morris, 2009; Schmidt, Strähle and Scholpp, 2013). Finally a literature search was undertaken and of the nine genes identified in (Hannon, Dempster, *et al.*, 2016) there was an enrichment in papers linking *AS3MT* and schizophrenia (Duarte *et al.*, 2016; L. Li *et al.*, 2016; M. Li *et al.*, 2016).

3.3.2. CRISPR/Cas9 knockout of *AS3MT*

3.3.2.1. Guide RNA design

To maximise the efficiency and to knockout both transcripts of *AS3MT* two gRNA were designed, one to target exon 4 and the other to target exon 6 with a targeted deletion of 2kb expected. These gRNA will also knock out the main methyltransferase protein domain in *AS3MT* resulting in full functional knockout. The guide sequences were selected using the online CRISPR design tool available at http://crispr.mit.edu/ from the Zhang lab (Ran *et al.*, 2013). Guides with high on target score and low off target

score were selected to minimise the chance of off target effects and are listed in **Table 3.1**. For cloning the complementary sequence of the gRNA was calculated and sticky ends added to the 5' ends of the forward and complementary sequences to allow insertion at the BbsI cloning site. A CACC sticky end was added to the forward gRNA and an AAAC added to the reverse gRNA as outlined in **Table 3.2**. gRNA were then ordered from Integrated DNA Technologies (https://www.idtdna.com/).

3.3.2.2. Cloning of gRNA into CRISPR vectors

3.3.2.2.1. CRISPR vectors

Two commercially available genomic CRISPR plasmids (pSpCas9(BB)-2A-GFP (PX458) and pU6-(BbsI)_CBh-Cas9-T2A-mCherry) were purchased from the plasmid depository service Addgene (www.addgene.org). These plasmids express gRNA, *Streptococcus pyogenes* Cas9 and EGFP or mCherry under control of the mammalian U6 promotor. The fluorescence markers are used for positive selection of transfected cells. Each plasmid also contains resistance to ampicillin for use in cloning and are first described in (Ran *et al.*, 2013; Chu *et al.*, 2015).

3.3.2.2.2. Oligo annealing

gRNA was purchased as standard desalted pelleted single stranded DNA (ssDNA) oligos from Integrated DNA Technologies, these were reconstituted to a concentration of 100µM using ddH_2O. The ssDNA gRNA oligo and its reverse complement were annealed together to form a dsDNA insert containing 5' sticky ends to enable cloning into CRISPR vectors. Furthermore the 5' ends were phosphorylated using polynucleotide 5'-hydroxyl-kinase (PNK) to allow for ligation of the sugar phosphate backbone. The reactions were set up as listed in **Table 3.3** and run in a thermocycler as listed in **Table 3.4**. The annealed gRNA were then diluted down 1 in 200 with ddH_2O before being ligated into the CRISPR vectors.

Table 3.1: gRNA used in the generation of the AS3MT knockout cell line. PAM sequences are identified in red.

gRNA	Sequence (5' – 3')	Chromosomal Location	Strand
AS3MT_Exon_4	GGTGAAAAAGGACACGTGACtgg	Chr10:102872554-102872576	+
AS3MT_Exon_6	GGAGGCATATCGGGTGCTGAagg	Chr10: 102874640-102874662	+

Table 3.2: gRNA modifications for cloning.

gRNA	Step	Sequence (5' – 3')
AS3MT_Exon_4	Raw gRNA	GGTGAAAAAGGACACGTGAC
	Reverse complement	GTCACGTGTCCTTTTTCACC
	Addition of sticky ends	CACCGGTGAAAAAGGACACGTGAC AAACGTCACGTGTCCTTTTTCACC
	Final insert	CACCGGTGAAAAAGGACACGTGAC CCACTTTTTCCTGTGCACTGCAAA
AS3MT_Exon_6	Raw gRNA	GGAGGCATATCGGGTGCTGA
	Reverse complement	TCAGCACCCGATATGCCTCC
	Addition of sticky ends	CACCGGAGGCATATCGGGTGCTGA AAACTCAGCACCCGATATGCCTCC
	Final insert	CACCGGAGGCATATCGGGTGCTGA CCTCCGTATAGCCCACGACTCAAA

Table 3.3: Annealing and phosphorylation of the gRNA to its complement for cloning.

Reagent	Volume (µl)	Final Concentration
gRNA (100µM)	1	10µM
Complementary gRNA (100µM)	1	10µM
10x T4 Ligation Buffer	1	1x
T4 PNK	1	10U
ddH_2O	6	
Total	**10**	

Table 3.4: Thermocycling conditions to anneal gRNA oligos.

Temperature (°C)	Time	Cycles
37	30 min	1
90	5 min	1
Ramp down to 25 at 5°C per min		1
25	∞	1

3.3.2.2.3. Single digestion and ligation reaction

This was carried out in a single digestion ligation reaction as described in (Ran *et al.*, 2013), where the backbone is digested and gRNA ligated into the backbone in a single reaction. The reaction is listed in **Table 3.5**, the reaction was incubated for six cycles of 37°C for 5 minutes followed by 21°C for 5 minutes.

3.3.2.2.4. *Transformation of recombinants into DH5α E. coli*

Transformation of the ligation reactions from **3.3.2.2.3.** into DH5α *E. coli* was carried out as described in **2.6.4.** . Colony counts are listed in **Table 3.6**

3.3.2.2.5. Overnight bacterial cultures

The following afternoon three clones from **3.3.2.2.4.** were selected at random from each plate and taken into 5ml overnight cultures as described in **2.6.6.**

3.3.2.2.6. Purification of plasmid DNA

The following morning plasmid DNA was purified using the GeneJET Plasmid Miniprep Kit (ThermoFisher Scientific) as described in **2.6.7.1.**

3.3.2.2.7. Diagnostic digest

A diagnostic digest was then performed on 100ng of each miniprepped plasmids as described in **2.6.8.** to confirm if gRNA has been correctly inserted into the plasmid by using the restriction enzymes BbsI and EcoRI. All miniprepped plasmids contained the gRNA insert (**Figure 3.3**).

3.3.2.2.8. Sanger sequencing of cloned plasmids

Sequencing was carried out as described in **2.6.9.** . Positive plasmids from the double diagnostic digest were sent off for Sanger sequencing (Genewiz, UK LTD) using the LKO.1 5' sequencing primer which is in the U6 promoter of the constructs. Sequence data was analysed in Jalview (https://www.jalview.org/) and confirmed the insertion of the gRNA in the correct orientation into the all of the miniprepped plasmids (**Figure 3.4**). Recombinant plasmids will now be referred to as indicated in **Table 3.7**.

Table 3.5: Single digestion ligation reaction used to clone gRNA into pSpCas9(BB)-2A-EGFP and pU6-(BbsI)_CBh-Cas9-T2A-mCherry.

Reagent	Volume (µl)
pSpCas9(BB)-2A-EGFP / pU6-(BbsI)_CBh-Cas9-T2A-mCherry	100ng
Diluted gRNA	2
10x Fast Digest Buffer	2
10mM DTT	1
10mM ATP	1
BbsI	1
T4 Ligase	0.5
ddH_2O	To 20
Total	**20**

Table 3.6: Colony counts following spread plating of transformed recombinants.

Plate	Colony number
AS3MT_Exon_4_mCherry	113
AS3MT_Exon_4_EGFP	76
AS3MT_Exon_6_mCherry	133
AS3MT_Exon_6_EGFP	104
pU6-(BbsI)_CBh-Cas9-T2A-mCherry (- ligation control)	142
pSpCas9(BB)-2A-GFP (PX458) (- ligation control)	44
pUC19 (transformation control)	>500
Empty plate (negative control)	0

Table 3.7: Nomenclature of constructed CRISPR plasmids for AS3MT knockout in this thesis.

gRNA	Backbone	Henceforth
Exon 4	pSpCas9(BB)-2A-EGFP	Exon 4 GFP
Exon 4	pU6_(BbsI)_CBhCas9-T2A-mCherry	Exon 4 mCh
Exon 6	pSpCas9(BB)-2A-EGFP	Exon 6 GFP
Exon 6	pU6_(BbsI)_CBhCas9-T2A-mCherry	Exon 4 mCh

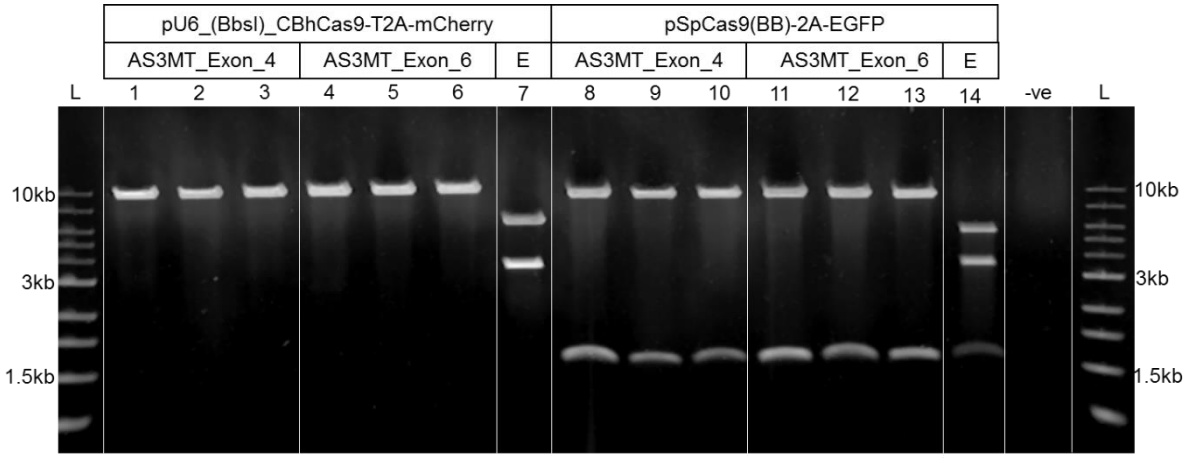


Figure 3.3: Double diagnostic digest to confirm correct insertion of gRNA into plasmid backbone in miniprepped plasmids following overnight inoculation into liquid broth*. Lane L, 1kb ladder. Lanes 1-7 are the pU6_(BbsI)_CBhCas9-T2A-mCherry backbone plasmid, with AS3MT_Exon_4 colonies in lanes 1-3 and AS3MT_Exon_6 colonies in lanes 4-6, lane 7 is a pU6_(BbsI)_CBhCas9-T2A-mCherry plasmid with no gRNA inserted as a digestion control. Lanes 8-14 are the same layout as lanes 1-7 only with pSpCas9(BB)-2A-EGFP as the backbone plasmid. Lane 14 is a digestion negative control with plasmid replaced with water. E is the empty backbone vector with no gRNA cloned insert. –ve is plasmid replaced with water.*

Figure 3.4: Sanger sequencing confirmed cloning of gRNA into CRISPR vectors. *Sequences analysed in Jalview and coloured by percentage base identity with blue indicating 100% identity and white 0% identity.*

3.3.2.3. Nucleofection of CRISPR plasmids into SH-SY5Y cells

Passage 10 SH-SY5Y cells were seeded in T175 flasks a week before nucleofection and allowed to reach 80% confluency. Cells were then collected and counted using a haemocytometer. 2µg total CRISPR plasmid were transfected into $1x10^6$ cells via nucleofection following the manufacturers protocol (Lonza). Six nucleofection reactions were carried out in total as listed in **Table 3.8**. Following nucleofection, cells were incubated at 37°C and 5% CO_2 for 24 hours. Cells were imaged for fluorescence using a Lecia DMi8 inverted widefield microscope (Leica) after 24 hours to confirm nucleofection efficiency (**Figure 3.5**). After confirmation of transfection success, GFP and mCherry double positive single cells were selected by fluorescent activated cell sorting (FACS) (see section **2.2.**).

Table 3.8: The six nucleofection reactions of CRISPR plasmids into SH-SY5Y cells undertaken. pMax GFP was used as a nucleofection positive control, programme and substance negative were nucleofection negative controls.

Nucleofection	Plasmid	Mass of plasmid	Total plasmid
Exon 4	Exon 4 GFP	2µg	2µg
Exon 6	Exon 6 mCh	2µg	2µg
Exon 4 + Exon 6	Exon 4 GFP	1µg	2µg
	Exon 6 mCh	1µg	
pMax GFP	pMax GFP	1µg	1µg
Programme Negative	pMax GFP	1µg	1µg
Substance Negative	-	-	-

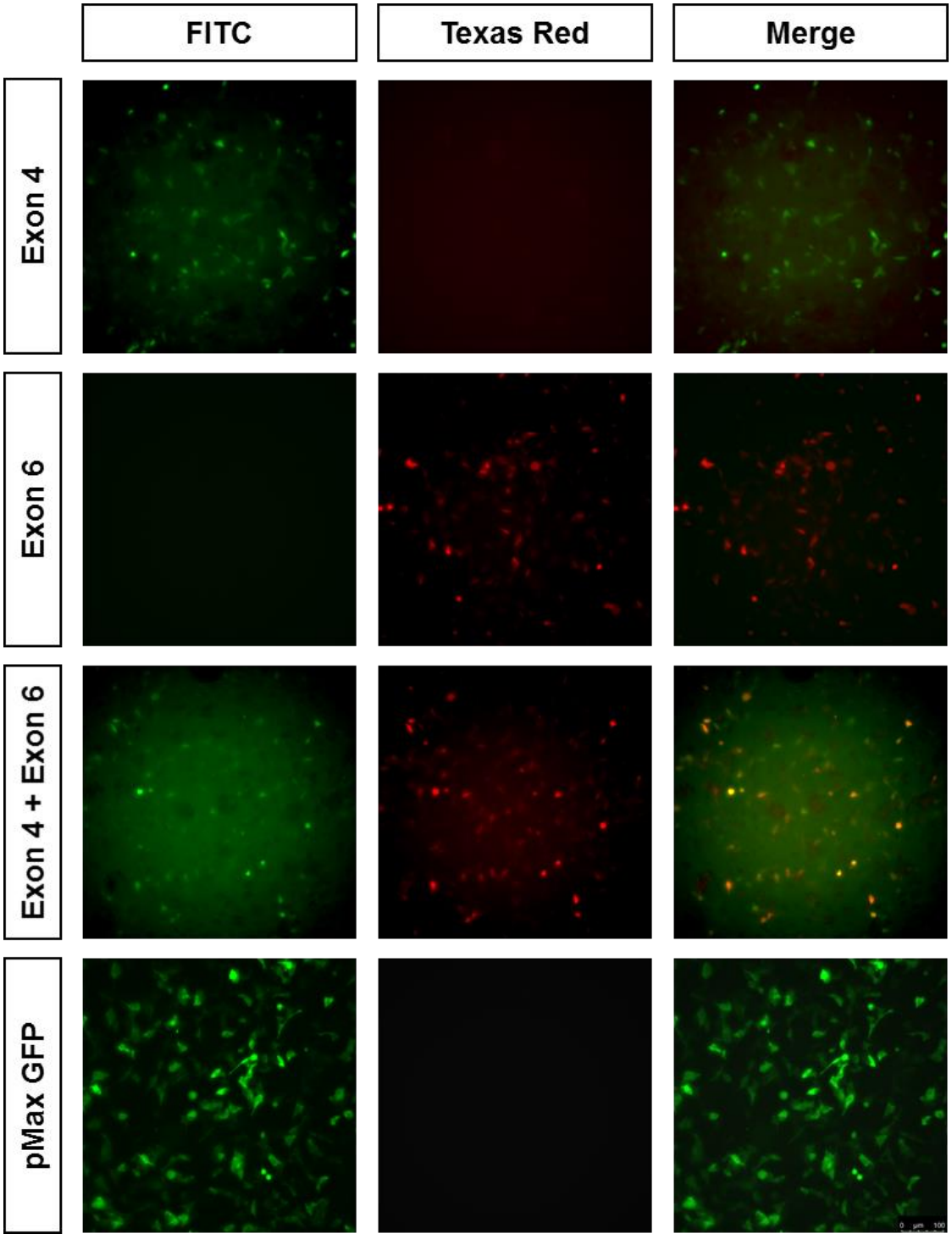


Figure 3.5: Fluorescent microscope images 24 hours following CRISPR nucleofection indicates cell which have up taken the CRISPR/Cas9 constructs. *In the single nucleofected Exon 4 GFP we see GFP fluorescence in the FITC channel only and no mCherry expression. With the single nucleofected Exon 6 mCherry we see mCherry expression only in the Texas Red channel. Following double nucleofection with both Exon 4 GFP and Exon 6 mCherry there are a proportion of cells which show double positive GFP and mCherry selection, these cells were selected for using FACS. The pMax GFP is a nucleofection positive control. Images taken on the DMi8 inverted widefield microscope (Leica) at 20x magnification. Scale bar 100µm.*

3.3.2.4. Fluorescent activated cell sorting

Once transfection was confirmed through microscopy cells were sorted by FACS. Cells were harvested from 6 well plates as described in **2.1.3.** and resuspended in FACS sorting medium (**Table 3.9**). FACS gates were set up to select for single double positive cells. The gates were selected by sorting populations negative control cells (no fluorescence) followed by Exon 4 GFP transfected cells for FITC gating and finally Exon 6 mCh for PE-CF594 gating. This allowed stringent gates to be set up reducing the risk of false positives. Single cells were sorted into 96 well plates containing growth media supplemented with antibiotics. Five 96 well plates of single cells were collected:

- One plate of substance negative controls (Wild Type Cells)
- One plate of single FITC positive (containing Exon 4 only)
- One plate of single PE-CF594 positive (containing Exon 6 only)
- Two plates double positive FITC and PE-CF594 (Exon 4 and Exon 6)

An Eppendorf of double positive FITC and PE-CF594 cells from the double transfection of Exon 4 and 6 was collected, these cells were then rerun in the FACS to confirm the selection and are henceforth known as enriched, this reduces the likelihood of a false positive cell being sorted. A plate of enriched Exon 4 and Exon 6 double positive cells were collected. Along with sorting single cells into individual wells a well of each plate was filled with 100 cells, this is to reduce clonal expansion time and allow genotyping to occur earlier to confirm the experiment has been successful. Cells were then returned to the incubator and allowed to expand by clonal expansion.

Table 3.9: FACS Sorting Media Composition.

Reagent	Volume (ml)	Concentration
DMEM/F-12 without phenol red	445	N/A
Foetal Bovine Serum	50	10%
Penicillin/Streptomycin (10,000 U/mL)	5	100 U/mL

3.3.2.5. Clonal expansion

Following sorting, cells were cultured for several weeks to expand numbers before genotyping. Plates were imaged using a Lecia DMi8 inverted wide field microscope (Leica) after seven days to confirm survival efficiency. Two weeks after FACS DNA was extracted from the 96 well seeded with 100 cells as described in **2.3.** to carry out genotyping and confirm CRISPR mediated disruption.

After four weeks of clonal expansion in 96 well plates cells were passaged into 24 well plates as described in **2.1.3.** the following week DNA was extracted from the 24 well plates for genotyping as described in **2.3.** . Cells which showed a knockout or heterozygous genotype were frozen down as described in **2.1.5.**

3.3.2.6. Genotyping

Genotyping was carried out using a PCR based approach. Two sets of PCR primers were designed to distinguish between heterozygous and homozygous *AS3MT* knockout lines. One set of primers amplified DNA within the two CRISPR gRNA (internal) and one set of primers amplified DNA outside of the deletion region (external), the primer sequences are listed in **Table 3.10**.

In a wild type sample the internal PCR product is present and the external PCR product is absent due to the large size, whereas in a homozygous knockout allele Exons 4 and 7 become closer together due to NHEJ, this then allows amplification of the external PCR product, these samples also lack the internal PCR product which has been deleted. In a heterozygous knockout both the internal and external PCR bands are visible, this allows for ease of detection of homozygous and heterozygous mutants. A visual representation is shown in **Figure 3.6**.

100ng of extracted DNA was run in a standard PCR reaction as described in **2.4.** .The thermal cycling conditions are listed in **Table 3.11** and **Table 3.12**. 2µl of each PCR product was loaded into a 0.5% agarose gel containing cyto60 and run for 120v for 45 minutes and imaged using the Licor Odyssey imaging system as described in **2.4.1.** (**Figure 3.11**).

External PCR products of samples identified as heterozygous or homozygous knockouts of *AS3MT* were cleaned using ExoSap. ExoSap inactivates unincorporated nucleotides and primers (as listed in **Table 3.13**) by incubating at 37°C for 1 hour. Cleaned PCR samples were quantified using the nanodrop spectrophotometer and

1µg of PCR products were sent for Sanger sequencing with the External Forward PCR Primer to confirm the genomic sequence of the knockout (Genewiz, Tackley, UK). Cell lines which were homozygous or heterozygous for *AS3MT* knockout along with three wildtypes (Substance negative) were allowed to clonally expand before being cryogenically frozen and stored for later experiments.

Table 3.10: PCR primers designed for genotyping AS3MT knockout cell lines.

Primer	Sequence (5' – 3')	Chromosomal Location	Exon
Internal_Forward	GAGAAGTTGGGAGAGGCTGG	Chr10:102873184-102873203	5
Internal_Reverse	AGCACTTGTTGTTTATCAGGCAC	Chr10: 102874614-102874636	6
External_Forward	CTGGGTAGTGGAAGTGGCAG	Chr10:102872506-102872525	4
External_Reverse	CTGGCAGTTCAAGGCTCGTA	Chr10:102876986-102877005	7

Table 3.11: Internal primer set thermocycling conditions for AS3MT genotyping.

Step	Temperature (°C)	Time	Number of Cycles
Hotstart	95	5 min	1
Denaturation	94	30 sec	
Annealing	64	30 sec	35
Extension	72	2 min	
Final extension	72	5 min	1

Table 3.12: External primer set thermocycling conditions for AS3MT genotyping.

Step	Temperature (°C)	Time	Number of Cycles
Hotstart	95	5 min	1
Denaturation	94	30 sec	
Annealing	64	30 sec	35
Extension	72	5 min	
Final extension	72	5 min	1

Table 3.13: Exosap clean-up of PCR reactions for Sanger sequencing.

Reagent	Volume (µl)	Concentration
PCR Product	18	
Exonuclease I	0.5	10U
FastAP	1	1U
ddH_2O	1	
Total	**20**	

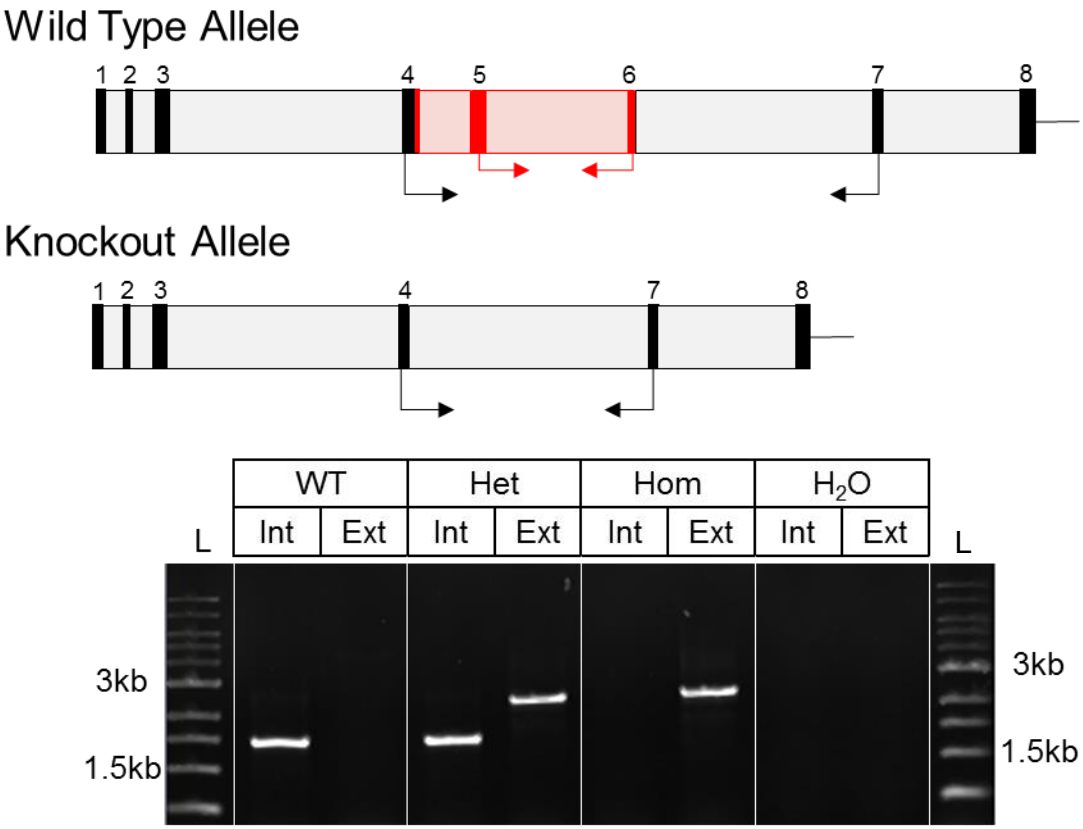


Figure 3.6: Example agarose gel showing the three different genotypes that can arise from knockout of AS3MT using this CRISPR methodology. *Lane L: 1kb DNA ladder. Lanes WT: AS3MT wildtype genotype shows an internal PCR product but no external product as the primers are too far apart to allow for successful amplification. Lanes Het: AS3MT heterozygous genotype produces an internal band indicative of a wildtype allele, however an additional external band of a knockout AS3MT allele is present because non-homologous end joining following disruption allows for successful amplification. Lanes Hom: AS3MT homozygous knockout shows no internal band indicating that deletion of the target region has occurred, confirmation of non-homologous end joining is shown by the presence of the external PCR product. Lanes H_2O are PCR negative controls with water replacing DNA. 2µl PCR product loaded, 0.5% agarose gel, 120v running voltage, 45 minutes running time. Het = Heterozygous, Hom = Homozygous, Kb = Kilobases, Int = Internal primer set, Ext = External primer set, WT = Wildtype.*

3.3.3. Protein extractions

Total cellular protein was extracted from cells to examine protein expression of *AS3MT* in the selected cell lines. Once 80% confluent in 6 well plates the cells were washed with ice cold phosphate buffered saline (PBS) before 100µl of ice cold lysis buffer (**Appendix A**) containing protease inhibitor (cOmplete Mini EDTA Free Protease Inhibitor Cocktail, Roche) was added to each well. Plates were left on ice for 10 minutes before active cell lysis using a cell scraper. Lysate was transferred to ice cold Eppendorfs before being vortexed for one minute to completely lyse the cells. Samples were centrifuged at 5000rpm at 4°C for 10 minutes to remove cell debris and the protein containing supernatant was collected for quantification.

3.3.4. Protein quantification

Protein concentration was calculated using a bicinchoninic acid assay (ThermoFisher Scientific) following manufacturer's instructions. Known bovine serum albumin (BSA) protein standards ranging from 2000µg/ml to 0µg/ml were used to calculate a standard protein absorbance curve. Absorbance was measured at 562nm using a PheraSTAR plate reader and unknown samples compared to the BSA standard curve.

3.3.5. SDS-Polyacrylamide gel electrophoresis

12% SDS-PAGE gels were used for all experiments and made on the day using the TGX stain free Fast cast acrylamide kit 12% (Biorad, Hercules, California, USA) following manufacturer's instructions. 40µg/ml of protein was mixed in a 1:1 ratio with laemmli sample buffer (Biorad, Hercules, California, USA) containing 2-mercaptoethanol (Biorad, Hercules, California, USA). This was incubated at 95˚C for 2 minutes before being loaded into the gel. 4µl of prestained blue protein ladder (GeneFlow Ltd, Lichfield, UK) was also loaded. The gels were run using a miniprotean tetra system at a constant voltage of 180V for 45 minutes in 1x running buffer (TGS Buffer, Biorad, Hercules, California, USA). Proteins were then transferred to a nitrocellulose transfer membrane (Biorad, Hercules, California, USA) at a constant voltage of 100V for 75 minutes. Upon completion of the run the membranes were removed and blocked in 5% skimmed milk 1x Tris Buffered Saline-Tween (TBST) for 1 hour at room temperature on a lab rotator before western blotting. Buffer compositions for Tris Buffered Saline (TBS) and TBST are shown in **Appendix A – Buffer Compositions**.

3.3.6. Western blotting

After blocking the membranes were washed four times for 5 minutes each with 1xTBST. They were then incubated with one of the following primary antibodies overnight at 4˚C on a lab rotator:

- Anti-β-Actin (4967S, New England Biolabs, Massachusetts, USA): monoclonal rabbit antibody, 1:5000 dilution in 2% skimmed milk 1xTBST.
- Anti Cyt19 (sc-377436, Santa Cruz Biotechnology, Texas, USA): monoclonal mouse antibody, 1:100 dilution in 2% skimmed milk 1xTBST.

The following morning each membrane was washed six times for 5 minutes with 1xTBST. The membrane was then incubated with the appropriate secondary antibody listed below for 1 hour at room temperature on a lab rotator:

- Goat anti-Rabbit IgG (H+L) Cross-Adsorbed Secondary Antibody, DyLight 800 (SA5-10036, Invitrogen, Carlsbad, California, USA): polyclonal goat antibody, 1:5000 dilution in 5% skimmed milk 1xTBST.
- Goat anti-Mouse IgG (H+L) Cross-Adsorbed Secondary Antibody, DyLight 680 (35519, Invitrogen, Carlsbad, California, USA): polyclonal goat antibody, 1:5000 dilution in 5% skimmed milk 1xTBST.

After incubation, membranes underwent six 5 minute washes with 1xTBST to remove any residual antibody before protein detection.

3.3.7. Protein detection and quantification

Protein detection was carried out using the Licor Odyssey imaging system utilising the 700nm and 800nm detection wavelengths of the secondary antibodies.

3.3.8. Differentiation of the SH-SY5Y cell line for *AS3MT* qPCR

$1x10^6$ P14 SY-SY5Y cells were seeded in six well plates prior to differentiation. 12 hours after seeding RNA was extracted from cells as described in **2.3.3.** for the 0 day time point. Remaining SH-SY5Y cells were then differentiated with differentiation media (DMEM/F-12 Glutamax, 1% FBS, 10µM RA) for 8 days, with media changes every 48 hours. RNA was collected at days 2, 4, 6 and 8 prior to media changes to fresh differentiation media.

3.3.9. qPCR of *AS3MT* isoforms throughout differentiation

100ng of RNA extracted (as described in 2.3.1.) from the differentiated SH-SY5Y cells was converted into cDNA using the SuperScript VILO cDNA Synthesis Kit as described in **2.5.1.** qPCR for $AS3MT^{full}$ and $AS3MT^{d2d3}$ was carried out as described in **2.5.2.** in a 384 well plate using the QuantStudio 6 Flex system. The qPCR primers for the *AS3MT* transcripts are listed in **Table 3.14**. Analysis was carried out as described in **2.5.4.** using 0 day sample as the baseline, all fold changes are shown as fold changes relative to 0 days.

Table 3.14: qPCR primers for the AS3MT isoforms.

Transcript	Primer Name	Sequence (5'–3')
AS3MT^{full}	qPCR_AS3MT_Full_F	CGAGGAGACATGGCTGCAC
	qPCR_AS3MT_Full_R	GTGGTGACACAGCCGTTGG
AS3MT^{d2d3}	qPCR_AS3MT_d2d3_F	GCCGAGGAGACAATATTATGGCT
	qPCR_AS3MT_d2d3_R	TGGTCATGTCTATTCCAGTCACGT

3.3.10. Differentiation of SH-SY5Y *AS3MT* cell lines for immunocytochemistry

$2x10^5$ of *AS3MT* cell lines B6, C6 (*AS3MT*$^{-/-}$) and D2, D4 (*AS3MT*$^{+/+}$) were seeded into a 24 well plates containing acid etched glass coverslips which had been pre-treated for 12 hours with 200µg/ml of poly-D-lysine (PDL) followed by treating for 12 hours with 20µg/ml laminin. 12 hours after seeding, growth media was changed to differentiation media (DMEM/F-12 Glutamax, 1% FBS, 10µM RA). Differentiation media was changed every 48 hours for seven days before cells were used for downstream experiments (Jahn, Blumer and Pathak, 2017).

3.3.11. Immunocytochemistry

Following differentiation for seven days differentiation media was removed and cells were washed with 1ml sterile PBS to remove any remaining media. Cells were then fixed to coverslips by adding 1ml of 4% paraformaldehyde (PFA) (diluted in 1x PBS) and incubating at room temperature for 10 minutes. Following incubation in PFA cells were washed three times with PBS to remove any residual PFA. Cell membranes were permeabalised by adding 1ml of 0.16% Triton-X PBS to each coverslip at room temperature for 20 minutes. Coverslips were then washed three times with 1ml 1x PBS to remove and residual Triton-X. Coverslips were then blocked for 1 hour at room temperature using 1ml blocking buffer (1x PBS, 1% BSA, 0.1% Tween-20, 300mM glycine) to prevent non-specific antibody binding. Coverslips were then moved to a black humidified chamber and were incubated with the following primary antibody overnight at 4°C:

- Anti-MAP-2 (AP20) (sc-32791, Santa Cruz Biotechnology, Texas, USA): monoclonal mouse antibody, 1:200 dilution in PBST (1x PBS, 1% BSA, 0.1% Tween-20).

The following morning, primary antibody was removed and coverslips washed three times for five minutes each with 1x PBS. Coverslips were then incubated with the following secondary antibody for 1 hour at room temperature in the dark:

- Alexa Fluor® 555 Goat Anti Mouse IgG (H+L) Cross-Adsorbed Secondary Antibody (A-21422, Invitrogen, California, USA): polyclonal goat antibody, 1:400 dilution in PBST (1x PBS, 1% BSA, 0.1% Tween-20).

Secondary antibody was then removed and coverslips washed three times for five minutes with 1x PBS to remove any residual secondary antibody. A final 1x PBS wash for 10 minute wash ensured full removal of unbound secondary antibody before mounting. Coverslips were mounted to glass slides, cell side down, using 10µl of VECTASHIELD HardSet Antifade Mounting Medium with DAPI (Vector Laboratories, Burlingame, CA, USA), which counterstains the nuclei. Coverslips were left to set in the dark at room temperature overnight before imaging the following day using a Leica DM4000 B LED Fluorescence microscope (**Figure 3.7**)

3.3.12. Imaging and analysis

Three coverslips per genotype were imaged using a Leica DM4000 B LED Fluorescence microscope at 20 x magnification. Images were exported and analysed using ImageJ. Prior to analysis, a scale of 185 pixels to 50 µm (micrometres) was applied to each image using the *FIJI* software. The "Polygon selections" tool was used to trace the borders of MAP2-positive cells (in red) and the area enclosed within the polygon (in μm^2) was measured to determine the cell area. To determine neurite lengths, the "Segmented Line" tool was used to measure the distance (in µm) from the border of the nucleus (in blue) to the distal end of neurites. The sum of the lengths of all the neurites for each cell was used to provide a total neurite length measurement for each MAP2-positive cell. Statistical analysis was undertaken using R. P values shown are from a Welch's t-test with degrees of freedom (DF), t-statistic, and p-value reported.

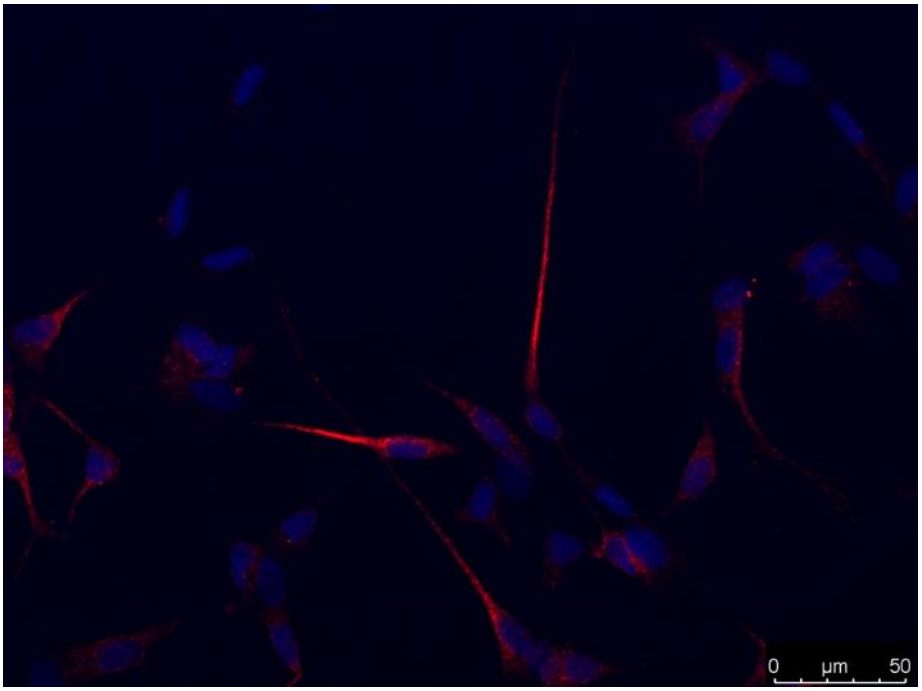


Figure 3.7: Example image of MAP2 immunocytochemistry on differentiated SH-SY5Y cells. *Imaged on Leica DM4000 B LED, 20x magnification. Red MAP2 staining, Blue DAPI nuclear stain. Note the neurite projections, these were measured using ImageJ.*

3.4. Results

3.4.1. The *AS3MT^{d2d3}* but not the *AS3MT^{full}* transcript is differentially expressed throughout SH-SY5Y differentiation towards neuronal fates

The first aim of this chapter was to confirm that the AS3MT^{d2d3} transcript is differentially expressed throughout neuronal differentiation in SH-SY5Y cells as previously reported in iPSC lines (M. Li *et al.*, 2016). To achieve this, SH-SY5Y cells were seeded in six well plates and differentiated into neurons with RA over eight days. RNA was extracted from the cells at two day intervals and qPCR was carried out for the AS3MT^{full} and AS3MT^{d2d3} transcripts. The fold change expression of the two isoforms was calculated and compared to the zero day samples, prior to the addition of RA. All time points have three biological replicates.

There was a relative fold change of 2.49 in the expression of the AS3MT^{d2d3} transcript after differentiation initiation compared to the zero day sample (One-way ANOVA, $F(4,10)=4.82$, p-value=0.02). A pairwise t-test comparison identified significant changes between zero days and two days (p-value=0.02), zero days and four days (p-value=0.02), zero days and six days (p-value=0.03) and zero days and eight days (p-value=0.02). There were no significant changes between the other time points, indicating that expression is stable through differentiation (**Figure 3.8**.). I further validated that there is no significant change in the expression of the AS3MT^{full} transcript (**Figure 3.9**) through differentiation in SH-SY5Y cells (One-way ANOVA, $F(4,10)=1.49$, p-value=0.28). There is also no significant difference in the relative ratio of the two isoforms through differentiation (One-way ANOVA, $F(4.10)=1.64$, p-value=0.24) (**Figure 3.10**).

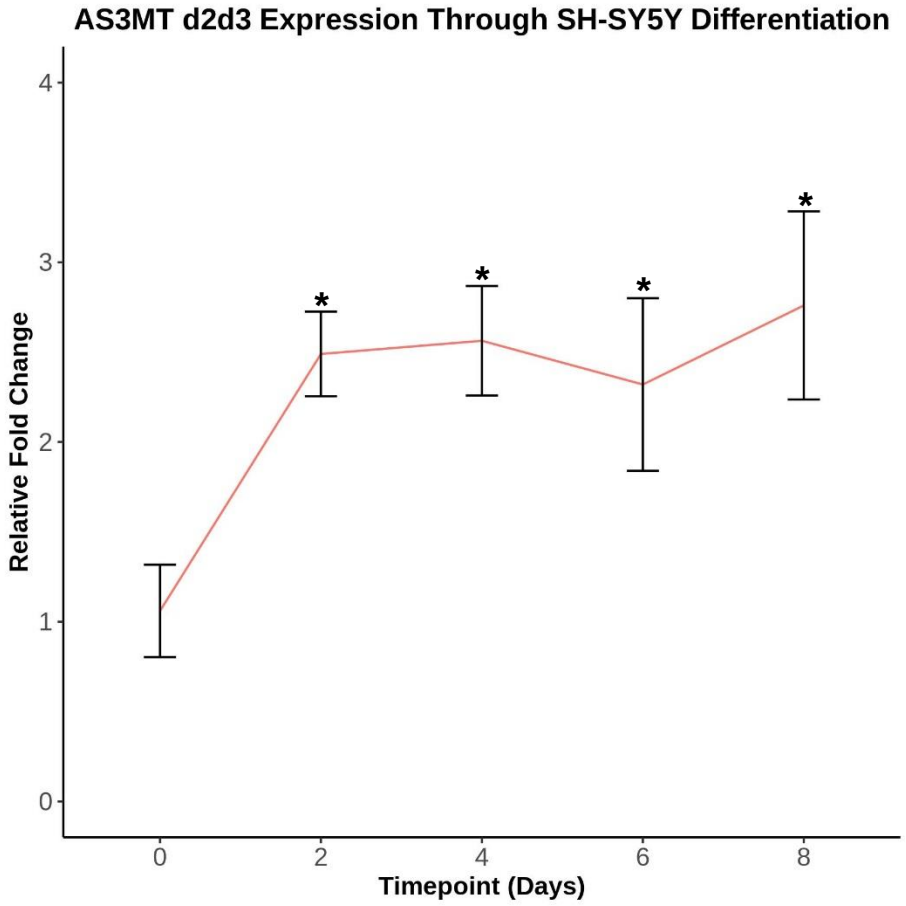


Figure 3.8: There is a significant increase in the expression of the AS3MT^{d2d3} transcript through SH-SY5Y differentiation towards neuronal fates. *One way ANOVA, F(4,10)=4.82, p-value=0.02. Significance between groups identified by pairwise t-test. n=3 biological repeats, error bars are +/- the SE.*

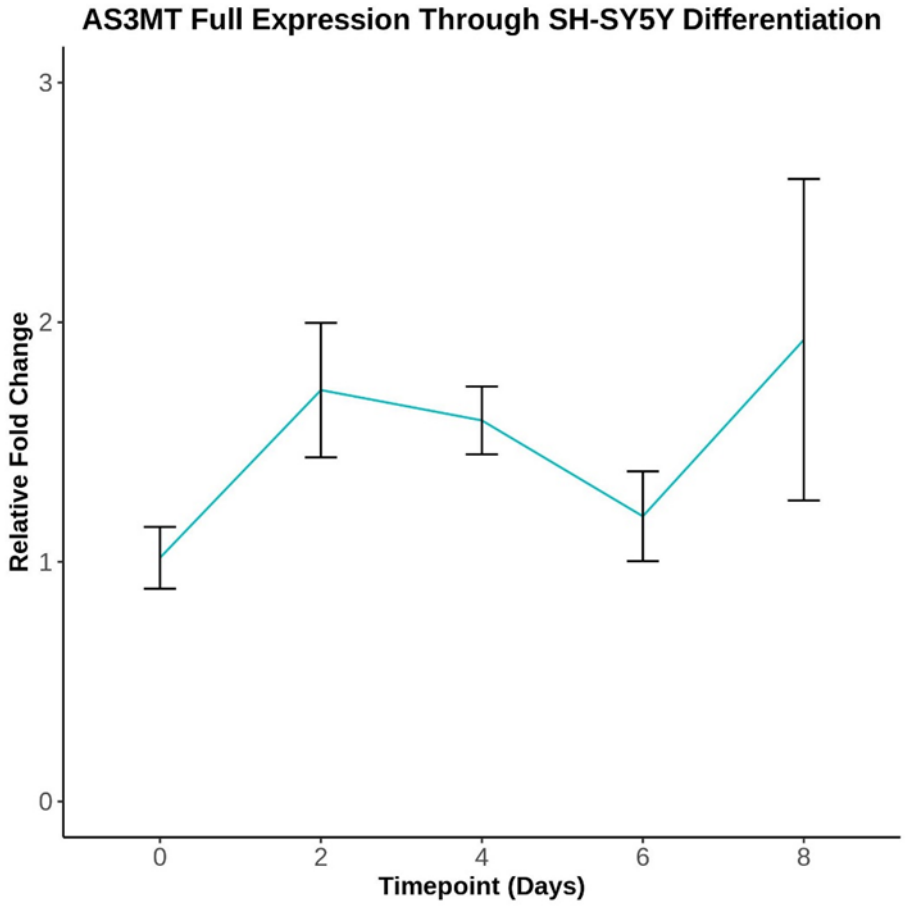


Figure 3.9: There is no change in the expression of the AS3MT^{full} transcript through SH-SY5Y differentiation towards neuronal fates. *One way ANOVA, F(4,10)=1.49, p-value=0.28. No significance between groups identified by pairwise t-test. n=3 biological repeats, error bars are +/- the SE.*

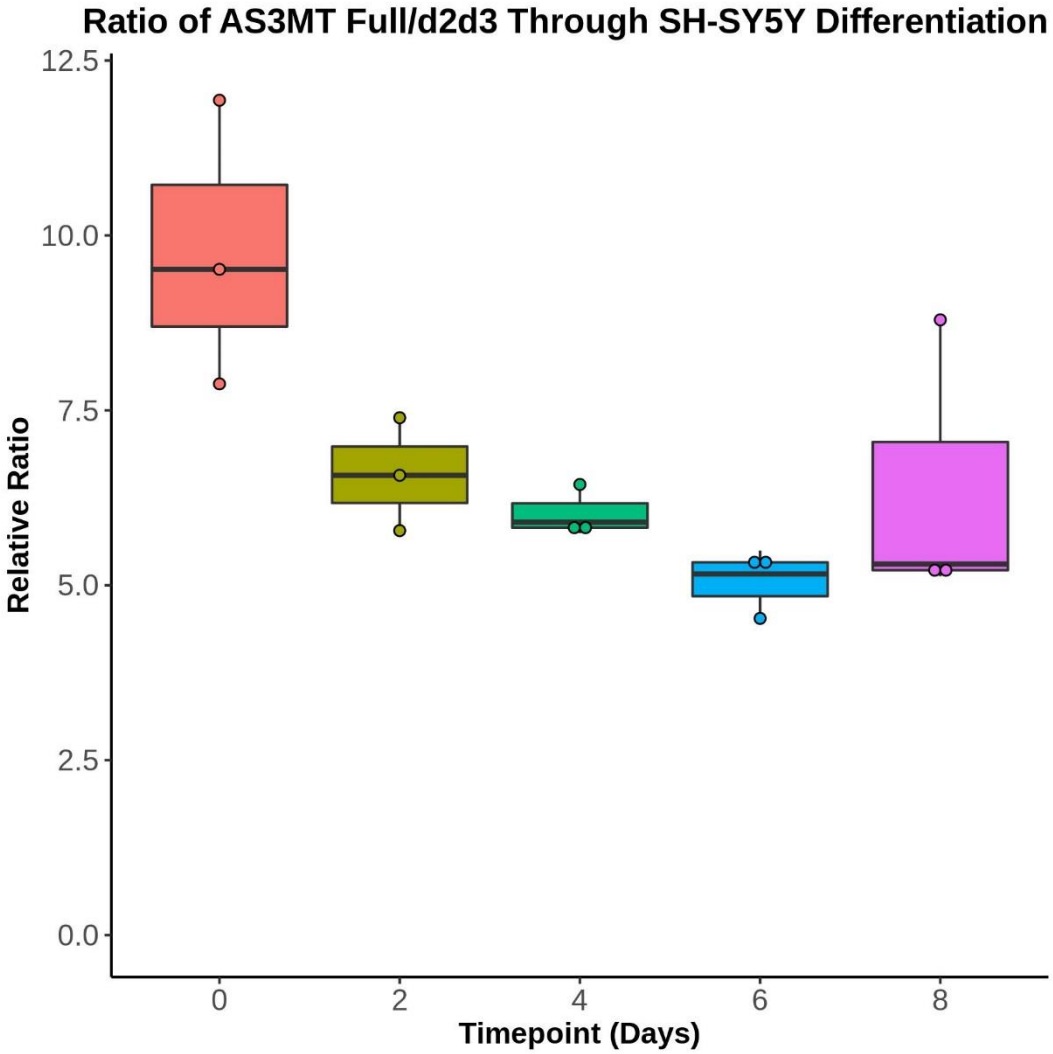


Figure 3.10: The relative ratio of the two AS3MT transcripts does not change through SH-SY5Y differentiation. *One way ANOVA F(4,10)=1.64, p-value=0.24. n=3 biological repeats.*

3.4.2. Overview of the experimental strategy for knockout of *AS3MT* in SH-SY5Y neuroblastoma cells

The second aim of this chapter was to generate a methodology for knocking out *AS3MT* using CRISPR-Cas9 genomic engineering techniques in a previously non CRISPR edited cell line. Two separate gRNA targeting the main methyltransferase domain of AS3MT were designed using the online CRISPR design tool (http://crispr.mit.edu), and guides were selected with the lowest off target score and highest on target scores **Table 3.1**. These gRNA were then cloned into commercially available CRISPR vectors which, when transfected into human cells, express the gRNA, active Cas9 and a fluorescent reporter (GFP or mCherry) and cause site specific disruption. Sanger sequencing confirmed correct cloning of gRNA into CRISPR vectors as shown in **Figure 3.4**. These CRISPR plasmids were then transfected into P11 SH-SY5Y neuroblastoma cells using through Nucleofection using a Nucleofection kit from Lonza as described in **Table 3.8**. Transfection efficiency was observed at 24 hours by fluorescence imaging on a Leica DMi8 inverted microscope as shown in **Figure 3.5**. Following transfection cells were then sorted into single cell populations using FACS before being clonally expanded for 2 months to give monoclonal populations. Monocultures of cells were then genotyped by PCR as described in **Figure 3.6** to confirm sequence disruption of the gene and the genotype of the cells. Finally total cellular protein was extracted from homozygous and wildtype cell lines and a western blot was undertaken to identify if disruption of the *AS3MT* sequence resulted in a knockdown of the AS3MT protein. A flowchart of the experimental procedures are shown in **Figure 3.10**.

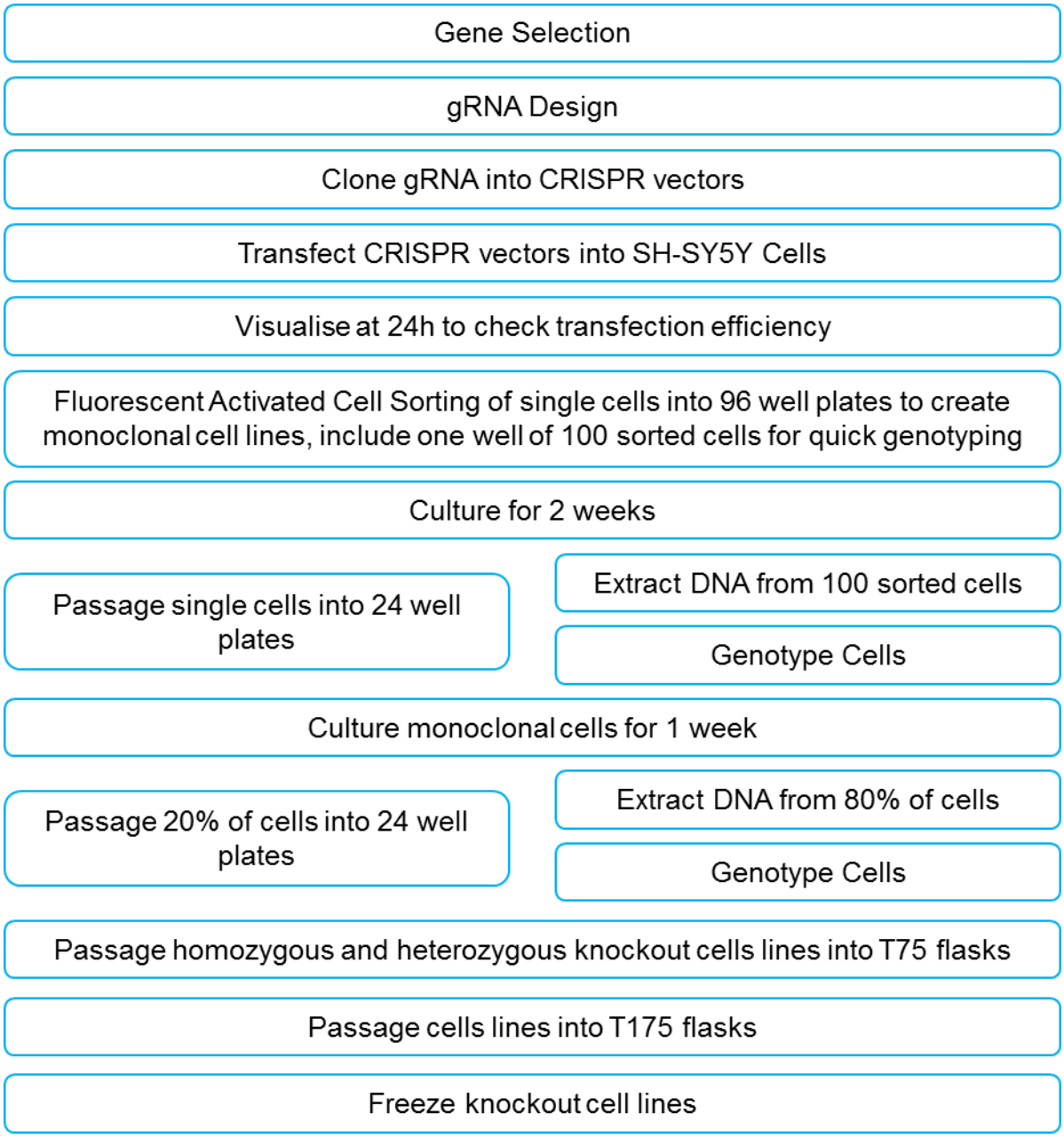


Figure 3.10: Flowchart identifying the process of generating the knockout AS3MT cell line using CRISPR/Cas9.

3.4.3. FACS of CRISPR transfected cells can be used to generate monoclonal populations

Following successful transfection of both the CRISPR plasmids into SH-SY5Y cells the cells were single cell sorted into a single well of a cell culture plate to develop monoclonal populations of cells for further experiments. As the transfection efficiency of SH-SY5Y cells is low there is a high probability that the identified transfected cells may still not have undergone disruption. Transfected cells which have undergone disruption will have different genotypes amongst themselves as the Cas9 cuts randomly on each allele, as shown in **Figure 3.11**. In order to increase the likelihood of screening knockout cell lines, the CRISPR plasmid include fluorescent markers of GFP or mCherry. Successfully transfected cells therefore express this fluorescent marker and FACS can be used to isolate this population of cells. We isolated single cells from this transfected population of cells into 96 well plates which we clonally expanded over several weeks to give a monoclonal populations of cells with the same genotype.

For FACS the Exon 4, Exon 6, Exon 4 + Exon 6, and substance negative transfections from **Table 3.8** were used. The Exon 4 CRISPR plasmid contained the GFP selection marker whereas the Exon 6 CRISPR plasmid contained the mCherry selection marker. Flurochrome specific wavelengths were used to set the gates on the FACS machine to allow selection of true positive cells, the substance negative cells were sorted to allow for a matched control for downstream experiments which are of the same age and has undergone the same experimental procedures. **Figure 3.12** shows the substance negative transfected cells (or wild type (WT) cells) and the gating strategy used to identify show where FITC and PE-CF594 positive cells would appear.

Following gating using the substance negative control the single transfections were FACS into 96 well plates. The Exon 4 samples were sorted first, 37.3% of the single cells sorted were GFP positive. 95 of these GFP positive cells were each sorted into a well of a 96 well plate as to have a single cell per well, the final well contained 100 GFP positive cells. The Exon 6 sample showed 33.3% of the single cells were mCherry positive. Again 95 single cells were sorted into a 96 well plate with the final well containing 100 mCherry positive cells. Once the single transfections were FACS sorted the double transfection containing both Exon 4 and Exon 6 were sorted for

double positive FITC and PE-CF594 cells. As shown in **Figure 3.13** there is a clear positive correlation between uptake of mCherry and GFP plasmids, where the fluorescence of FITC is higher there is also a higher fluorescence of PE-CF594. 19.5% of single cells were GFP positive, 22.2% were mCherry positive and 15.5% were double positive for GFP and mCherry. From these double positive cells we sorted three 96 well plates containing 95 single cells which were double positive and one well containing 100 cells.

Once two plates of double positive single cells were collected the remaining double positive cells were sorted into an Eppendorf, these were then rerun through the FACS machine to give an “enriched” sample removing any false positives. Unsurprisingly 87% were resorted as double positive cells. However this does give 13% of cells which underwent the first sort were false positives. One 96 well plate of enriched double positive single cells were collected along with a well containing 100 cells. A summary of all the FACS results is listed in **Table 3.15**. The % of wells with single cells observed at 7 days following sorting is listed in **Table 3.16**.

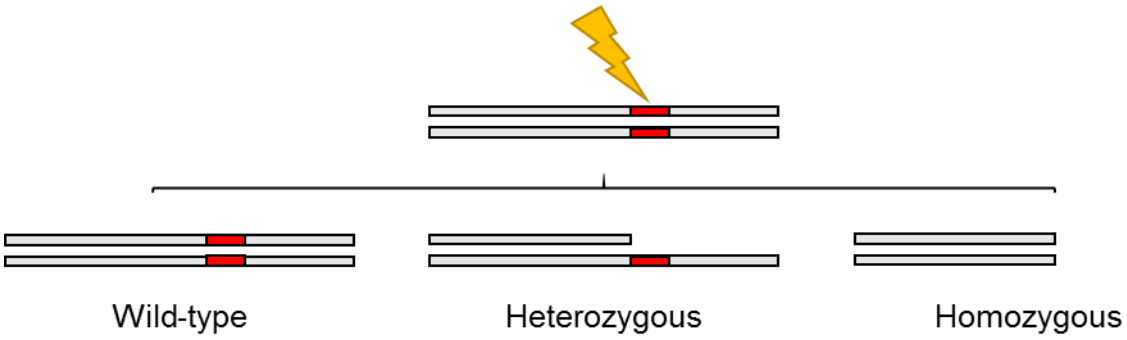


Figure 3.11: The different genotypes which can arise from a CRISPR experiment. Following disruption at the PAM sequence (Red) by the Cas9 (lightening) three distinct genotypes can occur in a diploid organism. *The first a full wild-type where the Cas9 has not cut either allele, this is most likely due to either a poorly designed gRNA or too low concentration of CRISPR plasmid. The second genotype is a heterozygous knockout, this is where only one allele of the gene has been disrupted by the CRISPR, and here the CRISPR has only cut at one PAM sequence. The final genotype is a full homozygous knockout, where disruption has occurred at the PAM sequence on both alleles. Identification of a homozygous population is advantageous as these are most likely to have loss of function.*

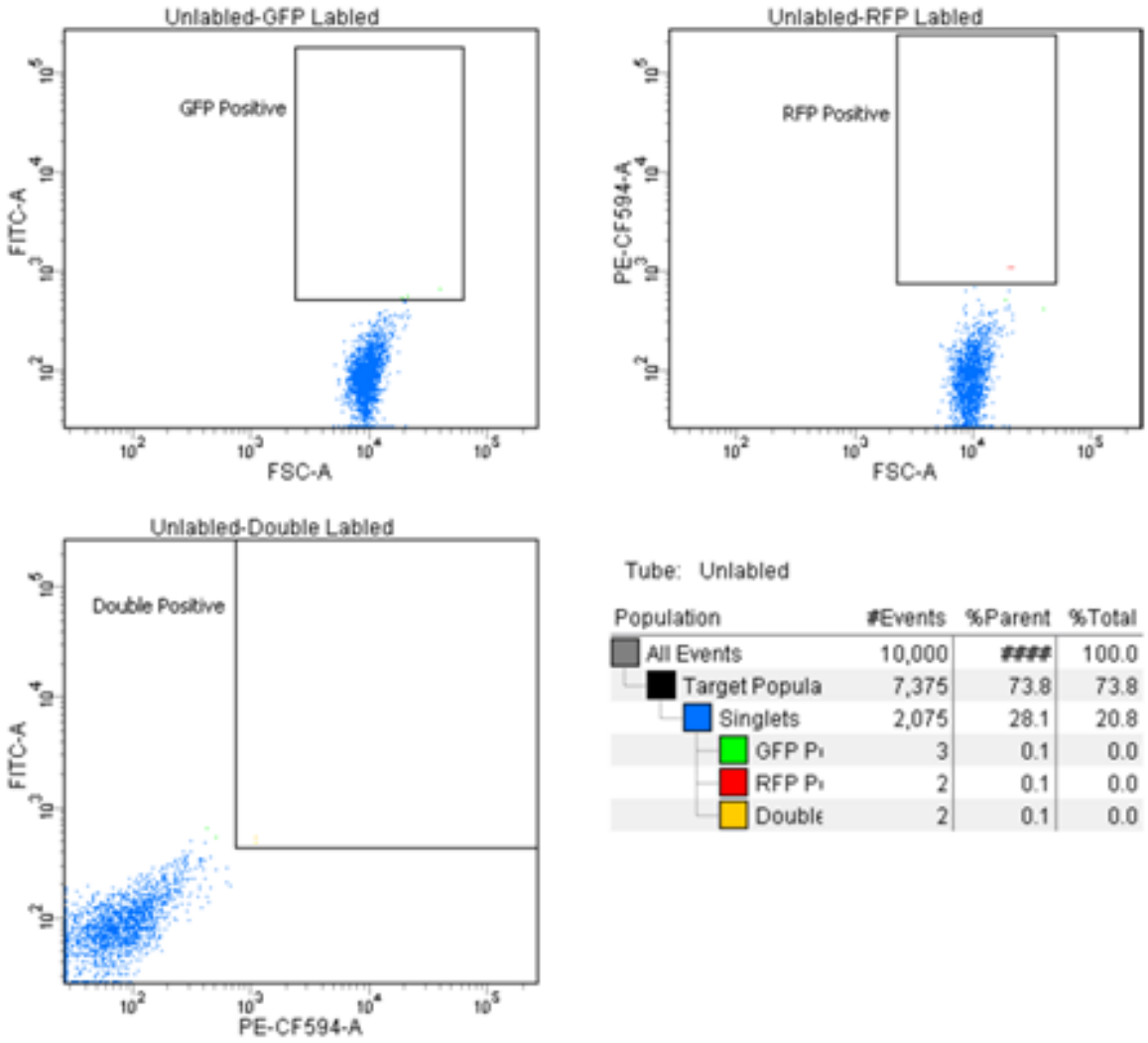


Figure 3.12: FACS results from the Substance Negative transfection. *Each blue dot within the three graphs are an event (cell) identified through FACS. The top two graphs plot the forward scatter on the X axis, a marker of cell size, with GFP intensity on the Y axis of the left graph and mCherry intensity on the Y axis of the right graph. Running the negative control through the FACS allows us to draw the gate for GFP and mCherry fluorescence, these are the boxes within the graphs. From this when we run experimental samples through the FACS any cells which are within this box are considered to be GFP or mCherry positive. The bottom left graph is the mCherry intensity on the X axis plotted against the GFP intensity on the Y axis, this is used for identifying double fluorescent positive cells which we want to select for. The bottom right is a table recording the events (cells) which occur in each subset.*

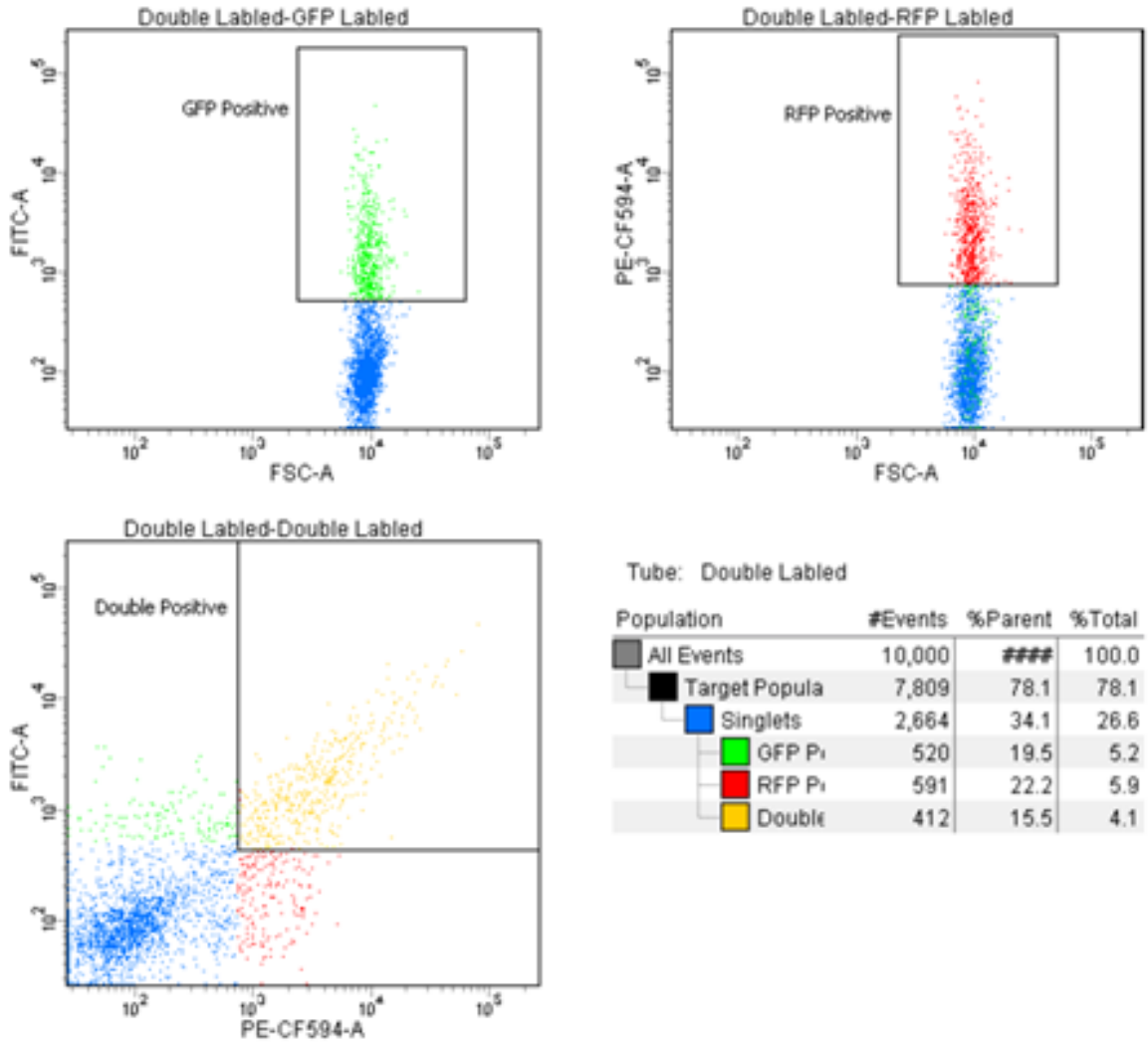


Figure 3.13: FACS results from the Exon 4 + Exon 6 double transfection. *Following gating using the substance negative control the double transfection of Exon 4 + Exon 6 was run through the FACS to select for double positive GFP and mCherry positive cells. From the FACS we identified that 19.5% of single cells were single GFP positive, 22.2% of sorted cells were mCherry positive and 15.5% were double positive. The double positive cells were sorted into three 96 well plates with a single cell in each well, apart from well 96 where 100 cells were seeded for quicker genotyping.*

Table 3.15: Reported fluorescence of each transfection following FACS sorting.

Transfection	GFP Positive Single Cells (%)	mCherry Positive Single Cells (%)	Double Positive Single Cells (%)
Substance negative	0.1	0.1	0.1
Exon 4	37.3	0.0	0.0
Exon 6	0.1	33.3	0.2
Exon 4 + Exon 6	19.5	22.2	15.5
Exon 4 + Exon 6 Enriched	87.8	88.3	87.0

Table 3.16: Survival efficiency of FACS sorted CRISPR transfected SH-SY5Y. Efficiency was calculated as the percentage of wells showing clonal growth after seven days.

Plate	% wells with clonal growth
Exon 4 + Exon 6 1	35
Exon 4 + Exon 6 2	39
Exon 4 + Exon 6 Enriched	21
Exon 4	42
Exon 6	32
Substance Negative	41

3.4.4. Genotypes of the CRISPR transfected cells identify homozygous and heterozygous knockout cell lines of *AS3MT*

Following FACS sorting the single cells were left to form monocultures over two weeks. After one week the 100 cells which were seeded in the 96th well of the plate were at confluency and DNA was extracted to be used in the genotype assay described in **3.3.2.6.** and **Figure 3.6**. While this is a heterogeneous population of cells genotyping at this stage can identify if the Cas9 has caused disruption and confirm the gRNA designed are targeting correctly. This can save time downstream and prevent the unnecessary passaging of unwanted cells.

The results from the diagnostic PCR from the 100 sorted cells are shown in **Figure 3.14**. From this diagnostic PCR we can see that the double transfected cells and enriched double transfection have a heterozygous mix of cells containing knockout and wildtype DNA, compared to just wildtype DNA in the substance –ve control, which are wildtype cells. Exon 4 and Exon 6 single transfections appear to show a wildtype genotype, however this screening method is only applicable to detect the large deletions between the two exons. Further analysis could have been undertaken on these samples through high resolution melt analysis or T7 endonuclease assay to identify if disruption has occurred in these single knockout lines. This would have been undertaken for troubleshooting if the knockout genotype was not observed in the double transfections as there could be a poor gRNA inhibiting knockout.

Following successful genotyping of the 100 cell stage single cell colonies were left to expand for two weeks before being passaged into 24 well plates. DNA was then extracted from the cells at the 24 well plate stage and was used for genotyping the monocultures of cells from the single FACS sorted cells. DNA extracted from a single 24 well plate which contained cells as outlined in **Table 3.17** underwent the diagnostic PCR as outlined in **3.3.2.6.** and the result is shown in **Figure 3.15** with the resultant genotypes listed in **Table 3.18**. From this approach of the 18 cell lines genotyped, 5/18 were heterozygous, 3/18 were homozygous. The cell lines kept for further functional analysis are listed in **Table 3.19**.

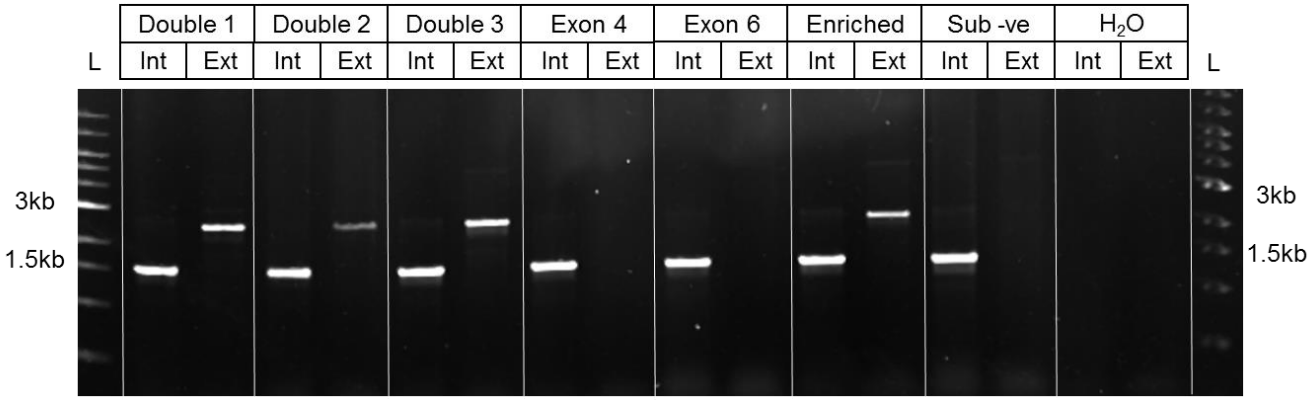


Figure 3.14: Diagnostic PCR of 100 FACS sorted cells following transfection with CRISPR plasmids. *Each of the 7 plates of cells sorted from FACS contained a single well with 100 cells to aid with genotyping as these reach confluency much faster than single cells. DNA was extracted and the diagnostic PCR as outlined in 3.3.2.6. was undertaken. In the double and enriched samples we observe the expected PCR products for a heterozygous knockout cell line, indicated by the presence of both the internal (Int) and external (Ext) PCR bands. The substance negative control are FACS sorted cells containing no CRISPR plasmids so will have a wildtype genotype as indicated by the presence of a single Int band. The Exon 4 and Exon 6 single transfection and sorted cells show only a single Int band indicating a wild-type genotype, however with this screening methodology we cannot rule out disruption has occurred, further analysis through high resolution melt would be required. L=1kb DNA Ladder (Solis Biodyne), Int = Internal primer set, Ext = External primer set.*

Table 3.17: The source of the monoclones used for the diagnostic PCR. *18 clones are from double CRISPR transfections from three different FACS sorts with the final six representing wild type controls.*

	1	**2**	**3**	**4**	**5**	**6**
A	Double 1	Double 1	Double 1	Double 1	Double 1	Double 1
B	Double 2	Double 2	Double 2	Double 2	Double 2	Double 2
C	Enriched	Enriched	Enriched	Enriched	Enriched	Enriched
D	Sub -ve	Sub -ve	Sub -ve	Sub -ve	Sub -ve	Sub -ve

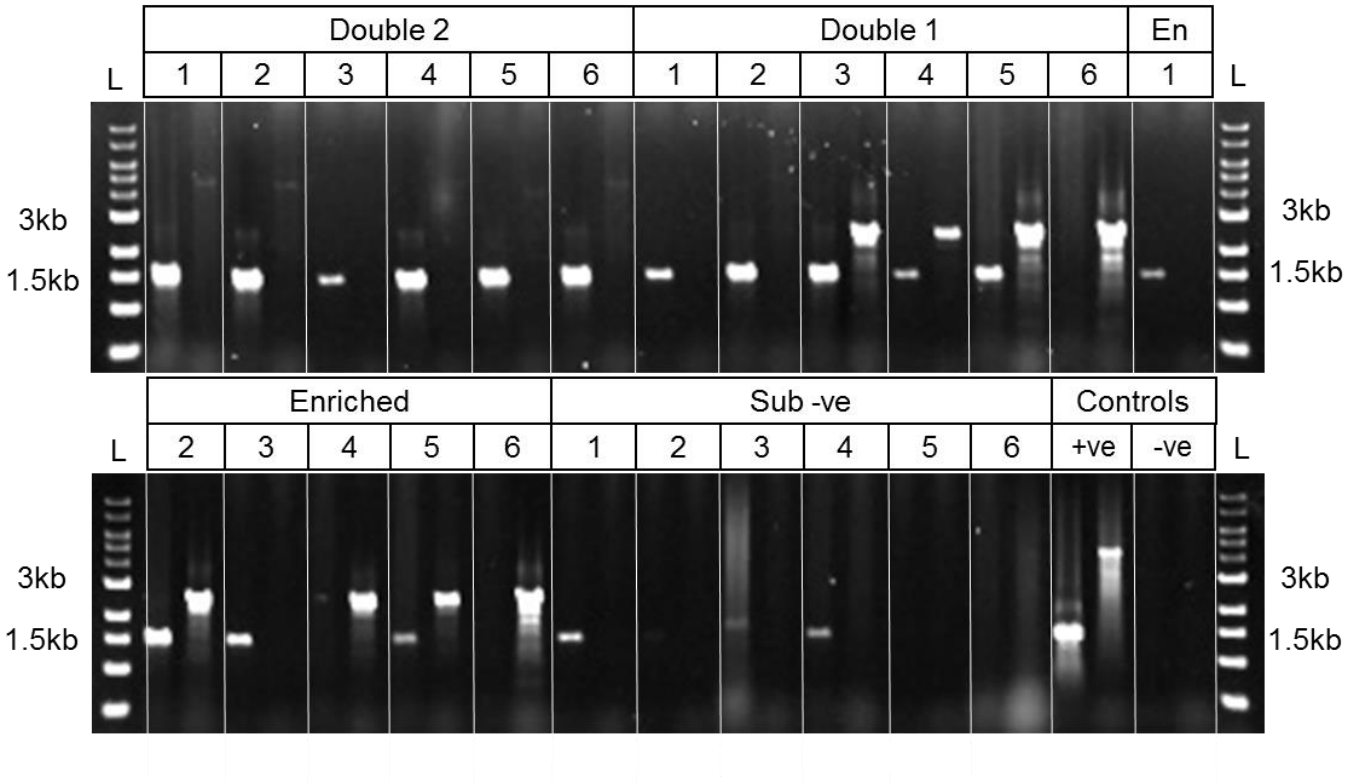


Figure 3.15: Genotyping PCR result from 24 well plate listed in Table 3.17*. Each sample has an internal PCR (left band) and external PCR (right band). Presence of left band only indicates WT genotype, presence of right band only indicates homozygous KO, both bands indicated heterozygous knockout. L= 1kb DNA Ladder (Solis Biodyne). +ve control, DNA extracted from 100 cell well, -ve control water replacing DNA.*

Table 3.18: AS3MT genotype of the 24 well plate cells following genotyping PCR. *Yellow indicates heterozygous knockout, green indicates homozygous knockout, white indicates wild type.*

	1	2	3	4	5	6
A	Double 1	Double 1	Double 1	Double 1	Double 1	Double 1
B	Double 2	Double 2	Double 2	Double 2	Double 2	Double 2
C	Enriched	Enriched	Enriched	Enriched	Enriched	Enriched
D	Sub -ve	Sub -ve	Sub -ve	Sub -ve	Sub -ve	Sub -ve

Table 3.19: AS3MT cell lines used for functional analysis.

Cell Line Name	Genotype	FACS Plate
B6		Double 2
C4	*AS3MT*$^{-/-}$	Enriched
C6		Enriched
B3		Double 2
B5	*AS3MT*$^{+/-}$	Double 2
C2		Enriched
D2		Sub –ve
D3	*AS3MT*$^{+/+}$	Sub –ve
D4		Sub -ve

3.4.5. Sanger sequencing of genotyped clones confirms disruption of the genomic sequence of *AS3MT*

While the diagnostic PCR can identify the genotype of the cells it does not give the exact genomic sequence. While we can distinguish whether or not the targeted sequence has been removed it is unclear if this will cause a frameshift mutation and disrupt the protein. In order to ascertain if the disruption has caused a frameshift the PCR products from the External PCR reactions were cleaned with Exosap and sequenced as described in **3.3.2.6.** .

Example sequencing chromatograms are shown in **Figure 3.16**, with a heterozygous and homozygous example. The three homozygous lines identified by the diagnostic PCR (B6, C4, and C6) have been confirmed to have homozygous *AS3MT* knockout mutations, where both alleles have been disrupted. The heterozygous cell lines (B3, B5, C2) show disruption in only one allele where the other allele is identified as WT. Unsurprisingly we see an increased rate of deletion compared to insertion following non-homologous end joining. Interestingly in all samples which contain a disruption it is within 5-10bp of the PAM sequence, showing how crucial the PAM sequence is for identifying where the cut will occur.

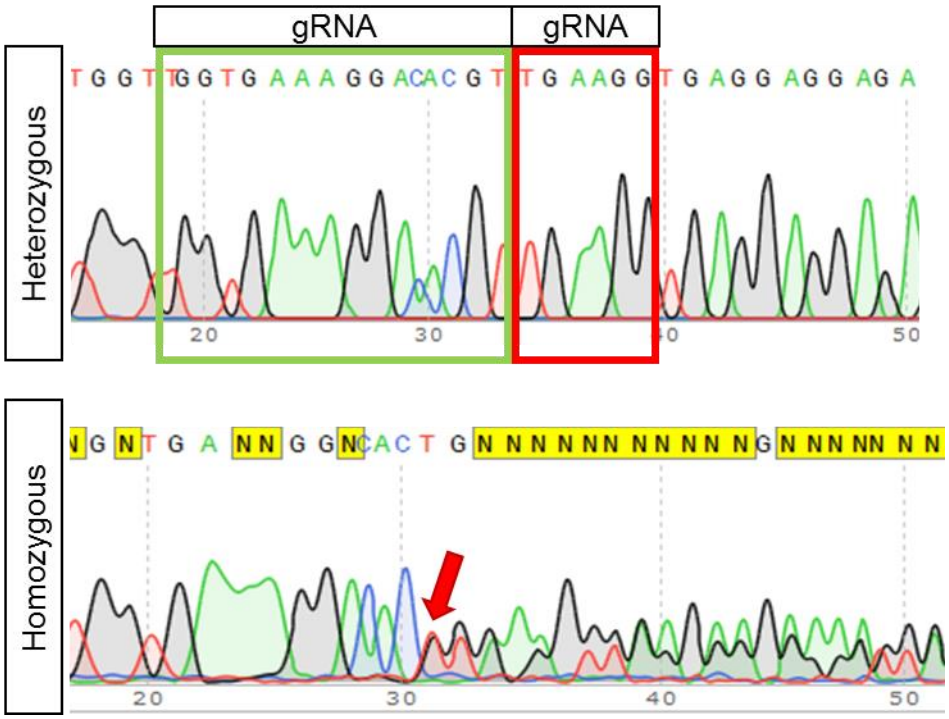


Figure 3.16: Sanger sequencing examples of the AS3MT external PCR products distinguishes between heterozygous and homozygous knockout. *The heterozygous samples you get a clean chromatogram where the start sequence of the first gRNA (green) and the end of the second gRNA (red) meet, indicating the region between has been deleted. In the homozygous the chromatograms overlay at the join point (red arrow) as both alleles are cut randomly by Cas9 resulting in two different lengths of deletion.*

3.4.6. Western blotting for AS3MT confirms knockout of the AS3MT protein

While it is clear that the Cas9 has cut and caused disruption at the target sites this may not have had an effect on the functionality of the protein. Because of this, a western blot was performed on whole cell lysate extracted from the homozygous *AS3MT*$^{-/-}$ and WT *AS3MT*$^{+/+}$ cell lines as described in **3.3.5.** , **3.3.6.** , and **3.3.7.** . As shown in **Figure 3.17** all three of the *AS3MT*$^{-/-}$ cell lines (B6, C4, and C6) have complete disruption of the AS3MT protein as no protein is detected, compared to the *AS3MT*$^{+/+}$ cell lines (D2, D3, and D4), β-Actin was used as a loading control. Therefore I can conclude that I have successful generated several knockout *AS3MT* cell lines for functional analysis.

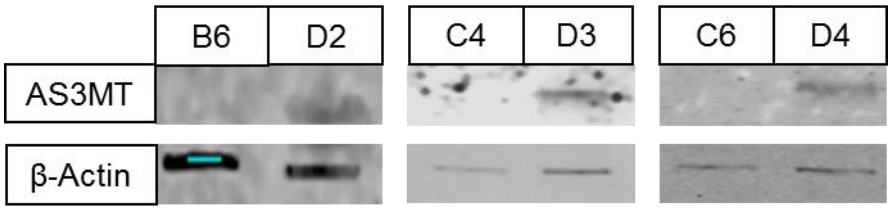


Figure 3.17: Western blot for AS3MT in the genotyped cell lines. *The homozygous AS3MT knockout cell lines B6, C4 and C6 have no expression of the AS3MT protein compared to the wildtype cell line D2, D3, and D4. AS3MT 1:100 dilution, B-Actin 1:5000 dilution, images from three separate western blots.*

3.4.7. MAP2 staining identifies increased cellular area and longest neurite length in differentiated *AS3MT* knockout cells compared to wild type

In order to examine the effect of *AS3MT* deletion on differentiation, two *AS3MT* WT cell lines (D2 and D4) and two *AS3MT* homozygous knockout cell line (B6 and C6) underwent differentiation with 10µM of RA on acid etched, PDL and laminin coated coverslips. These cells were differentiated for a week before being fixed with 4% PFA and immunocytochemistry undertaken for the neuronal marker MAP2. Cell lines C4 and D3 were excluded from analysis as these lines had begun to spontaneously differentiate under normal culturing conditions, a characteristic of aged SH-SY5Y cell line (Xicoy, Wieringa and Martens, 2017). Representative cell pictures of the six cell lines are shown in **Figure 3.18**.

MAP2 is a cytoskeletal marker which is mainly expressed in neurones, it is localised to the dendrites of terminally differentiated neurones and therefore is a marker of neuronal differentiation (Harada *et al.*, 2001; Korzhevskii, Karpenko and Kirik, 2012). MAP2 is a defining marker of neuronal cells in cultures of SH-SY5Y cells (Jahn, Blumer and Pathak, 2017; Shipley, Mangold and Szpara, 2017). Interestingly, MAP2 has been shown to be decreased in the auditory cortex of Schizophrenic patients with neurones showing smaller spines at a lower density (Shelton *et al.*, 2016; Mckinney *et al.*, 2019). As *AS3MT* is expressed though development and is thought to have an effect on neuronal differentiation we set out to identify if neurite length or total neurite length was altered in the knockout cell lines.

Following immunocytochemistry MAP2 positive cells were analysed from three different coverslips per cell line and cell area, longest neurite length and total neurite length was calculated using neurite tracing in ImageJ (National Institute of Health, University of Wisconsin, USA). From a total of 118 knockout neurones (from two different cell lines) and 76 WT neurones (from two different cell lines), there is an approximate increase of 33.9% (1500µm^2) in cellular area in the *AS3MT* knockout lines compared to the WT lines (Welch's T.test, p-value=2.30e-07) (**Figure 3.19**, **Table 3.20**). There is an average 25µm increase in the longest neurite length in *AS3MT* knockout lines (Welch's T.test, p-value=0.0498) (**Figure 3.20**, **Table 3.21**), however there is no significant difference in the total neurite length extending from neuronal cells (Welch's T.test, p-value=0.0563) (**Figure 3.21**, **Table 3.22**).

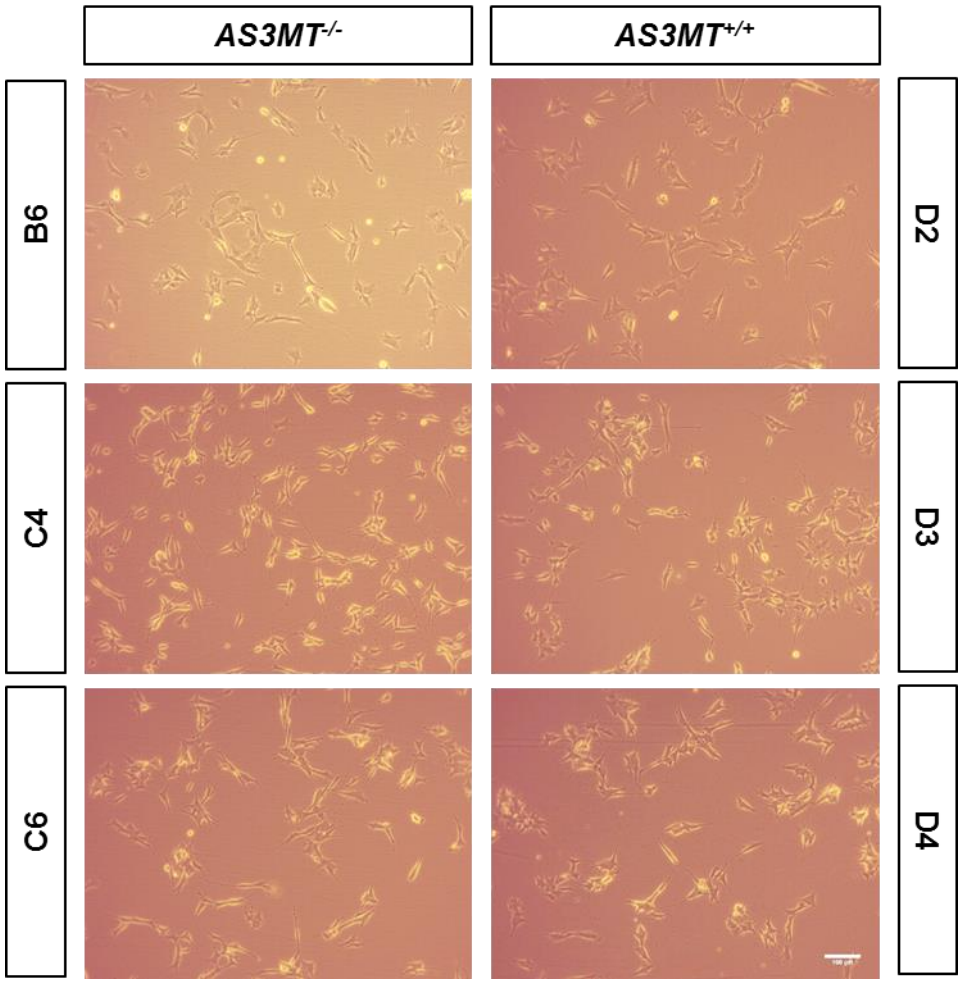


Figure 3.18: Bright-field Images of the six AS3MT KO and WT cell lines used for morphological analysis. *Cell lines C4 and D3 undergo spontaneous differentiation under normal culturing conditions and were excluded from morphological analysis. Scale bar 100µm.*

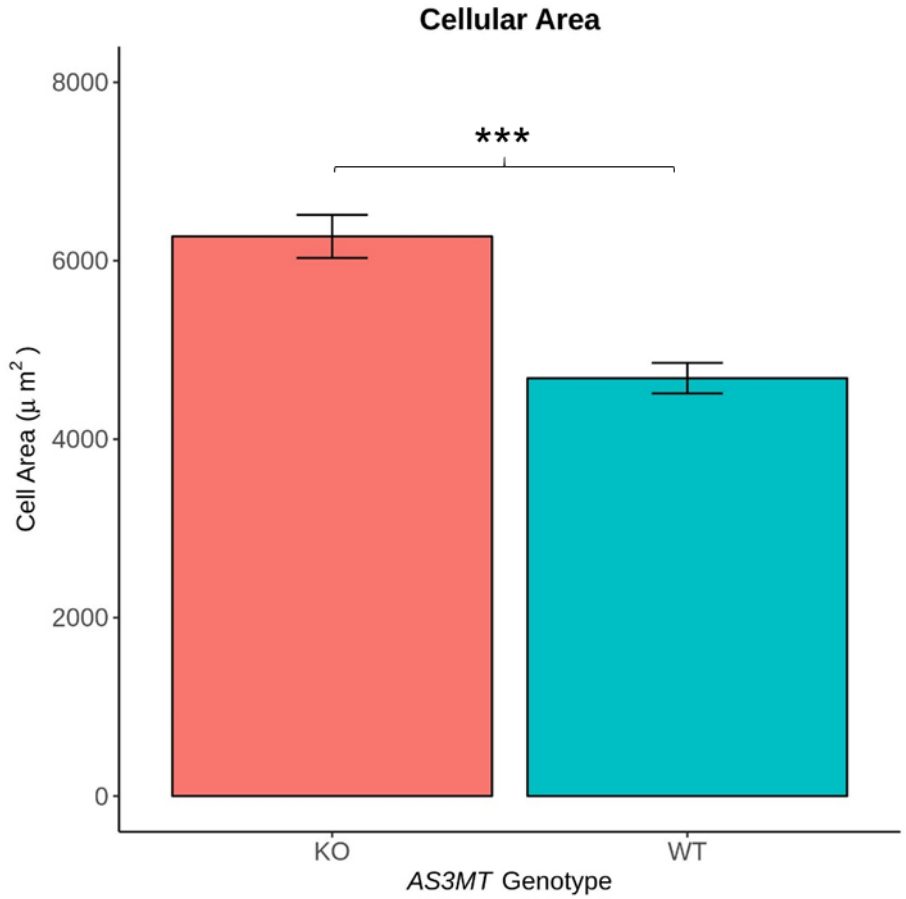


Figure 3.19: Differentiated AS3MT knockout cells show an increased cellular area compared to wild type. *Two-sample t(189.43)=5.369, p-value=2.30E-07*

Table 3.20: Summary statistics of cellular area

Genotype	n	Cellular Area (µm^2)	SD	SE
KO	118	6273.09	2624.45	241.60
WT	76	4683.88	1490.85	171.01

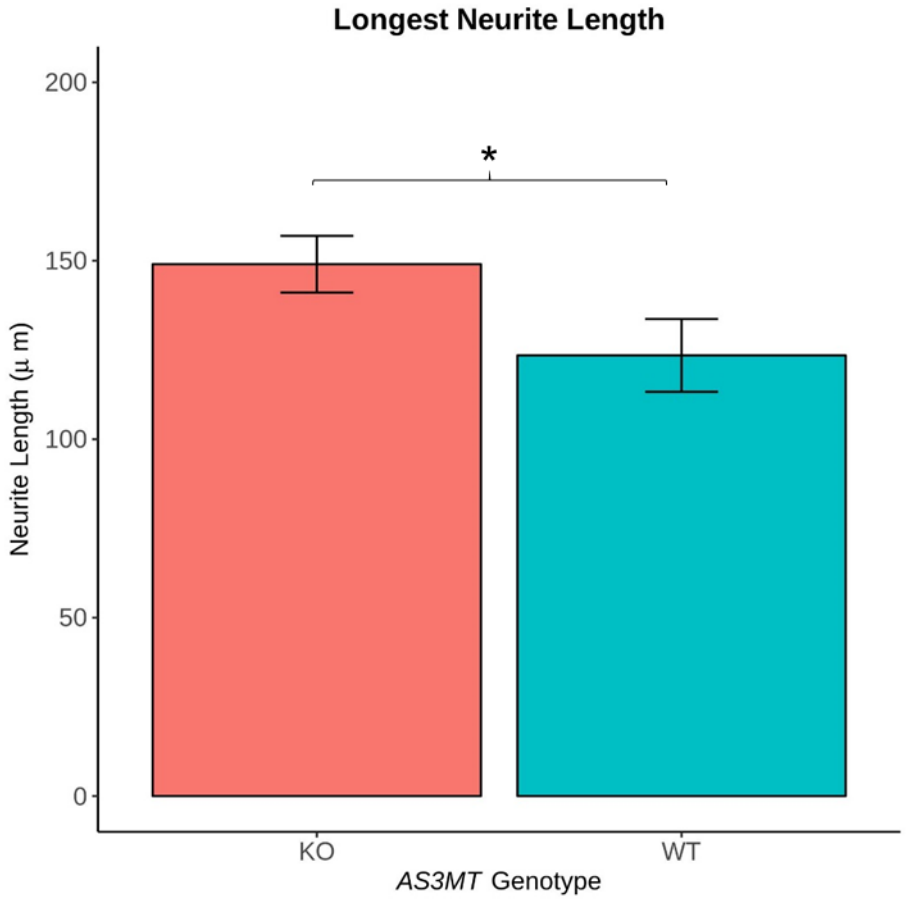


Figure 3.20: Differentiated AS3MT knockout cells show a significant increase in their longest neurite length. *Two-sample t(156.9)=1.977, p-value=0.050.*

Table 3.21: Summary statistics for longest neurite length

Genotype	n	Longest Neurite Length (µm)	SD	SE
KO	118	149.04	86.43	7.96
WT	76	123.48	88.86	10.20

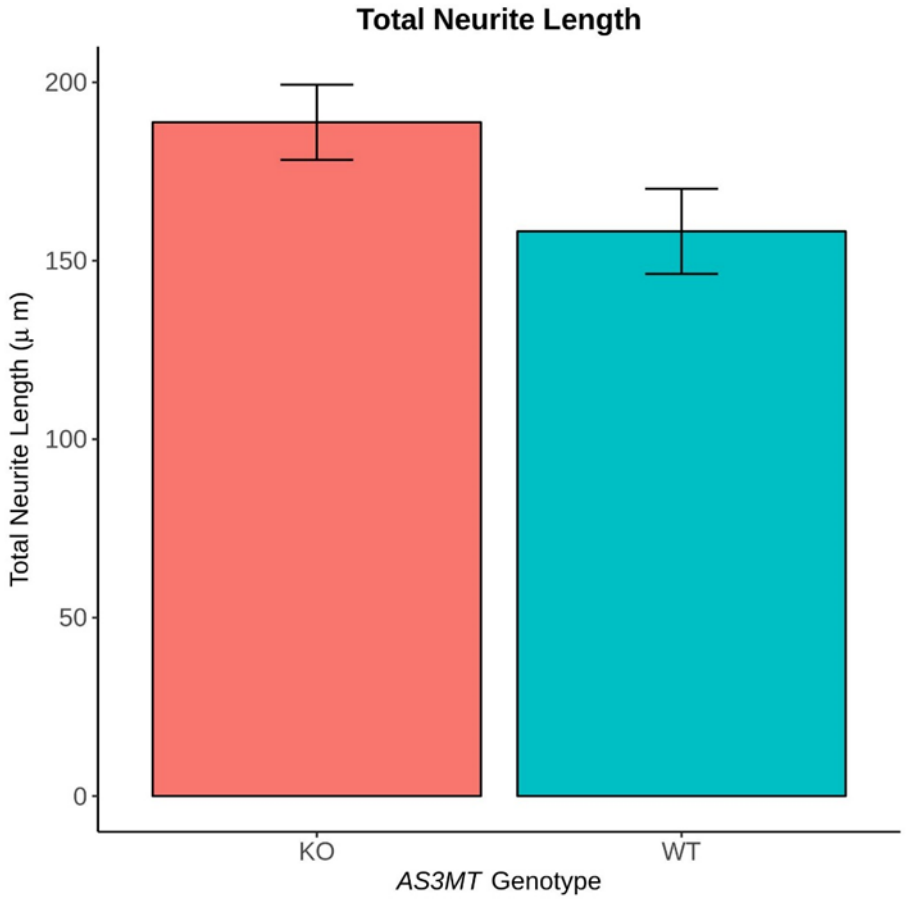


Figure 3.21: Differentiated AS3MT knockout cells show no difference in their total neurite length. *Two-sample t(171.19)=1.922, p=0.056.*

Table 3.22: Summary statistics for total neurite length

Genotype	n	Total Neurite Length (µm)	SD	SE
KO	118	188.80	114.50	10.54
WT	76	158.23	103.82	11.91

3.5. Discussion

In this chapter I first set out identify if *AS3MT* is expressed in SH-SY5Y cells and that expression does change in an isoform dependent manner throughout differentiation towards neuronal fates as previously described in iPSC cells by (M. Li *et al.*, 2016). Secondly, I set out to generate a protocol for knocking out the schizophrenia associated gene *AS3MT* in SH-SY5Y cells using CRISPR-Cas9, at the time there was limited literature using SH-SY5Y cells as a cell line for CRISPR-Cas9 experiments. Using CRISPR-Cas9 I generated three monoclonal *AS3MT* homozygous knockouts, three monoclonal *AS3MT* deletion heterozygotes, and have collected three monoclonal matched control cell lines. Finally, I characterised morphological changes using MAP2 staining, a marker of neuronal cells, resulting from *AS3MT* knockout in differentiated SH-SY5Y cells compared to matched controls. Further work in this thesis examines the transcriptional changes which result from knockout of *AS3MT* (See Chapter 4.).

I have successful identified that *AS3MT* is expressed in SH-SY5Y cells and that during neuronal differentiation with RA there is differential isoform expression of *AS3MT.* I observed a significant 2.5 fold increase in expression of the $AS3MT^{d2d3}$ isoform between pre-differentiation (zero days) and two days post differentiation commencing (ANOVA, $F(4,10)=4.82$, p-value=0.02, n=3), these levels then remain stable through differentiation over the eight days. We do not see any change in the expression of the $AS3MT^{full}$ isoform through differentiation (ANOVA, $F(4,10)=1.49$, p-value=0.28, n=3). These results support the previous evidence in iPSC cell lines (M. Li *et al.*, 2016) that $AS3MT^{d2d3}$ expression is involved with neural differentiation whereas the $AS3MT^{full}$ is not. However, the increase in expression at two days with the stability through days two to eight means expression could be a response to differentiation agents, in order to test this hypothesis other SH-SY5Y differentiation agents could be used such as brain derived neurotrophic factor (BDNF) or a combination of RA and BDNF (Encinas *et al.*, 2000).

The second aim was to use CRISPR-Cas9 in the SH-SY5Y cell line to generate homozygous knockout cell lines of *AS3MT* to elucidate any neuronal phenotype. I used an adapted protocol optimised for HEK293 cells generated by (Hsu, Lander and Zhang, 2014). Following successful design of gRNA I aimed to delete a 1.5kb region encompassing exons 4-6 of *AS3MT* and, more crucially, the main methyltransferase

domain in SH-SY5Y cells. Following successful monoculturing of cells which had up-taken both gRNA-CRISPR plasmids I identified three monoclonal knockout cell lines of *AS3MT* and three heterozygous knockout cell lines of *AS3MT* through PCR based approaches. Full protein knockout of *AS3MT* was confirmed through western blotting and genomic sequences confirmed through Sanger sequencing. In this chapter I have identified that the SH-SY5Y cell line can be used in CRISPR experiments for the generation of models for use in psychiatric research.

Neurite outgrowth has been shown to be dysregulated in schizophrenia (Bellon, 2007). *DISC1*, one of the most well-known schizophrenia genes, is involved in neurite outgrowth (Kamiya *et al.*, 2006; Lee *et al.*, 2015), it regulates neurogenesis (Kim *et al.*, 2012), and when knocked down in mouse ES cells causes increased in neurite outgrowth (Ren *et al.*, 2016), although in human iPSCs the opposite was observed (Srikanth *et al.*, 2018). As deficits in neurite outgrowth and integrity are impaired in schizophrenia I set out to examine if there are any changes in the neurite length, and cell size of differentiated knockout *AS3MT* SH-SY5Y cells using MAP2 staining, a known marker of mature neurons (Harada *et al.*, 2001). Following differentiation for seven days towards a neuronal like phenotype the cell size of MAP2 positive mature neurons in the $AS3MT^{-/-}$ lines was an average of 1500µm^2 larger than that of the control lines (two-sample t(189.43)=5.369, p-value=2.30E-07). The larger cell size might make it more difficult for the cell to migrate during development, as large structural changes occur during the earliest stages of development (Stiles and Jernigan, 2010). Or larger cells could result in a reduction in the number of cells in a given space, which could correspond to the neuronal loss seen in schizophrenics (Andreasen *et al.*, 2011). The increase in cell size could further result in dysregulated neural connections and an enhancement of specific neuronal signals which could be detrimental. Analysis of the neurite lengths from the cell lines identified an average 25µm increase in length of the longest neurite in the $AS3MT^{-/-}$ cell lines (two-sample t(156.9)=1.977, p-value=0.050) but no significant difference was found in the total neurite length between $AS3MT^{-/-}$ and $AS3MT^{+/+}$. Neurite length is important during neurodevelopment as neurones need to communicate with each other, if these connections are impaired this could result in impaired neuronal development. These results indicate that *AS3MT* plays a role in neurite outgrowth in the SH-SY5Y cellular model indicating that *AS3MT* is involved in neurodevelopment.

There are some important caveats to working with SH-SY5Y cells in CRISPR pipelines.

1) SH-SY5Y cells tend to spontaneously differentiate at higher passages, as a result the *AS3MT*$^{-/-}$ cell line C6 and *AS3MT*$^{+/+}$ cell line D3 were not used for morphological analysis as these had already undergone spontaneous differentiation. This is due to the length of time the cells spend in culture following cell sorting and generating monoclonal lines, therefore it would be imperative for the SH-SY5Y cells to be of the earliest passage possible before commencing work to generate knockout cell lines.
2) There are also concerns that FACS sorting used in this protocol damages cells and changes their cellular metabolome (Nun *et al.*, 2019) and REDOX state (Llufrio *et al.*, 2018). In order to minimise these effects we also FACS sorted SH-SY5Y cells which were not edited to expose them to the same stressors and time in culture, currently this is the best method in order to get around these limitations.
3) The PCR based approach for genotyping used in this chapter is not specific enough to detect deletions at single gRNA sites, only the whole 1.5kb deletion can be detected. It is possible that the three heterozygous KO cell lines that appeared heterozygous based on the PCR assay could in fact still be KO as the Cas9 could have caused disruption of sequence at one of the gRNA sites resulting in indels and a full knockout. There are various other methods used to genotype CRISPR edited cells these further including TA cloning (Hsu, Lander and Zhang, 2014), T7 endonuclease assays (Hoon *et al.*, 2014), and high resolution melt analysis (Montgomery *et al.*, 2007), however as I was only interested in homozygous knockouts these approaches were not used.
4) One of the largest caveats of CRISPR based approaches is the risk of off target mutations. One of the best ways to combat this is through stringent guide design approaches. Specificity of gRNA is dependent on the sequence 10-12 bp proximal of the PAM (Hsu *et al.*, 2013), GC content (Tsai *et al.*, 2015) and length (Sugano *et al.*, 2018). Many algorithms are now available within software's to provide on-target and off-target scores. There is always a risk of off target effects, and until we understand more about the CRISPR/Cas9

pathways, the easiest way to reduce off target effects is with stringent gRNA design.

3.6. Conclusion

I have identified that *AS3MT* isoforms are differentially expressed throughout SH-SY5Y differentiation towards neuronal fates, supporting previous evidence. I have also identified that knockout of *AS3MT* results in increased cellular size and longest neurite length in differentiated SH-SY5Y cells. This provides us with a functional link between how *AS3MT* could contribute so schizophrenia and other psychotic disorders. This chapter has also shown that the SH-SY5Y neuroblastoma cell line can be used for CRISPR-Cas9 based approached of genetic engineering to functionally validate genetic risk variants in psychiatric diseases, providing cellular models for functional analysis and validation.

Chapter 4. Transcriptomic profiling of the *AS3MT* knockout cell lines

4.1. Introduction

Recently RNA-sequencing (RNAseq) has replaced gene expression microarrays for profiling the transcriptome. The transcriptome refers to the complete set of transcripts expressed within a cell and their quantity at a given developmental time point. The ability to catalogue the transcriptome is vital in identifying functional elements of the genome and understanding the role of gene regulation and expression in different diseases and conditions. RNAseq for strand specific sequencing was first described by (Lister *et al.*, 2008) in *Arabidopsis thaliana,* and was first described in mammals, specifically mouse tissue, by (Mortazavi *et al.*, 2008). RNAseq not only allows for quantification of gene expression levels but also allows profiling and discovery of new transcripts. This provides it with a distinct advantage over previous technologies such as the microarray which are limited to known transcripts. As well as gene level quantification, RNAseq allows for transcript and isoform level quantification, as long as the sequencing depth is sufficient (Mortazavi *et al.*, 2008). Briefly, to perform an RNAseq experiment, whole RNA from a sample is converted into cDNA, which is then amplified and adapters incorporated which allows for high-throughput analysis as each read can be traced to its original sample. For the experiments in this chapter, the Illumina HiSeq 2500 system was used, which uses sequencing by synthesis technology to generate reads (further information is provided in **4.3.**). Once the reads are acquired they are then mapped to the reference human genome and quantified to obtain the number of reads corresponding to mRNA from each known gene, exon, splice event, or candidate gene, which can then be compared between samples (Mortazavi *et al.*, 2008). There are a multitude of different RNAseq analysis pipelines available and their utility is dependent on the research question (Conesa *et al.*, 2016).

In this chapter I describe my analysis of transcriptional changes associated with knockout of *AS3MT* in two of the cell lines I generated in Chapter 3. using highly-parallel RNAseq. For this analysis, two *AS3MT* knockout cell lines were sequenced alongside two experimentally matched *AS3MT* wild type cell lines, all reactions were performed at least in triplicate. Two independent sequencing runs were carried out, and the differentially expressed genes (DEGs) were compared between the two sequencing runs to identify robust changes in the *AS3MT* knockout lines. Finally, I used systems-level analysis to identify whether or not specific functional pathways have been altered through knockout of *AS3MT.*

4.2. Aims

Little is known regarding the postulated function of the schizophrenia associated gene *AS3MT* in the developing brain. To elucidate downstream genes and gene pathways involved in brain expressed *AS3MT* I undertook transcriptional profiling of *AS3MT* knockout neuronal cell lines. My aims were to identify what cellular pathways are altered in neuronal cells following knockout of *AS3MT* using CRISPR-Cas9 as described in Chapter 3. . This would help to identify the function of *AS3MT* in neuronal cells, independent of arsenic metabolism, and try to elucidate how disruption of normal *AS3MT* expression could contribute to schizophrenia. To do this I:

1. Examined gene expression changes between *AS3MT*$^{-/-}$ cell lines and control *AS3MT*$^{+/+}$ cell lines and identified biological pathways altered through gene ontology (GO) analysis.
2. Examined genes which are upregulated in the *AS3MT*$^{-/-}$ cell lines and identified pathways which are upregulated through knockout.
3. Examined genes which are downregulated in the *AS3MT*$^{-/-}$ cell lines and identified biological pathways which are downregulated due to knockout.

4.3. Methods

4.3.1. RNA extraction

Two of the AS3MT knockout SH-SY5Y cell lines (B6 and C6) and two wildtype control SH-SY5Y cell lines (D2 and D4) generated in Chapter 3. were cultured in 6 well plates until 80% confluent. Total RNA was extracted from cultured cells as described in **2.3.1.** and total RNA was isolated using the Zymo Directzol Miniprep kit (Cambridge Biosciences, Cambridge, UK) as described in **2.3.3.** . RNA quality, purity and integrity was assessed following isolation using an Agilent 2200 TapeStation (Agilent Technologies, California, USA) as outlined in **2.3.4.2.** . **Table 4.1** gives an overview of the RNA concentration and RIN results from the Agilent 2200 TapeStation. Details on these measures can be found in **2.3.4.** .

Table 4.1: RNA quantity and integrity of the RNA samples extracted before RNA sequencing.

Genotype	Cell Line	Sample	RIN	Concentration (ng/µl)	Sequenced
AS3MT$^{-/-}$	B6	1	8.6	29.1	Yes
		2	8.7	32.7	Yes
		3	8.3	30.2	No
		4	9.0	45.3	Yes
		5	8.8	34.1	Yes
		6	9.1	42.7	Yes
	C6	1	9.1	39.1	No
		2	9.3	30.7	Yes
		3	9.9	24.7	Yes
		4	9.9	27.1	Yes
		5	9.6	13.1	No
		6	10.0	17.5	No
AS3MT$^{+/+}$	D2	1	9.6	22.7	Yes
		2	9.8	26.7	Yes
		3	7.6	18.6	No
		4	8.7	23.2	Yes
		5	8.8	31.1	Yes
		6	8.3	20.6	Yes
	D4	2	8.9	33.5	Yes
		3	9.5	11.6	No
		4	7.2	46.6	No
		5	9.5	28.4	Yes
		6	9.4	21.3	Yes

4.3.2. Complementary DNA libraries preparation and RNA sequencing

Following on from quality checks using the Agilent 2200 Tapestation, five *AS3MT* KO B6 and five D2 WT samples with the highest RIN were selected for the primary RNAseq experiment. For the validation RNAseq experiment, three *AS3MT* KO C6 and three D2 WT samples with the highest RIN were selected. **Table 4.1** provides details about the samples selected for sequencing.

cDNA libraries were generated from 500ng of the selected RNA samples using the TruSeq DNA HT Library Preparation Kit using the 3' poly-A tail primer Oligo(dT) from (Illumina, San Diego, California, USA). Library preparations and sequencing were carried out by Audrey Farbos of the Exeter Sequencing Service following their in-house method of library preperation and sequencing.

Sequencing of the libraries was performed by the Illumina Hiseq 2500 Ultra-High-Throughput Sequencing System (Illumina, San Diego, California, USA), which performs next generation sequencing (NGS). NGS allows for parallel sequencing of millions of DNA molecules and is an adaptation of Sanger Sequencing method (Sanger and Nicklen, 1977). In brief, the libraries are washed over a flow cell where the cDNA fragments are captured on a lawn of surface bound probes complementary to the adapters added during library preparation. Each fragment is then amplified through bridge amplification to form clusters of clonal molecules. Each cluster acts as an individual sequencing reaction. These clusters are then sequenced using sequencing-by-synthesis. A fluorescent bound reversible terminator dNTP is incorporated into sequence, this is then imaged and then cleaved and the process repeats with the next dNTP. Each of the four dNTPs has its own fluorescent signal, allowing for the identification of the base incorporated during sequencing. I performed paired-end sequencing, which sequences both the forward and reverse template strand of the cDNA library, this improves accuracy when aligning to the genome.

Two separate sequencing runs were carried out. The primary RNA-seq experiment resulted in 50bp sequences whereas the validation sequencing resulted in 125bp sequences. This was a result of the sequencing runs available at the time provided by the University of Exeter Sequencing Service.

4.3.3. Quality control

The raw reads were downloaded from The Exeter Sequencing Service and the quality of the raw reads (.fastq files of both read 1 and read 2 from each sample) were inspected using MultiQC software version 1.5 (Ewels, Lundin and Max, 2016). This software integrates multiple quality control (QC) tools to provide one simple report of RNA-seq data quality. Below are some of the metrics provided by MultiQC and how these were used to inspect the quality of the RNAseq data.

MultiQC incorporates three different QC software for analysing the RNA-seq data:

1. **Cutadapt:** A tool used to find and remove adapter sequences, primers, poly-A tails and other unwanted sequences from the sequencing results.
2. **FastQ**: A screen to identify the organism which the sequences have been derived from.
3. **FastQC:** A multiple analysis tool developed by Simon Andrews of the Babraham Institute to perform multiple checks of sequencing data.

4.3.4. General statistics

The general statistics table generated by MultiQC is a summary of the QC undertaken on the sequenced reads. The general statistic table for the primary RNAseq experiment is shown in **Table 4.2** and for the replication experiment in **Table 4.3**.

Table 4.2: General statistics table generated by MultiQC for the primary RNAseq experiment of the B6 and D2 AS3MT cell lines.

Sample Name	% Trimmed	% Dups	% GC	Length	% Failed	Millions of Seqs
AS3MT_KO_1_r1	0.80%	51.80%	45%	50 bp	25%	9.7
AS3MT_KO_1_r2	0.80%	51.60%	45%	50 bp	17%	9.7
AS3MT_KO_1_trimmed_r1		50.80%	45%	50 bp	25%	9.7
AS3MT_KO_1_trimmed_r2		50.10%	45%	50 bp	17%	9.7
AS3MT_KO_2_r1	0.90%	61.70%	45%	50 bp	25%	9.1
AS3MT_KO_2_r2	0.90%	59.90%	46%	50 bp	17%	9.1
AS3MT_KO_2_trimmed_r1		60.60%	45%	50 bp	25%	9
AS3MT_KO_2_trimmed_r2		58.00%	46%	50 bp	8%	9
AS3MT_KO_4_r1	0.80%	55.50%	45%	50 bp	25%	10
AS3MT_KO_4_r2	0.80%	54.80%	46%	50 bp	17%	10
AS3MT_KO_4_trimmed_r1		54.50%	45%	50 bp	17%	9.9
AS3MT_KO_4_trimmed_r2		53.20%	46%	50 bp	8%	9.9
AS3MT_KO_5_r1	1.10%	56.50%	45%	50 bp	25%	6
AS3MT_KO_5_r2	1.10%	54.10%	45%	50 bp	17%	6
AS3MT_KO_5_trimmed_r1		55.40%	45%	50 bp	25%	5.9
AS3MT_KO_5_trimmed_r2		52.50%	45%	50 bp	17%	5.9
AS3MT_KO_6_r1	1.00%	62.60%	45%	50 bp	25%	13.2
AS3MT_KO_6_r2	1.00%	60.90%	45%	50 bp	17%	13.2
AS3MT_KO_6_trimmed_r1		61.50%	45%	50 bp	17%	13.1
AS3MT_KO_6_trimmed_r2		59.20%	45%	50 bp	17%	13.1
AS3MT_WT_1_r1	0.90%	60.30%	44%	50 bp	25%	13.6
AS3MT_WT_1_r2	0.90%	58.90%	45%	50 bp	17%	13.6
AS3MT_WT_1_trimmed_r1		59.20%	44%	50 bp	17%	13.5
AS3MT_WT_1_trimmed_r2		57.40%	45%	50 bp	17%	13.5
AS3MT_WT_2_r1	1.20%	67.10%	45%	50 bp	25%	15.4

AS3MT_WT_2_r2	1.20%	63.60%	45%	50 bp	17%	15.4
AS3MT_WT_2_trimmed_r1		65.90%	45%	50 bp	25%	15.3
AS3MT_WT_2_trimmed_r2		61.80%	45%	50 bp	17%	15.3
AS3MT_WT_4_r1	1.00%	65.20%	45%	50 bp	25%	22.6
AS3MT_WT_4_r2	1.00%	63.00%	45%	50 bp	17%	22.6
AS3MT_WT_4_trimmed_r1		64.10%	45%	50 bp	25%	22.4
AS3MT_WT_4_trimmed_r2		61.30%	45%	50 bp	8%	22.4
AS3MT_WT_5_r1	1.10%	55.30%	45%	50 bp	25%	12.6
AS3MT_WT_5_r2	1.10%	53.60%	46%	50 bp	17%	12.6
AS3MT_WT_5_trimmed_r1		54.10%	45%	50 bp	25%	12.4
AS3MT_WT_5_trimmed_r2		52.50%	46%	50 bp	17%	12.4
AS3MT_WT_6_r1	1.10%	64.30%	45%	50 bp	25%	14.9
AS3MT_WT_6_r2	1.10%	62.80%	45%	50 bp	17%	14.9
AS3MT_WT_6_trimmed_r1		63.10%	45%	50 bp	25%	14.8
AS3MT_WT_6_trimmed_r2		61.30%	45%	50 bp	17%	14.8
nc1_r1	2.00%	79.90%	43%	50 bp	25%	0.1
nc1_r2	2.00%	78.00%	43%	50 bp	17%	0.1
nc1_trimmed_r1		79.20%	43%	50 bp	33%	0.1
nc1_trimmed_r2		77.60%	43%	50 bp	17%	0.1

Table 4.3: General statistics table generated by MultiQC for the validation RNAseq experiment on the C6 and D4 cell AS3MT cell lines.

Sample Name	% Trimmed	% Dups	% GC	Length	% Failed	M Seqs
2907_AS3MT_KO_C6_1_r1	4.40%	61.70%	49%	125 bp	33%	26.9
2907_AS3MT_KO_C6_1_r2	4.40%	62.20%	49%	125 bp	25%	26.9
2907_AS3MT_KO_C6_1_trimmed_r1		61.10%	49%	121 bp	25%	26.4
2907_AS3MT_KO_C6_1_trimmed_r2		61.70%	49%	121 bp	25%	26.4
2907_AS3MT_KO_C6_3_r1	3.60%	61.20%	48%	125 bp	25%	25.6
2907_AS3MT_KO_C6_3_r2	3.60%	61.10%	48%	125 bp	25%	25.6
2907_AS3MT_KO_C6_3_trimmed_r1		60.80%	48%	121 bp	25%	25.3
2907_AS3MT_KO_C6_3_trimmed_r2		60.90%	48%	121 bp	33%	25.3
2907_AS3MT_KO_C6_4_r1	3.60%	60.50%	49%	125 bp	25%	25.3
2907_AS3MT_KO_C6_4_r2	3.60%	61.70%	49%	125 bp	25%	25.3
2907_AS3MT_KO_C6_4_trimmed_r1		60.10%	49%	121 bp	25%	25
2907_AS3MT_KO_C6_4_trimmed_r2		61.40%	49%	121 bp	33%	25
2907_AS3MT_WT_D4_2_r1	4.90%	60.10%	49%	125 bp	33%	25.6
2907_AS3MT_WT_D4_2_r2	4.90%	60.70%	49%	125 bp	42%	25.6
2907_AS3MT_WT_D4_2_trimmed_r1		59.30%	49%	121 bp	25%	25
2907_AS3MT_WT_D4_2_trimmed_r2		60.20%	49%	121 bp	33%	25
2907_AS3MT_WT_D4_5_r1	3.80%	60.50%	49%	125 bp	25%	23.7
2907_AS3MT_WT_D4_5_r2	3.80%	61.50%	49%	125 bp	33%	23.7
2907_AS3MT_WT_D4_5_trimmed_r1		60.10%	49%	121 bp	25%	23.4
2907_AS3MT_WT_D4_5_trimmed_r2		61.20%	49%	121 bp	33%	23.4
2907_AS3MT_WT_D4_6_r1	4.00%	61.10%	49%	125 bp	25%	27.2
2907_AS3MT_WT_D4_6_r2	4.00%	61.30%	49%	125 bp	33%	27.2
2907_AS3MT_WT_D4_6_trimmed_r1		60.60%	49%	121 bp	25%	26.8
2907_AS3MT_WT_D4_6_trimmed_r2		61.00%	49%	121 bp	33%	26.8

4.3.4.1. Cutadapt

During library preparation adaptor sequences are added at the 3' end of each molecule of interest. These usually are barcodes to identify the samples, allow sequencing primers to bind, and enable multiplexing. Consequently, the reads outputted from sequencing usually include these adaptor sequences, therefore it is paramount that these sequences be removed from the reads as only the sequences of interest (mRNA) are required for the downstream analysis. Inclusion of the adaptor sequences could result in false alignments, an increase in unassigned reads and false negative data. In order to remove these adaptor sequences the .fastq files were run through the cutadapt pipeline, a simple command line tool used to identify and remove adaptor sequences (Martin, 2011). All .fastq files were run through cutadapt to remove any adaptor sequences and the lengths trimmed and respective counts are shown in **Figure 4.1**.

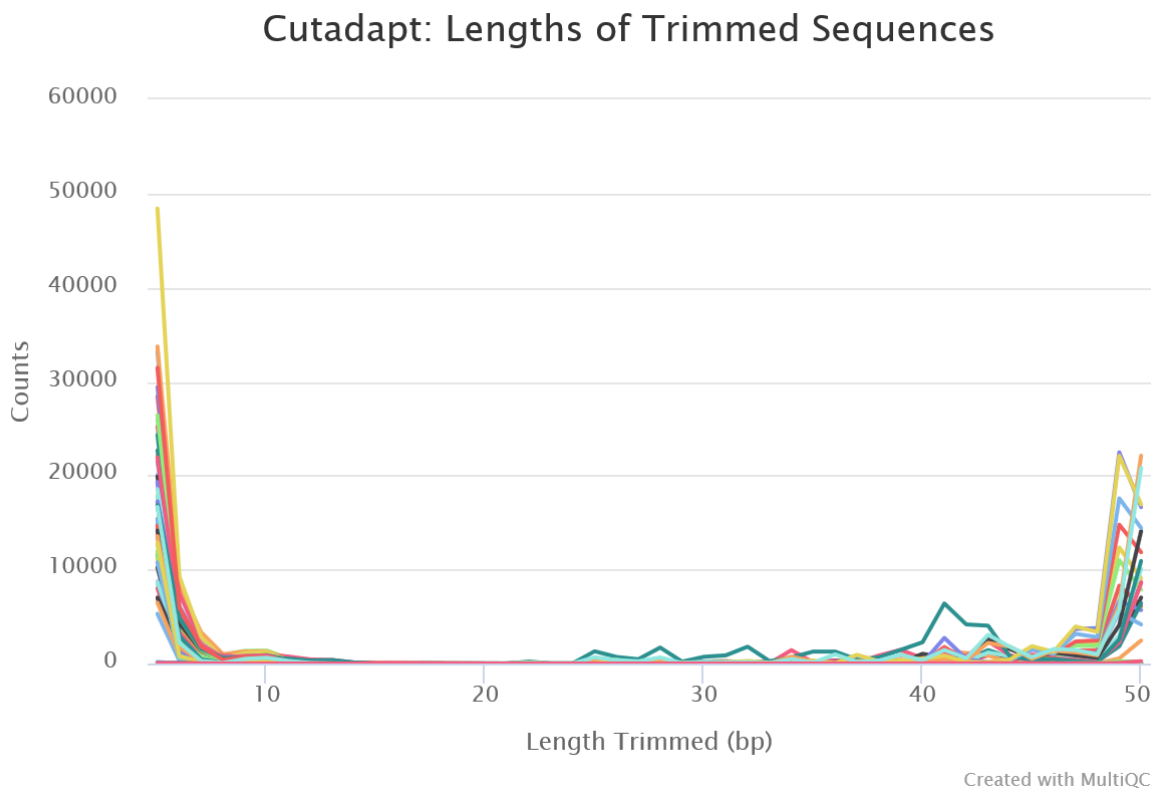


Figure 4.1: Lengths of sequence trimmed by Cutadapt and counts in each sample. *The highest number of counts are within 1-10bp and towards the end of the sequence.*

4.3.4.2. FastQ screen

In order to check for contamination of the samples, the percentage of reads mapped uniquely to the human genome, to other species, or to human ribosomal RNA were calculated. To do this the FastQ Screen software was used (Babraham Institute, 2015), which allows the screening of sequences against a database. The result from the primary RNAseq experiment FastQ screen is shown in **Figure 4.2**.

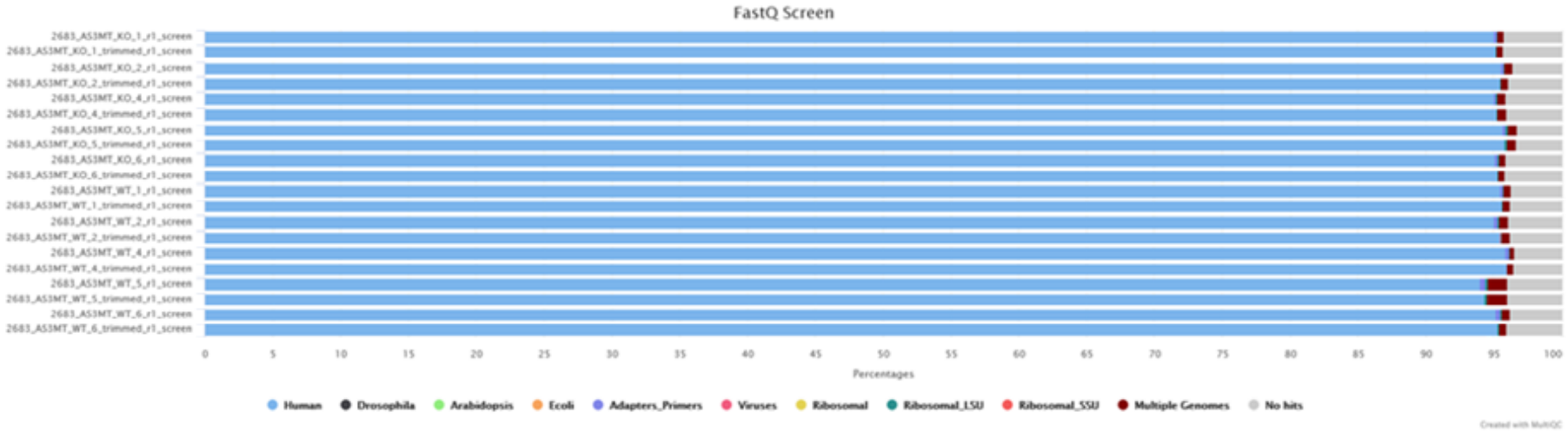


Figure 4.2: The output of the FastQ screen from the primary RNAseq experiment shows that the approximately 95% of the sequence reads map to the human genome, ensuring that the correct sequences have been sequenced.

4.3.4.3. FastQC

The final inspection of the QC pipeline involves detailed inspection of the raw and trimmed reads (both read 1 and read 2 in their .fastq file format) using FastQC software (Andrews, 2010). The software is modular and can provide in depth analysis for high throughput sequencing results. Below are some of the modular analyses in the FastQC software I used to assess the quality of the reads.

4.3.4.4. Phred quality score

The Phred score is usually the most important QC check for sequencing data. Through sequencing a Phred score is applied to each base of the read and is a marker of per base sequencing quality (Ewing and Green, 1998). The Phred score tells us the likelihood that the base called in the sequence is incorrect (**Table 4.4**). The Phred scores are given as a plot with an overview of the range of the Phred scores at an individual base of each sequence read. In **Figure 4.3** shown below, are the Phred scores for all strand 1 reads of sample AS3MT_WT_1. A score of above 30 is considered good and above 20 is acceptable, below 20 is considered inadequate and these reads would be discarded through trimming. The Phred scores peaked at Q39 indicating that the base call accuracy is 99.98% accurate and the reads are very high quality (**Figure 4.4**). All samples had a Phred score no lower than 25.

Table 4.4: Phred Quality Scores and their interpretation.

Phred Quality Score	Probability base incorrect	Base call accuracy
10	1 in 10	90%
20	1 in 100	99%
30	1 in 1000	99.90%
40	1 in 10,000	99.99%

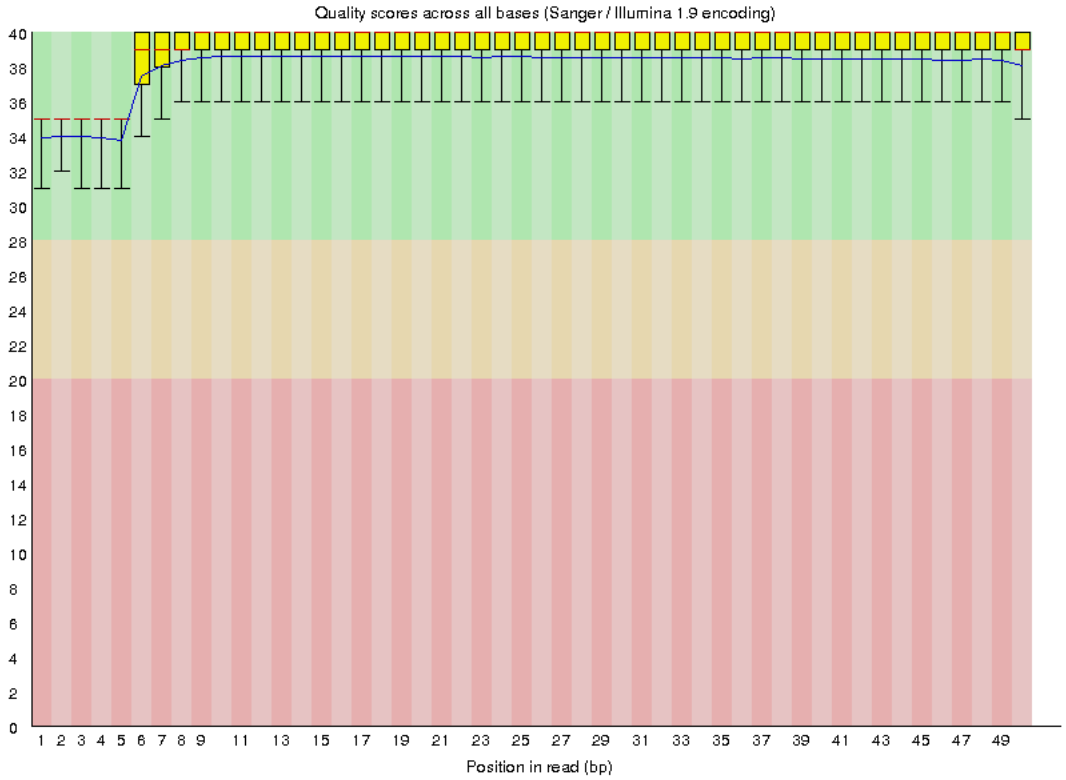


Figure 4.3: An example Phred quality score plot. *All the Phred scores are above 34, with a lower quality for the first five bases. The red line indicated median phred score over the reads, yellow boxes are the inter-quartile range, upper and lower whiskers represent the 10th and 90th percentile respectively and the blue line represents mean quality.*

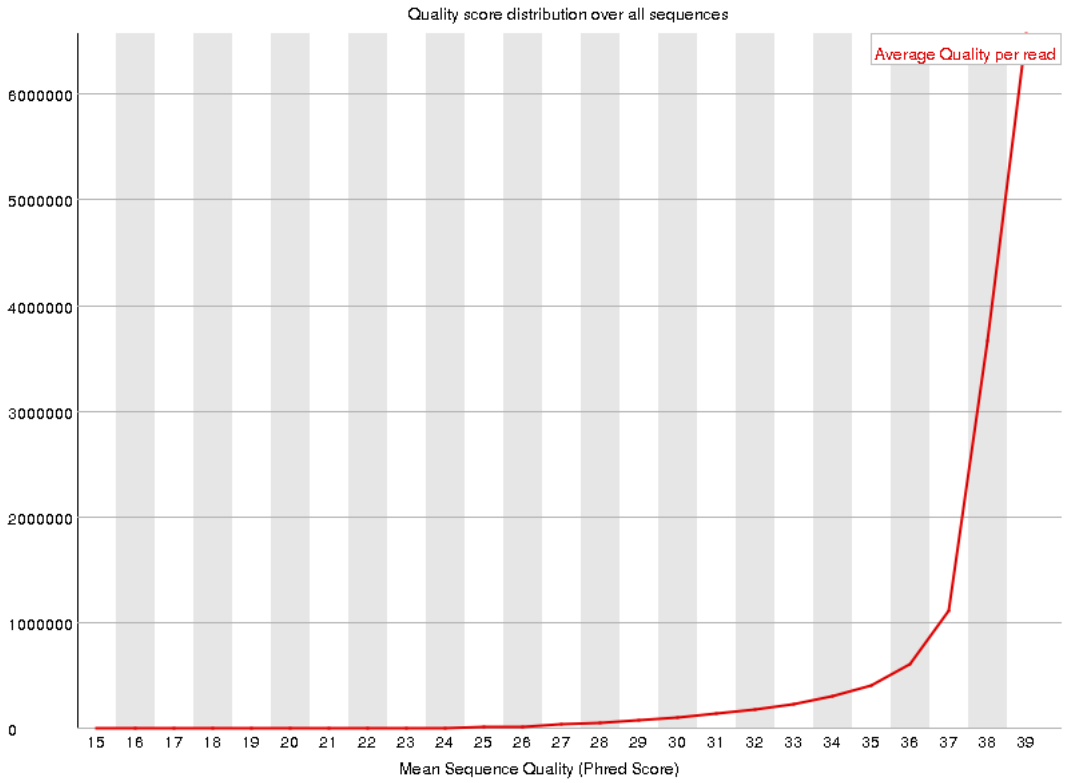


Figure 4.4: Example per sequence quality. *The Y Axis is read count, X axis is the Phred Score.*

4.3.4.4.1. Per Sequence quality scores

In an ideal RNAseq experiment there should be no variation in the bases called along the sequences. In order to do this I looked at the per base sequence content across the reads, which plots out the proportion of each base (A, C, T, or G) called at each position in a read. An example Per Base Sequence Content plot is shown in **Figure 4.5**. In our library there should be no difference between the different bases in the sequencing run, however the first 12 bases show some variation, this is due to the random hexamer primers used to produce the libraries (Andrews, 2010). This also explains the lower Phred scores towards the beginning of the sequences in **Figure 4.3**.

There should also be no variation in the GC content along the sequencing reads. In **Figure 4.6,** the GC content is shown for strand 1 reads of AS3MT_WT_1 in red, along with the theoretical distribution in blue. The distribution should be normally distributed if it is good sequencing data and match the theoretical distribution. If the curve has a high shoulder on the right side (indicating high GC content) this shows the sample is contaminated with ribosomal RNA (rRNA) or another species' RNA.

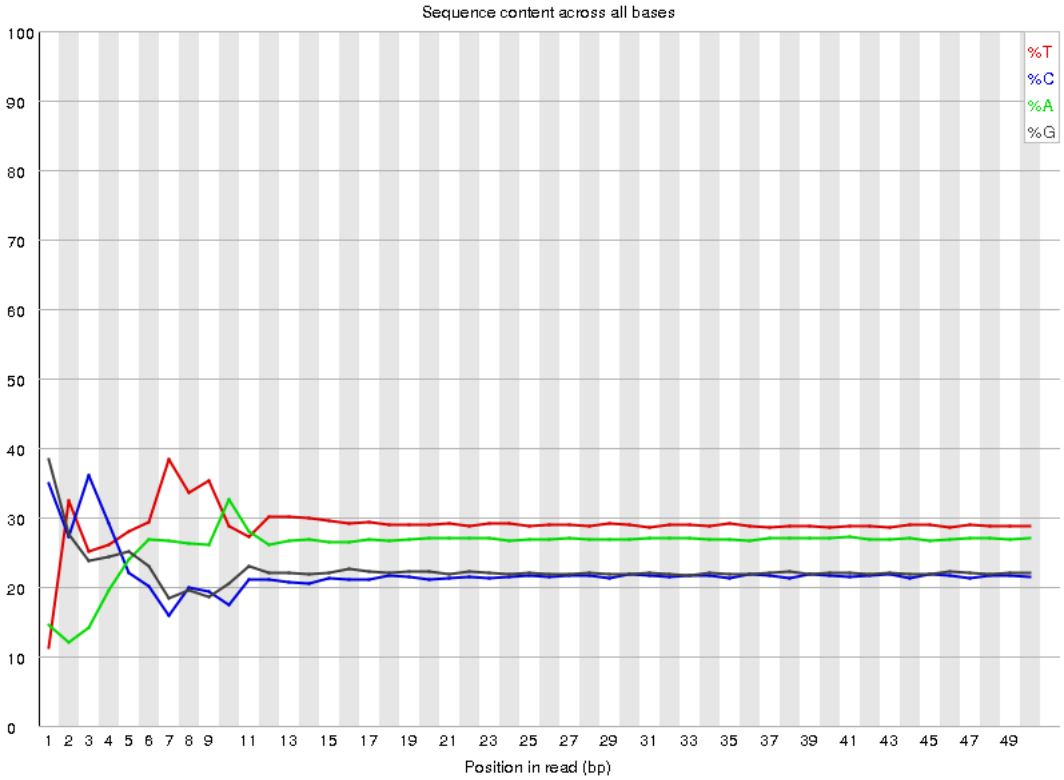


Figure 4.5: Example of a "Per Base Sequencing Content" plot. *The y-axis shows the base call content (%) along the length of the read (x-axis).*

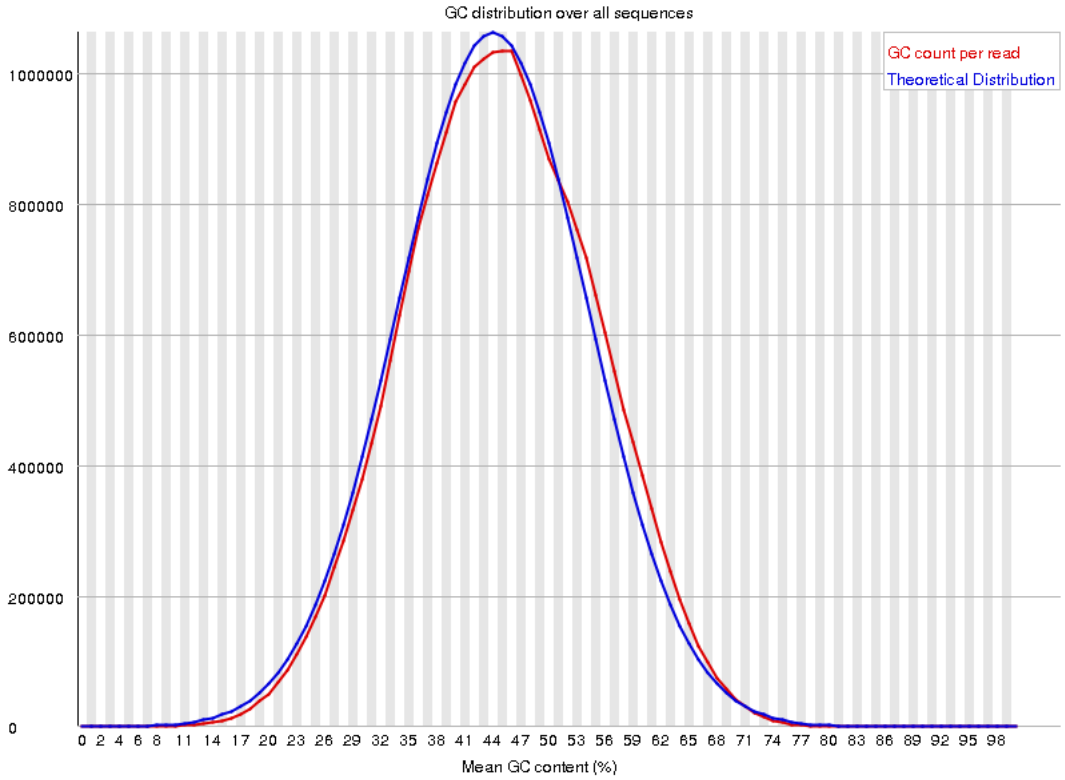


Figure 4.6: Example of a "Per Sequence GC Content" plot given for read 1 reads of AS3MT_WT_1. *The x-axis shows the mean GC Content (%) against the read count (y-axis). The theoretical distribution is shown in blue with the experimental sample shown in red. The high correlation indicates the sample is of high quality.*

4.3.4.4.2. Sequence duplication levels

The final check through FastQC is to check the sequence duplication levels. The plot represents the levels of duplicated sequences within the library. There is one limitation, in that FastQC assumes uniform coverage as it was developed for genomic DNA sequencing rather than RNA sequencing. In genomic sequences if a region has higher duplication this could indicate enrichment bias, but in RNAseq genes which are highly expressed have higher coverage and therefore more reads will be duplicated. **Figure 4.7** shows the sequence duplication levels for strand 1 reads of AS3MT_WT_1, the y-axis is percentage of sequences against the x-axis which is the sequence duplication level. This data is shifted to the left indicating the library is properly diverse, spikes to the right hand side indicate enrichments of subsets of data (Andrews, 2010).

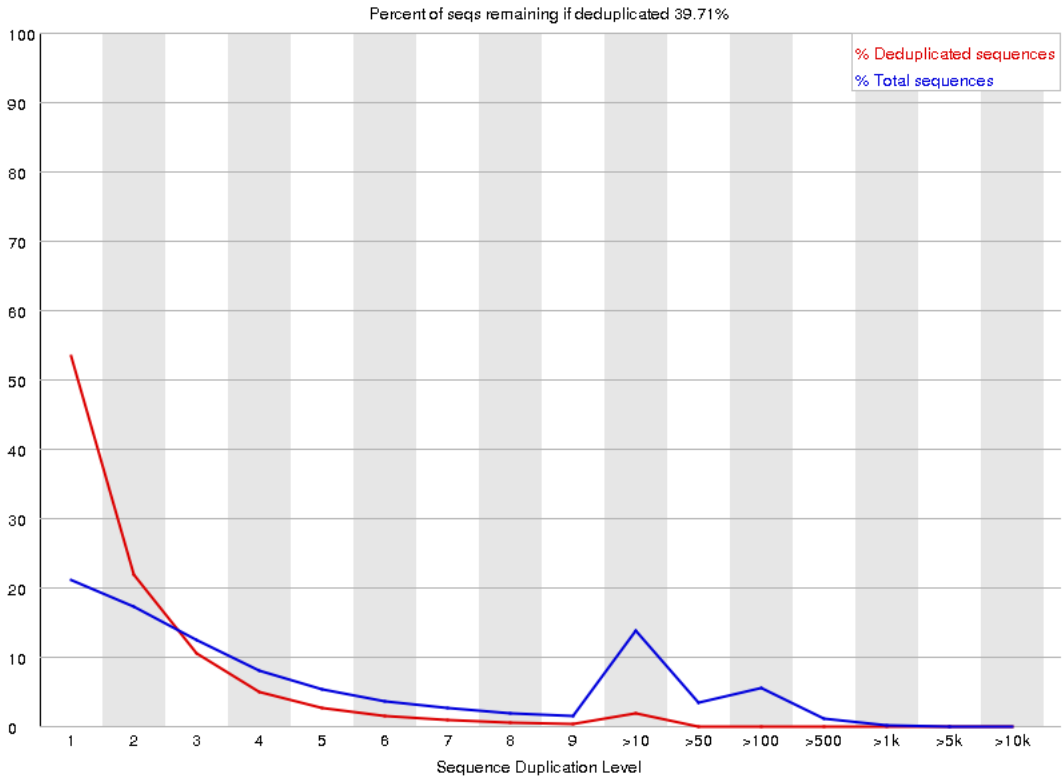


Figure 4.7: Example of a Sequence Duplication Levels plot. *The y-axis is percentage of sequences against the x-axis which is sequence duplication level. The blue line shows the distribution of the duplicated sequences and the red line plots sequences which are de-duplicated and the proportions shown are the proportions of the de-duplicated sets which come from different duplication levels in the original data.*

4.3.5. Aligning reads to the human reference genome using STAR

Before quantifying gene expression levels the reads from the sequencing experiment need to be mapped (or aligned) to a reference genome. Mapping to the genome has several benefits, it allows for identification of *de novo* transcripts and for in depth splicing analysis (Conesa *et al.*, 2016). The first part of analysis in this chapter was to identify DEGs between the *AS3MT*$^{-/-}$ and *AS3MT*$^{+/+}$ cell lines, therefore aligning to the latest human genome build was undertaken.

Alignment was performed using the Spliced Transcripts Alignment to a Reference (STAR) aligner which is publicly available (Dobin *et al.*, 2013). STAR has a built in function to run and create a reference genome index from publicly available data. Firstly, the latest human genome assembly sequence at the time (GRCh38.p12) was downloaded from GENCODE (https://www.gencodegenes.org/human/release_29.html). This included the nucleotide sequence of GRCh38.p12, reference chromosomes, scaffolds, patches, and haplotypes, to create the most up to date and detailed reference genome (Harrow *et al.*, 2012). The corresponding list of genomic features in Gene Transfer Format (GTF) (.gtf) for the human genome build GRCh38.p12 were also downloaded to provide gene names and transcript names from GENCODE. These files were compiled and used for alignment and gene annotation. To align my sequenced I used the parameters described below:

- Generating the reference genome

```
STAR --runThreadN --runMode genomeGenerate --genomeDir --genomeFastaFiles --sjdbGTFfile
```

This commands STAR to run in mode `--genomeGenerate` to create the reference genome index in the file (`--genomeDir`) from the supplied human genome GRCh38.12 .fasta files from gencode (`--genomeFastaFiles`) and the corresponding .gtf annotation file (`--sjdbGTFfile`). The `–runThreadN` option allows for multiple computer processors to be used to speed up the function.

- Aligning RNA-seq reads to the reference genome

```
STAR --runThreadN --genomeDir --readFilesIn --readFilesCommand --outFileNamePrefix --outSAMtype
```

This commands STAR to run in the alignment setting to align the paired trimmed reads from a samples together (`--readFilesIn`). The `–genomeDir` option tells STAR where the reference genome generated from `–genomeGenerate` is stored. The `–readFilesIn` are the trimmed .fastq file read 1 reads and read 2 reads (as the reads are paired-end) and the `–readFilesCommand` allows the files to be unzipped from their compressed state before alignment. The `–outFileNamePrefix` allows for the aligned file to be named. The final option (`--outSAMtype`) allows the aligned output file format to be selected, in this case the output was selected to be a BAM file with reads sorted by coordinate.

Table 4.5 presents the number of reads mapped by STAR for each sample, along with the number (and percentage) of uniquely mapped reads, number (and percentage) of multi-mapping reads, and the number (and percentage) of unmapped reads. It is important to note that after mapping reads to a reference genome they fall into three broad categories:

- Unique Reads: Reads which are mapped to only one position on the reference library.
- Multi-mapping Reads: Reads which are mapped to multiple positions on the reference library.
- Unmapped Reads: Reads which do not map at any position of the reference library.

In RNA-seq experiments mapping to the reference genome results in multi-mapping reads due to repetitive sequences or paralogous genes (Conesa *et al.*, 2016). Reads which are unmapped mean that these are not present in the reference genome, these could be reads from the library which come from other organisms such as bacteria or viruses. The mapping results from both experiments are outlined in **Table 4.5** and **Table 4.6**.

The output from STAR is a .bam file for each sample with the aligned reads organised by their coordinate in the reference genome. This .bam file can then be used in downstream processes including visualisation of the alignment in Integrative Genomics Viewer (IGV) (Robinson *et al.*, 2011; Thorvaldsdóttir, Robinson and Mesirov, 2013) and for calculating gene expression through counting reads mapped to a particular locus using FeatureCounts (Liao, Smyth and Shi, 2014).

Table 4.5: The output file from STAR when the sequenced reads from the primary RNAseq experiment are mapped to the human genome. *Over 90% of the reads are uniquely mapped to the human genome, indicating the alignment has worked correctly.*

		Unique Reads			Multi-mapping Reads				Unmapped Reads	
Sample	**Number of input reads**	**Uniquely mapped reads number**	**Uniquely mapped reads (%)**	**Average mapped length**	**Number of reads mapped to multiple loci**	**% of reads mapped to multiple loci**	**Number of reads mapped to too many loci**	**% of reads mapped to too many loci**	**% of reads unmapped: too short**	**% of reads unmapped: other**
AS3MT_KO_1	9653399	8928001	92.49%	99.39	525987	5.45%	3716	0.04%	1.98%	0.04%
AS3MT_KO_2	8997656	8296882	92.21%	99.42	532088	5.91%	3330	0.04%	1.80%	0.04%
AS3MT_KO_4	9923261	9158010	92.29%	99.42	569812	5.74%	3778	0.04%	1.89%	0.04%
AS3MT_KO_5	5946333	5419252	91.14%	99.48	391702	6.59%	1831	0.03%	2.21%	0.03%
AS3MT_KO_6	13095595	12054637	92.05%	99.42	722496	5.52%	5704	0.04%	2.35%	0.04%
AS3MT_WT_1	13482261	12527506	92.92%	99.43	702463	5.21%	4626	0.03%	1.80%	0.03%
AS3MT_WT_2	15279600	14049679	91.95%	99.47	787985	5.16%	5728	0.04%	2.82%	0.04%
AS3MT_WT_4	22391519	20759970	92.71%	99.47	1154561	5.16%	7789	0.03%	2.06%	0.04%
AS3MT_WT_5	12438799	11293753	90.79%	99.51	768515	6.18%	4191	0.03%	2.96%	0.03%
AS3MT_WT_6	14796052	13683603	92.48%	99.38	782871	5.29%	5524	0.04%	2.15%	0.04%

Table 4.6: The output file from STAR when the sequenced reads from the replication RNAseq experiment are mapped to the human genome. *Over 90% of reads are uniquely mapped to the human genome indicating the alignment has worked correctly.*

Sample	Number of input reads	Unique Reads			Multi-mapping Reads				Unmapped Reads	
		Uniquely mapped reads number	Uniquely mapped reads (%)	Average mapped length	Number of reads mapped to multiple loci	% of reads mapped to multiple loci	Number of reads mapped to too many loci	% of reads mapped to too many loci	% of reads unmapped: too short	% of reads unmapped: other
AS3MT_KO_1	26414673	24259483	91.84	243.15	1367498	5.18	4830	0.02	2.95	0.01
AS3MT_KO_3	25278352	23277632	92.09	243.43	1181399	4.67	3873	0.02	3.21	0.01
AS3MT_KO_4	24964717	23034578	92.39	243.82	1217355	4.88	4198	0.02	2.71	0.01
AS3MT_WT_2	25045889	22749670	90.83	242.96	1280106	5.11	4720	0.02	4.03	0.01
AS3MT_WT_5	23363725	21388322	91.54	243.96	1247307	5.34	3998	0.02	3.09	0.01
AS3MT_WT_6	26783400	24699613	92.22	242.99	1275724	4.76	4571	0.02	2.99	0.01

4.3.6. Gene expression quantification using FeatureCounts

In order to calculate changes in gene expression the number of reads mapping to a particular gene or exon, henceforth known as a feature, in the reference genome need to be counted. Common methods count the number of reads which overlap any exon within a gene (Anders *et al.*, 2013). Counting reads at the gene level provides an overall summary for the expression level of the gene however this method cannot distinguish differential isoform expression. This is due to reads overlapping multiple exons due to alternative splicing. Various different tools have been developed to quantify read counts at the transcript level and gene level, however there are very few general purpose softwares available which can do both gene and transcript level analysis. The most common general purpose software, HT-Seq, is well tested however it is not optimised for efficiency or speed (Anders, Pyl and Huber, 2015). Therefore *featureCounts* (version 1.5.2), another general purpose software, was used to count mapped reads to genomic features such as genes and exons. Furthermore, the gene or exon level can be specified by the user (Liao, Smyth and Shi, 2014). The advantage of *featureCounts* compared to other software is that the counting can be completed on a desktop computer within minutes, mitigating the use of a supercomputer without a lapse in efficiency or accuracy of counting reads (Liao, Smyth and Shi, 2014).

In order to calculate changes in gene expression, reads were counted to the meta-feature level, or gene level, using *featureCounts. featureCounts* uses the BAM files generated from STAR and a list of genomic features in GTF format, ideally the same GTF file which has been used to generate the reference genome. The output is given as a counts table, in which rows correspond to genes listed in the GTF file and the columns correspond to the sample. For counting reads to the gene level the following code was used in a UNIX operating system:

```
featureCounts -a annotation.gtf -T 8 -p -F GTF -O -g gene_name -o gene_counts_output Alignment_file.bam
```

Where –a is the annotation GTF file, -T is the number of threads used, -p indicates paired-end sequencing data, -F indicated the annotation file is in GTF format. –O allows for the counting of reads which overlap multiple metafeatures, this allows reads which are mapped to read through genes to be counted to each individual gene. -g is the attribute to group the sequences to, in this case gene_name. –o is the output table

name, and the input file is a list of aligned BAM files. **Table 4.7** shows the FeatureCounts summary report file following counting for each sample. Differential gene expression was then analysed from the counts table using the R package *DESeq2* (Love, Huber and Anders, 2014).

Table 4.7: The FeatureCounts summary report shows that over 75% of read fragments aligned using STAR are assigned to an annotated gene and counted.

Sample	Experiment	Number of fragments	Assigned fragments		Unassigned fragments			
			Number Assigned	Assigned (%)	Number multi mapping	Multi mapping (%)	Number with no feature	No feature (%)
AS3MT_KO_1	Primary	10380542	8266364	79.63	1452541	13.99	661637	6.38
AS3MT_KO_2		9807988	7781439	79.34	1511106	15.41	515443	5.26
AS3MT_KO_4		10815590	8555805	79.11	1657580	15.33	602205	5.57
AS3MT_KO_5		6579605	5152507	78.31	1160353	17.64	266745	4.05
AS3MT_KO_6		14086597	11140182	79.08	2031960	14.42	914455	6.49
AS3MT_WT_1		14553699	11755000	80.77	2026193	13.92	772506	5.31
AS3MT_WT_2		16338876	13102156	80.19	2289197	14.01	947523	5.80
AS3MT_WT_4		23899221	19524247	81.69	3139251	13.14	1235723	5.17
AS3MT_WT_5		14047853	10623184	75.62	2754100	19.61	670569	4.77
AS3MT_WT_6		15971942	12809119	80.20	2288339	14.33	874484	5.48
AS3MT_KO_1	Validation	30219875	23153664	76.62	5960392	19.72	1105819	3.66
AS3MT_KO_3		28115900	22319885	79.39	4838268	17.20	957747	3.41
AS3MT_KO_4		28211823	22104279	78.35	5147245	18.24	960299	3.40
AS3MT_WT_2		28547684	21721524	76.09	5798014	20.31	1028146	3.60
AS3MT_WT_5		27132231	20532537	75.68	5743909	21.17	855785	3.15
AS3MT_WT_6		30185869	23653336	78.36	5486256	18.17	1046277	3.47

4.3.7. Gene expression analysis using DESeq2

DESeq2 uses the raw read counts imported from the *featureCounts* count table and applies normalisation methods, estimates the library size, estimates dispersion, and runs a negative binomial generalised linear model fit. The first step is normalisation, which consists of calculating the geometric mean for each gene across all samples, the gene counts for each gene in each sample are then divided by the geometric mean. The median-of-ratios is calculated by calculating the ratio of expression for each gene between each sample, the median ratio is then calculated across all expressed genes and is used as the normalization scale otherwise known as size factor for the sample. The size factor is used to correct for the library size and RNA composition bias. *DESeq2* then uses shrinkage estimation for dispersions and fold changes, with a dispersion value estimated for each gene through fitting an empirical bayes model. Then, for every gene and sample, Cook's distance is calculated which is used as a test for outliers, immediately flagging genes which contain a Cook's distance above a cutoff for samples with three or more replicates. The p-values for these genes are set to "NA", finally the p-values are adjusted for multiple testing using the Benjamini-Hochberg correction (Love, Anders and Huber, 2016).

Prior to running the *DESeq2* pipeline, the counts table was filtered for non-expressed and lowly expressed genes, by removing any gene with less counts than the number of samples, these were 10 counts in the primary experiment and six counts in the validation experiment. Similarities in the genome-wide expression profiles between the samples were visualised by principal component analysis (PCA) by plotting the first two principle components (**Figure 4.8**) and a heatmap clustered by Euclidean distance (**Figure 4.9**) using all inputed gene counts, post removal of lowly expressed genes. These two visual analysis allowed for identification of mislabelled samples of which there were none.

To identify genotype specific gene expression changes a Wald test was used. P-values were adjusted for multiple testing using the Benjamini and Hochberg correction, otherwise known as the False Discovery Rate (FDR) (Benjamini and Hochberg, 1995). FDR-adjusted p-values less than 0.05 were defined as significant. Fold changes are shown with wild type cell lines used as baseline. Significant changes in gene expression between the two sequencing experiments were compared for overlapping

genes (FDR-adjusted p-value <0.05), shared upregulated genes (FDR-adjusted p-value <0.05, $\log_2$ fold change >1), and shared downregulated genes (FDR-adjusted p-value <0.05, $\log_2$ fold change <1).

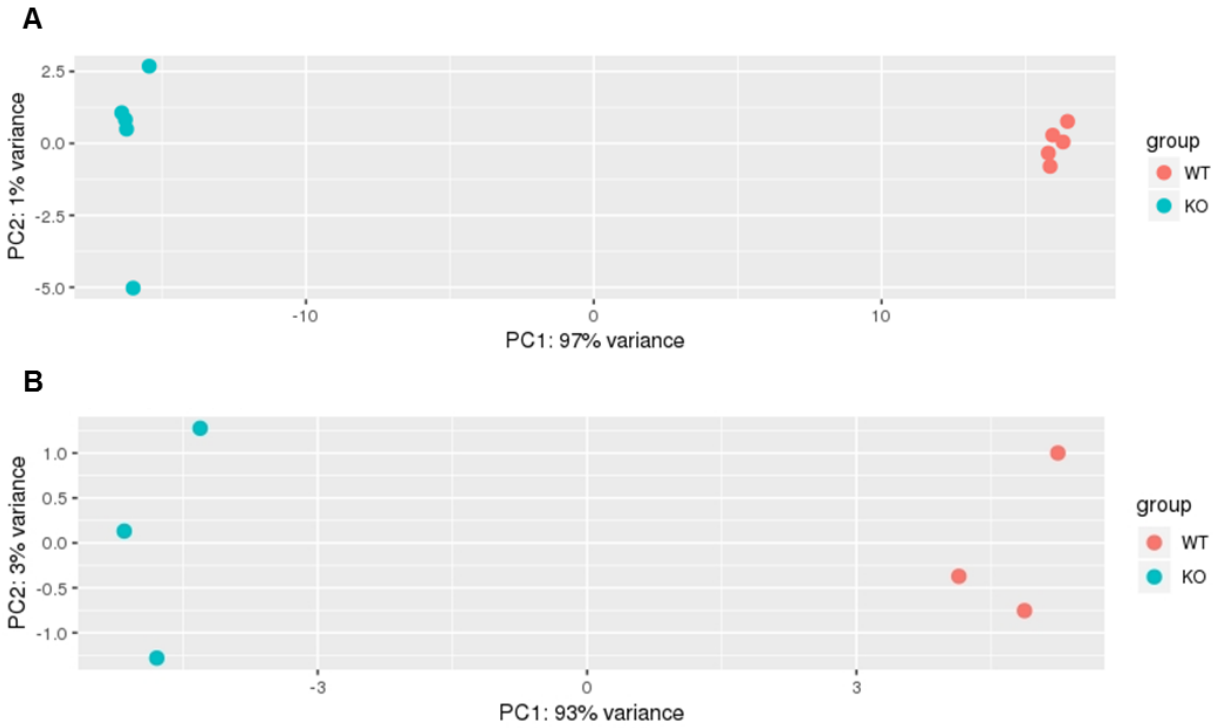


Figure 4.8: Principal Component Analysis (PCA) plot of the first two principle components in the primary RNAseq experiment (A) and the validation RNAseq experiment (B). *From this it is clear that the variance between the samples in principle component 1 is being driven by the genotype of the samples.*

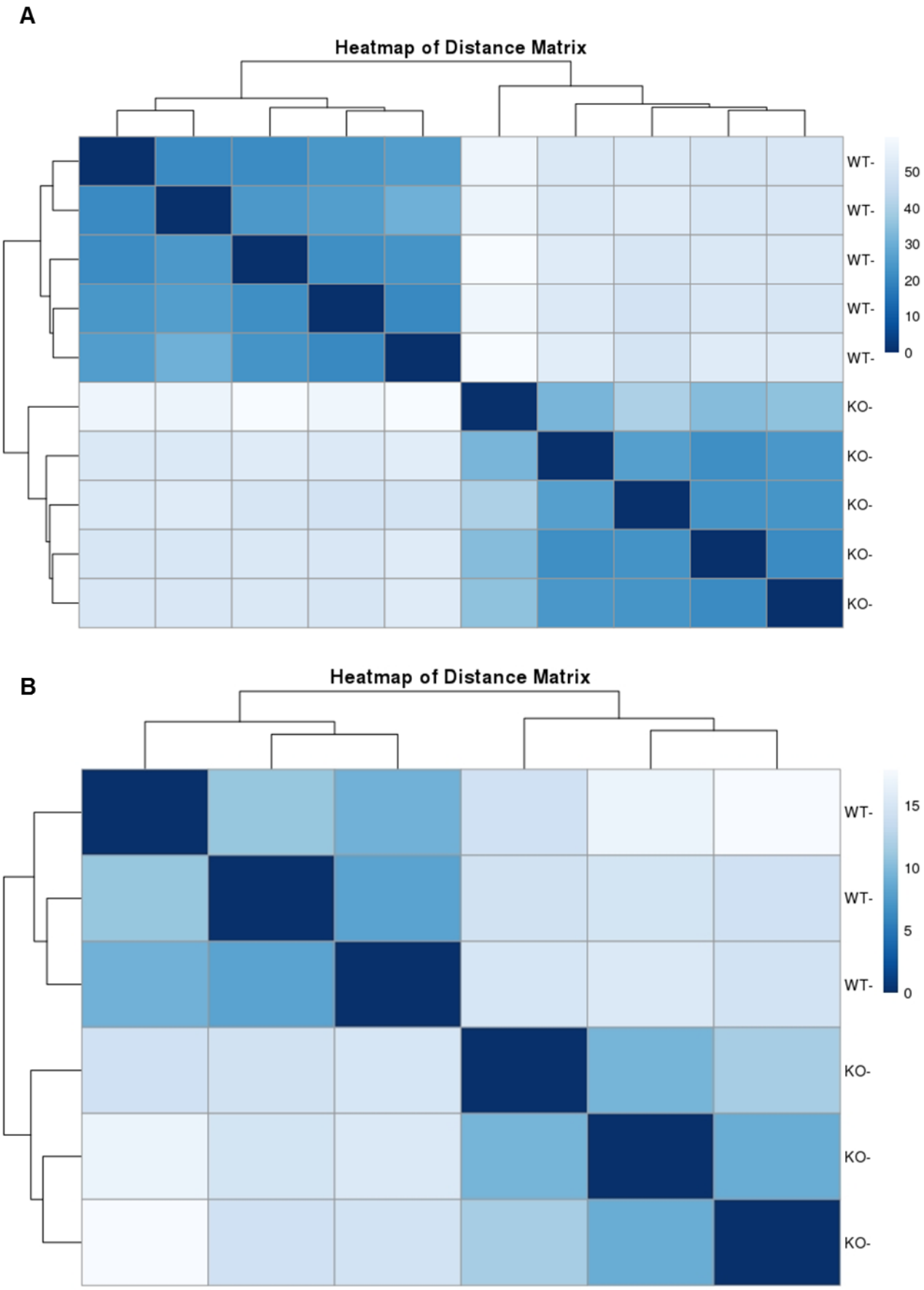


Figure 4.9: Clustering by Euclidean Distance between samples in the primary RNAseq experiment (a) and the validation RNAseq experiment (b) shows the WT and KO samples cluster together, further showing that samples are not mislabelled.

4.3.8. Functional gene annotation and gene ontology analysis using GOseq

Lists of differentially expressed genes generally do not provide much information regarding what is functionally happening within the cell. Genes are usually co-expressed in networks and the genes in these networks provide similar phenotypic functions. These gene functions have been collated into Gene Ontology (GO) (Ashburner *et al.*, 2000), an open data set listing all known gene functions which is constantly expanding. Using this database the significant genes identified through *DESeq2* can then be analysed for over-enrichment in their GO terms to identify if fundamental shared pathways have been altered, providing more power than a list of gene names.

In order to annotate the significant differentially expressed genes from *DESeq2* the *GOseq* package was used (Young *et al.*, 2010). *GOseq* was selected as it has functionality to correct for gene length biases, which can confound RNAseq data (Oshlack and Wakefield, 2009). This is because longer genes will more likely have reads which map by chance, and genes with higher expression will therefore have more power to detect a significant difference between groups. This is further confounded in studies involving the brain tissue as the brain transcriptome contains large numbers of larger genes, therefore it is important to correct for gene length bias when annotating functions (Raychaudhuri *et al*., 2010). First *GOseq* takes the list of genes which have been identified as DEGs from *DESeq2* then a probability weighting function is estimated from the data to calculate how the probability of a DEG changes as a function of its transcript length. Finally, resampling of the data is performed by randomly selecting a set of genes the same size as the DEGs and counting the number of genes associated with the GO category of interest. This continues many times and the resulting distribution of GO category membership is the approximate shape of the true distribution. This then allows for the calculation of a p-value for each GO category which is over-represented in the set of DEGs using the Wallenius approximation which uses hypergeometric distributions resulting in computationally light analysis compared to traditional random sampling methods with minimal loss of accuracy. The output from *GOseq* is a list of GO terms and their corresponding ontology, which could be either a biological process, cellular component or molecular function. An enrichment can identify fundamental pathways which have been altered between the two genotypes.

This image has been removed by the author of this thesis for copyright reasons.

4.4. Results

4.4.1. Sequencing metrics

There was a significant difference in read-depth identified in raw reads between the *AS3MT* KO B6 and *AS3MT* WT D2 (n=5 samples each group, two-tailed unpaired t-test, t(6.88)=-2.94, p-value=0.02). This could have been caused by differences in library preparation or through sequencing, however the *DESeq2* pipeline corrects for this variation when calculating DEGs. There was no difference in the read-depth observed between the *AS3MT* KO C6 and *AS3MT* WT D4 cell lines (n=3 samples, two-tailed unpaired t-test, t(2.77)=0.452, p-value=0.68). In addition, there were no significant differences in the number of sequences mapped to unique regions of the human genome when comparing *AS3MT* KO and *AS3MT* WT samples in both the primary experiment (B6 and D2, n=5 samples, two-tailed unpaired t-test, t(6.66)=-0.30, p-value=0.77) and the validation experiment (C6 and D4, n=3 samples, two-tailed unpaired t-test, t(1.34)=2.61, p-value=0.29) **Figure 4.11**.

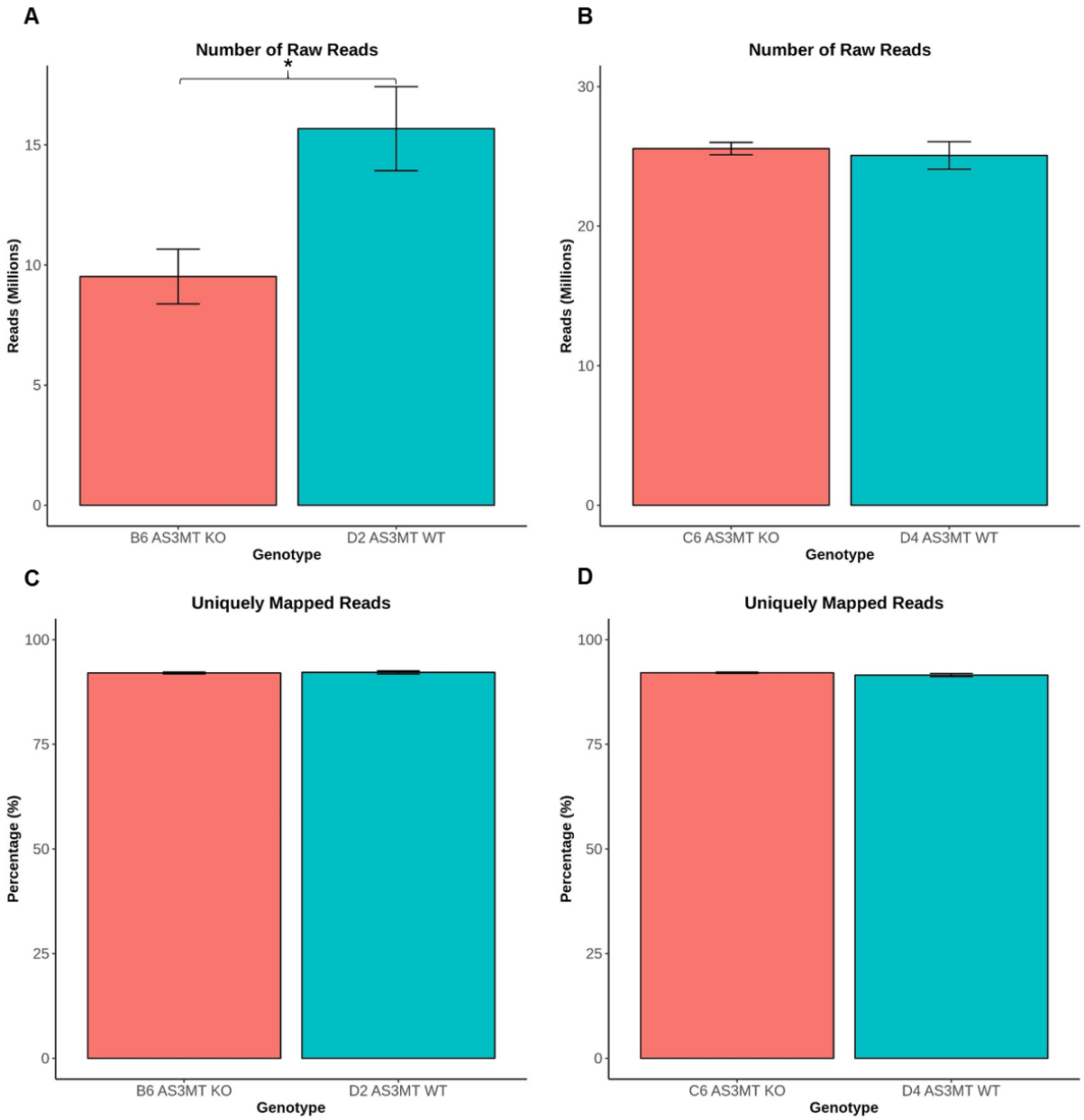


Figure 4.10: Comparison of the RNA-seq reads in the AS3MT knockout and WT cell lines in each experiment. *The number of raw-reads differed between the B6 AS3MT KO and D2 AS3MT WT cell lines (n=5 samples each group, two-tailed unpaired t-test, t(6.88)=-2.94, p-value=0.02) (A) but did not differ between the C6 AS3MT KO and D4 AS3MT WT cell lines (B). There was no difference in the percentage of reads mapped uniquely to the human genome in both the B6 AS3MT KO and D2 AS3MT WT cell lines (C) or the C6 AS3MT KO and D4 AS3MT WT cell lines (D).*

4.4.2. *AS3MT* expression is significantly reduced in both knockout cell lines

From previous results in Chapter 3. where knockout of the AS3MT protein has been confirmed with western blot (**3.4.6.**) and sequence disruption identified through sequencing (**3.4.5.**) I expected there to be a reduction of *AS3MT* expression in the knockout cell lines compared to control. To examine this I first merged all of the alignment files from *STAR* for each of the biological repeats of each cell line into single .bam files for each line (B6, C6, D2, and D4), the alignments were then visualised using IGV. From Chapter 3. I expected there to be reduced read coverage of exons 4, 5, and 6 of *AS3MT* in the knockout lines compared to wild type, as the gRNA result in deletion of this region. When viewed in IGV there is complete removal of exon 5 and 6 and reduction towards the 3' end of exon 4 in the knockout cell lines but no removal in the wild type lines (**Figure 4.12**). When the expression of *AS3MT* was quantified using *DESeq2* I identified a significant decrease in the gene expression of *AS3MT* in both knockout cell lines versus their control (log_2 fold change of -0.566 (padj =2.13E-05) and log_2 fold change of -0.891 (padj = 2.24E-24) **Table 4.7**). The normalised read counts from *DESeq2* for each replicate are shown in **Figure 4.13**.

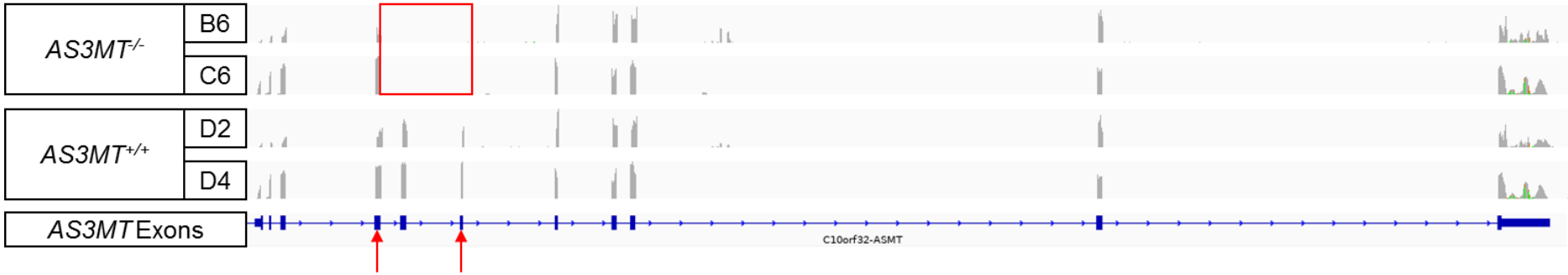


Figure 4.11: Knockout of AS3MT results in removal of sequence between exons 4 and 6. This is confirmed by reduced read coverage of exons 4, 5, and 6 in the knockout cell lines (B6 and C6) compared to controls (D2 and D4) (red box). *The two gRNA used to generate the cell lines in CHAPTER are shown as red arrows on the exon track.*

Table 4.7: DESeq2 results for AS3MT in the primary and validation experiments show a log2 fold change of -0.566 and -0.891 in AS3MT in the knockout cell lines.

	baseMean	Log2Fold Change	lfcSE	Stat	p-value	p-adj
Primary	365	-0.566	0.120	-4.71	2.43E-06	2.13E-05
Validation	722	-0.891	0.0831	-10.7	8.38E-27	2.24E-24

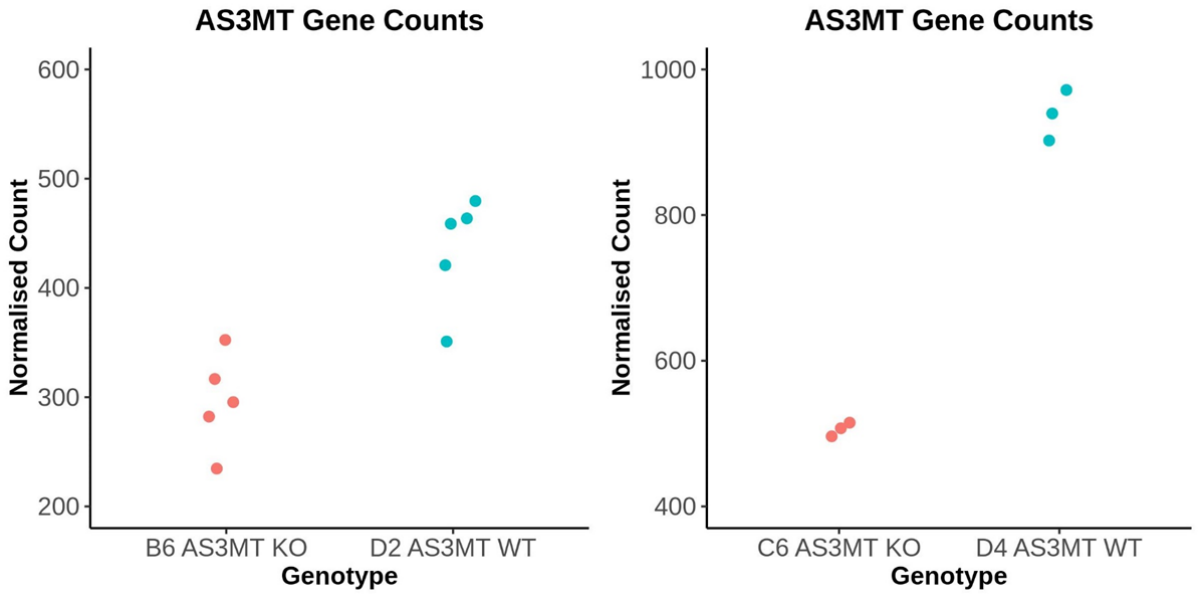


Figure 4.12: Normalised counts of AS3MT mapped reads for the primary and validation RNAseq both show a significant decrease in AS3MT expression in knockout cell lines compared to wild type.

4.4.3. There are 973 genes differentially expressed genes which overlap in the *AS3MT*$^{-/-}$ B6 and C6 cell lines

The first aim of this chapter was to identify differentially expressed genes (DEGs) between the *AS3MT*$^{-/-}$ cell lines and the *AS3MT*$^{+/+}$ cell lines. In order to do this, RNA from two homozygous *AS3MT* knockout cell lines generated in Chapter 3. were sequenced alongside two *AS3MT* wild type control cell lines in two separate RNAseq experiments. From this, two lists of DEGs were generated using *DESeq2*, these two lists were then compared for overlapping DEGs which are differentially expressed in both *AS3MT* knockout cell lines compared to their respective controls. The gene list was then analysed using *GOseq* to identify any biological pathways overrepresented by DEGs, providing a potential functional link for the *AS3MT* knockout.

From *DESeq2* I identified 7097 DEGs (defined as an FDR-adjusted p-value<0.05) in the *AS3MT*$^{-/-}$ B6 cell line compared to *AS3MT*$^{+/+}$ D2 cell line. 1556 DEGs were identified in the *AS3MT*$^{-/-}$ C6 cell line compared to *AS3MT*$^{+/+}$ D4 cell line. Comparison of the two lists of DEGs identified 973 DEGs present in both data sets (**Figure 4.14**). The direction of effect of the DEGs was similar in both experiments. 500 of the overlapping DEGs are downregulated ($\log_2$ fold change <0) and 264 are upregulated ($\log_2$ fold change >0) in both experiments. 92 DEGs are upregulated in *AS3MT*$^{-/-}$ C6 but downregulated in *AS3MT*$^{-/-}$ B6 and 117 are upregulated in *AS3MT*$^{-/-}$ B6 and downregulated in *AS3MT*$^{-/-}$ C6 (**Figure 4.15**). Pearson's correlation coefficient shows a positive correlation between the $\log_2$ fold change values of the overlapping DEGs between both experiments (Pearson's correlation coefficient, r =0.519, p-value=2.2e-16).

Multiple pathways involved in protein targeting, translation, nonsense mediated decay, cell adhesion and cell-cell interactions, neuron development and immune system responses were identified from the list of DEGs when run through *GOseq* **Table 4.8**. The top pathway enriched in the DEGs was SRP-dependent cotranslational protein targeting to membrane (GO:0006614, number DEGs = 50, total genes in category = 88, corrected p-value=3.99e-37). This pathway is involved with targeting proteins to membranes during translation and could indicate that protein production or secretion is altered in *AS3MT* knockout cell lines (Luirink and Sinning, 2004). The second highest enriched GO term is for nuclear-transcribed mRNA catabolic process,

nonsense mediated decay (GO:0000184, number DEGs=51, total genes in category=115, corrected p-value=8.74e-30). This is not surprising, as CRISPR-Cas9 mutations result in a frameshift and resultant premature stop codons within the mRNA transcript, if nonsense mediated decay pathways are altered then it indicates that the transcripts produced in the knockout *AS3MT* gene are being destroyed before they are translated (Lykke-andersen and Jensen, 2015; Popp and Maquat, 2016). There is also enrichment within several broad cellular morphological pathways such as focal and cell adhesion, extracellular space, neuronal cell body, cell junctions, growth cones, axon guidance and dendrites. This is an indication that the general morphology of the SH-SY5Y cells are altered once *AS3MT* is knocked out. This supports previous evidence I have shown in Chapter 3. where the neurite length and cell size are increased in knockout *AS3MT* cell lines.

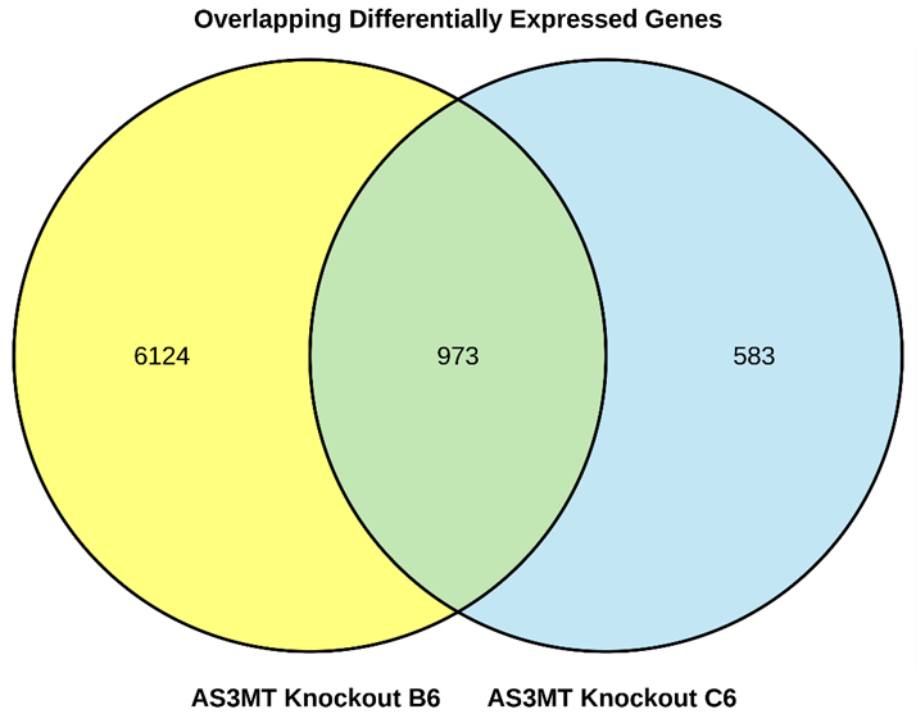


Figure 4.13: Venn diagram showing that 973 differentially expressed genes identified by DESeq2 overlap between the two AS3MT knockout cell lines. *6124 are specific to the AS3MT B6 knockout cell line and 583 to the AS3MT knockout C6.*

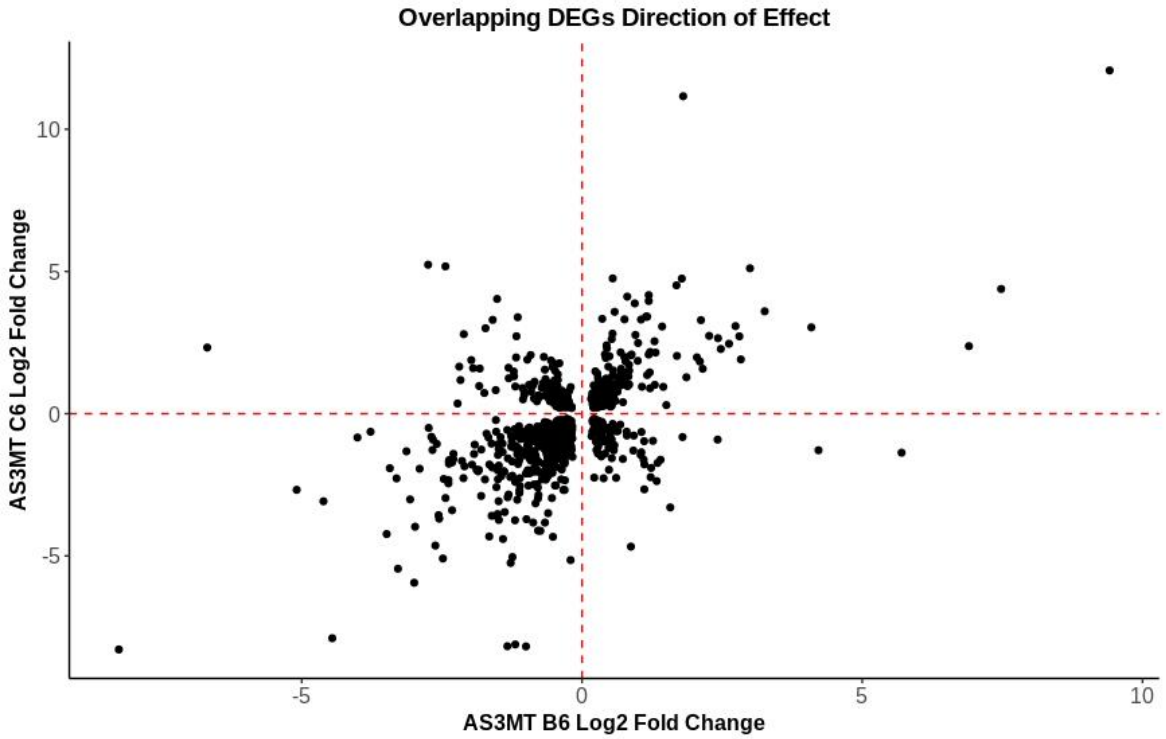


Figure 4.14: Effect sizes of differentially expressed genes in the AS3MT B6 knockout line are correlated with those in the AS3MT C6 knockout cell line. *There is a positive correlation for effect size (Log2 Fold Changes are compared to AS3MT WT cell lines). Pearson's correlation coefficient, r =0.52, p-value=2.2e-16.*

Table 4.8: GO terms for overlapping significant genes between the two RNAseq experiments from GOseq.

Category	over_represented_pvalue	under_represented_pvalue	numDEInCat	numInCat	term	ontology	correctedP
GO:0006614	2.43E-41	1	50	88	SRP-dependent cotranslational protein targeting to membrane	BP	3.99E-37
GO:0000184	1.06E-33	1	51	115	nuclear-transcribed mRNA catabolic process, nonsense-mediated decay	BP	8.71E-30
GO:0006413	4.65E-33	1	53	128	translational initiation	BP	2.55E-29
GO:0003735	2.09E-32	1	54	154	structural constituent of ribosome	MF	8.58E-29
GO:0006412	2.12E-30	1	57	172	translation	BP	6.98E-27
GO:0022625	1.49E-24	1	31	61	cytosolic large ribosomal subunit	CC	4.08E-21
GO:0022627	2.62E-18	1	22	40	cytosolic small ribosomal subunit	CC	6.16E-15
GO:0005615	1.70E-17	1	106	806	extracellular space	CC	3.50E-14
GO:0002181	1.27E-14	1	18	37	cytoplasmic translation	BP	2.32E-11
GO:0042788	7.03E-11	1	14	30	polysomal ribosome	CC	1.16E-07
GO:0005925	9.50E-11	1	60	380	focal adhesion	CC	1.42E-07
GO:0005886	5.83E-10	1	246	2756	plasma membrane	CC	7.98E-07
GO:0007155	7.51E-10	1	56	354	cell adhesion	BP	9.50E-07
GO:0005840	1.38E-09	1	16	45	ribosome	CC	1.62E-06
GO:0005576	2.57E-09	1	102	992	extracellular region	CC	2.82E-06
GO:0062023	4.73E-09	0.999	35	181	collagen-containing extracellular matrix	CC	4.86E-06
GO:0005887	1.09E-08	1	95	845	integral component of plasma membrane	CC	1.05E-05
GO:0070062	1.62E-08	1	152	1649	extracellular exosome	CC	1.48E-05
GO:0044267	2.77E-08	0.999	27	140	cellular protein metabolic process	BP	2.39E-05
GO:0015935	3.57E-08	0.999	8	13	small ribosomal subunit	CC	2.94E-05
GO:0014069	4.25E-08	0.999	38	217	postsynaptic density	CC	3.33E-05
GO:0005788	7.30E-07	0.999	34	220	endoplasmic reticulum lumen	CC	0.000545
GO:0005178	1.21E-06	0.999	22	103	integrin binding	MF	0.000861
GO:0043025	2.37E-06	0.999	42	302	neuronal cell body	CC	0.00162
GO:0000027	2.52E-06	0.999	9	25	ribosomal large subunit assembly	BP	0.00166

GO:0030424	3.09E-06	0.999	37	241	axon	CC	0.00196
GO:0007411	4.23E-06	0.999	28	160	axon guidance	BP	0.00257
GO:0010976	5.07E-06	0.999	19	91	positive regulation of neuron projection development	BP	0.00297
GO:0045202	6.98E-06	0.999	32	208	synapse	CC	0.00395
GO:0030054	9.37E-06	0.999	53	433	cell junction	CC	0.00513
GO:0030426	1.02E-05	0.999	22	116	growth cone	CC	0.00540
GO:0030425	1.12E-05	0.999	43	323	dendrite	CC	0.00575
GO:0005883	1.99E-05	0.999	5	7	neurofilament	CC	0.00992
GO:0007417	2.18E-05	0.999	17	83	central nervous system development	BP	0.0105
GO:0051823	2.65E-05	1	4	4	regulation of synapse structural plasticity	BP	0.0124
GO:0001525	2.80E-05	0.999	27	178	angiogenesis	BP	0.0128
GO:0008237	2.91E-05	0.999	13	51	metallopeptidase activity	MF	0.0129
GO:0009986	5.19E-05	0.999	45	385	cell surface	CC	0.0225
GO:0005604	5.67E-05	0.999	15	69	basement membrane	CC	0.0239
GO:0007399	5.96E-05	0.999	33	245	nervous system development	BP	0.0245
GO:0031290	6.88E-05	0.999	7	17	retinal ganglion cell axon guidance	BP	0.0276
GO:0060384	7.61E-05	0.999	7	17	innervation	BP	0.0298
GO:0019955	8.57E-05	0.999	8	24	cytokine binding	MF	0.0328
GO:0004896	0.000101	0.999	7	19	cytokine receptor activity	MF	0.0377
GO:0048245	0.000106	0.999	4	7	eosinophil chemotaxis	BP	0.0387
GO:0009897	0.000120	0.999	22	148	external side of plasma membrane	CC	0.0425
GO:0097187	0.000122	1	3	3	dentinogenesis	BP	0.0425
GO:0010469	0.000124	0.999	23	196	regulation of signalling receptor activity	BP	0.0425
GO:0004861	0.000146	0.999	5	12	cyclin-dependent protein serine/threonine kinase inhibitor activity	MF	0.0488

4.4.4. There is an enrichment of schizophrenia GWAS genes which are differentially expressed following knockout of *AS3MT*

I then set out to identify if there are any DEGs which have been previously associated with schizophrenia and if there is an enrichment following knockout of *AS3MT.* To do this the gene list from the most recent GWAS of schizophrenia was gathered from (Pardiñas *et al.*, 2018). I then examined for these genes in the individual DEG list for each experiment (defined as FDR <0.05) and performed an enrichment analysis using fishers exact test. I then extracted the schizophrenia DEGs which overlapped in each experiment to examine if they have the same direction of effect.

In the primary experiment, 23,783 genes were tested for differential expression using *DESeq2.* Of these, 7097 reached significance and were defined as a DEG. 533 genes from the schizophrenia GWAS were tested for differential expression and of these 192 were DEGs with FDR <0.05. This results in a significant enrichment of genes involved in schizophrenia being DEGs following knockout of *AS3MT* in the B6 line (Fishers exact, p-value=0.0146). In the validation experiment, 25,719 genes were tested for differential expression using *DESeq2* and 1556 reached significance and were defined as a DEG. Within the tested gene list, 492 genes were in the schizophrenia GWAS and of these 44 were DEGs. Following a Fishers exact test there is a significant enrichment in genes associated with schizophrenia being DEGs (Fishers exact, p-value=0.0129). Examining the overlap there are 27 genes shared between both datasets (**Table 4.9**). All bar two (*MGAT3* and *SNAP91*) have the same direction of effect in both knockout cell lines compared to controls (**Figure 4.15**).

Table 4.9: There are 27 schizophrenia associated genes which are DEGs within both datasets. Of these 27, 25 have the same direction of effect in both AS3MT knockout cell lines compared to controls.

Gene Name	baseMean D4	log2 Fold Change C6	lfcSE	stat	pvalue	Padj C6	baseMean D2	log2 Fold Change B6	lfcSE	stat	pvalue	Padj B6
ABCB1	3835.26	-0.18	0.06	-2.99	0.00278	0.0371	2483.36	-0.56	0.08	-7.31	2.63E-13	5.50E-12
AS3MT	721.51	-0.89	0.08	-10.72	8.38E-27	2.24E-24	364.98	-0.57	0.12	-4.71	2.43E-06	2.13E-05
ASCL1	816.95	1.43	0.09	15.96	2.36E-57	2.83E-54	21.08	3.06	0.41	7.52	5.32E-14	1.18E-12
DPYD	285.13	-0.35	0.11	-3.08	0.00208	0.0289	302.33	-0.76	0.11	-6.92	4.37E-12	8.16E-11
DRD2	109.06	-1.03	0.18	-5.74	9.57E-09	5.20E-07	29.87	-1.46	0.29	-5.01	5.39E-07	5.32E-06
INA	1308.61	-0.86	0.08	-11.23	2.80E-29	8.55E-27	363.04	-0.73	0.10	-7.17	7.56E-13	1.51E-11
KIAA1324L	1553.23	-0.28	0.07	-4.19	2.83E-05	0.000694	1055.33	-0.25	0.07	-3.50	0.000471	0.00252
KLC1	3932.00	-0.27	0.06	-4.82	1.45E-06	4.95E-05	1721.93	-0.17	0.06	-2.86	0.00422	0.0169
LRP4	1152.99	0.35	0.07	4.89	1.00E-06	3.56E-05	518.70	0.32	0.11	2.81	0.00496	0.0194
LRRN3	3061.04	0.28	0.09	3.32	0.000912	0.0146	563.86	0.37	0.10	3.51	0.000444	0.00240
MAPK7	1039.16	0.28	0.08	3.75	0.000175	0.00349	292.20	1.25	0.11	11.71	1.18E-31	7.70E-30
MFAP4	8155.06	0.87	0.07	12.80	1.54E-37	6.90E-35	1544.12	2.05	0.09	23.06	1.31E-117	6.21E-115
MGAT3	154.36	-0.52	0.16	-3.35	0.000807	0.0131	25.28	1.64	0.31	5.26	1.42E-07	1.53E-06
NAB2	672.15	-0.30	0.09	-3.34	0.000841	0.0136	323.58	-0.46	0.12	-3.77	0.000166	0.000988
NDRG4	724.51	-0.34	0.08	-4.06	4.88E-05	0.00113	238.53	-0.88	0.12	-7.57	3.75E-14	8.44E-13
NDUFAF2	698.31	0.28	0.08	3.27	0.00107	0.0168	203.50	0.63	0.15	4.26	2.05E-05	0.000149
OSBPL3	289.38	-0.33	0.11	-2.91	0.00362	0.0460	509.69	-1.70	0.12	-14.42	3.89E-47	4.70E-45
PRKD1	43.03	1.29	0.29	4.46	8.27E-06	0.000233	22.18	2.54	0.42	6.01	1.81E-09	2.53E-08
PTN	3967.55	-0.51	0.06	-7.99	1.32E-15	1.75E-13	7146.68	-0.98	0.08	-12.42	1.96E-35	1.49E-33
REEP2	826.75	-0.36	0.08	-4.50	6.67E-06	0.000192	241.09	-0.74	0.12	-6.14	8.09E-10	1.19E-08
RPS17	23850.23	0.36	0.07	5.08	3.82E-07	1.53E-05	3892.45	0.36	0.12	3.12	0.00179	0.00811
RTN1	2382.90	-0.38	0.07	-5.81	6.27E-09	3.59E-07	1345.22	-0.70	0.06	-10.85	1.95E-27	1.04E-25
SNAP91	200.44	-1.19	0.14	-8.70	3.37E-18	5.41E-16	53.30	0.96	0.21	4.54	5.53E-06	4.54E-05
SPATS2L	933.77	-0.32	0.08	-4.12	3.74E-05	0.000888	922.43	-0.43	0.07	-6.37	1.94E-10	3.05E-09
TCF4	13453.99	0.23	0.06	3.91	9.39E-05	0.00200	6496.09	0.23	0.07	3.22	0.00127	0.00604

USMG5	3146.55	0.37	0.09	3.95	7.85E-05	0.00171	1221.08	0.33	0.13	2.59	0.00954	0.0337
ZNF804A	205.27	1.21	0.15	7.89	2.94E-15	3.77E-13	25.80	1.45	0.28	5.10	3.41E-07	3.49E-06

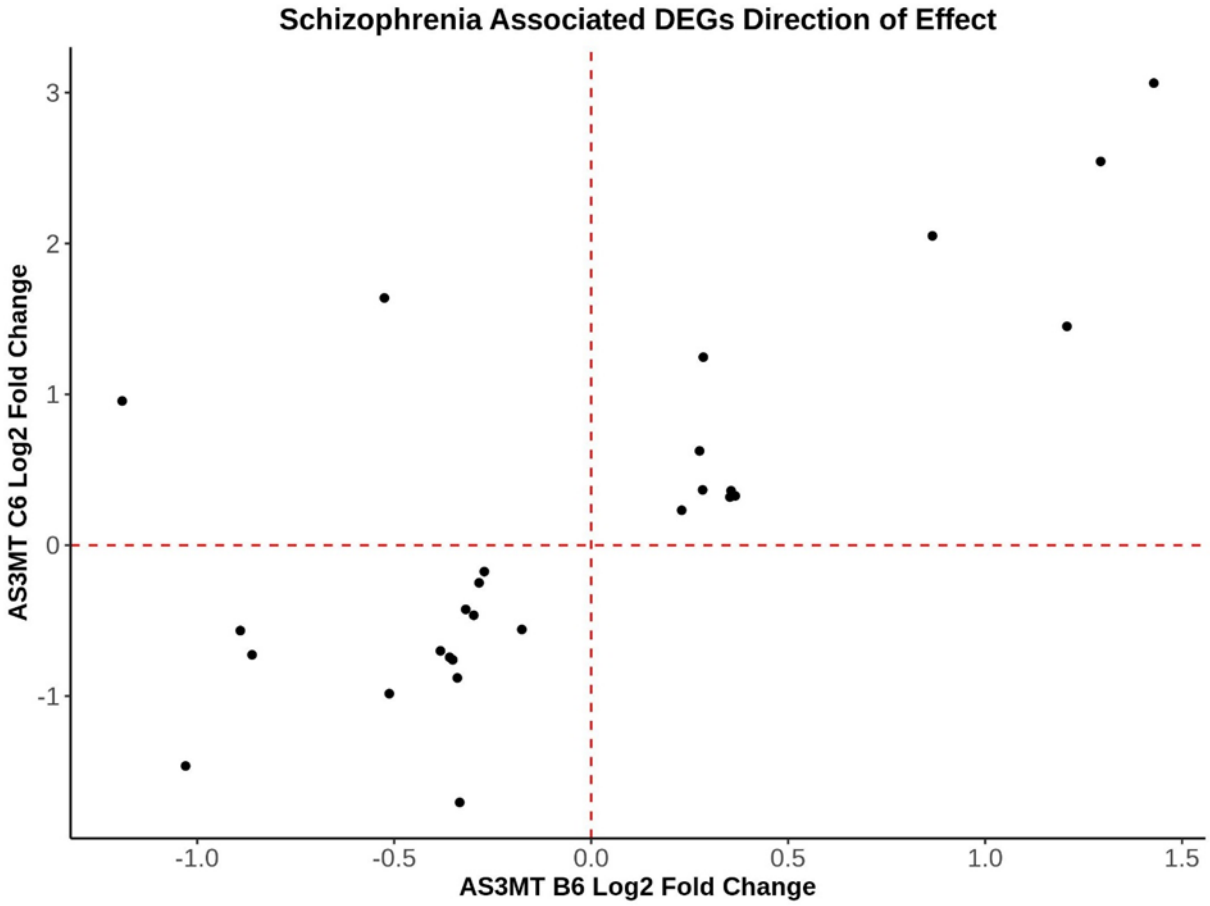


Figure 4.15: 25 of the schizophrenia GWAS genes that are DEGs have the same direction of effect in both knockout cell lines compared to control. *Two genes, MGAT3 and SNAP91 have increased expression in the C6 knockout line and decreased within the B6 knockout line.*

4.4.5. 36 of the overlapping differentially expressed genes are upregulated with a log_2 fold change>1 and FDR <0.05

I next set out to narrow down the list of 973 overlapping differentially expressed genes to identify a list of genes with a log_2 fold change of >1 and FDR<0.05. These would be genes which have shared increased expression between the two KO cell lines. I then set out to identify which pathways (if any) are altered by the upregulation of these genes.

Only 663 of the 7097 FDR <0.05 genes in the B6 *AS3MT*$^{-/-}$ and 105 of the 1556 FDR <0.05 genes in the C6 *AS3MT*$^{-/-}$ cell line show a log_2 fold change increase >1 (**Figure 4.17**). This corresponds to 9.34% and 6.75% of the DEGs identified which have a log_2 fold change of >1 in both cell lines respectively. Of these genes, only 36 of the 973 overlapping genes have a log_2 fold change >1 (**Figure 4.18**) these are listed in **Table 4.10**.

GOseq analysis of the 36 overlapping genes resulted in no significant molecular or cellular GOterms being enriched, this is probably due to the low gene number input.

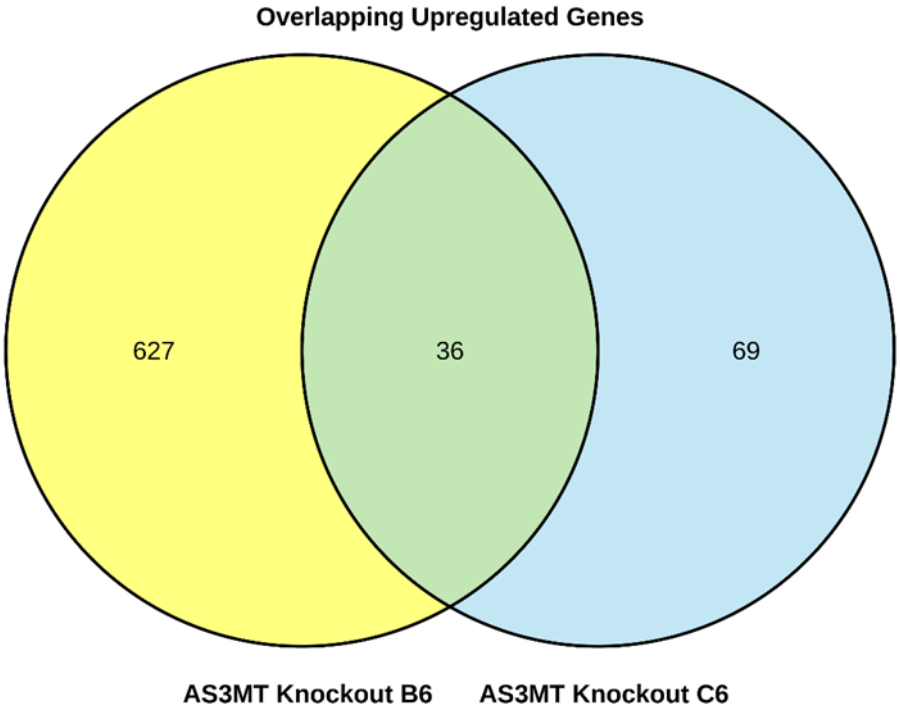


Figure 4.16: Venn diagram shows that there are 36 differentially expressed upregulated (log_2 fold change >1) genes which overlap between the two AS3MT knockout cell lines. *627 are specific to the AS3MT B6 knockout cell line and 69 to the AS3MT knockout C6*

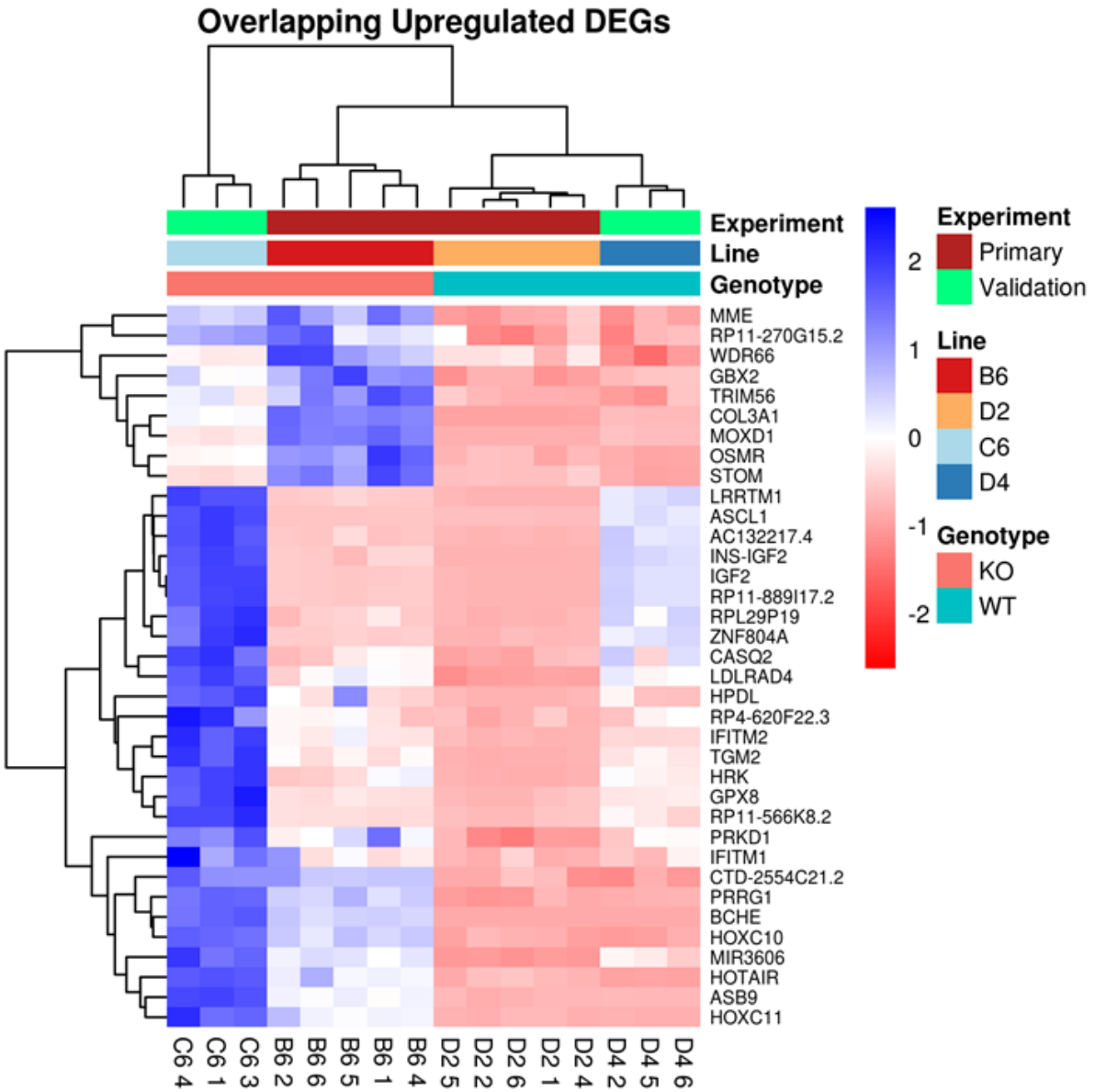


Figure 4.17: Heatmap showing the normalised read count of the 36 upregulated DEGs with log_2 fold change >1 which overlap between the two datasets. *The DEGs group firstly by AS3MT genotype and then by cell line. Each column is a sample, each row is a gene and each row has been scaled where blue is increased read count and red decreased.*

Table 4.10: Overlapping upregulated DEGs between the two AS3MT knockout lines (log_2 fold change >1 and FDR p-adj <0.05). *Values provided are the base expression, log_2 fold change, log_2 fold change standard error (lfcSE), wald statistic (stat), raw p-value (pvalue) and BH corrected p-value (padj). Values for both experiments are shown. Genes are orders by rank of ranks for average padj and average log_2 fold change.*

Gene Name	baseMean D4	log_2 Fold Change C6	lfcSE	stat	pvalue	Padj C6	baseMean D2	log_2 Fold Change B6	lfcSE	stat	pvalue	Padj B6
BCHE	567.44	9.41	0.65	14.47	2.01E-47	1.29E-44	312.23	12.07	0.92	13.16	1.55E-39	1.43E-37
COL3A1	2596.74	1.68	0.07	24.28	3.58E-130	2.29E-126	4474.4	4.51	0.07	66.83	0	0
ASB9	134.95	4.09	0.25	16.16	9.14E-59	1.17E-55	51.91	3.03	0.25	12.02	2.72E-33	1.89E-31
HOTAIR	80.25	6.9	0.74	9.32	1.13E-20	2.14E-18	43.83	2.37	0.28	8.49	2.07E-17	6.11E-16
PRRG1	38.62	3.26	0.39	8.42	3.67E-17	5.60E-15	23.16	3.6	0.43	8.43	3.42E-17	9.92E-16
MOXD1	54.5	1.8	0.27	6.78	1.16E-11	9.94E-10	166.31	11.16	0.92	12.14	6.22E-34	4.42E-32
TGM2	105.31	2.12	0.2	10.41	2.34E-25	5.91E-23	23.08	3.29	0.39	8.51	1.80E-17	5.35E-16
MIR3606	85.24	1.78	0.22	7.95	1.91E-15	2.48E-13	32.47	4.75	0.43	11.04	2.54E-28	1.40E-26
IGF2	2409.08	1.16	0.08	15.37	2.80E-53	2.69E-50	159.95	3.41	0.18	19.26	1.21E-82	3.04E-80
HOXC11	32.06	7.48	1.21	6.19	5.85E-10	3.91E-08	13.87	4.38	0.61	7.16	8.03E-13	1.60E-11
IFITM2	334.73	2.48	0.13	18.88	1.75E-79	4.20E-76	93.77	2.28	0.2	11.2	4.15E-29	2.37E-27
RP11-889I17.2	2331.16	1.14	0.08	15.13	9.68E-52	8.10E-49	157.18	3.41	0.18	19.27	9.91E-83	2.51E-80
HOXC10	107.68	2.83	0.22	12.79	1.88E-37	8.21E-35	76.1	1.91	0.18	10.37	3.39E-25	1.61E-23
AC132217.4	356.27	1.19	0.12	9.99	1.65E-23	3.52E-21	21.03	3.96	0.5	7.88	3.27E-15	8.19E-14
OSMR	58.84	2.27	0.27	8.43	3.36E-17	5.16E-15	129.46	2.74	0.2	13.79	3.02E-43	3.19E-41
HPDL	20.91	3	0.51	5.83	5.67E-09	3.27E-07	5.96	5.11	1.08	4.75	2.01E-06	1.79E-05
INS-IGF2	378.93	1.05	0.11	9.83	8.43E-23	1.74E-20	29.64	3.31	0.42	7.91	2.51E-15	6.34E-14
STOM	15.93	2.81	0.56	5.04	4.66E-07	1.82E-05	59.21	2.72	0.22	12.29	1.04E-34	7.64E-33
LRRTM1	76.8	1.19	0.21	5.62	1.88E-08	9.72E-07	6.18	4.17	0.86	4.84	1.30E-06	1.20E-05
ASCL1	816.95	1.43	0.09	15.96	2.36E-57	2.83E-54	21.08	3.06	0.41	7.52	5.32E-14	1.18E-12
IFITM1	15.09	2.74	0.59	4.66	3.16E-06	9.94E-05	5.24	3.08	0.83	3.7	0.000217	0.00127
TRIM56	10.33	2.43	0.68	3.58	0.000345	0.00635	19.96	2.65	0.37	7.26	3.77E-13	7.78E-12
GPX8	241.49	2.15	0.15	14.67	9.47E-49	6.74E-46	51.78	1.58	0.24	6.54	6.14E-11	1.01E-09
LDLRAD4	149.28	1.21	0.16	7.61	2.82E-14	3.30E-12	49.84	2.17	0.26	8.42	3.80E-17	1.10E-15

RP11-566K8.2	25.11	2.62	0.43	6.07	1.27E-09	8.15E-08	3.38	2.46	0.84	2.93	0.00344	0.0142
HRK	112.7	1.69	0.19	9.01	1.99E-19	3.47E-17	27.85	2.03	0.34	5.92	3.16E-09	4.31E-08
PRKD1	43.03	1.29	0.29	4.46	8.27E-06	0.000233	22.18	2.54	0.42	6.01	1.81E-09	2.53E-08
MME	111.72	1.16	0.18	6.38	1.79E-10	1.31E-08	132.18	1.36	0.15	8.82	1.14E-18	3.65E-17
RP4-620F22.3	23.83	2.05	0.43	4.8	1.59E-06	5.39E-05	6.08	1.98	0.65	3.03	0.00246	0.0107
CTD-2554C21.2	44.06	1.86	0.3	6.27	3.50E-10	2.41E-08	39.23	1.28	0.23	5.6	2.16E-08	2.61E-07
ZNF804A	205.27	1.21	0.15	7.89	2.94E-15	3.77E-13	25.8	1.45	0.28	5.1	3.41E-07	3.49E-06
GBX2	45.36	1	0.27	3.64	0.000273	0.00518	60.21	2.48	0.23	10.92	9.34E-28	5.01E-26
RP11-270G15.2	8.79	2.1	0.67	3.13	0.00177	0.0253	8.71	1.83	0.55	3.36	0.000777	0.00393
CASQ2	20.61	1.31	0.42	3.13	0.00173	0.0248	3.94	2.15	0.82	2.61	0.00907	0.0323
RPL29P19	37.36	1.19	0.31	3.88	0.000106	0.00223	4.31	2.08	0.72	2.87	0.00405	0.0163
WDR66	21.65	1.3	0.4	3.26	0.00112	0.0174	41.75	1.01	0.23	4.45	8.42E-06	6.66E-05

4.4.6. 96 of the overlapping differentially expressed genes are downregulated with a log_2 fold change<-1 and FDR <0.05

I finally set out to narrow down the list of 973 overlapping differentially expressed genes to identify a list of genes with a log_2 fold change of <-1 and FDR<0.05. These would be genes which have shared decreased expression within the two knockout cell lines compared to controls. I then set out to identify which pathways are altered by the downregulation of these genes, if any.

Only 1279 of the 7097 FDR <0.05 genes in the B6 *AS3MT*$^{-/-}$ and 211 of the 1556 FDR <0.05 genes in the C6 *AS3MT*$^{-/-}$ cell line show a log_2 fold change decrease <-1. This corresponds to 18.02% and 13.56% of the DEGs identified have a log_2 fold change of <-1 in both cell lines respectively (**Figure 4.19**). Of these genes, only 96 of the 973 (9.87%) overlapping genes have a log_2 fold change <-1 (**Figure 4.20**), these are listed in **Table 4.11**.

Given the functions of the downregulated genes identified in both knockout lines it is no surprise that there is an enrichment in axon, synapse assembly and collagen containing extracellular matrix GO Terms (**Table 4.12**). These results indicate that the *AS3MT* knockouts have decrease in gene expression in genes that are involved in synapse formation and axon development and thus the formation of functional neurons.

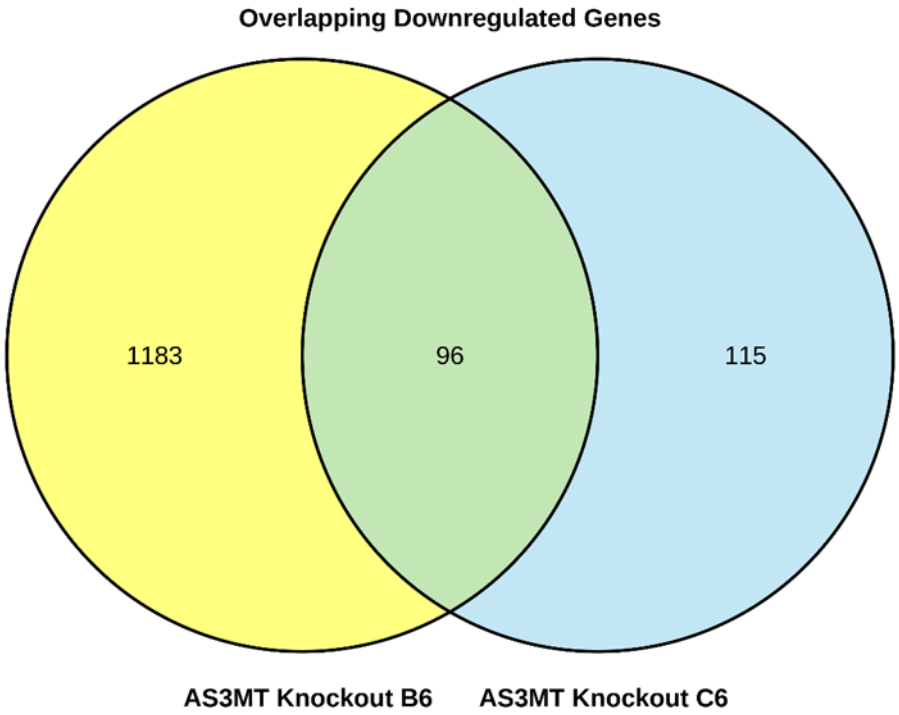


Figure 4.18: Venn diagram shows that there are 96 differentially expressed downregulated genes (log_2 fold change <-1) which overlap between the two AS3MT knockout cell lines. *1183 are specific to the AS3MT B6 knockout cell line and 115 to the AS3MT knockout C6*

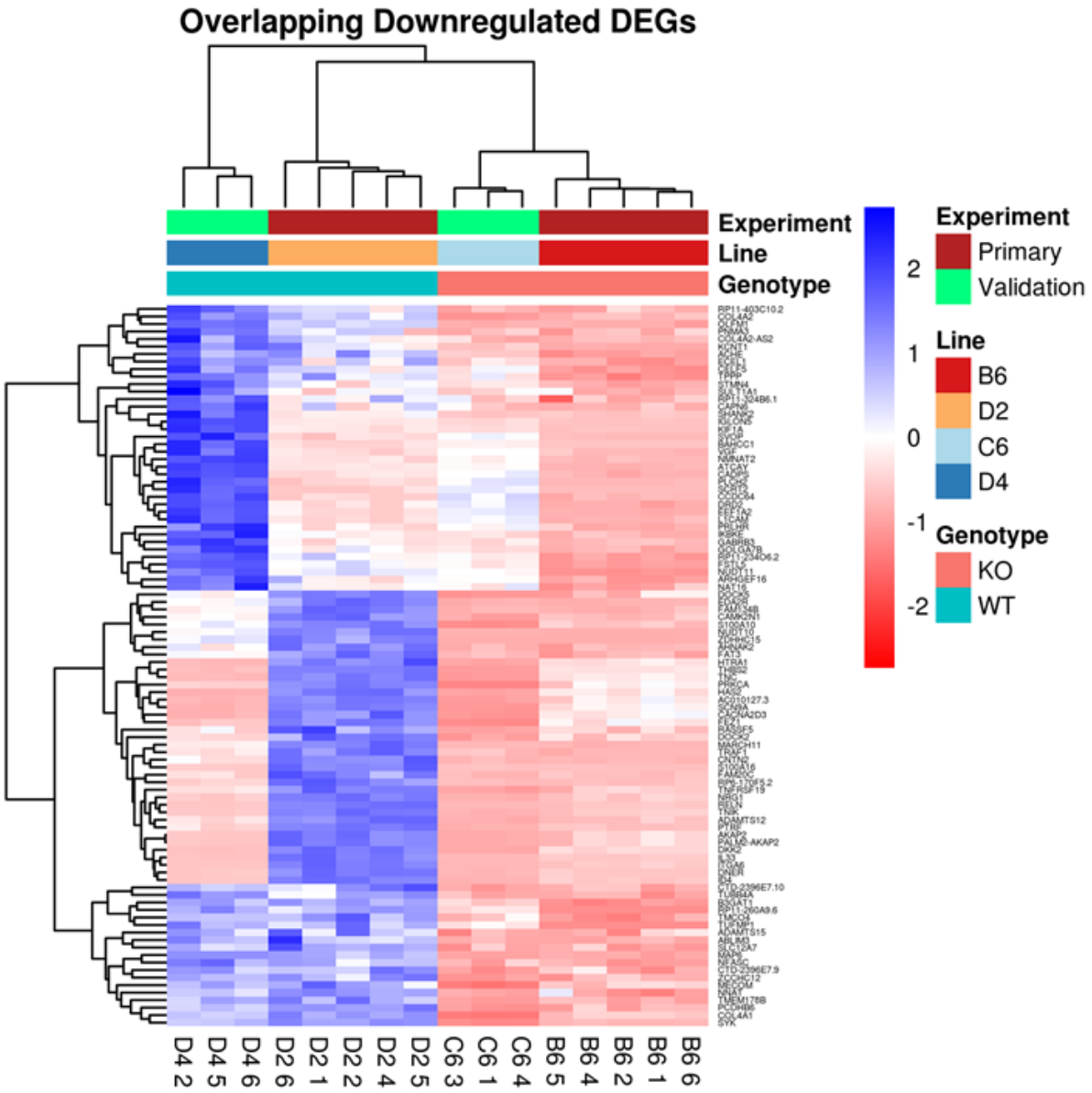


Figure 4.19: Heatmap showing the normalised read count of the 96 downregulated DEGs with log$_2$ fold change <-1 which overlap between the two datasets. *The DEGs group firstly by AS3MT genotype and then by cell line. Each column is a sample, each row is a gene and each row has been scaled where blue is increased read count and red decreased.*

Table 4.11: Overlapping downregulated DEGs between the two AS3MT knockout lines (log_2 fold change <-1 and FDR p-adj <0.05). *Values provided are the base expression, log_2 fold change, log_2 fold change standard error (lfcSE), wald statistic (stat), raw p-value (pvalue) and BH corrected p-value (padj). Values for both experiments are shown. Genes are orders by rank of ranks for average padj and average log_2 fold change.*

Gene Name	baseMean D4	log_2FoldChange C6	lfcSE	stat	pvalue	Padj C6	baseMean D2	log_2FoldChange B6	lfcSE	stat	pvalue	Padj B6
SYK	152.25	-3.43	0.21	-16.43	1.17E-60	1.61E-57	217.19	-1.92	0.12	-16.35	4.46E-60	7.64E-58
IGLON5	222.69	-3.49	0.18	-19.35	1.88E-83	6.03E-80	33.57	-4.23	0.43	-9.79	1.22E-22	4.95E-21
KIF1A	814.31	-3.07	0.11	-28.57	1.80E-179	3.46E-175	127.71	-3.01	0.24	-12.44	1.66E-35	1.27E-33
RELN	478.06	-1.51	0.1	-15.11	1.34E-51	1.07E-48	4004.87	-3.53	0.2	-17.78	1.02E-70	2.18E-68
IL33	87.99	-2.56	0.24	-10.81	3.19E-27	8.76E-25	1684.17	-3.57	0.1	-35.41	1.32E-274	3.07E-271
ZDHHC15	28.14	-8.26	1.21	-6.82	8.96E-12	7.80E-10	53.37	-8.29	0.93	-8.92	4.55E-19	1.50E-17
EDA2R	42.61	-2.98	0.36	-8.35	7.08E-17	1.04E-14	91.4	-3.97	0.26	-15.03	4.96E-51	6.80E-49
FSTL5	214.85	-1.61	0.14	-11.37	5.68E-30	1.79E-27	76.6	-3.59	0.28	-12.84	9.39E-38	8.11E-36
COL4A1	518.51	-3.13	0.12	-27.17	1.65E-162	1.59E-158	804.95	-1.32	0.09	-14.36	8.94E-47	1.07E-44
NUDT11	136.42	-1.27	0.17	-7.6	2.88E-14	3.36E-12	51.43	-5.25	0.47	-11.13	9.19E-29	5.18E-27
MARCH11	329.97	-1.49	0.11	-12.96	2.03E-38	9.53E-36	757.02	-3.08	0.1	-30.45	1.15E-203	1.58E-200
S100A10	210.71	-2.37	0.16	-15.27	1.17E-52	1.02E-49	403.83	-1.67	0.11	-15.88	8.64E-57	1.40E-54
TMEM178B	21.15	-4.61	0.72	-6.4	1.60E-10	1.19E-08	25.78	-3.08	0.41	-7.6	3.05E-14	6.96E-13
S100A16	119.59	-1.24	0.19	-6.57	4.95E-11	3.97E-09	960.62	-5.04	0.13	-37.66	0	0
KCNT1	22.3	-2.99	0.49	-6.13	8.54E-10	5.57E-08	10.58	-5.94	1	-5.92	3.24E-09	4.41E-08
ATCAY	519.06	-1.73	0.1	-17.54	7.13E-69	1.25E-65	109.35	-2.31	0.18	-12.65	1.06E-36	8.64E-35
CAMK2N1	84.89	-2.38	0.25	-9.57	1.09E-21	2.14E-19	198.3	-2.33	0.16	-14.82	1.05E-49	1.38E-47
THBS2	185.56	-2.37	0.17	-14.13	2.56E-45	1.54E-42	1192.31	-1.63	0.07	-24.48	2.17E-132	1.29E-129
RP11-260A9.6	50.18	-1.33	0.27	-5.01	5.37E-07	2.05E-05	32.95	-8.18	0.94	-8.7	3.24E-18	1.00E-16
COL4A2	621.8	-2.67	0.11	-23.35	1.32E-120	6.36E-117	527.73	-1.28	0.12	-10.49	9.67E-26	4.70E-24
ZCCHC12	75.96	-2.37	0.25	-9.69	3.40E-22	6.89E-20	88.15	-1.76	0.22	-8.11	5.06E-16	1.35E-14
BAHCC1	81.77	-2.48	0.24	-10.31	6.57E-25	1.58E-22	5.94	-5.09	1.05	-4.87	1.13E-06	1.06E-05
ID4	57.15	-1.8	0.26	-6.84	7.71E-12	6.74E-10	382.65	-2.89	0.12	-24.6	1.12E-133	7.27E-131
NUDT10	11.4	-4.46	0.94	-4.76	1.95E-06	6.44E-05	27.15	-7.89	0.94	-8.41	4.00E-17	1.15E-15

RP11-234O6.2	59.06	-1.52	0.25	-6.13	8.72E-10	5.67E-08	20.65	-3.6	0.51	-7.1	1.25E-12	2.44E-11
ABLIM3	31.9	-2.38	0.37	-6.38	1.74E-10	1.28E-08	30.53	-2.45	0.35	-6.96	3.49E-12	6.59E-11
TUFMP1	51.04	-1.19	0.27	-4.49	7.27E-06	0.000208	31.68	-8.11	0.94	-8.59	9.01E-18	2.71E-16
SHANK2	34.89	-5.09	0.64	-7.91	2.67E-15	3.45E-13	8.13	-2.68	0.63	-4.26	2.06E-05	0.00015
ITGA6	181.34	-1.12	0.14	-7.79	6.50E-15	8.22E-13	1212.15	-2.79	0.11	-26.41	1.19E-153	9.56E-151
OLFM1	478.65	-1.93	0.1	-18.82	5.09E-79	1.09E-75	347.49	-1.42	0.1	-14.27	3.59E-46	4.17E-44
DKK2	232.66	-1.5	0.13	-11.28	1.62E-29	5.01E-27	1397.05	-1.99	0.09	-22.89	5.85E-116	2.68E-113
DNER	19.92	-2.32	0.47	-4.97	6.86E-07	2.56E-05	169.17	-3.39	0.17	-20.07	1.24E-89	3.86E-87
CNTN2	10.59	-3.28	0.77	-4.27	1.97E-05	0.000504	37.91	-5.45	0.58	-9.44	3.68E-21	1.38E-19
NMNAT2	228.77	-1.86	0.14	-13.13	2.36E-39	1.16E-36	41.32	-2.03	0.28	-7.15	8.90E-13	1.76E-11
GABRB3	78.21	-2.47	0.25	-10	1.56E-23	3.37E-21	16.82	-2.3	0.46	-5.03	4.82E-07	4.80E-06
FAM20C	16.04	-2.55	0.56	-4.51	6.39E-06	0.000185	73.59	-3.69	0.29	-12.84	9.99E-38	8.59E-36
GOLGA7B	47.82	-2.12	0.3	-7.14	9.15E-13	9.02E-11	15.81	-2.27	0.43	-5.24	1.60E-07	1.71E-06
TNFRSF19	51.97	-1.88	0.28	-6.8	1.05E-11	9.02E-10	145.77	-1.97	0.16	-12.44	1.59E-35	1.23E-33
RP11-403C10.2	21.65	-3.31	0.53	-6.26	3.77E-10	2.58E-08	11.98	-2.28	0.52	-4.36	1.28E-05	9.71E-05
TNIK	177.65	-1.19	0.15	-8.19	2.70E-16	3.82E-14	797.66	-2.4	0.11	-22.47	7.83E-112	3.26E-109
SVOP	55.8	-1.66	0.26	-6.31	2.84E-10	2.01E-08	5.35	-4.32	1.06	-4.07	4.75E-05	0.000319
B3GAT1	43.75	-1.38	0.29	-4.83	1.37E-06	4.73E-05	35.89	-3.46	0.35	-9.8	1.13E-22	4.61E-21
RP6-170F5.2	10.62	-2.62	0.66	-3.96	7.43E-05	0.00163	72.81	-4.64	0.34	-13.46	2.58E-41	2.56E-39
PLCH2	43.02	-1.41	0.29	-4.82	1.40E-06	4.79E-05	3.67	-4.4	1.1	-4.01	6.10E-05	0.0004
EEF1A2	1842.63	-1.14	0.08	-14.58	3.65E-48	2.51E-45	390.19	-1.77	0.16	-11.15	6.98E-29	3.96E-27
MAP6	273.06	-1.49	0.12	-12.25	1.71E-34	6.57E-32	254.01	-1.25	0.11	-11.46	2.20E-30	1.34E-28
TUBB4A	47.17	-1.96	0.29	-6.74	1.59E-11	1.34E-09	38.1	-1.79	0.29	-6.22	4.87E-10	7.31E-09
PRKCA	500.03	-1.91	0.1	-19.12	1.76E-81	4.84E-78	1568.82	-1.09	0.12	-8.98	2.64E-19	8.88E-18
MECOM	74.76	-2.59	0.25	-10.23	1.39E-24	3.25E-22	99.69	-1.06	0.18	-5.96	2.58E-09	3.55E-08
TNC	160.67	-1.42	0.15	-9.2	3.58E-20	6.37E-18	795.9	-1.35	0.08	-16.57	1.22E-61	2.19E-59
CTD-2396E7.10	24.57	-2.1	0.4	-5.22	1.77E-07	7.65E-06	29.81	-1.85	0.33	-5.67	1.41E-08	1.75E-07
CADPS	120.21	-1.44	0.18	-8.15	3.66E-16	5.10E-14	23.17	-2.2	0.37	-5.95	2.61E-09	3.59E-08
SCN9A	549.49	-1.18	0.1	-11.53	9.57E-31	3.17E-28	2893.69	-1.21	0.09	-13.11	2.91E-39	2.66E-37
PTRF	122.88	-1.18	0.19	-6.1	1.07E-09	6.90E-08	352.59	-2.36	0.1	-22.49	5.81E-112	2.46E-109

HAS2	17.91	-2.32	0.48	-4.79	1.68E-06	5.64E-05	286.79	-1.71	0.13	-13.04	6.98E-39	6.28E-37
ADAMTS12	69.24	-1.28	0.23	-5.55	2.90E-08	1.44E-06	179.08	-2.22	0.14	-15.43	9.86E-54	1.43E-51
L1CAM	1762.8	-1.24	0.09	-14.47	1.93E-47	1.28E-44	435.35	-1.18	0.14	-8.32	8.81E-17	2.49E-15
NFASC	29.79	-1.84	0.37	-5.01	5.36E-07	2.05E-05	23.48	-1.98	0.33	-5.95	2.60E-09	3.58E-08
IKBKE	12.71	-2.44	0.58	-4.19	2.76E-05	0.000678	3.21	-2.97	1.01	-2.92	0.00345	0.0143
COL4A2-AS2	14.02	-2.9	0.62	-4.64	3.46E-06	0.000108	7.46	-1.94	0.58	-3.36	0.000772	0.00391
FEZ1	118.15	-1.42	0.18	-7.97	1.58E-15	2.09E-13	380.82	-1.05	0.11	-9.53	1.60E-21	6.11E-20
TPPP	42.68	-1.15	0.29	-4.04	5.44E-05	0.00124	22.08	-3.03	0.42	-7.18	6.79E-13	1.36E-11
SCRT2	30.18	-1.48	0.34	-4.36	1.30E-05	0.000349	1.54	-3.74	1.31	-2.85	0.004345	0.017364
DOCK5	313.87	-1.13	0.12	-9.69	3.28E-22	6.71E-20	514.01	-1.22	0.15	-7.99	1.38E-15	3.57E-14
AHNAK2	78.91	-1.24	0.23	-5.31	1.12E-07	5.08E-06	135.12	-2.18	0.17	-13.07	4.95E-39	4.47E-37
CTD-2396E7.9	16.74	-2.3	0.5	-4.58	4.70E-06	0.000142	18.44	-1.6	0.37	-4.32	1.53E-05	0.000114
PRLHR	33.52	-1.53	0.33	-4.56	5.05E-06	0.000151	7.13	-2.58	0.7	-3.72	0.000203	0.00119
FAM134B	109.4	-1	0.18	-5.54	3.02E-08	1.49E-06	191.71	-2.1	0.14	-14.69	7.17E-49	9.33E-47
VGF	1787.47	-1.49	0.1	-15.29	9.00E-53	8.24E-50	450.72	-1.07	0.17	-6.25	4.15E-10	6.28E-09
PALM2-AKAP2	45.69	-1.39	0.27	-5.07	3.95E-07	1.57E-05	221.1	-2.04	0.16	-12.95	2.21E-38	1.96E-36
TRAF1	22.26	-1.2	0.4	-2.98	0.00284	0.037741	59.62	-3.75	0.29	-12.74	3.42E-37	2.85E-35
ECEL1	158.6	-1.09	0.16	-6.68	2.44E-11	2.00E-09	109.22	-1.3	0.2	-6.45	1.14E-10	1.85E-09
ARHGEF16	21.48	-1.33	0.4	-3.36	0.000786	0.0128	7.26	-2.94	0.73	-4.05	5.19E-05	0.000346
PCDHB6	140.41	-1.18	0.16	-7.34	2.18E-13	2.26E-11	171.04	-1.13	0.13	-8.81	1.31E-18	4.18E-17
AKAP2	45.69	-1.39	0.27	-5.07	3.95E-07	1.57E-05	217.39	-2.02	0.16	-12.77	2.51E-37	2.09E-35
STMN4	73.02	-1.5	0.23	-6.45	1.09E-10	8.29E-09	33.31	-1.45	0.29	-4.99	5.93E-07	5.81E-06
ACHE	26	-1.5	0.37	-4.06	4.98E-05	0.00115	19.8	-2.31	0.38	-6.06	1.37E-09	1.96E-08
NNAT	30.82	-1.6	0.34	-4.72	2.39E-06	7.74E-05	38.58	-1.89	0.31	-6.17	7.04E-10	1.05E-08
NRG1	68.38	-1.11	0.23	-4.83	1.39E-06	4.76E-05	270.81	-2.27	0.13	-18.08	4.27E-73	9.67E-71
CELF5	53.24	-1.11	0.27	-4.12	3.72E-05	0.000887	24.49	-2.54	0.37	-6.88	5.81E-12	1.07E-10
FAT3	17.3	-1.31	0.44	-2.99	0.00278	0.0370	28.51	-2.84	0.35	-8.16	3.32E-16	8.95E-15
DRD2	109.06	-1.03	0.18	-5.74	9.57E-09	5.20E-07	29.87	-1.46	0.29	-5.01	5.39E-07	5.32E-06
TMCO4	32.06	-1.11	0.33	-3.34	0.00083	0.0134	29.81	-2.69	0.35	-7.62	2.53E-14	5.81E-13
SLC12A7	31.95	-1.48	0.35	-4.27	1.93E-05	0.000495	37.41	-1.82	0.27	-6.73	1.66E-11	2.89E-10

AC010127.3	105.23	-1	0.19	-5.27	1.40E-07	6.18E-06	497.92	-1.31	0.14	-9.53	1.54E-21	5.88E-20
PNMA3	19.58	-2.29	0.46	-4.94	7.85E-07	2.89E-05	11.18	-1.41	0.48	-2.93	0.00334	0.0139
ADAMTS15	9.84	-2.14	0.65	-3.31	0.00094	0.0150	11.04	-1.66	0.48	-3.46	0.000542	0.00285
NAT16	34.69	-1.25	0.32	-3.86	0.000114	0.00238	13.2	-1.68	0.44	-3.83	0.000127	0.000776
RASSF5	13.12	-1.65	0.53	-3.12	0.00182	0.0259	32.58	-1.79	0.29	-6.1	1.05E-09	1.52E-08
DOCK2	26.05	-1.32	0.36	-3.64	0.000269	0.00511	68.8	-1.57	0.2	-7.74	1.01E-14	2.40E-13
CAPN6	49.83	-1.61	0.28	-5.74	9.49E-09	5.20E-07	26.78	-1.05	0.31	-3.42	0.000633	0.00327
CCDC64	92.2	-1.16	0.2	-5.76	8.48E-09	4.73E-07	13.55	-1.43	0.42	-3.4	0.000666	0.00342
HTRA1	30.1	-1.12	0.33	-3.36	0.000784	0.0128	189.33	-1.65	0.14	-11.75	6.98E-32	4.61E-30
SULT1A1	26.06	-1.8	0.4	-4.45	8.70E-06	0.000244	11.04	-1.26	0.49	-2.56	0.0106	0.0368
CACNA2D3	30.78	-1.25	0.33	-3.77	0.000162	0.00326	158.43	-1.12	0.15	-7.7	1.38E-14	3.26E-13
RP11-324B6.1	37.29	-1.26	0.31	-4.03	5.54E-05	0.00126	21.2	-1.15	0.37	-3.16	0.00159	0.00731

Table 4.12: GO terms identified in the overlapping downregulated genes.

category	over_represented_pvalue	under_represented_pvalue	numDEInCat	numInCat	term	ontology	correctedP
GO:0030424	6.43E-07	1	11	241	axon	CC	0.0106
GO:0062023	1.85E-06	1	9	181	collagen-containing extracellular matrix	CC	0.0110
GO:0007416	2.01E-06	1	6	53	synapse assembly	BP	0.0110

4.5. Discussion

In this chapter, I quantified gene counts and profiled gene expression within the knockout *AS3MT* cell lines generated in Chapter 3. compared to matched WT cell lines. To my knowledge, this is the first experiment to quantify gene expression changes resultant of knockout of *AS3MT* in a neuronal cell line to elucidate a role for *AS3MT* in neuronal function. In two separate experiments using different *AS3MT* KO cell lines I have confirmed a significant reduction of gene expression of *AS3MT,* I then identified a list of DEGs in both knockout lines and then cross-referenced these lists to find overlapping gene expression changes, with further analysis into overlapping upregulated and downregulated genes with a $\log_2$ fold change +/-1. I found a positive correlation in the direction of effect of significant DEGs in both knockout cell lines, identifying robust changes in gene expression resulting from different knockout lines of *AS3MT* (Pearson's correlation coefficient, r=0.52, p-value=2.2e-16). Analysis identified that the significant set of overlapping DEGs are enriched for functions involved in protein targeting, translation, nonsense mediated decay, cell adhesion and cell-cell interactions, neuron development and immune system responses. The top pathway enriched in the DEGs was SRP-dependent cotranslational protein targeting to membranes and the second highest enriched GO term is for nuclear-transcribed mRNA catabolic process, nonsense mediated decay.

Analysis into the list of DEGs identified genes which have previously been associated with schizophrenia through GWAS. To test the theory that schizophrenia genes are enriched in the DEG list following knockout of *AS3MT* I undertook an enrichment analysis by examining the number of genes in the DEG list which have also been identified in the most recent GWAS of schizophrenia (Pardiñas *et al.*, 2018). I identified that in both *AS3MT* knockout lines that there is an enrichment on genes associated with schizophrenia in the list of DEGs produced by *DESeq2* (Fishers exact test primary experiment p-value=0.0146, validation experiment p-value=0.0129). Furthermore, I identified that of these 27 schizophrenia genes, 25 have the same direction of effect in both cell lines. This could signify that *AS3MT* sits within a network of genes involved in neurodevelopment all of which are dysregulated following knockout. In the wider field this could be instrumental as current methods to identify genetic risk do not include gene networks and only examine genes or SNPs on a gene by gene basis.

Because of this, methods to identify individual genetic risk should consider the knock on effect of gene networks in risk gene in an additive model rather than single genes.

Further analysis into the overlapping genes identified upregulation of genes involved in neurotransmitter metabolism, differentiation, cellular morphology and homeobox activity. There was no enrichment in any specific functional pathway, indicating the list upregulated genes have no shared function but instead a broad range of functions. One result that is clear is that there is upregulation of genes responsible for differentiation and cellular metabolism, these genes could be counter mechanisms to restore normal cellular function following knockout or could be expressed due to uncontrolled regulation by *AS3MT.* Analysis into the downregulated genes identified several pathways involved in axon development, synapse assembly and extracellular matrix formation being downregulated. One of the top downregulated genes in both cell lines is *ZDHHC15* which has previously been associated with neuronal development, schizophrenia models, and immature dopaminergic neurone formation in both cell lines and zebrafish (Mukai *et al.*, 2008; Fukata and Fukata, 2010; Wang *et al.*, 2015; Shah, Shimell and Bamji, 2019). Further genes involved with vesicle tracking, translation, potassium channels, synapse formation and adhesion molecules identifies that there are global changes in genes involved with cellular morphology and programming.

The gene with the largest combined $\log_2$ fold change increase between the two experiments is *BCHE*, which encodes for serine hydrolase enzyme butyrylcholinesterase (BChE) (average $\log_2$ fold change +10.74). BChE co-regulates the expression of the neurotransmitter acetylcholine in the brain by hydrolysing acetylcholine from within the synaptic cleft (Darvesh, Hopkins and Geula, 2003). BChE is expressed in the thalamus, amygdala and hippocampus and is expressed in neurons (Darvesh, Grantham and Hopkins, 1998; Darvesh and Hopkins, 2003) and is thought to play a role in nervous system development (Layer, 1991), this is supported by its upregulation during iPSC differentiation to neuronal fates (Tiethof, Richardson and Hart, 2018). BChE has also been implicated in schizophrenia, where increased expression of BChE was observed in ketamine dosed rats, a known model of schizophrenia, and following chronic exposure to cigarette smoke during pregnancy (Zugno *et al.*, 2013). Furthermore, BChE has been identified to be upregulated in schizophrenic patients, and when patients are treated with rivastigmine, an inhibitor of

BChE and acetylcholinesterase, their quality of life is improved identified through improvements in their satisfaction with life domains scale scores (Lenzi *et al.*, 2003). Therefore it is interesting that knockout of *AS3MT* causes upregulation of BChE in SH-SY5Y cells. Other genes highly upregulated between the two cell lines include, *MOXD1* (log_2 fold change +6.48), a gene encoding monooxygenase DBH-like 1, which has been identified to be upregulated in the chemical conversion of astrocytes into neurons (Ma, Yin and Chen, 2019). *HOXC11* (log_2 fold change +5.93), a member of the homeobox gene superfamily which encode for transcription factors involved in morphogenesis and differentiation during embryogenesis, *HOXC11* has been found to be upregulated during differentiation of GOTO neuroblastoma cells into Schwannian cells *in vitro* (Zhang *et al.*, 2007). There is also upregulation of *HOTAIR* (log_2 fold change +4.64), a non-coding RNA which is co-expressed with *HOXC* genes to repress *HOXD* via histone trimethylation at H3-K27 (Rinn *et al.*, 2007). *HPDL* (log_2 fold change +4.05) encodes for 4-hydroxypheylpryuvate dioxygenase like protein, which is involved with oxidoreductase activity. Interestingly there is upregulation of a schizophrenia risk gene, *ZNF804A* (log_2 fold change +1.33) in both knockout cell lines. *ZNF804A* has been identified in multiple GWAS studies of schizophrenia and is a robust finding between different ethnicities (O'Donovan *et al.*, 2008; Ripke *et al.*, 2013). Knockdown on *ZNF804A* in SH-SY5Y cells has been shown to alter mRNA processing involved in nervous system development (Chapman *et al.*, 2018).

The gene with the largest combination log_2 fold change decrease across both experiments is *ZDHHC15*, which encodes for zinc finger DHHC-Type containing 15 (zDHHC15) (average log_2 fold change -8.28). zDHH enzymes are involved in the post translational modification by S-acetylation (palmitoylation), which is important for shuttling proteins towards membranes, regulation of cellular signalling and neuronal development (Fukata and Fukata, 2010). *ZDHHC15* expression is highly expressed in the brain compared to other tissues (Fukata *et al.*, 2004). Palmitoylation deficits in have been identified in the 22q11 mouse model of schizophrenia (Mukai *et al.*, 2008), and knockdown of the zDHHC15 homolog in zebrafish results in immature dopaminergic neuron formation (Wang *et al.*, 2015), and one report identified loss of *ZDHH15* in a female patient with X-linked intellectual disability (Mansouri *et al.*, 2005). Furthermore, *ZDHH15* knockdown in primary hippocampal cultures showed inhibition of dendritic outgrowth, formation of mature spines, and reduction in excitatory synapse

density (Shah, Shimell and Bamji, 2019). The reduction of *ZDHH15* in the *AS3MT* knockout cell line indicates that there is a change in the signalling reducing dendritic outgrowth and the formation of mature neurones. Other genes of interest that are highly downregulated in the two KO cell lines include *NUDT10* (log_2 fold change -6.17) which encodes for Nudix Hydrolase 10, a protein involved in metabolism of diphosphoinositol polyphosphate a molecule involved in apoptosis, vesicle trafficking and stress responses (Thomas and Potter, 2014; Carreras-puigvert *et al.*, 2017). *TUFMP1* (log_2 fold change -4.65), a pseudogene of *TUFM* (log_2 fold change -4.65), a mitochondrial gene involved in translation of mitochondrial genes. *KCNT1* (log_2 fold change -4.47), which encodes for potassium channel and has been associated with epilepsy (Mctague *et al.*, 2018). *CNTN2* (log_2 fold change -4.37), a member of the contactin family of cell adhesion molecules, involved in outgrowth of neurites and has been associated with Alzheimer's Disease (Chatterjee, Schild and Teunissen, 2019). There are also a number of genes down regulated in both datasets which have been previously implicated in schizophrenia. *CACNA2D3* (log_2 fold change -1.19) is a voltage gated calcium channel which has been identified in the Chinese population (Zhang *et al.*, 2018). *DRD2* (log_2 fold change -1.25), or dopamine receptor D2 is one of the most robust genes associated with schizophrenia and the dopamine hypothesis (Schizophrenia Working Group of the Psychiatric Genomics Consortium, 2014). Dopamine receptor D2 (D2R) is one of the key targets for pharamacological intervention in schizophrenia, most current antipsychotics are antagonists of D2R, this is of no surprise as dopaminergic pathways are disrupted in schizophrenia (for a review on antipsychotics and dopamine signalling see (Rampino *et al*., 2019)). *GABRB3* (log_2 fold change -2.39) the GABA receptor beta 3 receptor expression has been predicted to be reduced following luciferase assays in SH-SY5Y cells with the schizophrenia risk alleles (Liu *et al.*, 2019). *NRG1* (log_2 fold change -1.75) and *RELN* (log_2 fold change -2.52) have both been extensively studied in schizophrenia, with knockout animal models used as a models for schizophrenia (Stefansson *et al.*, 2002). Finally *SHANK2* (log_2 fold change -3.89), a gene which encodes for the SHANK scaffolding proteins in glutamatergic synapses, and has been previously associated with schizophrenia and autism (Guilmatre *et al.*, 2014).

There are some limitations to this study, which should be taken into consideration when interpreting the results. First of all, the two sequencing experiments were done

at separate times introducing batch effects with regards to culturing conditions. While every effort was attempted to keep the cell lines at the same or similar confluency (80%) there might be some experimental variation, which in turn could impact on the resultant gene expression profiles. Secondly, this analysis was performed on bulk SH-SY5Y cells, while these cells have been genotyped and are monoclonal for *AS3MT* there are known to be two distinct sub-populations of SH-SY5Y cells, consequently the composition of the culture of each of these cell populations would contribute to the observed outcomes, if they differ between cell lines. Equally the passage of the cell lines, SH-SY5Y cells are known to spontaneously differentiate at higher passages, these cells were cultured for a long period of time and are reaching the point of spontaneous differentiation, this is why the *AS3MT* knockout cell line C4 and wildtype cell line D3 were not selected for sequencing as they have already undergone differentiation. With regards to analysis I only looked at total gene expression changes, while I have identified significant gene expression changes there could be transcripts within non DEGs which have been differentially expressed which I have not identified, thus missing out on a range of transcripts. The final point of note is that, while the wild-type *AS3MT* cell lines should be identical, I have not carried out analysis to identify gene expression changes between these two cell lines, there are too many confounding factors including culturing conditions, confluency, library preparations and sequencing run, which could all impact on the accuracy of the results. In the future it would be beneficial to run experiments at the same time to provide multiple comparisons to identify if the CRISPR methodology results in any significant changes.

4.6. Conclusion

These results provide support that *AS3MT* is not only involved with arsenic metabolism but it has an important role in neuronal cells during brain development, with a particular focus on the formation neurons. These experiments show that SH-SY5Y cells lacking AS3MT have altered neuronal differentiation and are unlikely to form functional axons and synapses to result in successful neurotransmission. However the mechanism in which *AS3MT* is involved in these processes still needs to be elucidated. It could be that the eQTL for *AS3MT* at rs7085103, which has been associated with schizophrenia and increased expression of the $AS3MT^{d2/d3}$ isoform (Duarte *et al.*, 2016; L. Li *et al.*, 2016; M. Li *et al.*, 2016) results in a non-functional isoform which acts as a repressor

of *AS3MT*^{full} expression. Or it could be that this isoform has an independent function in neuronal development and is required for normal neuronal differentiation.

Chapter 5. Functional analysis of differentially methylated regions using luciferase reporters

The work presented in this chapter is based on the functional validation work I carried out with Stefania Policicchio, a fellow PhD student in our lab (Policicchio *et al.* 2020). A copy of the published manuscript can be found in **Appendix B**.

5.1. Introduction

EWAS studies result in hundreds or thousands of CpG sites associated with disease phenotypes and assessing the functional relevance of these changes is critical to understanding disease aetiology. Further pipelines in EWAS aim to detect differentially methylated regions (DMRs), these are contiguous genomic regions which have differential methylation between phenotypes (Rakyan *et al.*, 2011). DMRs are thought to be more functionally relevant than single methylation changes as they are usually identified around promoters, within gene bodies, and at intergenic regulatory regions (Jones, Baylin and Kimmel, 2002; Suzuki and Bird, 2008). An abundance of software has been developed to statistically identify DMRs from EWAS data, such as Comb-p, DMRcate, bumphunter, and dmrff (Jaffe *et al.*, 2012; Pedersen *et al.*, 2012; Peters *et al.*, 2015; Suderman, French and Arathimos, 2018). A recent study by our lab (Murphy *et al.*, 2017) set out to identify DMRs within depressed suicide completers compared to non-psychiatric sudden-death controls. DNA methylation was profiled in two cortical brain regions (Brodmann Area 11 and 25) from 20 depressed suicide completers and 20 non-psychiatric sudden-death controls using the Illumina HumanMethylation450 BeadChip array. Comb-p was then used to identify DMRs between depressed suicide completers and healthy controls. The top ranking DMR from this study was a region within the promoter of *PSORS1C3* non-coding gene where 13 different CpGs are all hypomethylated in depressed suicide completers compared to non-psychiatric sudden-death controls (p-value=8.02×10^{-6}) (**Figure 5.1**). Interestingly, this same DMR was also identified in an analysis of discordant schizophrenia twins (Dempster et al in prep). While the methylation state of the *PSORS1C3* DMR was validated using pyrosequencing there was no further work to confirm if this DMR is an active promoter and able to initiate gene expression. One way in which we can identify this is through a luciferase reporter.

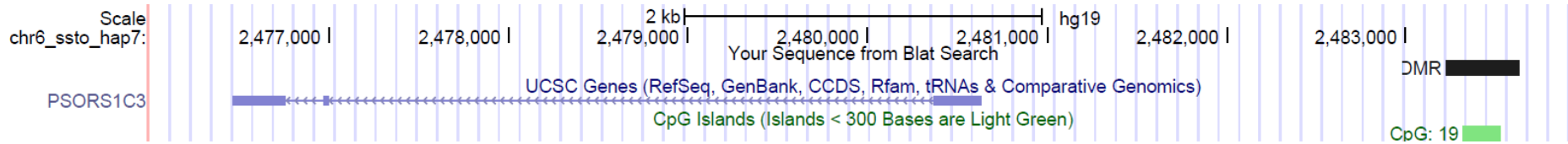


Figure 5.1: The genomic context of the PSORS1C3 DMR identified by Murphy et al 2017*. The DMR was identified to be hypomethylated in suicide completers and sits within a CpG island upstream of the PSORS1C3 gene. Image from UCSC genome browser from hg19 human genome build.*

Luciferase reporter assays emerged in the late 1980's following the isolation and cloning of firefly luciferase from the common firefly (*Photinus pyralis)* (Wet *et al.*, 1985). Firefly luciferase converts luciferin to oxyluciferin in the presence of ATP, oxygen, and Mg^{2+} and releases light as a by-product, therefore a measure of activity can be gauged as activity is directly proportional to light emitted (Marques and Esteves, 2009). Further to this, another luciferase extracted from the sea pansy, *Renilla reniformis,* is also regularly used in reporter assays (Roda *et al.*, 2009). The *Renilla* luciferase uses coelenterazine as a substrate and is converted into coelenteramide in the presence of oxygen, again releasing light as a by-product. These two luciferases have been used in tandem to develop a dual luciferase assay, where two distinct signals can be measured. In most cases the reporter construct contains Firefly luciferase which, when a promoter is cloned upstream, results in expression of the firefly luciferase. The *Renilla* luciferase is used as a transfection control, to normalise the firefly luciferase activity based on differing transfection efficiencies per experiment (Stables *et al.*, 2009). The luciferase assay is therefore an important tool in molecular biology, and has also been developed for use in detection of ATP levels through fusion of luciferase to biotin acceptor peptide (Nakamura *et al.*, 2006), *in vivo* imaging (Iii and Szalay, 2002) and protein-protein interactions (Massoud *et al.*, 2007) and can be used to examine the effect of promoter methylation on luciferase activity (Klug and Rehli, 2006).

Promoter regions are important regulators of gene expression as they harbour binding sites for transcription factors, and methylation of these regions is well known to control gene expression (Smale and Kadonaga, 2003; Juven-gershon and Kadonaga, 2010). The luciferase approach for examining promoter activity involves the cloning promoter fragments upstream of the luciferase gene to drive expression of luciferase (Klug and Rehli, 2006). Therefore by cloning a DMR of interest into the luciferase reporter, if an increase in luciferase activity is seen, then the DMR has promoter activity. Further, to determine if DNA methylation is able to silence the promoter, the CpGs within the DMR can be artificially methylated *in vitro* using the commercial enzyme *M.sssI*. This would results in a reduction of the firefly luciferase activity if DNA methylation is silencing expression (Klug and Rehli, 2006). One limitation of this methodology is that early reporter vectors contained high numbers of CpG dinucleotides which can be methylated and could impact on the activity of firefly luciferase. To reduce this the CpG

free firefly luciferase reporter, henceforth known as pCpGL, was developed (Klug and Rehli, 2006). These luciferase assays have previously been used to validate DMRs of interest in suicide (Taylor *et al.*, 2014), pre-natal famine (Tobi *et al.*, 2014), to examine the effect of promoter SNPs on promoter activity (Jeonghae *et al.*, 2012), and to identify enhancer element methylation (Andres *et al.*, 2014). Furthermore it has been shown that luciferase activity is dependent on sequence of known promoter activity as cloning of random DNA sequences has no effect on luciferase activity (Taylor *et al.*, 2014). I have adopted this into our functional epigenetic toolbox and have used this to functionally examine the promoter ability of the *PSORS1C3* DMR identified in (Murphy *et al.*, 2017).

The *PSORS1C3* DMR was PCR cloned into pCpGL upstream of the firefly luciferase gene, to determine if this DMR acts as a promoter (**Figure 5.1**). The recombinant construct was then methylated *in vitro* using *M.sssI.* The methylated and unmethylated constructs were then transfected into HEK293 cells, along with a *Renilla* luciferase transfection control, and dual luciferase expression was recorded 24 hours post transfection. The same procedure was carried out with the empty pCpGL vector, this was to confirm the backbone vector is having no effect on luciferase activity.

5.2. Aims

The general aims of this chapter were to develop a luciferase based assay for analysing the promotional ability of the *PSORS1C3* DMR identified through EWAS and to confirm that methylation alters the expression of luciferase. In order to achieve this there were several aims:

1. To clone the *PSORS1C3 DMR* into pCpGL Basic.
2. To assess if pCpGL, when methylated or unmethylated, has any effect on luciferase expression.
3. To identify if the *PSORS1C3* DMR identified in (Murphy *et al.*, 2017) has promoter activity by examining luciferase expression.
4. To examine if the *PSORS1C3* DMR identified in (Murphy *et al.*, 2017) is methylation sensitive by methylating the region in pCpGL Basic *in vitro* and examining the effect on luciferase expression.

5.3. Methods

5.3.1. Primer design

Primers were designed to amplify around the *PSORS1C3* associated DMR to encompass as many of the 13 CpG probes identified in (Murphy *et al.*, 2017). Two sets of primers were designed and used for generation of multiple constructs. The primers were 5' tagged with the restriction sites for BglII (AGATCT) on the forward primer and NcoI (CCATGG) on the reverse primer, so they can be cloned into pCpGL basic backbone vector. Six extra bases were added at the 5' end to aid in restriction enzyme binding and cleavage (**Table 5.1**). The resultant sequences, including all the 5' tags, are shown in **Table 5.2**. The primers were then paired in different combinations for cloning (**Table 5.3**), and the primer locations of primer set A relative to the hypomethylated probes are shown in **Figure 5.2**.

CATAAAGGCCACAGTTTGGGGAAGGTTATGGCTCAGGGGAAGGGGAGAGGTG
CTAAATAATTAAGCCCCCCTACTACTCAGCACCCGCGTGAGGCATCGTCAGGC
ATCGTCAGGCCTCCAGTGGTGGTGGTGGCACCGGGCCTCAACCTCCCCGGAG
GGCTGGACTCTCGCTGCCAGGCTGTGGGGATCAGGCGTTGTGGGGGAGGGG
GACACTTAACAGGTATGGAGGGCGGAGCAGAGCCCCGCAGTCACTGGCCTGA
CTTCCGGAACGAACCGTCGCCAGCAAGCACAGCAGTAGGACCAGGGGGATGC
AAGAGCGGGGGCGGCCGGGGATCGTGCTTCTCGCTCAGGTCCAGATTCCCG
GCAACCAGGCCGGCGGAATCACGTGCCATGCTCCAGGCCAGCGTAGTCCCG
CCCATCTTCCAGCTGAGCGTACCGGGAGGCTCCCATTGGACTGGAGCTGCTA
CGGAGGCGGGACTTTCCCTTTTTCTTGAACCCCATTGGGTTAAGTCCAGTCCG
AGACAAGCGTCTCTCCTCAGCAGTGGGAGGGGTGATTTGGCT

Figure 5.2: Region 6:31,148,140-31,148,700. ***Purple*** *are the 13 450k probes identified in Murphy et al as hypomethylated in depression and suicide patients compared to healthy controls.* ***Bold*** *CG are other CG sites within this region not covered by the 450K array. Yellow Indicates primer sequences used to generate recombinant PSORS1C3-A-pCpGL-basic*

Table 5.1: Primer sequences and locations for cloning PSORS1C3 DMR into pCpGL. *Including additional tags to be added at the 5' end to aid in cloning.*

Primer Name	Sequence (5'-3')	Chromosomal Location	Strand	Added restriction sites	Added Random bases
PSORS1C3_F1_BglII	GCTTGTCTCGGACTGGACTT	6:31148647-31148666	-	AGATCT	ATGCTG
PSORS1C3_F2_BglII	GTCTCGGACTGGACTTAACCCA	6:31148641-31148662	-	AGATCT	CTAGCT
PSORS1C3_R1_NcoI	GGAAGGGGAGAGGTGCTAAA	6:31148177-31148196	+	CCATGG	GCTTAC
PSORS1C3_R2_NcoI	AGGTTATGGCTCAGGGGAAG	6:31148162-31148181	+	CCATGG	GAACTG

Table 5.2: Final primer sequences cloning PSORS1C3 DMR into pCpGL*. Yellow highlights random six bases to aid with enzyme binding and green shows the restriction sites.*

Primer Name	Final Primer Sequence (5'-3')
PSORS1C3_F1_BglII	ATGCTGAGATCTGCTTGTCTCGGACTGGACTT
PSORS1C3_F2_BglII	CTAGCTAGATCTGTCTCGGACTGGACTTAACCCA
PSORS1C3_R1_NcoI	GCTTACCCATGGGGAAGGGGAGAGGTGCTAAA
PSORS1C3_R2_NcoI	GAACTGCCATGGAGGTTATGGCTCAGGGGAAG

Table 5.3: Combinations of primers used to generate the luciferase constructs used in this chapter.

Primer Set	Construct Name	Forward Primer	Reverse Primer	Amplicon Size (bp)
A	PSORS1C3_pCpGL_A	PSORS1C3_F1_BglII	PSORS1C3_R1_NcoI	490
B	PSORS1C3_pCpGL_B	PSORS1C3_F1_BglII	PSORS1C3_R2_NcoI	505
C	PSORS1C3_pCpGL_C	PSORS1C3_F2_BglII	PSORS1C3_R1_NcoI	486
D	PSORS1C3_pCpGL_D	PSORS1C3_F2_BglII	PSORS1C3_R2_NcoI	501

5.3.2. High-fidelity PCR amplification

100ng of human male DNA (Promega, Wisconsin, USA) was used as a template for amplifying the *PSORS1C3* region in a high fidelity PCR reaction, each primer set had a water negative control. The reactions were set up as outlined in **2.6.1.** . The PCR reaction was run for 35 cycles at temperatures and times outlined in **Table 5.4**. Following PCR amplification products were visualised on a 2% Agarose gel as outlined in **2.4.1.** (**Figure 5.3**). Samples were run alongside 2µl of 100bp ladder (Solis Biodyne, Teaduspargi, Estonia).

Table 5.4: PCR conditions for PSORS1C3 high-fidelity PCR with Phusion polymerase.

Step	Temperature (°C)	Time	Number of Cycles
Hotstart	98	2 min	1
Denaturation	98	30 sec	
Annealing	61	30 sec	35
Extension	72	1 min	
Final Extension	72	10 min	1
Storage	4	∞	1

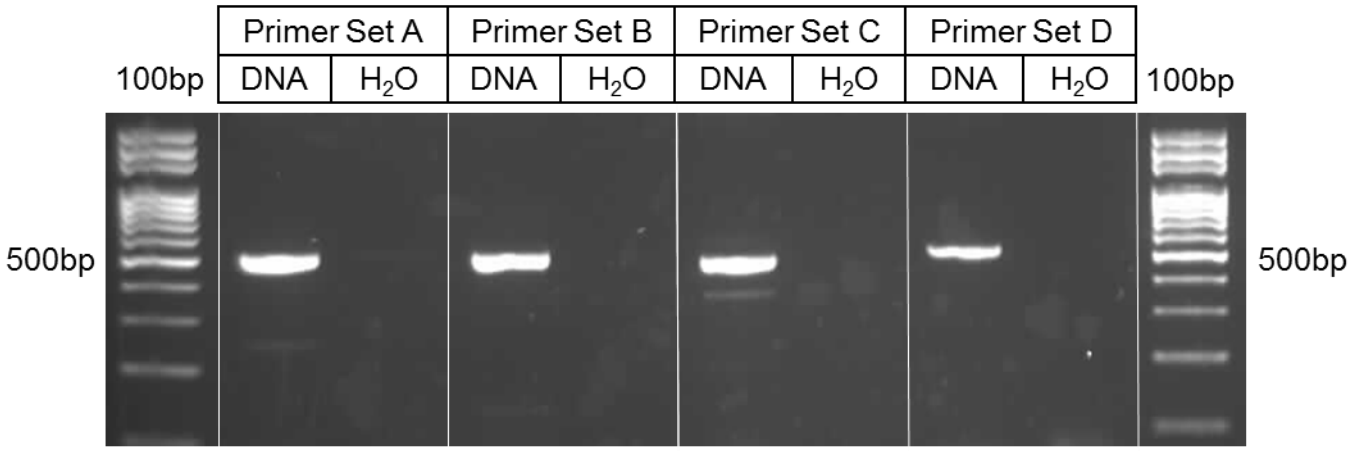


Figure 5.3: Amplification of the PSORS1C3 DMR using Phusion Polymerase. All four different primer combinations resulted in clear bands which were excised and extracted for cloning into pCpGL basic*. Lane 1, 100bp ladder. Lanes 2 and 3 are primer set A, lanes 4 and 5 are primer set B, lanes 6 and 7 are primer set C and lanes 8 and 9 are primer set D. All primer sets contain DNA template on the left and water control on the right. Lane 10 is another 100bp ladder. Samples run on 2% Agarose with Ethidium Bromide for 30 minutes at 120 v in 1 x TBE. Gel was visualised under UV light.*

5.3.3. Gel extraction of PCR product

As shown in **Figure 5.3** primer sets A, B, and C contain multiple amplicons. In order to prevent these amplicons being cloned into pCpGL-basic a gel extraction was performed to isolate the correct size amplicons as described in **2.4.3.** . The optional isopropanol step was included and the samples were eluted in 50µl of elution buffer. Final PCR product concentrations were then calculated using Nanodrop and visualised using agarose gel electrophoresis (**Figure 5.4**).

5.3.4. Digestion of pCpGL-basic and *PSORS1C3* PCR amplicons

Following successful clean up and purification the PCR amplicons and backbone vector, pCpGL-basic, were cut with NcoI and BglII. The use of two different restriction enzymes would allow cloning of the DMR in the correct orientation into the luciferase promoter within pCpGL-basic. 500ng of PCR amplicons and 200ng of pCpGL-basic backbone vector were digested in separate 20µl reactions as outlined in **2.6.2.** . Reactions were incubated in a thermal cycler at 37°C for 1 hour, followed by inactivation at 65°C for 15 minutes. Reactions were then cleaned using the GeneJET PCR clean up kit (ThermoFisher Scientific, Massachusetts, USA) as outlined in **2.4.2.** , with the optional isopropanol step included for the *PSORS1C3* PCR amplicons, reactions were eluted in 25µl elution buffer.

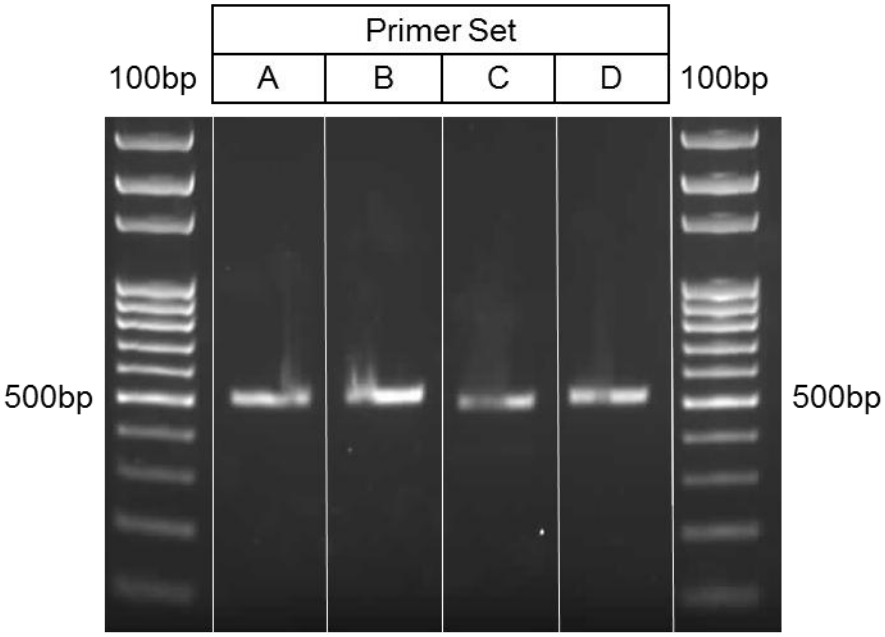


Figure 5.4: Recovered PSORS1C3 PCR product following clean up with the GeneJET gel extraction kit. *Lane 1, 100bp ladder, lane 2, primer set A, lane 3 primer set 3 lane 4 primer set C lane 5 primer set D, lane 6 100bp ladder. Samples run on 1.5% agarose gel with ethidium bromide for 45 minutes at 120v.*

5.3.5. Ligation of *PSORS1C3* PCR amplicons into pCpGL-basic

Five individual ligation reactions were set up for each *PSORS1C3* amplicon, along with a water control. Each reaction contained one digested and cleaned *PSORS1C3* PCR amplicon, digested pCpGL-basic, T4 ligase and buffer (Invitrogen, California, USA), as listed in **Table 5.5**. These were incubated at room temperature for one hour, then left at 14°C overnight in a thermocycler. Following ligation, reactions were stored at 4°C, if transformed on the same day, otherwise they were stored at -20°C. The names of the recombinant plasmids, and the amplicons used are listed in **Table 5.6**.

Table 5.5: Ligation reaction for ligating NcoI and BglII digested PSORS1C3 amplicons and pCpGL-basic.

Component	Volume (µl)
Digested_PSORS1C3 amplicon	7
Digested_pCpGL-basic	1
T4 DNA Ligase (1U/µl)	1
5x T4 DNA Ligase Buffer	2
Total	11

Table 5.6: Names of the recombinant plasmids made through ligating PSORS1C3 PCR amplicons with the pCpGL-basic backbone.

Recombinant Name	PSORS1C3 PCR Amplicon	Plasmid Backbone
PSORS1C3-A-pCpGL-basic	PSORS1C3-A	pCpGL-basic
PSORS1C3-B-pCpGL-basic	PSORS1C3-B	pCpGL-basic
PSORS1C3-C-pCpGL-basic	PSORS1C3-C	pCpGL-basic
PSORS1C3-D-pCpGL-basic	PSORS1C3-D	pCpGL-basic

5.3.6. Transformation of ligated plasmids

Ligated plasmids were transformed into One Shot PIR1 *E. coli* (ThermoFisher Scientific, Massachusetts, USA) to allow for monoclonal amplification of the recombinant plasmids. For the transformation, 50µl of One Shot PIR1 *E. coli* were used for each transformation, including a negative ligation control and a positive transformation control using the plasmid pUC19. The cells were incubated on ice for one hour prior to transformation to allow a slow thaw. Once thawed, 5µl of each ligation reaction were added to individual 50µl vials of One Shot PIR1 cells, 2µl of control plasmid pUC19 was added to the control vial. Vials were mixed by gentle flicking, not by gentle pipetting, as pipetting would damage the fragile cells. Vials were incubated on ice for 30 minutes before being heat shocked for 20 seconds in a 42°C water bath. Following heat-shock vials were placed on ice for two minutes before 250µl of 37°C SOC media was added to each vial. The vials were then incubated at 37°C for one hour, shaking at 225rpm. This incubation allows for the transcription of antibiotic resistance genes before plating onto zeocin containing agar plates. After one hour, 20µl and 200µl of One Shot PIR1 *E.coli* with PSORS1C3-X-pCpGL-basic were spread plated on 50µg/ml zeocin agar plates, and the positive pUC19 *E. coli* were plated on a 100µg/ml ampicillin agar plate. These plates were left inverted overnight at 37°C, a negative control plate containing no spread plated bacteria was also included. An outline of all the transformations are shown in **Table 5.7**.

Table 5.7: Transformations into One-Shot PIR1 E.coli following ligation of recombinant plasmids.

Name	Plasmid	One-Shot PIR1 *E.coli*	Antibiotic Resistance
Ligation A	PSORS1C3-A-pCpGL-basic	YES	Zeocin
Ligation B	PSORS1C3-B-pCpGL-basic	YES	Zeocin
Ligation C	PSORS1C3-C-pCpGL-basic	YES	Zeocin
Ligation D	PSORS1C3-D-pCpGL-basic	YES	Zeocin
Ligation –ve	pCpGL-basic	YES	Zeocin
Transformation +ve	pUC19	YES	Ampicillin
Amp Plate –ve	-	NO	Ampicillin
Zeo Plate –ve	-	NO	Zeocin

5.3.7. Preparation of overnight cultures

The following morning transformed plates were examined for the presence of single colonies as listed in **Table 5.8.** Single colonies from each transformation (minus the controls) were inoculated into 3ml of liquid broth containing 50µg/ml zeocin overnight at 37°C shaking at 225 rpm as described in **2.6.6.** . The number of colonies selected for overnight culturing are listed in **Table 5.9**.

5.3.8. Miniprep of overnight cultures

Overnight cultures which appear cloudy indicate that the *E.coli* have successfully been inoculated and have divided. All 10 of the inoculations showed bacterial growth, with only the negative control remaining clear. 500µl of the inoculant was mixed with 50% glycerol to create glycerol stocks, these were stored at -80°C. Samples were miniprepped using the GeneJET Plasmid Miniprep Kit as outlined in **2.6.7.1.** Samples were eluted in 50µl elution buffer and were then quantified using Nanodrop and plasmids were then stored at -20°C.

5.3.9. Diagnostic digest of miniprepped plasmids

A diagnostic restriction digest, as outlined in **2.6.8.** , was used to confirm that the plasmids isolated have successfully ligated the *PSORS1C3* amplicons into pCpGL-basic. To do this, 100ng of plasmid was digested using BglII and NcoI as listed in **2.6.8.** . The expected digest pattern for the ligation of *PSORS1C3* into pCpGL-basic is two bands, one at ~500bp and another at ~4Kb. Of the 10 miniprepped plasmids, eight contained the insert with A3 and D1 not containing the insert as shown in **Figure 5.5** (these were discarded).

Table 5.8: Colony counts from 20µl and 200µl transformation reactions following overnight incubation at 37°C.

Transformation	20µl plate	200µl plate
Ligation A	1	3
Ligation B	1	17
Ligation C	>100	>100
Ligation D	1	0
Ligation –ve	2	15
Transformation +ve	>100	>100
Amp Plate –ve	0	-
Zeo Plate –ve	0	-

Table 5.9: The number of single colonies selected for inoculation into liquid broth.

Transformation	20µl plate	200µl plate	Total Colonies
Ligation A	1	2	3
Ligation B	1	2	3
Ligation C	3	-	3
Ligation D	1	0	1

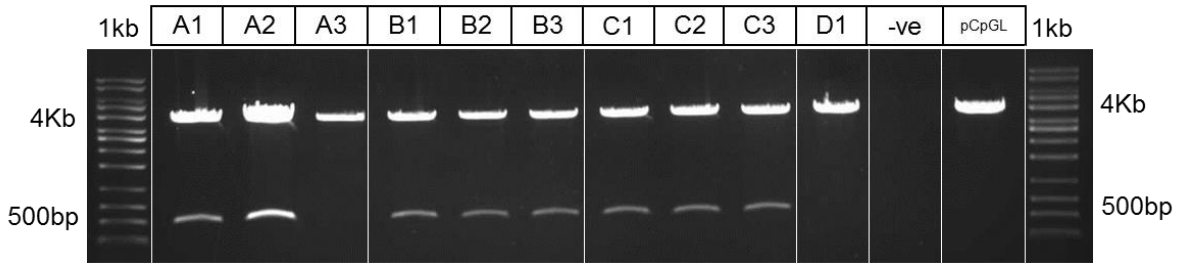


Figure 5.5: Diagnostic restriction digest to confirm successful ligation of PSORS1C3 DMR into pCpGL-basic. *The expected digest pattern of the correct recombinant plasmid is two fragments of 4Kb and 500bp size. Of all the recombinant plasmids tested, A3 and D1 are the only two which do not contain the insert. –ve control is water replacing pCpGL-basic and pCpGL is the empty pCpGL-basic vector as a positive control.*

5.3.10. Sequencing of miniprepped plasmids A1, B1 and C1 identified *PSORS1C3* insert

500ng of plasmids A1, B1, and C1 were Sanger Sequenced using 5uM of LucNRev sequencing primer by Genewiz (Tackley, Essex, United Kingdom). The resultant sequences are shown in **Figure 5.6**. All three sequences contain the correct *PSORS1C3* region amplified through PCR, however constructs PSORS1C3-B1-pCpGL-basic and PSORS1C3-C1-pCpGL-basic contain a C/T SNP (**Figure 5.6**, red bases). These plasmids were therefore not used in downstream experiments. All further experiments used the PSORS1C3-A1-pCpGL-basic construct.

PSORS1C3-A1-pCpGL-basic

ATTCTTGGCATCCTCCATGGGGAAGGGGAGAGGTGCTAAATAATTAAGCCCCC
CTACTACTCAGCACCCGCGTGAGGCATCGTCAGGCATCGTCAGGCCTCCAGT
GGTGGTGGTGGCACCGGGCCTCAACCTCCCCGGAGGGCTGGACTCTCGCTG
CCAGGCTGTGGGGATCAGGCGTTGTGGGGGAGGGGGACACTTAACAGGTAT
GGAGGGCGGAGCAGAGCCCCGCAGTCACTGGCCTGACTTCCGGAACGAAC**C**
GT**CG**CCAGCAAGCACAGCAGTAGGACCAGGGGGATGCAAGAG**CG**GGGGCGG
C**CG**GGGATCGTGCTTCTCGCTCAGGTCCAGATTCC**CG**GCAACCAGGC**CG**GCG
GAATCACGTGCCATGCTCCAGGCCAGCGTAGTCCCGCCCATCTTCCAGCTGA
G**CG**TACCGGGAGGCTCCCATTGGACTGGAGCTGCTACGGAGGCGGGACTTTC
CCTTTTTCTTGAACCCCATTGGGTTAAGTCCAGTCCGAGACAAGCAGATCTGG
ATCCACTAGTCC

PSORS1C3-B1-pCpGL-basic

ATTCTTGGCATCCTCCATGGAGGTTATGGCTCAGGGGAAGGGGAGAGGTGCT
AAATAATTAAGCCCCCCTACTACTCAGCACCCGCGTGAGGCATCGTCAGGCAT
CGTCAGGCCTCCAGTGGTGGTGGTGGCACCGGGCCTCAACCTCCCCGGAGG
GCTGGACTCTCGCTGCCAGGCTGTGGGGATCAGGCGTTGTGGGGGAGGGGG
ACACTTAACAGGTATGGAGGGCGGAGCAGAGCCCCGCAGTCACTGGCCTGAC
TTCCGGAACGAACCGTCGCCAGCAAGCACAGCAGTAGGACCAGGGGGATGCA
AGAGCGGGGGCGGCCGGGGATCGTGCTTCTCGCTCAGGTCCAGATTCCCGG
CAACCAGGCCGGCGGAATCACGTGCCATGCTCCAGGCCAGCGTAGTCCCGCC
CATCTTCCAGCTGAGCGTACCGGGAGGCTCCCATTGGACTGGAGCTGCTATG
GAGGCGGGACTTTCCCTTTTTCTTGAACCCCATTGGGTTAAGTCCAGTCCGAG
ACAAGCAGATCTGGATCCACTAGTCC

PSORS1C3-C1-pCpGL-basic

ATTCTTGGCATCCTCCATGGGGAAGGGGAGAGGTGCTAAATAATTAAGCCCCC
CTACTACTCAGCACCCGCGTGAGGCATCGTCAGGCATCGTCAGGCCTCCAGT
GGTGGTGGTGGCACCGGGCCTCAACCTCCCCGGAGGGCTGGACTCTCGCTG
CCAGGCTGTGGGGATCAGGCGTTGTGGGGGAGGGGGACACTTAACAGGTAT
GGAGGGCGGAGCAGAGCCCCGCAGTCACTGGCCTGACTTCCGGAACGAAC**C**
GT**CG**CCAGCAAGCACAGCAGTAGGACCAGGGGGATGCAAGAG**CG**GGGGCGG
C**CG**GGGATCGTGCTTCTCGCTCAGGTCCAGATTCC**CG**GCAACCAGGC**CG**GCG
GAATCACGTGCCATGCTCCAGGCCAGCGTAGTCCCGCCCATCTTCCAGCTGA
G**CG**TACCGGGAGGCTCCCATTGGACTGGAGCTGCTATGGAGGCGGGACTTTC
CCTTTTTCTTGAACCCCATTGGGTTAAGTCCAGTCCGAGACAAGCAGATCTGG
ATCCACTAGTCC

Figure 5.6: Sanger sequencing results of PSORS1C3-X-pCpGL-basic plasmids. *Green bases are pCpGL-basic sequences, yellow are the PCR primers used for amplifying the PSORS1C3 region, purple are the CpG sites hypomethylated in depression and suicide completers, the red base in B1 and C1 is a SNP introduced through PCR. Construct PSORS1C3-A1-pCpGL-basic was used for further experiments.*

5.3.11. In vitro methylation of the PSORS1C3-pCpGL-basic construct

The *PSORS1C3* DMR was methylated *in vitro* using *M.sssI* methyltransferase (New England Biolabs, Massachusetts, USA) in the presence of S-adenosylmethionine (SAM). This would methylate any CpG within the construct, and in theory, if the sequences of interest possesses promoter activity that is methylation sensitive, repress the expression of luciferase once transfected into HEK293 cells. The empty pCpGL-basic vector was also methylated to act as a control, as the pCpGL vector is CpG free there might be background expression of luciferase once transfected into human cells.

Briefly, 1µg of plasmid was methylated using *M.sssI* with 1600µM SAM as a methyl donor at 37°C for 4 hours, with an additional 1600µM SAM added after 2 hours. An unmethylated control was run under the same conditions with *M.sssI* and SAM replaced with ddH_2O. The outline of the reaction is provided in **Table 5.10**. The reactions were then cleaned to remove buffer salts, *M.sssI* and SAM using the GeneJET PCR Clean up kit, following manufacturer's instructions. The plasmids were eluted in 40µl of elution buffer and quantified using Nanodrop (ThermoFisher Scientific, Massachusetts, USA).

To confirm successful methylation, the plasmids underwent digestion with the methylation sensitive enzyme HpaII, which will only cut and linearize the plasmid if it is unmethylated, thus rendering the plasmid the correct size on an agarose gel following electrophoresis. If the plasmid is methylated then it will run as its supercoiled or nicked state, and will not be the correct size on an agarose gel. 100ng of methylated and unmethylated PSORS1C3-A1-pCpGL-basic and pCpGL-basic were digested with HpaII at 37°C for 1 hour followed with a 65°C inactivation step for 15 minutes, as outlined in **Table 5.11**. The samples were then run on a 0.5% agarose gel for 40 minutes at 120V as outlined in **2.4.1.** . The expected sizes for unmethylated PSORS1C3-A1-pCpGL-basic is 4.1kb, the pCpGL basic does not undergo digestion with HpaII. The resultant gel is shown in **Figure 5.7**.

Table 5.10: The reaction for in vitro methylation of pCpGL constructs.

Component	PSORS1C3		pCpGL-basic		Final Conc
	Methylated	Unmethylated	Methylated	Unmethylated	
Miniprepped Plasmid	1µg	1µg	1µg	1µg	200ng/µl
10x NEB buffer 2	2µl	2µl	2µl	2µl	1x
M.sssI	1µl	-	1µl	-	4 units
1600µM SAM	2µl	-	2µl	-	80µM/µl
ddH_2O	To 20µl	To 20µl	To 20µl	To 20µl	
Total	**20**	**20**	**20**	**20**	

Table 5.11: To detect in vitro methylation, methylated and unmethylated plasmids were digested with the methylation sensitive enzyme HpaII, which only cuts at an unmethylated CCGG site only present in the PSORS1C3-A1-pCpGL-basic construct.

Component	Volume (µl)	Final Concentration
Plasmid	100ng	5ng/µl
HpaII	1	10 units
10x Tango Buffer	2	1x
ddH_2O	To 20	
Total	20	

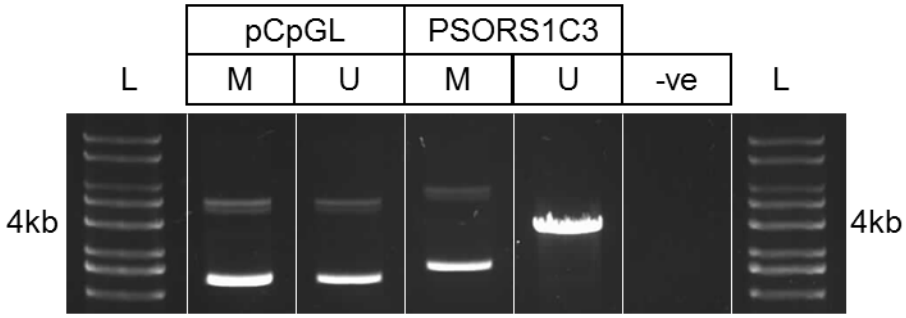


Figure 5.7: Agarose gel electrophoresis image of pCpGL-basic (pCpGL) and PSORS1C3A-A1-pCpGL-basic (PSORS1C3) digested by HpaII following methylation (M) and unmethylated control (U). *The unmethylated PSORS1C3 plasmid has been successfully cut by HpaII rendering it linearized, it therefore runs at the correct size of 4Kb. The pCpGL-basic and M-PSORS1C3 are not cut by HpaII and run in their supercoiled and nicked states. –ve is ddH_2O replacing plasmid.*

5.3.12. Transfection of vectors into HEK293 cells

HEK293 cells were maintained as outlined in **2.1.2.** . The day prior to transfection, 2 x 10^5 cells were seeded in six well plates. The following day media was removed and the cells washed with PBS to remove cell debris, and 1.5ml of fresh growth media was added. 500ng of *PSORS1C3* or pCpGL plasmid and 100ng of pGL4.74[hRluc/TK] reporter control vector (Promega, Wisconsin, USA) were diluted in 500µl of Opti-MEM reduced serum media (Gibco, Massachusetts, USA) in an Eppendorf and left to equilibrate for five minutes at room temperature. 4.5µl of Lipofectamine LTX Reagent (ThermoFisher Scientific, Massachusetts, USA) was then added to each Eppendorf and incubated for 30 minutes at room temperature. Following incubation, 500µl of LTX plasmid mix was added, in a dropwise circular motion, to the HEK293 cells in six well plates. Cells were then incubated at 37°C 5% CO_2 for 24 hours to allow expression of firefly and *Renilla* luciferases.

5.3.13. Dual luciferase assay

24 hours following transfection a dual-luciferase reporter assay (Promega, Wisconsin, USA) was carried out to measure the expression of firefly luciferase and *Renilla* luciferase in the transfected cells. Cell lysates were prepared using 100µl of the supplied passive lysis buffer (PLB) at 1x concentration. Transfected cells had media removed and were washed with 1x PBS to remove debris. 100µl of 1x PLB were added to each well and cells detached using a plastic cell scraper. Following scraping, lysates were transferred into ice cold Eppendorf tubes and 20µl of lysate was then transferred into a white walled 96 well plate for quantification of luciferase expression. Each experiment contained three technical repeats and there are three experimental repeats.

Quantifying luciferase expression requires the addition of two reagents from the dual-luciferase reporter assay kit. Firstly, 80µl of LARII is added to each well containing sample, LARII contains beetle luciferin which is converted to oxyluciferin in the presence of Firefly luciferase (expressed from the pCpGL backbone) which releases light. The absorbance of light at 525nm is measured for 10 seconds to quantify the expression of Firefly Luciferase. After 10 seconds 80µl of Stop & Glo buffer is added to each reaction. The Stop & Glo reaction is twofold, the first reaction quenches the conversion of beetle luciferin to oxyluciferin preventing the detection of Firefly

luciferase activity. Secondly, the reagent contains coelenterazine which is converted to coelenteramide in the presence of *Renilla* luciferase (expressed by the pGL4.74[hRluc/TK] reporter control vector). The absorbance is then measured for another 10 seconds to record the activity of the *Renilla* luciferase before the reaction is terminated. An outline of the chemistry is shown in **Figure 5.8**. The injections and light absorbance were carried out automatically using a Pherastar Plate Reader. **Figure 5.9** shows an example luciferase activity report following the dual-luciferase assay.

5.3.14. Data analysis

The average firefly luciferase activity was calculated by averaging 525nm absorbance readings between 2 and 10 seconds. The average *Renilla* luciferase activity was calculated by averaging absorbance readings between 14 and 22 seconds. Absorbance readings between 0-1.5 and 10-13.5 seconds were excluded as the absorbance reading is skewed by the pipetting of the LARII and Stop & Glo into the plate.

Data analysis was carried out as described in (Jacobs and Dinman, 2004). Briefly, firefly luciferase activity was normalised by dividing expression by its corresponding *Renilla* luciferase activity, this gives the firefly luciferase activity relative to *Renilla* luciferase. The median normalised firefly luciferase activity was calculated across all three technical repeats and the upper and lower quartiles calculated. The interquartile range (IQR) is then calculated as the difference between the upper and lower quartiles. To exclude any outliers the IQR is then multiplied by 1.5 and added to the median to calculate the upper limit or subtracted to calculate the lower limit. Any normalised firefly luciferase activity greater than the upper limit or lower than the lower limit were excluded from downstream analysis. Four outliers were removed, the averages were recalculated and the variance, standard deviation, and standard error were calculated for each experiment (n=3). All data is presented as a normalised firefly luciferase activity relative to *Renilla* luciferase. Fold change expression was calculated by dividing the unmethylated normalised firefly luciferase activity by the respective methylated normalised luciferase activity, to give fold change versus methylated.

A paired T-test was then used to compare the pCpGL or *PSORS1C3* methylated Firefly Luciferase Activity to the unmethylated pCpGL or *PSORS1C3*. A one-way

ANOVA was used to compare between all four groups and a pairwise t-test post hoc analysis was undertaken to compare between groups.

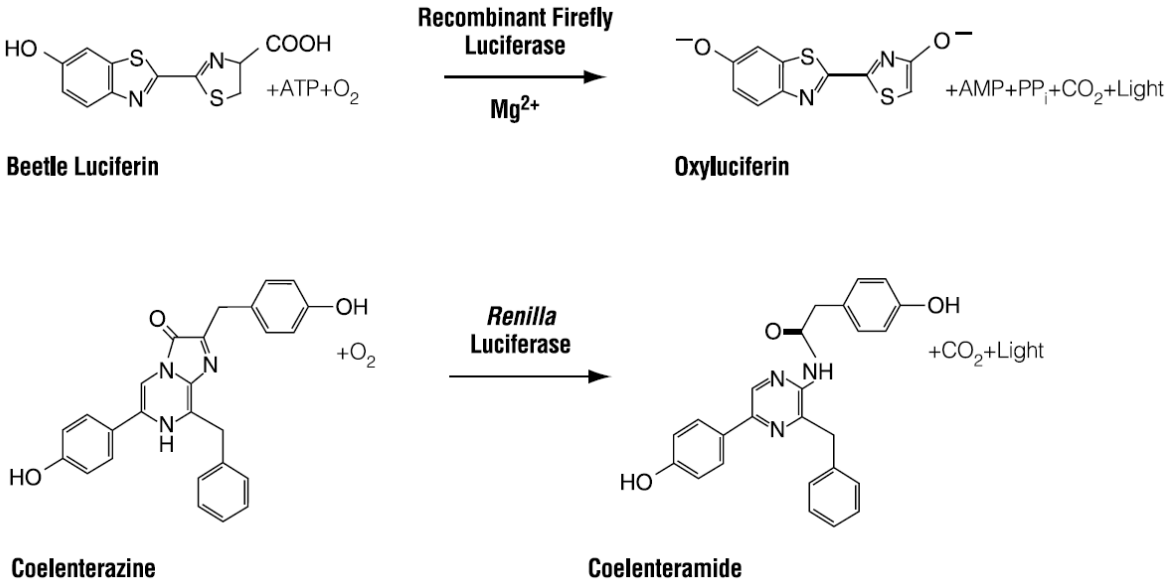


Figure 5.8: The chemistry behind the Dual-Luciferase Assay*. LARII contains Beetle Luciferin which is converted into oxyluciferin in the presence of Firefly Luciferase, releasing light as a by-product. The pCpGL-basic backbone vector contains the Firefly Luciferase gene and will be expressed in the presence of an unmethylated promoter. The Stop & Glo buffer contains Coelenterazine which is converted to Coelenteramide in the presence of Renilla Luciferase, releasing light as a by-product. The Renilla Luciferase is expressed by the pGL4.74[hRluc/TK] reporter control vector.*

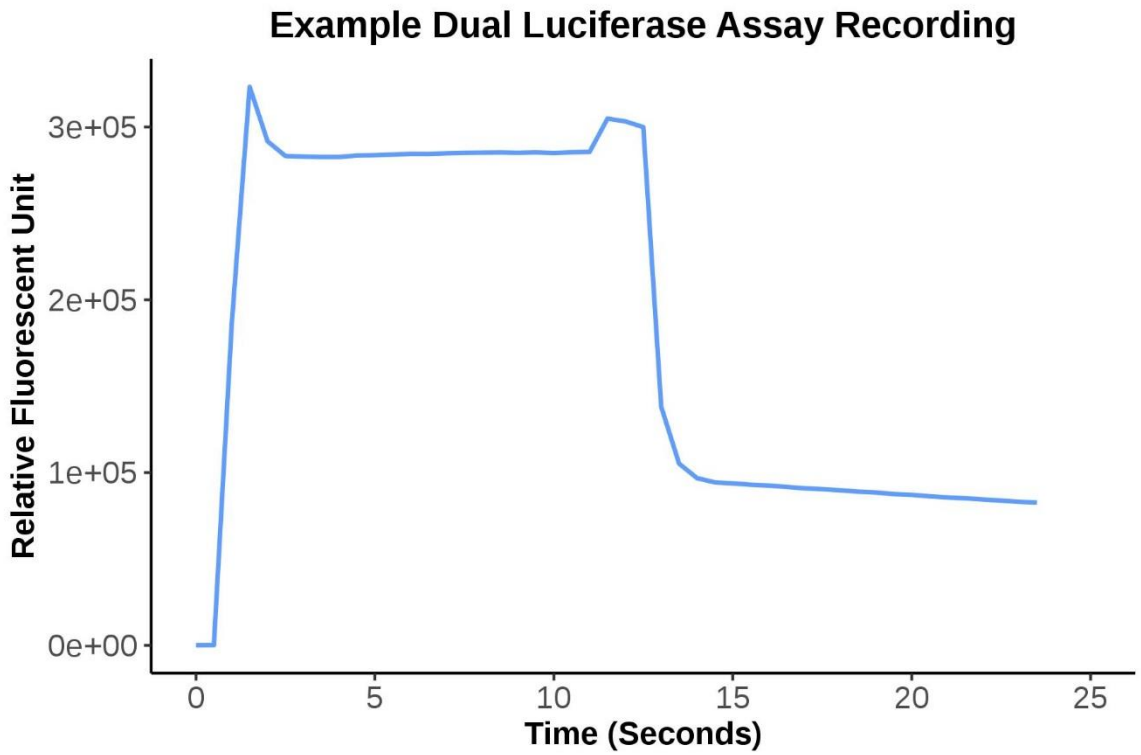


Figure 5.9: An example dual-luciferase assay recording. *X axis is the time of the reaction in seconds and the Y axis is the relative fluorescence unit (RFU) at 525nm. Absorbance readings at 525nm are used to detect luciferase activity and are recorded every 0.5 seconds. At 0 seconds LARII is injected into the cell lysates, the absorbance measured between 0 and 12 seconds represents the Firefly luciferase activity. At 12 seconds Stop & Glo is injected to each sample, this quenches Firefly activity and activates Renilla luciferase. Absorbance readings between 12 and 24 seconds represents Renilla luciferase activity. RFU between 2 and 10 were averaged for Firefly activity and between 14 and 22 for Renilla.*

5.4. Results

5.4.1. The methylation state of the pCpGL-basic vector has no effect on luciferase activity

First I set out to examine whether or not the pCpGL-basic vector has any background effect on luciferase activity. Previous studies have identified no difference in luciferase activity within the methylated and unmethylated pCpGL-basic vector.

To do this, pCpGL was methylated in the presence of SAM and *M.sssI*, then both methylated and unmethylated pCpGL vectors were transfected into HEK293 cells along with a *Renilla* luciferase transfection control. Luciferase activity was examined 24 hours post transfection. In three experimental repeats I see no change in luciferase activity between methylated pCpGL and unmethylated pCpGL (**Figure 5.10**). Relative expression ratio of Firefly to *Renilla* increases from 0.0016 in methylated pCpGL to 0.0019 in unmethylated pCpGL however this is not significant (n=3 experimental repeats, two-tailed paired t-test, $t(2)=0.87$, p-value=0.53). While there is a 1.3 fold change increase in luciferase activity in the unmethylated pCpGL versus methylated pCpGL this is not significant (n=3 experimental repeats, two-tailed paired t-test, $t(2)=0.76$, p-value=0.48). This indicates any change in luciferase assay is dependent on the methylation state of the genomic region cloned into the vector and not through methylation of the vector backbone supporting evidence provided by the vector creator (Klug and Rehli, 2006).

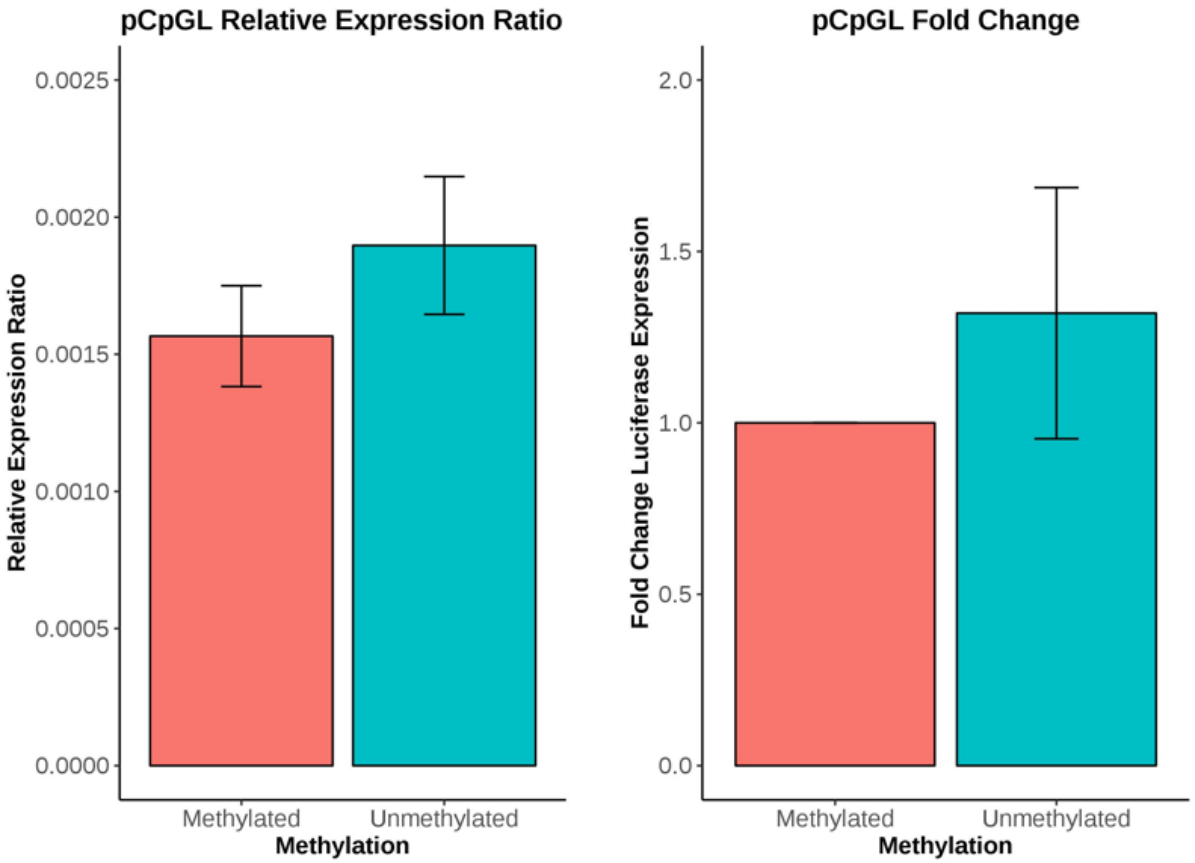


Figure 5.10: There is no significant difference in luciferase activity between methylated and unmethylated pCpGL. *There is no significant difference in relative expression of Firefly luciferase activity when normalised to activity of the Renilla luciferase control (two-tailed paired t-test, t(2)=0.87, p-value=0.53). There is no significant difference in Fold change induction of Firefly luciferase in the unmethylated pCpGL plasmid versus methylated pCpGL per experiment (two-tailed paired t-test, t(2)=0.76, p-value=0.48). n=3 experimental repeats, with three technical repeats per experiment. Error bars represent +/- standard error.*

5.4.2. Methylation of the *PSORS1C3* DMR decreases firefly luciferase activity

I set out to identify if the DMR from Murphy *et al* implicated in depression and suicide is an active promoter or enhancer. To do this the identified DMR was PCR cloned into the pCpGL vector using the BglII and NcoI restriction sites. This construct was then methylated *in vitro* using SAM and MsssI. Unmethylated and methylated *PSORS1C3* vectors were transfected into HEK293 cells along with a *Renilla* luciferase transfection control. 24 hours post transfection cells were harvested and luciferase activity quantified by a dual luciferase assay.

There is a marked increase in the relative expression of firefly luciferase activity in the unmethylated *PSORS1C3* cloned pCpGL vector versus methylated (**Figure 5.11**). The average relative expression ratio of Firefly to *Renilla* in the methylated vector was 0.0269, this increased to 3.81 with the unmethylated vector (n=3 experimental repeats, two-tailed paired t-test, t(2)=13.14, p-value=0.0057). The increase in activity is more pronounced when converted to a fold change, where there is a 206 fold change increase of firefly luciferase activity in the unmethylated *PSORS1C3* plasmid versus the methylated *PSORS1C3* plasmid.

This has identified that the DMR from Murphy et al, when differentially methylated *in vitro*, can act as a promoter for luciferase activity. Translating that in *vivo* could indicate it is a regulatory region for gene expression and genes involved in suicide and depression.

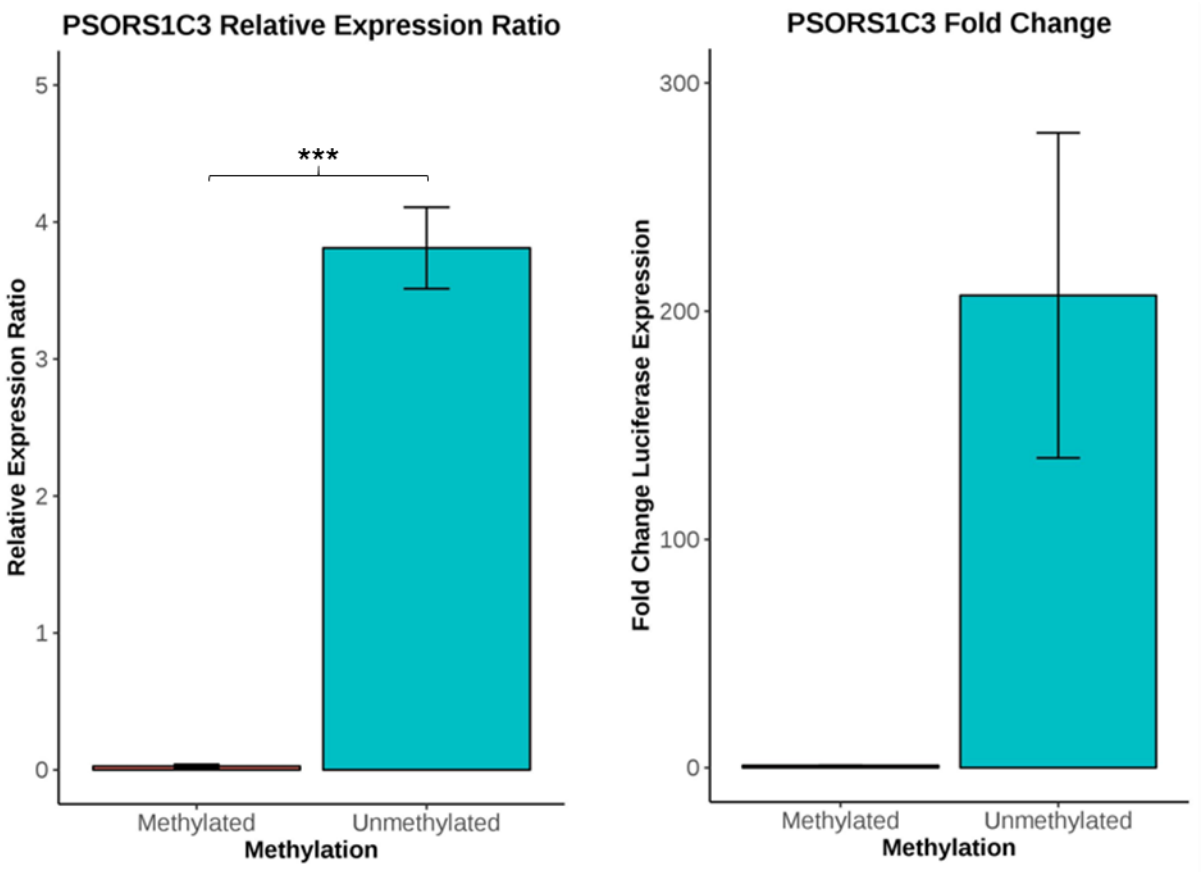


Figure 5.11: Methylation of the PSORS1C3 cloned pCpGL vector reduces luciferase activity to near negligible. *There is a significant increase in firefly luciferase expression when normalised to Renilla luciferase, the normalised expression ratios of firefly activity increases to 3.81 when unmethylated verses 0.0269 when methylated in vitro (two-tailed paired t-test, t(2)=13.14, p-value=0.0057). There is a 206 fold change increase in luciferase activity in the unmethylated PSORS1C3 cloned vector versus the methylated. n=3 experimental repeats with each experiment containing three technical repeats. Error bars represent +/- the standard error.*

5.4.3. Methylation of the PSORS1C3 DMR reduces luciferase activity to that seen in the pCpGL basic vector

Promoters are known to be “leaky” and result in low basal expression despite being methylated (Huang *et al.*, 2015). I set out to identify if methylation of the *PSORS1C3* DMR completely reduces luciferase expression to that of the empty pCpGL basic vector or if leakage is observed.

The relative expression ratio results in a significant difference between the four groups (pCpGL M/U and *PSORS1C3* M/U) (One-way ANOVA (F(3,8)=162.94, p-value=1.64E-07)). Pairwise t-test comparisons between groups identified that there was no significant difference between each group, except between the *PSORS1C3* unmethylated plasmid and all other samples (p-value=1.64e-07, Relative Expression Ratio 3.81). The Methylated *PSORS1C3* (Relative Expression Ratio 0.027) was not significantly different when compared to both the methylated and unmethylated pCpGL vector (Relative Expression Ratio 0.0016 and 0.0019 respectively) (Student’s t-test, p-value=1 and p-value=1 respectively). Indicating there is no promoter leakage, in the *PSORS1C3* DMR. Relative expression ratios are shown in **Figure 5.12** and **Table 5.12**.

Fold changes were calculated using the methylated pCpGL plasmid as baseline. The fold change in luciferase expression supports the relative expression ratios where there was a significant difference between the four groups (One-way ANOVA (F(2,8)=42.05, p-value=3.04E-05)). Pairwise t-test comparisons between groups identified that there was no significant difference between the groups, except between the *PSORS1C3* unmethylated plasmid (p-value=3.3E-05, +2421 Fold Change Expression). The methylated *PSORS1C3* (+16 Fold Change Expression) was not significantly different when compared to both the methylated (+1 Fold Change Expression) and unmethylated pCpGL plasmid (+1.3 Fold Change Expression) (pairwise t-test, p-value=1 and p-value=1 respectively). This further supports there is no promoter leakage in the *PSORS1C3* DMR. Fold change in expression values are shown in **Figure 5.13** and **Table 5.13**.

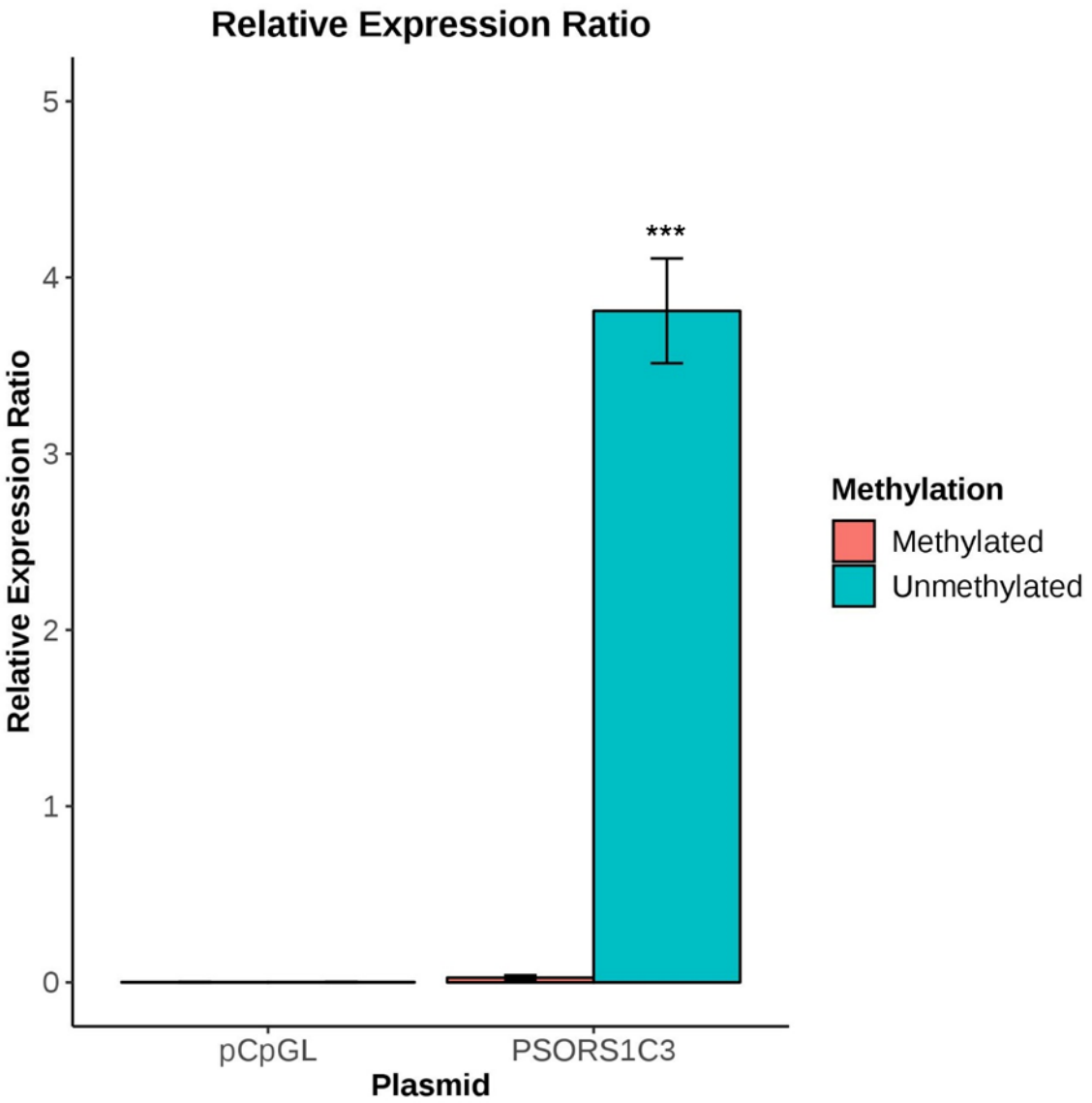


Figure 5.12: Methylation of the PSORS1C3 DMR reduces the relative expression ratio of luciferase to that seen in the empty pCpGL plasmid. *One-way ANOVA (F(3,8)=162.94, p-value=1.64E-07). Pairwise T-test comparisons between groups identified that there was a significant difference between the PSORS1C3 unmethylated plasmid and all other samples (p-value=1.64e-07). n=3 experimental repeats with each experiment containing three technical repeats. Error bars represent +/- the standard error.*

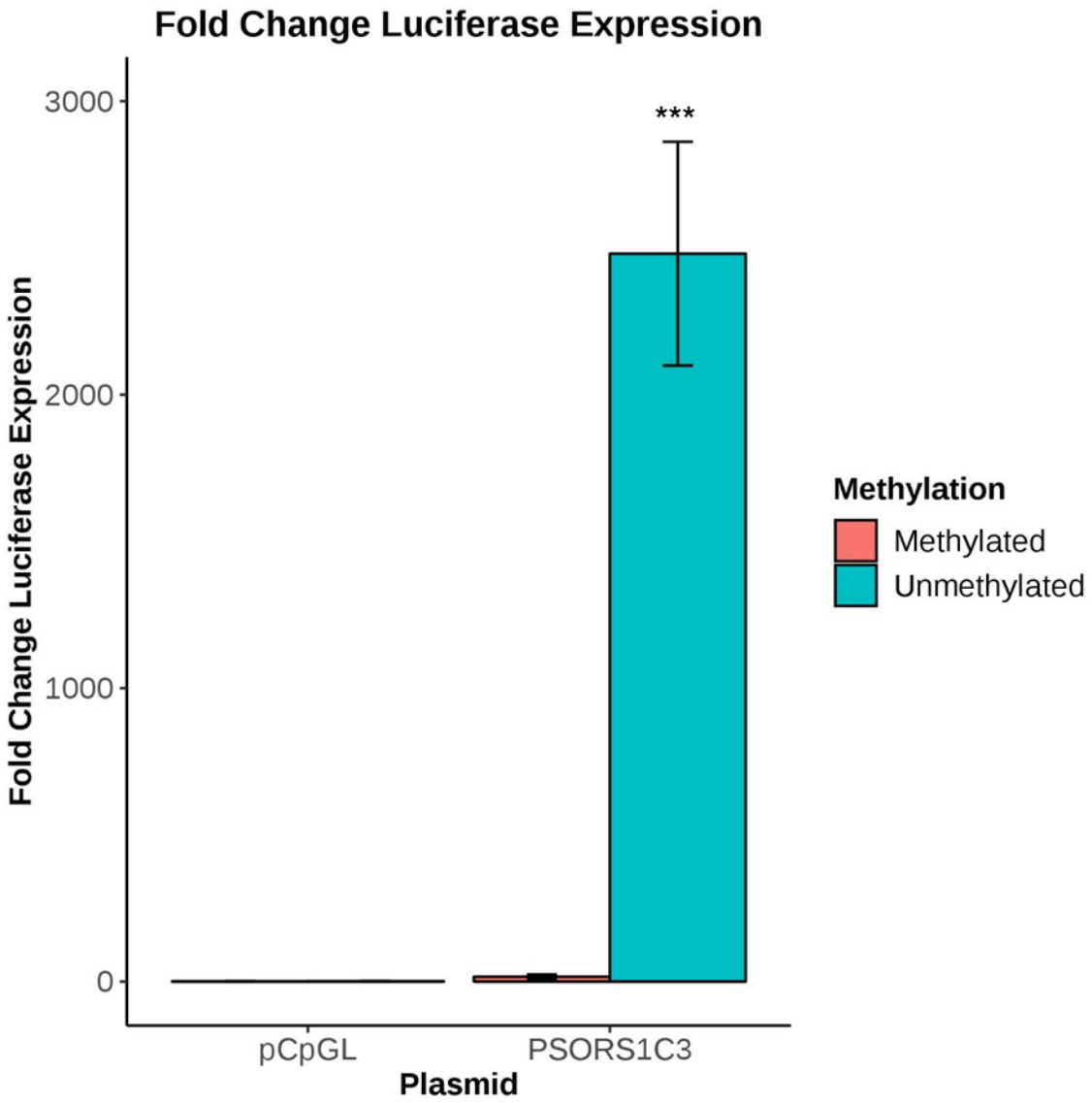


Figure 5.13: Methylation of the PSORS1C3 DMR reduces Fold Change Luciferase Expression to that seen in the empty pCpGL plasmid. *Fold changes were calculated using the methylated pCpGL plasmid as baseline. There was a significant difference between the four groups (One-way ANOVA (F(2,8)=42.05, p-value=3.04E-05)). Pairwise T-test comparisons between groups identified a significant difference between the PSORS1C3 unmethylated plasmid (+2421 Fold Change Expression, p-value=3.3E-05). n=3 experimental repeats with each experiment containing three technical repeats. Error bars represent +/- the standard error.*

Table 5.12: Relative expression ratio of firefly luciferase to renilla luciferase.

Plasmid	Methylation	n	Relative Expression Ratio	SE
pCpGL	Methylated	3	0.0016	0.00018
pCpGL	Unmethylated	3	0.0019	0.00025
PSORS1C3	Methylated	3	0.027	0.014
PSORS1C3	Unmethylated	3	3.81	0.30

Table 5.13: Fold change in luciferase expression compared to methylated pCpGL.

Plasmid	Methylation	n	Fold Change Luciferase Expression	SE
pCpGL	Methylated	3	1	0
pCpGL	Unmethylated	3	1.32	0.36
PSORS1C3	Methylated	3	16.89	7.73
PSORS1C3	Unmethylated	3	2421.15	425.07

5.5. Discussion

In this chapter I set out to examine if a luciferase assay could be used for the functional validation of identified DMRs as active promoters or enhancers. To do this I cloned a 500bp region containing 13 CpG sites upstream of the *PSORS1C3* gene identified in Murphy *et al* into the commercially available CpG free vector pCpGL-basic (Klug and Rehli, 2006). These 13 probes are hypomethylated in suicide and major depression disorder, and it is unknown if they are within a promoter or enhancer region or have any functional link to gene expression.

Following a dual luciferase assay, I can confirm that the *PSORS1C3* region identified in Murphy *et al* is methylation sensitive promoter region. When cloned into pCpGL basic, the relative expression of luciferase increased 2400 fold when compared to the empty pCpGL basic control *(*pairwise t-test comparison, p-value=3.3E-05). When this plasmid is methylated only a 16 fold increase in luciferase activity is observed, while this is still higher than the unmethylated pCpGL plasmid it does not reach significance. This is also common as promoters are usually "leaky", where you cannot fully suppress the gene expression only reduce it (Huang *et al.*, 2015). I also identified that the changes in luciferase expression are caused by the *PSORS1C3* region and not the pCpGL-basic backbone vector. There is no difference in the luciferase activity reported as either relative expression ratio (two-tailed paired t-test, t(2)=0.87, p-value=0.53) or fold change (two-tailed paired t-test, t(2)=0.76, p-value=0.48) between the methylated and unmethylated pCpGL. This matches with previously reported data using this plasmid backbone (Klug and Rehli, 2006; Andres *et al.*, 2014) This is crucial, as if change was observed, this would have to be controlled into all further analysis.

While I have identified that this *PSORS1C3* region is methylation sensitive we cannot identify whether this is within an active promoter or enhancer region. Further analysis would have to be carried out where the region is cloned upstream of a known promoter within pCpGL Basic and then luciferase activity examined following methylation (Andres *et al.*, 2014). Further, I cannot identify which gene the DMR is controlling through this assay, it would be interesting to look at gene expression in cis/trans of the region in suicide completers. Furthermore, new methods have been developed to look at methylation *in vitro* as (discussed later in Chapter 6.), where we can change the methylation state of these CpGs *in vitro* and examine gene expression in cis/trans as a result. This would provide a clear insight into which gene(s) these CpGs control

expression of. While only 13 probes within these region have been identified in Murphy et al, due to the limitations of the array used to identify the DMR, there might be a wider window than the 500bp used in this study which is differentially methylated within suicide completers. There could be numerous CpG sites which I have not included which could further increase the activity of the luciferase if these were included. I have fully methylated the promoter using *M.SssI*, whereas Murphy et al only identified a moderate decrease in methylation, it is unknown whether these small decreases in methylation translate into gene expression changes. Because we did not bisulphite convert and pyrosequence the cloned plasmid we are unable to draw conclusions as to if there are a subset of the 13 probes which are driving expression of the luciferase or if there is an additive effect. While the luciferase assay is a good tool to have in the functional epigenetics toolbox, it still cannot answer fundamental questions however this tool will be of use to determine if DMR regions are within active promoters or enhancers. The crucial limitation is that this assay provides no clue to which gene that expression is being driven, if any.

5.6. Conclusions

This study has identified that the hypomethylated DMR identified in Murphy *et al* has promoter activity which is sensitive to DNA methylation, this provides a functional link to identified DMRs and their role in regulating activity of gene expression. The simplistic approach of cloning of the DMR identified into a luciferase reporter is a method which can be adopted into our EWAS pipelines to functionally validate if the findings are within an area of active transcriptional control. However, while this study has identified the *PSORS1C3* DMR is within a methylation sensitive promoter region, we cannot draw conclusions as to which gene expression changes are being driven. Further work should be carried out to identify the genes which this DMR is controlling expression of and how these link to suicide, depression, and schizophrenia.

Chapter 6. Functional epigenetic editing using CRISPR/Cas9

6.1. Introduction

Characterising DNA methylation changes is complex, while environmental factors are often thought to exert a substantial influence on the methylome the underlying genetic architecture actually are just as important (Hannon, Dempster, *et al.*, 2016). DNA methylation has shown to be highly heritable and numerous studies have identified SNP/DNA methylation correlations or mQTLs (Bell *et al.*, 2011; Gutierrez-Arcelus *et al.*, 2015). One hypothesis in how SNPs contribute to disease is through altering DNA methylation, which in turn alters gene expression. This theory is supported by a number of GWAS of complex diseases (Wijsman *et al.*, 2011; Schizophrenia Working Group of the Psychiatric Genomics Consortium, 2014; The Autism Spectrum Disorders Working Group of The Psychiatric Genomics Consortium, 2017; Huckins *et al.*, 2019) that have identified very few variants associated with disease in coding regions, and the majority of associated variants occur in intronic regions such as enhancers and regions of open chromatin (Maurano *et al.*, 2012; Schaub *et al.*, 2012). Therefore it is important to combine GWAS, EWAS, and the more recent TWAS studies to understand how genetic risk can first influence DNA methylation and in turn gene expression and establish a causal link to disease. There are a number of theories for how genetic risk can influence gene expression and methylation and these are shown in **Figure 6.1**. There could be a single SNP which acts on both gene expression as an eQTL or on DNA methylation as an mQTL and these two associations are independent of each other (**Figure 6.1A**). Another theory is that the SNP changes DNA methylation as an mQTL which then alters gene expression, so the SNP changes gene expression through DNA methylation as an intermediate step (**Figure 6.1B**). This can also be argued to be the reverse, where the SNP alters gene expression as an eQTL where gene expression then changes DNA methylation (**Figure 6.1C**). A final theory is that rather than a single SNP controlling both DNA methylation and gene expression there are in fact two SNPs within linkage disequilibrium which are acting on DNA methylation and expression independently (**Figure 6.1D**). It is therefore very difficult to dissect the direction of effect and the causal relationship between genetics, epigenetics, and gene expression. In order to examine if DNA methylation results in gene expression changes a recent analysis was carried out in our lab by (Hannon *et al.*, 2018) using summary-based mendellian randomisation (SMR) (Zhu *et al.*, 2016),

to examine for CpG sites which are associated with specific gene expression changes. From this statistical approach you can infer the direction of causality, ie whether or not DNA methylation is mediating gene expression (**Figure 6.1B**). However these statistical approaches can result in many false positives due to the number of statistical tests carried out and multiple corrections undertaken. Therefore these statistical approaches need to be validated by changing the DNA methylation at these identified CpG sites *in vitro* and confirming gene expression changes in the associated genes.

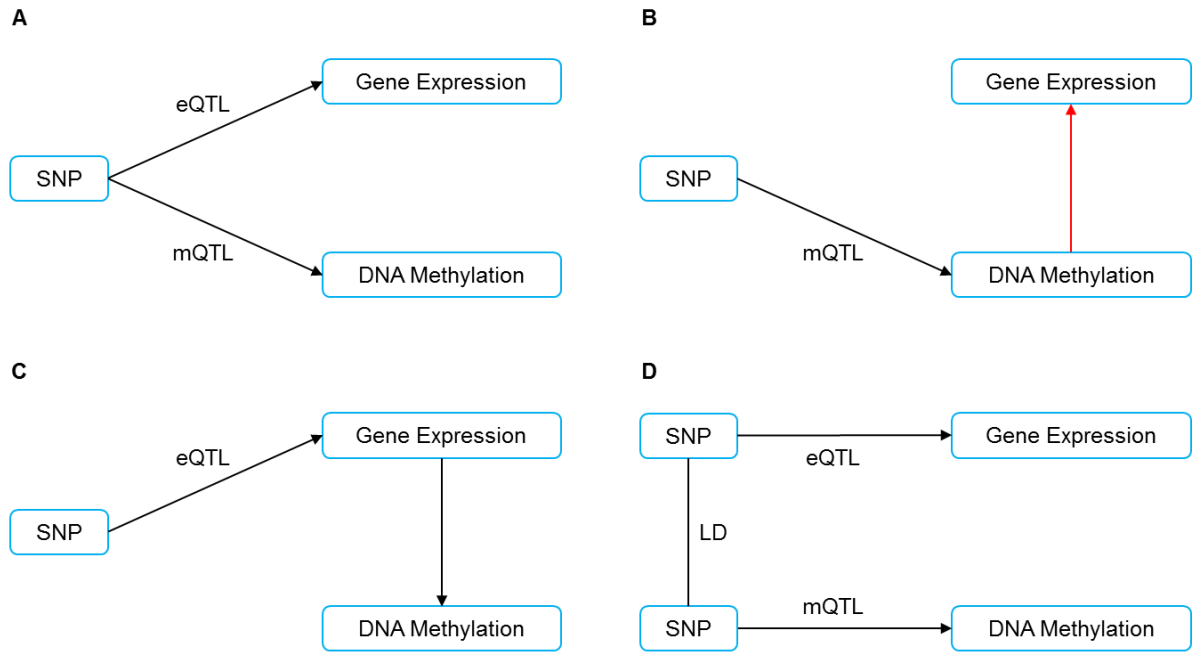


Figure 6.1: The pleiotropic association between SNPs, gene expression, and DNA methylation. *From statistical approaches we can determine whether a disease associated SNP is associated with gene expression (as an eQTL) or DNA methylation (as an mQTL). In scenario (A) the SNP action on DNA methylation and gene expression are acting independent of each other. The hypothesis suggested in (B) is that the SNP could be changing gene expression by altering DNA methylation as an intermediate step. Another hypothesis is that the SNP could change gene expression which results in DNA methylation changes (C). Or in scenario (D) there are two SNPs within linkage disequilibrium (LD) acting independently with one driving gene expression another driving DNA methylation. This chapter aims to develop methods to determine if changing DNA methylation at specific SNP associated CpG sites results in gene expression changes in SNP associated genes identified through eQTL analysis (red arrow).*

Current functional validation of DNA methylation is limited. As shown in Chapter 5. we can confirm if a DMR is an active promoter or enhancer using a luciferase assay based approach, with *in vitro* methylation reducing luciferase activity (Klug and Rehli, 2006). However, linking methylation at individual CpG sites to gene function and expression is difficult due to a lack of appropriate molecular tools. Recently new methods have been developed based on CRISPR/Cas9 that activate gene expression (Aach *et al.*, 2013), repress gene expression (Qi *et al.*, 2013), edit single bases (Komor *et al.*, 2016), and edit the epigenome (Liu *et al.*, 2016; Tadi *et al.*, 2016; Vojta *et al.*, 2016; Kwon *et al.*, 2017). These methodologies utilise catalytically dead Cas9 (dCas9), where the Cas9 contains mutations which prevent cleavage of dsDNA but still possesses the ability to form complexes with gRNA and target the genome (Doudna and Charpentier, 2014). The dCas9 is then fused to a variety of proteins which can be targeted to specific regions using a gRNA to result in modifications as mentioned above. In this chapter I set out to develop a pipeline to functionally validate if CpG sites identified from my labs previous SMR analysis (Hannon *et al.*, 2018) are altering gene expression at predicted genes. To do this I used a modified form of CRISPR/Cas9 first described in (Liu *et al.*, 2016) which has been designed to edit the epigenome by allowing methylation and demethylation at specific sites using dCas9 fused to DNMT3a and dCas9 fused to Tet1CD (Liu *et al.*, 2016, 2018). The differences between genomic CRISPR/Cas9 systems and DNA methylation CRISPR/Cas9 systems are shown in **Error! Reference source not found.** and have been previously discussed in **1.6.1.3.**.


This image has been removed by the author of this thesis for copyright reasons.

The (Hannon *et al.*, 2018) SMR study identified a multitude of CpG sites that were putatively associated with gene expression in known GWAS regions. To do this, mQTLs were identified using blood samples from the 1,193 individuals from the Understanding Society study. This data was then compared to eQTLs derived from microarray studies on blood samples (Westra *et al.*, 2013). (Hannon *et al.*, 2018) were able to establish CpG sites associated with gene expression by using SMR with the underlying genetics as a proxy for this association. This identified over 5,000 CpG sites which were associated with gene expression changes. For functional analysis and validation I first reduced the dataset of associated CpGs using specific criteria. The first criteria was to select a DMR, so only sites with more than one significant CpG within a 1kb window were selected. Next, expression of the nearest gene to the DMR and the genes identified though eQTL analysis were examined using the Human Protein Atlas (www.proteinatlas.org) to confirm adequate gene expression in SH-SY5Y cells (Schwenk, Asplund and Edqvist, 2011). A literature search was then carried out on these genes to identify whether or not the genes have been associated with any disease or known function. Finally the baseline methylation of the CpGs within the DMR was examined in SH-SY5Y cells using already available data generated in Chapter 7. , this step was to identify in which direction I could alter the DNA methylation.

From this analysis we identified two DMRs to characterise using epigenetic editing. First a DNA methylation probe within the transcription start site of the *FANCG* gene on chromosome 9, where methylation of cg05293216 has been associated with expression of *PIGO* and *STOML2* through SMR analysis (**Figure 6.3**). *PIGO* is a gene involved in the synthesis of a cell membrane anchor and disruption results in mental retardation (Krawitz *et al.*, 2012), epilepsy (Nakamura *et al.*, 2014) and developmental disorders (Pagnamenta *et al.*, 2017). *STOML2* is a mitochondrial gene required for the creation of the respiratory chain and ATP generation (Mitsopoulos *et al.*, 2015). Reduction of *STOML2* expression has been identified in iPSC derived fibroblasts from patients with sporadic Alzheimer's disease (Martín-Maestro *et al.*, 2017). Furthermore, *STOML2* is a risk gene for asthma where increased expression increases risk of asthma (Revez *et al.*, 2016). Around this CpG site there are four other CpG dinucleotides which are thought to be associated with the same causal variant (mQTL) identified through Bayesian co-localisation analysis (Hannon *et al.*, 2018). These CpG

sites are within a window of 16bp, all of which are hyper-methylated in SH-SY5Y cells, with an average DNA methylation of 73.2% across the four. There is also another CpG dinucleotide which is not present on the methylation arrays or used in SMR analysis within the DMR. Therefore to examine if this DMR is controlling the expression of *PIGO* and *STOML2,* as predicted by SMR, I de-methylated these five CpG sites with Tet1CD fused to dCas9. I then examined changes in DNA methylation and gene expression changes in *PIGO, STOML2* and *FANCG*.

The second DMR selected was located in the 5' untranslated region of the *SNCA* gene (**Figure 6.4**), which encodes for α-synuclein, and is a well-known and characterised Parkinson's risk gene (Siddiqui, Pervaiz and Abbasi, 2016). This DMR consists of three CpG dinucleotides within 182bp which are hypomethylated in SH-SH5Y cells, with an average DNA methylation of 10.94% all of which are associated with *SNCA* expression through SMR analysis. Furthermore there are another 12 CpG dinucleotides within this region not included on the methylation array for SMR. Therefore to examine if this DMR is controlling *SNCA* expression I decided to methylate this 128bp region using DNMT3a fused to dCas9 and characterise DNA methylation changes and *SNCA* expression. The genomic context of both DMR sites targeted in this chapter is shown in **Figure 6.5**.

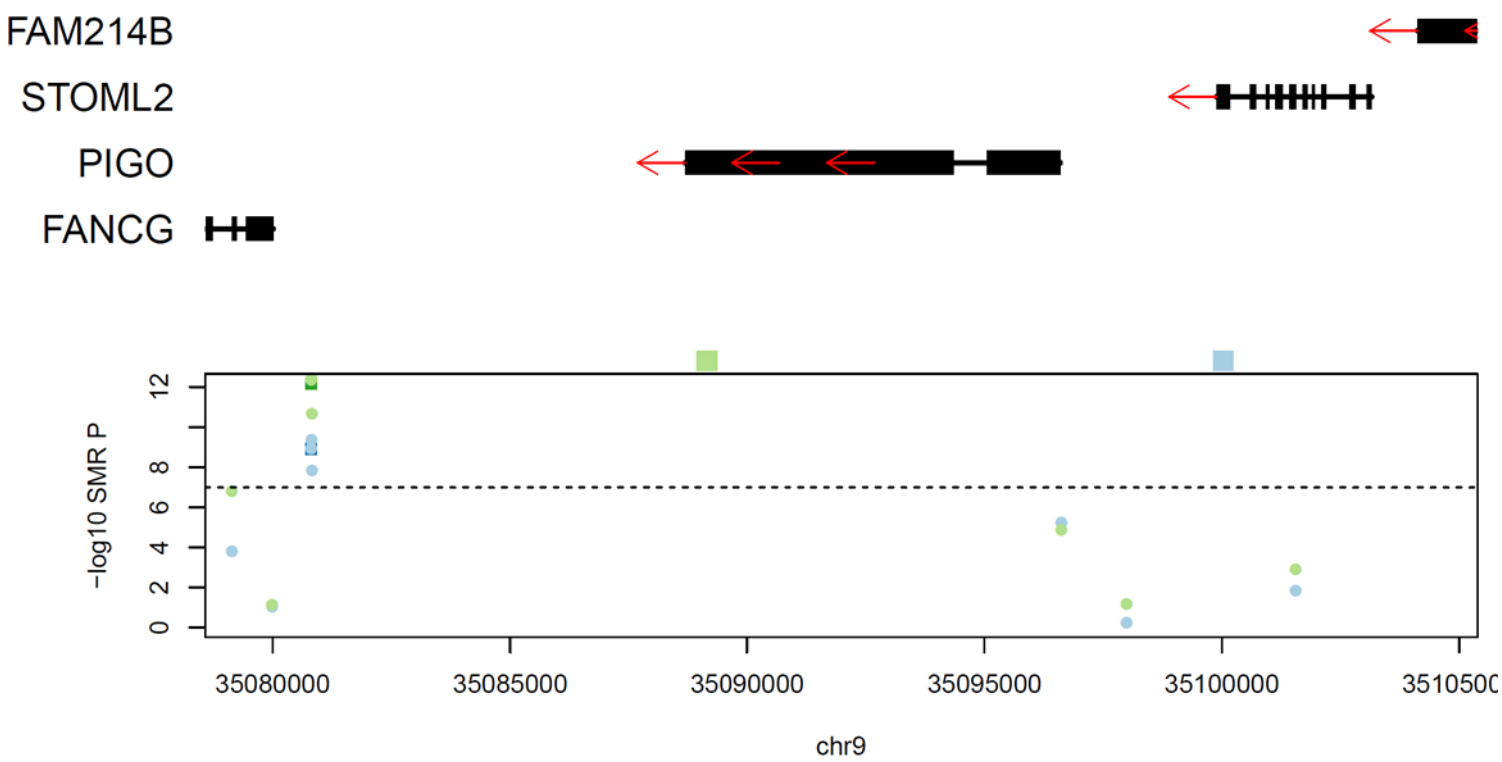


Figure 6.2: SMR Plot from (Hannon et al 18) for the FANCG DMR identifying DNA methylation sites associated with gene expression. *Squares indicate location of gene expression probes in PIGO and STOML2. The SMR Manhattan plot below shows DNA methylation sites which were tested for gene expression against the probes. Colour indicates which gene expression probe was tested against. Dashed line indicates the significance threshold. Squares within the SMR Manhattan are the CpGs targeted by dCas9-Tet1CD.*

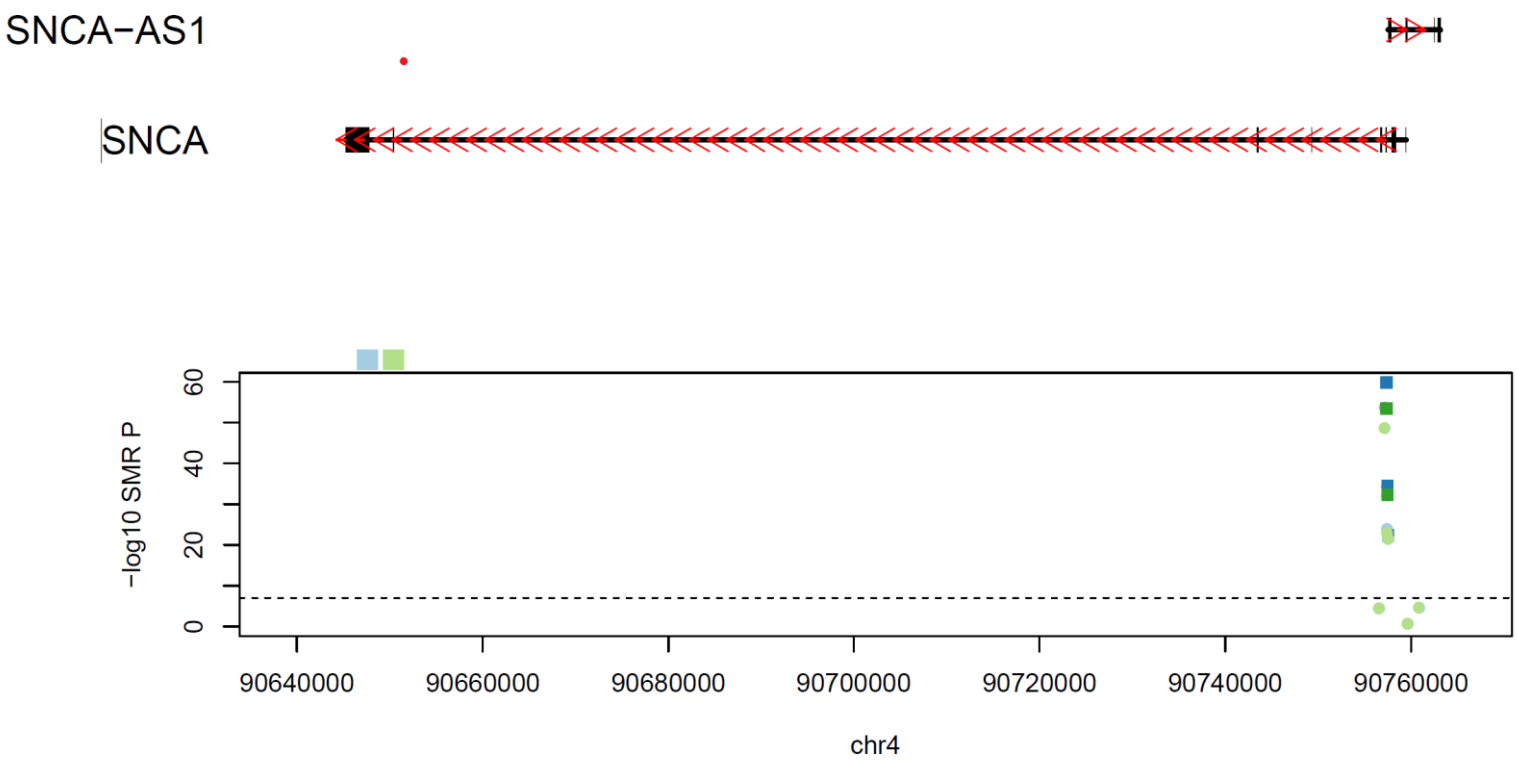


Figure 6.3: SMR plot from (Hannon et al 18) for the SNCA DMR identifying DNA methylation sites associated with gene expression. *Squares indicate location of gene expression probes in SNCA. The SMR Manhattan plot below shows DNA methylation sites which were tested for gene expression against the probes. Colour indicates which gene expression probe was tested against. Dashed line indicates the significance threshold. Squares within the SMR Manhattan are the CpGs targeted by dCas9-DNMT3a.*

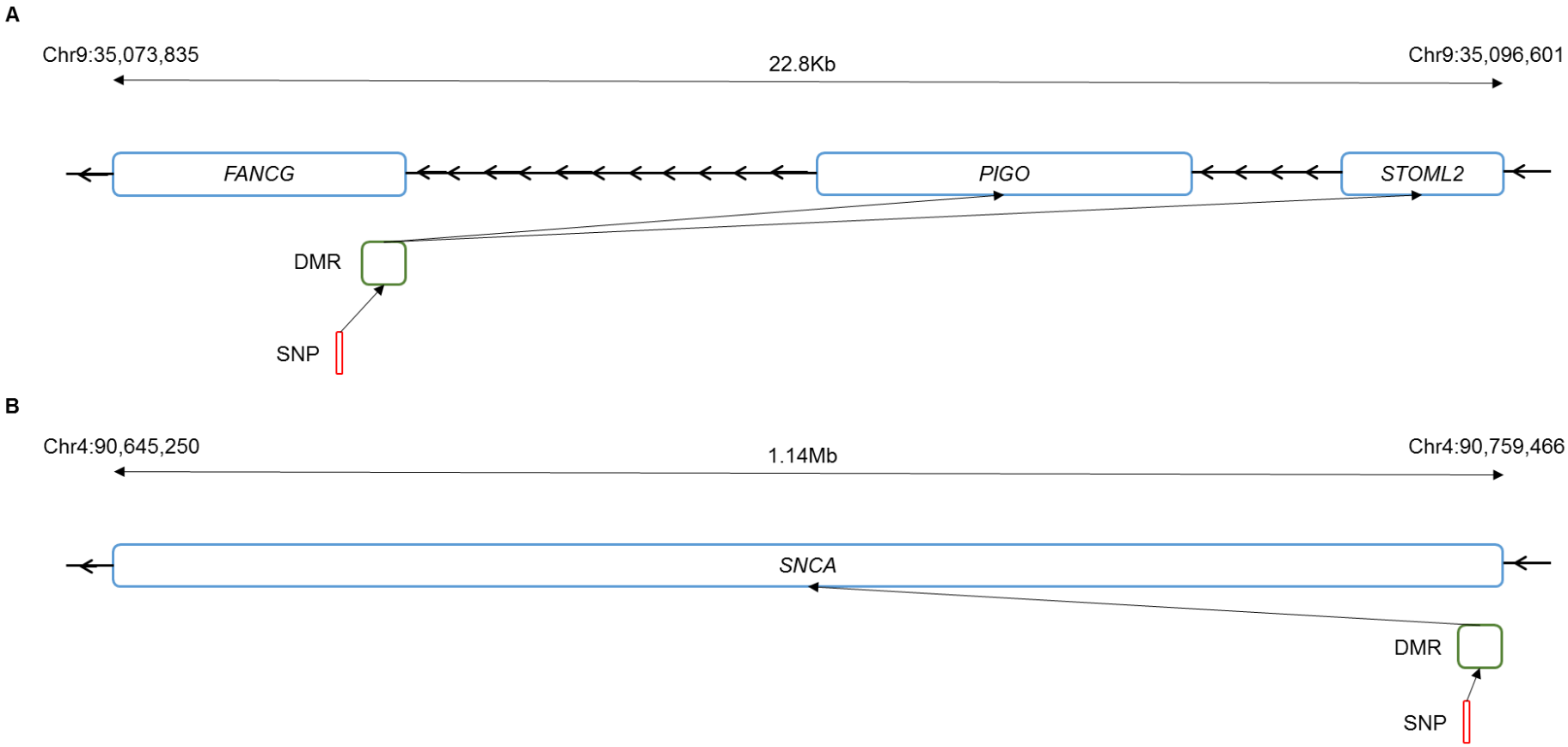


Figure 6.4: The genomic context of the two DMRs identified by SMR for functional validation using epigenetic editing*. The first DMR is located in the promoter of FANCG and contains four CpG sites associated with gene expression of PIGO and STOML2 through SMR analysis. These sites are hypermethylated in SH-SY5Y cells so were demethylated using Tet1CD (A). The second DMR is located in the 5' UTR of SNCA where three CpG sites are associated with the gene expression of SNCA through SMR. This region is hypomethylated in SH-SY5Y cells so were methylated using DNMT3a (B).*

6.2. Aims

The overall aim of this chapter was to develop a method to functionally characterise the relationship between DNA methylation and gene expression at disease associated loci identified through SMR (Hannon *et al.*, 2018). To do this I used a SMR dataset which used eQTL, mQTL, and GWAS data to identify regions of DNA methylation that are affecting gene expression at specific target genes. In order to validate this experimentally I had several aims:

1. To generate DNMT3a and Tet1CD lentiviral constructs which contained the selection marker blue fluorescent protein (BFP) for selection of positively integrated cells.
2. To generate a pipeline for editing methylation using a modified CRISPR/Cas9.
3. To target the hypermethylated *FANCG* DMR identified by (Hannon *et al.*, 2018), using epigenetic CRISPR to reduce DNA methylation using dCas9 fused to Tet1CD and examine the resultant effect on the gene expression of *FANCG, STOML2* and *PIGO*.
4. To target the hypomethylated *SNCA* DMR identified by (Hannon *et al.*, 2018) using epigenetic CRISPR to increase DNA methylation using dCas9 fused to Tet1CD and investigate gene expression of *SNCA*.

6.3. Methods

6.3.1. Generation of control constructs

Seven commercially available lentiviral constructs for editing the epigenome were purchased from the plasmid depository service Addgene (www.addgene.org) and are listed in **Table 6.1**. These plasmids contain the machinery for production of lentiviral particles containing DNMT3a, Tet1CD, both active and inactive, and the targeting gRNA. The DNMT3a and Tet1CD inactive mutation (IM) plasmids did not contain a fluorescent selection marker of BFP, therefore the first step was to clone the inactive forms of DNMT3a and Tet1CD into their respective active plasmids, replacing the active enzyme. These controls were selected as to mimic the action of the active forms of DNMT3a and Tet1CD with infection, integration, and dCas9 binding, as it is unknown if any of these steps could confound the results by altering DNA methylation. The Tet1CD_IM and DNMT3a_IM regions of the Fuw-dCas9-Tet1CD_IM and Fuw-dCas9-dnmt3a_IM plasmids were amplified with PCR containing the restriction sites BamHI and EcoRI. The Fuw-dCas9-DNMT3a-P2A-tagBFP and Fuw-dCas9-Tet1C-P2A-BFP were then digested with BamHI and EcoRI to remove the functional DNMT3a/Tet1CD. The PCR amplified DNMT3a_IM and Tet1CD_IM amplicons were then ligated into the backbones for fusion to BFP. The method is shown in **Figure 6.6**.

Table 6.1: Lentiviral plasmids used in this chapter

Addgene ID	Plasmid Name	Function	Depositing Lab
84569	Fuw-dCas9-DNMT3a-P2A-tagBFP	Lentiviral construct to expressed dead Cas9 fused with active Dnmt3a catalytic domain tagged to BFP	Jaenisch Lab (Liu *et al.*, 2016)
84478	Fuw-dCas9-dnmt3a_IM	Lentiviral construct to expressed dead Cas9 fused with inactive Dnmt3a catalytic domain with mutations E664A and E756A	Jaenisch Lab (Liu *et al.*, 2016)
NA	Fuw-dCas9-DNMT3a_IM-P2A-tagBFP	Lentiviral construct to expressed dead Cas9 fused with inactive Dnmt3a catalytic domain with mutations E664A and E756A tagged to BFP via P2A	Generated in house
108245	Fuw-dCas9-Tet1CD-P2A-BFP	Lentiviral construct to express dead Cas9 fused with active Tet1CD catalytic domain tagged to BFP	Jaenisch Lab (Liu *et al.*, 2018)
84479	Fuw-dCas9-Tet1CD_IM	Lentiviral construct to express dead Cas9 fused with inactive Tet1CD catalytic domain with mutations H1672Y and D1674A	Jaenisch Lab (Liu *et al.*, 2016)
NA	Fuw-dCas9-Tet1CD_IM-P2A-BFP	Lentiviral construct to express dead Cas9 fused with inactive Tet1CD catalytic domain with mutations H1672Y and D1674A tagged to BFP via P2A	Generated in house
84477	pgRNA-modified	Lentiviral construct to express gRNA with scaffold and mCherry and puromycin selection	Jaenisch Lab (Liu *et al.*, 2016)
12259	pMD2.G	Lentiviral construct to expressed the VSV-G gene required for lentiviral envelope	Trono Lab
12260	psPAX2	Lentiviral construct required for expression of packaging genes pol, gag and rev	Trono Lab

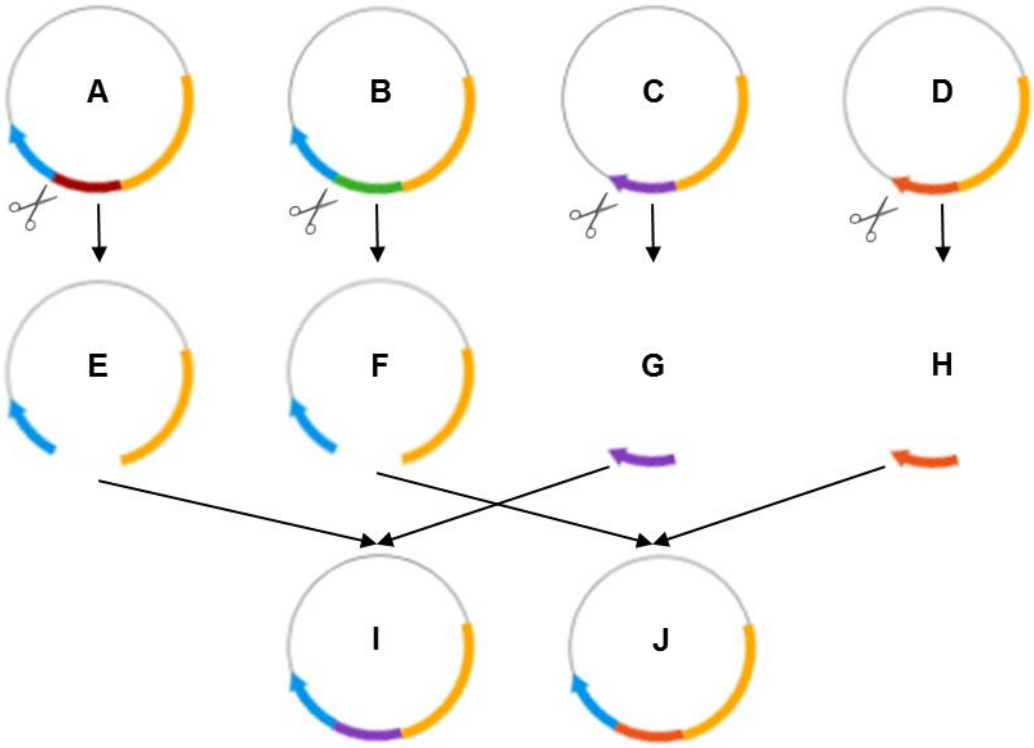


Figure 6.5: The cloning strategy to create the BFP fused control vectors. *The active DNMT3a (red) and Tet1CD (green) enzymes are removed from the Fuw-dCas9-DNMT3a-P2A-tagBFP (A) and Fuw-dCas9-Tet1CD-P2A-BFP (B) plasmids using BamHI and EcoRI. The inactive DNMT3a (purple) and Tet1CD (orange) are PCR amplified with primers containing the BamHI and EcoRI restriction sites from the Fuw-dCas9-dnmt3a_IM (C) and Fuw-dCas9-tet1CD_IM (D) plasmids. The digested inactive DNMT3a (G) and Tet1CD (H) are ligated into the digested backbones (E and F) containing dCas9 (yellow) and BFP (blue). This results in the final constructs of Fuw-dCas9-DNMT3a_IM-tagBFP (I) and Fuw-dCas9-Tet1CD_IM-tagBFP (J).*

6.3.1.1. Generation of DNMT3a and Tet1CD inactive insert by PCR

Primers were designed to amplify around the DNMT3a and Tet1CD inactive inserts of the Fuw-dCas9-dnmt3a_IM and Fuw-dCas9-tet1CD_IM constructs available from (Liu *et al.*, 2016). The primers were 5' tagged with the restriction sites for BamHI (GGATCC) on the forward primer and EcoRI (GAATTC) on the reverse primer, so they can be cloned into the Fuw-dCas9-EMPTY-tagBFP backbone vector. Six extra bases were added at the 5' end to aid in restriction enzyme binding and cleavage **Table 6.2**. The resultant sequences, including the 5' tags are shown in **Table 6.3**. The combinations of primers to generate the amplicons for cloning are shown in **Table 6.4** including the expected amplicon size.

100ng of plasmid DNA (Fuw-dCas9-dnmt3a_IM or Fuw-dCas9-tet1CD_IM) was used as a template for amplifying the dnmt3a_IM or tet1CD_IM region in a high fidelity PCR reaction, where each primer set had a water negative control. The reactions were set up as outlined in **2.6.1.** . The PCR reaction was run for 35 cycles at temperatures outlined in **Table 6.5**. Following amplification each PCR product was visualised on a 1% agarose gel as outlined in **2.4.1.** , resultant gel shown in **Figure 6.7**.

Following successful PCR amplification a gel extraction was performed using the GeneJET Gel Extraction Kit (Thermofisher Scientific, Massachusetts, USA) as described in **2.4.3.** to isolate only the correct amplicon sizes. Samples were eluted in 50µl of elution buffer. Final PCR amplicon concentrations were calculated using Nanodrop as described in **2.3.4.1.** . 500ng of amplicon were then digested in separate 20µl double digest reactions as outlined in methods using BamHI and EcoRI. Reactions were incubated at 37°C for 1 hour, followed by inactivation at 65°C for 15 minutes. Reactions were then cleaned using the GeneJET PCR clean up kit (Thermofisher Scientific, Massachusetts, USA) as outlined in **2.4.2.** , and eluted in 10µl.

Table 6.2: Primer sequences and locations for cloning inactive DNMT3a and Tet1CD into the Fuw-dCas9-EMPTY-tagBFP vector. *Including addition tags to be added at the 5' end to aid with cloning.*

Primer Name	Sequence (5'-3')	Plasmid Location	Strand	Added restriction sites	Added Random bases
DNMT3a_IM_F_BamHI	AACCACGACCAGGAATTTGAC	10028-10048	+	GGATCC	CATGCA
Tet1CD_IM_F_BamHI	CTGCCCACCTGCAGCTGTCTT	4966-4986	+	GGATCC	GTACGT
DNMT3a_IM_R_EcoRI	CACACACGCAAAATACTCC	10912-10930	-	GAATTC	CGTACT
Tet1CD_IM_R_EcoRI	GACCCAATGGTTATAGGGCC	7100-7119	-	GAATTC	CGTACT

Table 6.3: Final primer sequences for cloning inactive DNMT3a and Tet1CD. *Yellow highlights random six bases to aid with restriction enzyme binding and green shows the restriction sites.*

Primer Name	Final Primer Sequence (5'-3')
DNMT3a_IM_F_BamHI	CATGCAGGATCCAACCACGACCAGGAATTTGAC
Tet1CD_IM_F_BamHI	GTACGTGGATCCCTGCCCACCTGCAGCTGTCTT
DNMT3a_IM_R_EcoRI	CGTACTGAATTCCACACACGCAAAATACTCC
Tet1CD_IM_R_EcoRI	CGTACTGAATTCGACCCAATGGTTATAGGGCC

Table 6.4: Combination of primers used to generate the constructs used in this chapter.

Primer Set	Construct Name	Forward Primer	Reverse Primer	Amplicon Size (bp)
A	BamHI-DNMT3a_IM-EcoRI	DNMT3a_IM_F_BamHI	DNMT3a_IM_R_EcoRI	2166
B	BamHI-Tet1CD_IM-EcoRI	Tet1CD_IM_F_BamHI	Tet1CD_IM_R_EcoRI	915

Table 6.5: Cycling conditions for DNMT3a_IM and Tet1CD_IM high-fidelity PCR with Phusion polymerase.

Step	Temperature (°C)	Time	Number of Cycles
Hotstart	98	2 min	1
Denaturation	98	30 sec	
Annealing	62	30 sec	35
Extension	72	2 min	
Final Extension	72	10 min	1
Storage	4	∞	1

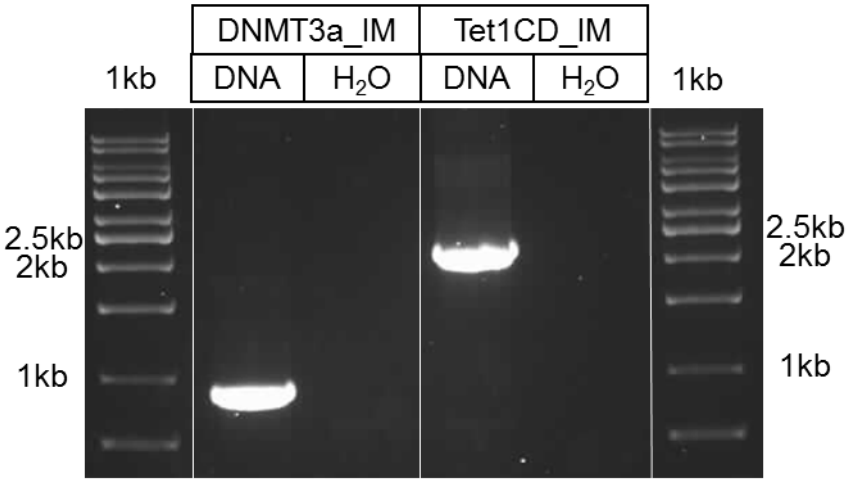


Figure 6.6: Amplification of the DNMT3a_IM and Tet1CD_IM regions using Phusion polymerase. *Lanes 1 and 7, 1kb ladder, lane 2 Fuw-dCas9-dnmt3a_IM plasmid template DNA, lane 3 negative control. Lane 4 fuw-dCas9-tet1CD_IM template DNA, lane 5 negative control.* 1% Agarose gel with ethidium bromide, 45 minutes in 1x TBE.

6.3.1.2. Digestion of Fuw-dCas9-DNMT3a-P2A-tagBFP and Fuw-dCas9-Tet1CD-P2A-BFP

The backbone vectors Fuw-dCas9-DNMT3a-P2A-tagBFP and Fuw-dCas9-Tet1CD-P2A-BFP were digested with BamHI and EcoRI to remove the functional dnmt3a and tet1CD regions to be replaced with the inactive inserts generated in **6.3.1.1.** . To do this, 500ng of vectors were digested in a 20µl double digest reaction with EcoRI and BamHI as described in **2.6.2.** . Following digestion each reaction was run on a 0.5% agarose gel containing ethidium bromide for 1 hour at 100v to separate out the digested and undigested plasmids. The bands of interest were then extracted using the GeneJET Gel Extraction Kit (Thermofisher Scientific, Massachusetts, USA) as described in **2.4.3.** , with the additional dilution step as the vectors are greater than 10Kb. Digested backbone vectors were eluted in 10µl of elution buffer.

6.3.1.3. Ligation of DNMT3a_IM and Tet1CD_IM into Fuw-dCas9-EMPTY-P2A-tagBFP and Fuw-dCas9-EMPTY-P2A-BFP

Ligations were carried out as described in **2.6.3.** using T4 ligase with a molar concentration of 1:3 backbone to insert, to maximise the cloning efficiency. Two ligation reactions were set up including both backbone and the amplicon (A and C), two reactions were set up without the amplicon to check background undigested plasmid (B and D) and one water control was set up as listed in **Table 6.6**. Reactions were ligated at 23° for one hour then left overnight at 16°C in a thermocycler. Following ligation, reactions were stored at 4°C until transformation. The names of the recombinant plasmids generated are listed in **Table 6.7**.

Table 6.6: Ligation reactions for ligating DNMT3a_IM and Tet1CD_IM amplicons into Fuw-dCas9-EMPTY-P2A-tagBFP. *A contains BamHI-DNMT3a_IM-EcoRI insert with Fuw-dCas9-EMPTY-P2A-tagBFP backbone, B is Fuw-dCas9-EMPTY-P2A-tagBFP backbone only, C is BamHI-Tet1CD_IM-EcoRI insert with Fuw-dCas9-EMPTY-P2A-BFP backbone, and D is Fuw-dCas9-EMPTY-P2A-BFP backbone only.*

Component	A	B	C	D	Water
Digested IM insert	1µl (2ng)	-	1µl (4.47ng)	-	-
Digested Fuw-dCas9-EMPTY-P2A-BFP	1.87µl (10ng)	1.87µl (10ng)	1.09µl (10ng)	1.09µl (10ng)	-
T4 DNA Ligase (1U/µl)	0.5µl	0.5µl	0.5µl	0.5µl	0.5µl
5x T4 Ligase Buffer	2µl	2µl	2µl	2µl	2µl
ddH_2O	4.63µl	5.63µl	5.41µl	6.41µl	7.5µl
Total	**10µl**	**10µl**	**10µl**	**10µl**	**10µl**

Table 6.7: Names of the recombinant plasmids generated.

Recombinant Name	Amplicon	Backbone
Fuw-dCas9-DNMT3a_IM-P2A-tagBFP	BamHI-DNMT3a_IM-EcoRI	Fuw-dCas9-EMPTY-P2A-tagBFP
Fuw-dCas9-Tet1CD_IM-P2A-BFP	BamHI-Tet1CD_IM-EcoRI	Fuw-dCas9-EMPTY-P2A-BFP

6.3.1.4. Transformation of ligated plasmids

Transformation of the ligation reactions from **6.3.1.3.** into DH5α *E. coli* was carried out as described in **2.6.4.**

6.3.1.5. Overnight bacterial cultures

The following afternoon, six colonies from **6.3.1.4.** were selected at random from ligation reaction A and C taken into overnight cultures as described in **2.6.6.** .

6.3.1.6. Purification of plasmid DNA

The following morning, plasmid DNA was purified using the GeneJET Plasmid Miniprep Kit (Thermofisher Scientific, Massachusetts, USA) as described in **2.6.7.1.** and eluted in 30µl of elution buffer.

6.3.1.7. Diagnostic digest

A diagnostic digest was then performed on 100ng of each miniprepped plasmids as described in **2.6.8.** to confirm if the DNMT3a_IM or Tet1CD_IM has been correctly inserted into the backbone using the restriction enzyme XhoI. The expected digest patterns for Fuw-dCas9-DNMT3a_IM-P2A-tagBFP are 8.5Kb, 4.7Kb and 1.8Kb and for Fuw-dCas9-Tet1CD_IM-P2A-BFP are 8.5Kb, 5Kb and 2.8Kb. The results of the digest are shown in **Figure 6.8** with only sample Fuw-dCas9-Tet1CD_IM-P2A-BFP (4) showing the incorrect digest pattern.

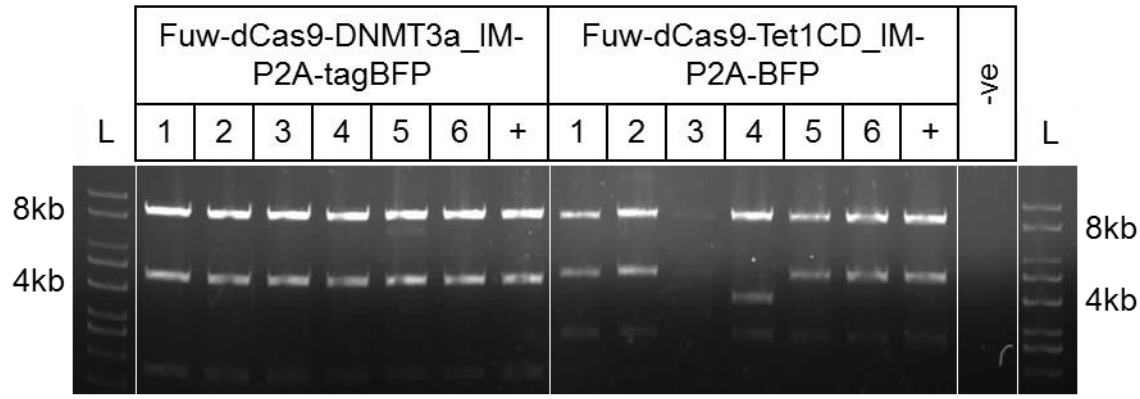


Figure 6.7: XhoI digest of miniprepped plasmids. *+ are the positive controls Fuw-dCas9-dnmt3a_IM and Fuw-dCas9-tet1CD_IM plasmids. -ve is negative control with plasmid replaced with water. L 1Kb ladder.*

6.3.1.8. Sanger sequencing of cloned plasmids

Sanger sequencing was carried out as described in **2.6.9.** . Positive plasmids from the XhoI digest in **6.3.1.7.** underwent Sanger sequencing using the sequencing primers listed in **Table 6.8** by Genewiz. Two sequencing runs were carried out, one to confirm BFP cloning and removal of the stop codon, and another to confirm that the inactive mutations were still present following PCR amplification. Sequence data was analysed and confirmed insertion of the inactive constructs fused to BFP.

Table 6.8: Sanger sequencing primers used for lentiviral constructs.

Sequencing Primer Name	Sequence (5’ – 3’)	Sequence targeted
Lenti_dnmt3a_sequencing	GGCAAAGACCAGCATTTTCC	Stop codon removal
Lenti_tet1CD_sequencing	GAAAATGAAGGCCTCAGAGC	Stop codon removal
DNMT3a_IM_F_BamHI	CATGCAGGATCCAACCACGACCAGGAATTTGAC	IM mutations
Tet1CD_IM_F_BamHI	GTACGTGGATCCCTGCCCACCTGCAGCTGTCTT	IM mutations
mU6 Forward	CAGCACAAAAGGAAACTCACC	gRNA cloning site

6.3.2. Generation of gRNA constructs

In order to target the DNMT3a or Tet1CD enzyme to the target CpGs a lentiviral gRNA construct is required. For this we use the pgRNA-modified plasmid first described in (Liu *et al.*, 2016). This plasmid is designed for use with lentivirus and contains the AarI cloning site for insertion of gRNA, and the positive selection markers of puromycin and mCherry.

6.3.2.1. gRNA design

To maximise the efficiency of epigenetic editing the gRNA to *FANCG* and *SNCA* was carried out using the online software provided by Benchling (www.benchling.com). Benchling's gRNA design software uses two algorithms to predict on target and off target scores for each gRNA (Ran *et al.*, 2013; Hsu, Lander and Zhang, 2014; Doench *et al.*, 2016). gRNA were selected with high on target and high off target scores, on target scores less than 45 and off target scores less than 60 were not considered. Selected gRNA are listed in **Table 6.9** and the locations of the *FANCG* and *SNCA* gRNA relative to CpG sites are shown in **Figure 6.9** and **Figure 6.10**. For cloning the complementary sequence of the gRNA is also included to allow cloning via the AarI restriction site into pgRNA-modified. A TTGG sticky end was added at the 5' end of the forward gRNA and an AAAC sticky end was added to the reverse gRNA as outlined in **Table 6.10**. gRNA were then ordered from Integrated DNA Technologies (www.idtdna.com).

Table 6.9: gRNA for epigenetic editing of CpG within FANCG and SNCA.

gRNA	Sequence (5' – 3')	Chromosomal Location	Strand
FANCG_gRNA_1	CATTAAGAAGTCATACGCTGGGG	Chr9:35080798-35080818	+
FANCG_gRNA_2	ACCTCGTGATCCGCCCGCCTCGG	Chr9:35080854-35080874	-
FANCG_gRNA_3	TTTCACCGTGTTAGCCATGATGG	Chr9:35080890-35080910	-
SNCA_gRNA_1	GCTGAGAACGCCCCCTCGGGTGG	Chr4:90757421-90757440	-
SNCA_gRNA_2	AGACGGCCCGCGAGTGTGAGCGG	Chr4:90757383-90757402	-
SNCA_gRNA_3	AAGGAGGCCAAGTCAACAGGTGG	Chr4:90757487-90757506	-

Table 6.10: Final gRNA sequences including reverse complements and added sticky ends for cloning using AarI into pgRNA-modified.

gRNA	Final Sequence
FANCG_gRNA_1	TTGGCATTAAGAAGTCATACGCTG
FANCG_gRNA_1_Rev	AAACCAGCGTATGACTTCTTAATG
FANCG_gRNA_2	TTGGACCTCGTGATCCGCCCGCCT
FANCG_gRNA_2_Rev	AAACAGGCGGGAGGATCACGAGGT
FANCG_gRNA_3	TTGGTTTCACCGTGTTAGCCATGA
FANCG_gRNA_3_Rev	AAACTCATGGCTAACACGGTGAAA
SNCA_gRNA_1	TTGGGCTGAGAACGCCCCCTCGGG
SNCA_gRNA_1_Rev	AAACCCCGAGGGGGCGTTCTCAGC
SNCA_gRNA_2	TTGGAGACGGCCCGCGAGTGTGAG
SNCA_gRNA_2_Rev	AAACCTCACACTCGCGGGCCGTCT
SNCA_gRNA_3	TTGGAAGGAGGCCAAGTCAACAGG
SNCA_gRNA_3_Rev	AAACCCTGTTGACTTGGCCTCCTT

ATTAAAATTAAAAACTCTTGTACTACAAAGACATCATTAAGAAGTCATA**CG**CTGG
GGC**CG**GG**CG**C**CG**TGGCTCA**CG**CCTGTGATCCCAGCACTTTGGGAGAC**CG**AG
G**CG**GG**CG**GATCA**CG**AGGTCAGGAGAT**CG**AGACCATCATGGCTAACA**CG**GTGA
AACCC**CG**TCTCTACTAAAAATACAAAAAATTAGC**CG**GG**CG**ATG

Figure 6.8: FANCG region 9:35,080,750-35,080,950. ***Purple*** *are probes identified by Hannon et al 2018 as being under the control of an mQTL.* ***Bold*** *are other CG sites within the region. Yellow indicates gRNA sequences 1, 2 and 3 respectively.*

C**CG**CCCTCAGCTATCTACCCTGAGCAGG**CG**C**CG**CTCACACT**CGCG**GGC**CG**TC
TCCAACCC**CGCG**CCAGCCACC**CG**AGGGGG**CG**TTCTCAGCCTCCACCCTAG**CG**
GACCC**CGCG**CCAGCACTTGTTAACC**CG**TTACCACCTGTTGACTTGGCCTCCTT
GCACTTCCATTTCATTATTTGCCAAA**CG**CAGCTTGCTTACTTTA

Figure 6.9: SNCA region 4:90,757,350-90,757,550. ***Purple*** *are probes identified by Hannon et al 2018 as being under the control of an mQTL.* ***Bold*** *are other CG sites within the region. Yellow indicates gRNA sequences 2, 1 and 3 respectively.*

6.3.2.2. Oligo annealing

gRNA arrived as pelleted ssDNA oligos from Integrated DNA Technologies, these were reconstituted to a concentration of 100µM using ddH_2O. The ssDNA gRNA oligo and its reverse complement were annealed together to form a dsDNA insert containing 5' sticky ends to enable cloning into pgRNA-modified. Furthermore the 5' ends were phosphorylated using PNK to allow for ligation of the sugar phosphate backbone. The reactions were set up as listed in **Table 6.11** and run in a thermocycler as listed in **Table 6.12**. The annealed gRNA were then diluted down 1 in 200 with ddH_2O before ligation.

Table 6.11: Annealing and phosphorylation of the gRNA to its complement for cloning.

Reagent	Volume (µl)	Final Concentration
gRNA (100µM)	1	10µM
Complementary gRNA (100µM)	1	10µM
10x T4 Ligation Buffer	1	1x
T4 PNK	1	10U
ddH_2O	6	-
Total	**10**	

Table 6.12: Thermocycling conditions to anneal gRNA oligos.

Temperature (°C)	Time	Cycles
37	30 min	1
90	5 min	1
Ramp down to 25 at 5°C per min		1
25	∞	1

6.3.2.3. Digestion of pgRNA-modified

1µg of pgRNA-modified was digested in a single digest reaction for 1 hours at 37°C with AarI as outlined in **Table 6.13**. Following digestion reactions were cleaned up using the GeneJET PCR Clean up kit as outline in **2.4.2.** . Digested plasmid was eluted in 20µl of elution buffer before being quantified on the Nanodrop as described in **2.3.4.1.** . Samples were stored at 4°C until the ligation reaction.

Table 6.13: Digestion of pgRNA-modified with AarI.

Reagent	Volume (µl)
pgRNA-modified	100ng
10x AarI Buffer	2
Oligo	0.4
AarI	1
ddH_2O	To 20

6.3.2.4. Ligation of gRNA into AarI digested pgRNA-modified

Ligations were carried out as described in **2.6.3.** using 2µl of diluted annealed oligo gRNA from **6.3.2.2.** and 100ng of digested pgRNA-modified from **6.3.2.3.** using T4 ligase. A background control consisting of only pgRNA-modified and no gRNA insert was included along with a water negative control. Reactions were ligated at 23° for one hour then left overnight at 16°C in a thermocycler. Following ligation, reactions were stored at 4°C until transformation. The names of the recombinant plasmids generated are listed in **Table 6.14**.

Table 6.14: Recombinant plasmids created for expression of gRNA for use in lentiviruses.

Recombinant Name	gRNA	Backbone
FANCG_1_pgRNA-modified	FANCG_gRNA_1	pgRNA-modified
FANCG_2_pgRNA-modified	FANCG_gRNA_2	pgRNA-modified
FANCG_3_pgRNA-modified	FANCG_gRNA_3	pgRNA-modified
SNCA_1_pgRNA-modified	SNCA_gRNA_1	pgRNA-modified
SNCA_2_pgRNA-modified	SNCA_gRNA_2	pgRNA-modified
SNCA_3_pgRNA-modified	SNCA_gRNA_3	pgRNA-modified

6.3.2.5. Transformation of ligated plasmids

5µl of ligation reactions were transformed from **6.3.2.4.** into DH5α as described in **2.6.4.** .

6.3.2.6. Colony PCR

The following morning, eight colonies were selected and colony PCR was carried out as described in **2.6.5.** . The mU6 sequencing primer was used as the forward primer (**Table 6.8**) and the reverse primers are the gRNA reverse sequences listed in **Table 6.10**. The expectant product size is 100bp. The resultant gels for *FANCG* gRNA cloning is shown in **Figure 6.11** and for *SNCA* gRNA cloning is shown in **Figure 6.12**.

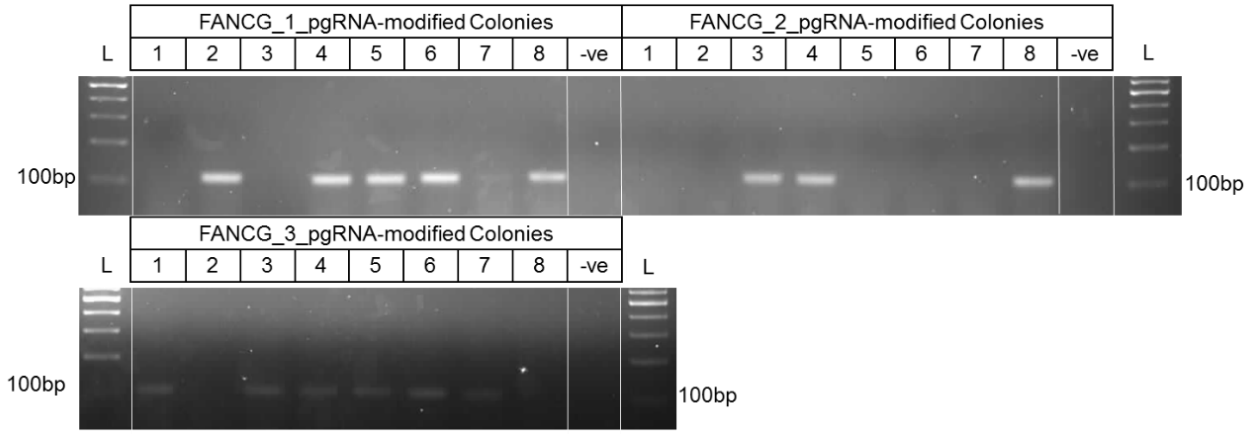


Figure 6.10: Colony PCR result for FANCG gRNA cloning. *–ve is negative control with no bacterial stab, L is 100bp ladder (solis biodyne).*

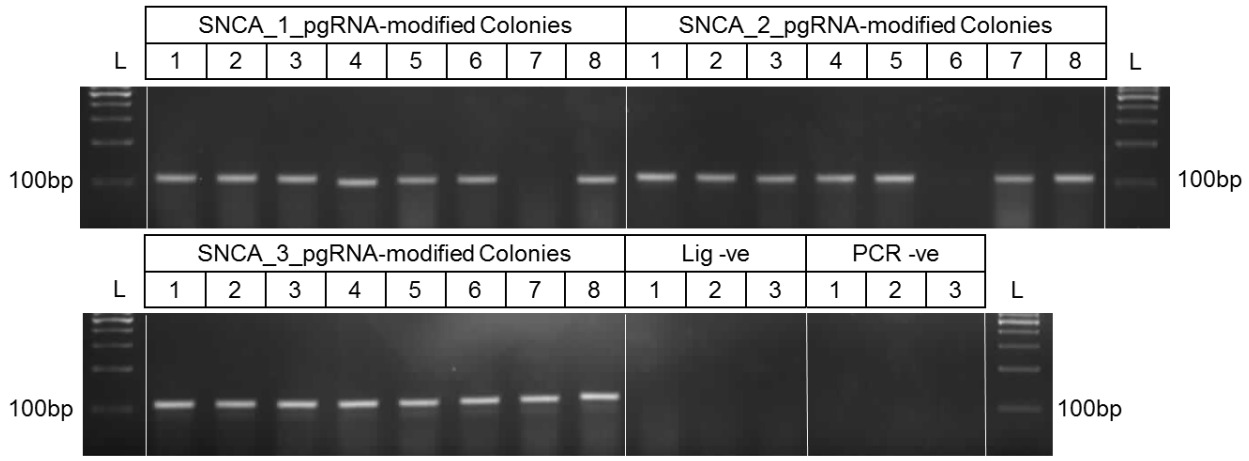


Figure 6.11: Colony PCR result for SNCA gRNA cloning. *Lig –ve is one colony from the ligation negative control reaction, PCR –ve is the negative PCR control with no bacterial stab. L is 100bp ladder (solis biodyne).*

6.3.2.7. Overnight bacterial cultures

The following afternoon, two positive colonies from the colony PCR were selected at taken into 5ml overnight cultures as described in **2.6.6.**

6.3.2.8. Purification of plasmid DNA

The following morning, plasmid DNA was purified using the GeneJET Plasmid Miniprep Kit (Thermofisher Scientific, Massachusetts, USA) as described in **2.6.7.1.** and eluted in 20µl of elution buffer.

6.3.2.9. Diagnostic digest

A diagnostic double digest was then performed on 100ng of each miniprepped plasmids with EcoRI and AarI as described in **2.6.8.** to confirm if gRNA has been correctly inserted into pgRNA-modified. The results of the *FANCG* digest are shown in **Figure 6.13** and the *SNCA* digest are shown in **Figure 6.14**.

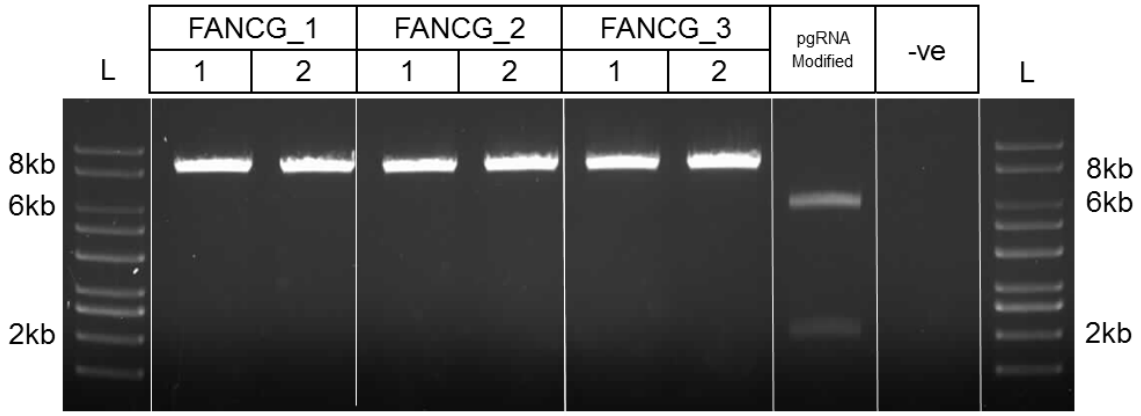


Figure 6.12: Double diagnostic digest with AarI and EcoRI to confirm insertion of FANCG gRNA into pgRNA-modified. *L is 1kb DNA ladder (solis biodyne), pgRNA-modified is the empty pgRNA-modified backbone vector, -ve is the digestion negative control where water replaces plasmid.*

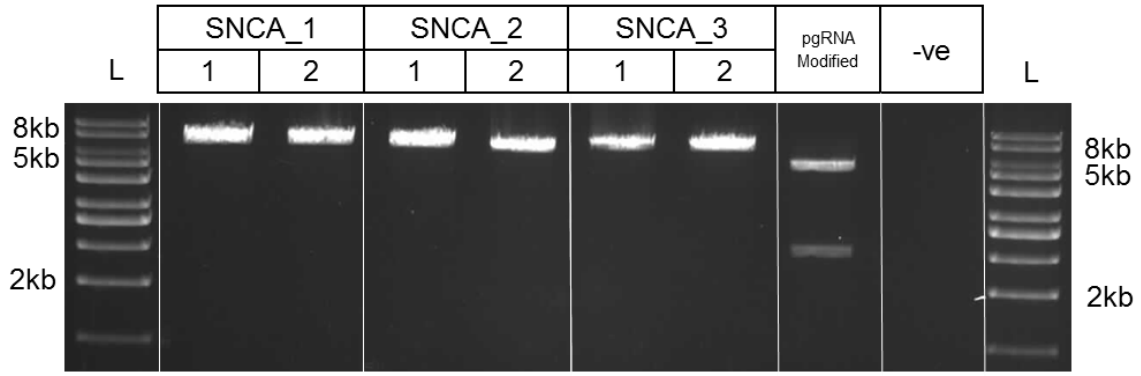


Figure 6.13: Double diagnostic digest with AarI and EcoRI to confirm insertion of SNCA gRNA into pgRNA-modified. *L is 1kb DNA ladder (solis biodyne), pgRNA-modified is the empty pgRNA-modified backbone vector, -ve is the digestion negative control where water replaces plasmid.*

6.3.2.10. Sanger sequencing of cloned plasmids

Sanger sequencing was carried out as described in **2.6.9.** . Positive plasmids from the EcoRI and AarI digest in **6.3.2.9.** were sent for Sanger sequencing by Genewiz using the common mU6 forward sequencing primer (**Table 6.8**). Sequence data was analysed in Jalview and confirmed insertion of gRNA constructs in the correct orientation as shown in **Figure 6.14**.

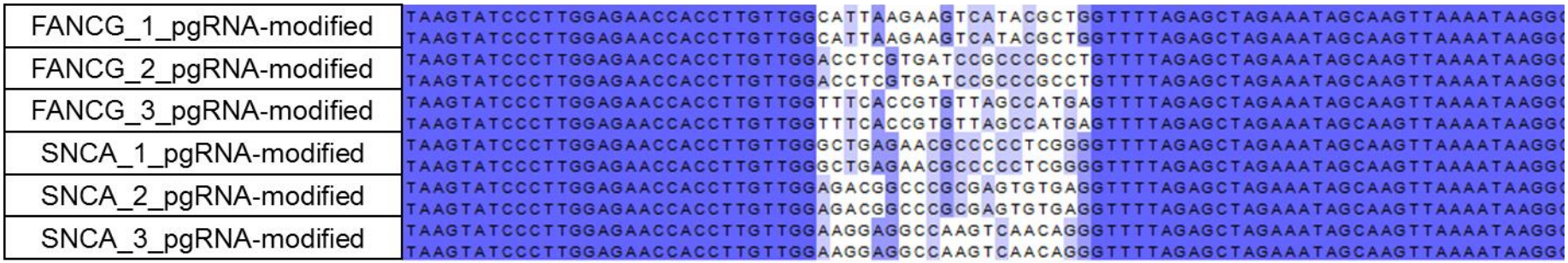


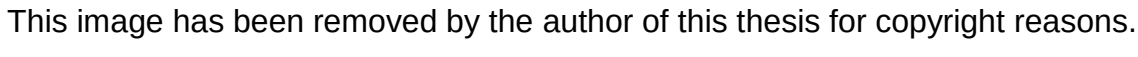
Figure 6.14: Sanger sequencing of miniprepped plasmids confirms insertion of gRNA in the correct orientation into the pgRNA-modified backbone vector*. Sequences analysed in Jalview and coloured by percentage base identity with blue indicating 100% identity and white 0% identity.*

6.3.3. Generation of lentiviruses

As we have seen in Chapter 3. the transfection efficiency of SH-SY5Y cells is low, and resulted in 15.5% of cells being double positive for two plasmids following Nucleofection. In order to increase the efficiency of delivery of CRISPR constructs into SH-SY5Y cells a lentiviral method of delivery was selected. Lentiviral generation for stable transfection of cell lines was first described by (Naldini *et al.*, 1996) and involves the generation of lentiviral particles to package and deliver DNA sequences into cells. The lentiviral system is based on the human immunodeficiency virus (HIV), as a result the lentivirus can infect human cells with relative ease. To increase biosafety of lentiviral vectors the genes required for particle generation are split over several plasmids. When all of these plasmids are transfected into HEK293T cells the individual lentiviral genes are transcribed and results in the assembly of functional lentiviral particles. These then package the CRISPR machinery required for epigenetic editing at the cell membrane when the lentivirus buds off of the cell into the growth media. This growth media is then harvested at several time points and the lentiviral particles are collected. These viruses can then be added to any mammalian cell line and results in the transduction of the CRISPR-Cas9 machinery (**Error! Reference source not found.**). The packaged genes are then reverse transcribed into cDNA and these then integrate into the host cell genome. Once integrated the host cell machinery results in the transcription of the dCas9-DNMT3a/Tet1CD and gRNA and results in epigenetic editing.

Lentiviruses are beneficial as they can integrate into non-dividing cells as well as dividing cells, this integration is then kept once the cell divides, thus generating a stable cell line. This then results in a high level of integrated protein expression compared to transient transfection, as the cell divides the integrated virus is copied along with the constitutive DNA, whereas plasmids are not duplicated in cell division and are subsequently lost. The lentiviruses also contain engineered mutations to prevent them replicating once they have integrated, this prevents consistent generation and infection of cell cultures and prevents saturation of integration into the cells. However there are several negatives of using lentiviruses, the integration of lentiviral vectors into the host genome is random and can result in non-stable cell lines, and, as the cell line continues to divide, the integrated sequences can be lost. There is also the risk of recombination leading to replication competent lentiviruses, however

the risk of this is reduced due to the introduction of self-inactivating sequences (Zufferey *et al.*, 1998).


This image has been removed by the author of this thesis for copyright reasons.

6.3.3.1. Transfection of vectors into HEK293T cells

HEK293T cells were maintained as described in **2.1.2.** . The day prior to transfection, early passage HEK293T cells were seeded at a density of $4x10^6$ cells in 10cm dishes with growth media (DMEM/F-12, with Glutamax, 10% FBS) as described in **2.1.** . The following morning a transfection mixture containing the plasmids required for lentiviral generation, Lipofectamine LTX (ThermoFisher Scientific, Massachusetts, USA) and Opti-MEM reduced serum media (Gibco. Massachusetts, USA) as listed in **Table 6.15**. The mixture was left to equilibrate at room temperature for 30 minutes at room temperature. Following incubation growth media was removed from the HEK293T cells and replaced with 5ml of pre-warmed Opti-MEM reduced serum media. The LTX plasmid mix was then added, in a dropwise circular motion, to the HEK293T cells in 10cm dishes. Cells were then incubated at 37°C 5% CO_2 for six hours before the media was removed and replaced with 10mls of fresh growth media.

24 hours post transfection cells were checked for BFP and RFP fluorescence using a EVOS FLoid Cell Imaging Station (Thermofisher, Massachusetts, USA) as the lentiviral transfer plasmids contain a fluorescent marker to check transfection efficiency (**Figure 6.17**). Growth media was removed and replaced with 6mls of fresh growth media. Cells were returned to the incubator and incubated at 37°C 5% CO_2.

Table 6.15: Transfection mixture to generate lentivirus in HEK293T cells.

Component	Volume (µl)
Fuw-dCas9-DNMT3a/Tet1CD-P2A-tagBFP OR pgRNA-modified	4µg
psPAX2	2µg
pmD2.G	2µg
Lipofectamine LTX (ThermoFisher Scientific, Massachusetts, USA)	15
Opti-MEM (Gibco, Massachusetts, USA)	500

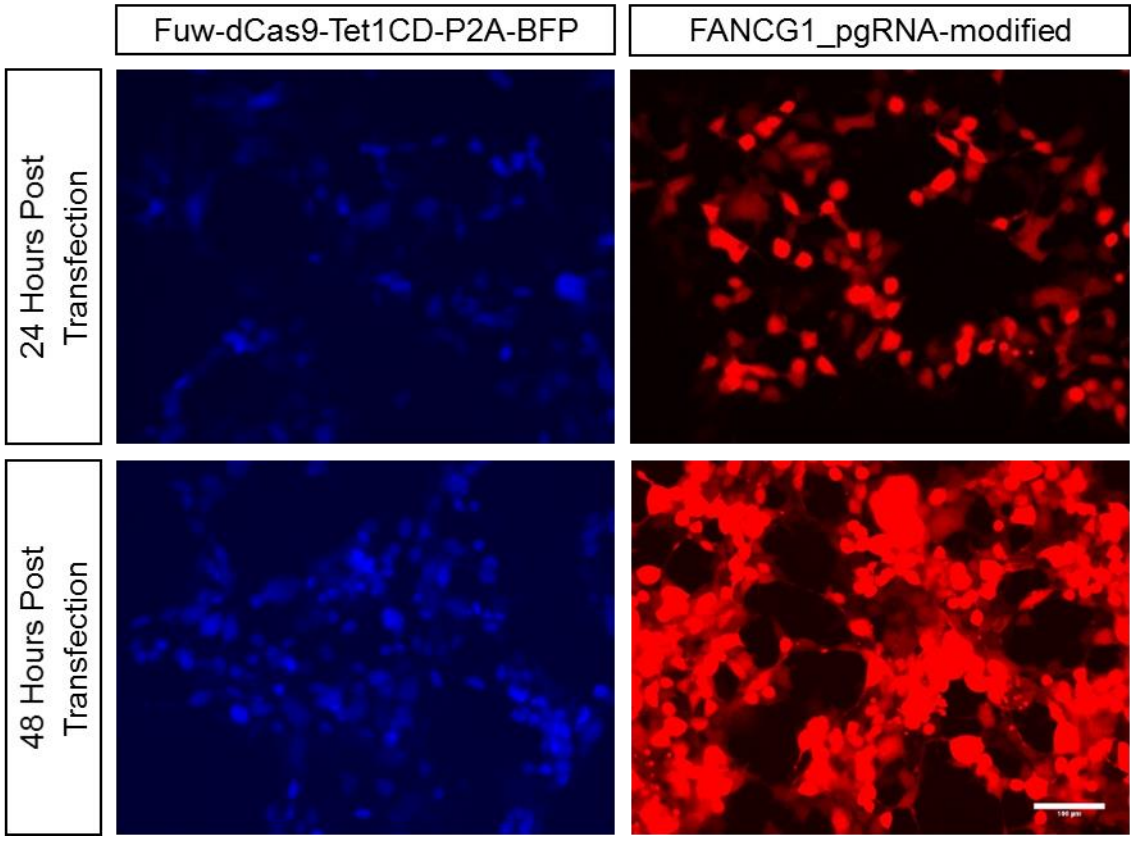


Figure 6.15: Example fluorescence images 24 and 48 hours post transfection of lentiviral vectors into HEK293T cells. *The DNMT3a and Tet1CD vectors contain BFP selection and the pgRNA-modified vector contains mCherry selection. Scale bar 100µm.*

6.3.3.2. Lentiviral harvesting

Lentiviral particles were harvested at 48 and 72 hours post transfection from the HEK293T cell media. At 48 hours post transfection 6mls of growth media are collected from the cells into a 15ml falcon and 3µl of 10mg/ml of polybrene solution (final concentration 5µg/ml) was added to each sample and mixed. Polybrene is a known to increase the transduction of retroviruses into the host cell line (Davis, Morgan and Yarmush, 2002). These were then stored at 4°C until the 72 hour collection. 6mls of fresh growth media was added to the HEK293T cells and these were incubated at 37°C and 5% CO_2 for another 24 hours. At 72 hours the process was repeated, with the 6mls of media added to the 6mls of media previously collected, another 3µl of 10mg/ml polybrene solution was added to the falcon tubes. The media was then centrifuged at 3000rpm for five minutes to pellet any cells that have been collected. Following centrifugation 11mls of media was collected and filtered through a 0.45µm filter into a clean 15ml falcon. Viral titre was then undertaken or viruses stored at -80°C for long term storage.

6.3.3.3. Rough titration of lentivirus

Following collection lentiviruses underwent titration to quantify the number of viral particles created and to calculate the correct volume of virus to add to cells for successful transduction. Rough viral titre was calculated using Lenti-X GoStix Plus (Takara Bioscience, California, USA) following manufacturer's instructions. This is a lateral flow based method which quantifies the lentiviral antigen p24, the binding of this antibody allows for detection of a titre based on the intensity of the band compared to the control band. These intensities are measured by a mobile phone application and gives a corresponding GO value (GV) between 0 and 1000 which positively correlates to lentiviral titre. An example of a good titre and bad titre from the Lenti-X GoStix Plus is shown in **Figure 6.16**.

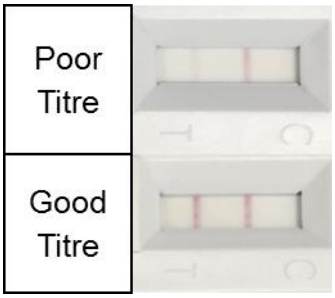


Figure 6.16: Example results from the Lenti-X GoStix Plus showing a poor and good lentiviral titre. *The test band (T) is much stronger compared to the control band (C) in the good titre.*

6.3.4. Infection of SH-SY5Y cells

6.3.4.1. Infection of SH-SY5Y cells for accurate titration

If the GV from **6.3.3.3.** was high then further titration was undertaken using FACS. $1x10^5$ SH-SY5Y cells were seeded in a 6 well plate and 12 hours post seeding different dilutions of lentivirus was added to the wells. 24 hours post infection media on the cells was replaced to remove any remaining lentiviral particles and imaged using the EVOS FLoid Imaging Station. At 72 hours post infection, cells were harvested as described in **2.1.** and resuspended in DMEM/F-12 with no phenol red (Gibco, Massachusetts, USA) and FACS was undertaken as described in **2.2.** to quantify the percentage of BFP or mCherry positive cells. The volume of virus was then correlated to the percentage of fluorescent cells to calculate the volume of virus required to achieve maximal infection efficiency in the cell line of choice. There are stark differences in the lentiviral titres of the four dCas9 vectors as shown in **Figure 6.19**. The Fuw-dCas9-DNMT3a-P2A-tagBFP showed the lowest titre of the four due to the larger size (16kb vs 14kb). To achieve maximal editing, 600µl of Fuw-dCas9-Tet1CD-P2A-BFP and Fuw-dCas9-Tet1CD_IM-P2A-BFP virus, 500µl Fuw-dCas9-DNMT3a_IM-P2A-tagBFP, and 3ml of Fuw-dCas9-DNMT3a-P2A-tagBFP were used per transduction. When the SNCA_1_pgRNA-modified was titrated much better transduction efficiencies were observed (**Figure 6.20**), however as the volume of virus increased as did the bleed through of fluorescence into the DAPI channel used for BFP selection. Therefore a total volume of 25µl of gRNA virus was used per transduction to minimise the risk of sorting false positives.

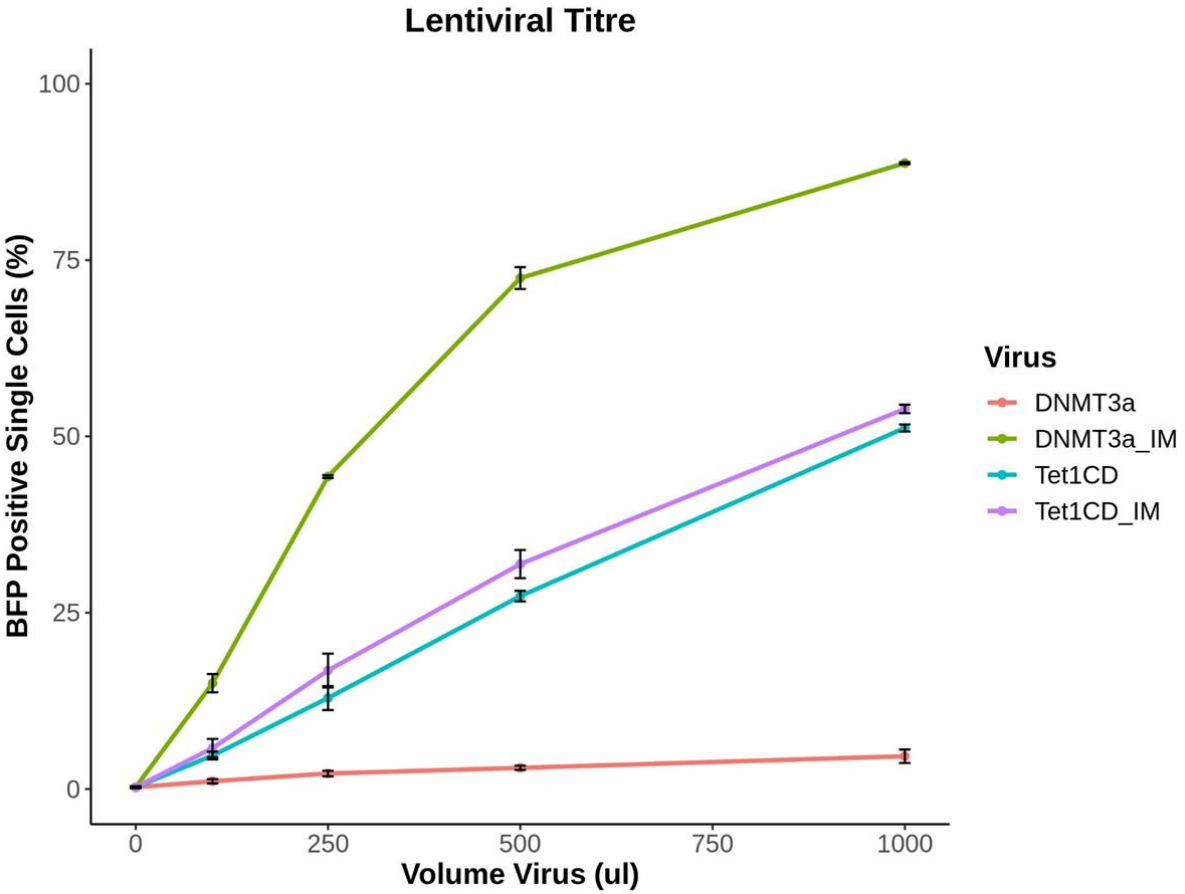


***Figure 6.17: Viral titre is dependent on transfer plasmid size**. Fuw-dCas9-DNMT3a-P2A-tagBFP has the lowest titre compared to the other transfer plasmids, this is due to its increased size. N=2 transductions per virus.*

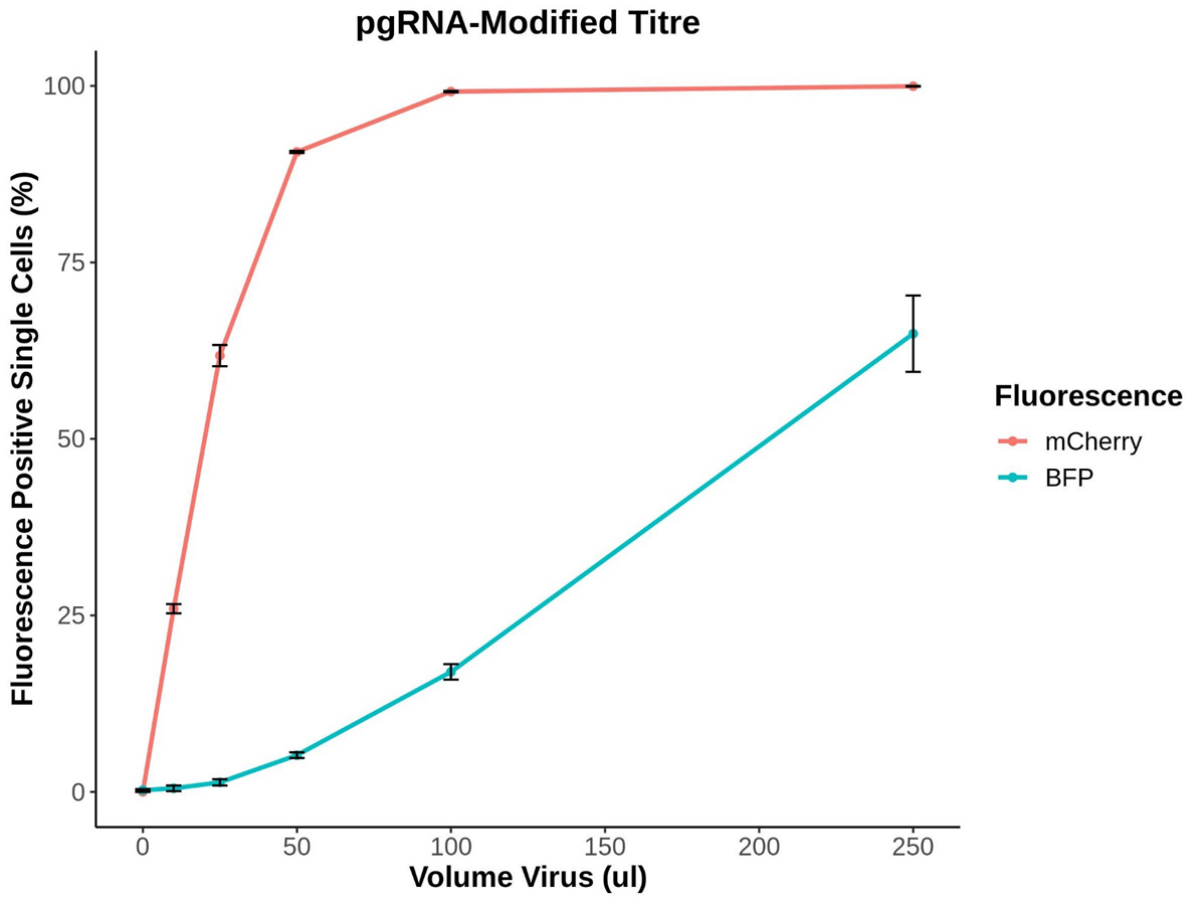


Figure 6.18: pgRNA-modified has a much better viral titre however excess volumes of virus leads to bleed through into the DAPI channel on the FACS. *25µl of pgRNA-modified virus with gRNA was used per transduction to avoid fluorescent bleed through into the DAPI channel and the sorting of false positives. N=2 transductions.*

6.3.4.2. Infection of SH-SY5Y for epigenetic editing

Once the viral titre was calculated the SH-SY5Y cell could be cotransduced with gRNA and Tet1CD/DNMT3a viruses. To do this, $1x10^5$ SH-SY5Y cells were seeded in a 6 well plate the night prior to infection. The following morning, media was removed from the cells and cells washed with 1ml HBSS, 2mls of fresh growth media was then added to each cell. For each infection, a total of 25µl pgRNA-modified virus and 500µl of one of the following were used: Fuw-dCas9-DNMT3a-P2A-tagBFP, Fuw-dCas9-DNMT3a_IM-P2A-tagBFP, Fuw-dCas9-Tet1CD-P2A-BFP, Fuw-dCas9-Tet1CD_IM-P2A-BFP. Single infection positive controls and a non-infection negative control were also included as standard. Viruses were added in a circular dropwise motion to the cells. Final media volumes were made up to 3ml using extra growth media. Plates were then agitated in a cross pattern to distribute the virus. Plates were then incubated at 37°C and 5% CO_2.

24 hours post infection media was removed and 3mls of fresh growth media was added. Fluorescence was checked using the EVOS FLoid microscope as lentiviral integration results in a mCherry fluorescent signal for the gRNA and BFP signal for DNMT3a and Tet1CD constructs, however BFP is not visible under the microscope available. 72 hours post infection cells were passaged at a ratio of 1:1 into T75 flasks as described in **2.1.3.** . 7 days post infection cells were harvested as described in **2.1.3.** and resuspended in 500µl of phenol-free DMEM/F-12 before FACS.

6.3.5. Fluorescent activated cell sorting

FACS was carried out to isolate pure cell populations containing cells which have integrated both the DNMT3a or Tet1CD lentivirus and pgRNA-modified lentivirus. This was done to reduce the false positive rate as not all the bulk cells will contain the CRISPR machinery to edit the epigenome. FACS sorting was carried out as described in **2.2.** . DNA and RNA samples were collected as described in **2.2.** . For RNA samples, the cells were sorted directly into 500µl of TRIzol LS (ThermoFisher Scientific, Massachusetts, USA). For mCherry identification the PE-CF594 laser was used, and for BFP identification the DAPI laser was used.

6.3.6. DNA and RNA extractions

DNA and RNA extractions were carried out as described in **2.3.** on sorted populations of cells with some slight modifications. For DNA extractions the

Slagboom buffer was made up in the Eppendorf of sorted cells by making the sorted cell volume to 400µl with ddH_2O then adding 50µl of 10x STE and 50µl of 5% SDS. For the RNA extractions the Direct-zol™ RNA MicroPrep kit (Zymo Research, California, USA) was used. This protocol is the same as the one described in **2.3.3.** .

6.3.7. Bisulphite conversion (BC)

BC was carried out as outlined in **2.7.1.** . However all DNA extracted from **6.3.6.** was converted instead of 500ng due to the low amounts recovered from FACS. This was to maximise BC DNA output.

6.3.8. Pyrosequencing

Pyrosequencing is designed to detect base changes of DNA prepared from biological samples. The principle is based around the formation of the sugar phosphate backbone of DNA. This process releases a pyrophosphate group following addition of a free dNTP to the growing DNA strand. This pyrophosphate is then converted into ATP, which is a cofactor for luciferase activity. As a result, when dNTP are added sequentially and the correct base is incorporated into the backbone, light is released and recorded. This allows for the detection of modified bases at specific genomic locations (Harrington *et al.*, 2013). As a result, pyrosequencing has become the gold standard for detection of DNA methylation at CpG sites following BC. To do this, the region to be examined is first amplified by PCR with biotinylated PCR primers as described in **2.4.** with the cycling conditions listed in **Table 6.16**. In addition sequencing primers were also designed to the region of interest. All primers for PCR and pyrosequencing were designed using the PyroMark Assay Design Software (Qiagen, Venlo, Netherlands). Pyromark assays were created using the PyroMark Q24 software (Qiagen, Venlo, Netherlands) using standard conditions. All the primers for pyrosequencing are listed in **Table 6.17**. The following sections describe the protocol for pyrosequencing as per the manufacturer's instructions.

6.3.8.1. Immobilisation of PCR product

Sepharose beads (Steptavidin Sepharose High Performance, GE Healthcare) were gently shaken to resuspend before being combined with binding buffer, ddH_2O, and biotinylated PCR product in a 96 well plate as described in **Table 6.18**. The plate was sealed and agitated for 10 minutes at 1,400rpm.

Table 6.16: Pyrosequencing PCR cycling conditions for FANCG and SNCA.

Step	Temperature (°C)	Time	Number of Cycles
Hotstart	95	15 min	1
Denaturation	95	30 sec	
Annealing	62	30 sec	35
Extension	72	60 sec	
Final Extension	72	10 min	1

Table 6.17: Pyrosequencing PCR and sequencing primers for FANCG and SNCA.

PCR Reaction	Primer Name	Sequence (5' – 3')
FANCG	FANCG_pyro_F	ATGATTTTGGGTTTGGTAATAAGTTGT
	FANCG_pyro_R	[Btn]CCTTTCTCCTACCTCAACCT
	FANCG_pyro_S	ATAAAGATATTATTAAGAAGTTATA
SNCA	SNCA_pyro_F	GTTTGGTAAATAATGAAATGGAAGTGTA
	SNCA_pyro_R	[Btn]AACAACAAACCCAAATATAATAAT
	SNCA_pyro_S	GGTGGAGGTTGAGAA

Table 6.18: Components to immobilise PCR product onto sepharose beads. Volumes given here are per sample.

Reagent	Volume (µl)
Sepharose beads (Streptavidin Sepharose High Performance, GE Healthcare)	2
Binding Buffer (Qiagen, Venlo, Netherlands)	40
Biotinylated PCR product	20
ddH_2O	18
Total	**80**

6.3.8.2. Preparing the workstation

Before beginning a pyrorun the vacuum workstation needs to be prepared. The appropriate troughs were filled with the following reagents:

a. 50ml 70% Ethanol
b. 50ml Denaturation Solution (0.2M Sodium Hydroxide (NaOH)), (Qiagen, Venlo, Netherlands)
c. 70ml 1x Wash Buffer (Qiagen, Venlo, Netherlands)
d. 70ml ddH_2O
e. 70ml ddH_2O

Once all troughs were filled the vacuum pump was switched on and filter probes were lowered into trough e. The water was washed through the probes and the trough was then refilled. The vacuum pump was then turned off and returned to the "parked" position.

6.3.8.3. Sequencing primer set up

Before sequencing the sequencing primers from (**Table 6.17**) were diluted to a final concentration of 0.3µM in Annealing Buffer (Qiagen, Venlo, Netherlands). 25µl of the diluted sequencing primer was then pipetted into each well of the Q24 plate.

6.3.8.4. Combining the plate

Once the plate from **6.3.8.1.** had finished agitation it was moved to the workstation. The vacuum tool was switched on and the probes immersed into the wells of the PCR plate. The tool was then moved to the ethanol trough (a) and the ethanol was aspirated for five seconds to wash the DNA bound sepharose beads. The tool was then moved to the denaturation trough (b) and the solution aspirated for five seconds. The tool was then moved to the wash trough (c) and the solution aspirated for 10 seconds. The vacuum tool was then raised to 90° and the probes left to dry for 20 seconds. The vacuum tool was then switched off and lowered into the Q24 plate containing sequencing primer from **6.3.8.3.** . The vacuum tool was then agitated side to side to dislodge the DNA bound sepharose beads into the Q24 well. The Q24 well was then transferred to a heat block and incubated at 80°C for two minutes before being left to stand at room temperature for five minutes before being loaded onto the pyrosequencer.

6.3.8.5. Loading the pyrosequencer

The sequencing cartridge was loaded with the reagents as described for each assay in the pre-run information produced by the PyroMark Q24 software in **6.3.8.** . The loaded cartridge and Q24 plate from **6.3.8.4.** were loaded into the sequencer and the appropriate run was loaded onto the machine using a USB flash drive. The program was then left to run until completed.

6.3.8.6. Data analysis

Analysis was performed using the PyroMark Q24 software. Using the inbuilt quality checker, traces were included or excluded for downstream analysis. The 5mC values are reported as a percentages of modifications at each CpG site and these were used for downstream analysis.

6.3.9. qPCR for genes of interest

100ng of RNA extracted from the sorted SH-SY5Y cells was converted into cDNA using the SuperScript VILO cDNA Synthesis Kit as described in **2.5.1.** . qPCR was carried out for *FANCG, PIGO,* and *STOML2* as described in **2.5.2.** . The housekeeping genes *β-Actin* and *GAPDH* were used for normalising. Analysis was carried out as described in **2.5.4.** using the Tet1CD_IM transduced cells as baseline, all fold changes are shown as fold changes relative to Tet1CD_IM. The primers used for qPCR are listed in **Table 6.19**.

Table 6.19: qPCR primers for gene expression analysis.

Gene	Primer Name	Sequence (5' – 3')
FANCG	FANCG_qPCR_F	TAGTTGAGGCCTTGAATGTC
	FANCG_qPCR_R	CTTGCTAGTATGTGCTTGGT
PIGO	PIGO_qPCR_F	CCCTGGTCCTTTTACTGTGC
	PIGO_qPCR_R	GCCAATGGATAGCTGGAAAG
STOML2	STOML2_qPCR_F	GTGACTCTCGACAATGTAAC
	STOML2_qPCR_R	TGATCTCATAACGGAGGCAG

6.4. Results

6.4.1. Engineered control plasmids show BFP expression and retain their point mutations

The first aim of this chapter was to generate control vectors which contain a fluorescent selection marker. At the time of this work there were no lentiviral vectors on Addgene which contained the inactive forms of DNMT3a and Tet1CD fused to a fluorescent selection marker. In order to achieve this I PCR cloned the DNMT3a_IM and Tet1CD_IM genes from the available Fuw-dCas9-dnmt3a_IM and Fuw-dCas9-Tet1CD_IM vectors that had no fluorescence selection markers into the commercially available Fuw-dCas9-DNMT3a-P2A-tagBFP and Fuw-dCas9-Tet1CD-P2A-BFP vectors and in the process replacing the catalytically active forms of DNMT3a and Tet1CD with the inactive forms.

Following Sanger sequencing the E664A and E756A mutations that render the DNMT3a protein inactive are still present in the recombinant DNMT3a_IM plasmid and not in the active version of the plasmid Fuw-dCas9-DNMT3a-P2A-tagBFP (**Figure 6.21**). The H1672Y and D1674A mutations are still present in the Tet1CD_IM plasmids and are not in the active plasmid Fuw-dCas9-Tet1CD-P2A-BFP (**Figure 6.22**). These results confirm that the recombinant IM plasmids generated are inactive and can be used as a control and will not alter the methylation state of the targeted CpGs. The stop codon at the end of the DNMT3a_IM/Tet1CD_IM sequence has been removed following cloning to allow a clean open reading frame through the dCas9, Tet1CD/DNMT3a, and the P2A linker to the end of the BFP gene. Transfection of 1µg of these plasmids into HEK293 cells confirms that these inactive vectors successfully express BFP after 24 hours and can now be used as a matched control for lentiviral work **Figure 6.23**.

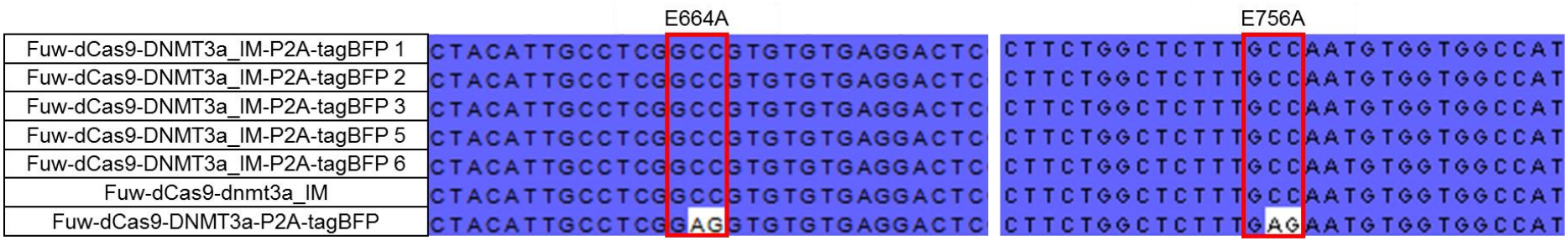


Figure 6.19: Sanger sequencing confirms that the point mutations which render the DNMT3a enzyme inactive remain following cloning. *These point mutations are in the codons coding for glutamic acids at 664 and 756 and result in a translation change to alanine, rendering the protein inactive. Sequences analysed in Jalview and coloured by percentage base identity with blue indicating 100% identity and white 0% identity. The sequencing reaction for Fuw-dCas9-DNTM3a_IM-P2A-tagBFP 4 terminated early.*

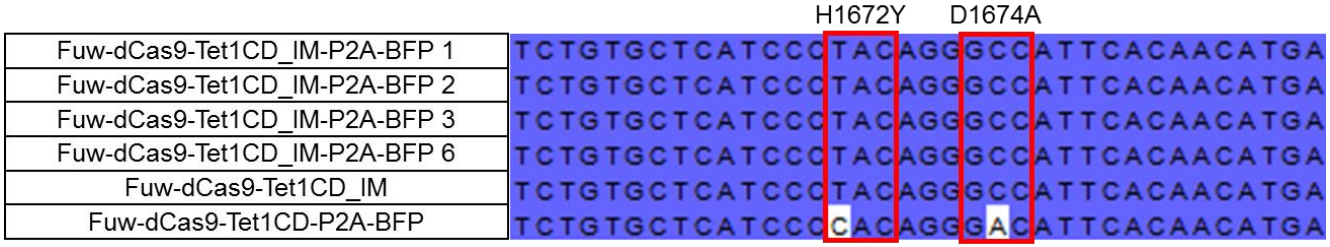


Figure 6.20: Sanger sequencing confirms that the point mutations which render the Tet1CD enzyme inactive remain following cloning. *These result in the codons coding for histadine at 1672 and aspartic acid at 1674 being translated as tyrosine and alanine respectively, rendering the protein inactive. Sequences analysed in Jalview and coloured by percentage base identity with blue indicating 100% identity and white 0% identity. The sequencing reaction for Fuw-dCas9-Tet1CD_IM-P2A-BFP 5 failed to prime and was not used for subsequent experiments.*

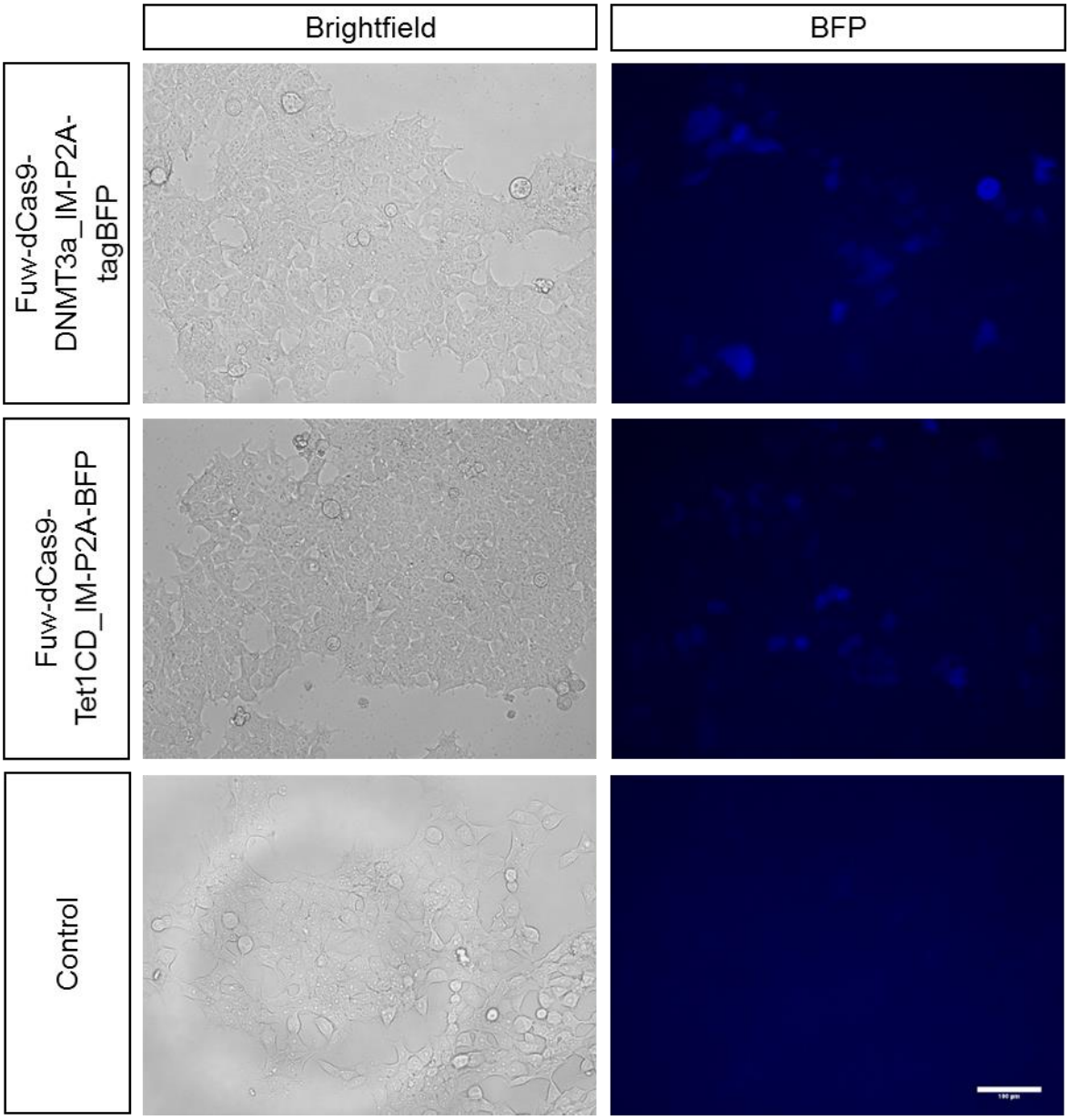


Figure 6.21: Fluorescent images 24 hours post transfection of 1µg of plasmid into HEK293 cells. *The constructed created in this chapter show BFP expression. Scale bar 100µm.*

6.4.2. An overview of the epigenetic editing protocol

There are currently limited methodologies available to change methylation at a specific CpG sites to uncover functional roles of differential methylation. However recent advances in CRISPR/Cas9 technology have allowed us to edit methylation at specific CpG sites. Therefore the first aim of this chapter was to adopt the methods first described by (Liu *et al.*, 2016) to develop a protocol to edit DNA methylation at specific CpG sites *in vitro* and to determine if these changes results in targeted gene expression changes. A simplified overview of the final protocol is shown in **Figure 6.24**.

Firstly site selection was carried out as described in **6.1.** to identify regions to target containing the most CpG probes in the highest density which are which have shared mQTLs. A panel of gRNA were then designed over these CpG sites using the online tool available from www.benchling.com. These gRNA were then cloned into a lentiviral gRNA construct, pgRNA-modifed, using the AarI cloning site. Once successfully cloned we then developed a method for generating lentiviruses by transfecting transfer (either pgRNA-modified, or dCas9 fused to Tet1CD/Tet1CDIM/DNMT3a/DNMT3aIM), packaging, and envelop plasmids in HEK293T cells. These plasmids express the genes required for successful lentiviral generation. Viral particles were collected from the HEK293T cells at 48 and 72 hours post transfection and purified by centrifugation and filtration. Once viruses were collected they were quickly titred using Lenti-X GoStix to check sufficient virus was available for precise titration. To perform precise titration, 100,000 SH-SY5Y cells were seeded in six well plates and transduced with dilutions of each virus. 72 hours after being transduced the cells underwent FACS to determine how much virus was required for a 60% transduction rate based on the incorporated fluorescent reporters.

Following titration, another 100,000 SH-SY5Y cells were seeded in six well plates and transduced with lentiviruses containing gRNA and lentivirus containing dCas9 fused to either Tet1CD/Tet1CDIM/DNMT3a/DNMT3aIM. We aimed for a 60% transduction efficiency of gRNA and 60% transduction efficiency of dCas9-effector. After 72 hours cells were passaged into T75 flasks at a ratio of 1:1, transduction rates could then be roughly measured using a fluorescent EVOS FLoid Cell Imaging Station (ThermoFisher Scientific, Massachusetts, USA)**.** At seven days post transduction the

cells underwent FACS for double positive mCherry (gRNA) and BFP (dCas9-effector). The double positive fractions were collected and DNA and RNA extracted. DNA was then BC and pyrosequenced to profile specific DNA methylation changes. The RNA was converted to cDNA and used for gene expression profiling to examine if any genes in *cis* of the CpG sites targeted have resultant changes in gene expression.

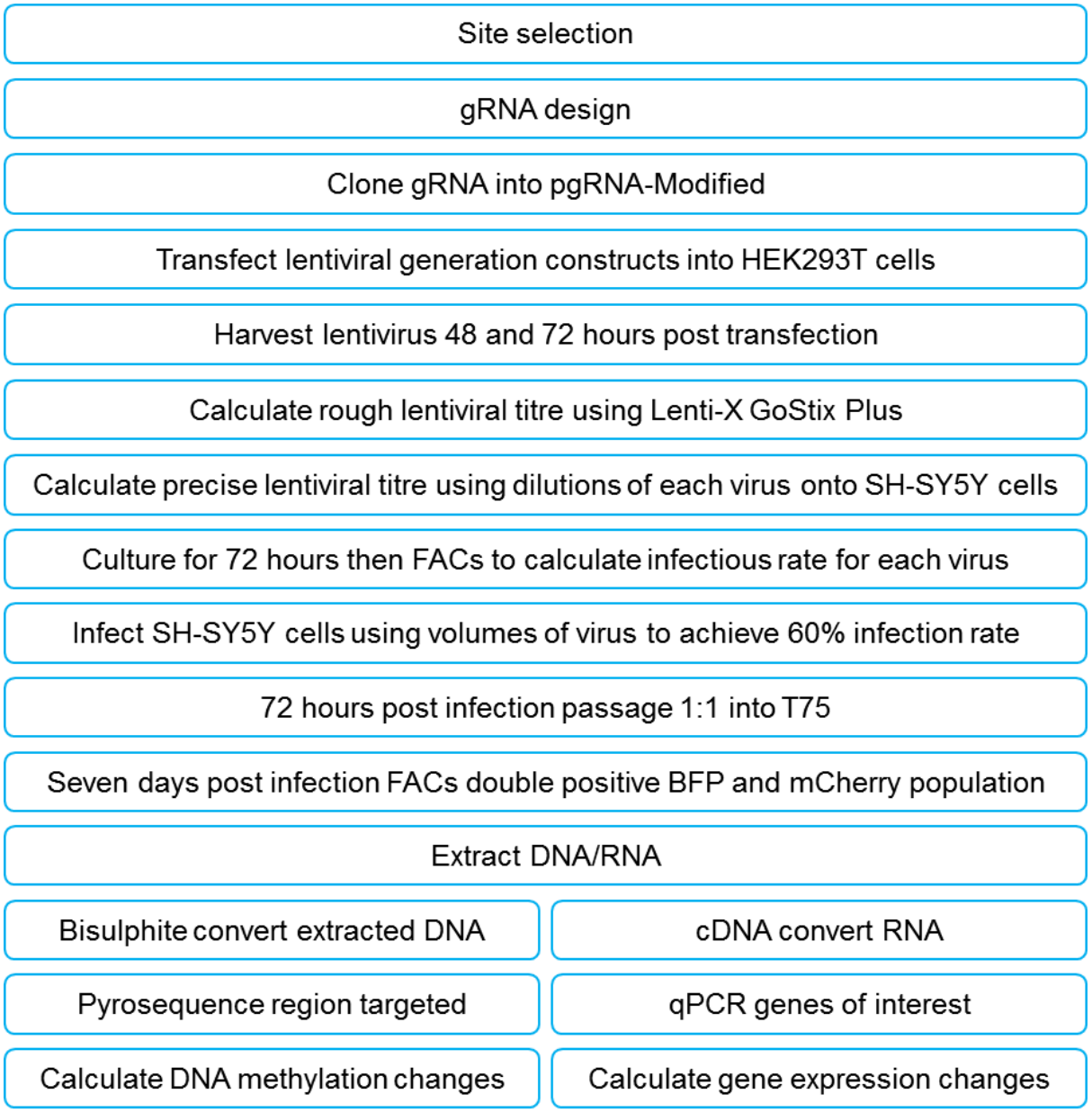


Figure 6.22: An overview of the protocol developed for functional validation of differentially methylated CpG sites in vitro.

6.4.3. Methylation is reduced at the *FANCG* locus using Tet1CD in a gRNA dependent manner

The third objective of this chapter was to identify if we can reduce the methylation of specific hypermethylated CpGs using a modified form of CRISPR/Cas9 with Tet1CD. To do this we targeted a DMR within the *FANCG* promoter which contains four CpG sites thought to be associated gene expression of *STOML2* and *PIGO* (Hannon *et al.*, 2018). I believed that reducing the DNA methylation I would see a change in the gene expression of *STOML2* or *PIGO* as predicted by SMR. To reduce the DNA methylation in these probes I designed two gRNA flanking the four CpGs and generated lentiviruses containing each of these gRNAs a lentivirus containing dCas9 fused to Tet1CD. I then transduced SH-SY5Y cells with both Tet1CD and each individual gRNA or a combination of both gRNAs. After seven days I FACS double positive mCherry and BFP cells and extracted DNA. The DNA was then bisulphite converted and pyrosequenced. I hypothesised that we would see a decrease in the DNA methylation state of the targeted CpGs by Tet1CD.

The DNA methylation change results are summarised in **Table 6.20**. In all three experimental conditions I observe a reduction of DNA methylation in all five probes within the region when targeted with Tet1CD compared to Tet1CD IM control. There is large variation in the effect on DNA methylation between the three experimental conditions. This can be explained by the differences in gRNA quality. In the samples transduced with gRNA 1 only we see large changes in DNA methylation. The largest change in DNA methylation is at CpG 9:35080824 (-16.5% decrease) and the smallest at CpG 9:35080812 (-3% decrease) with an average methylation decrease across all probes at -7.5% (**Figure 6.25**). Interestingly the CpG site with the largest decrease in DNA methylation is the furthest away from the PAM sequence (CpG 9:35080824). In contrast, the samples transduced with only the gRNA 2 I only see a modest decrease in DNA methylation over the sites targeted compared to gRNA 1 (**Figure 6.26**). The average change in DNA methylation over the five sites is only -2.4% with gRNA 2 compared to -7.5% with gRNA 1. This could be due to the distance from the PAM sequence or the quality of the gRNA. This effect can improved by using a combination of both gRNA 1 and gRNA 2 which increases the efficiency of the DNA methylation reduction. However this does not reach the levels of using gRNA 1 only, this could be due to decreased volumes of gRNA 1 (**Figure 6.27**).

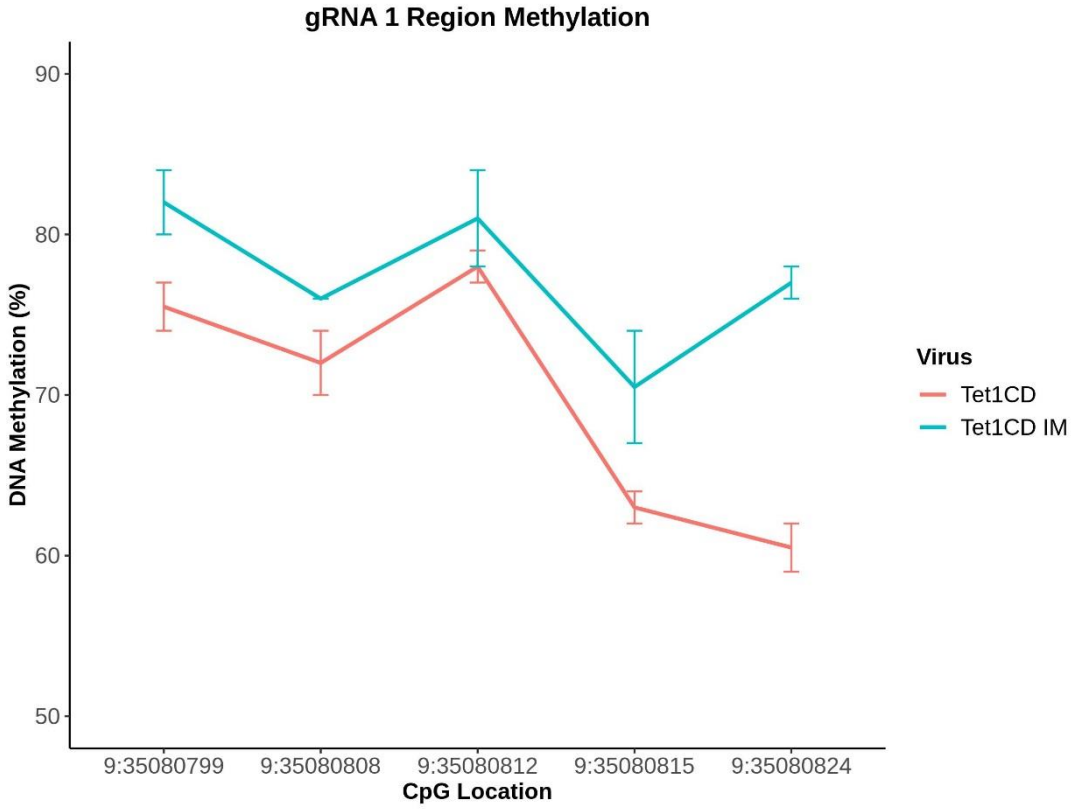


Figure 6.23: Using gRNA 1 only we observe a large decrease in DNA methylation when co-transduced with Tet1CD compared to Tet1CD IM control. *N=2 biological repeats, error bars are +/- SEM.*

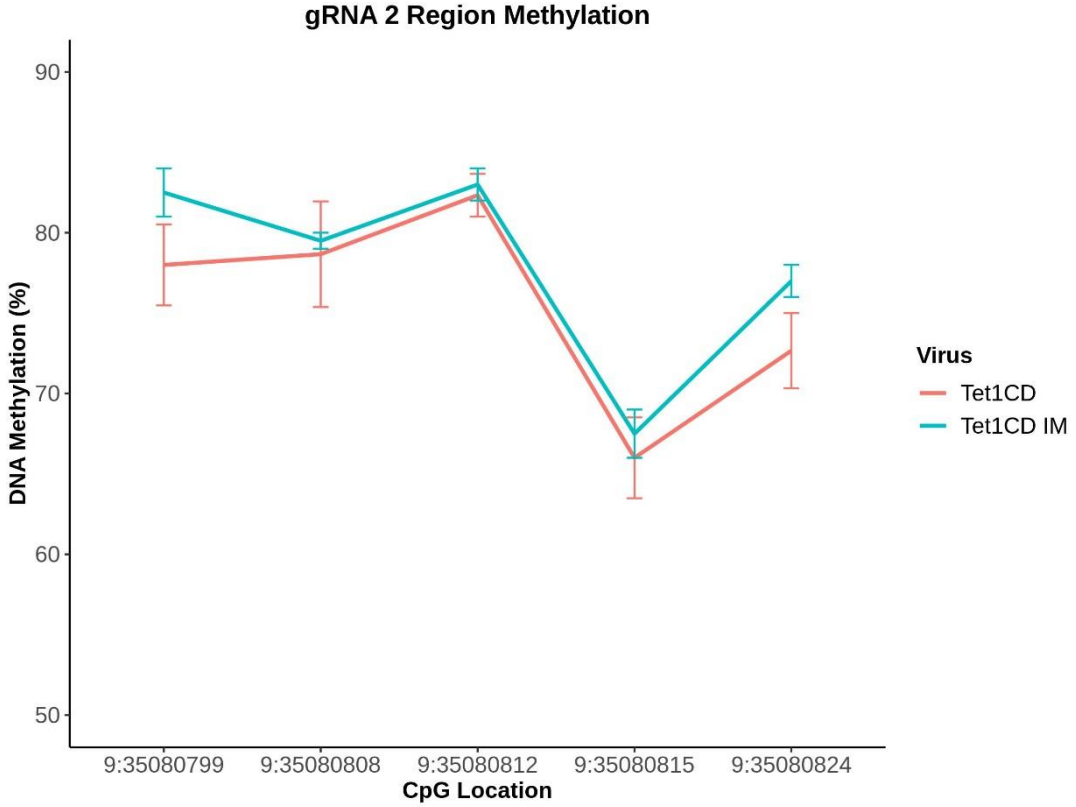


Figure 6.24: When using gRNA 2 only there is only a modest change in DNA methylation compared to gRNA 1. *N=2 biological repeats, error bars are +/- SEM.*

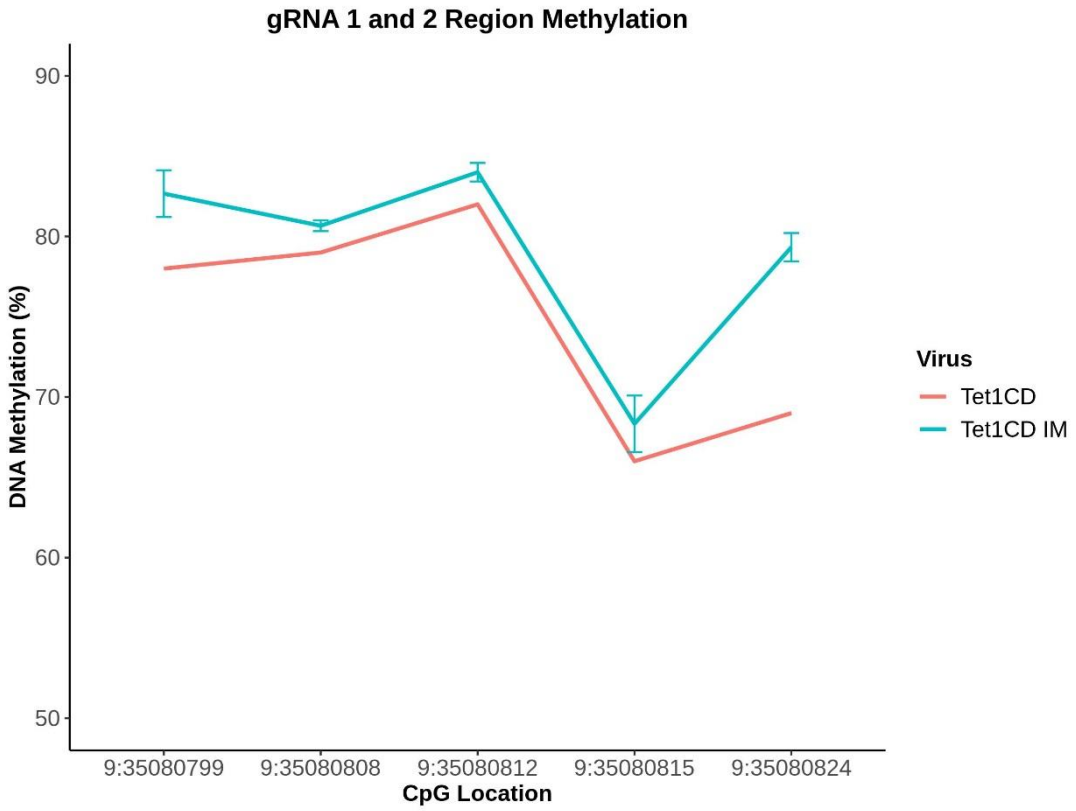


Figure 6.25: When using a combination of both gRNA we see a decrease in DNA methylation. *N=2 for Tet1CD IM, N=1 for Tet1CD, error bars are +/- SEM.*

Table 6.20: A summary of the DNA methylation changes observed in each experimental condition of each CpG in the region.

CpG Site	Average DNA Methylation Change (%)		
	gRNA 1	gRNA 2	gRNA 1/gRNA 2
9:35080799	-6.5	-4.5	-4.7
9:35080808	-4.0	-0.83	-1.7
9:35080812	-3.0	-0.67	-2.0
9:35080815	-7.5	-1.5	-2.3
9:35080824	-16.5	-4.3	-10.3

6.4.4. Gene expression changes are observed in *FANCG* but not *STOML2* or *PIGO* in samples targeted with FANCG_gRNA_1

Following successful confirmation of reduced DNA methylation at the targeted CpGs I next set out to examine if there are any changes in the gene expression of *FANCG, STOML2,* or *PIGO.* Following qPCR I identified that the changes in expression are minimal and are dependent on the level of demethylation. The level of demethylation is determined by the quality and location of the gRNA. From section **6.4.2.** I observed that gRNA 1 had the greatest demethylation changes therefore I would expect to see the largest changes in expression in gRNA 1 targeted samples.

In cells targeted with gRNA 1 only, with the largest DNA methylation changes, I observed a 1.184 fold increase in *FANCG*, no change in *STOML2,* and 0.869 fold decrease in *PIGO* expression when compared to control cells (**Figure 6.28**). However none of these resulted in statistically significant changes of gene expression due to the small sample size (n=2). In cells targeted with gRNA 2 only I observe no changes in expression of *FANCG, STOML2,* or *PIGO* (**Figure 6.29**). This could be due to there being very limited changes in DNA methylation at the targeted CpGs shown in section **6.4.2.** . Using a combination approach of both gRNA 1 and 2 also resulted in no changes in gene expression of *FANCG, STOML2,* or *PIGO* (**Figure 6.30**). Again this is probably due to the modest changes in DNA methylation observed. The gene expression changes are summarised in **Table 6.21** along with p-values from Welch's two sample t-test.

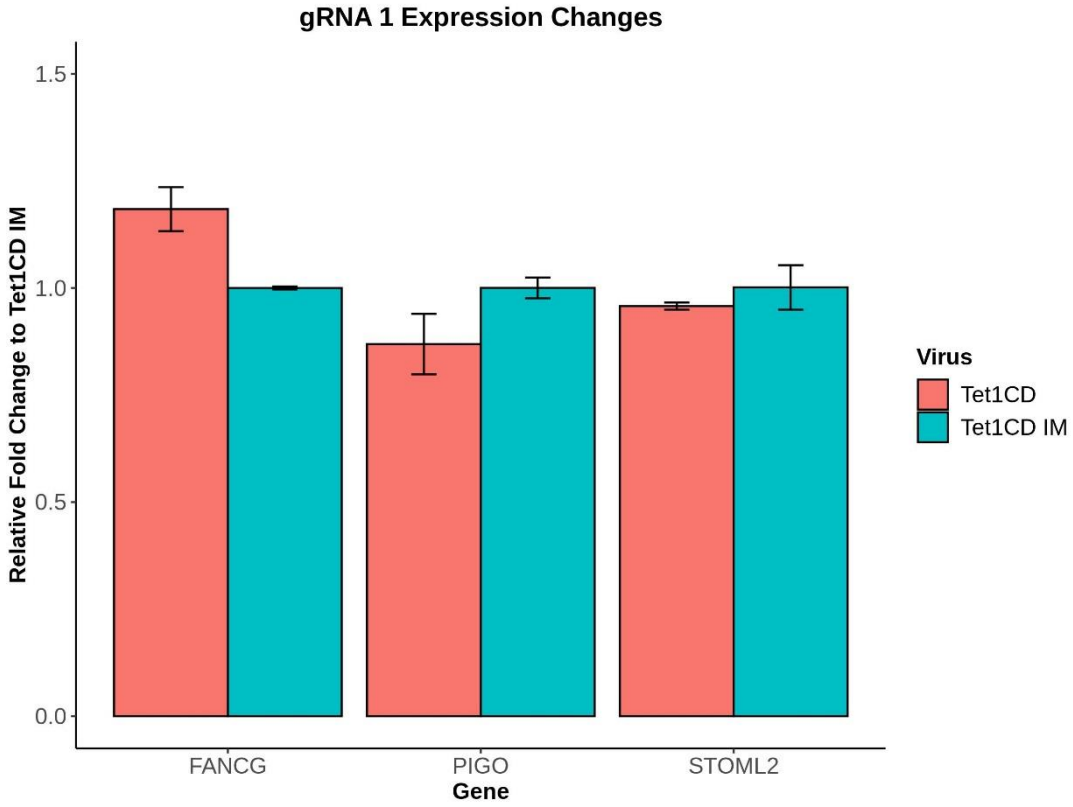


Figure 6.26: FANCG expression is slightly increased following demethylation with Tet1CD using gRNA 1 only. *There is no change in PIGO or STOML2. n=2, error bars are +/- SE.*

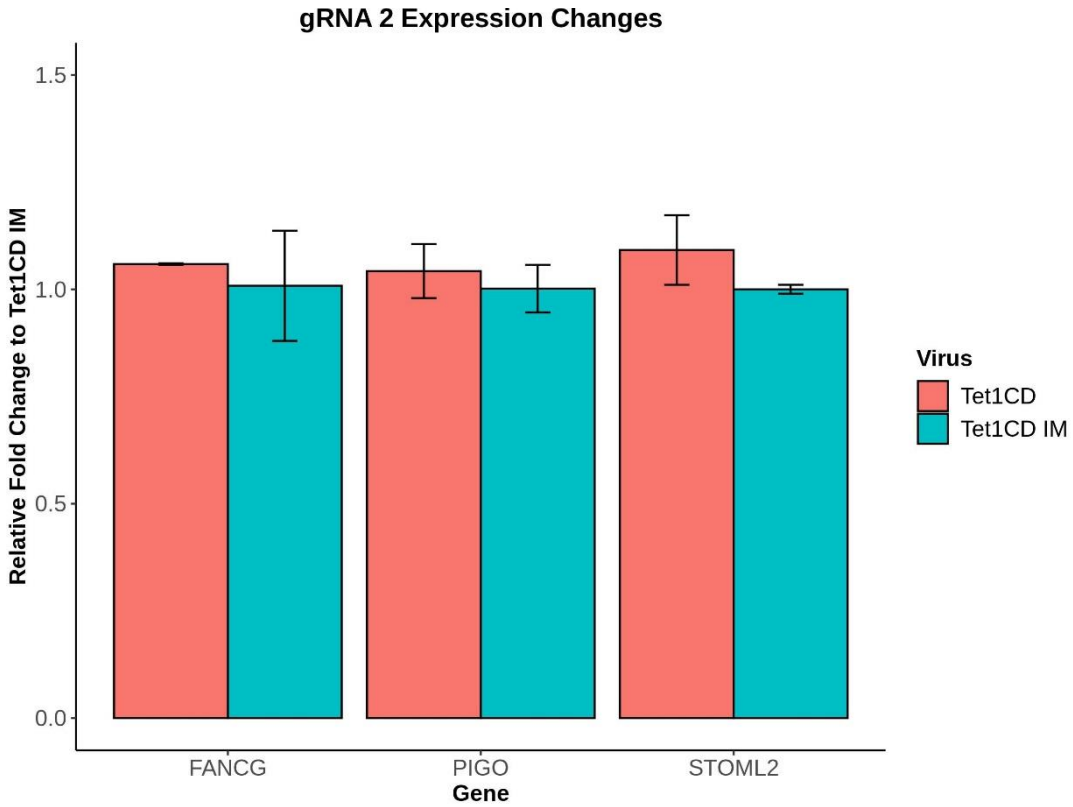


Figure 6.27: There is no change in expression in FANCG, PIGO, or STOML2 following targeting using gRNA 2 only. *n=2, error bars are +/- SE.*

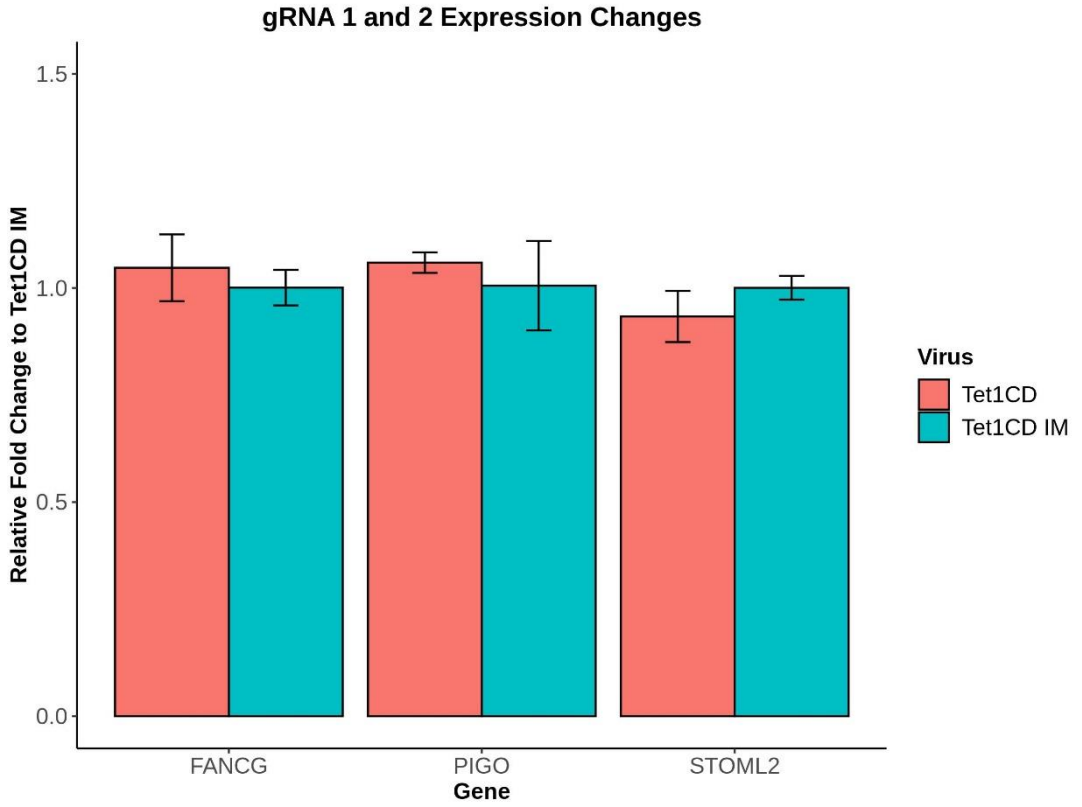


Figure 6.28: There is no change in gene expression of FANCG, PIGO, or STOML2 following targeting with both gRNA 1 and 2. *n=2, error bars= +/- SE.*

Table 6.21: Relative fold changes of FANCG, PIGO, and STOML2 in Tet1CD transduced cells. p-values are from a Welch's two sample t-test.

gRNA	Gene	Relative Fold Change to Tet1CD IM	p-value
1	*FANCG*	1.184	0.171
	PIGO	0.869	0.292
	STOML2	0.958	0.553
2	*FANCG*	1.059	0.761
	PIGO	1.043	0.674
	STOML2	1.092	0.460
1 and 2	*FANCG*	1.047	0.666
	PIGO	1.059	0.697
	STOML2	0.933	0.452

6.4.5. Methylation is increased at the *SNCA* locus using DNMT3a

The final objective of this chapter was to identify if I can increase the methylation of specific hypomethylated CpGs using a dCas9 fused to DNMT3a. To do this we targeted a DMR that was hypomethylated in SH-SY5Y cells within the *5'* UTR of *SNCA* which contains 15 CpG sites, of which three have been shown to be associated with gene expression changes in *SCNA* (Hannon *et al.*, 2018). To increase the DNA methylation at these CpG sites I designed three gRNA flanking the two of the three CpGs and generated lentiviruses containing these gRNAs and dCas9 fused to DNTM3a. I then transduced SH-SY5Y cells with both DNMT3a and pooled gRNA viruses. After seven days I FACS double positive mCherry and BFP cells and extracted DNA. The DNA was then bisulphite converted and pyrosequenced. I hypothesised that I would see an increase in the DNA methylation state of the targeted CpGs by DNMT3a. Unfortunately I was unable to harvest any RNA from the DNMT3a transduced cells to examine gene expression changes.

I collected DNA from two biological repeats of each DNMT3a and DNMT3a IM transduced SH-SY5Y cells. However following extraction and pyrosequencing the data from one DNMT3a sample was unreliable due to insufficient material. The average change of DNA methylation over the targeted sites is summarised in **Table 6.22**. The pyro assay designed covered nine CpGs within the 5' UTR of *SNCA* but we could not design an assay which included any of the three CpG sites identified by (Hannon et al 2018). From the pyrosequencing assay we can see there is an increase in DNA methylation over the nine CpG sites when transduced with active DNMT3a compared to inactive DNMT3a (**Figure 6.10**). The largest increase of DNA methylation was at CpG 4:90757378 where an 8% increase in methylation is observed. The smallest change in DNA methylation is at CpG 4:90757423 where only a 1% methylation increase is observed. There is an average methylation change across all nine CpG sites of +4.67%. I have therefore developed a method to successfully increase DNA methylation at specific CpG sites *in vitro* using a modified form of dCas9 fused to DNMT3a.

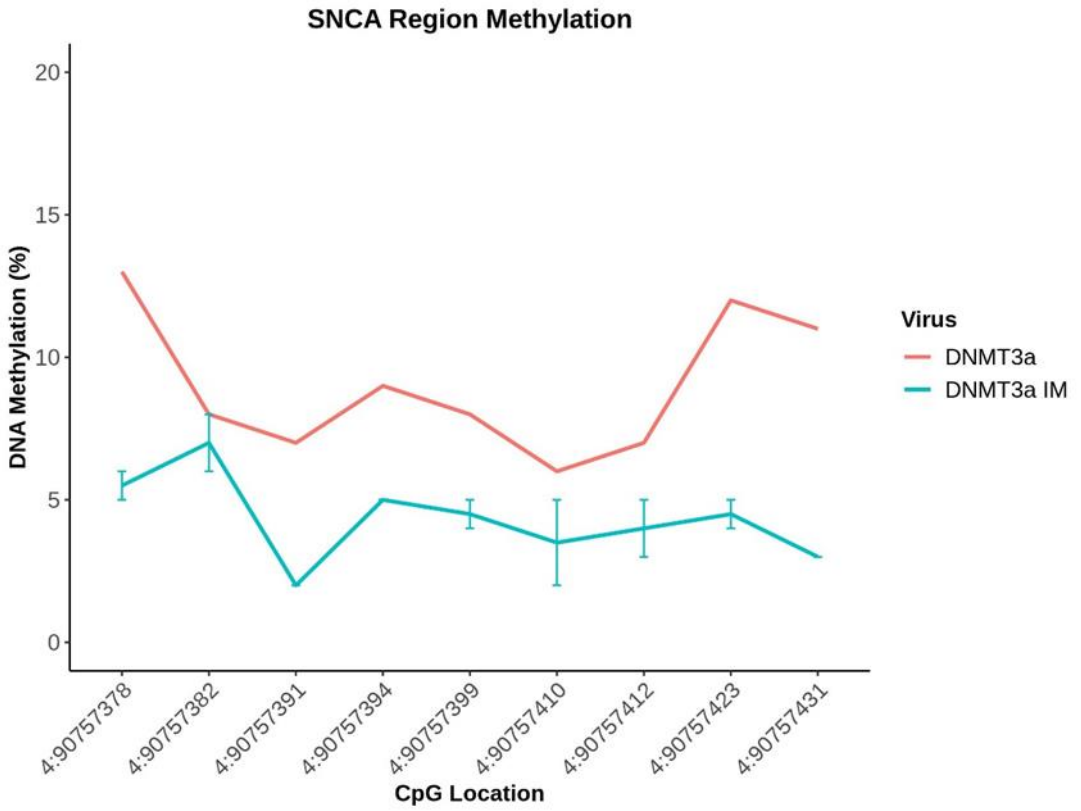


Figure 6.29: Modest increases in DNA methylation are observed at each of the CpGs within the SNCA 5' UTR targeted with a panel of gRNA and DNMT3a in SH-SY5Y cells*. N=2 for DNMT3a IM and N=1 for DNMT3a. Error bars are +/- the SEM.*

Table 6.22: A summary of the average DNA methylation change at each CpG assayed within the SNCA region.

CpG Site	Average DNA Methylation Change (%)
4:90757431	+7.5
4:90757423	+1
4:90757412	+5
4:90757410	+4
4:90757399	+3.5
4:90757394	+2.5
4:90757391	+3
4:90757382	+7.5
4:90757378	+8

6.5. Discussion

In this chapter I have developed a pipeline for editing DNA methylation *in vitro* to examine how DNA methylation changes can impact on cellular function in a model system. This methodology can now be used to functionally validate and characterise DNA methylation changes in response to EWAS studies. The simplicity of the design and delivery makes this method incredibly easy and translatable to a number of different diseases and conditions. The use of lentiviral delivery means that any cell type can be transduced as you remove the limitation of having a transfectable cell line. I have also developed two methods, one to increase and one to decrease DNA methylation. I therefore have control over the direction of effect of DNA methylation. This will be critical at sites of hemimethylation where we are unsure whether decrease or increase DNA methylation will have the greatest effect.

I first set out to decrease DNA methylation within the *FANCG* promoter where a CpG has been associated with gene expression of *STOML2* and *PIGO* through SMR analysis (Hannon *et al.*, 2018). Furthermore, there are four CpGs within the *FANCG* promoter thought to be under control of the same mQTL and it is unknown if the genetic component or methylation is responsible for gene expression changes. The aim was to examine if DNA methylation is driving these gene expression changes rather than the underlying genetic component. To do this, I designed three gRNA to surround these four CpGs and used two to edit the DNA methylation in SH-SY5Y cells using dCas9 fused to Tet1CD. Demethylation was selected as the baseline methylation of these four CpGs in SH-SY5Y cells is 73.2%. Following transduction with Tet1CD and gRNA 1, gRNA 2 or gRNA 1 and gRNA 2 I identified a reduction of DNA methylation over the five CpG sites within the region in a guide dependent manner. The largest decrease in DNA methylation was in samples transduced with gRNA 1 only (average -7.5%) compared to gRNA 2 only (average -2.4%). When the two gRNA were used in conjunction the average methylation change of the five CpG was -4.2% showing that inclusion of both gRNA is near average of the single gRNA transductions. The largest decrease in DNA methylation was observed in samples transduced with Tet1CD and gRNA 1 only where a -16.5% decrease is seen in the CpG furthest away from the PAM sequence of the gRNA. This indicates that there could be spatial limitations between where the PAM is located and the site of action by Tet1CD. This has previously been observed in a DNMT3a plasmid where the site

of maximal activity was 27bp downstream of the PAM sequence (Vojta *et al.*, 2016). When I examined the gene expression changes in *FANCG, STOML2* and *PIGO* I did not observe any significant changes in gene expression. However in the samples transduced with gRNA 1 only, with the largest DNA methylation decrease, there is a slight increase in *FANCG* expression (1.184 fold increase). It could be that the methylation changes engineered are not enough to exert large changes in gene expression. It could be that the size of the region targeted is also too small and a larger region should be edited to observe expression changes.

I then set out to examine if we can increase DNA methylation within the 5' UTR of the *SNCA* gene where CpG sites have been associated with an increase in *SNCA* expression through SMR (Hannon *et al.*, 2018). Again it is thought that the three CpG sites identified are under the control of the same mQTL, and it is unknown whether DNA methylation or the genetic component is responsible for the change in expression. The aim was to examine if increasing DNA methylation using DNMT3a results in gene expression changes of *SNCA* independent of the underlying genetic component. To do this, I designed three gRNA to target 14 CpG sites within the *SNCA* 5' UTR including one identified by (Hannon *et al.,* 2018) and used these three in conjunction with dCas9 fused to DNMT3a to increase the DNA methylation at these sites. Increased methylation was selected as the baseline DNA methylation in the three CpG sites identified by (Hannon *et al.,* 2018) is 10.94% in SH-SY5Y cells. Following transduction with DNMT3a and a pool of three gRNA lentiviruses I only managed to collect limited number of cells due to reduced transduction efficiency. When the DNA underwent extraction and BC one of the DNMT3a samples yielded insufficient material therefore only one sample was used for further analysis. In this sample I observed an increase in all of the nine CpG sites within the pyrosequencing assay designed, with an average DNA methylation increase of +4.67% compared to the inactive control. Unfortunately I was unable to collect RNA for gene expression analysis due to the limited number of cells transduced.

While the methylation changes I have introduced in both *FANCG* and *SNCA* seem modest, these small changes are what is often observed in our EWAS studies. Therefore these can provide a more translatable method for examining small DNA methylation changes rather than larger changes seen in other diseases such as cancer. These changes are also similar to what have been seen in other experiments

using DNMT3a and Tet1CD where only modest changes are observed (Liu *et al.*, 2016; Vojta *et al.*, 2016). I also observed gRNA specific DNA methylation changes with the *FANCG* experiment, this shows how important guide design and quality is when designing CRISPR experiments. Interestingly when we used more gRNA in the SNCA experiments we see an improvement in the methylation differences. It could be that more gRNAs are required over a given region to maximise the efficiency of methylation changes or these gRNA are of better quality or addition of methyl groups is easier than removal.

It is also worth noting that these experiments were done using FACS sorted cells rather than bulk cells. The advantage of this is that we can select for cells which have been transduced with both the dCas9 and gRNA viruses, however the accuracy of FACS for detecting single cell events has been debated. It could be that I had a number of false positive cells within the sorted population which could be distorting the results, given methylation is shown as a proportion of methylated vs unmethylated at each site. To improve this single cell populations could be sorted and allowed to clonally expand, the DNA methylation profile should be maintained by DNMT1 following cell division and then would allow for a clonal population to be analysed. This approach would further increase the number of samples for analysis from two biological repeats to a large number of single cell analyses.

There is a further limitation regarding the DNMT3a edited samples, where the methylation is only increased very modestly compared to previous reports of up to 50% methylation increases (Liu *et al.*, 2016). This could be as a result of multiple factors. The first being the limited transduction success of virus containing dCas9-DNTM3a (4% transduced with 1ml virus vs 80% for DNMT3a IM). This means that there were fewer cells which were successfully transduced with the DNMT3a machinery resulting in fewer FACS sorted cells and as a result limited DNA was harvested and no RNA was harvested to examine *SNCA* expression changes. It has previously been shown that increasing the size of the packaging component drastically decreases the viral titre received from HEK293T cells (Kumar *et al.*, 2001). As the DNMT3a lentiviral plasmid is approximately 17Kb in size the reduction in transduction efficiency is most probably due to the vector size, which is well outside the optimal packaging capacity of a lentivirus. One way to overcome this would be to reduce the size of the DNMT3a plasmid, this could be done by removing unnecessary sequence

within the DNMT3a enzyme to include only the methyltransferase domain, however further cloning would need to be done. Another way to overcome this is to FACS sort and generate integrated stably expressing cell lines. However, with the vectors used in this chapter this would prove problematic, as the DNMT3a/Tet1CD would be constitutively expressed within the transduced cells which could result in non-specific DNA methylation changes. To overcome this a doxycycline inducible promoter within these vectors could be introduced to allow further control.

Off target effects are greater in methylation editing pipelines due to the design (Lin *et al.*, 2018). Most differentially methylated CpGs occur in intronic regions and regions of high sequence homology, therefore gRNA specificity is increasingly difficult due to the repetitive nature of these regions. Indeed, when (Lin *et al.*, 2018) examined off target effects of DNMT3a in HEK293T cells they found increased DNA methylation at promoters, 5' UTRs, CpG Islands and DNase hypersensitivity sites. However this paper used different vectors to the ones used in this thesis. Further implications for off target effects of epigenetic editors have been shown in CRISPR screening experiments with gRNA designed to CTCF loops, with poorly designed gRNA shown to be large cofounders in these studies (Tycko *et al.*, 2019). The current methods to reduce off-target effects, much with the genomic CRISPR discussed in Chapter 3, is good gRNA design and selection. Further to this, newer gRNA design tools are being released every year to combat these problems with improved algorithms, however currently these are mostly optimised for exonic regions and genomic CRISPR/Cas9.

Finally, as discussed in the introduction, demethylation from 5mC to C is a multistep processes with lots of intermediate steps, which we still do not fully understand. Addition of other cofactors could improve the demethylation efficiency, specific chromatin decompactors or histone modifiers could be included to relax the chromatin to facilitate access of the dCas9 and Tet1CD machinery allowing more efficient editing of CpG sites. Recently, (Taghbalout *et al.*, 2019) examined this theory by developing new methodologies using GADD45A, an enhancer of Tet1CD activity, and NEIL2, a co-factor for the final BER mechanism from 5fC and 5caC to C. They found that using these factors in conjunction with Tet1CD increased demethylation at targeted CpGs in HEK293T cells. Furthermore they also noted increased rates of other modified C markers such as 5hmC. This can open the avenue for development of other epigenetic editors to examine the functional effects of other cytosine modifications.

6.6. Conclusion

I have developed a protocol to edit specific CpG sites *in vitro* using a modified form of CRISPR/Cas9 that can both increase and decrease DNA methylation using either dCas9 fused to DNMT3a or Tet1CD as previously described by (Liu *et al.*, 2016). I used these methods to decrease DNA methylation at CpG sites within the *FANCG* promoter and identified a slight increase in *FANCG* expression following targeted demethylation. I then used these methods to successfully increase the methylation at the *SNCA* 5' UTR but unfortunately could not collect RNA for expression quantification. Moving forward we can use these methods to functionally validate how increases or decreases in DNA methylation identified through EWAS studies relate to cellular transcription levels.

Chapter 7. Examining the effect of Δ^9-tetrahydrocannabinol on the methylome.

7.1. Introduction

As discussed previously there are not only genetic factors which contribute to schizophrenia (**1.1.4**), there are also a number of environmental factors (**1.1.3**). The main environmental factors robustly associated with schizophrenia are maternal stressors during pregnancy such as maternal stress (Khashan *et al.*, 2008), maternal infection (Brown, 2011), and maternal famine (Wang and Zhang, 2017). Equally there are a number of environmental factors that act postnatally which infer increased risk to developing schizophrenia. These are particularly important at adolescence which is a vulnerable stage of brain development. One such risk factor is cannabis exposure and has previously been discussed in **1.1.3.4**.

Cannabis is the most widely used illicit drug with an estimate usage in 192 million people worldwide, which will increase as more countries move towards commercial legalised sales (United Nations Office on Drugs and Crime, 2018). The two key components of cannabis are THC and CBD. THC is the known psychoactive competent of cannabis and has been shown to induce psychotic symptoms and impaired memory in non-psychiatric individuals in a dose dependent manner (D'Souza *et al.*, 2004; Murray *et al.*, 2007). CBD however has not been shown to be psychoactive and does not induce hallucinations or delusions but has been proposed to counteract the psychoactive effects of THC (D'Souza *et al.*, 2004; Waldo Zuardi *et al.*, 2012). However the exact mechanisms of the THC and CBD interaction are still unknown, as some studies show CBD counteracting the effect of THC whereas others show CBD potentiate the action of THC (Boggs *et al.*, 2018). The causal relationship between cannabis exposure and changes to brain morphology have been previously shown (for review see (Panlilio and Justinova, 2018)). In a three year longitudinal study, grey matter reductions were observed in the hippocampus, amygdala, and superior temporal gyrus of heavy cannabis users compared to healthy controls (Koenders *et al.*, 2016). Interestingly, even only one or two doses of cannabis can exert changes in grey matter volume in adolescents (Orr *et al.*, 2019). The age of first cannabis usage also impacts on grey matter volume, with subjects' who's first usage before the age of 17 show reduced grey mater volume compared to those after 17 (Wilson *et al.*, 2000). It is therefore no surprise that cannabis exposure has been linked to schizophrenia and other psychosis.

The first longitudinal study into cannabis exposure and schizophrenia risk was carried out in Sweden in the 1980's. This study identified a six fold relative risk of developing schizophrenia with frequent use of cannabis (defined as patients with over 50 exposures to cannabis) (Andreasson *et al.*, 1987). Since then other studies have been published examining a causal link between cannabis usage and schizophrenia. A meta-analysis of cannabis use and psychosis from a multitude of different studies found the overall risk of developing schizophrenia was two-fold higher with cannabis usage (Arseneault *et al.*, 2004). More recently, a larger meta-analysis over multiple centres examined the effect of THC content in cannabis on psychosis, this study showed a three-fold increase in psychosis with low potency cannabis (<10% THC content) use but this rose to five-fold with high potency cannabis (>10% THC) use (Di Forti *et al.*, 2019). This study is of importance as the potency of cannabis has been increasing over the last decade with an increase in THC content compared to CBD (Freeman *et al.*, 2019). Suggesting that there could be an increase in psychosis related to excess THC exposure. While cannabis exposure is generally thought to be an entirely environmental risk, it has been shown that there is a genetic link between cannabis use and schizophrenia. In a recent GWAS on cannabis use it was shown that the genetic signatures for cannabis use overlaps with the genetic regions associated with schizophrenia from the latest GWAS (Pasman *et al.*, 2018). These observations suggest that schizophrenia patients are genetically predisposed to taking cannabis, whether this is causal or as a result of the disease is still unknown. There is further evidence to suggest that drug and environmental exposure effects arise through changes in DNA methylation and this could be how THC exerts its psychoactive effect (Meehan *et al.*, 2018).

There have been a number of review papers that suggest cannabis may have an effect on the methylome (Szutorisz and Hurd, 2016, 2018). Further to this, there have been limited studies into the link between cannabis exposure and DNA methylation in different tissues and model systems each with various levels of findings (**Table 7.1**). The first study undertaken examined the effect of THC exposure on the DNA methylation status of the orexin promoter, a neuropeptide involved in drug craving, in DNA derived from human blood samples and found no significant differences in DNA methylation between THC dependent subjects, nicotine only smokers, and non-smoker controls (Rotter *et al.*, 2012). Another study in mouse myeloid-derived

suppressor cells found increased methylation at *DNMT3a* and *DNMT3b* in THC exposed mice, however no global changes in DNA methylation were observed (Sido *et al.*, 2015). These studies could have been limited by their choice of cells, responses in blood may be reduced due to low expression of cannabinoid receptors (The Genotype Tissue Expression Consortium, 2019a). However, when DNA methylation has been examined in iPSC cells differentiated into neuronal precursors, a more relevant cell type than blood, no methylation changes in *CB_1R* were observed following exposure to 10μM THC (Stanslowsky *et al.*, 2017). Some of the most interesting findings from the literature suggesting THC affects the epigenome comes from studies investigating transgenerational effects of THC exposure in rodents. The first genome-wide experiment examining THC effects was focussed on the nuclear accumbens of rats born to parents who were exposed to THC throughout their lives. This study identified over 1000 DMRs present in rats born to THC exposed parents versus vehicle controls (Watson *et al.*, 2015). This study implies that there are molecular signatures of cannabis exposure and, in particular, how THC can alter the methylome in neuronal tissue.

To date, studies examining the effect of THC on the methylome have been mainly targeted approaches investigating candidate genes with limited studies examining the genome-wide effect on neuronal cell lines. Because of the importance of cannabis usage globally, the increase in THC content, and the increased risk of developing schizophrenia when exposed to THC, it is important to understand what effect, if any, THC has on the neuronal methylome.

Table 7.1: Summary of published literature examining THC exposure and DNA methylation.

Reference	Tissue	Samples	Method	Targeted/ Genome-wide	Key Finding
(Rotter *et al.*, 2012)	Peripheral blood lymphocytes	36 THC dependency sufferers, 20 cigarette smoker controls, 21 never smoker controls	Digestion of DNA with HpaII and MspI then qPCR of Orexin promoter.	Targeted to *HRCT*	Changes in Orexin expression observed however no change in methylation state of promoter between groups
(Sido *et al.*, 2015)	Myeloid-derived suppressor cells	Five THC exposed female BL6 mice, five control female BL6 mice	Bisulphite conversion of DNA then PCR	Targeted to *DNMT3a* and *DNMT3b*	Methylation increased at *DNMT3a* and *DNMT3b* however no global changes in DNA methylation observed
(Watson *et al.*, 2015)	Nuclear Accumbens	32 F1 adult rats following parental life exposure to THC or Vehicle	Enhanced Reduced Representation Bisulphite Sequencing	Genome-wide	Identified 406 hypermethylated DMRs and 621 hypomethylated DMRs
(Stanslowsky *et al.*, 2017)	Small Molecule Neuronal Precursor Cells	Five different experiments from small molecule neuronal precursor cells derived from human core blood-derived iPSCs	Bisulphite conversion of DNA then PCR	Targeted to *CB_1R*, *MAP2* and *DAT*	No significant differences in DNA methylation

(Murphy *et al.*, 2018)	Human Sperm	24 males, 12 weekly cannabis users for at least 6 months, 12 matched controls	Reduced Representation Bisulphite sequencing	Genome-wide	6640 CpGs with altered methylation, Hippo signalling and cancer pathways enriched
	Rat Sperm	17 9 week old rats, 9 exposed to THC for 12 days, 8 exposed to vehicle			Hippo signalling and cancer pathways enriched
(Gerra *et al.*, 2018)	Human Whole Blood	40 cannabis users, 96 control subjects	Methylated DNA immunoprecipitation PCR	Targeted to *ANNK1, CNR1, DRD2, NCAM1*	Hypermethylation at *DRD2* and *NCAM1*
(Schrott *et al.*, 2019)	Human Sperm	24 males, 12 weekly cannabis users for at least 6 months, 12 matched controls	Targeted Bisulphite Pyrosequencing	Targeted to *DLGAP2*	Hypomethylation observed in THC exposed samples
	Rat Sperm, Hippocampus, Nuclear Accumbens	17 9 week old rats, 9 exposed to THC for 12 days, 8 exposed to vehicle			Hypomethylation observed in sperm samples but not in hippocampus or nuclear accumbens.

7.2. Aims

There have been no genome-wide DNA methylation profiling studies investigating the effect of direct THC exposure to neuronal cells *in vitro*. My aim in this chapter was to identify whether or not exposure to THC results in altered DNA methylation patterns in neuronal cells and to identify which cellular pathways are altered, if any. My subsequent aim was to elucidate if there are different responses depending on duration of THC exposure. In order to do this:

1. SH-SY5Y cells were exposed to 5μM of THC or vehicle control for either one day, or seven days. All cells were harvested after seven days in culture, DNA was extracted and genome-wide DNA methylation were measured using the Infinium HumanMethylationEPIC BeadChip Array.
2. Changes in DNA methylation were identified following continuous or singular exposure to THC.
3. Pathway analysis was performed and I examined any overlap between the two treatment duration datasets.

7.3. Methods

7.3.1. THC dosage selection

To determine the minimal dosage of THC to elicit a response in SH-SY5Y cells a literature search was conducted on Pubmed with the search terms “SH-SY5Y AND tetrahydrocannabinol”. A selection of the published literature at the time is shown in **Table 7.2**. Based on the current literature, I selected a dosage of 5µM THC, this dose was selected to prevent decreased viability of the SH-SY5Y cells over the seven day treatment.

Table 7.2: Published literature on SH-SY5Y cells exposed to THC.

Dosage THC	Exposure Time	Toxic Dose	Paper
10^{-5}M, 10^{-7}M, 10^{-9}M,	48 hours	-	(Lew, 1996)
1µM, 5µM, 10µM	72 hours	Dosage at 10µM resulted in 25% viability	(Klegeris, Bissonnette and Mcgeer, 2003)
50nM, 100nM, 150nM, 200nM	24 hours	No toxic dose	(Cannarsa *et al.*, 2012)
0.1µM, 1µM, 5µM, 10µM	24 hours	-	(Carroll *et al.*, 2012)
10µM	48 hours	-	(Zeissler *et al.*, 2016)

7.3.2. Cell culture

SH-SY5Y cells were maintained as described in **2.1.2.** . The day prior to dosage with THC, $2x10^5$ cells were seeded in six well plates with growth media (DMEM/F-12, with Glutamax, 10% FBS) as described in **2.1.** . THC solution was acquired in ethanol solution from Sigma (Sigma-Aldrich, Missouri, USA). THC solution was diluted down to a working solution of 5µM in growth media (DMEM/F-12, with Glutamax, 10% FBS) prior to adding to cells. A vehicle control of 100% ethanol replacing THC was also created. Media was removed from six well plates and replaced with 3mls of either THC containing media or ethanol vehicle control. THC media was replaced every 48 hours until the end of the experiment. Cells were split at 48 hours into fresh 6 well plates. Cells were harvested at seven days post first dosage with THC as described in **2.1.3.** . Each condition had six biological repeats. An overview of the experiment is shown in **Figure 7.1**.

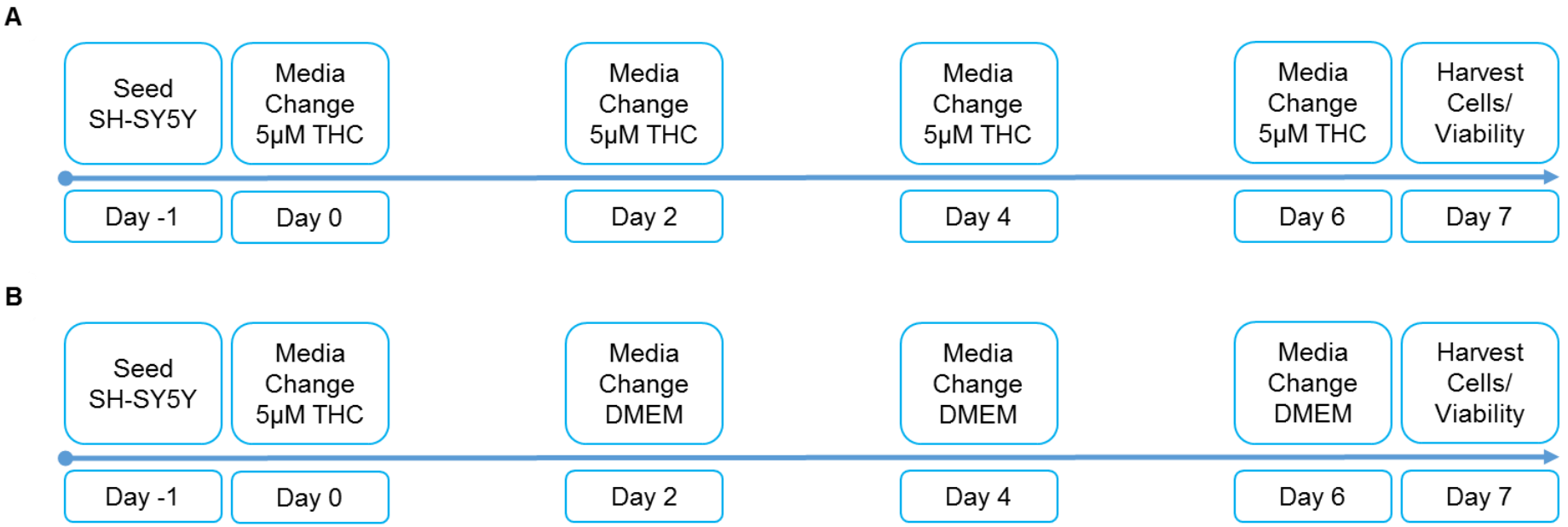


Figure 7.1: Experimental overview for SH-SY5Y exposure to THC. *SH-SY5Y cells were seeded in six well plates at day -1. Two pipelines were set up, either a continuous exposure to THC over seven days with fresh THC or ethanol vehicle control media changes every 48 hours (a) or single exposure to THC or ethanol vehicle control media at day zero with media replacements with normal DMEM/F-12 with Glutamax 10% FBS every 48 hours (b). At 7 days cells were harvested and a viability assay undertaken. Each condition had six biological replicates.*

7.3.3. Viability assay

In order to assess whether THC dosage was affecting cell viability a viability assay was undertaken during cell harvesting. After disassociation with accutase 10µl of resuspended cells were diluted in trypan blue 0.4% solution in an Eppendorf (Invitrogen, Carlsbad, California, USA). The number of unstained (alive) cells were counted using a haemocytometer alongside the positively stained (dead) cells to calculate the percentage viability. Percentage viability was calculated by:

$$Percentage\ Viability = \frac{Number\ of\ Unstained\ Cells}{Total\ Number\ of\ Cells}$$

7.3.4. DNA extractions

Once all samples were collected DNA was extracted from the cell pellets as described in **2.3.** . Samples were randomised during extraction as to reduce the batch effect. DNA was resuspended in 25µl of TE.

7.3.5. DNA bisulphite conversion

500ng of DNA underwent BC as outlined in **2.7.1.** .

7.3.6. DNA methylation profiling

Whole DNA methylation profiling was undertaken by Dr Joe Burrage using the Infinium HumanMethylationEPIC BeadChip Array as described in **2.7.2.** .

7.3.7. Data analysis

Data from the Infinium HumanMethylationEPIC BeadChip Array underwent stringent QC using our developed in house pipeline to generate β values for each probe representing level of DNA methylation where 0 is fully unmethylated and 1 is fully methylated, this is described in **2.7.2.** . The QC and normalising processes were from the WateRmelon package developed by our lab (Pidsley *et al.*, 2013). For QC first of all, the median signal intensities are calculated for both the methylated and unmethylated signals for each sample across all sites. Any samples with a median intensity of less than 2000 units is removed from further analysis, no samples had intensities less than 2000. The efficiency of bisulphite conversion is then calculated using data from control probes present on the array and any sample with a conversion less than 80% is excluded, all samples had a conversion rate >86%. A pfilter is then

applied to remove any probes which do not reach a beadcount or p-value threshold. For beadcounts, if the percentage of samples with a beadcount less than three is greater than 5% for any probe the probe is then excluded. If the percentage of probes with a detection p-value > 0.05 is greater than 1% for any samples the sample is removed. If the percentage of samples with a detection p-value > 0.05 is greater than 1% for any probe then the probe is removed. All samples passed this pfiltering step. Samples are then normalised using the internal normalisation function *dasen* which normalises for background differences between type I and type II probes. Following normalisation known cross-hybridizing and SNP probes are removed from the data before performing analysis.

To detect changes in DNA methylation a linear regression model was used across every CpG site with the following formula:

$$Methylation\ Change = lm(dasen\ normalised\ \beta \sim Condition)$$

Where β represents the *dasen* normalised beta values for each probe and condition represents 5µM THC or matched ethanol vehicle. A quantile-quantile (QQ) plots for both the single dose and continuous dosages were assessed for p-value inflation (**Figure 7.2** and **Figure 7.3**). All p-values were corrected for multiple testing using Bonferroni correction. For GO analysis using missMethyl a significance value of $5x10^{-3}$ was selected as this provided enough probes to run the pipeline (Phipson and Maksimovic, 2016). All analysis was performed in R.

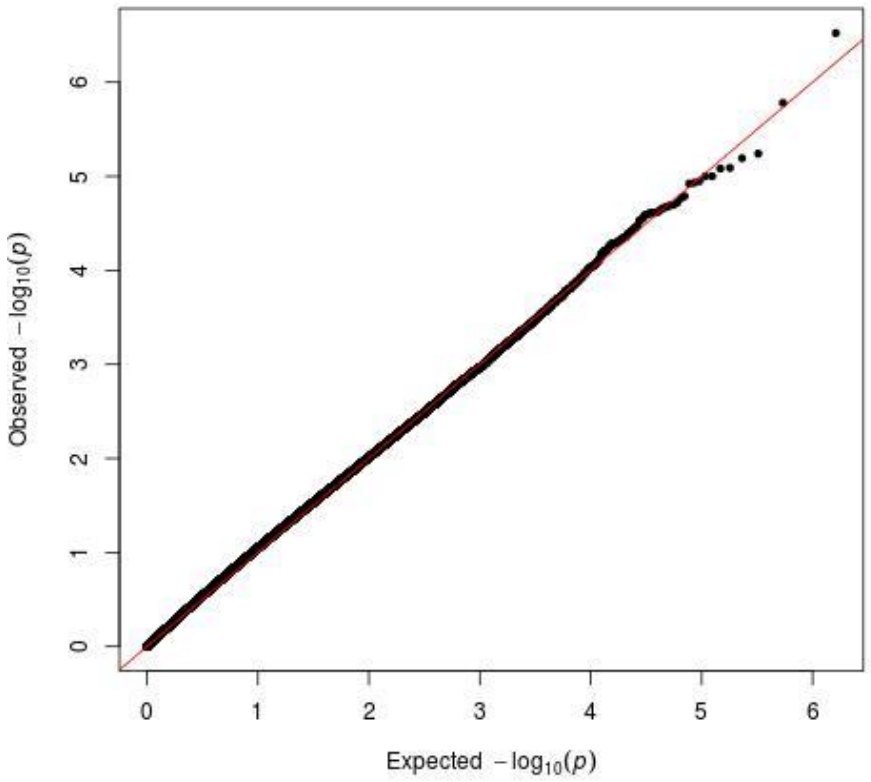


Figure 7.2: Quantile-quantile plot of expected versus observed p-value to check for inflation in the linear regression analysis for the single dose of THC EWAS. *Shown are the expected (x-axis) and observed (y-axis) p-values following linear regression analysis. There is no inflation within this analysis as the p-values are correlated.*

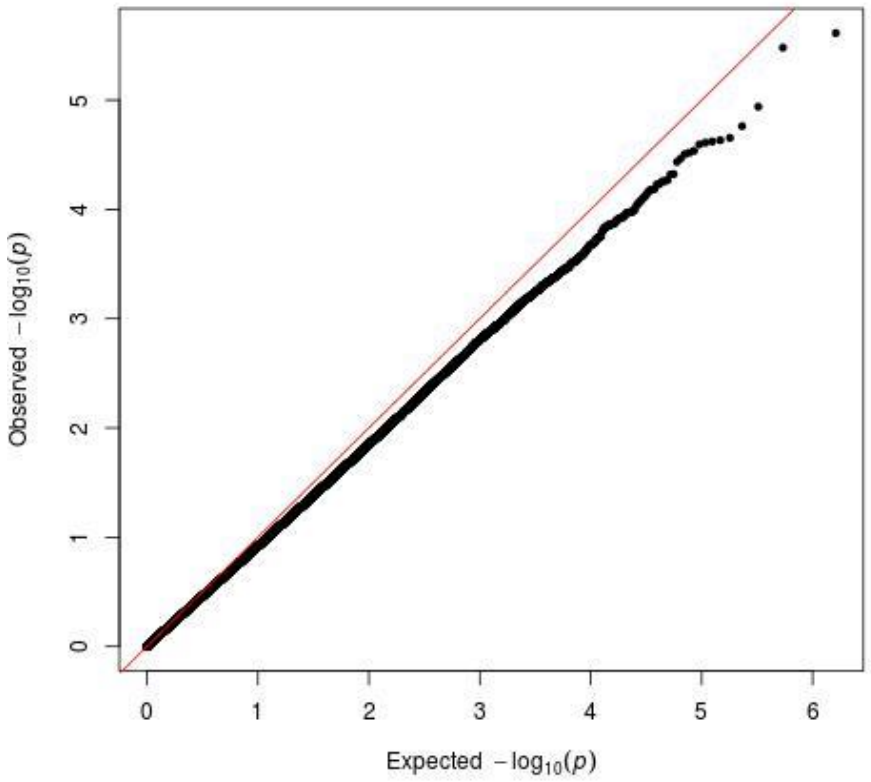


Figure 7.3: Quantile-quantile plot of expected versus observed p-value to check for inflation in the linear regression analysis for the continuous dose of THC EWAS. *Shown are the expected (x-axis) and observed (y-axis) p-values following linear regression analysis. There is no inflation within this analysis as the p-values are correlated.*

7.4. Results

7.4.1. 5µM dosage of THC does not affect SH-SY5Y cell viability

The dose selection of 5µM THC was selected as this dosage was not found to be toxic to the SH-SY5Y cells. In order to confirm this, I performed a viability assay while harvesting cells for DNA at day seven following first addition of 5µM THC or ethanol vehicle control. From our results I confirmed that 5µM of THC has no effect on the viability of the SH-SY5Y cells following either single or continuous exposure over the seven days, one-way ANOVA ($F(3,20)=0.25$, p-value=0.86) **Figure 7.4**.

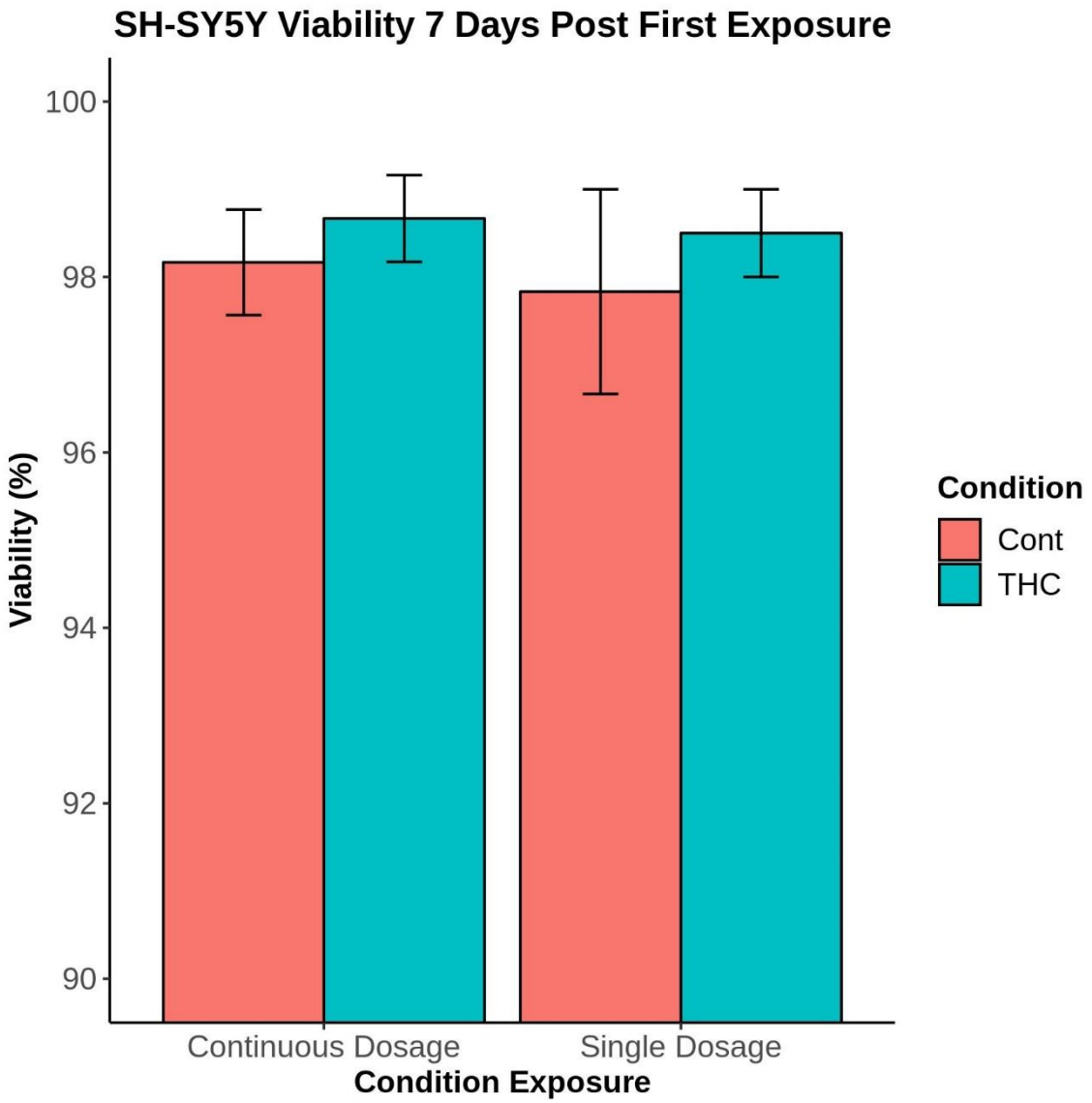


Figure 7.4: Continuous or single exposure to 5µM THC (THC) or ethanol vehicle control (Cont) had no effect on SH-SY5Y cell viability after 7 days. *N=6 per condition and dosage. One-way ANOVA (F(3,20)=0.25, p-value=0.86).*

7.4.2. Continuous exposure to THC results in nominal DNA methylation changes

I set out to identify if long term exposure to THC results in any genome-wide changes to DNA methylation. To date, this is the first study undertaken to examine genome-wide methylation changes in a neuronal cell line exposed to THC. In order to achieve this, I exposed SH-SY5Y cells to 5µM THC over seven days, with fresh THC media given every 48 hours. At seven days DNA was harvested, BC, and run on the Infinium HumanMethylationEPIC Array to profile genome-wide DNA methylation changes.

Following QC and analysis no probes reached genome-wide significance following correction for multiple testing. This is probably due to the small size of the dataset that was underpowered to detect small changes in DNA methylation. I therefore used a nominally significant significance threshold to include any probe with an uncorrected p-value less than or equal to $5x10^{-5}$. When this threshold is selected I identified 16 differentially methylated probes (DMP) associated with THC exposure compared to ethanol vehicle control, these probes are listed in **Table 7.3**. Of these 16 probes, six showed an increase in DNA methylation and 10 showed decreased DNA methylation following THC exposure. The probes with largest increase and decrease in DNA methylation are shown in **Figure 7.5** and **Figure 7.6** respectively.

Table 7.3: Top ranked THC-associated differentially methylated probes (DMPs) identified following seven days of continuous 5µm THC treatment.

Probe ID	Methylation Difference (%)	SE	P	Chr	Genomic Location	Gene Annotation	Gene Region
cg13841476	-2.80	0.29	2.44E-06	22	19036432	DGCR11;DGCR2	TSS1500;Body
cg11981204	-1.98	0.21	3.31E-06	4	7423056	SORCS2	Body
cg26667821	1.54	0.19	1.15E-05	1	155099774	EFNA1	TSS1500
cg23194532	-5.46	0.71	1.73E-05	3	38192265	-	-
cg09618674	-5.56	0.75	2.22E-05	2	142629557	LRP1B	Body
cg15120122	6.13	0.83	2.33E-05	19	17203122	MYO9B	5'UTR
cg21635789	-1.80	0.24	2.39E-05	12	11639516	-	-
cg07563873	-1.20	0.16	2.45E-05	12	51818501	SLC4A8	TSS200
cg15149294	-2.42	0.33	2.54E-05	20	45960088	ZMYND8	5'UTR;Body
cg05093728	-2.97	0.41	2.92E-05	2	129069094	HS6ST1	Body
cg13120954	-3.44	0.48	3.04E-05	6	153473676	-	-
cg05020081	3.43	0.48	3.12E-05	14	102263662	PPP2R5C	Body
cg06387943	3.78	0.54	3.43E-05	12	50306291	LOC283332;LOC101927292	TSS1500;Body
cg12821675	3.56	0.51	3.68E-05	17	30117392	-	-
cg15763628	2.98	0.44	4.75E-05	10	641167	DIP2C	Body
cg02462962	-2.25	0.33	4.80E-05	5	154320236	MRPL22	TSS1500

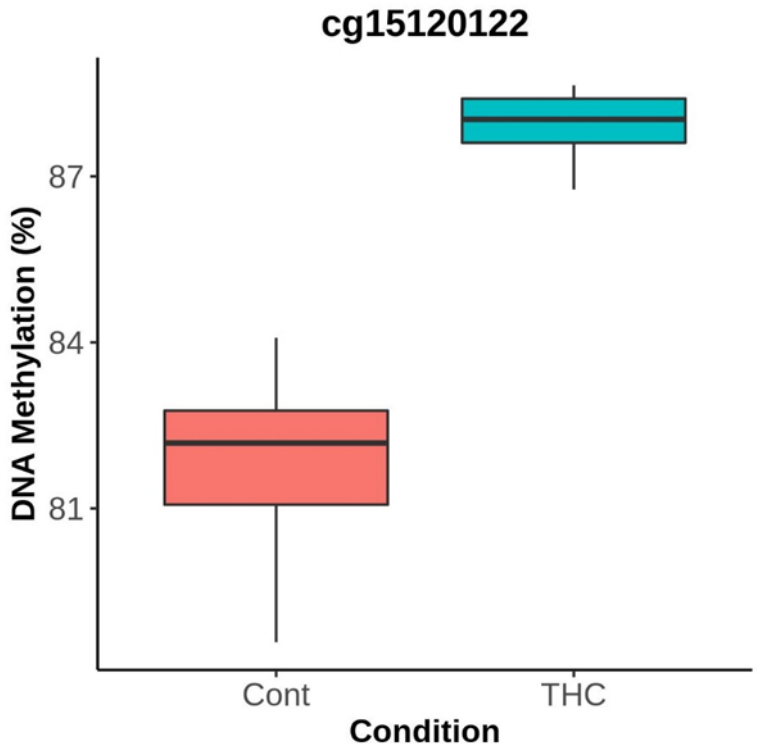


Figure 7.5: cg15120122 shows the largest increase in DNA methylation following continuous THC exposure. *P=2.33E-05, n=6.*

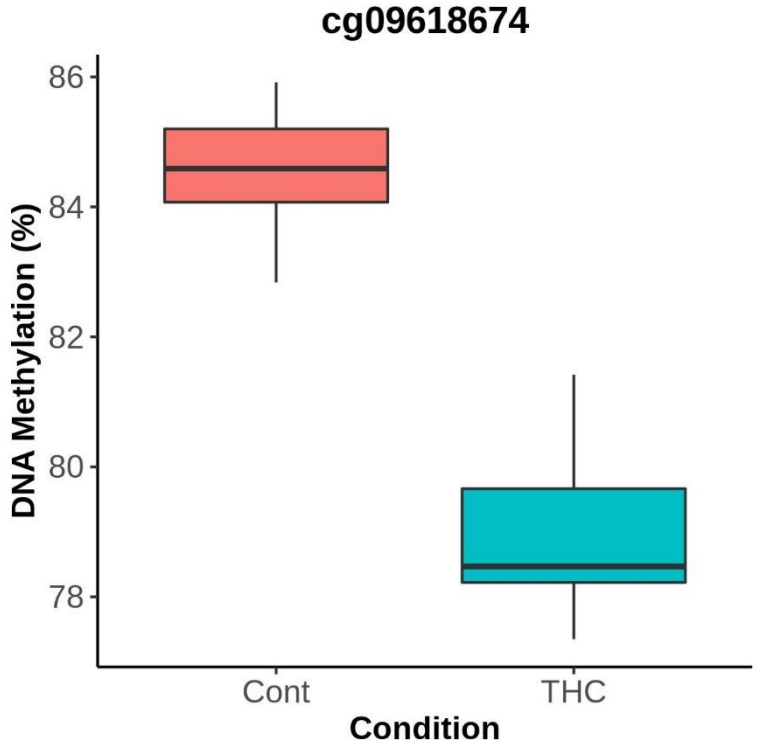


Figure 7.6: cg09618674 shows the largest decrease in DNA methylation following continuous THC exposure. *P=2.22E-05, n=6.*

7.4.3. GO Analysis identified differential methylation in pathways involved in neuronal development are affected following continuous THC exposure.

As probes only provide information related to single CpG sites I wanted to identify if there are any biological pathways altered following THC exposure over seven days. To do this I used the missMethyl package to identify GO terms associated with some of the top DMPs of continuous cannabis exposure (Phipson and Maksimovic, 2016). Because GO analysis requires several thousand probes to provide accurate results I relaxed the genome wide significance threshold used to call DMP to allow more probes to be included in GO analysis. I selected a p-value threshold of $5x10^{-3}$ which included 2730 DMP for GO analysis.

As with DMP calling, due to the small sample size, none of the GO terms reach a stringent multiple correction testing significance (0.05< FDR). The top 20 GO terms are listed in **Table 7.4**. These probes show an enrichment in pathways involved in neurogenesis, neuron development, axon outgrowth, ion channels, and cell adhesion. These observations indicate that continuous exposure of neuronal cells to THC impact genes involved with neuronal development, however as none of these GO terms reached FDR <0.05 this conclusion should be treated with caution.

Table 7.4: Top 20 GO terms associated with seven days of continuous 5μM THC treatment.

Category	Term	Ontology	Total Number of Probes in Category	Number DMP in Category	P	FDR
GO:0022008	neurogenesis	BP	1552	235	0.000598	1
GO:0048699	generation of neurons	BP	1456	221	0.000943	1
GO:0001933	negative regulation of protein phosphorylation	BP	407	66	0.00106	1
GO:0031102	neuron projection regeneration	BP	54	15	0.00154	1
GO:0033686	positive regulation of luteinizing hormone secretion	BP	4	3	0.00159	1
GO:0031115	negative regulation of microtubule polymerization	BP	12	6	0.00164	1
GO:0010632	regulation of epithelial cell migration	BP	284	45	0.00165	1
GO:0055088	lipid homeostasis	BP	144	24	0.00180	1
GO:0048008	platelet-derived growth factor receptor signaling pathway	BP	56	16	0.00185	1
GO:0031103	axon regeneration	BP	45	13	0.00204	1
GO:0051270	regulation of cellular component movement	BP	1044	144	0.00206	1
GO:0015119	hexose phosphate transmembrane transporter activity	MF	3	3	0.00226	1
GO:0015315	organophosphate:inorganic phosphate antiporter activity	MF	3	3	0.00226	1
GO:0015526	hexose-phosphate:inorganic phosphate antiporter activity	MF	3	3	0.00226	1
GO:0061513	glucose 6-phosphate:inorganic phosphate antiporter activity	MF	3	3	0.00226	1
GO:0051893	regulation of focal adhesion assembly	BP	60	18	0.00231	1
GO:0090109	regulation of cell-substrate junction assembly	BP	60	18	0.00231	1
GO:0034594	phosphatidylinositol trisphosphate phosphatase activity	MF	6	4	0.00243	1
GO:0042326	negative regulation of phosphorylation	BP	446	69	0.00252	1
GO:0007399	nervous system development	BP	2281	319	0.00266	1

7.4.4. Single exposure to THC resulted in DNA methylation changes

I set out to identify if a single exposure to THC results in DNA methylation changes following a recovery period in a neuronal cell-line. To achieve this, SH-SY5Y cells were exposed to 5µM of THC for 24 hours before having the THC media removed and a recovery period of six days with normal growth media (**Figure 7.1B**). At day seven, cells were harvested, DNA extracted and BC, then genome-wide DNA methylation profiling was carried out using the Infinium HumanMethylationEPIC Array.

Following QC and analysis, as described earlier, no probes reached genome wide significance following correction for multiple testing. I therefore lowered the threshold to include any probe with an uncorrected p-value less than or equal to $5x10^{-5}$. When this "nominally significant" threshold is selected I identified 46 differentially methylated probes (DMP) associated with a single exposure to THC compared to ethanol vehicle control, these probes are listed in **Table 7.5**. Of these 46 probes, 24 showed an increase in DNA methylation and 22 showed decreased DNA methylation following THC exposure. None of these top probes overlap with the top probes of the continuous dose THC experiment. The probes with the largest increase and decrease in DNA methylation following THC exposure are shown in **Figure 7.7** and **Figure 7.8**.

Table 7.5: Top ranked THC-associated differentially methylated probes (DMPs) identified following a single 5µm THC treatment and recovery period.

Probe ID	Methylation Difference (%)	SE	P	Chr	Genomic Location	Gene Annotation	Gene Region
cg24043594	-2.88	0.24	3.00E-07	17	13494282	HS3ST3A1	Body
cg00555085	-3.75	0.38	1.66E-06	16	87616248	-	-
cg09207718	-3.99	0.46	5.75E-06	15	75041386	CYP1A2	5'UTR
cg05375699	-2.88	0.34	6.43E-06	1	3245451	PRDM16	Body
cg14083651	-3.75	0.45	8.14E-06	17	78313883	RNF213	Body
cg03526920	2.86	0.34	8.29E-06	14	77291706	C14orf166B	TSS1500
cg00669024	3.26	0.40	9.98E-06	13	24318208	MIPEP	Body
cg21336456	3.46	0.42	1.00E-05	2	1.01E+08	CHST10	Body
cg09579787	-3.94	0.49	1.13E-05	10	73358161	CDH23	Body
cg13510420	-3.18	0.40	1.18E-05	16	74930837	WDR59	Body
cg02924237	-5.15	0.65	1.20E-05	10	1.17E+08	ATRNL1	ExonBnd;Body
cg09919577	-1.78	0.23	1.63E-05	12	1.24E+08	DNAH10	ExonBnd;Body
cg20643897	-4.68	0.61	1.72E-05	8	1.43E+08	-	-
cg03892569	2.20	0.29	1.91E-05	6	12942040	PHACTR1	Body
cg14474880	1.55	0.21	2.02E-05	1	93544497	MTF2	TSS1500
cg18497587	2.12	0.28	2.03E-05	19	8212367	FBN3	1stExon;5'UTR
cg04032707	-4.05	0.54	2.11E-05	19	41938946	ATP5SL	Body
cg11075227	1.59	0.21	2.15E-05	23	7066076	HDHD1A	Body
cg00722188	-3.81	0.51	2.21E-05	14	91806251	CCDC88C	Body
cg07805040	-3.11	0.42	2.35E-05	13	77297851	-	-
cg21064632	4.22	0.57	2.40E-05	15	59268225	-	-
cg25624488	-2.74	0.37	2.45E-05	4	1.86E+08	-	-

cg12714395	7.02	0.95	2.45E-05	13	83631684	-	-
cg13882200	-3.42	0.47	2.46E-05	3	9166661	SRGAP3	Body
cg26857965	-4.10	0.56	2.51E-05	15	90787578	CIB1	Body
cg22707857	1.47	0.20	2.57E-05	6	32369453	BTNL2	Body
cg24330729	2.00	0.27	2.60E-05	20	61641993	LOC63930	Body
cg01469661	6.49	0.90	2.81E-05	8	62899765	-	-
cg15946779	3.88	0.54	2.88E-05	2	42395805	EML4;LOC102723824	TSS1500;Body
cg02511333	2.46	0.34	2.97E-05	10	98158479	TLL2	Body
cg18966570	-2.16	0.30	3.33E-05	19	3007682	TLE2	Body
cg08474775	-1.94	0.27	3.51E-05	4	1.55E+08	RNF175	Body
cg16192707	6.75	0.96	3.52E-05	6	99292347	-	-
cg26835683	2.48	0.35	3.69E-05	1	1.6E+08	PEA15	5'UTR
cg08901704	1.62	0.23	3.80E-05	4	8023314	ABLIM2	Body
cg14250737	-2.77	0.40	3.89E-05	14	1.02E+08	-	-
cg02449608	2.96	0.43	4.06E-05	19	58486600	C19orf18	TSS1500
cg21492942	1.57	0.23	4.09E-05	20	44718714	NCOA5	TSS200
cg10311217	3.23	0.47	4.38E-05	6	47171647	-	-
cg15306636	4.11	0.60	4.40E-05	7	1.44E+08	-	-
cg07527625	-2.75	0.40	4.54E-05	16	70808289	VAC14	Body
cg06597593	5.33	0.78	4.59E-05	9	1.32E+08	FAM73B	Body
cg03898766	-2.62	0.38	4.61E-05	12	56590689	-	-
cg26682900	2.77	0.41	4.74E-05	12	1.23E+08	HIP1R	Body
cg00386663	1.87	0.28	4.87E-05	16	692659	FAM195A	Body
cg13419032	-3.22	0.47	4.91E-05	1	1.11E+08	SLC6A17	5'UTR

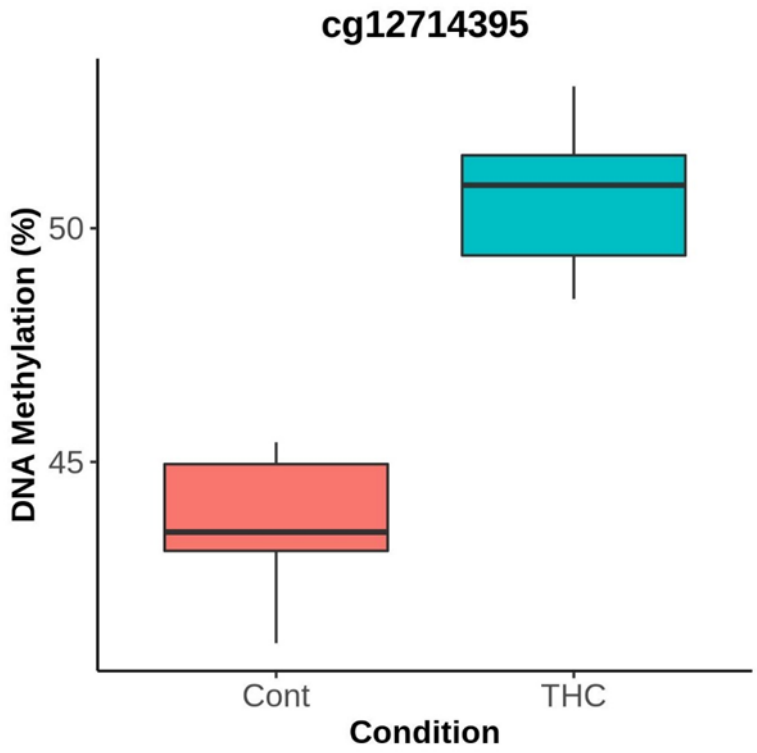


Figure 7.7: cg12714395 shows the largest increase in DNA methylation following a single dose of THC and recovery period. *P-value=2.45E-05, n=6.*

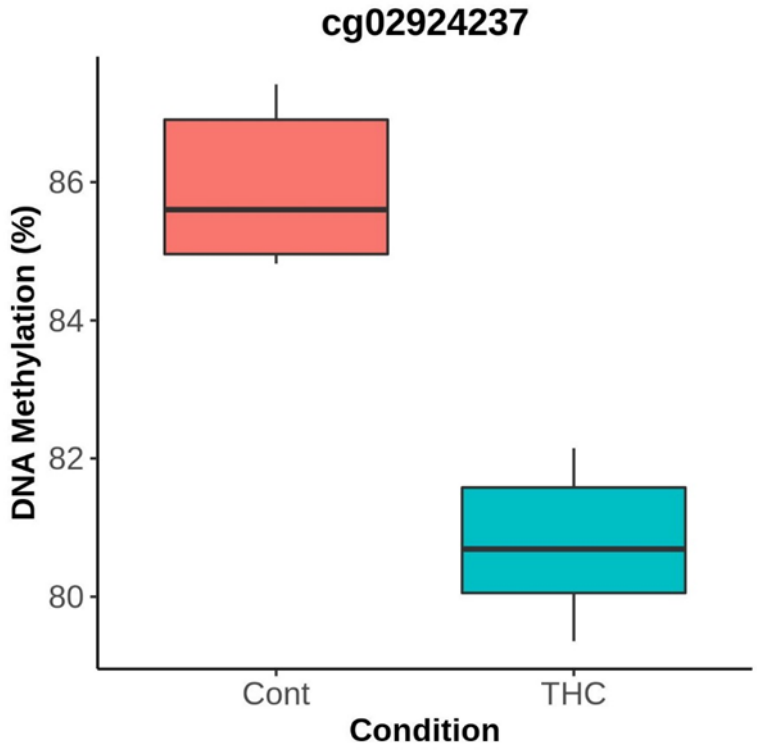


Figure 7.8: cg02924237 shows the largest decrease in DNA methylation following a single dose of THC and recovery period. *P-value=1.20E-05, n=6.*

7.4.5. GO Analysis identified differential methylated pathways involved in epigenetic function and tyrosine kinase activity following single THC exposure and recovery.

As previously stated in **7.4.3.** , due to lack of power I used a nominal significance threshold of $5x10^{-3}$ to increase the number of DMPs for GO analysis. Increasing the p-value threshold resulted in 4009 DMP which were used to identify enriched GO terms and functions using the missMethyl package.

As previously reported no, GO term reached FDR significance following GO analysis with missMethyl. **Table 7.6** lists the top 20 GO terms enriched in the 4009 DMP selected. The top GO term which is enriched for an epigenetic marker of histone H4 acetylation (FDR<0.1).

Table 7.6: Top 20 GO terms associated with a single exposure of 5μM THC treatment.

Category	Term	Ontology	Total Number of Probes in Category	Number DMP in Category	P	FDR
GO:0090239	regulation of histone H4 acetylation	BP	13	10	4.10E-06	0.0932
GO:1903997	positive regulation of non-membrane spanning protein tyrosine kinase activity	BP	7	7	1.65E-05	0.188
GO:1903995	regulation of non-membrane spanning protein tyrosine kinase activity	BP	8	7	0.000115	0.872
GO:0071012	catalytic step 1 spliceosome	CC	5	4	0.000524	1
GO:0005272	sodium channel activity	MF	41	16	0.000647	1
GO:2000618	regulation of histone H4-K16 acetylation	BP	6	5	0.000868	1
GO:0050662	coenzyme binding	MF	281	59	0.00142	1
GO:0005828	kinetochore microtubule	CC	7	5	0.00144	1
GO:0050717	positive regulation of interleukin-1 alpha secretion	BP	3	3	0.00144	1
GO:0098742	cell-cell adhesion via plasma-membrane adhesion molecules	BP	270	65	0.00179	1
GO:0097264	self proteolysis	BP	8	6	0.00184	1
GO:0043087	regulation of GTPase activity	BP	469	102	0.00205	1
GO:0009403	toxin biosynthetic process	BP	6	4	0.00207	1
GO:2000820	negative regulation of transcription from RNA polymerase II promoter involved in smooth muscle cell differentiation	BP	3	3	0.00214	1
GO:0007528	neuromuscular junction development	BP	47	17	0.00228	1
GO:0044459	plasma membrane part	CC	2732	456	0.00235	1
GO:0022610	biological adhesion	BP	1391	259	0.00237	1

GO:0048566	embryonic digestive tract development	BP	34	13	0.00246	1
GO:0010719	negative regulation of epithelial to mesenchymal transition	BP	30	11	0.00249	1
GO:0071006	U2-type catalytic step 1 spliceosome	CC	4	3	0.00253	1

7.4.6. <u>There are shared DMP between the continuous and singular dosage THC cells.</u>

In order to examine shared functions between the continuous dosage and singular dosage of THC I set out to examine if there are any DMP shared between the two data sets. Following on from DMP identification I then set out to perform GO analysis using missMethyl to identify if there are any shared pathways which are common to both doses.

To do this I selected DMPs from both datasets with a p-value less than $5x10^{-2}$, this would allow for enough probes which are overlapping to perform meaningful GO analysis. I identified 44,414 DMP p-value <0.05 in the single THC dose samples, and 31,046 DMP in the continuous dose samples. Of these probes, I found that 1639 were shared between both the single dose and continuous dose datasets (**Figure 7.9**). There is no correlation in the direction of effect between the shared DMPs (Pearson's correlation coefficient, p-value=0.81) (**Figure 7.10**). 407 DMP have increased DNA methylation in both datasets. **Figure 7.11** shows the top 10 DMP with the largest average increase in DNA methylation between the two groups and their associated genes are shown in **Table 7.7**. 425 DMP have decreased DNA methylation in both datasets, with **Figure 7.12** showing the top 10 DMP with the largest average decrease in DNA methylation between the two groups and their associated genes listed in **Table 7.8**. I also identified 428 DMPs which have increased DNA methylation in single THC dose and decreased DNA methylation in continuous dose, and 379 DMP with decreased DNA methylation in single THC and increased DNA methylation in the continuous dose.

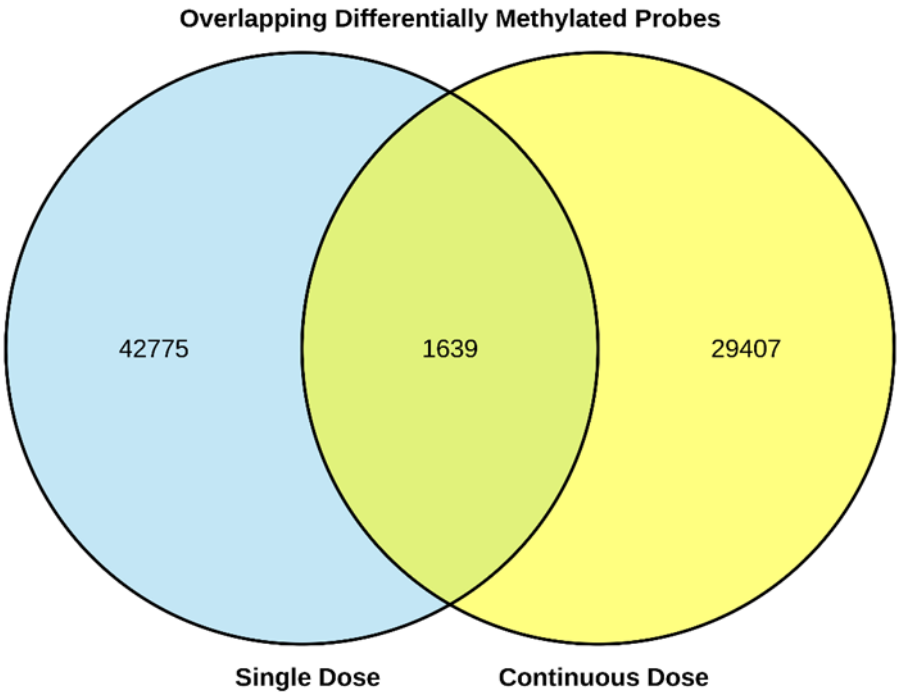


Figure 7.9: 1639 probes are differentially methylated in both datasets (P<0.05). *42775 probes which are differentially methylated are unique to the single THC dose (blue) and 29407 probes are unique to the continuous dose (yellow).*

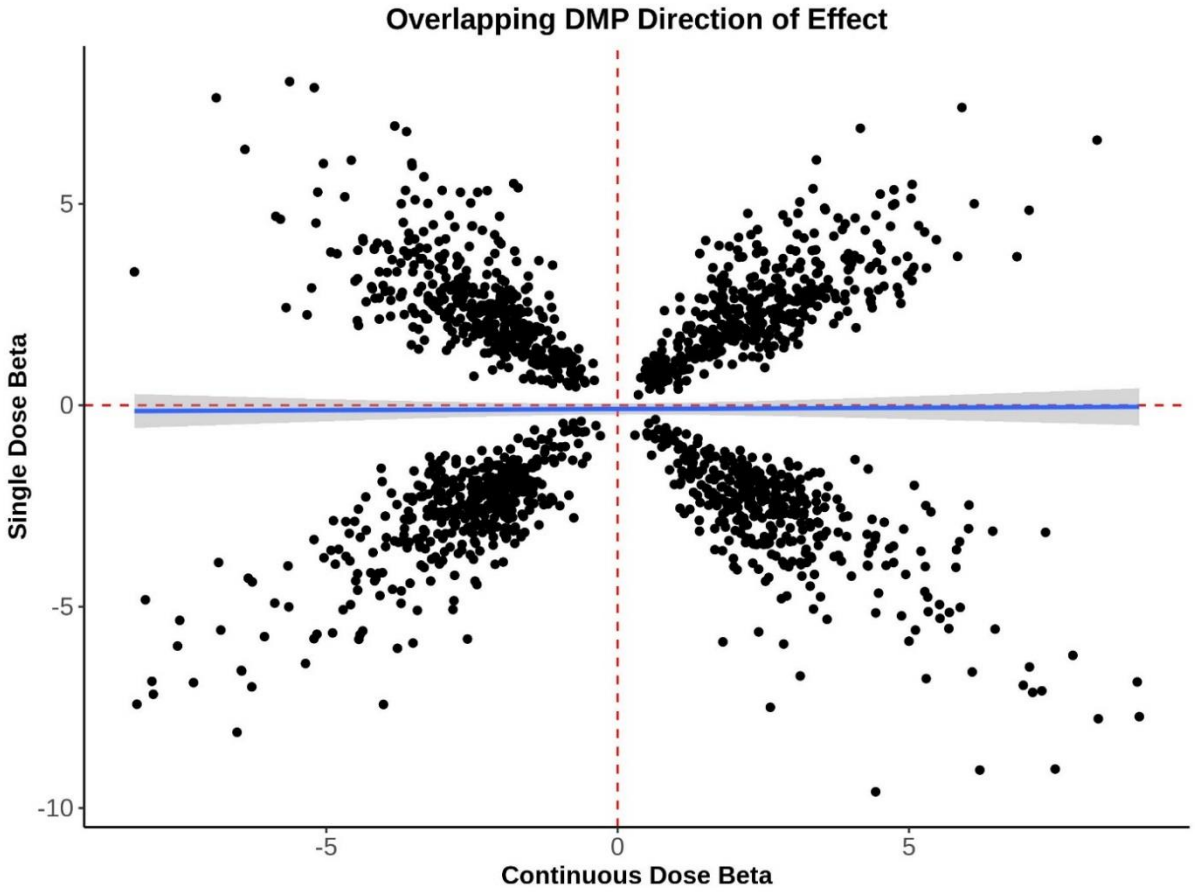


Figure 7.10: There is no correlation between the direction of effect in the shared DMPs between single and continuous dosage of THC. *Pearson's correlation coefficient, P=0.81).*

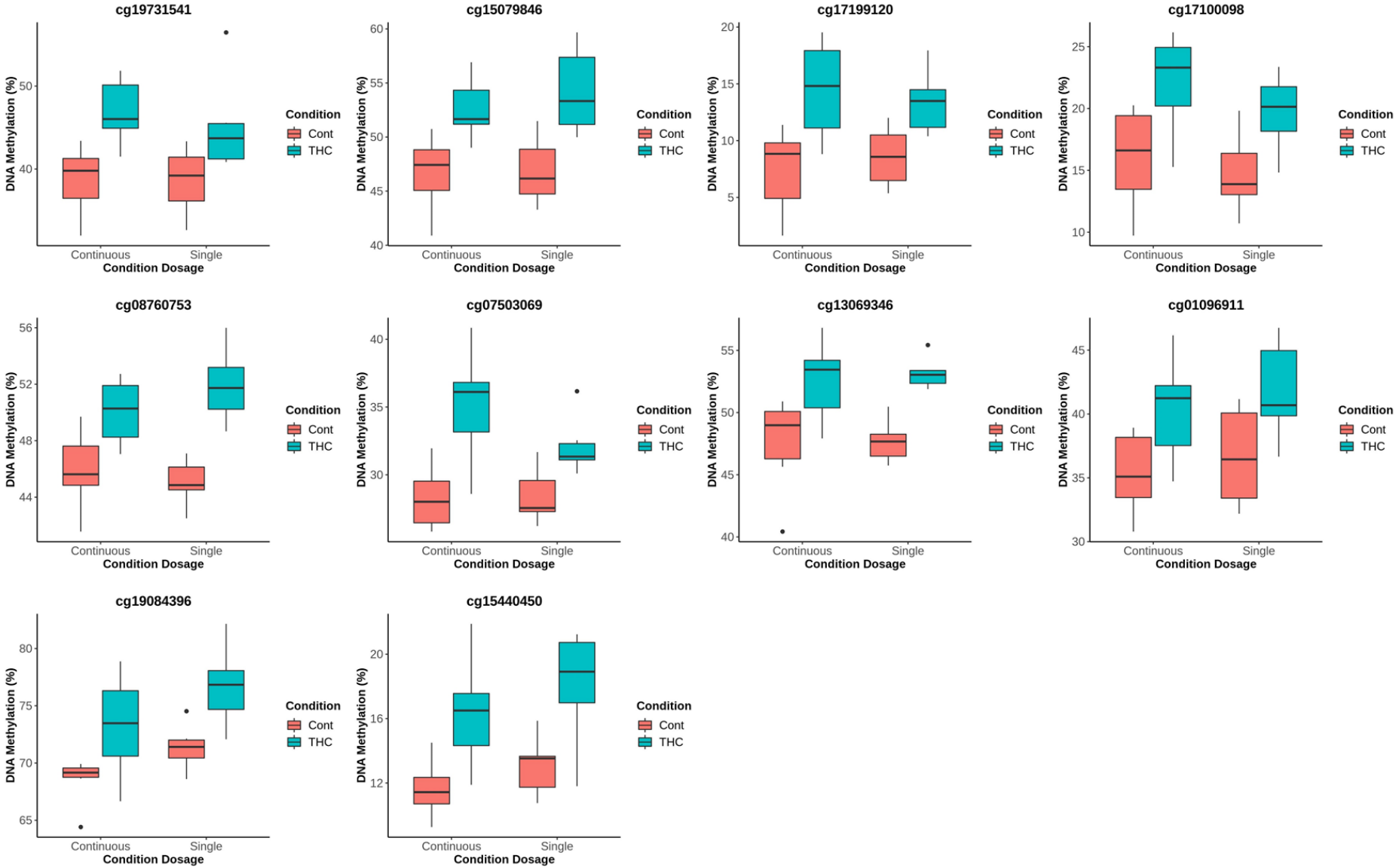


Figure 7.11: The top 10 shared probes with the largest average increase in DNA methylation between the continuous and single dosage of THC. These are the top 10 probes with the largest average increase in DNA methylation in the THC treated samples in both the continuous and single doses of THC compared to their controls. N=6 for each group.

Table 7.7: The top 10 shared probes with the largest average increase in DNA methylation between the continuous and single dosage of THC and their associate genomic location and genes.

Probe ID	Average Methylation Difference (%)	Continuous Methylation Difference (%)	Continuous P	Single Methylation Difference (%)	Single P	Chr	Genomic Location	Gene	Gene Region
cg19731541	7.40	8.23	0.00596	6.58	0.0475	2	206553344	NRP2	Body
cg15079846	6.65	5.91	0.0103	7.39	0.00522	6	169513125		
cg17199120	5.95	7.06	0.0139	4.84	0.0126	5	124244323		
cg17100098	5.56	6.12	0.0284	5.00	0.0212	4	111544781	PITX2	TSS1500;Body
cg08760753	5.52	4.17	0.0197	6.88	0.000283	22	49069623	FAM19A5	Body
cg07503069	5.27	6.85	0.00605	3.69	0.0127	2	241084049		
cg13069346	5.27	5.05	0.0371	5.48	8.29E-05	13	66437506		
cg01096911	5.09	5.03	0.0421	5.14	0.0488	15	60763888	ICE2	Body;5'UTR
cg19084396	5.05	4.74	0.0407	5.35	0.00866	6	41237112		
cg15440450	4.88	4.76	0.0138	5.00	0.0124	9	129606471		

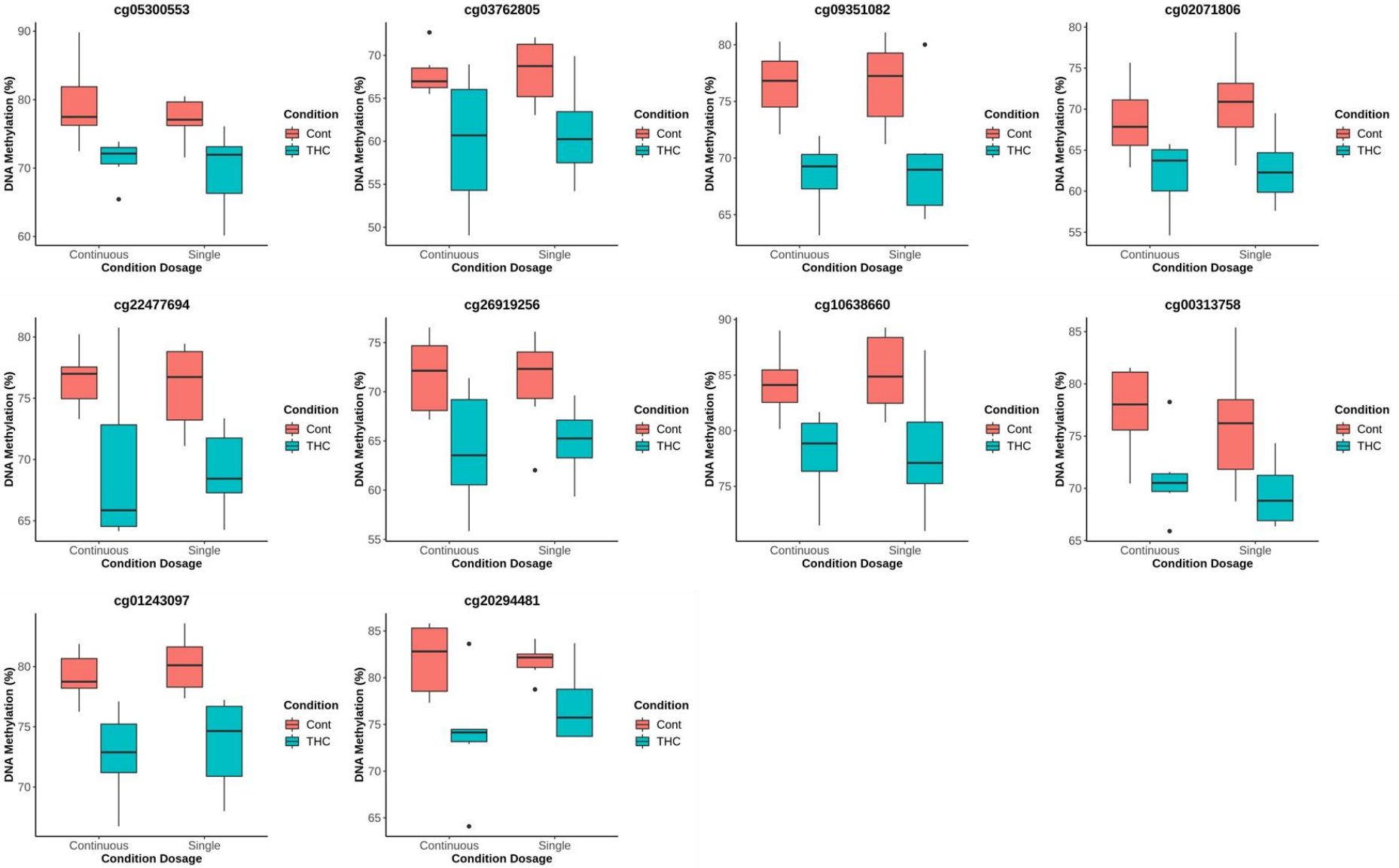


Figure 7.12: The top 10 shared probes with the largest average decrease in DNA methylation between the continuous and single dosage of THC. These are the top 10 probes with the largest average decrease in DNA methylation in the THC treated samples in both the continuous and single doses of THC compared to their controls. N=6 for each group.

Table 7.8: The top 10 shared probes with the largest average decrease in DNA methylation between the continuous and single dosage of THC and their associate genomic location and genes.

Probe ID	Average Methylation Difference (%)	Continuous Methylation Difference (%)	Continuous P	Single Methylation Difference (%)	Single P	Chr	Genomic Location	Gene	Gene Region
cg05300553	-7.83	-8.25	0.0150	-7.42	0.0248	2	181049644		
cg03762805	-7.57	-7.97	0.0415	-7.17	0.0266	17	30040601		
cg09351082	-7.42	-7.99	0.00125	-6.85	0.0335	10	97319133	SORBS1	5'UTR
cg02071806	-7.32	-6.53	0.0319	-8.12	0.0181	15	69857379	DRAIC	Body
cg22477694	-7.08	-7.28	0.0351	-6.89	0.00660	5	74011256	HEXB	Body
cg26919256	-6.76	-7.55	0.0293	-5.98	0.0412	15	61709352		
cg10638660	-6.63	-6.28	0.0105	-6.99	0.0294	18	8717365	MTCL1	TSS200
cg00313758	-6.52	-6.45	0.0245	-6.60	0.0382	11	73599569	PAAF1	5'UTR;Body
cg01243097	-6.52	-6.46	0.00411	-6.58	0.00531	21	35146511	ITSN1	Body
cg20294481	-6.47	-8.10	0.0217	-4.83	0.0240	20	3524996	ATRN	Body

7.4.7. GO analysis of the overlapping DMP identifies pathways involved in endoplasmic reticulum stress, oestrogen binding, structural organisation and brain-derived neurotropic factor binding.

I then set out to examine if there are any pathways or GO terms enriched within the set of overlapping probes between the continuous and single dosage of THC. I then examined the subset of shared probes with increased and decreased methylation in both data sets to identify any pathways which are altered and could be robust changes. As with previous GO analysis, due to the small sample size and number of probes included no GO term reaches FDR significance.

Overall in the entire set of 1639 overlapping significant probes (considered p-value<$5x10^{-2}$ in both data sets) I identified pathways involved in oestrogen response binding (p-value=8.70-E05), cellular components (p-value=9.62E-05), and apoptotic signalling following endoplasmic reticulum stress (p-value=$1.23x10^{-4}$). The top 20 GO terms for the overlapping DMP are listed in **Table 7.9**.

When I examined the 407 overlapping probes which have increased DNA methylation in both data sets I identified GO terms involved in anatomical structural arrangement (p-value=1.62E-06) and cranial nerve structural organisation (p-value=8.78E-05). The top 20 GO terms for the increased methylation probes are listed in **Table 7.10**.

When examining to 425 DMP which have decreased DNA methylation within the two datasets I identified changes to the GO term associated with brain-derived neurotropic factor binding (p-value=2.34E-05). The top 20 GO terms for the shared decrease in DNA methylation are shown in **Table 7.11**.

Table 7.9: Top 20 GO terms in overlapping DMP between the single and continuous dosage of THC.

Category	Term	Ontology	Total Number of Probes in Category	Number DMP in Category	P	FDR
GO:0034056	estrogen response element binding	MF	4	4	8.70E-05	0.935
GO:0005575	cellular_component	CC	18546	1088	9.62E-05	0.935
GO:0070059	intrinsic apoptotic signalling pathway in response to endoplasmic reticulum stress	BP	58	13	0.000123	0.935
GO:0005622	intracellular	CC	14678	929	0.000533	1
GO:0006305	DNA alkylation	BP	71	13	0.000760	1
GO:0006306	DNA methylation	BP	71	13	0.000760	1
GO:0044424	intracellular part	CC	14371	911	0.000841	1
GO:0031400	negative regulation of protein modification process	BP	590	61	0.000880	1
GO:1902236	negative regulation of endoplasmic reticulum stress-induced intrinsic apoptotic signaling pathway	BP	18	6	0.000889	1
GO:0048403	brain-derived neurotrophic factor binding	MF	3	3	0.00104	1
GO:0009165	nucleotide biosynthetic process	BP	338	37	0.00105	1
GO:0055086	nucleobase-containing small molecule metabolic process	BP	718	64	0.00106	1
GO:0006999	nuclear pore organization	BP	14	5	0.00119	1
GO:1903573	negative regulation of response to endoplasmic reticulum stress	BP	42	9	0.00125	1
GO:0045197	establishment or maintenance of epithelial cell apical/basal polarity	BP	38	10	0.00126	1
GO:1901293	nucleoside phosphate biosynthetic process	BP	341	37	0.00126	1
GO:0045668	negative regulation of osteoblast differentiation	BP	50	10	0.00133	1

GO:2000378	negative regulation of reactive oxygen species metabolic process	BP	58	11	0.00135	1
GO:0045936	negative regulation of phosphate metabolic process	BP	563	58	0.00144	1
GO:0010563	negative regulation of phosphorus metabolic process	BP	564	58	0.00147	1

Table 7.10: Top 20 GO terms for the shared probes with increased DNA methylation between continuous and single dosage of THC.

Category	Term	Ontology	Total Number of Probes in Category	Number DMP in Category	P	FDR
GO:0048532	anatomical structure arrangement	BP	17	6	1.62E-06	0.070
GO:0021604	cranial nerve structural organization	BP	11	4	8.78E-05	0.999
GO:0061004	pattern specification involved in kidney development	BP	9	3	0.000318	1
GO:0072048	renal system pattern specification	BP	9	3	0.000318	1
GO:0021602	cranial nerve morphogenesis	BP	28	5	0.000339	1
GO:0021783	preganglionic parasympathetic fiber development	BP	17	4	0.000533	1
GO:0016266	O-glycan processing	BP	60	6	0.00063	1
GO:0072086	specification of loop of Henle identity	BP	4	2	0.000710	1
GO:0043457	regulation of cellular respiration	BP	24	4	0.000801	1
GO:0048486	parasympathetic nervous system development	BP	19	4	0.000831	1
GO:0021612	facial nerve structural organization	BP	9	3	0.00101	1
GO:1901135	carbohydrate derivative metabolic process	BP	1163	33	0.00108	1
GO:0072047	proximal/distal pattern formation involved in nephron development	BP	5	2	0.00138	1
GO:0072081	specification of nephron tubule identity	BP	5	2	0.00138	1
GO:0071442	positive regulation of histone H3-K14 acetylation	BP	4	2	0.00138	1
GO:1903135	cupric ion binding	MF	4	2	0.00166	1
GO:0032580	Golgi cisterna membrane	CC	77	6	0.00168	1
GO:0021561	facial nerve development	BP	11	3	0.00186	1
GO:0021610	facial nerve morphogenesis	BP	11	3	0.00186	1

GO:0042339	keratan sulfate metabolic process	BP	33	4	0.00196	1

Table 7.11: Top 20 GO terms for the shared probes with decreased DNA methylation between continuous and single dosage of THC.

Category	Term	Ontology	Total Number of Probes in Category	Number DMP in Category	P	FDR
GO:0048403	brain-derived neurotrophic factor binding	MF	3	3	2.34E-05	0.532
GO:0044424	intracellular part	CC	14371	259	0.00028	1
GO:0032899	regulation of neurotrophin production	BP	2	2	0.000384	1
GO:0006641	triglyceride metabolic process	BP	100	7	0.000561	1
GO:0005622	intracellular	CC	14678	261	0.000786	1
GO:0006639	acylglycerol metabolic process	BP	122	8	0.000919	1
GO:0009214	cyclic nucleotide catabolic process	BP	14	4	0.000929	1
GO:0006638	neutral lipid metabolic process	BP	123	8	0.000998	1
GO:0004115	3',5'-cyclic-AMP phosphodiesterase activity	MF	15	4	0.001213	1
GO:2000620	positive regulation of histone H4-K16 acetylation	BP	2	2	0.001316	1
GO:0061357	positive regulation of Wnt protein secretion	BP	3	2	0.001531	1
GO:0051569	regulation of histone H3-K4 methylation	BP	26	4	0.001753	1
GO:0043121	neurotrophin binding	MF	9	3	0.002261	1
GO:0016874	ligase activity	MF	150	9	0.002379	1
GO:0015849	organic acid transport	BP	315	13	0.002583	1
GO:0046942	carboxylic acid transport	BP	315	13	0.002583	1
GO:0072523	purine-containing compound catabolic process	BP	51	5	0.002954	1
GO:0019432	triglyceride biosynthetic process	BP	38	4	0.003013	1
GO:0001866	NK T cell proliferation	BP	7	2	0.003211	1
GO:0006198	cAMP catabolic process	BP	9	3	0.003252	1

7.5. Discussion

In this chapter I set out to identify if exposure to THC, the psychoactive component of cannabis, alters DNA methylation in a cell model. Cannabis exposure has been associated with an increased risk of developing psychosis and schizophrenia however the aetiology of how it does this remains unknown. To date there have been no genome wide methylation studies of THC exposure on a cell line. Previous studies investigating the effect of THC exposure on the methylome in both targeted and genome-wide studies have produced conflicting results (see **Table 7.1**). In this study I exposed SH-SY5Y neuroblastoma cells to continuous or single 5µM doses of THC over a period of seven days. After exposure, DNA methylation was measured using the Illumina Infinium HumanMethylationEPIC BeadChip array. DNA methylation changes were investigated in both the single and continuously exposed cells compared to vehicle. The overlap between the two datasets was investigated to see if there are any robust and shared DNA methylation changes which are due to THC exposure.

I identified nominal DMPs in both the single and continuous dosage of THC in a SH-SY5Y cell line compared to ethanol vehicle control. When examining the continuous dosage of THC the largest increase in DNA methylation was +6.13% in cg15120122 located in the 5' UTR of *MYO9B* (**Figure 7.5**). *MYO9B* has been shown to repress RhoA/ROCK signalling involved in dendrite morphogenesis in cortical neurones both *in vitro* cell culture and *in vivo* rat models (Long *et al.*, 2013). Another study identified that *MYO9B* deficiency in mouse model impairs cellular motility, taxis, and protrusion formation in macrophages (Hanley *et al.*, 2010). The largest decrease in DNA methylation was -5.56% in cg09618674 located in the body of *LRP1B* (**Figure 7.6**). *LRP1B* is a tumour suppressor gene and a miRNA targeting *LRP1B* has been shown to be downregulated in the striatum of macaques following THC exposure (Liu *et al.*, 2000; Simon *et al.*, 2016). Interestingly *DGCR11* a gene associated with schizophrenia through 22q11 deletion syndrome is hypomethylated by -2.80% at cg13841476 (Bassett, Chow and Weksberg, 2000; Bassett and Chow, 2008; Rees *et al.*, 2014). The 22q11 region has previously been associated as a region of interest linking cannabis and schizophrenia, with *COMT* deletion highly associated with cannabis dependence (Tunbridge, Harrison and Weinberger, 2006; Williams, Owen and O'Donovan, 2007). In the single dosage of THC the largest increase in DNA

methylation was +7.02 observed in cg12714395 (**Figure 7.7**). Cg12714395 is located in an intronic region within a DNAse hypersensitivity site. The nearest gene is *SLITRK1* which is approximately 8.2Mb downstream and encodes for a transmembrane protein involved in synapse development (Beaubien *et al.*, 2016). However from this data I cannot conclude that this CpG is directly involved in expression of *SLITRK1.* The largest decrease in DNA methylation observed is -5.15% in cg02924237 located in the exon body of *ATRNL1* (**Figure 7.8**). *ATRNL1* is a paralog of *ATRN* which produces attractin which is involved in myelination, however not much is known about *ATRNL1* function (Kuramoto *et al.*, 2001). *ATRNL1* transcripts are highly expressed in the mouse dentate gyrus and hippocampus and is thought to be implicated in energy homeostasis (Walker *et al.*, 2006). Interestingly a patient suffering from a microdeletion of 10q25.3, which includes *ATRNL1*, presented with neurodevelopmental abnormalities, however the role of *ATRNL1* in development has not been explored further (Stark *et al.*, 2010). I identified more DMPs in the single dosage compared to the continuous dosage. This could be due to response or stress pathways activated in first exposure where in the case of the continuous dose the cells could have entered a phase of tolerance where DNA methylation changes are limited at day seven following recurrent dosages.

When I examined the GO pathways altered in the single dose THC DMP I identified pathways involved in epigenetic mechanisms and tyrosine kinase signalling (**Table 7.6**). Histone H4 acetylation is associated with chromatin decompaction and allows for transcription factor access and is thought to play a role in DNA replication (Vettese-Dadey *et al.*, 1996; Ruan *et al.*, 2015). Global histone modifications, including acetylation and methylation, have been previously associated in T cells exposed to THC compared to vehicle (Yang *et al.*, 2014). Another GO term which is enriched is the regulation of tyrosine kinase activity from non-membrane spanning tyrosine kinases. Tyrosine kinases possess a broad range of activities, the common theme however is the transfer of phosphate groups from ATP to other proteins. This usually results in protein activation and is involved in numerous protein signalling cascades (Hubbard and Till, 2000). These observations are suggestive of differences in cellular signalling following a single dose of THC. This shows that single dosage of THC could be enough to cause global changes in gene expression through epigenetic mechanisms independent of DNA methylation, such as histone modifications. In order

to examine this, RNAseq or chromatin immunoprecipitation sequencing (ChIP-seq) could be undertaken on samples exposed to THC to see if transcription has changed independent of DNA methylation through chromatin modifications. Interestingly when I examined the continuous dose of THC GO terms there was an enrichment in probes associated with neurogenesis and neurodevelopment (**Table 7.4**). This supports the hypothesis that THC is impacting on neurodevelopment (Panlilio and Justinova, 2018). It could be that continuous exposure to THC results in delayed neurodevelopment or impairs normal neurodevelopment. This is crucial as the brain is continually developing through adolescence, and this is when first exposure to THC mostly occurs.

When I examined overlapping effects between single and continuous dosage of THC I observed no correlation in the direction of effect of the DMP. However I do see an enrichment of some GO terms such as oestrogen response and endoplasmic reticulum stress (**Table 7.9**). Oestrogen response elements are short sequences within DNA which oestrogen hormones bind to and initiate transcription (Klinge, 2001). Oestrogen hormones have been shown to promote synaptogenesis and synaptic plasticity in both brain slices and cultured neurones (Jelks *et al.*, 2007; Babayan and Kramár, 2013). Oestrogen has further been shown to promote differentiation, and formation of dendrites in SH-SY5Y cells (Teppola *et al.*, 2016). Therefore dysregulation of oestrogen signalling could disrupt normal neuronal development. Endoplasmic reticulum stress can be caused due to a build-up of misfolded proteins which can then result in apoptosis to protect the neighbouring cells (Lin, Walter and Yen, 2007). Previous research has shown that THC can induce autophagy through endoplasmic reticulum stress in cancer cell lines and can also impair mitochondrial function resulting in further stress (Salazar *et al.*, 2009; Lojpur *et al.*, 2019).

However there are a subset of DMP which do share the same direction of effect, 407 DMP have shared increases in DNA methylation and 425 DMP have shared decreases in DNA methylation in both. The shared probes with an increase of DNA methylation are involved with structural arrangement and nerve structure (**Table 7.10**). THC has been shown to alter the morphology of neuronal dendrites in animal models (Fernández-Cabrera *et al.*, 2018; Kolb *et al.*, 2018). However the effect on dendrite spine density was dependent on the brain region examined with decrease in density observed in the frontal cortex, and increases identified in the medial prefrontal cortex and nucleus accumbens (Kolb *et al.*, 2018). When I examined the GO terms in the

shared decrease in DNA methylation there is an enrichment for brain-derived neurotropic factor (BDNF) signalling (**Table 7.11**). BDNF is a known neurotrophin which is involved in neuronal development and synaptic plasticity (Kowiański *et al.*, 2018). Increases in BDNF expression is observed following acute and chronic THC treatment in rats (Suliman *et al.*, 2018). Increases in BDNF have also been observed in humans, where chronic exposure to THC over 12 months in adolescents increases in plasma BDNF (Miguez *et al.*, 2019). When BDNF was overexpressed in mice memory impairments were prevented following dosage with THC, identifying a protective role of BDNF (Segal-Gavish *et al.*, 2017). It is possible that increased BDNF expression results in the synaptic changes observed given its key role in neurogenesis. These results indicate that there are shared common functions and responses between the two exposures of THC, however these could be confounded by the cell lines rather than the exposure, however none of these GO terms reached FDR<0.05 these results should be treated with caution.

There are several important considerations to take into account when regarding these results and how this experiment could be improved in the future:

1) The most important note is that none of the DMP or GO terms reached significance following correction for multiple testing. This is simply because the sample size is too small and the dataset too large. In order to combat this either the sample size should be increased or the number of probes examined reduced to target relevant probes. The former could be achieved using single cell genomics by looking at single cell methylation patterns from exposed cells and control cells rather than bulk analysis from a well of cells. For the latter, targeted approaches could be problematic due to bias brought about in the target selection process and the bulk of data missed by this approach.
2) The dosage selected was based on previous research within SH-SY5Y cells. Although the DNA methylation changes observed were small this could be due to the dosage of THC used. In future experiments it would be beneficial to perform a dose viability assay to calculate the EC50 of THC on SH-SY5Y cells. The reduced viability at 10µM observed in previous studies could have been due to differences in culturing techniques, cell age, or sample preparation and not just dosage of THC. It is possible that the dose was insufficient to mimic exposure to THC in brain as it is currently unknown how much THC reaches

the brain, and this could be highly variable between location, method of ingestion, and THC content.

3) In this experiment I used undifferentiated SH-SY5Y cells. This could have confounded the results as SH-SY5Y cells are a bone marrow derived neuroblastoma cell line. Therefore the response observed may be muted or exacerbated in different cell lines. Future experiments could examine the effect on differentiated SH-SY5Y cells towards neuronal fates (as previously described in Chapter 3). Further research could have been carried out on other cell lines or even differentiated iPSC cells. This would then make the results more robust if the same DMP are identified in multiple cell types over multiple experiments.
4) Although I state that seven days of exposure to THC is continuous it would be beneficial to increase the time course of exposure to examine trans-generational effects following increased passaging and aging of cells. There could be changes in DNA methylation which are yet to be observed due to the short exposure time to THC. Further to this it could be that the cell line could become tolerant to THC after a certain time, resulting in a reversion or no further changes in DNA methylation. These would all be critical in analysing the overall effect of THC on the methylome.
5) Finally there is limited translational ability of these findings to the cannabis available to the general population. There are other components of cannabis, namely CBD, and the interaction between THC and CBD has still yet to be elucidated. Further experiments would need to carefully examine what effect the CBD:THC ratio has on DNA methylation. These doses are also difficult to translate between human exposure and cell model exposure. In population based studies the concentration of THC and CBD vary greatly between countries and even within regions (Di Forti *et al.*, 2019), therefore it is very difficult to quantify THC exposure in different populations. We still do not know how much THC (ingested or smoked) reaches the brain through the bloodstream, until this is further examined we will not know to what extent THC has an effect on DNA methylation in the brain.

7.6. Conclusion

I have identified that exposure to THC results in nominal changes in DNA methylation across the genome. However further work will have to be done in other cell lines with increased dosages and time courses to identify if these changes are robust and translational between different cell types. This has opened the door for further work into how THC can result in psychotic episodes through alterations in DNA methylation and the potential longitudinal effect of continuous cannabis usage on the biology of neurons.

Chapter 8. General Discussion

8.1. Summary of results

From previous research we know that there is strong evidence that genetics, epigenetics, and the environment are all involved in the aetiology of schizophrenia. More and more (epi)genomic risk loci are being uncovered as GWAS and EWAS studies become more powered to identify smaller effect sized by increased cohort sizes, however very few of these risk variants have a clear functional link to how they contribute to disease due to lack of appropriate molecular tools and methods to reliably fine map regions of LD. In this PhD I set out to develop molecular biology tools and pipelines to be able to functionally validate genetic, epigenetic, and environmental risks for schizophrenia in cellular models. To do this, I used cutting edge molecular biology techniques to generate knockout cell lines, edit the epigenome at specific risk sites, examine promoter activity of DMRs, and examine methylomic changes following drug exposure.

8.1.1. Chapter 3: Generation of the knockout cell line *AS3MT*

In Chapter 3. I describe my methodology for generating knockout SH-SY5Y cell lines, a previously unreported CRISPR edited cell line. I then used this pipeline to knockout the robust schizophrenia risk gene *AS3MT* and then characterise changes in cellular morphology following differentiation towards neuronal-like fates. The SH-SY5Y cell line, at the time of experimenting, had not been used as a cell line for CRISPR/Cas9 experiments, therefore there was no protocol established for successful editing of this cell line. I therefore set out to generate a method to knockout *AS3MT* using a modified CRISPR pipeline first described by (Ran *et al.*, 2013). My methodology utilised two gRNA to knockout the main methyltransferase domain of *AS3MT* between exons 4 and 6 of the gene, thus rendering the two known transcripts of *AS3MT* non-functional and redundant. Through this protocol I generated three homozygous knockout *AS3MT* cell lines confirmed by PCR, Sanger sequencing, and western blotting, and three heterozygous knockout *AS3MT* cell lines confirmed by PCR.

I then set out to identify if *AS3MT* plays a role in neuronal development as previously hypothesised through both statistical and laboratory experimental analysis (Hannon, Dempster, *et al.*, 2016; Hannon, Spiers, *et al.*, 2016; M. Li *et al.*, 2016). I identified that indeed there is differential isoform expression of *AS3MT* throughout SH-SY5Y differentiation towards neuronal fates. I then further identified that knockout of *AS3MT*

results in a significant increase in the longest neurite length, increase in cellular size, and a trend towards an increase in total neurite length. These results suggest that *AS3MT* is involved in neuronal development and regulates normal neurite outgrowth in differentiated neurones. Thus changes to *AS3MT* expression through changes in gene regulation mediated by schizophrenia associated eQTLs or mQTLs, may contribute to a dysregulation of normal neurogenesis.

The analysis performed in this chapter was carried out on two independent knockout *AS3MT* cell lines and two independent controls which were passage and experimentally matched to identify if these changes are robust. One caveat of this research is that the work was carried out in SH-SY5Y neuroblastoma cells. While these cells have often been used as a model system for schizophrenia and neuroscience research (Xicoy, Wieringa and Martens, 2017) they are not strictly neuronal as they were originally bone marrow derived. Therefore it would be prudent to validate these findings in either human derived iPSC differentiated neurones or an animal model for neurodevelopment such as zebrafish.

This protocol which has been developed can be easily adjusted for different diseases, cell lines, genes, and can even be used to knockout specific risk conferring SNPs. This should hopefully pave the way for functional validation of genomic risk loci and genes in a number of different diseases.

8.1.2. Chapter 4: Transcriptional profiling of the *AS3MT* knockout cell line

In Chapter 4. I describe analysis of the transcriptional changes following knockout of *AS3MT* in the SH-SY5Y cells. At the time there were currently no transcriptional studies examining the effect of *AS3MT* knockout in cell lines or animal models. I wanted to examine how knockout of *AS3MT* impacts the transcriptome and how these changes could correlate to neuronal development. To do this I used RNAseq to identify genes of interest in the *AS3MT* signalling cascade that could be involved in neuronal development and not the canonical arsenic metabolism pathways, to open up new avenues of exploration.

Using these methodologies I have identified robust changes in gene expression in two independent knockout cell lines compared to matched controls. Further to this I identified over 900 robust overlapping changes in both cell lines which are strongly correlated in terms of their direction of effect when compared to matched controls.

These 900 genes were identified to have functions related to neuronal development, and immunological responses, both of which have been proposed as mechanisms for schizophrenia development (Strous and Shoenfeld, 2006; Owen *et al.*, 2011). When I examined the overlapping dataset more closely, I identified 27 genes which are differentially expressed in both knockout cell lines which have been previously associated with schizophrenia through GWAS (Pardiñas *et al.*, 2018). Furthermore, I identified 36 shared genes which are upregulated by a fold change >1 in both knockout cell lines which included genes involved in neuronal development and neuronal function. Further, when I examined the overlapping genes which are downregulated in both datasets (fold change <-1), I identified a downregulation in genes involved in axon development and extracellular matrix formation, further strengthening the previous findings in Chapter 3. that *AS3MT* is involved in neuronal development.

One clear caveat in this chapter is the use of undifferentiated *AS3MT* knockout cell lines. It could be that differentiation and sequencing would identify more changes in gene expression in the neuronal population versus the non-neuronal. I also used two different sequencing runs with the experiments run at different times, it would have been beneficial to run both experiments at once to reduce the batch effect, however the fact I am seeing overlapping genes in a completely independent experiment with two different knockout and control cell lines makes these findings more robust and are more likely linked to *AS3MT* knockout than other confounders.

In summary I identified that knockout of *AS3MT* results in transcriptional changes in genes involved in ribosomal activity, neurogenesis, and neurodevelopment with enrichment in previously associated schizophrenia genes. This further supports the hypothesis, and the results in Chapter 3. that *AS3MT* is involved in neuron development, and that dysregulation through genetic variation at this locus can contribute to schizophrenia.

8.1.3. Chapter 5: Functional analysis of differentially methylated regions using luciferase reporters

In Chapter 5. I describe how we can examine the promotional ability and the effect of DNA methylation on identified DMRs from EWAS studies using a luciferase based assay in cellular models. To do this I used a CpG free luciferase vector from (Klug and Rehli, 2006) to establish if a DMR within *PSORS1C3,* identified through an EWAS

study into suicide completers from (Murphy *et al.*, 2017; Policicchio *et al., 2020*), is in an active promoter, and if methylation of the region *in vitro* has an effect on its promotional ability.

Using this method, I cloned in the DMR into the CpG free vector pCpGL-basic and methylated it *in vitro* using *M.sssI* and SAM. Following transfection of the methylated and unmethylated vectors into HEK293 cells I identified that the *PSORS1C3* DMR sits within an active promoter as luciferase expression is observed. Following methylation the expression of luciferase is quenched back to that of the empty pCpGL-basic vector. Using this I have identified that the DMR identified by (Murphy *et al., 2017)* is within an active promoter which can be controlled by DNA methylation. This experimental approach represents the first step to functional validate DMRs identified from EWAS studies.

The major caveat of this approach is that I have targeted a whole DMR. Within that DMR there may be one causal CpG which is responsible for the change in luciferase activity which I cannot detect with this approach. To combat this, constructs could be created with different lengths or section of the DMR to fine map the causal locus. Furthermore, I do not know which genes the target DMR is driving the expression of, I only know that they are under the control of methylation state. Further work would need to be done, either computationally or laboratory based to examine which genes expression is being changed as a result of a change of DNA methylation status at this DMR.

In summary, I developed and demonstrated utility of a protocol for functionally determining if methylation state of DMRs identified through EWAS influence expression of a reporter gene *in vitro*. This straightforward assay can be easily applied to elucidate a functional role of DMRs identified through EWAS.

8.1.4. Chapter 6: Functional epigenetic editing using CRISPR/Cas9

In Chapter 6. I describe my methodology for editing specific CpG sites using a modified form of CRISPR/Cas9 to examine the resultant transcriptional changes following altered DNA methylation *in vitro.* These pipelines can be used to functionally validate individual CpGs or DMR risk loci identified through EWAS in a number of different diseases. Unlike the luciferase assay provided in Chapter 5. for functional validation

of DMRs, this methodology can determine which gene expression is being controlled by the risk DMP and DMR and provides fine mapping for the direction of causality.

Using the methods first described by (Liu *et al.*, 2016, 2018) I set out to identify if specific CpG sites identified through SMR analysis in a DNA methylation and gene expression dataset by (Hannon *et al.*, 2018) are controlling the expression of hypothesised genes. To do these I selected CpG sites which have a known mQTL which is associated with gene expression through the (Hannon *et al.*, 2018) SMR analysis. I selected two DMRs, one which is hypermethylated and another which is hypomethylated in SH-SY5Y cells to test the hypothesis that DNA methylation is mediating gene expression. I identified four CpG sites within the promoter of *FANCG* which are thought to be under the control of an mQTL which is predicted to drive expression of the nearby genes *PIGO* and *STOML2*, furthermore *STOML2* has been associated with Alzheimer's disease. In order to test the hypothesis that DNA methylation at this genetic region influences gene expression of *PIGO* and *STOML2* I targeted these four CpG using a modified form of dCas9 fused to Tet1CD to reduce DNA methylation. I identified a decrease in DNA methylation of up to 16.5% in the targeted CpGs compared to controls and identified a slight increase in *FANCG* expression but found no change in *PIGO* or *STOML2* expression, suggesting that their expression is not driven by differential methylation of these identified CpGs. I then went on to examine if I can increase DNA methylation using a DMR located in the *SNCA* promoter identified by SMR analysis, which is a gene associated with Parkinson's disease. In this case I used a different modified form of dCas9 this time fused to DNMT3a and DNA methyltransferase. I identified a slight increase in DNA methylation of the *SNCA* promotor compared to controls, however I was unable to establish if gene expression changed as a result due to insufficient edited cells.

One limitation of this study is that I used bulk cells transduced with lentiviruses. Lentiviruses randomly integrate and are constitutively expressed and result in increased protein production, therefore there could be an increase in off target effects. This, coupled with poor gRNA specificity for intronic regions, means that this data could be confounded. In order to account for some of these factors, in the future cell lines could be generated with doxycycline inducible promoters to reduce the constitutive expression, further increasing the control over the system.

In summary, I have developed a protocol for editing DNA methylation, resulting in both and increase and decrease in DNA methylation, at risk CpG sites within SH-SY5Y cells. I managed to engineer a decreases in DNA methylation in the *FANCG* promoter using dCas9 fused to Tet1CD and an increase in DNA methylation in the *SNCA* promoter using dCas9 fused to DNMT3a. The high tissue tropism of lentiviral systems means that this method can easily be applied to other cell lines and diseases to functionally elucidate the effects of altered DNA methylation and their direction of effect on gene expression.

8.1.5. Chapter 7: Examining the effect of THC on the DNA methylome

Finally, in Chapter 7. I describe the effect that THC, the psychoactive constituent of cannabis, a known environmental risk for schizophrenia, has on DNA methylation of the human SH-SY5Y cell line using the Human Infinium MethylationEPIC BeadChip Array. No studies to date have examined the effect of THC exposure on a neuronal cell line to profile genome-wide methylation changes. Doing this, I hoped to identify a functional link between THC exposure and DNA methylation which could explain how cannabis usage relates to increased risk of schizophrenia.

Following exposure of SH-SY5Y cells to 5µM of THC, either as a single dose or constant dosage, over seven days I identified changes in DNA methylation in both data sets. We identified more changes in the single dose compared to the constant dose dataset, this could be due to the continuous cells reaching a level of tolerance to the dosage over the seven days and DNA methylation has stabilised. I then examined the overlapping probes seen between both datasets and identified pathways involved in oestrogen signalling, endoplasmic reticulum stress, structural organisation, and BDNF signalling being disrupted. These have all been previously associated with THC exposure in cell, animal, and human studies but have also been related to neurogenesis and development. This provides insight into how THC impacts on DNA methylation to result in changes to neurodevelopment, which could result in increased risk to developing schizophrenia.

The biggest caveat with this study is that due to the small sample size no associations between THC exposure and DNA methylation reached statistical significant after correcting for multiple testing. To address these limitations, in the future, larger sample sizes should be used, however larger sample sizes would require hundreds of samples

per condition to identify subtle changes in DNA methylation based on power calculations (Mansell *et al.*, 2019). It would be beneficial, as research progresses, to examine specific CpG or genes of interest which have been previously identified to have been changed in response to THC exposure. This could then elucidate changes which would have much more statistical power. Furthermore the dosage of THC used in these experiments could have been too low to result in DNA methylation changes. In future a dose dependent survival assay should be carried out before commencing the experiment to work out the maximal dosage required to give a response without reducing cellular survival.

In summary, I have identified DNA methylation changes in SH-SY5Y cells which are as a result of exposure to 5μM of THC, irrespective of length of exposure. I identified pathways involved in neuronal development and signalling which have robust changes in both single and continuous dosages of THC which are the shared effect of THC exposure. From this, I hope to open avenues for future research into validating these findings and have performed the first DNA methylomic profiling of a neuronal cell line following exposure to THC.

8.2. Current challenges and future directions

As discussed above and in previous chapters, there are a number of limitations associated with studying genetic, epigenetic, and environmental risk factors and their relation to schizophrenia. This next section will discuss some of the common problems associated with these studies and how I have attempted to address these issues in this thesis.

8.2.1. Cell models

In these experiments I have utilised the SH-SY5Y neuroblastoma cell line as the model of choice. SH-SY5Y have been shown to have a relatively normal karyotype, is human derived, and is dopaminergic in nature (Xicoy, Wieringa and Martens, 2017). However the cell line is derived from a bone marrow biopsy and subcloned from the SK-N-SH cell line, therefore the true neuronal characteristics of this cell line are debated (Biedler *et al.*, 1978). While they can form neurites following differentiation with retinoic acid the translational ability to that of stem cell derived neurones is debated. A comparison between neurones from SH-SY5Y cells and mouse P19 embryonal carcinoma cells differentiated showed that the neuronal networks formed by the P19 cells were more complex than SH-SY5Y (Popova, Karlsson and Jacobsson, 2017). There is also the heterogeneity of the SH-SY5Y cell line which results in two subpopulations of cells with differing functions. The “N” are neuroblastomic compared to the “S” which are considered more epithelial in nature and do not exhibit neuronal properties compared to the “N” subtype (Ciccarone *et al.*, 1989; Kovalevich and Langford, 2013). The proportions of these populations could result in differences within cultures, however I minimised these effects by using the same batches of SH-SY5Y cells from the same initial flask for all experiments within this thesis.

Furthermore these cells are grown in 2D culture with only one cell type, the human brain is far more complex than a single monolayer of cells, and there are a multitude of different cells working in harmony signalling to each other which can result in the phenotype. It would therefore be beneficial to study changes within whole organisms such as mouse or zebrafish, which can be used to study neuronal development. However these have the caveat of not being human and the translational ability to human brain development is limited. One new technique which has recently been developed is the creation of 3D cultures from human derived iPSC cells known as

organoids. These organoids transcriptionally resemble human foetal development and allow for the development of multiple different cell types that are seen throughout development such as neurones and astrocytes (Paşca *et al.*, 2015; Sloan *et al.*, 2017, 2018). The development of these technologies paired with CRISPR/Cas9 technologies will pave the way to examine how genetic and epigenetic effects can alter normal neuronal development in 3D space in a human derived cell line.

8.2.2. QTL translation

As previously discussed the genetic makeup of an individual can results in correlation with quantitative traits (QTLs). In humans we know that genetic variation has an effect on gene expression (eQTL) (Aguet *et al.*, 2017) and have been shown to impact on DNA methylation (mQTL) (Hannon *et al.*, 2018). We also know that gene expression changes are tissue specific by nature. In mouse models 12% of transcripts have tissues specific expression bias based on tissue and alleles (Keane *et al.*, 2011). Therefore the landscape is more complex than a disease associated SNP causing global gene expression changes. Furthermore, these changes could be tissue specific and not in the tissues that are currently being used for validation. In order to combat this the GTEx consortium are currently paving the way in identifying eQTLs for 54 different tissues from 948 donors (The Genotype Tissue Expression Consortium, 2019b). Integration of these datasets with already published and publically available data will allow for fine mapping of the causal variant within larger regions of LD in diseased individuals. The primary issue with eQTL identification is that RNA profiles used to generate eQTLs are often from post-mortem tissue, where the interval between death and RNA collection can result in different transcriptomes, or different cell composition. This therefore, would alter the transcriptome and the resultant analyses could be biased, however with current technologies this the best we can do. In addition to the caveats discussed above with eQTLs, identifying the gene or genes that mQTLs control expression of is even more complex due to the 3D nature of DNA. mQTLs could be working *in cis* on the nearest gene or could be working *in trans* on different chromosomes or on genes Mb away. This makes it ever more important to develop methodologies to tackle and validate the complexity of QTL association to disease.

8.2.3. Epigenetic causality

Epigenetic changes in disease have the potential to provide treatment due to their dynamic state, however this dynamism is also detrimental to epigenetic studies. Changes to cytosine modifications could be as a result of disease progression rather than the cause. While this thesis has sought to establish molecular biology tools to functionally validate the direction of causality we are still using risk loci from datasets gathered from post-mortem tissues, where disease has progressed and epigenetic changes are already well defined. Further to this, there are many confounders to epigenetic studies, such as smoking, which can drastically alter DNA methylation. If statistical approaches are not robust enough to detect significant disease specific CpG sites then functional validation will take much longer due to false positives. This is already being improved with fine-mapping techniques to try and identify these causal variants.

As previously discussed, the expression profiles of certain cell types can be different for the same QTL, therefore using the correct cell line is of utmost importance for the translation of these findings. This is further complicated by the use of bulk post-mortem tissue used in epigenetic profiling experiments. EWAS studies in psychiatric illnesses are carried out using heterogeneous tissues from different brain regions, and as discussing in previous chapters, methylation is specific to different cell types. It could be that subtle EWAS findings are being diluted by alterations in different cell type abundance or exacerbated by dramatic increases in rare cell types that we are not able to identify through bulk analysis. In future, EWAS studies should be carried out on FACS or Fluorescence activated nuclei sorting of tissues or nuclei where individual cell types are isolated and characterised separately. This would then allow us to identify the specific cell types which are involved in disease progression and allow for functional follow-up in the relevant cell type. This would then hopefully make the validation more translatable and relevant to that in diseased individuals.

8.2.4. Off target effects of CRISPR/Cas9

Within the CRISPR field there is growing uncertainty around validity of findings presented within CRISPR/Cas9 experiments. This is due to the risk of off target effects introduced by mismatching of the gRNA to non-specific sequences of the genome, resulting in editing and mutations. However the rate of off target discovery is

dependent on the gRNA sequence. Off targets have shown to be dependent on mismatches at the 5' end of the gRNA whereas random point mutations within the gRNA have negligible effect on off targets (Kim *et al.*, 2015). In order to minimise the risk of off target effects in our experiments I used the gRNA algorithm provided by (Hsu *et al.*, 2013) for designing the *AS3MT* gRNA and then the improved (Doench *et al.*, 2016) algorithm provided through benchling for the epigenetic editing gRNA. At the time these were the golden standard of tools for CRISPR gRNA design. I further improved the specificity of the *AS3MT* targeting gRNA by insuring that the gRNA do not have known off-targets within exonic regions. Current work is being undertaken to profile the off-target sites for the *AS3MT* paper, although results from the RNAseq data showed no change in gene expression of genes *in cis* of the off target sites. There are further issues when designing gRNA to intronic regions for epigenetic editing as these are highly repetitive sequences, currently there is no tool perfected for gRNA design for intronic regions. In the last few years newer tools have been developed to both optimise gRNA design (Tycko *et al.*, 2019) and modify Cas9 (Kleinstiver *et al.*, 2016) to reduce the off target effects previously seen. These advances will become more relevant as the CRISPR field moves from singleplex to multiplexing and screening where multiple gRNA are used simultaneously.

8.2.5. Other cytosine modifications and epigenetic marks

In the epigenetic work presented in this thesis is that we have profiled DNA methylation (5mC) whereas discussed in Chapter 1. there are other cytosine modifications present, namely 5hmC. A limitation of the sodium BC used for profiling of DNA methylation is that it cannot distinguish between changes in 5mC and 5hmC. Any 5hmC modification is also protected from deamination when BC and results in an unmodified C which will be interpreted as a 5mC. This is incredibly important as 5hmC is abundant in the brain through development (Spiers *et al.*, 2017). The exact function of 5hmC is still to be elucidated, although it is of no surprise that it is abundant in gene promoters and enhancers, so could play a crucial role in gene regulation or splice expression (Shi *et al.*, 2017). Already 5hmC has been identified to be correlated with age in schizophrenia patients compared to controls in blood samples (Jiang *et al.*, 2017). Recent advances have only just started to be able to profile 5hmC from 5mC so in the coming years there could be more validation to be done on not only 5mC modification but also 5hmC in the brains of schizophrenic patients.

DNA methylation is only one of the major epigenetic markers involved in schizophrenia. This thesis did not divulge into other epigenetic markers such as histone modifications or non-coding RNA in schizophrenia aetiology. It would be naïve to assume that all epigenetic modifications work independently of each other. Examining the effect of genetic variation on histone activity or non-coding RNA would be of interest given that most QTL and GWAS signals occur in regions of gene regulation. This would be relatively easy to combine genomic CRISPR to modify bases and examine the accessibility of chromatin through Assay for Transposase-Accessible Chromatic using sequencing (Buenrostro *et al.*, 2013). Other research should be done to examine how schizophrenia risk loci effect gene enhancers and promoters. While CRIPSR based pipelines have currently been developed to functionally examine enhancer CTCF loops and their role in gene expression and gene control very few have started to analyse the effect of QTLs in CTCF binding and organisation (Tarjan, Flavahan and Bernstein, 2019).

8.2.6. Beyond schizophrenia

The most powerful part of the research presented in this thesis is its translational ability to not only other psychiatric diseases such as dementia, depression, suicide, autism etc. but also to other diseases in general. The methods presented can be easily adopted to different fields, cell, and animal models with relative ease to help with functional validation of (epi)genomic risk loci. There is also considerable overlap with the risk variants for schizophrenia and neuronal development, therefore these research mechanisms can help to elucidate how genetic and epigenetic variation can contribute to normal neurodevelopment. Ultimately the goal would be to profile these individual genetic and epigenetic changes in multiple cell types and models to identify robust changes which can be correlated to disease pathology. From this research I have developed a proposed pipeline for functional validation of GWAS and EWAS risk loci in cellular models.

8.3. Future research directions

The results presented in this PhD thesis represent different methods we can use to functionally validate both GWAS, EWAS, and environmental risks in schizophrenia in cell models, to elucidate a function in disease aetiology. While I have examined genetic knockout of a whole gene it would be of interest to knockout or edit risk SNPs and examine their effect on cellular function. This work is already being carried out on eQTLs identified through GWAS and GTEx datasets (Brandt *et al.*, 2019). Moving from whole gene knockout to SNP editing would be more relevant as we would then be able to identify the causal SNP within regions of LD. Our methods are currently simplex, moving from simplex to multiplex would improve the productivity, this can easily be achieved by modifying forms of CRISPR screening protocols along with advances in single cell sequencing (Dixit *et al.*, 2016; Alda-Catalinas *et al.*, 2019). Further work should be done examining the effect of QTLs in different tissues, in order to do this the pipelines can be used in human derived iPSC lines which can be differentiated into different cell fates. This would then provide comprehensive analysis of cell specific transcript and methylome changes as a response to risk SNPs. Furthermore these iPSC lines can then be used to generate 3D cultures (Sloan *et al.*, 2018) which can then be used as a model of early development. This then allows us to move from a 2D culture of cells to a 3D model and examine how this risk SNPs can impact on neuronal development and migration when surrounded by other cell types required for normal development which are not present in 2D cultures. The protocols can also be adapted to examine how specific DNA methylation changes can result in changes to neuronal development using iPSC and organoid production.

The most important aspect of this thesis is that development of these methods can help us to identify potential causal variants associated with disease and their mechanism of action. If these variants are genomic, it can provide laboratory based screens for diseases or can be used to refine already developed polygenic risk scores to improve diagnostics. If these causal variants result in changes to DNA methylation and we can infer the direction of causality we can then develop targeted approaches to reverse these changes using pharmaceutical approaches due to the dynamic nature of epigenetic modifications. The methods shown in this thesis are the next logical step to follow on from computational and statistical approaches to epigenetic epidemiology. We already might have identified the causal variants in a number of complex diseases

but it is only until now that we have the toolkit to functional validate these variants in cellular models.

8.4. Conclusions

The studies provided in this thesis have provided protocols and methods we can use to functionally elucidate roles for genetic, epigenetic, and environmental risk involvement in the aetiology of schizophrenia. I have demonstrated that CRISPR/Cas9 can be used to identify the function of schizophrenia risk genes. I identified that knockout of *AS3MT* resulted in impairments to neuronal differentiation and development and could be a crucial regulator of neurodevelopment. I have also developed luciferase based mechanisms to infer DNA methylation dependent function to DMRs associated with disease, however we are unable to determine the actual gene (if any) this functionality relates to. To address these issues a CRISPR/Cas9 based method has been developed to edit specific DNA methylation sites *in vitro* to enable DNA methylation induced gene expression to be elucidated. Using this approach I identified that changing DNA methylation results in changes of gene expression but not necessarily in the genes predicted by the statistical analysis (Hannon et al 2018). Finally, I examined how environmental risk factors can result in changes in DNA methylation, and that these methylation changes link to pathways associated with schizophrenia.

However there are many questions which still need to be answered. I have only looked at the resultant DNA methylation changes and their links to gene expression. Looking to the future it would be of importance to edit risk SNPs rather than risk genes to elucidate roles of genetic variants in complex disease, as they could be operating independent of DNA methylation. What is of importance is that we now have the available molecular tools to move from computational models of validating risk loci to laboratory models to validate findings in cell models with relative ease.

Appendix A – Buffer Compositions

Supplementary Table 1: 10x STE Buffer.

Component	Quantity	Final Concentration
NaCl	29.22g	1M
Tris-HCl	6.06g	100mM
EDTA	1.46g	10mM
ddH_2O	500ml	

pH to 8.0

Supplementary Table 2: Slagboom Buffer.

Component	Volume	Final Concentration
10x STE	5ml	100mM NaCl, 10mM Tris, 1mM EDTA
5% SDS	5ml	0.5% SDS
ddH_2O	40ml	

Supplementary Table 3: Lysis Buffer for Protein Extractions.

Reagent	Quantity	Final Concentration
Tris-Base	1.21g	20mM
NaCl	4.38g	150mM
EDTA	0.186g	1.25mM
Triton-X	5ml	1%
ddH_2O	500ml	

Supplementary Table 3: 10x TBS.

Reagent	Quantity	Final Concentration
Tris-Base	24g	200mM
NaCl	88g	1.5M
ddH_2O	1L	

pH to 7.6

Supplementary Table 4: 1x TBST.

Reagent	Volume	Final Concentration
10x TBS	100ml	20mM Tris-Base, 150mM NaCl
10% Tween-20	10ml	0.1%
ddH_2O	890ml	

Appendix B – Genome-wide DNA methylation meta-analysis in the brains of suicide completers

Title Page

Title: Genome-Wide DNA Methylation Meta-analysis in the Brains of Suicide Completers

Stefania Policicchio[1], **Sam Washer[1]**, Joana Viana[1], Artemis Iatrou[5], Joe Burrage[1], Eilis Hannon[1], Gustavo Turecki[2], Zachary Kaminsky[3,4], Jonathan Mill[1,] Emma L Dempster[1*], Therese M Murphy[1,6*§]

[1] University of Exeter Medical School, University of Exeter, Exeter, UK.

[2]Douglas Institute, Department of Psychiatry, McGill University, Verdun, QC H4H 1R3, Canada

[3]Department of Psychiatry, School of Medicine, Johns Hopkins University, Baltimore, MD, USA

[4]Department of Mental Health, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD, USA

[5]Rush Alzheimer's neuroDisease Center, Rush University Medical Center, 600 South Paulina Street, Chicago, IL 60612, USA

[6]School of Biological and Health Sciences, Technological University Dublin, City Campus, Dublin 2, Ireland.

* These authors contributed equally to this work.

[§]Corresponding author

Dr Therese M Murphy (PhD)

School of Biological and Health Sciences,

Technological University Dublin – City Campus

Kevin Street, Dublin 2, D08 X622

Email: therese.murphy@TUDublin.ie

Keywords: DNA methylation, Major Depression, Suicide, Epigenetics, Brain

Abstract

Suicide is the second leading cause of death globally among young people representing a significant global health burden. Although the molecular correlates of suicide remains poorly understood, it has been hypothesised that epigenomic processes may play a role. The objective of this study was to identify suicide-associated DNA methylation changes in the human brain by utilising previously published and unpublished methylomic datasets. We analyzed prefrontal cortex (PFC, n = 211) and cerebellum (CER, n = 114) DNA methylation profiles from suicide completers and non-psychiatric, sudden-death controls, meta-analyzing data from independent cohorts for each brain region separately. We report evidence for altered DNA methylation at several genetic loci in suicide cases compared to controls in both brain regions with suicide-associated differentially methylated positions enriched among functional pathways relevant to psychiatric phenotypes and suicidality, including nervous system development (PFC) and regulation of long-term synaptic depression (CER). In addition, we examined the functional consequences of variable DNA methylation within a PFC suicide-associated differentially methylated region (*PSORS1C3 DMR*) using a dual luciferase assay and examined expression of nearby genes. DNA methylation within this region was associated with decreased expression of firefly luciferase but was not associated with expression of nearby genes, *PSORS1C3* and *POU5F1.* Our data suggest that suicide is associated with DNA methylation, offering novel insights into the molecular pathology associated with suicidality.

Introduction

Suicide represents a global public health problem, with approximately 800,000 people dying worldwide from suicide annually and suicide attempts up to 20 times more frequent than completed suicide [1]. Moreover, suicide is the second leading cause of death among young people worldwide and ranks amongst the 20th leading causes of death across all ages [2]. The risk for suicidal acts is multifactorial, and consists of a range of biological, psychiatric, psychosocial and cultural risk factors [3]. Despite its economic and social burden, the underlying biological aetiology of suicidal behaviour (SB) remains poorly understood.

To date, large-scale genome-wide association studies (GWAS) [4-7] have failed to identify robust associations suggesting that the risk of SB is highly polygenic in nature and that individual gene variants are likely to account only for a small proportion of the total phenotypic variability [8]. Other factors, such as the environment, behavioural traits, psychiatric diagnosis, lifestyle and coping mechanisms, are essential regulators of suicide risk and likely to account for more sizeable effects [9]. Recently, increased understanding of epigenetic processes that occur in

the brain has opened promising avenues in suicide research. The epigenome is potentially malleable—changing with age [10], in response to specific environmental [11] and psychosocial factors [12] —providing a mechanism for the interaction between genotype and the environment [13]. Epigenetic processes, including DNA methylation, have recently been implicated in the aetiology of numerous mental health disorders [14-21] and SB [22, 23].

In the last decade, research aiming to understand the contribution of epigenetic mechanisms to SB has implicated the role for key biological pathways, including hypothalamic pituitary adrenal (HPA) axis, stress response, polyamine system, neurotrophic signalling and lipid metabolism [8]. However, studies examining DNA methylation differences associated with SB have primarily focussed on candidate genes [24-27] and few have examined genome-wide DNA methylation changes in the brains of suicide completers [22, 28]. The availability of brain samples is a major challenge for psychiatric research and many previous studies examining DNA methylation variation in suicide are performed on a limited number of post-mortem brain samples [22, 29, 30]. Such small studies have reduced statistical power to detect small changes in DNA methylation. The objective of this study was to identify suicide-associated DNA methylation changes in the human brain by utilising previously published and unpublished methylomic datasets.

Genome-wide DNA methylation profiles were available from post-mortem brain samples of suicide completers and non-psychiatric, sudden-death controls for a total of 7 cohorts. Methylomic data available for two different brain regions - Prefrontal cortex (PFC) and Cerebellum (CER), (PFC: 4 cohorts, n = 211; CER: 3 cohorts, n= 114) - were meta-analysed across the suicide cohorts for each brain region separately. We report evidence for altered DNA methylation in suicide cases compared to non-psychiatric controls in both the PFC and CER and examined the functional implications of a top-ranked PFC suicide-associated differentially methylated region (DMR) on gene expression levels in that region. Finally, gene ontology enrichment analysis was performed in each brain region separately to identify pathways of genes associated with suicide completion.

Materials and Methods

Sample collection / data recruitment

For the PFC meta-analysis, we included 4 independent previously published studies [22, 31-33] aimed at profiling DNA methylation in human PFC in individuals with a diagnosed axis I psychiatric disorder and healthy non-psychiatric controls. Only data from individuals who died by suicide and non-psychiatric controls were included for the initial meta-analysis. In two of the four studies selected [31, 32], DNA methylation was profiled from fluorescence-activated

nuclei sorted (FANS) neurones, with the remaining two studies performed in bulk tissue [22, 33]. Raw DNA methylation data for all 4 studies are deposited in the Gene Expression Omnibus (GEO) database (accession number: GSE89707, GSE88890, GSE98203, GSE41826) and full details of the sample cohort can be obtained from the original studies [22, 31-33]. For the CER meta-analysis, three DNA methylation datasets were included two of which are currently unpublished (GSE137222 and GSE137223). Raw DNA methylation data for the CER published EWAS study [33] is deposited in GEO database (accession number: GSE89702). The unpublished studies were approved by the University of Exeter Medical School Research Ethics Board (REB). In all three CER cohort's DNA methylation profiles were derived from bulk brain tissue and cases were individuals who died by suicide (hanging, jumping from height, intentional poisoning, self-harm/bleeding). Cause and manner of death as well as joint presence of psychiatric diagnosis were determined by a forensic pathologist after evaluating autopsy results, circumstances of death, data from extensive toxicological testing, police reports, family interviews and medical records. Controls were individuals who died suddenly (e.g. cardiac failure, viral infection, or accidents) and did not have evidence of axis I disorders. See ***Supplementary Table S1*** for a complete description of sample selection, numbers and demographic characteristics of each cohort.

DNA Methylation analysis

DNA methylation was measured using the Illumina HumanMethylation450K BeadChip ("Illumina 450K array") or Infinium MethylationEPIC BeadChip ("Illumina EPIC array", one CER cohort) platform (Illumina Inc., San Diego, CA, USA). To ensure consistency of the methodological approach, raw DNA methylation data (idat files) were recovered and each cohort was independently reanalysed, applying the same quality control (QC) and pre-processing pipelines. Briefly, QC checks, quantile normalization, and separate background adjustment of methylated and unmethylated intensities of type I and II probes were employed using the wateRmelon package in R[34]. Probes on the X- and Y-chromosomes were used to confirm sample sex. Only samples which passed stringent QC measures (>1% of sites with a detection *P* value (*P*) >0.01) were included. Probes with a detection *P* >0.01 in at least 1% of samples and/or a beadcount <3 in 5% of samples, non-specific probes, potentially cross-reactive probes or probes near SNPs [35, 36] were removed across all samples. Only probes common to both the 450k Array and EPIC array were included in downstream analyses for the CER. For the annotation of probes, the University of California, Santa Cruz (UCSC) RefGene name from Illumina's annotation file and enhanced annotation to the UCSC Known Gene were used. All annotations use the human February 2009 (GRCh37/hg19) assembly.

Estimating differential neuronal proportions

The R package (available at www.cran.r-project.org), Cell EpigenoType Specific (CETS) mapper, designed for the quantification and normalization of differing neuronal proportions in genome-wide DNA methylation datasets was used as previously described [32] to estimate brain cellular heterogeneity in each of the 4 PFC cohorts. Similar estimates could not be obtained for the CER cohorts as the algorithm for the correction of brain cellular heterogeneity bias was developed using post-mortem frontal cortex data and NeuN is not expressed in CER purkinje neurons[32].

Data analysis

Statistical analyses were performed using R statistical package (version 3.4.3). The β value is a ratio between methylated probe intensity and total probe intensities (sum of methylated and unmethylated probe intensities) and ranges from 0 to 1. Linear regression was used to examine differences in DNA methylation scores (reported as change in β value (Δβ)) between suicide cases and controls at each CpG site, controlling for potential confounders. Covariates included in all models were: age, sex and chip. We also included ethnicity or brain bank as covariates in the model for those cohorts where that information was available and represented a potential source of variation. In the PFC cohorts only, we also adjusted for estimated neuronal proportions. For one study[22], DNA methylation differences were investigated across individual matched cortical regions (Brodmann area 11 (BA11), Brodmann area 25 (BA25)) by fitting a linear mixed-effect model using the lme4 R package (available at https://cran.r-project.org [37]). Whereby, brain region and sample ID were included in the model as random effects ('within participants' factors) while diagnosis, age, sex, PH and cellular composition were included in the model as fixed effects.

Meta-analyses

Suicide completers versus non-psychiatric controls

The results obtained from the linear regression were then meta-analysed for each brain region independently. A fixed-effect model, using the 'metagen' function in the R package 'meta', was applied by providing the regression coefficients and standard errors from each individual cohort to calculate weighted pooled estimates and to test for significance. Experiment-wide significance (*P* <1E-07) (threshold estimated from permutation analysis in a larger dataset (N= 675 individuals) generated previously by our group [38]) was chosen as a multiple testing threshold to determine statistically significant DNA methylation changes.

Suicide completers versus non-suicide psychiatric controls

In order to assess whether the observed suicide-associated DNA methylation changes identified in our original meta-analysis were driven by the psychiatric disorder comorbidity rather than being suicide-specific changes, a second exploratory analysis was performed in additional samples obtained from the CER datasets only, which had additional non-suicide psychiatric samples with DNA methylation data available (not included in the primary analysis report here). In the second meta-analysis, each CER cohort consisted of suicide cases (that were included in the original meta-analysis) and psychiatric controls, where individuals had a diagnosed axis I disorders (Major Depressive Disorders (MDD), Schizophrenia (SZ), Bipolar disorder (BD)) but had no documented evidence of suicidal behaviour. In total, the secondary meta-analysis included 130 samples (case group, N =50; psychiatric control group, N =80). Results obtained from the linear regression were then meta-analysed using a fixed-effect model as described previously.

Region based analysis

The results obtained from both the PFC and CER meta-analyses were used to perform a regional-based analysis using the Python module Comb-p [39], to identify suicide-associated DMRs. The Comb-p software groups spatially correlated DMPs (seed $P < 1E-03$, minimum of 3 probes) at a maximum distance of 500 bp in each brain region. DMR *P* were corrected for multiple testing using Šidák correction [40].

Gene ontology term enrichment analysis

A previously described logistic regression approach [41] was used to test if genes (Illumina UCSC gene annotation) annotated to probes in our PFC and CER meta-analyses (DMPs with $P \leq 1E-04$) predicted pathway membership, while controlling for the number of probes annotated to each gene. Briefly, pathways were downloaded from the Gene Ontology (GO) website (http://geneontology.org/) and all genes annotated to parent terms were also included. Genes containing at least one Illumina probe and annotated to at least one GO pathway were considered. Pathways were filtered to those containing between 10 and 2,000 genes and a list of significant (after correction for multiple testing - Bonferroni correction) pathways were identified as previously described [41].

Functional follow-up of significant DNA methylation findings

Tissue (N=71) from two regions of the cortex, BA11 (N=38) and BA25 (N=33), collected from 20 MDD suicide cases and 20 non-psychiatric sudden death controls was obtained from the Douglas Bell-Canada Brain Bank (DBCBB) (http://douglasbrainbank.ca/), further details are available in [22]. Previously, our group performed DNA methylation profiling in these samples[22]

and the results of that study were included in this meta-analysis study. To examine whether our identified suicide-associated DMR (*PSORS1C3 DMR;* Chr6:31,148,370-31,148,553 (Hg19)*, 2694 bp downstream the TSS of PSORS1C3 gene)* is associated with the expression of nearby genes, we measured expression levels of two nearby genes in these brain tissue samples. We tested for an association with gene expression firstly at the closest transcription start site (TSS) gene - the lncRNA gene*, PSORS1C3* - and then at the second closest gene*, POU5F1*.

Gene expression analysis

Thirty mg of frozen PFC tissue from each brain sample was homogenized with Qiazol Lysis Reagent (Qiagen, Valencia, CA, USA), as per the manufacturer's instructions, before running it through a QIAshredder (Qiagen, Valencia, CA, USA). Total RNA was extracted using the Qiagen miRNeasy Mini column-purification system and treated with DNase I as outlined by the manufacturer. The Agilent 2100 Bioanalyzer was used to check the quality and concentration of the extracted RNA samples. 1µg of total RNA was reverse transcribed into complementary DNA (cDNA) (20-µL reactions) according to the manufacturer's instructions using the Invitrogen VILO cDNA synthesis kit (Life Technologies Ltd, Paisley, UK). Three housekeeping genes Ubiquitin Conjugating Enzyme E2 D2 (*UBE2D2*), Cytochrome C1 (*CYC1*), Ribosomal Protein L13 (*RPL13)* identified previously [42] as being among the most stably expressed in the brain were selected to normalise the target gene expression. Next, quantitative RT-PCR was performed in triplicate for each assay using the StepOnePlus Real-Time PCR machine (Applied Biosystems, Foster City, Calif) and pre-optimized Taqman gene expression assays (Applied Biosystems, Foster City, Calif). A full list of the qPCR assays used is given in ***Supplementary Table S2***. PCR cycling conditions were as follows: 50°C for 2 min, 95°C for 20 sec, and 40 cycles of 95°C for 10 sec and 60°C for 20 sec. We undertook stringent QC of raw qPCR data, repeating samples where there was high variability between triplicates (Ct > 0.5). The abundance of each test gene was determined by the comparative Ct method [43], expressed relative to the geometric mean of the three housekeeping genes. Data were log2-transformed to ensure normal distribution and presented as a fold-difference in expression of suicide cases relative to controls using the $2^{-\Delta\Delta CT}$ method. To assess whether *POU5F1* expression levels were associated with a history of suicide, we used a linear mixed-effect model (LMM) using the lme4 R package (available at https://cran.r-project.org) [37] where ΔCt values of the target gene (*POU5F1*) was the response variable. Brain region and sample ID were included in the model as random effects ('within participants' factors) while diagnosis, age, sex and neuronal proportion were included in the model as fixed effects. Finally, since 450K array data were available from the same individuals, we examined the correlation between gene expression levels and mean DNA methylation levels at the DMR.

Reporter constructs

The *PSORS1C3 DMR* sequence was inserted into the pCpGL-basic vector (see [44] for details), which is devoid of CpG sites and was generously provided by the Rehli laboratory [44]. Briefly, the cleaned *PSORS1C3* PCR amplicon was inserted into a digested pCpGL-basic plasmid using T4 ligase and buffer (Invitrogen, California, USA). Ligated plasmids were transformed into One Shot PIR1 *E.coli* (ThermoFisher Scientific, Massachusetts, USA) to allow for monoclonal amplification of the recombinant plasmids. For the transformation, 50µl of One Shot PIR1 *E.coli* were used, including a negative ligation control and a positive transformation control (pUC19) using standard procedures.

Clones were subsequently checked by clonal PCR, restriction digest using BglII and NcoI and Sanger sequencing (see ***Supplementary Figure S1***) to confirm the DMR had been inserted in the correct orientation. The pCpGL construct was methylated *in vitro* using M.sssI methyltransferase (New England Biolabs, Massachusetts, USA) in the presence of S-adenosylmethionine (SAM) following the manufacture's protocol. An empty pCpGL-basic vector was also methylated to act as a control. To confirm successful methylation, the plasmids underwent digestion with the methylation sensitive enzyme HpaII (see ***Supplementary Figure S2***).

Cell culture and transfections in HEK293 cells were cultured in Dulbecco's modified Eagle's medium (þ4.5 g l_1 D-glucose, L-glutamate, pyruvate) (Gibco) with 10% fetal bovine serum (Gibco) at 37°C and 5% CO_2. Briefly, 2 x 10^5 cells were seeded in six well plates. The following day media was removed, the cells washed with PBS and 1.5ml of fresh growth media was added. 500ng of PSORS1C3 or pCpGL plasmid and 100ng of pGL4.74[hRluc/TK] reporter control vector (Promega, Wisconsin, USA) were diluted in 500µl of Opti-MEM reduced serum media (Gibco, Massachusetts, USA) in an Eppendorf and left to equilibrate for five minutes at room temperature. 4.5µl of Lipofectamine LTX Reagent (ThermoFisher Scientific, Massachusetts, USA) was then added to each Eppendorf and incubated for 30 minutes at room temperature. Following incubation, 500µl of LTX plasmid mix was added to the HEK293 cells in six well plates. Cells were then incubated at 37°C, 5% CO_2, for 24 hours to allow expression of firefly and *Renilla* luciferases.

Dual Luciferase Assay

24 hours following transfection a dual-luciferase reporter assay (Promega, Wisconsin, USA) was carried out to measure the expression of firefly luciferase and *Renilla* luciferase in the transfected cells as per manufacturers' instructions. Each experiment contained three technical repeats and the experiment was repeated three times. The injections and light absorbance were carried out automatically using the pherastar plate reader. The average

firefly luciferase activity was calculated by averaging absorbance readings between 2 and 10 seconds. The average *Renilla* luciferase activity was calculated by averaging absorbance readings between 14 and 22 seconds. Data analysis was carried out as described in [45]. All data is presented as a normalised firefly luciferase activity relative to *Renilla* luciferase. Fold change expression was calculated by dividing the unmethylated normalised firefly luciferase activity by the respective methylated normalised luciferase activity. A student's T-test was used to compare the methylated versus unmethylated vectors.

Results

Suicide-associated DMPs in human cortex and cerebellum

An overview of the methodological approach used in this study is given in ***Supplementary Figure S3***. We identified one DMP (cg00963169) in the PFC, which reached experiment-wide significance (*P* =3.30E-08 (***Figure 1A***). The effect size at this CpG site, located downstream of exon 1 of the neuron-specific protein coding gene, ELAV-like RNA binding protein 4 (*ELAVL4),* was largely consistent across all cohorts included (see **Supplementary *Figure S4***), showing hypomethylation in suicide cases relative to controls. Interestingly, the 20 most significant (*P* <5E-05) suicide-associated differentially methylated loci identified in the PFC, listed in ***Supplementary Table S3***, include probes in the vicinity of several loci previously implicated in psychiatric phenotypes. In the CER, 6 probes (cg14392966, cg17855963, cg25590492, cg12284382, cg10757978, cg04525580) passed the experiment-wide significance threshold (*P* <1E-07) (***Figure 1B***). Of interest the top-ranked DMP, cg14392966 (*P* =3.06E-11), which is located within the coding region (exon1) of the *PUS3* gene on Chr11, has been previously associated with severe neurodevelopmental disorders [46]. **Supplementary *Figure S5*** shows that this DMP is hypomethylated in two of the 3 cohorts in suicide cases compared to healthy controls. A list of the top 20 DMPs in the CER is provided in ***Supplementary Table S4.***

Region-based analysis of altered DNA methylation in suicide completers

We used the python module, Comb-p [39], to identify DMRs in suicide cases compared to controls in each brain region. The regional analysis identified 3 and 8 significant (Sidak-corrected *P* <0.05) DMRs in the PFC and CER, respectively (See ***Table 1*** for details). In the PFC, the top-ranked DMR was found within the *WRB* gene (Sidak-corrected *P* =5.11E-06) and was consistently hypomethylated across all 5 CpG sites in suicide cases relative to controls, in all 4 PFC methylomic studies (***Figure 2***). Of interest, the second top-ranked suicide-associated DMR in the PFC (Sidak-corrected *P* = 3.81E-05) was located downstream

of the promoter region of the *PSORS1C3* non-coding gene, a DMR previously reported by our group as associated with MDD suicide completers [22].

In the CER, the top-ranked suicide-associated DMR was identified on chromosome 22, distributed along the intronic region of the *CERC2* gene and spanning 4 CpG sites. The *CERC2*-associated DMR (***Figure 3***) showed significant hypermethylation (Sidak-corrected P =5.68E-07) across all 4 CpG sites within the region in suicide cases compared with controls. The direction of this change was found to be consistent across all 3 CER methylomic studies.

Pathway Analysis

The biological relevance of our findings was investigated through gene ontology analysis on genes annotated to suicide-associated DMPs (*P* ≤1E-04). Results revealed an enrichment of DNA methylation alterations in genes involved in cognitive processes such as long term synaptic depression and brain development (See ***Supplementary Table S5 and S6***).

Suicide–associated DMPs identified in the CER are largely independent of comorbid psychiatric disorders.

In order to disentangle the relative contribution of psychopathologies from DNA methylation changes specific to suicide, we performed an additional meta-analysis in the CER whereby all the non-psychiatric controls initially included in our meta-analysis were excluded and individuals with documented Axis I psychiatric disorders (MDD, SZ, BD) and who died by suicide were compared to psychiatric cases without a documented history of suicidal behaviour/suicide fatalities (3 cohorts, N=130, cases N=50, controls N=80). Comparison of the results (effect sizes at the top 500 (*P* < 0.05) DMPs) from our original CER meta-analysis and the secondary analysis (***Supplementary Figure S6***) revealed a strong positive correlation (*P*=2.2E-16; R=0.89). Moreover, 2 of the 6 CER-associated DMPs (P <1 E-07) were nominally significantly differentially methylated in our suicide cases versus psychiatric controls analysis (cg10757978, P=1.19E-04; cg04525580, P=0.017) (see ***Supplementary Table S7*** for details) and a similar direction of effect was observed for both analyses for the remaining 4 CER-associated DMPs. We were unable to perform the same analysis in the PFC due to lack of samples from individuals with an axis-I diagnosis who did not die by suicide.

Functional validation and gene expression analysis of the PSORS1C3 DMR

Given our replication of a *PSORS1C3 DMR* in suicide [22] we aimed to functionally evaluate its effect on the expression of nearby genes. First, we examined the effect of DNA methylation at the suicide-associated DMR on nearby gene expression using a CpG-Free Luciferase Reporter (pCpGL vector) gene assay [47]. Next, we examined gene expression levels of nearby annotated genes (*PSORS1C3* and *POU5F1)* in a subset of samples for which brain tissue was available (two brain regions (BA11 and BA25, N=70; suicide cases (N=36), non-

psychiatric controls (N=34)) and examined the correlation between expression levels of our selected target genes and mean DNA methylation at the suicide-associated DMR.

We found a marked increase in the relative expression of firefly luciferase activity normalised to *Renilla luciferase* in the unmethylated *PSORS1C3* cloned pCpGL vector compared to the methylated vector (***Figure 4***) (fold change = 206, P = 0.006, N=3). Next, we quantified gene expression levels of the *PSORS1C3* long non-coding gene in a cohort of 70 post-mortem brain samples (BA11, N=38; BA25, N=32, both regions obtained from the same individuals). Our analysis showed that in our sample set, *PSORS1C3* lncRNA was not expressed in the PFC (Ct Value >31 or undetermined). This result is consistent with findings in Genotype-Tissue Expression (GTEx) portal[48] (https://www.gtexportal.org/home/gene/PSORS1C3), which shows little to no expression for this gene in brain samples examined. Next, we examined expression levels of the second closest gene (*POU5F1*) to the suicide-associated DMR. A linear mixed-effect model was used to compare mean dCt values between suicide cases and non-psychiatric controls and the analysis showed no significant difference in gene expression levels between the two groups (*P* = 0.598; ***Supplementary Figure S7A)***. Furthermore, we found no significant correlation between mean DNA methylation at the suicide-associated region *PSORS1C3* DMR and *POU5F1* gene expression levels (Pearson's R =-0.04, *P* =0.67; ***Supplementary Figure S7B***).

Discussion

In this study, we utilised previously published and unpublished methylomic datasets to perform a meta-analysis of variable DNA methylation in the brain of suicide completers. DNA methylation data were available for two different brain regions (PFC: 4 cohorts, N = 211; CER: 3 cohorts, N= 114) and data were meta-analysed across the suicide cohorts for each brain region separately. To our knowledge, this represents the most extensive methylomic study of suicide completers using post-mortem brain tissue to date.

While several studies suggest the involvement of the PFC in suicidal behaviour [22, 49-51]), suicide associated epigenetic changes in the CER have not been investigated. However, the CER is known to play an important role in motor control, cognition, and emotional processing and is involved in a variety of psychiatric disorders, including depression, bipolar disorder, and schizophrenia[52]. Suicidal behaviour in those with depression has been associated with a decreased cerebellar volume[53] and low regional cerebral blood flow in the cerebellum[54]. Moreover imaging studies have reported structural abnormalities associated to suicide attempt in MDD in cortical and subcortical regions, including cerebellum[55] and alterations in functional cerebellum networks were found in depressed patients with a suicide attempt history[56]. Altogether these studies support the hypothesis of a potential involvement of CER in the

psychopathology of attempted suicide in patients with MDD. We set out to further explore the role of DNA methylation and suicidality in this region as well as to determine if suicidality associated DNA methylation changes are brain region specific.

We first examined site-specific genome-wide patterns of DNA methylation in suicide cases compared with controls in the PFC and CER separately. We identified one DMP (*cg00963169, P* = 3.30E-08) in the PFC, which reached experiment-wide significance. This probe, located in the intronic region of the *ELAVL4* gene, shows consistent hypomethylation in suicide cases compared to controls. The *ELAVL4* gene has a known role in translation and stabilization of mRNA, especially in the brain, and acts as a negative regulator of proliferation, activity and differentiation in neural stem cells [57, 58]. Through their mRNA stabilizing activities, this family of proteins modulate neuronal development and maintenance, and their altered activity has been implicated in neurological conditions [59] and disorders including Alzheimer's disease [60] schizophrenia [61] and autism [62]. Of interest, the association with suicide at this site was found to be largely driven by PFC cohorts derived from neuronal nuclei (see ***Supplementary Figure S4***) and thus future replication of this finding in sorted neuronal cells may yield more significant associations with suicide.

In the CER, we identified 6 probes (cg14392966, cg17855963, cg25590492, cg12284382, cg10757978, cg04525580) at experiment-wide significance threshold (*P* < 1E-07). The top-ranked DMP, cg14392966 (*P* = 3.06E-11), is located in exon 1 of the *PUS3* gene. *PUS3* encodes a highly conserved enzyme responsible for post-transcriptional modification of tRNA and has previously been associated with intellectual disability [63] and severe neurodevelopmental disorders [46]. The remaining DMPs include probes in the vicinity of several loci previously implicated in pathways relevant to psychiatric phenotypes. For example, *ZIC1* is thought to play an important role in neurogenesis and cerebellum differentiation [64], whereas *RASD2* is known to modulate dopaminergic neurotransmission [65]. Furthermore, the probe cg04525580 *(P* = 9.08E-08*),* is located at the 5'UTR of the interferon regulatory factor 2 (*IRF2)* gene. This locus plays an important role in transcriptional activation at promoters [66] and regulates the expression of a variety of genes involved in immune responses in the brain [67], further supporting a role for immune-related pathways in suicide.

To increase the power of our study to identify changes in DNA methylation between cases and controls and given that DNA methylation at adjacent probes is often correlated, we employed the regional-based analysis, Comb-p, to identify DMRs. Our analysis identified 3 and 8 significant (Sidak-corrected *P* <0.05) DMRs in the PFC and CER, respectively. In the PFC, the top-ranked DMR, located in intron 1 of the *WRB* gene is consistently hypomethylated across all 5 CpG sites in suicide cases relative to controls, in all cohorts. Recent studies suggest a role for Wrb in photoreceptor synaptic transmission in zebrafish [68] and the *WRB*

locus was reported among the differentially expressed genes in a mouse model study looking at cognitive impairment and neuropathology in Down syndrome brain [69] further supporting the hypothesis of its involvement in the correct development and functioning of the CNS.

The second top-ranked suicide-associated DMR in the PFC (Sidak-corrected *P* = 3.81E-05) was located downstream the promoter region of the *PSORS1C3* non-coding gene, a DMR previously reported by our group as associated with MDD suicide completers [22]. Although the function of this gene product remains unclear, it is thought to be a potential regulator of nearby immune-related genes [70] and is a known risk gene for psoriasis [71, 72], supporting a role in immune system regulation. To gain further insight into the role of this suicide-associated *PSORS1C3 DMR* on nearby gene expression we first used a dual luciferase assay to determine if DNA methylation at this region decreases expression of firefly luciferase in a CpG free vector. The methylated *PSORS1C3 DMR* construct was significantly associated with decreased expression of the reporter gene product, indicating that the methylation status of this DMR has the potential to modify promoter activity, however the identity of the modified gene product is unknown. We found no evidence of suicide-associated differential gene expression of nearby genes, *PSORS1C3* and *POU5F1* in the PFC. We hypothesize that DNA methylation changes at this suicide-associated DMR are associated with a different nearby gene or an unknown splice variant of either the *PSORS1C3* or *POU5F1* genes. Finally, an additional DMR was identified in the PFC; located on chromosome 22, in the promoter region of the *LGALS1* gene. This gene is thought to play a role in immune system functioning [65] and DNA methylation changes at this locus have previously been implicated in schizophrenia [33].

In the CER, the top-ranked suicide-associated DMR, located in the intronic region of the *CERC2* gene on chromosome 22, spans 4 CpG sites. This locus is known to be involved in the control of the periodic oscillation of cyclin E expression in proliferating cells likely through its histone deacetylase activity [73]. The *CERC2*-associated DMR showed significant hypermethylation (Sidak-corrected P =5.68E-07) across all 4 CpG sites within the region in suicide cases compared with controls and the direction of this change was found to be consistent across all 3 independent CER methylomic studies. To the best of our knowledge, this gene has not been previously implicated in the pathology of SB. Seven additional suicide-associated DMRs were identified in the CER. Of interest is the DMR located in exon 10 of the *SLC44A4* gene; a gene recently implicated in a study looking at the role of the major histocompatibility complex region in schizophrenia susceptibility [74]. An additional suicide-associated DMR, worthy of further investigation, is located in exon 3 of the *WWTR1* gene, a transcriptional coactivator known for its role in preserving neuronal health [75]. Furthermore, a missense variant in this gene, was recently associated with lower cognitive ability in a GWAS study for infant mental and motor ability [76]. Finally, we identified a suicide-associated DMR

located downstream the promoter region of *MED13L* gene and genetic variants at this locus have been widely reported as associated with intellectual disability [77, 78] suggesting that this gene may play an important role in neurological development.

Since SB is often a complication of a psychiatric disorder, distinguishing suicide diathesis-related DNA methylation changes from those associated with mood disorders and other psychiatric disorders has remained a challenge. In order to unravel the relative contribution of psychopathologies from DNA methylation changes specific to suicide, we performed an additional meta-analysis in the CER whereby individuals with documented Axis I psychiatric disorders (MDD, SZ, BD) and who died by suicide, were compared to psychiatric cases without a documented history of SB. Comparison of the results (effect size of the top 500 nominally significant (P<0.05) DMPs) from our original CER meta-analysis and the secondary analysis revealed a strong positive correlation (P=2.2E-16; R=0.89). Moreover, we replicated our findings for 2 of the 6 CER-associated DMPs, which reached multiple testing threshold in our suicide cases versus psychiatric controls meta-analysis (cg10757978, P=1.19E-04; cg04525580, P=0.017) and we observed similar direction of effect for the remaining loci. Taken together these findings suggest that suicide-associated DMPs identified in the CER are largely independent of comorbid psychiatric disorders. Unfortunately, we were unable to perform the same analysis in the PFC due to the limited number of samples that did not die by suicide but had an Axis-I diagnosis.

Despite the power of the methodological approaches used in this study, there are several caveats. First, the modest number of studies included made this meta-analysis relatively underpowered to detect small changes in DNA methylation. Despite this we were able to identify several statistically significant DMPs and DMRs in both brain regions. Another major limitation is that bulk brain tissue was used in most of the studies included in our meta-analysis and cellular heterogeneity is a well-known confounder in DNA methylation studies. In order to bypass this issue, we used a previously reported *in silico* method to estimate the neuronal proportion in each sample in bulk PFC cohorts and included these estimates in the statistical models [32]. This method could not be applied to our analysis of the CER and thus it is plausible that cellular heterogeneity is confounding some of our CER results. Third, recent research has implicated the importance of other DNA modifications (i.e., 5-hydroxymethyl cytosine) in the brain [79]. Our measure of DNA methylation in this study cannot be distinguished from 5-hydroxymethyl cytosine (5hmC). Of interest, we examined the presence of detectable 5hmC levels at statistical significant DMPs identified in this study in their respective brain regions using the Hydroxymethylation Annotation in Brain Integrative Tool (HABIT) tool (http://epigenetics.iop.kcl.ac.uk/HMC/ [41]). This tool identified detectable levels of 5hmC at the

following suicide-associated DMPs cg00963169 (PFC) and cg17855963, cg04525580 (CER) suggesting that the majority of DMPs identified in this study are not confounded by 5hmC. However, future studies should attempt to examine the role of 5hmC in SB. Fourth, medication data, smoking information and method of suicide were not available for all individuals; thus, we cannot rule out the possibility that the observed DNA methylation changes are influenced by these potential confounders.

Fifth, we acknowledge the possibility that many of the associations reported (DMPs/DMRs) could be related to the severity and/or duration of the mental health disorder. From our secondary analysis in the cerebellum we show that for certain top-ranked DMPs the association appears to be suicide associated rather than associated with the underlying mental health disorder. Given the lack of information related to severity and/or duration of mental illness for samples included in this meta-analysis we can't rule out the contribution of the above mentioned confounders.

Finally, although our study presents evidence for novel DNA methylation changes associated with suicide, further replication using a larger sample size is required to support these results. In addition, future studies could also examine the transcriptional consequences of the observed DNA methylation changes at the *PSORS1C3 DMR* on additional nearby genes and/or novel splice variants in the region. There is considerable interest in using DNA methylation based biomarkers as predictors for suicide risk and previous studies[30, 80-84] have identified polymorphic CpGs that can act as a unique molecular signature for suicide prediction. The data from this study provide many more candidate regions as potential biomarkers for suicide risk and also identifies genes/networks potentially dysregulated in suicidal brain.

In summary, our data, which utilises several published and unpublished suicide cohorts, have identified DMPs and several DMRs associated with suicide in both the PFC and CER, including the previously identified DMR upstream of the *PSORS1C3* non-coding gene. We show that this DMR can influence gene expression using a dual luciferase assay, but we have yet to identify its target gene.

Acknowledgements

We are grateful to all the patients and control subjects who contributed to this study. The authors would like to acknowledge support of the Brain and Behaviour Research Foundation through a NARSAD Young Investigator Grant to TMM and from the UK Medical Research Council (MRC) (grant number MR/K013807/1) to JM. ZK would like to acknowledge funding from the NIH grant (NIMH 1R21MH094771). The Douglas Bell Canada Brain Bank is supported by the FRQS through the Quebec Network on Suicide, Mood Disorders and Related

Disorders, and by Brain Canada through an infrastructure grant. ELD would like to acknowledge the AMS Springboard scheme that supported SW studentship

Conflict of Interest

The authors declare no conflict of interest.

Figure Legends

Figure 1. Suicide-associated DMPs in human cortex and cerebellum. Manhattan plot showing site-specific genome wide pattern of DNA methylation in **A)** the human Prefrontal Cortex (PFC) and **B)** the human Cerebellum (CER). One CpG site was identified as differentially methylated between suicide cases and healthy controls at experiment-wide significant (*P*=1E-07) in the PFC. Six CpG sites were identified as differentially methylated between suicide cases and healthy controls at experiment-wide significant (*P*=1E-07) in the CER.

Figure 2. Suicide-associated differentially methylated region (DMR) in human Prefrontal cortex (PFC). Plot showing the top-ranked DMR in the PFC. This DMR, spanning 5 CpG sites and located in the promoter region of the *WRB* gene (Sidak corrected *P* = 5.11 E-06), was found consistently hypomethylated in suicide cases compared to healthy controls, across all 4 suicide brain cohorts. The solid line is for illustration purposes and not indicative that the CpG sites between sites are also methylated.

Figure 3. Suicide-associated differentially methylated region (DMR) in human Cerebellum (CER). Plot showing the top-ranked DMR in the CER. This DMR, spanning 4 CpG sites and located within the coding region of the *CECR2* gene (Sidak corrected *P* = 5.68 E-07), was found consistently hypermethylated in suicide cases compared to healthy controls, across 3 suicide brain cohorts. The solid line is for illustration purposes and not indicative that the CpG sites between sites are also methylated.

Figure 4: Methylation of the *PSORS1C3* cloned pCpGL vector significantly reduces luciferase activity. There is a significant increase in firefly luciferase expression when normalised to Renilla luciferase, the normalised expression ratios of firefly activity increases to 3.81 when unmethylated versus 0.0269 when methylated *in vitro* (students t-test, *P*=0.0057, n=3) which corresponds to a 206-fold change increase in luciferase activity in the unmethylated *PSORS1C3* cloned vector compared to the methylated identical vector. N=3 experimental repeats with each experiment containing three technical repeats. Error bars represent +/- the standard error.

***Table 1*. Comb-p differentially methylated region (DMR) analysis**

Brain region	Hg19	Annotated gene (UCSC)	No. Of probes	Slk *P value*	Sidak *P value*
PFC	chr21:40759534-40759695	*WRB*	5	2.02E-09	5.11E-06
	chr6:31148370-31148553	*PSORS1C3*	10	1.71E-08	3.81E-05
	chr22:38071168-38071189	*LGALS1*	3	3.37E-08	0.0006529
CER	chr22:17956453-17956561	*CECR2*	4	1.55E-10	5.68E-07
	chrX:79590789-79590956	*CHMP1B2P*	4	3.74E-09	8.84E-06
	chr13:99100506-99100587	*FARP1*	3	2.68E-09	1.31E-05
	chr6:31838402-31838529	*SLC44A4*	5	2.68E-08	8.32E-05
	chr3:149374761-149374915	*WWTR1*	3	7.56E-08	0.0001938
	chr12:116756805-116756949	*MED13L*	3	8.64E-08	0.0002369
	chr1:1846046-1846155	*CALML6*	3	9.40E-08	0.0003406
	chr11:2397486-2397686	*CD81-AS1*	4	2.10E-06	0.004138

Abbreviations: CER, Cerebellum; PFC, prefrontal cortex; Hg19,Human Genome version 19; UCSC, University of California, Santa Cruz Human Genome Browser. Stouffer-Liptak-Kechris correction (slk); one-step Siidak (1967) multiple-testing correction.

References

1. 2017 CDC. Suicide Mortality by State report [https://www.cdc.gov/nchs/pressroom/sosmap/suicide-mortality/suicide.htm].

2. 2018 WHO. Global Health Observatory (GHO) data [Available from: https://www.who.int/news-room/fact-sheets/detail/suicide].

3. Malone KM, Oquendo MA, Haas GL, Ellis SP, Li S, Mann JJ. Protective factors against suicidal acts in major depression: reasons for living. *Am J Psychiatry* 2000; **157**(7)**:** 1084-1088.

4. Schosser A, Butler AW, Ising M, Perroud N, Uher R, Ng MY *et al.* Genomewide association scan of suicidal thoughts and behaviour in major depression. *PLoS One* 2011; **6**(7)**:** e20690.

5. Willour VL, Seifuddin F, Mahon PB, Jancic D, Pirooznia M, Steele J *et al.* A genome-wide association study of attempted suicide. *Mol Psychiatry* 2012; **17**(4)**:** 433-444.

6. Galfalvy H, Zalsman G, Huang YY, Murphy L, Rosoklija G, Dwork AJ *et al.* A pilot genome wide association and gene expression array study of suicide with and without major depression. *World J Biol Psychiatry* 2013; **14**(8)**:** 574-582.

7. Mullins N, Bigdeli TB, Borglum AD, Coleman JRI, Demontis D, Mehta D *et al.* GWAS of Suicide Attempt in Psychiatric Disorders and Association With Major Depression Polygenic Risk Scores. *Am J Psychiatry* 2019**:** appiajp201918080957.

8. Policicchio S, Dempster EL, Murphy TM. Deciphering the Epigenetic Landscape of Suicidal Behaviour: A Review of Current Findings, Caveats and Future Directions. 2(4):039: OBM Genetics, 2018.

9. Turecki G, Brent DA. Suicide and suicidal behaviour. *Lancet* 2016; **387**(10024)**:** 1227-1239.

10. Bell JT, Tsai PC, Yang TP, Pidsley R, Nisbet J, Glass D *et al.* Epigenome-wide scans identify differentially methylated regions for age and age-related phenotypes in a healthy ageing population. *PLoS Genet* 2012; **8**(4)**:** e1002629.

11. Feil R, Fraga MF. Epigenetics and the environment: emerging patterns and implications. *Nat Rev Genet* 2012; **13**(2)**:** 97-109.

12. Burns SB, Szyszkowicz JK, Luheshi GN, Lutz PE, Turecki G. Plasticity of the epigenome during early-life stress. *Semin Cell Dev Biol* 2018; **77:** 115-132.

13. Mill J, Heijmans BT. From promises to practical strategies in epigenetic epidemiology. *Nat Rev Genet* 2013; **14**(8)**:** 585-594.

14. Wong CC, Meaburn EL, Ronald A, Price TS, Jeffries AR, Schalkwyk LC *et al.* Methylomic analysis of monozygotic twins discordant for autism spectrum disorder and related behavioural traits. *Mol Psychiatry* 2014; **19**(4)**:** 495-503.

15. Pidsley R, Viana J, Hannon E, Spiers H, Troakes C, Al-Saraj S *et al.* Methylomic profiling of human brain tissue supports a neurodevelopmental origin for schizophrenia. *Genome Biol* 2014; **15**(10)**:** 483.

16. Uddin M, Aiello AE, Wildman DE, Koenen KC, Pawelec G, de Los Santos R *et al.* Epigenetic and immune function profiles associated with posttraumatic stress disorder. *Proc Natl Acad Sci U S A* 2010; **107**(20)**:** 9470-9475.

17. Fisher HL, Murphy TM, Arseneault L, Caspi A, Moffitt TE, Viana J *et al.* Methylomic analysis of monozygotic twins discordant for childhood psychotic symptoms. *Epigenetics* 2015; **10**(11)**:** 1014-1023.

18. Murphy TM, O'Donovan A, Mullins N, O'Farrelly C, McCann A, Malone K. Anxiety is associated with higher levels of global DNA methylation and altered expression of epigenetic and interleukin-6 genes. *Psychiatr Genet* 2015; **25**(2)**:** 71-78.

19. Dempster EL, Wong CC, Lester KJ, Burrage J, Gregory AM, Mill J *et al.* Genome-wide methylomic analysis of monozygotic twins discordant for adolescent depression. *Biol Psychiatry* 2014; **76**(12)**:** 977-983.

20. Davies MN, Krause L, Bell JT, Gao F, Ward KJ, Wu H *et al.* Hypermethylation in the ZBTB20 gene is associated with major depressive disorder. *Genome Biol* 2014; **15**(4)**:** R56.

21. Uddin M, Koenen KC, Aiello AE, Wildman DE, de los Santos R, Galea S. Epigenetic and inflammatory marker profiles associated with depression in a community-based epidemiologic sample. *Psychol Med* 2011; **41**(5)**:** 997-1007.

22. Murphy TM, Crawford B, Dempster EL, Hannon E, Burrage J, Turecki G *et al.* Methylomic profiling of cortex samples from completed suicide cases implicates a role for PSORS1C3 in major depression and suicide. *Transl Psychiatry* 2017; **7**(1)**:** e989.

23. Murphy TM, Mullins N, Ryan M, Foster T, Kelly C, McClelland R *et al.* Genetic variation in DNMT3B and increased global DNA methylation is associated with suicide attempts in psychiatric patients. *Genes Brain Behav* 2013; **12**(1)**:** 125-132.

24. Labonte B, Suderman M, Maussion G, Navaro L, Yerko V, Mahar I *et al.* Genome-wide epigenetic regulation by early-life trauma. *Arch Gen Psychiatry* 2012; **69**(7)**:** 722-731.

25. Ernst C, Deleva V, Deng X, Sequeira A, Pomarenski A, Klempan T *et al.* Alternative splicing, methylation state, and expression profile of tropomyosin-related kinase B in the frontal cortex of suicide completers. *Arch Gen Psychiatry* 2009; **66**(1)**:** 22-32.

26. Fiori LM, Turecki G. Epigenetic regulation of spermidine/spermine N1-acetyltransferase (SAT1) in suicide. *J Psychiatr Res* 2011; **45**(9)**:** 1229-1235.

27. Keller S, Sarchiapone M, Zarrilli F, Videtic A, Ferraro A, Carli V *et al.* Increased BDNF promoter methylation in the Wernicke area of suicide subjects. *Arch Gen Psychiatry* 2010; **67**(3)**:** 258-267.

28. Haghighi F, Xin Y, Chanrion B, O'Donnell AH, Ge Y, Dwork AJ *et al.* Increased DNA methylation in the suicide brain. *Dialogues Clin Neurosci* 2014; **16**(3)**:** 430-438.

29. Fiori LM, Gross JA, Turecki G. Effects of histone modifications on increased expression of polyamine biosynthetic genes in suicide. *Int J Neuropsychopharmacol* 2012; **15**(8)**:** 1161-1166.

30. Maussion G, Yang J, Suderman M, Diallo A, Nagy C, Arnovitz M *et al.* Functional DNA methylation in a transcript specific 3'UTR region of TrkB associates with suicide. *Epigenetics* 2014; **9**(8)**:** 1061-1070.

31. Kozlenkov A, Jaffe AE, Timashpolsky A, Apontes P, Rudchenko S, Barbu M *et al.* DNA Methylation Profiling of Human Prefrontal Cortex Neurons in Heroin Users Shows Significant Difference between Genomic Contexts of Hyper- and Hypomethylation and a Younger Epigenetic Age. *Genes (Basel)* 2017; **8**(6).

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

33. Viana J, Hannon E, Dempster E, Pidsley R, Macdonald R, Knox O *et al.* Schizophrenia-associated methylomic variation: molecular signatures of disease and polygenic risk burden across multiple brain regions. *Hum Mol Genet* 2017; **26**(1)**:** 210-225.

34. Pidsley R, CC YW, Volta M, Lunnon K, Mill J, Schalkwyk LC. A data-driven approach to preprocessing Illumina 450K methylation array data. *BMC Genomics* 2013; **14:** 293.

35. Price ME, Cotton AM, Lam LL, Farre P, Emberly E, Brown CJ *et al.* Additional annotation enhances potential for biologically-relevant analysis of the Illumina Infinium HumanMethylation450 BeadChip array. *Epigenetics Chromatin* 2013; **6**(1)**:** 4.

36. Chen YA, Lemire M, Choufani S, Butcher DT, Grafodatskaya D, Zanke BW *et al.* Discovery of cross-reactive probes and polymorphic CpGs in the Illumina Infinium HumanMethylation450 microarray. *Epigenetics* 2013; **8**(2)**:** 203-209.

37. Multivariate Data Visualization with R. http://lmdvrr-forger-projectorg, 2008., 2008, Accessed Date Accessed 2008 Accessed.

38. Hannon E, Dempster E, Viana J, Burrage J, Smith AR, Macdonald R *et al.* An integrated genetic-epigenetic analysis of schizophrenia: evidence for co-localization of genetic associations and differential DNA methylation. *Genome Biol* 2016; **17**(1)**:** 176.

39. Pedersen BS, Schwartz DA, Yang IV, Kechris KJ. Comb-p: software for combining, analyzing, grouping and correcting spatially correlated P-values. *Bioinformatics* 2012; **28**(22)**:** 2986-2988.

40. Šidák Z. Rectangular Confidence Regions for the Means of Multivariate Normal Distributions. *Journal of the American Statistical Association* 1967; **62:318:** 626-633.

41. Lunnon K, Hannon E, Smith RG, Dempster E, Wong C, Burrage J *et al.* Variation in 5-hydroxymethylcytosine across human cortex and cerebellum. *Genome Biol* 2016; **17:** 27.

42. Rydbirk R, Folke J, Winge K, Aznar S, Pakkenberg B, Brudek T. Assessment of brain reference genes for RT-qPCR studies in neurodegenerative diseases. *Sci Rep* 2016; **6:** 37116.

43. Pfaffl MW. A new mathematical model for relative quantification in real-time RT-PCR. *Nucleic Acids Res* 2001; **29**(9)**:** e45.

44. Rehli MKM. Functional Analysis of Promoter CPG-Methylation using a CpG-Free Luciferase Reporter Vector. *Epigenetics* 2006; **1**(3)**:** 4.

45. Jacobs JL, Dinman JD. Systematic analysis of bicistronic reporter assay data. *Nucleic Acids Res* 2004; **32**(20)**:** e160.

46. Abdelrahman HA, Al-Shamsi AM, Ali BR, Al-Gazali L. A null variant in PUS3 confirms its involvement in intellectual disability and further delineates the associated neurodevelopmental disease. *Clin Genet* 2018; **94**(6)**:** 586-587.

47. Klug M, Rehli M. Functional Analysis of Promoter CPGMethylation using a CpG-Free Luciferase Reporter Vector. *Epigenetics* 2006; **1:3:** 127-130.

48. Consortium GT. The Genotype-Tissue Expression (GTEx) project. *Nat Genet* 2013; **45**(6)**:** 580-585.

49. Schneider E, El Hajj N, Muller F, Navarro B, Haaf T. Epigenetic Dysregulation in the Prefrontal Cortex of Suicide Completers. *Cytogenet Genome Res* 2015; **146**(1)**:** 19-27.

50. Nagy C, Suderman M, Yang J, Szyf M, Mechawar N, Ernst C *et al.* Astrocytic abnormalities and global DNA methylation patterns in depression and suicide. *Mol Psychiatry* 2015; **20**(3)**:** 320-328.

51. Nagy C, Torres-Platas SG, Mechawar N, Turecki G. Repression of Astrocytic Connexins in Cortical and Subcortical Brain Regions and Prefrontal Enrichment of H3K9me3 in Depression and Suicide. *Int J Neuropsychopharmacol* 2017; **20**(1)**:** 50-57.

52. Phillips JR, Hewedi DH, Eissa AM, Moustafa AA. The cerebellum and psychiatric disorders. *Front Public Health* 2015; **3:** 66.

53. Hwang JP, Lee TW, Tsai SJ, Chen TJ, Yang CH, Lirng JF *et al.* Cortical and subcortical abnormalities in late-onset depression with history of suicide attempts investigated with MRI and voxel-based morphometry. *J Geriatr Psychiatry Neurol* 2010; **23**(3)**:** 171-184.

54. Amen DG, Prunella JR, Fallon JH, Amen B, Hanks C. A comparative analysis of completed suicide using high resolution brain SPECT imaging. *J Neuropsychiatry Clin Neurosci* 2009; **21**(4)**:** 430-439.

55. Lee YJ, Kim S, Gwak AR, Kim SJ, Kang SG, Na KS *et al.* Decreased regional gray matter volume in suicide attempters compared to suicide non-attempters with major depressive disorders. *Compr Psychiatry* 2016; **67:** 59-65.

56. Jung J, Choi S, Han KM, Kim A, Kang W, Paik JW *et al.* Alterations in functional brain networks in depressed patients with a suicide attempt history. *Neuropsychopharmacology* 2019.

57. Stawski R, Piaskowski S, Stoczynska-Fidelus E, Wozniak K, Bienkowski M, Zakrzewska M *et al.* Reduced expression of ELAVL4 in male meningioma patients. *Brain Tumor Pathol* 2013; **30**(3)**:** 160-166.

58. Akamatsu W, Fujihara H, Mitsuhashi T, Yano M, Shibata S, Hayakawa Y *et al.* The RNA-binding protein HuD regulates neuronal cell identity and maturation. *Proc Natl Acad Sci U S A* 2005; **102**(12)**:** 4625-4630.

59. Ince-Dunn G, Okano HJ, Jensen KB, Park WY, Zhong R, Ule J *et al.* Neuronal Elav-like (Hu) proteins regulate RNA splicing and abundance to control glutamate levels and neuronal excitability. *Neuron* 2012; **75**(6)**:** 1067-1080.

60. Talman V, Pascale A, Jantti M, Amadio M, Tuominen RK. Protein Kinase C Activation as a Potential Therapeutic Strategy in Alzheimer's Disease: Is there a Role for Embryonic Lethal Abnormal Vision-like Proteins? *Basic Clin Pharmacol Toxicol* 2016; **119**(2)**:** 149-160.

61. Yamada K, Iwayama Y, Hattori E, Iwamoto K, Toyota T, Ohnishi T *et al.* Genome-wide association study of schizophrenia in Japanese population. *PLoS One* 2011; **6**(6)**:** e20468.

62. Berto S, Usui N, Konopka G, Fogel BL. ELAVL2-regulated transcriptional and splicing networks in human neurons link neurodevelopment and autism. *Hum Mol Genet* 2016; **25**(12)**:** 2451-2464.

63. de Paiva ARB, Lynch DS, Melo US, Lucato LT, Freua F, de Assis BDR *et al.* PUS3 mutations are associated with intellectual disability, leukoencephalopathy, and nephropathy. *Neurol Genet* 2019; **5**(1)**:** e306.

64. Yokota N, Aruga J, Takai S, Yamada K, Hamazaki M, Iwase T *et al.* Predominant expression of human zic in cerebellar granule cell lineage and medulloblastoma. *Cancer Res* 1996; **56**(2)**:** 377-383.

65. Vitucci D, Di Giorgio A, Napolitano F, Pelosi B, Blasi G, Errico F *et al.* Rasd2 Modulates Prefronto-Striatal Phenotypes in Humans and 'Schizophrenia-Like Behaviors' in Mice. *Neuropsychopharmacology* 2016; **41**(3)**:** 916-927.

66. Sun H, Wang J, Gong Z, Yao J, Wang Y, Xu J *et al.* Quantitative integration of epigenomic variation and transcription factor binding using MAmotif toolkit identifies an important role of IRF2 as transcription activator at gene promoters. *Cell Discov* 2018; **4:** 38.

67. Drew PD, Franzoso G, Carlson LM, Biddison WE, Siebenlist U, Ozato K. Interferon regulatory factor-2 physically interacts with NF-kappa B in vitro and inhibits NF-kappa B induction of major histocompatibility class I and beta 2-microglobulin gene expression in transfected human neuroblastoma cells. *J Neuroimmunol* 1995; **63**(2)**:** 157-162.

68. Daniele LL, Emran F, Lobo GP, Gaivin RJ, Perkins BD. Mutation of wrb, a Component of the Guided Entry of Tail-Anchored Protein Pathway, Disrupts Photoreceptor Synapse Structure and Function. *Invest Ophthalmol Vis Sci* 2016; **57**(7)**:** 2942-2954.

69. Ling KH, Hewitt CA, Tan KL, Cheah PS, Vidyadaran S, Lai MI *et al.* Functional transcriptome analysis of the postnatal brain of the Ts1Cje mouse model for Down syndrome reveals global disruption of interferon-related molecular networks. *BMC Genomics* 2014; **15:** 624.

70. Nair RP, Stuart PE, Nistor I, Hiremagalore R, Chia NVC, Jenisch S *et al.* Sequence and haplotype analysis supports HLA-C as the psoriasis susceptibility 1 gene. *Am J Hum Genet* 2006; **78**(5)**:** 827-851.

71. Chang YT, Chou CT, Shiao YM, Lin MW, Yu CW, Chen CC *et al.* Psoriasis vulgaris in Chinese individuals is associated with PSORS1C3 and CDSN genes. *Br J Dermatol* 2006; **155**(4)**:** 663-669.

72. Wisniewski A, Matusiak L, Szczerkowska-Dobosz A, Nowak I, Kusnierczyk P. HLA-C*06:02-independent, gender-related association of PSORS1C3 and PSORS1C1/CDSN single-nucleotide polymorphisms with risk and severity of psoriasis. *Mol Genet Genomics* 2018; **293**(4)**:** 957-966.

73. Polanowska J, Fabbrizio E, Le Cam L, Trouche D, Emiliani S, Herrera R *et al.* The periodic down regulation of Cyclin E gene expression from exit of mitosis to end of G(1) is controlled by a deacetylase- and E2F-associated bipartite repressor element. *Oncogene* 2001; **20**(31)**:** 4115-4127.

74. Yamada K, Hattori E, Iwayama Y, Toyota T, Iwata Y, Suzuki K *et al.* Population-dependent contribution of the major histocompatibility complex region to schizophrenia susceptibility. *Schizophr Res* 2015; **168**(1-2)**:** 444-449.

75. Pfleger CM. The Hippo Pathway: A Master Regulatory Network Important in Development and Dysregulated in Disease. *Curr Top Dev Biol* 2017; **123:** 181-228.

76. Sun R, Wang Z, Claus Henn B, Su L, Lu Q, Lin X *et al.* Identification of novel loci associated with infant cognitive ability. *Mol Psychiatry* 2018.

77. Hamdan FF, Srour M, Capo-Chichi JM, Daoud H, Nassif C, Patry L *et al.* De novo mutations in moderate or severe intellectual disability. *PLoS Genet* 2014; **10**(10)**:** e1004772.

78. Nizon M, Laugel V, Flanigan KM, Pastore M, Waldrop MA, Rosenfeld JA *et al.* Variants in MED12L, encoding a subunit of the mediator kinase module, are responsible for intellectual disability associated with transcriptional defect. *Genet Med* 2019.

79. Branco MR, Ficz G, Reik W. Uncovering the role of 5-hydroxymethylcytosine in the epigenome. *Nat Rev Genet* 2011; **13**(1)**:** 7-13.

80. Clive ML, Boks MP, Vinkers CH, Osborne LM, Payne JL, Ressler KJ *et al.* Discovery and replication of a peripheral tissue DNA methylation biosignature to augment a suicide prediction model. *Clin Epigenetics* 2016; **8:** 113.

81. Sadeh N, Wolf EJ, Logue MW, Hayes JP, Stone A, Griffin LM *et al.* Epigenetic Variation at Ska2 Predicts Suicide Phenotypes and Internalizing Psychopathology. *Depress Anxiety* 2016; **33**(4)**:** 308-315.

82. Kaminsky Z, Wilcox HC, Eaton WW, Van Eck K, Kilaru V, Jovanovic T *et al.* Epigenetic and genetic variation at SKA2 predict suicidal behavior and post-traumatic stress disorder. *Transl Psychiatry* 2015; **5:** e627.

83. Lockwood LE, Su S, Youssef NA. The role of epigenetics in depression and suicide: A platform for gene-environment interactions. *Psychiatry Res* 2015; **228**(3)**:** 235-242.

84. Guintivano J, Brown T, Newcomer A, Jones M, Cox O, Maher BS *et al.* Identification and replication of a combined epigenetic and genetic biomarker predicting suicide and suicidal behaviors. *Am J Psychiatry* 2014; **171**(12)**:** 1287-1296.

Figures

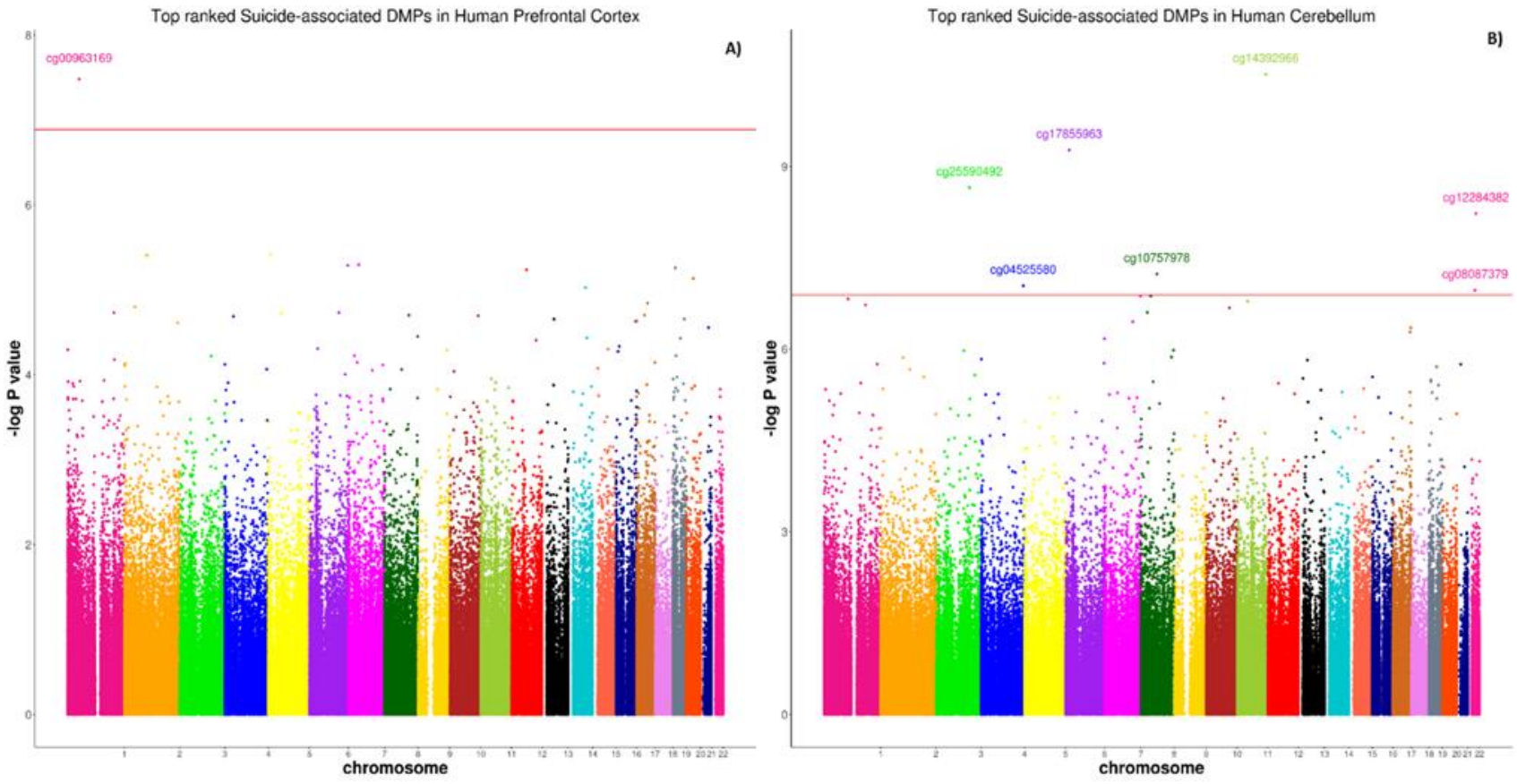


Figure 1. Suicide-associated DMPs in human cortex and cerebellu

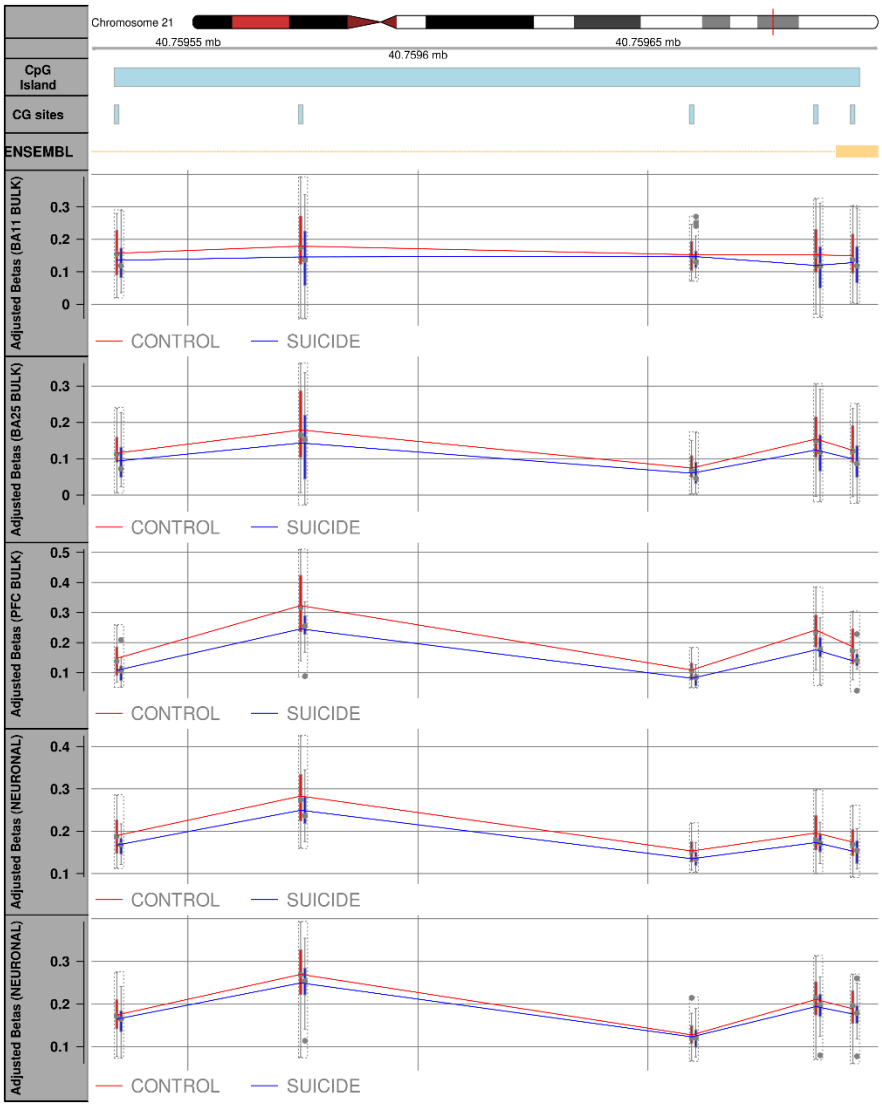


Figure 2. Suicide-associated differentially methylated region (DMR) in human Prefrontal cortex (PFC).

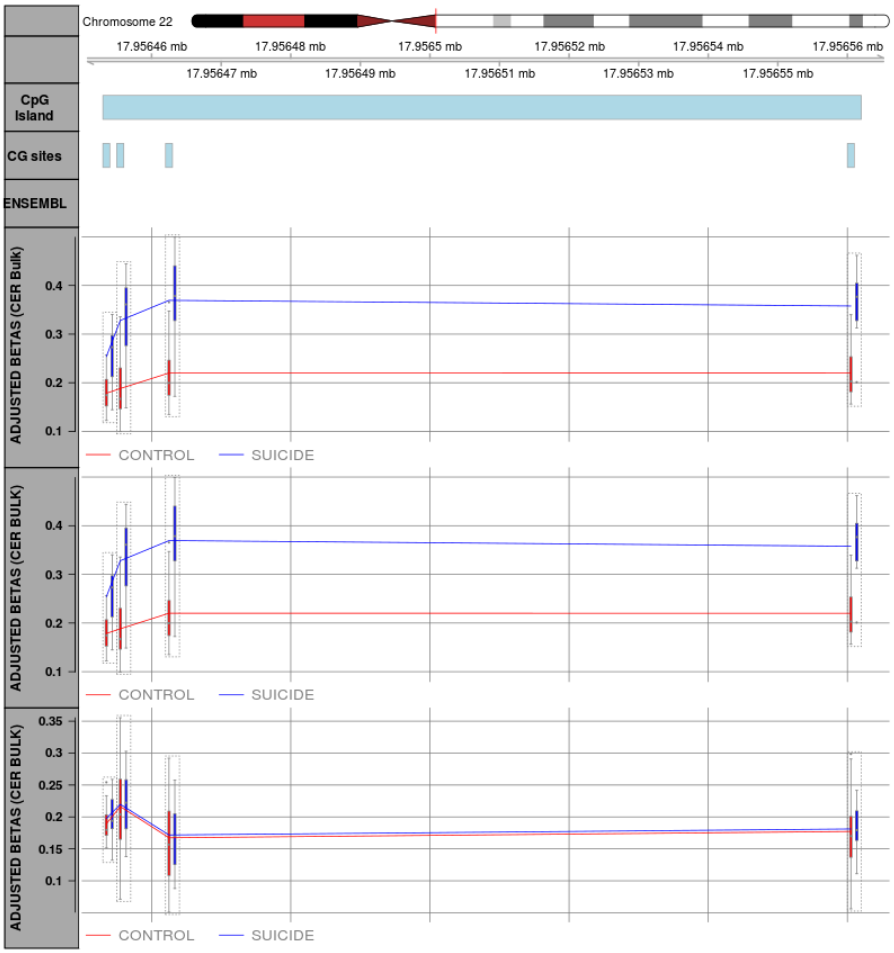


Figure 3. Suicide-associated differentially methylated region (DMR) in human Cerebellum (CER).

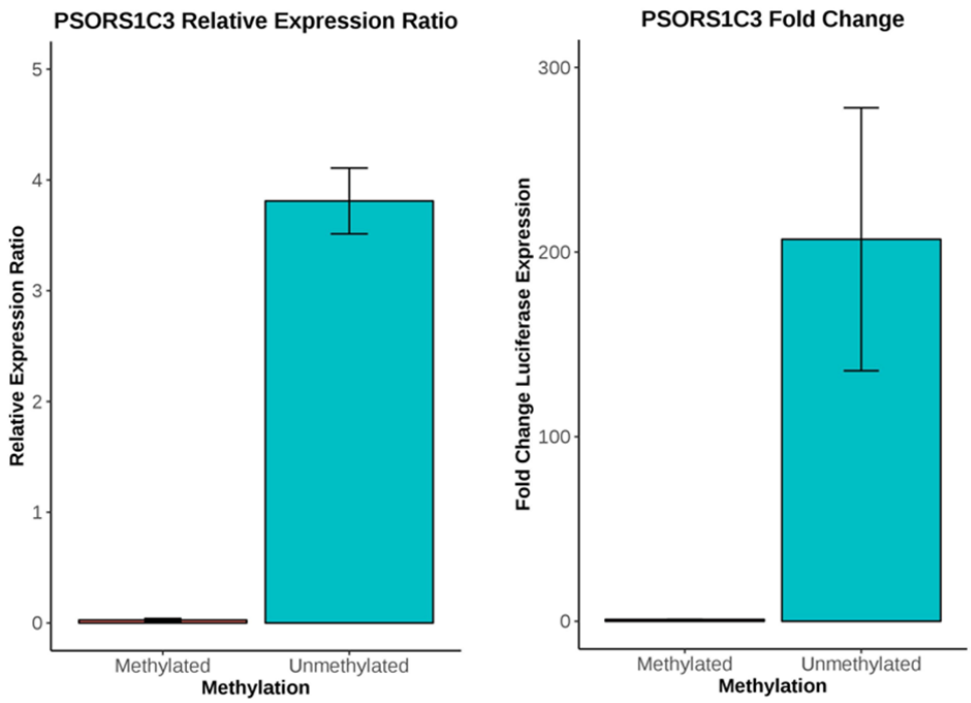
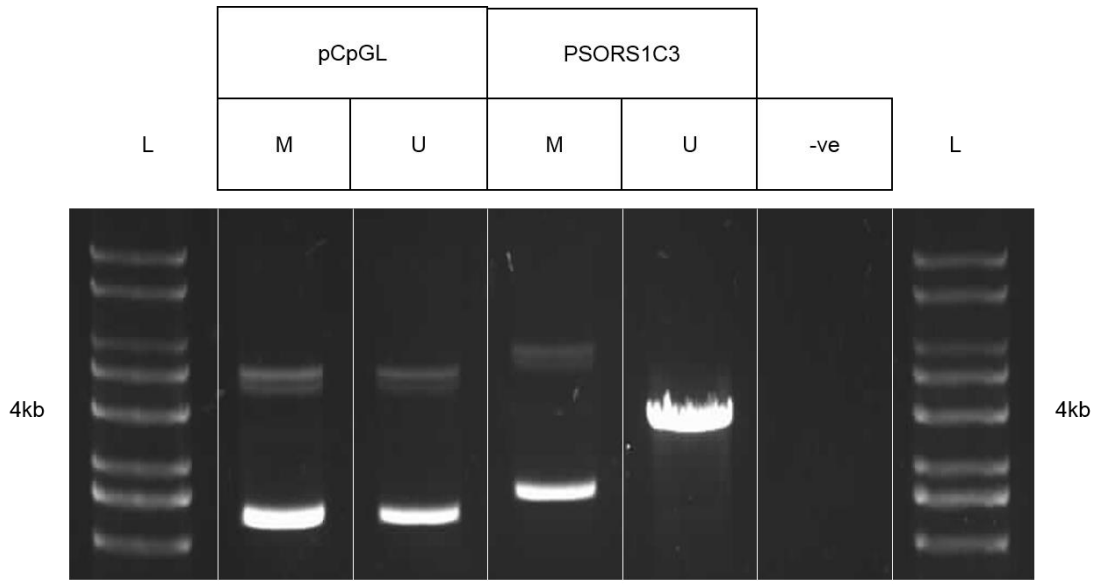


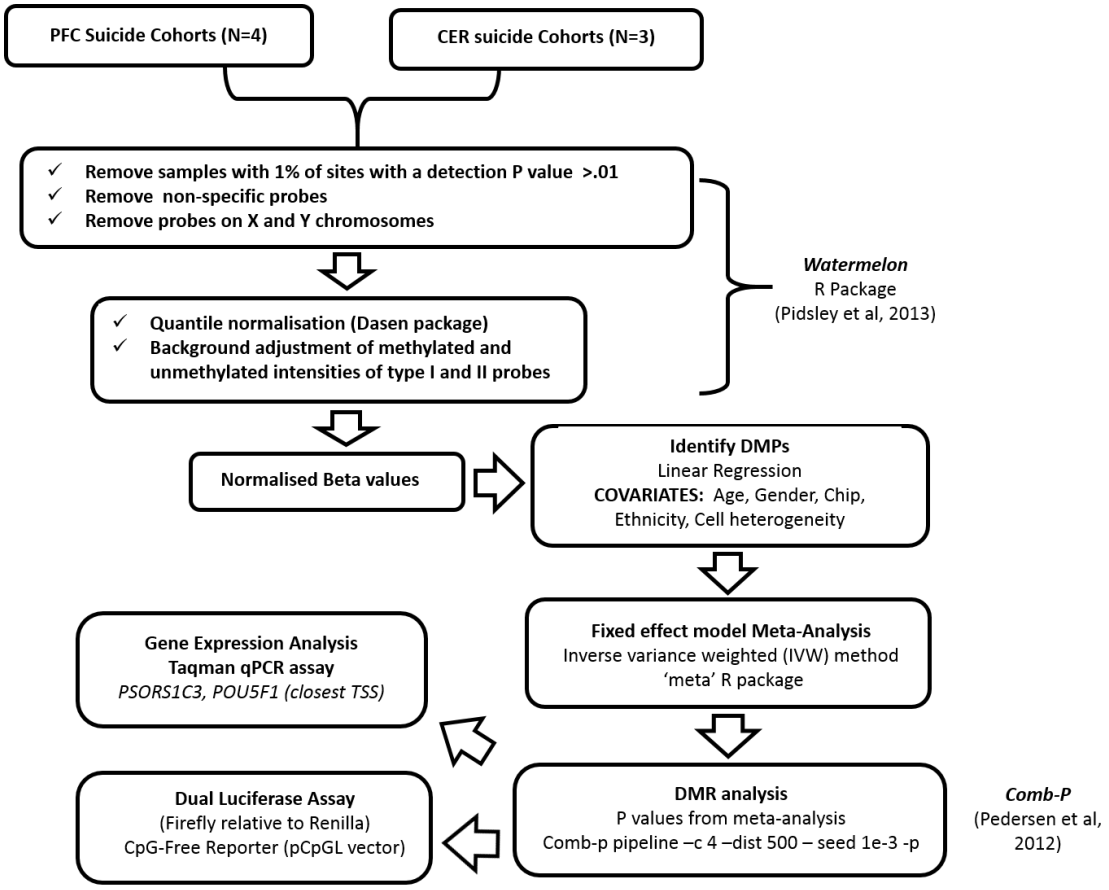
Figure 4: Methylation of the PSORS1C3 cloned pCpGL vector significantly reduces luciferase activity

- **PSORS1C3-A1-pCpGL-basic**
- ATTCTTGGCATCCTCCATGGGGAAGGGGAGAGGTGCTAAATAATTAAGCCCC
  CCTACTACTCAGCACCCGCGTGAGGCATCGTCAGGCATCGTCAGGCCTCCA
  GTGGTGGTGGTGGCACCGGGCCTCAACCTCCCCGGAGGGCTGGACTCTCG
  CTGCCAGGCTGTGGGGATCAGGCGTTGTGGGGGAGGGGGACACTTAACAG
  GTATGGAGGGCGGAGCAGAGCCCCGCAGTCACTGGCCTGACTTCCGGAAC
  GAACCGTCGCCAGCAAGCACAGCAGTAGGACCAGGGGGATGCAAGAGCGG
  GGGCGGCCGGGGATCGTGCTTCTCGCTCAGGTCCAGATTCCCGGCAACCA
  GGCCGGCGGAATCACGTGCCATGCTCCAGGCCAGCGTAGTCCCGCCCATCT
  TCCAGCTGAGCGTACCGGGAGGCTCCCATTGGACTGGAGCTGCTACGGAG
  GCGGGACTTTCCCTTTTTCTTGAACCCCATTGGGTTAAGTCCAGTCCGAGAC
  AAGCAGATCTGGATCCACTAGTCC

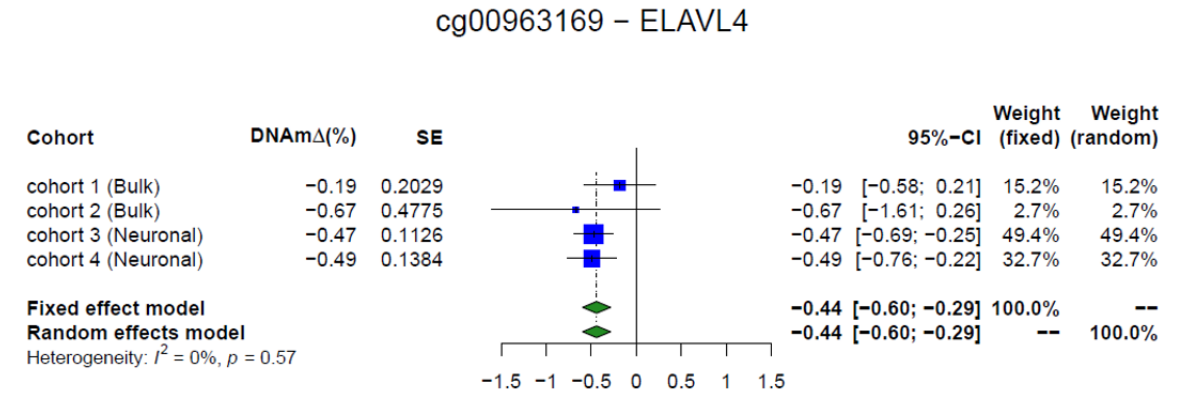
Supplementary Figure S1. *Sanger sequencing results of PSORS1C3-pCpGL-basic plasmids. Green bases are pCpGL-basic sequences, yellow are the PCR primers used for amplifying the PSORS1C3 region, purple are the CpG sites represented on the 450K array.*



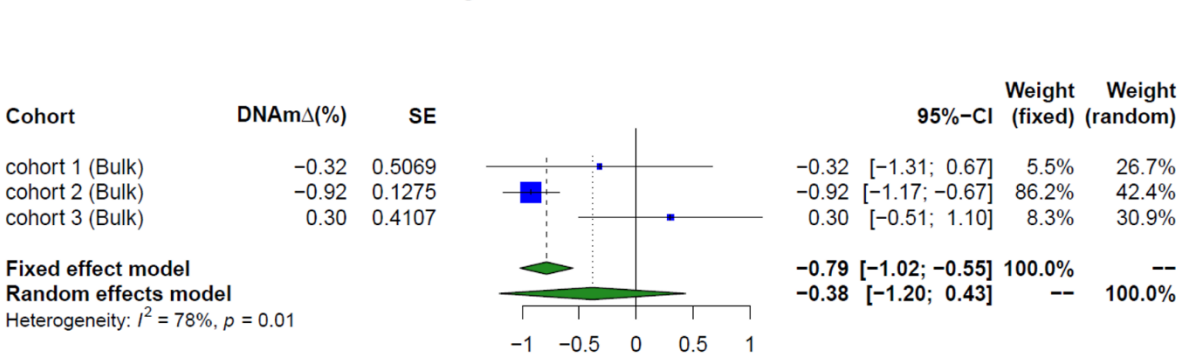
Supplementary Figure S2



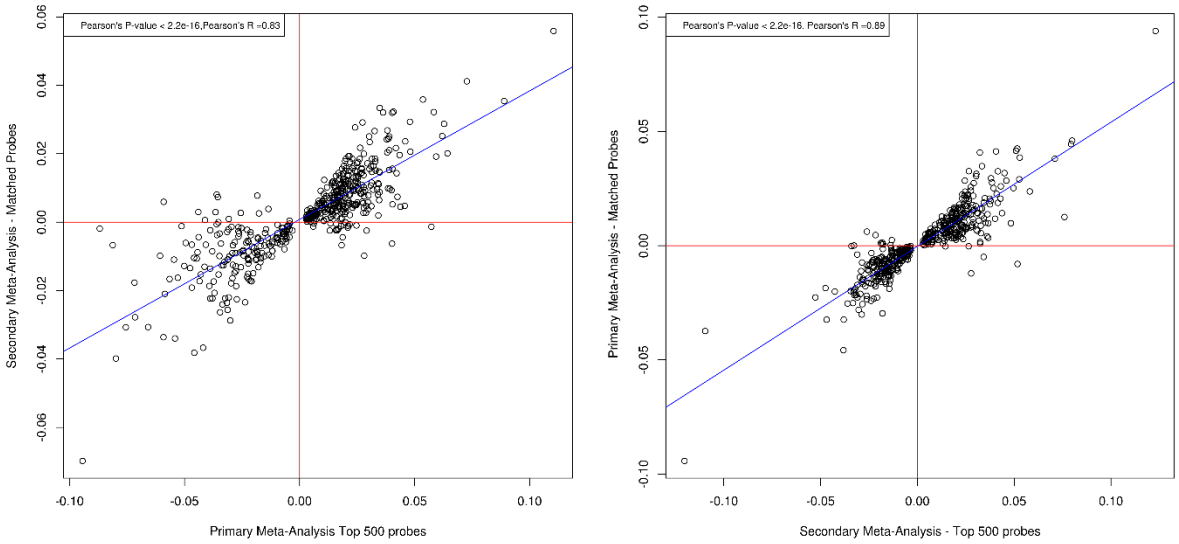
Supplementary Figure S3. Methodological pipeline overview. *Flow chart describing the methodological meta-analysis approach used in the study. Initial quality control was performed on signal intensities for each probe and all samples underwent to further quality-control checks. Stringent quality control measures were employed. Non-specific probes and probes on the X and Y chromosomes were removed. Quantile normalisation and background adjustment of M and UNM intensities of probes were employed using Watermelon R package. Linear regression was used to examine differences in DNA methylation scores (reported as changes in beta values (Δβ) between suicide completers and CTRs at each CpGs site, controlling for potential confounders. The results from the linear model were then met-analysed for each brain region separately. A fixed effects meta-analysis of 4 PFC suicide cohorts was performed using the 'metagen' function in the R package 'meta'. The results obtained from both the PFC and CER meta-analysis were used to perform a regional-based analysis using the Python module Comb-p, to identify suicide-associated differentially methylated regions (DMRs). In addition, functional consequences of variable DNA methylation within a PFC suicide-associated differentially methylated region (PSORS1C3 DMR) were examined using a dual luciferase assay (CpG-Free Luciferase Reporter (pCpGL vector) gene assay) and assessing expression levels of nearby genes (real-time qPCR using pre-optimized Taqman assays). Abbreviations: Prefrontal Cortex, PFC; Cerebellum, CER.*



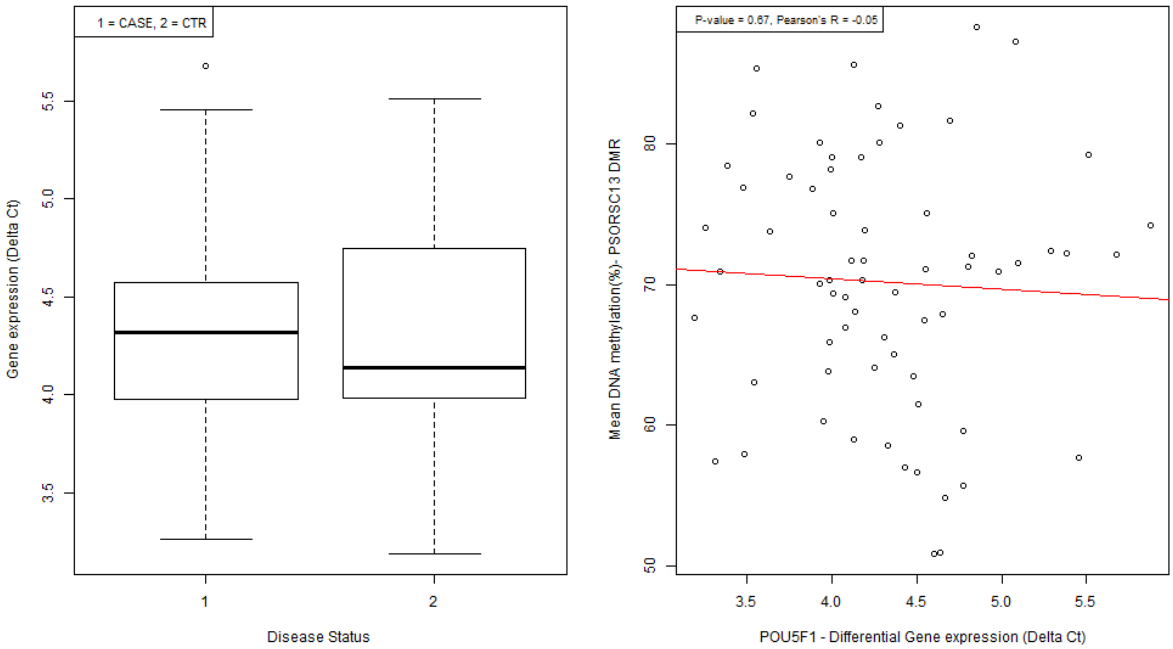
Supplementary Figure S4.Top-ranked differentially methylated probe (DMP) in human Prefrontal Cortex (PFC). *Forest plot showing the effect size at the top ranked CpG site (cg00963169) which is largely consistent across all the independent studies included. Abbreviations: DNAmΔ, DNA methylation changes (%); SE, standard errors; CI, confidence interval.*



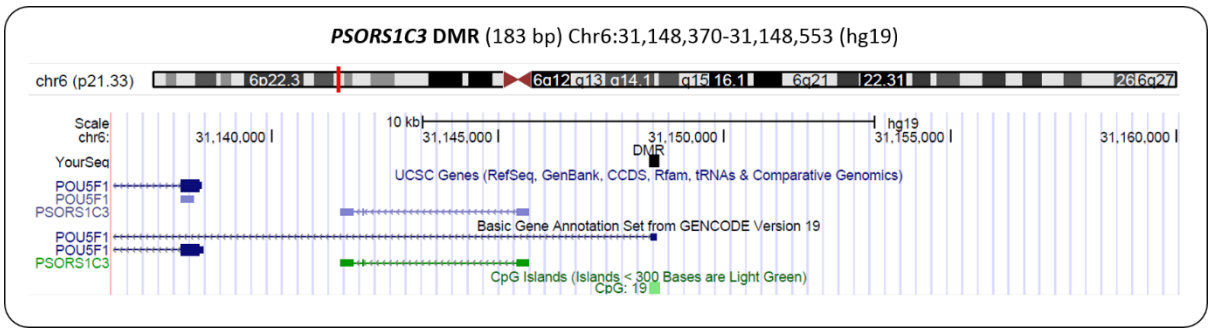
Supplementary Figure S5. Top-ranked differentially methylated probe (DMP) in human Cerebellum (CER). *Forest plot showing the effect size at the top ranked CpG site (cg14392966) which is largely consistent across all the independent studies included. This DMP, which was found hypomethylated in two of the 3 cohorts in suicide cases compared to healthy controls, is located in the promoter region of the pseudouridine synthase 3 (PUS3) gene encoding for a highly conserved enzyme responsible for posttranscriptional modification of tRNA and known for being involved in intellectual disability and previously associated to neurodevelopmental disease (PMID: 30308082, PMID: 30697592). Abbreviations: DNAmΔ, DNA methylation changes (%); SE, standard errors; CI, confidence interval.*



Supplementary Figure S6. Suicide-specific DNA methylation changes in Human Cerebellum. *Correlation of effect sizes (adjusted beta difference) across the two suicide meta-analyses: a) Primary Meta-analysis (psychiatric suicide cases versus non-psychiatric controls) and (b) Secondary Meta-analysis (psychiatric suicide cases versus non-suicide psychiatric controls). Both plots showed a significant positive correlation between effect size at the DMPs (Pearson's R=0.89, Pearson's R=0.83, respectively). PFC, Prefrontal cortex; CER, Cerebellum; DMP, differentially methylated position.*



Supplementary Figure S7. A) Gene expression levels of POU5F1 gene in suicide cases and healthy controls. *We analysed the association between gene-expression and disease status for our target gene, POU5F1 using a linear mixed effects model. No statistically significant difference in the expression of POU5F1 was observed (P = 0.5).* ***B) Correlation between DNA methylation levels at PSORSC13 DMR and POU5F1 gene expression levels.*** *As DNA methylation is generally associated to down-regulated gene expression we tested whether a negative correlation was present between gene expression levels (Delta Ct) of the POU5F1 gene and mean DNA methylation (%) measured at the nearby suicide-associated DMR (). No significant correlation was found (cor test= -0.05).*



Supplementary Figure S8. *Figure showing the location within the human genome of the suicide-associated, PFC specific, PSORS1C3 DMR. The two genes examined as part of the functional validation of the meta-analysis findings are also shown. Source: UCSC Genome Browser (hg19).*

Bibliography

Aach, J. *et al.* (2013) 'CAS9 transcriptional activators for target specificity screening and paired nickases for cooperative genome engineering', *Nature Biotechnology*, 31(9), pp. 883–838. doi: 10.1038/nbt.2675.

Abdolmaleky, H. M. *et al.* (2005) 'Hypermethylation of the reelin (RELN) promoter in the brain of schizophrenic patients: A preliminary report', *American Journal of Medical Genetics - Neuropsychiatric Genetics*, 134 B(1), pp. 60–66. doi: 10.1002/ajmg.b.30140.

Abdolmaleky, H. M. *et al.* (2009) 'Hypomethylation of MB-COMT promoter is a major risk factor for schizophrenia and bipolar disorder', *Human Molecular Genetics*, 15(21), pp. 3132–3145. doi: 10.1093/hmg/ddl253.Hypomethylation.

Agid, O. *et al.* (1999) 'Environment and vulnerability to major psychiatric illness: A case control study of early parental loss in major depression, bipolar disorder and schizophrenia', *Molecular Psychiatry*, 4(2), pp. 163–172. doi: 10.1038/sj.mp.4000473.

Aguet, F. *et al.* (2017) 'Genetic effects on gene expression across human tissues', *Nature*, 550(7675), pp. 204–213. doi: 10.1038/nature24277.

Alda-Catalinas, C. *et al.* (2019) 'A single-cell transcriptomics CRISPR-activation screen identifies new epigenetic regulators of zygotic genome activation', *bioRxiv*, p. 741371. doi: 10.1101/741371.

Allis, C. D. and Jenuwein, T. (2016) 'The molecular hallmarks of epigenetic control', *Nature Reviews Genetics*. Nature Publishing Group, 17(8), pp. 487–500. doi: 10.1038/nrg.2016.59.

Ambrosio, E. D. *et al.* (2019) 'The effect of a genetic variant at the schizophrenia associated AS3MT/BORCS7 locus on striatal dopamine function: A PET imaging study', *Psychiatry Research: Neuroimaging*, 291(July), pp. 34–41. doi: 10.1016/j.pscychresns.2019.07.005.

Anders, S. *et al.* (2013) 'Count-based differential expression analysis of RNA sequencing data using R and Bioconductor', *Nature Protocols*, 8(9), pp. 1765–1786. doi: 10.1038/nprot.2013.099.

Anders, S., Pyl, P. T. and Huber, W. (2015) 'Genome analysis HTSeq — a Python framework to work with high-throughput sequencing data', *Bioinformatics*, 31(2), pp. 166–169. doi: 10.1093/bioinformatics/btu638.

Andreasen, N. C. *et al.* (2011) 'Progressive brain change in schizophrenia: A prospective longitudinal study of first-episode schizophrenia', *Biological Psychiatry*, 70(7), pp. 672–679. doi: 10.1016/j.biopsych.2011.05.017.

Andreasson, S. *et al.* (1987) 'Cannabis and Schizophrenia: A Longitudinal Study of Swedish Conscripts', *The Lancet*, pp. 1483–1486. doi: 10.1016/S0140-6736(87)92620-1.

Andres, M. C. *et al.* (2014) 'Loss of Methylation in CpG Sites in the NF-kB Enhancer Elements of Inducible Nitric Oxide Synthase Is Responsible for Gene Induction in Human Articular Chondrocytes', *Arthritis & Rheumatology*, 65(3), pp. 732–742. doi:

10.1002/art.37806.Loss.

Aran, D. *et al.* (2011) 'Replication timing-related and gene body-specific methylation of active human genes', *Human Molecular Genetics*, 20(4), pp. 670–680. doi: 10.1093/hmg/ddq513.

Arseneault, L. *et al.* (2004) 'Causal association between cannabis and psychosis: examination of the evidence', *British Journal of Psychiatry*, 184, pp. 110–117.

Ashburner, M. *et al.* (2000) 'Gene Ontology: tool for the unification of biology', *Nature Genetics*, 25(1), pp. 25–29. doi: 10.1038/75556.Gene.

Babayan, A. H. and Kramár, E. A. (2013) 'Rapid Effects of Oestrogen on Synaptic Plasticity: Interactions with Actin and its Signaling Proteins', *Journal of Neuroendocrinology*, 25(5), pp. 1163–1172. doi: 10.1111/mec.13536.Application.

Banerjee, D. and Slack, F. (2002) 'Control of developmental timing by small temporal RNAs: A paradigm for RNA-mediated regulation of gene expression', *BioEssays*, 24(2), pp. 119–129. doi: 10.1002/bies.10046.

Bannister, A. J. and Kouzarides, T. (2011) 'Regulation of chromatin by histone modifications', *Cell Research*. Nature Publishing Group, 21(3), pp. 381–395. doi: 10.1038/cr.2011.22.

Barrangou, R. and Marraffini, L. A. (2014) 'CRISPR-Cas Systems: Prokaryotes Upgrade to Adaptive Immunity', *Molecular Cell*. Elsevier Inc., 54(2), pp. 234–244. doi: 10.1016/j.molcel.2014.03.011.

Bartel, D. P. (2004) 'MicroRNAs：Genomics, Biogenesis, Mechanism, and Function', *Cell*, 116(2), pp. 281–297. Available at: http://www.ncbi.nlm.nih.gov/pubmed/14744438.

Bassett, A. S. *et al.* (2005) 'Clinical features of 78 adults with 22q11 deletion syndrome', *American Journal of Medical Genetics*, 138 A(4), pp. 307–313. doi: 10.1002/ajmg.a.30984.

Bassett, A. S. *et al.* (2017) 'Rare Genome-Wide Copy Number Variation and Expression of Schizophrenia in 22q11.2 Deletion Syndrome', *American Journal of Psychiatry*, (6), pp. 0–4. doi: 10.1176/appi.ajp.2017.16121417.

Bassett, A. S. and Chow, E. W. C. (2008) 'Schizophrenia and 22q11.2 Deletion Syndrome', *Current Psychiatry Reports*, 10(2), pp. 148–157.

Bassett, A. S., Chow, E. W. and Weksberg, R. (2000) 'Chromosomal abnormalities and schizophrenia', *American Journal of Medical Genetics - Seminars in Medical Genetics*, 97(1), pp. 45–51. doi: 10.1002/(SICI)1096-8628(200021)97:1<45::AID-AJMG6>3.0.CO;2-9.

Bassett, A. S., Scherer, S. W. and Brzustowicz, L. M. (2010) 'Copy Number Variations in Schizophrenia: Critical Review and New Perspectives on Concepts of Genetics and Disease', *American Journal of Psychiatry*, 167(8), pp. 899–914. doi: 10.1176/appi.ajp.2009.09071016.Copy.

Beaubien, F. *et al.* (2016) 'Slitrk1 is localized to excitatory synapses and promotes their development', *Scientific Reports*, 6(July 2015), pp. 1–10. doi: 10.1038/srep27343.

Bell, J. T. *et al.* (2011) 'DNA methylation patterns associate with genetic and gene expression variation in HapMap cell lines', *Genome Biology*. BioMed Central Ltd, 12(1), p. R10. doi: 10.1186/gb-2011-12-1-r10.

Bellon, A. (2007) 'New genes associated with schizophrenia in neurite formation: a review of cell culture experiments', *Molecular Psychiatry*, 12, pp. 620–629. doi: 10.1038/sj.mp.4001985.

Benacerraf, B. (1981) 'Role of MHC gene products in immune regulation', *Science*, 212, pp. 1229–1238. doi: 10.1126/science.6165083.

Benjamini, Y. and Hochberg, Y. (1995) 'Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing', *Journal of the Royal Statistical Society*, 57(1), pp. 289–300.

Bibikova, M. *et al.* (2001) 'Stimulation of Homologous Recombination through Targeted Cleavage by Chimeric Nucleases', *Molecular and Cellular Biology*, 21(1), pp. 289–297. doi: 10.1128/mcb.21.1.289-297.2001.

Bibikova, M. *et al.* (2002) 'Targeted Chromosomal Cleavage and Mutagenesis in Drosophila Using Zinc-Finger Nucleases', *Genetics*, 1175(July), pp. 1169–1175.

Biedler, J. L. *et al.* (1978) 'Multiple Neurotransmitter Synthesis by Human Neuroblastoma Cell Lines and Clones', *Cancer Research*, 38(November), pp. 3751–3757.

Bird, A. P. (1986) 'CpG-rich islands and the function of DNA methylation', *Nature*, 321, pp. 1–5.

Bobilin, R. T. and Ching, Y. C. (2006) 'BEAMing up for detection and quantification of rare sequence variants', *Nature Methods*, 3(2), pp. 95–97. doi: 10.1038/nmeth850.

Boch, J. *et al.* (2009) 'Breaking the code of DNA binding specificity of TAL-type III effectors', *Science*, 326(5959), pp. 1509–1512. doi: 10.1126/science.1178811.

Boch, J. and Bonas, U. (2010) 'Xanthomonas AvrBs3 Family-Type III Effectors: Discovery and Function ', *Annual Review of Phytopathology*, 48(1), pp. 419–436. doi: 10.1146/annurev-phyto-080508-081936.

Van Den Bogaert, A. *et al.* (2003) 'The DTNBP1 (Dysbindin) Gene Contributes to Schizophrenia, Depending on Family History of the Disease', *American Journal of Human Genetics*, 73(6), pp. 1438–1443. doi: 10.1086/379928.

Boggs, D. L. *et al.* (2018) 'Clinical and Preclinical Evidence for Functional Interactions of Cannabidiol and delta9-Tetrahydrocannabinol', *Neuropsychopharmacology*. Nature Publishing Group, 43(1), pp. 142–154. doi: 10.1038/npp.2017.209.

Bolotin, A. *et al.* (2005) 'Clustered regularly interspaced short palindrome repeats (CRISPRs) have spacers of extrachromosomal origin', *Microbiology*, 151(8), pp. 2551–2561. doi: 10.1099/mic.0.28048-0.

Brandt, M. *et al.* (2019) 'A polyclonal allelic expression assay for detecting regulatory effects of transcript variants', *BioRxiv*.

Brennand, K. J. *et al.* (2011) 'Modelling schizophrenia using human induced

pluripotent stem cells', *Nature*. Nature Publishing Group, 473(7346), pp. 221–225. doi: 10.1038/nature09915.

Brown, A. S. *et al.* (2004) 'Elevated Maternal Interleukin-8 Levels and Risk of Schizophrenia in Adult Offspring', *American Journal of Psychiatry*, 161(5), pp. 889–895. doi: 10.1176/appi.ajp.161.5.889.

Brown, A. S. (2011) 'The environment and susceptibility to schizophrenia', *Progress in Neurobiology*, 93(1), pp. 23–58. doi: 10.1016/j.pneurobio.2010.09.003.The.

Brown, A. S. and Derkits, E. J. (2010) 'Prenatal Infection and Schizophrenia: A Review of Epidemiologic and Translational Studies', *American Journal of Psychiatry*, 167(3), pp. 261–280. Available at: https://ajp.psychiatryonline.org/doi/pdf/10.1176/appi.ajp.2009.09030361.

Brown, A. S. and Patterson, P. H. (2011) 'Maternal Infection and Schizophrenia : Implications for Prevention', *Schizophrenia Bulletin*, 37(2), pp. 284–290. doi: 10.1093/schbul/sbq146.

Buenrostro, J. D. *et al.* (2013) 'Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position', *Nature Methods*, 10(12), pp. 1213–1218. doi: 10.1038/nmeth.2688.

Burmeister, M., Mcinnis, M. G. and Zöllner, S. (2008) 'Psychiatric genetics: progress amid controversy', *Nature Reviews Genetics*, 9(6), p. 420. doi: 10.1038/nrg2381.

Canetta, S. and Kellendonk, C. (2018) 'Can we use mice to study schizophrenia?', *Philosophical Transactions of the Royal Society B: Biological Sciences*, 373(1742). doi: 10.1098/rstb.2017.0032.

Cannarsa, R. *et al.* (2012) 'Δ(9)-Tetrahydrocannabinol decreases NOP receptor density and mRNA levels in human SH-SY5Y cells.', *Journal of molecular neuroscience : MN*, 46(2), pp. 285–292. doi: 10.1007/s12031-011-9552-0.

Cannon, M., Jones, P. B. and Murray, R. M. (2002) 'Obstetric complications and schizophrenia: Historical and meta-analytic review', *American Journal of Psychiatry*, 159(7), pp. 1080–1092. doi: 10.1176/appi.ajp.159.7.1080.

Cantor-Graae, E. and Selten, J. P. (2005) 'Schizophrenia and migration: A meta-analysis and review', *American Journal of Psychiatry*, 162(1), pp. 12–24. doi: 10.1176/appi.ajp.162.1.12.

Cardno, A. G. and Gottesman, I. I. (2000) 'Twin studies of schizophrenia: From bow-and-arrow concordances to star wars Mx and functional genomics', *American Journal of Medical Genetics - Seminars in Medical Genetics*, 97(1), pp. 12–17. doi: 10.1002/(SICI)1096-8628(200021)97:1<12::AID-AJMG3>3.0.CO;2-U.

Carreras-puigvert, J. *et al.* (2017) 'A comprehensive structural, biochemical and biological profiling of the human NUDIX hydrolase family', *Nature Communications*, 8(1541), pp. 1–17. doi: 10.1038/s41467-017-01642-w.

Carroll, C. B. *et al.* (2012) 'Δ9-tetrahydrocannabinol (Δ9-THC) exerts a direct neuroprotective effect in a human cell culture model of Parkinson's disease', *Neuropathology and Applied Neurobiology*, 38(6), pp. 535–547. doi: 10.1111/j.1365-2990.2011.01248.x.

Chan, R. F. *et al.* (2019) 'Independent Methylome-Wide Association Studies of Schizophrenia Detect Consistent Case–Control Differences', *Schizophrenia Bulletin*, pp. 1–9. doi: 10.1093/schbul/sbz056.

Chapman, G. V (2000) 'Instrumentation for flow cytometry', *Journal of Immunological Methods*, 243, pp. 3–12.

Chapman, J. R., Taylor, M. R. and Boulton, S. J. (2012) 'Playing the End Game: DNA Double-Strand Break Repair Pathway Choice', *Molecular Cell*. Elsevier, 47(4), pp. 497–510. doi: 10.1016/j.molcel.2012.07.029.

Chapman, R. M. *et al.* (2018) 'Convergent Evidence That ZNF804A Is a Regulator of Pre-messenger RNA Processing and Gene Expression', *Schizophrenia Bulletin*, pp. 1–12. doi: 10.1093/schbul/sby183.

Chatterjee, M., Schild, D. and Teunissen, C. E. (2019) 'Contactins in the central nervous system: role in health and disease', *Neural Regeneration Research*, 14(2), pp. 206–216. doi: 10.4103/1673-5374.244776.

Chavez, A. *et al.* (2015) 'Highly efficient Cas9-mediated transcriptional programming', *Nature Methods*, 12(4), pp. 326–328. doi: 10.1038/nmeth.3312.

Chen, Y. *et al.* (2013) 'Discovery of cross-reactive probes and polymorphic CpGs in the Illumina Infinium HumanMethylation450 microarray', *Epigenetics*, 8(2), pp. 203–209.

Chowdari, K. V. (2003) 'Association and linkage analyses of RGS4 polymorphisms in schizophrenia', *Human Molecular Genetics*, 12(14), pp. 1781–1781. doi: 10.1093/hmg/ddg171.

Chu, V. T. *et al.* (2015) 'Increasing the efficiency of homology-directed repair for CRISPR-Cas9-induced precise gene editing in mammalian cells.', *Nature Biotechnology*, 33(5), pp. 543–548. doi: 10.1038/nbt.3198.

Chumakov, I. *et al.* (2002) 'Genetic and physiological data implicating the new human gene G72 and the gene for D-amino acid oxidase in schizophrenia', *Proceedings of the National Academy of Sciences of the United States of America*, 99(26), pp. 13675–13680.

Ciccarone, V. *et al.* (1989) 'Phenotypic Diversification in Human Neuroblastoma Cells: Expression of Distinct Neural Crest Lineages', *Cancer Research*, 49(1), pp. 219–225.

Clapcote, S. J. *et al.* (2007) 'Behavioral Phenotypes of Disc1 Missense Mutations in Mice', *Neuron*, 54(3), pp. 387–402. doi: 10.1016/j.neuron.2007.04.015.

Clowry, G., Molnár, Z. and Rakic, P. (2010) 'Renewed focus on the developing human neocortex', *Journal of Anatomy*, 217(4), pp. 276–288. doi: 10.1111/j.1469-7580.2010.01281.x.

Conesa, A. *et al.* (2016) 'A survey of best practices for RNA-seq data analysis', *Genome Biol*, 17(1), p. 13. doi: 10.1186/s13059-016-0881-8.

Costas, J. *et al.* (2011) 'Interaction between COMT haplotypes and cannabis in schizophrenia: A case-only study in two samples from Spain', *Schizophrenia Research*. Elsevier B.V., 127(1–3), pp. 22–27. doi: 10.1016/j.schres.2011.01.014.

Cox, M. M. *et al.* (2009) 'Neurobehavioral abnormalities in the dysbindin-1 mutant, sandy, on a C57BL/6J genetic background', *Genes, Brain and Behavior*, 8(4), pp. 390–397. doi: 10.1111/j.1601-183X.2009.00477.x.

Craddock, N., O'Donovan, M. and Owen, M. (2005) 'The genetics of schizophrenia and bipolar disorder: Dissecting psychosis', *Journal of Medical Genetics*, 42(3), pp. 193–204. doi: 10.1136/jmg.2005.030718.

Crick, F. (1970) 'Central Dogma of Molecular Biology', *Nature*, 227, pp. 561–563. doi: 10.1007/978-1-4020-6754-9_2672.

Cunningham, F. *et al.* (2019) 'Ensembl 2019', *Nucleic Acids Research*, 47(D1), pp. D745–D751. doi: 10.1093/nar/gky1113.

D'Souza, D. C. *et al.* (2004) 'The psychotomimetic effects of intravenous delta-9-tetrahydrocannabinol in healthy individuals: Implications for psychosis', *Neuropsychopharmacology*, 29(8), pp. 1558–1572. doi: 10.1038/sj.npp.1300496.

Dana, H. *et al.* (2017) 'Molecular Mechanisms and Biological Functions of siRNA.', *International journal of biomedical science : IJBS*, 13(2), pp. 48–57. Available at: http://www.ncbi.nlm.nih.gov/pubmed/28824341%0Ahttp://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC5542916.

Darvesh, S., Grantham, D. L. and Hopkins, D. A. (1998) 'Distribution of Butyrylcholinesterase in the Human Amygdala and Hippocampal Formation', *The Journal of Comparative Neurology*, 393(October 1997), pp. 374–390.

Darvesh, S. and Hopkins, D. A. (2003) 'Differential Distribution of Butyrylcholinesterase and Acetylcholinesterase in the Human Thalamus', *The Journal of Comparative Neurology*, 43(August 2002), pp. 25–43. doi: 10.1002/cne.10751.

Darvesh, S., Hopkins, D. A. and Geula, C. (2003) 'Neurobiology of butrylcholinesterase', *Nature reviews. Neuroscience*, 4(February), pp. 1–8. doi: 10.1038/nrn1035.

Datlinger, P. *et al.* (2017) 'Pooled CRISPR screening with single-cell transcriptome readout', *Nature Methods*, 14(3), pp. 297–301. doi: 10.1038/nmeth.4177.

Davis, H. E., Morgan, J. R. and Yarmush, M. L. (2002) 'Polybrene increases retrovirus gene transfer efficiency by enhancing receptor-independent virus adsorption on target cell membranes', *Biophysical Chemistry*, 97(2–3), pp. 159–172. doi: 10.1016/S0301-4622(02)00057-1.

Deltcheva, E. *et al.* (2011) 'CRISPR RNA maturation by trans-encoded small RNA and host factor RNase III', *Nature*, 471(7340), pp. 602–607. doi: 10.1038/nature09886.

Demichele-sweet, M. A. A. *et al.* (2018) 'Genetic Risk for Schizophrenia and Psychosis in Alzheimer Disease', *Molecular Psychiatry*, 23(4), pp. 963–972. doi: 10.1038/mp.2017.81.Genetic.

Denayer, T., Stöhrn, T. and Van Roy, M. (2014) 'Animal models in translational medicine: Validation and prediction', *New Horizons in Translational Medicine*. Elsevier, 2(1), pp. 5–11. doi: 10.1016/j.nhtm.2014.08.001.

Dixit, A. *et al.* (2016) 'Perturb-seq: Dissecting molecular circuits with scalable single cell RNA profiling of pooled genetic screens', *Cell*, 167(7), pp. 1853–1866. doi: 10.1016/j.physbeh.2017.03.040.

Dobbyn, A. *et al.* (2018) 'Landscape of Conditional eQTL in Dorsolateral Prefrontal Cortex and Co-localization with Schizophrenia GWAS', *The American Journal of Human Genetics*, 102, pp. 1–16. doi: 10.1016/j.ajhg.2018.04.011.

Dobin, A. *et al.* (2013) 'STAR: ultrafast universal RNA-seq aligner', *Bioinformatics*, 29(1), pp. 15–21. doi: 10.1093/bioinformatics/bts635.

Doench, J. G. *et al.* (2016) 'Optimized sgRNA design to maximize activity and minimize off-target effects of CRISPR-Cas9', *Nature Biotechnology*. Nature Publishing Group, 34(2), pp. 184–191. doi: 10.1038/nbt.3437.

Doudna, J. A. and Charpentier, E. (2014) 'The new frontier of genome engineering with CRISPR-Cas9', *Science*, 346(6213), pp. 1077–1087. doi: 10.1126/science.1258096.

Duarte, R. R. R. *et al.* (2016) 'Genome-Wide Significant Schizophrenia Risk Variation on Chromosome 10q24 Is Associated With Altered cis -Regulation of BORCS7 , AS3MT , and NT5C2 in the Human Brain', *American Journal of Medical Genetics Part B: Neuropsychiatric Genetics*, (March), pp. 806–814. doi: 10.1002/ajmg.b.32445.

Dubridge, R. B. *et al.* (1987) 'Analysis of Mutation in Human Cells by Using an Epstein-Barr Virus Shuttle System', *Molecular and Cellular Biology*, 7(1), pp. 379–387.

Egger, G. *et al.* (2004) 'Epigenetics in human disease and prospects for epigenetic therapy', *Nature*, 429(6990), pp. 457–463. doi: 10.1038/nature02625.

Ellegood, J. *et al.* (2014) 'Neuroanatomical phenotypes in a mouse model of the 22q11.2 microdeletion', *Molecular Psychiatry*, 19(1), pp. 99–107. doi: 10.1038/mp.2013.112.

Ellison-Wright, I. and Bullmore, E. (2009) 'Meta-analysis of diffusion tensor imaging studies in schizophrenia', *Schizophrenia Research*. Elsevier B.V., 108(1–3), pp. 3–10. doi: 10.1016/j.schres.2008.11.021.

Encinas, M. *et al.* (2000) 'Sequential Treatment of SH-SY5Y Cells with Retinoic Acid and Brain-Derived Neurotrophic Factor Gives Rise to Fully Differentiated, Neurotrophic Factor-Dependent, Human Neuron-Like Cells', *Journal of Neurochemistry*, 75(3), pp. 991–1003.

Eppig, J. T. and Eppig, J. T. (2017) 'Mouse Genome Informatics (MGI) Resource : Genetic, Genomic, and Biological Knowledgebase for the Laboratory Mouse', *ILAR*, 58(1), pp. 17–41. doi: 10.1093/ilar/ilx013.

Ewels, P., Lundin, S. and Max, K. (2016) 'MultiQC: summarize analysis results for multiple tools and samples in a single report', *Bioinformatics*, 32(19), pp. 3047–3048. doi: 10.1093/bioinformatics/btw354.

Ewing, B. and Green, P. (1998) 'Base-calling of automated sequencer traces using phred. I. Accuracy Assessment', *Genome Research*, 8(3), pp. 186–194. doi: 10.1101/gr.8.3.186.

Faris, R. and Dunham, H. (1939) ‘Mental disorders in urban areas: an ecological study of schizophrenia and other psychoses’, *University of Chicago Press*. doi: 10.5962/bhl.title.7369.

Fatemi, S. H. (2001) ‘Reelin mutations in mouse and man: From reeler mouse to schizophrenia, mood disorders, autism and lissencephaly’, *Molecular Psychiatry*, 6(2), pp. 129–133. doi: 10.1038/sj.mp.4000129.

Fatemi, S. H. and Folsom, T. D. (2009) ‘The Neurodevelopmental Hypothesis of Schizophrenia, Revisited’, *Schizophrenia Bulletin*, 35(3), pp. 528–548. doi: 10.1093/schbul/sbn187.

Fearon, P. *et al.* (2006) ‘Incidence of schizophrenia and other psychoses in ethnic minority groups: Results from the MRC AESOP Study’, *Psychological Medicine*, 36(11), pp. 1541–1550. doi: 10.1017/S0033291706008774.

Fernández-Cabrera, M. R. *et al.* (2018) ‘Selective effects of Δ9-tetrahydrocannabinol on medium spiny neurons in the striatum’, *PLoS ONE*, 13(7), pp. 1–20. doi: 10.1371/journal.pone.0200950.

Fisher, H. L. *et al.* (2015) ‘Methylomic analysis of monozygotic twins discordant for childhood psychotic symptoms’, *Epigenetics*, 10(11), pp. 1014–1023.

Flores, K. *et al.* (2012) ‘Genome-wide association between DNA methylation and alternative splicing in an invertebrate’, *BMC Genomics*. BMC Genomics, 13(1), p. 1. doi: 10.1186/1471-2164-13-480.

Fone, K. C. and Porkess, M. V. (2008) ‘Behavioural and neurochemical effects of post-weaning social isolation in rodents-Relevance to developmental neuropsychiatric disorders’, *Neuroscience and Biobehavioral Reviews*, 32(6), pp. 1087–1102. doi: 10.1016/j.neubiorev.2008.03.003.

Di Forti, M. *et al.* (2019) ‘The contribution of cannabis use to variation in the incidence of psychotic disorder across Europe (EU-GEI): a multicentre case-control study’, *The Lancet Psychiatry*, 6(5), pp. 427–436. doi: 10.1016/S2215-0366(19)30048-3.

Freeman, T. P. *et al.* (2019) ‘Increasing potency and price of cannabis in Europe, 2006–16’, *Addiction*, 114(6), pp. 1015–1023. doi: 10.1111/add.14525.

Fromer, M. *et al.* (2014) ‘De novo mutations in schizophrenia implicate synaptic networks’, *Nature*, 506(7487), pp. 179–184. doi: 10.1038/nature12929.

Fromer, M., Roussos, P., Sieberts, Solveig K., *et al.* (2016) ‘Gene expression elucidates functional impact of polygenic risk for schizophrenia’, *Nature Neuroscience*, 19(11), pp. 1442–1453. doi: 10.1038/nn.4399.

Fromer, M., Roussos, P., Sieberts, Solveig K, *et al.* (2016) ‘Gene expression elucidates functional impact of polygenic risk for schizophrenia’, *Nature Neuroscience*, (September). doi: 10.1038/nn.4399.

Frommer, M. *et al.* (1992) ‘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*, 89(March), pp. 1827–1831.

Fukata, M. *et al.* (2004) ‘Identification of PSD-95 Palmitoylating Enzymes’, *Neuron*, 44, pp. 987–996.

Fukata, Y. and Fukata, M. (2010) ‘Protein palmitoylation in neuronal development and synaptic plasticity’, *Nature reviews. Neuroscience*. Nature Publishing Group, 11, pp. 161–175. doi: 10.1038/nrn2788.

Gage, S. H., Hickman, M. and Zammit, S. (2016) ‘Association between cannabis and psychosis: Epidemiologic evidence’, *Biological Psychiatry*. Elsevier, 79(7), pp. 549–556. doi: 10.1016/j.biopsych.2015.08.001.

Garneau, J. E. *et al.* (2010) ‘The CRISPR/cas bacterial immune system cleaves bacteriophage and plasmid DNA’, *Nature*, 468(7320), pp. 67–71. doi: 10.1038/nature09523.

Gartler, S. M. and Riggs, A. D. (1983) ‘Mammalian X-Chromosome Inactivation’, *Annual Reviews Genetics*, 17, pp. 155–190. Available at: www.annualreviews.org.

Gaudelli, N. M. *et al.* (2017) ‘Programmable base editing of A T to G C in genomic DNA without DNA cleavage’, *Nature*. Nature Publishing Group, 551(7681), pp. 464–471. doi: 10.1038/nature24644.

Geddes, J. R. *et al.* (1999) ‘Schizophrenia and complications of pregnancy and labor: An individual patient data meta-analysis’, *Schizophrenia Bulletin*, 25(3), pp. 413–423. doi: 10.1093/oxfordjournals.schbul.a033389.

Gerra, M. C. *et al.* (2018) ‘Gene variants and educational attainment in cannabis use: Mediating role of DNA methylation’, *Translational Psychiatry*. Springer US, 8(1), pp. 1–11. doi: 10.1038/s41398-017-0087-1.

Gilbert, L. A. *et al.* (2013) ‘CRISPR-mediated modular RNA-guided regulation of transcription in eukaryotes’, *Cell*. Elsevier Inc., 154(2), p. 442. doi: 10.1016/j.cell.2013.06.044.

Grayson, D. R. *et al.* (2005) ‘Reelin promoter hypermethylation in schizophrenia’, *Proceedings of the National Academy of Sciences of the United States of America*, 102(26), pp. 9341–9346. doi: 10.1073/pnas.0503736102.

Guidotti, A. *et al.* (2000) ‘Decrease in Reelin and Glutamic Acid Decarboxylase 67 (GAD 67) Expression in Schizophrenia and Bipolar Disorder’, *Archives of general psychiatry*, 57, pp. 1061–1069.

Guilmatre, A. *et al.* (2014) ‘The Emerging Role of SHANK Genes in Neuropsychiatric Disorders’, *Developmental Neurobiology*, 74(2), pp. 113–122. doi: 10.1002/dneu.22128.

Guo, J. Y., Ragland, J. D. and Carter, C. S. (2019) ‘Memory and cognition in schizophrenia’, *Molecular Psychiatry*. Springer US, 24(5), pp. 633–642. doi: 10.1038/s41380-018-0231-1.

Gusev, A. *et al.* (2018) ‘Transcriptome-wide association study of schizophrenia and chromatin activity yields mechanistic disease insights’, *Nature Genetics*, 50, pp. 538–548. doi: 10.1038/s41588-018-0092-1.

Gutierrez-Arcelus, M. *et al.* (2015) ‘Tissue-Specific Effects of Genetic and Epigenetic Variation on Gene Regulation and Splicing’, *PLoS Genetics*, 11(1), pp. 1–25. doi:

10.1371/journal.pgen.1004958.

Häfner, H. *et al.* (1994) 'The Epidemiology of Early Schizophrenia: Influence of Age and Gender on Onset and Early Course', *British Journal of Psychiatry*, 164(S23), pp. 29–38. doi: 10.1192/s0007125000292714.

Hamosh, A. *et al.* (2002) 'Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders', *Nucleic Acids Research*, 30(1), pp. 52–55.

Hanley, P. J. *et al.* (2010) 'Motorized RhoGAP myosin IXb (Myo9b) controls cell shape and motility', *Proceedings of the National Academy of Sciences of the United States of America*, 107(27), pp. 12145–12150. doi: 10.1073/pnas.0911986107.

Hannon, E., Dempster, E., *et al.* (2016) 'An integrated genetic-epigenetic analysis of schizophrenia: evidence for co-localization of genetic associations and differential DNA methylation', *Genome Biology*. Genome Biology, 17(1), p. 176. doi: 10.1186/s13059-016-1041-x.

Hannon, E., Spiers, H., *et al.* (2016) 'Methylation quantitative trait loci in the developing brain and their enrichment in schizophrenia-associated genomic regions', *Nature neuroscience*, 19(1), pp. 48–54. doi: 10.4315/0362-028X.JFP-13-395.Knowledge.

Hannon, E. *et al.* (2018) 'Leveraging DNA methylation quantitative trait loci to characterize the relationship between methylomic variation, gene expression and complex traits.', *The American Journal of Human Genetics*. ElsevierCompany., 103, pp. 1–12. doi: 10.1101/297176.

Harada, A. *et al.* (2001) 'MAP2 is required for dendrite elongation , PKA anchoring in dendrites , and proper PKA signal transduction', *Journal of Cell Biology*, 158(3), pp. 541–549. doi: 10.1083/jcb.200110134.

Harrington, C. T. *et al.* (2013) 'Fundamentals of pyrosequencing', *Archives of Pathology and Laboratory Medicine*, 137(9), pp. 1296–1303. doi: 10.5858/arpa.2012-0463-RA.

Harrison, P. J. (2004) 'The hippocampus in schizophrenia : a review of the neuropathological evidence and its pathophysiological implications', *Psychopharmacology*, 174, pp. 151–162. doi: 10.1007/s00213-003-1761-y.

Harrow, J. *et al.* (2012) 'GENCODE: The reference human genome annotation for The ENCODE Project', *Genome Research*, 22(9), pp. 1760–1774. doi: 10.1101/gr.135350.111.

Haukvik, U. K. *et al.* (2018) 'Neuroimaging hippocampal subfields in schizophrenia and bipolar disorder: A systematic review and meta-analysis', *Journal of Psychiatric Research*, 104(July), pp. 217–226. doi: 10.1016/j.jpsychires.2018.08.012.

Haute, L. Van *et al.* (2013) 'Knockout mice created by TALEN-mediated gene targeting', *Nature Biotechnology*, 31(1), pp. 23–24.

He, Y.-F. *et al.* (2011) 'Tet-Mediated Formation of 5-Carboxylcytosine and Its Excision by TDG in Mammalian DNA', *Science*, 333(September), pp. 225–240. doi: 10.1016/b978-0-408-01434-2.50020-6.

Henikoff, S. and Matzke, M. (1997) 'Exploring and explaining epigenetic effects', *Trends in Genetics*, 13(8), pp. 293–295. doi: 10.1016/S0168-9525(97)01219-5.

Hill, M. and Iaa, N. W. (1977) 'Characteristics of a Human Cell Line Transformed by DNA from Human Adenovirus Type 5', *Journal of General Virology*, 36, pp. 59–74.

Hilton, I. B. *et al.* (2015) 'Epigenome editing by a CRISPR-Cas9-based acetyltransferase activates genes from promoters and enhancers', *Nature Biotechnology*, 33(5), pp. 510–517. doi: 10.1038/nbt.3199.

Hoon, Y. *et al.* (2014) 'Generation of knockout mice using engineered nucleases', *Methods*. Elsevier Inc., 69(1), pp. 85–93. doi: 10.1016/j.ymeth.2014.02.009.

Hsu, P. D. *et al.* (2013) 'DNA targeting specificity of RNA-guided Cas9 nucleases', *Nature Biotechnology*, 31(9), pp. 827–832. doi: 10.1038/nbt.2647.

Hsu, P. D., Lander, E. S. and Zhang, F. (2014) 'Development and applications of CRISPR-Cas9 for genome engineering', *Cell*. Elsevier, 157(6), pp. 1262–1278. doi: 10.1016/j.cell.2014.05.010.

Huang, L. *et al.* (2015) 'Effects of promoter leakage on dynamics of gene expression', *BMC Systems Biology*, 9(16), pp. 1–12. doi: 10.1186/s12918-015-0157-z.

Hubbard, S. R. and Till, J. H. (2000) 'Protein Tyrosine Kinase Structure and Function', *Annual Review of Biochemistry*, 69, pp. 373–98.

Huckins, L. M. *et al.* (2019) 'Gene expression imputation across multiple brain regions provides insights into schizophrenia risk', *Nature Genetics*, 51(4), pp. 659–674. doi: 10.1038/s41588-019-0364-4.

Hwang, W. Y. *et al.* (2013) 'Efficient genome editing in zebrafish using a CRISPR-Cas system', *Nat Biotechnol*, 31(3), pp. 227–229. doi: 10.1038/nbt.2501.

Iii, L. F. G. and Szalay, A. A. (2002) 'Imaging of light emission from the expression of luciferases in living cells and organisms: a review', *Luminescence*, 17, pp. 43–74.

Illumina (2015) 'Infinium ® HD Assay Methylation Protocol Guide', (November).

Impagnatiello, F. *et al.* (1998) 'A decrease of reelin expression as a putative vulnerability factor in schizophrenia', *Proceedings of the National Academy of Sciences of the United States of America*, 95(26), pp. 15718–15723. doi: 10.1073/pnas.95.26.15718.

Ishino, Y. *et al.* (1987) 'Nucleotide Sequence of the iap Gene , Responsible for Alkaline Phosphatase Isozyme Conversion in Escherichia coli , and Identification of the Gene Product', *Journal of Bacteriology*, 169(12), pp. 5429–5433.

Ito, S. *et al.* (2011) 'Tet proteins can convert 5-methylcytosine to 5-formylcytosine and 5-carboxylcytosine', *Science*, 333(6047), pp. 1300–1303. doi: 10.1126/science.1210597.

Iwamoto, K. *et al.* (2005) 'DNA methylation status of SOX10 correlates with its downregulation and oligodendrocyte dysfunction in schizophrenia', *Journal of Neuroscience*, 25(22), pp. 5376–5381. doi: 10.1523/JNEUROSCI.0766-05.2005.

Iwasaki, Y. W., Siomi, M. C. and Siomi, H. (2015) 'PIWI-Interacting RNA: Its

Biogenesis and Functions', *Annual Review of Biochemistry*, 84(1), pp. 405–433. doi: 10.1146/annurev-biochem-060614-034258.

Jacobs, J. L. and Dinman, J. D. (2004) 'Systematic analysis of bicistronic reporter assay data', *Nucleic acids research*, 32(20), pp. 1–10. doi: 10.1093/nar/gnh157.

Jaenisch, R. and Bird, A. (2003) 'Epigenetic regulation of gene expression: How the genome integrates intrinsic and environmental signals', *Nature Genetics*, 33(3S), pp. 245–254. doi: 10.1038/ng1089.

Jaffe, A. E. *et al.* (2012) 'Bump hunting to identify differentially methylated regions in epigenetic epidemiology studies', *International Journal of Epidemiology*, 41(1), pp. 200–209. doi: 10.1093/ije/dyr238.

Jaffe, A. E. *et al.* (2016) 'Mapping DNA methylation across development, genotype and schizophrenia in the human frontal cortex', *Nature Neuroscience*, 19(1546-1726 (Electronic)), pp. 40–47. doi: 10.1038/nn.4181.

Jager, P. L. De *et al.* (2015) 'Alzheimery's disease pathology is associated with early alterations in brain DNA methylation at ANK1, BIN1, RHBDF2 and other loci', *Nature Neuroscience*, 17(9), pp. 1156–1163. doi: 10.1038/nn.3786.Alzheimery.

Jahn, K., Blumer, C. W. N. and Pathak, M. M. H. (2017) 'A cell culture model for investigation of synapse influenceability: epigenetics , expression and function of gene targets important for synapse formation and preservation in SH-SY5Y neuroblastoma cells differentiated by retinoic acid', *Journal of Neural Transmission*. Springer Vienna, 124(11), pp. 1341–1367. doi: 10.1007/s00702-017-1769-9.

Janhunen, S. K. *et al.* (2015) 'The subchronic phencyclidine rat model: Relevance for the assessment of novel therapeutics for cognitive impairment associated with schizophrenia', *Psychopharmacology*, 232(21–22), pp. 4059–4083. doi: 10.1007/s00213-015-3954-6.

Jansen, R. *et al.* (2002) 'Identification of genes that are associated with DNA repeats in prokaryotes', 43, pp. 1565–1575.

Jeevakumar, V. *et al.* (2015) 'Ketamine administration during the second postnatal week induces enduring schizophrenia-like behavioral symptoms and reduces parvalbumin expression in the medial prefrontal cortex of adult mice', *Behavioural Brain Research*. Elsevier B.V., 282, pp. 165–175. doi: 10.1016/j.bbr.2015.01.010.

Jelks, K. B. *et al.* (2007) 'Estradiol targets synaptic proteins to induce glutamatergic synapse formation in cultured hippocampal neurons: Critical role of estrogen receptor-α', *Journal of Neuroscience*, 27(26), pp. 6903–6913. doi: 10.1523/JNEUROSCI.0909-07.2007.

Jeonghae, C. *et al.* (2012) 'DNA methylation of the 5'-untranslated region at +298 and +351 represses BACE1 expression in mouse BV-2 microglial cells', *Biochemical and Biophysical Research Communications*. Elsevier Inc., 417(1), pp. 387–392. doi: 10.1016/j.bbrc.2011.11.123.

Jiang, T. *et al.* (2017) 'Variation in global DNA hydroxymethylation with age associated with schizophrenia', *Psychiatry Research*. Elsevier Ireland Ltd, 257(April), pp. 497–500. doi: 10.1016/j.psychres.2017.08.022.

Jinek, M. *et al.* (2012) 'A Programmable Dual-RNA – Guided DNA Endonuclease in

Adaptive Bacterial Immunity', *Science (New York, N.Y.)*, 337(August), pp. 816–822. doi: 10.1126/science.1225829.

Jirtle, R. L. and Skinner, M. K. (2007) 'Environmental epigenomics and disease susceptibility', *Nature Review Genetics*, 8(8), pp. 253–262. doi: 10.1038/nrg2045.

Jones, P. A., Baylin, S. B. and Kimmel, S. (2002) 'The Fundamental Role of Epigenetic Events in Cancer', *Nature Reviews Genetics*, 3, pp. 415–428. doi: 10.1038/nrg816.

Juven-gershon, T. and Kadonaga, J. T. (2010) 'Regulation of gene expression via the core promoter and the basal transcriptional machinery', *Developmental Biology*. Elsevier Inc., 339(2), pp. 225–229. doi: 10.1016/j.ydbio.2009.08.009.

Kamiya, A. *et al.* (2006) 'DISC1–NDEL1/NUDEL protein interaction, an essential component for neurite outgrowth, is modulated by genetic variations of DISC1', *Human Molecular Genetics*, 15(22), pp. 3313–3323. doi: 10.1093/hmg/ddl407.

Kasowski, M. *et al.* (2013) 'Extensive variation in chromatin states across humans', *Science*, 342(6159), pp. 750–752. doi: 10.1126/science.1242510.

Kaufman, M. and Evans, M. (1981) 'Establishment in culture of pluripotential cells from mouse embryos', *Nature*, 292(July), pp. 154–156.

Keane, T. M. *et al.* (2011) 'Mouse genomic variation and its effect on phenotypes and gene regulation', *Nature*, 477(7364), pp. 289–294. doi: 10.1038/nature10413.

Kendell, R. E. (1987) 'Diagnosis and classification of functional psychoses', *British Medical Bulletin*, 43(3), pp. 499–513.

Khare, T. *et al.* (2012) '5-hmC in the brain is abundant in synaptic genes and shows differences at the exon-intron boundary', *Nature Structural and Molecular Biology*. Nature Publishing Group, 19(10), pp. 1037–1044. doi: 10.1038/nsmb.2372.

Khashan, A. S. *et al.* (2008) 'Higher Risk of Offspring Schizophrenia Following Antenatal Maternal Exposure to Severe Adverse Life Events', *Archives of general psychiatry*, 65(2), pp. 146–152.

Kim, D. *et al.* (2015) 'Digenome-seq: Genome-wide profiling of CRISPR-Cas9 off-target effects in human cells', *Nature Methods*, 12(3), pp. 237–243. doi: 10.1038/nmeth.3284.

Kim, J. Y. *et al.* (2012) 'Interplay between DISC1 and GABA Signaling Regulates Neurogenesis in Mice and Risk for Schizophrenia', *Cell*. Elsevier Inc., 148(5), pp. 1051–1064. doi: 10.1016/j.cell.2011.12.037.

Kim, Y. G., Cha, J. and Chandrasegaran, S. (1996) 'Hybrid restriction enzymes: Zinc finger fusions to Fok I cleavage domain', *Proceedings of the National Academy of Sciences of the United States of America*, 93(3), pp. 1156–1160. doi: 10.1073/pnas.93.3.1156.

Kirkbride, J. B. *et al.* (2006) 'Heterogeneity in incidence rates of schizophrenia and other psychotic syndromes: Findings from the 3-center ÆSOP study', *Archives of General Psychiatry*, 63(3), pp. 250–258. doi: 10.1001/archpsyc.63.3.250.

Klegeris, A., Bissonnette, C. J. and Mcgeer, P. L. (2003) 'Reduction of human

monocytic cell neurotoxicity and cytokine secretion by ligands of the cannabinoid-type CB2 receptor', *British Journal of Pharmacology*, 139, pp. 775–786. doi: 10.1038/sj.bjp.0705304.

Kleinstiver, B. P. *et al.* (2016) 'High-fidelity CRISPR-Cas9 variants with undetectable genome- wide off-targets', *Nature*, 528(7587), pp. 490–495. doi: 10.1001/jamasurg.2014.1086.Feasibility.

Klinge, C. M. (2001) 'Estrogen receptor interaction with estrogen response elements', *Nucleic Acids Research*, 29(14), pp. 2905–2919.

Klose, R. J. and Bird, A. P. (2006) 'Genomic DNA methylation: The mark and its mediators', *Trends in Biochemical Sciences*, 31(2), pp. 89–97. doi: 10.1016/j.tibs.2005.12.008.

Klug, M. and Rehli, M. (2006) 'Functional Analysis of Promoter CpG Methylation Using a CpG-Free Luciferase Reporter Vector', *Epigenetics*, 1(3), pp. 127–130. doi: 10.4161/epi.1.3.3327.

Koenders, L. *et al.* (2016) 'Grey matter changes associated with heavy cannabis use: A longitudinal sMRI study', *PLoS ONE*, 11(5), pp. 1–13. doi: 10.1371/journal.pone.0152482.

Kolb, B. *et al.* (2018) 'THC alters alters morphology of neurons in medial prefrontal cortex, orbital prefrontal cortex, and nucleus accumbens and alters the ability of later experience to promote structural plasticity', *Synapse*, 72(3). doi: 10.1002/syn.22020.

Komor, A. C. *et al.* (2016) 'Programmable editing of target base in genomic DNA without double-stranded DNA cleavage', *Nature*. Nature Publishing Group, 533(7603), pp. 420–424. doi: 10.1038/nature17946.

Korzhevskii, D. E., Karpenko, M. N. and Kirik, O. V (2012) 'Microtubule-Associated Proteins as Indicators of Differentiation and the Functional State of Nerve Cells', *Neuroscience and Behavioral Physiology*, 42(3), pp. 215–222.

Kovalevich, J. and Langford, D. (2013) 'Considerations for the Use of SH-SY5Y Neuroblastoma Cells in Neurobiology', *Methods in Molecular Biology*, 1078, pp. 9–21. doi: 10.1007/978-1-62703-640-5.

Kowiański, P. *et al.* (2018) 'BDNF: A Key Factor with Multipotent Impact on Brain Signaling and Synaptic Plasticity', *Cellular and Molecular Neurobiology*, 38(3), pp. 579–593. doi: 10.1007/s10571-017-0510-4.

Krawitz, P. M. *et al.* (2012) 'Mutations in PIGO, a member of the GPI-anchor-synthesis pathway, cause hyperphosphatasia with mental retardation', *American Journal of Human Genetics*, 91(1), pp. 146–151. doi: 10.1016/j.ajhg.2012.05.004.

Kriaucionis, S. and Heintz, N. (2009) 'The nuclear DNA base 5-hydroxymethylcytosine is present in purkinje neurons and the brain', *Science*, 324(5929), pp. 929–930. doi: 10.1126/science.1169786.

Kumar, M. *et al.* (2001) 'Systematic determination of the packaging limit of lentiviral vectors', *Human Gene Therapy*, 12(15), pp. 1893–1905. doi: 10.1089/104303401753153947.

Kumarasinghe, N., Tooney, P. A. and Schall, U. (2012) 'Finding the needle in the

haystack: A review of microarray gene expression research into schizophrenia', *Australian and New Zealand Journal of Psychiatry*, 46(7), pp. 598–610. doi: 10.1177/0004867412442405.

Kuramoto, T. *et al.* (2001) 'Attractin/mahogany/zitter plays a critical role in myelination of the central nervous system', *Proceedings of the National Academy of Sciences of the United States of America*, 98(2), pp. 559–564. doi: 10.1073/pnas.98.2.559.

Kwon, D. Y. *et al.* (2017) 'Locus-specific histone deacetylation using a synthetic CRISPR-Cas9-based HDAC', *Nature Communications*. Nature Publishing Group, 8(May), pp. 1–8. doi: 10.1038/ncomms15315.

Layer, P. G. (1991) 'Cholinesterases During Development of the Avian Nervous System', *Cellular and Molecular Neurobiology*, 11(1), pp. 7–33.

Lee, S. *et al.* (2015) 'Disrupted-in-schizophrenia 1 (DISC1) Regulates Dysbindin Function by Enhancing Its Stability', *Journal of Biological Chemistry*, 290(11), pp. 7087–7096. doi: 10.1074/jbc.M114.614750.

Lenzi, A. *et al.* (2003) 'Effects of Rivastigmine on Cognitive Function and Quality of Life in Patients With Schizophrenia', *Clinical Neuropharmacology*, 26(6), pp. 317–321.

Lerner, M. R. *et al.* (1980) 'Are snRNPs involved in splicing?', *Nature*, 283(5743), pp. 220–224. doi: 10.1038/283220a0.

Lew, G. M. (1996) 'Tau Protein after in a Human Neuroblastoma Cell Line', *General Pharmacology*, 27(7), pp. 1141–1143.

Li, L. *et al.* (2016) 'Evidence of AS3MT d2d3 -Associated Variants within 10q24.32–33 in the Genetic Risk of Major Affective Disorders', *Molecular Neuropsychiatry*, 2(4), pp. 213–218. doi: 10.1159/000452998.

Li, M. *et al.* (2016) 'A human-specific AS3MT isoform and BORCS7 are molecular risk factors in the 10q24.32 schizophrenia-associated locus.', *Nature medicine*. Nature Publishing Group, 22(6), pp. 649–656. doi: 10.1038/nm.4096.

Liao, Y., Smyth, G. K. and Shi, W. (2014) 'Sequence analysis featureCounts: an efficient general purpose program for assigning sequence reads to genomic features', *Bioinformatics*, 30(7), pp. 923–930. doi: 10.1093/bioinformatics/btt656.

Lin, J. H., Walter, P. and Yen, T. S. B. (2007) 'Endoplasmic Reticulum Stress in Disease Pathogenesis', *Annual Review of Pathology: Mechanisms of Disease*, 0(0), p. 071003161323003. doi: 10.1146/annurev.pathol.3.121806.151434.

Lin, L. *et al.* (2018) 'Genome-wide determination of on-target and off-target characteristics for RNA-guided DNA methylation by dCas9 methyltransferases', *GigaScience*, 7(3), pp. 1–19. doi: 10.1093/gigascience/giy011.

Lin, M. *et al.* (2016) 'Integrative transcriptome network analysis of iPSC-derived neurons from schizophrenia and schizoaffective disorder patients with 22q11.2 deletion', *BMC Systems Biology*. BMC Systems Biology, 10(1), pp. 1–20. doi: 10.1186/s12918-016-0366-0.

Lister, R. *et al.* (2008) 'Resource Highly Integrated Single-Base Resolution Maps of

the Epigenome in Arabidopsis', *Cell*, 133(May), pp. 523–536. doi: 10.1016/j.cell.2008.03.029.

Lister, R. *et al.* (2009) 'Human DNA methylomes at base resolution show widespread epigenomic differences', *Nature*. Nature Publishing Group, 462(7271), pp. 315–322. doi: 10.1038/nature08514.

Liu, C. X. *et al.* (2000) 'Genomic organization of a new candidate tumor suppressor gene, LRP1B', *Genomics*, 69(2), pp. 271–274. doi: 10.1006/geno.2000.6331.

Liu, J. *et al.* (2012) 'Efficient and Specific Modifications of the Drosophila Genome by Means of an Easy TALEN Strategy', *Journal of Genetics and Genomics*. Elsevier Limited and Science Press, 39(5), pp. 209–215. doi: 10.1016/j.jgg.2012.04.003.

Liu, X. S. *et al.* (2016) 'Editing DNA Methylation in the Mammalian Genome', *Cell*. Elsevier Inc., 167(1), pp. 233-247.e17. doi: 10.1016/j.cell.2016.08.056.

Liu, X. S. *et al.* (2018) 'Rescue of Fragile X Syndrome Neurons by DNA Methylation Editing of the FMR1 Gene Article Rescue of Fragile X Syndrome Neurons by DNA Methylation Editing of the FMR1 Gene', *Cell*. Elsevier Inc., pp. 1–14. doi: 10.1016/j.cell.2018.01.012.

Liu, Yi *et al.* (2019) 'Functional analysis of haplotypes and promoter activity at the 5′ region of the human GABRB3 gene and associations with schizophrenia', *Molecular Genetics and Genomic Medicine*, 7(5), pp. 1–10. doi: 10.1002/mgg3.652.

Livak, K. J. and Schmittgen, T. D. (2001) 'Analysis of Relative Gene Expression Data Using Real- Time Quantitative PCR and the 2-ΔΔCT Method', *Methods*, 28, pp. 402–408. doi: 10.1006/meth.2001.1262.

Lizio, M. *et al.* (2015) 'Gateways to the FANTOM5 promoter level mammalian expression atlas', *Genome Biology*, 16(1), pp. 1–14. doi: 10.1186/s13059-014-0560-6.

Llufrio, E. M. *et al.* (2018) 'Redox Biology Sorting cells alters their redox state and cellular metabolome', *Redox Biology*. Elsevier B.V., 16(February), pp. 381–387. doi: 10.1016/j.redox.2018.03.004.

Lodato, M. A. *et al.* (2015) 'Somatic mutation in single human neurons tracks developmental and transcriptional history', *Science*, 350(6256), pp. 94–98. doi: 10.1126/science.aab1785.

Lodge, D. J. and Grace, A. A. (2009) 'Gestational methylazoxymethanol acetate administration: A developmental disruption model of schizophrenia', *Behavioural Brain Research*, 204(2), pp. 306–312. doi: 10.1016/j.bbr.2009.01.031.

Lojpur, T. *et al.* (2019) 'Δ9-Tetrahydrocannabinol leads to endoplasmic reticulum stress and mitochondrial dysfunction in human BeWo trophoblasts', *Reproductive Toxicology*. Elsevier, 87(April), pp. 21–31. doi: 10.1016/j.reprotox.2019.04.008.

Long, H. *et al.* (2013) 'Myo9b and RICS modulate dendritic morphology of cortical neurons', *Cerebral Cortex*, 23(1), pp. 71–79. doi: 10.1093/cercor/bhr378.

Love, M. I., Anders, S. and Huber, W. (2016) *DESeq2 vignette*. doi: 110.1186/s13059-014-0550-8.

Love, M. I., Huber, W. and Anders, S. (2014) 'Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2', *Genome Biology*, 15(12), pp. 1–21. doi: 10.1186/s13059-014-0550-8.

Luirink, J. and Sinning, I. (2004) 'SRP-mediated protein targeting: structure and function revisited', *Biochimica et Biophysica Acta - Molecular Cell Research*, 1694(1–3), pp. 17–35. doi: 10.1016/j.bbamcr.2004.03.013.

Lunnon, K. and Mill, J. (2013) 'Epigenetic studies in Alzheimer's disease: current findings, caveats and considerations for future studies', *American Journal of Medical Genetics Part B: Neuropsychiatric Genetics*, 162(8), pp. 789–799. doi: 10.1002/ajmg.b.32201.Epigenetic.

Lykke-andersen, S. and Jensen, T. H. (2015) 'Nonsense-mediated mRNA decay : an intricate machinery that shapes transcriptomes', *Nature Publishing Group*. Nature Publishing Group, 16(11), pp. 665–677. doi: 10.1038/nrm4063.

Lyko, F. (2018) 'The DNA methyltransferase family: a versatile toolkit for epigenetic regulation', *Nature Reviews Genetics*. Nature Publishing Group, 19(2), pp. 81–92. doi: 10.1038/nrg.2017.80.

Lynch, M. A. (2004) 'Long-Term Potentiation and Memory', *Physiological Reviews*, 84(1), pp. 87–136. doi: 10.1152/physrev.00014.2003.

Ma, C. *et al.* (2018) 'The integrated landscape of causal genes and pathways in schizophrenia', *Translational Psychiatry*, 8(67), pp. 1–14. doi: 10.1038/s41398-018-0114-x.

Ma, N., Yin, J. and Chen, G. (2019) 'Transcriptome Analysis of Small Molecule–Mediated Astrocyte-to-Neuron Reprogramming', *Frontiers in Cell and Developmental Biology*, 7(May), pp. 1–17. doi: 10.3389/fcell.2019.00082.

Makarova, K. S. *et al.* (2011) 'Evolution and classification of the CRISPR-Cas systems', *Nature Reviews Microbiology*. Nature Publishing Group, 9(6), pp. 467–477. doi: 10.1038/nrmicro2577.

Malaspina, D. *et al.* (2001) 'Advancing paternal age and the risk of schizophrenia', *Archives of General Psychiatry*, 58(4), pp. 361–367. doi: 10.1001/archpsyc.58.4.361.

Mansell, G. *et al.* (2019) 'Guidance for DNA methylation studies : statistical insights from the Illumina EPIC array', *BMC Genomics*. BMC Genomics, 20(366), pp. 1–15.

Mansouri, M. R. *et al.* (2005) 'Loss of ZDHHC15 expression in a woman with a balanced translocation t(X;15)(q13.3;cen) and severe mental retardation', *European Journal of Human Genetics*, 13, pp. 970–977. doi: 10.1038/sj.ejhg.5201445.

Marques, S. M. and Esteves, J. C. G. (2009) 'Critical Review Firefly Bioluminescence: A Mechanistic Approach of Luciferase Catalyzed Reactions', *Life*, 61(January), pp. 6–17. doi: 10.1002/iub.134.

Marshall, C. R. *et al.* (2017) 'Contribution of copy number variants to schizophrenia from a genome-wide study of 41,321 subjects', *Nature Genetics*, 49(1), pp. 27–35. doi: 10.1038/ng.3725.

Martín-Maestro, P. *et al.* (2017) 'Slower Dynamics and Aged Mitochondria in Sporadic Alzheimer's Disease', *Oxidative Medicine and Cellular Longevity*, 2017.

doi: 10.1155/2017/9302761.

Martin, M. (2011) ‘Cutadapt removes adapter sequences from high-throughput sequencing reads’, *EMBnet.journal*, 17(1), pp. 10–12.

Massoud, T. F. *et al.* (2007) ‘Reporter gene imaging of protein–protein interactions in living subjects’, *Current Opinion in Biotechnology*, 18, pp. 31–37. doi: 10.1016/j.copbio.2007.01.007.

Matera, A. G., Terns, R. M. and Terns, M. P. (2007) ‘Non-coding RNAs: Lessons from the small nuclear and small nucleolar RNAs’, *Nature Reviews Molecular Cell Biology*, 8(3), pp. 209–220. doi: 10.1038/nrm2124.

Mattick, J. S. and Makunin, I. V (2006) ‘Non-coding RNA’, *Human molecular genetics*, 15 Spec No(1), pp. 17–29. doi: 10.1093/hmg/ddl046.

Maunakea, A. K. *et al.* (2013) ‘Intragenic DNA methylation modulates alternative splicing by recruiting MeCP2 to promote exon recognition’, *Cell Research*. Nature Publishing Group, 23(11), pp. 1256–1269. doi: 10.1038/cr.2013.110.

Maurano, M. T. *et al.* (2012) ‘Systematic Localization of Common Disease-Associated Variation in Regulatory DNA’, *Science*, 337(September), pp. 1190–1196.

McCarroll, S. A. and Altshuler, D. M. (2007) ‘Copy-number variation and association studies of human disease’, *Nature Genetics*, 39(7S), pp. S37–S42. doi: 10.1038/ng2080.

McCutcheon, R. A., Abi-Dargham, A. and Howes, O. D. (2019) ‘Schizophrenia, Dopamine and the Striatum: From Biology to Symptoms’, *Trends in Neurosciences*. Elsevier Ltd, 42(3), pp. 205–220. doi: 10.1016/j.tins.2018.12.004.

McGrath, J. J. (2005) ‘Myths and plain truths about schizophrenia epidemiology - The NAPE lecture 2004’, *Acta Psychiatrica Scandinavica*, 111(1), pp. 4–11. doi: 10.1111/j.1600-0447.2004.00467.x.

Mckinney, B. C. *et al.* (2019) ‘Density of small dendritic spines and microtubule-associated- protein-2 immunoreactivity in the primary auditory cortex of subjects with schizophrenia’, *Neuropsychopharmacology*. Springer US, (February). doi: 10.1038/s41386-019-0350-7.

Mctague, A. *et al.* (2018) ‘Clinical and molecular characterization of KCNT1-related severe early-onset epilepsy’, *Neurology*, 90(1), pp. 55–66. doi: 10.1212/WNL.0000000000004762.

Mcvicker, G. *et al.* (2013) ‘Identification of Genetic Variants That Accect Histone Modifications in Human Cells’, *Science*, 342(November), pp. 747–749.

Meehan, R. R. *et al.* (2018) ‘DNA methylation as a genomic marker of exposure to chemical and environmental agents’, *Current Opinion in Chemical Biology*. The Authors, 45, pp. 48–56. doi: 10.1016/j.cbpa.2018.02.006.

Mei, L. and Xiong, W. C. (2008) ‘Neuregulin 1 in neural development, synaptic plasticity and schizophrenia’, *Nature Reviews Neuroscience*, 9(6), pp. 437–452. doi: 10.1038/nrn2392.

Meltzer, H. Y. and Stahl, S. M. (1976) ‘The Dopamine Hypothesis of Schizophrenia:

a review', *Schizophrenia Bulletin*, 2(1), pp. 19–76. doi: 10.1111/j.1744-6163.1990.tb00312.x.

Members of the Complex Trait Consortium (2003) 'The nature and identification of quantitative trait loci: a community's view', *Nature Reviews Genetics*, 4(11), pp. 911–916. doi: 10.1007/s11103-011-9767-z.Plastid.

Meulenbellt, I. *et al.* (1995) 'High-Yield Noninvasive Human Genomic DNA Isolation Method for Genetic Studies in Geographically Dispersed Families and Populations', *American Journal of Human Genetics*, 57, pp. 1252–1254.

Meyer, M. *et al.* (2010) 'Gene targeting by homologous recombination in mouse zygotes mediated by zinc-finger nucleases', *Proceedings of the National Academy of Sciences of the United States of America*, 107(34), pp. 15022–15026. doi: 10.1073/pnas.1009424107.

Mi, H. *et al.* (2017) 'PANTHER version 11: expanded annotation data from Gene Ontology and Reactome pathways, and data analysis tool enhancements', *Nucleic Acids Research*, 45(November 2016), pp. 183–189. doi: 10.1093/nar/gkw1138.

Miguez, M. J. *et al.* (2019) 'Marijuana use among adolescents is associated with deleterious alterations in mature BDNF', *AIMS Public Health*, 6(1), pp. 4–14. doi: 10.3934/publichealth.2019.1.4.

Mill, J. *et al.* (2008) 'Epigenomic Profiling Reveals DNA-Methylation Changes Associated with Major Psychosis', *American Journal of Human Genetics*, 82(3), pp. 696–711. doi: 10.1016/j.ajhg.2008.01.008.

Millar, J. K. *et al.* (2000) 'Disruption of two novel genes by a translocation co-segregating with schizophrenia', *Human Molecular Genetics*, 9(9), pp. 1415–1423. doi: 10.1093/hmg/9.9.1415.

Miller, J. C. *et al.* (2011) 'A TALE nuclease architecture for efficient genome editing', *Nature Biotechnology*, 29(2), pp. 143–150. doi: 10.1038/nbt.1755.

Mitsopoulos, P. *et al.* (2015) 'Stomatin-Like Protein 2 Is Required for In Vivo Mitochondrial Respiratory Chain Supercomplex Formation and Optimal Cell Function', *Molecular and Cellular Biology*, 35(10), pp. 1838–1847. doi: 10.1128/mcb.00047-15.

Moehle, E. A. *et al.* (2007) 'Targeted gene addition into a specified location in the human genome using designed zinc finger nucleases', *Proceedings of the National Academy of Sciences of the United States of America*, 104(9), pp. 3055–3060. doi: 10.1073/pnas.0611478104.

Montgomery, J. *et al.* (2007) 'Simultaneous mutation scanning and genotyping by high-resolution DNA melting analysis', *Nature Protocols*, 2(1), pp. 59–66. doi: 10.1038/nprot.2007.10.

Morgan, C. *et al.* (2007) 'Parental separation, loss and psychosis in different ethnic groups: A case-control study', *Psychological Medicine*, 37(4), pp. 495–503. doi: 10.1017/S0033291706009330.

Morgan, C. and Fisher, H. (2007) 'Environment and schizophrenia: Environmental factors in schizophrenia: Childhood trauma - A critical review', *Schizophrenia Bulletin*, 33(1), pp. 3–10. doi: 10.1093/schbul/sbl053.

Morison, I. M., Ramsay, J. P. and Spencer, H. G. (2005) 'A census of mammalian imprinting', *Trends in Genetics*, 21(8), pp. 457–465. doi: 10.1016/j.tig.2005.06.008.

Morris, J. A. (2009) 'Zebrafish: A model system to examine the neurodevelopmental basis of schizophrenia', *Progress in Brain Research*, 179(C), pp. 97–106. doi: 10.1016/S0079-6123(09)17911-6.

Mortazavi, A. *et al.* (2008) 'Mapping and quantifying mammalian transcriptomes by RNA-Seq', *Nature Methods*, 5(7), pp. 1–8. doi: 10.1038/NMETH.1226.

Mothersill, O., Knee-Zaska, C. and Donohoe, G. (2016) 'Emotion and Theory of Mind in Schizophrenia—Investigating the Role of the Cerebellum', *Cerebellum*, 15(3), pp. 357–368. doi: 10.1007/s12311-015-0696-2.

Mukai, J. *et al.* (2008) 'Palmitoylation-dependent neurodevelopmental deficits in a mouse model of 22q11 microdeletion', *Nature Neuroscience*, 11(11), pp. 1302–1310. doi: 10.1038/nn.2204.

Murphy, S. K. *et al.* (2018) 'Cannabinoid exposure and altered DNA methylation in rat and human sperm', *Epigenetics*. Taylor & Francis, 13(12), pp. 1208–1221. doi: 10.1080/15592294.2018.1554521.

Murphy, T. M. *et al.* (2017) 'Methylomic profiling of cortex samples from completed suicide cases implicates a role for PSORS1C3 in major depression and suicide', *Nature Publishing Group*. Nature Publishing Group, 7(October 2016). doi: 10.1038/tp.2016.249.

Murray, R. M. *et al.* (2007) 'Cannabis, the mind and society: The hash realities', *Nature Reviews Neuroscience*, 8(11), pp. 885–895. doi: 10.1038/nrn2253.

Nakamura, K. *et al.* (2014) 'PIGO mutations in intractable epilepsy and severe developmental delay with mild elevation of alkaline phosphatase levels', *Epilepsia*, 55(2), pp. 13–17. doi: 10.1111/epi.12508.

Nakamura, M. *et al.* (2006) 'Cell-surface-localized ATP detection with immobilized firefly luciferase', *Analytical Biochemistry*, 352, pp. 61–67. doi: 10.1016/j.ab.2006.02.019.

Naldini, L. *et al.* (1996) 'In Vivo Gene Delivery and Stable Transduction of Nondividing Cells by a Lentiviral Vector', *Science*, 272(April), pp. 263–268.

Nestler, E. J. and Hyman, S. E. (2010) 'Animal models of neuropsychiatric disorders', *Nature Neuroscience*. Nature Publishing Group, 13(10), pp. 1161–1169. doi: 10.1038/nn.2647.

Nicholas, C. R. *et al.* (2013) 'Functional maturation of hPSC-derived forebrain interneurons requires an extended timeline and mimics human neural development', *Cell Stem Cell*. Elsevier Inc., 12(5), pp. 573–586. doi: 10.1016/j.stem.2013.04.005.

Nun, V. *et al.* (2019) 'Flow Cytometry Has a Significant Impact on the Cellular Metabolome', *Journal of Proteome Research*, 18(1), pp. 169–181. doi: 10.1021/acs.jproteome.8b00472.

O'Donovan, M. C. *et al.* (2008) 'Identification of loci associated with schizophrenia by genome-wide association and follow-up', *Nature Genetics*, 40(9), pp. 1053–1055. doi: 10.1038/ng.201.

Orr, C. *et al.* (2019) 'Grey Matter Volume Differences Associated with Extremely Low Levels of Cannabis Use in Adolescence', *The Journal of Neuroscience*, 10(10), pp. 1817–1827. doi: 10.1523/JNEUROSCI.3375-17.2018.

Oshlack, A. and Wakefield, M. J. (2009) 'Transcript length bias in RNA-seq data confounds systems biology', *Biology Direct*, 4(14), pp. 1–10. doi: 10.1186/1745-6150-4-14.

Owen, M. J. *et al.* (2011) 'Neurodevelopmental hypothesis of schizophrenia', *British Journal of Psychiatry*, 198(3), pp. 173–175. doi: 10.1192/bjp.bp.110.084384.

Ozsolak, F. and Milos, P. M. (2011) 'RNA sequencing: Advances, challenges and opportunities', *Nature Reviews Genetics*. Nature Publishing Group, 12(2), pp. 87–98. doi: 10.1038/nrg2934.

Pagnamenta, A. T. *et al.* (2017) 'Analysis of exome data for 4293 trios suggests GPI-anchor biogenesis defects are a rare cause of developmental disorders', *European Journal of Human Genetics*. Nature Publishing Group, 25(6), pp. 669–679. doi: 10.1038/ejhg.2017.32.

Panlilio, L. V. and Justinova, Z. (2018) 'Preclinical Studies of Cannabinoid Reward, Treatments for Cannabis Use Disorder, and Addiction-Related Effects of Cannabinoid Exposure', *Neuropsychopharmacology*. Nature Publishing Group, 43(1), pp. 116–141. doi: 10.1038/npp.2017.193.

Papatheodorou, I. *et al.* (2018) 'Expression Atlas: gene and protein expression across multiple studies and organisms', *Nucleic Acids Research*, 46(November 2017), pp. 246–251. doi: 10.1093/nar/gkx1158.

Pardiñas, A. F. *et al.* (2018) 'Common schizophrenia alleles are enriched in mutation-intolerant genes and in regions under strong background selection', *Nature Genetics*, 50(March), pp. 381–389. doi: 10.1038/s41588-018-0059-2.

Paşca, A. M. *et al.* (2015) 'Functional cortical neurones and astrocytes from human pluripotent stem cells in 3D culture', *Nature Methods*, 12(7), pp. 671–678. doi: 10.1038/nmeth.3415.Functional.

Pasman, J. A. *et al.* (2018) 'GWAS of lifetime cannabis use reveals new risk loci, genetic overlap with psychiatric traits, and a causal influence of schizophrenia', *Nature Neuroscience*.

Pear, W. S. *et al.* (1993) 'Production of high-titer helper-free retroviruses by transient transfection', *Proceedings of the National Academy of Sciences*, 90(September), pp. 8392–8396.

Pedersen, B. S. *et al.* (2012) 'Comb-p : software for combining , analyzing , grouping and correcting spatially correlated P -values', *Bioinformatics*, 28(22), pp. 2986–2988. doi: 10.1093/bioinformatics/bts545.

Peters, T. J. *et al.* (2015) 'De novo identification of differentially methylated regions in the human genome', *Epigenetics & Chromatin*, 8(6), pp. 1–16.

Pfaffeneder, T. *et al.* (2011) 'The discovery of 5-formylcytosine in embryonic stem cell DNA', *Angewandte Chemie - International Edition*, 50(31), pp. 7008–7012. doi: 10.1002/anie.201103899.

Phipson, B. and Maksimovic, J. (2016) 'missMethyl: an R package for analyzing data from Illumina's HumanMethylation4560 platform', *Bioinformatics*, 32(2), pp. 286–288.

Picot, J. *et al.* (2012) 'Flow cytometry: retrospective, fundamentals and recent instrumentation', *Cytotechnology*, 64(2), pp. 109–130. doi: 10.1007/s10616-011-9415-0.

Pidsley, R. *et al.* (2013) 'A data-driven approach to preprocessing Illumina 450K methylation array data', *BMC genomics*, 14(293), pp. 1–16.

Pidsley, R. *et al.* (2014) 'Methylomic profiling of human brain tissue supports a neurodevelopmental origin for schizophrenia', *Genome Biology*, 15(10), pp. 1–10. doi: 10.1186/s13059-014-0483-2.

Pidsley, R. *et al.* (2016) 'Critical evaluation of the Illumina MethylationEPIC BeadChip microarray for whole-genome DNA methylation profiling', *Genome Biology*. Genome Biology, 17(208), pp. 1–17. doi: 10.1186/s13059-016-1066-1.

Pidsley, R. and Mill, J. (2011) 'Epigenetic studies of psychosis: Current findings, methodological approaches, and implications for postmortem research', *Biological Psychiatry*. Elsevier Inc., 69(2), pp. 146–156. doi: 10.1016/j.biopsych.2010.03.029.

Popova, D., Karlsson, J. and Jacobsson, S. O. P. (2017) 'Comparison of neurons derived from mouse P19, rat PC12 and human SH-SY5Y cells in the assessment of chemical- and toxin-induced neurotoxicity', *BMC Pharmacology and Toxicology*. BMC Pharmacology and Toxicology, 18(1), pp. 1–11. doi: 10.1186/s40360-017-0151-8.

Popp, M. W. and Maquat, L. E. (2016) 'Leveraging Rules of Nonsense-Mediated mRNA Decay for Genome Engineering and Personalized Medicine', *Cell*, 165(6), pp. 1319–1322. doi: 10.1016/j.cell.2016.05.053.Leveraging.

Porteus, M. H. and Baltimore, D. (2003) 'Chimeric nucleases stimulate gene targeting in human cells', *Science*, 300(5620), p. 763. doi: 10.1126/science.1078395.

Price, E. M. *et al.* (2013) 'Additional annotation enhances potential for biologically-relevant analysis of the Illumina Infinium HumanMethylation450 BeadChip array', *Epigenetics & Chromatin*, 6(4), pp. 1–15. doi: 10.1186/1756-8935-6-4.

Qi, L. S. *et al.* (2013) 'Repurposing CRISPR as an RNA-Guided Platform for Sequence-Specific Control of Gene Expression', *Cell*. Elsevier, 152(5), pp. 1173–1183. doi: 10.1016/j.cell.2013.02.022.

Radhakrishnan, R., Wilkinson, S. T. and D'Souza, D. C. (2014) 'Gone to pot-a review of the association between cannabis and psychosis', *Frontiers in Psychiatry*, 5(MAY), pp. 1–24. doi: 10.3389/fpsyt.2014.00054.

Rakyan, V. K. *et al.* (2008) 'An integrated resource for genome-wide identification and analysis of human tissue-specific differentially methylated regions (tDMRs)', *Genome Research*, 18(9), pp. 1518–1529. doi: 10.1101/gr.077479.108.

Rakyan, V. K. *et al.* (2011) 'Epigenome-wide association studies for common human diseases', *Nature Reviews Genetics*. Nature Publishing Group, 12(8), pp. 529–541. doi: 10.1038/nrg3000.

Ran, F. A. *et al.* (2013) 'Genome engineering using the CRISPR-Cas9 system',

Nature protocols, 8(11), pp. 2281–2308. doi: 10.1038/nprot.2013.143.Genome.

Rath, D. *et al.* (2015) 'The CRISPR-Cas immune system: Biology, mechanisms and applications', *Biochimie*. Elsevier B.V, 117, pp. 119–128. doi: 10.1016/j.biochi.2015.03.025.

Rees, E. *et al.* (2012) 'De novo mutation in schizophrenia', *Schizophrenia Bulletin*, 38(3), pp. 377–381. doi: 10.1093/schbul/sbs047.

Rees, E. *et al.* (2014) 'Evidence that duplications of 22q11.2 protect against schizophrenia', *Molecular Psychiatry*, 19(1), pp. 37–40. doi: 10.1038/mp.2013.156.

Rees, H. A. and Liu, D. R. (2018) 'Base editing: precision chemistry on the genome and transcriptome of living cells', *Nature Reviews Genetics*. Springer US, 19(12), pp. 770–788. doi: 10.1038/s41576-018-0059-1.

Ren, J. *et al.* (2016) 'Interaction between DISC1 and CHL1 in regulation of neurite outgrowth', *Brain Research*. Elsevier, 1648, pp. 290–297. doi: 10.1016/j.brainres.2016.06.033.

Revez, J. A. *et al.* (2016) 'Identification of STOML2 as a putative novel asthma risk gene associated with IL6R', *Allergy*, 71(7), pp. 1020–1030. doi: 10.1016/j.physbeh.2017.03.040.

Reynolds, O. (1883) 'An Experimental Investigation of the Circumstances which determine whether the Motion of Water shall be Direct or Sinuous, and of the Law of Resistance in Parallel Channels.', *Philosophical Transactions of the Royal Society of London*, 174, pp. 935–982.

Rinn, J. L. *et al.* (2007) 'Functional Demarcation of Active and Silent Chromatin Domains in Human HOX Loci by Noncoding RNAs', *Cell*, 129, pp. 1311–1323. doi: 10.1016/j.cell.2007.05.022.

Rio, D. C., Clark, S. G. and Tjian, R. (1985) 'A Mammalian Host-Vector System That Regulates Expression and Amplification of Transfected Genes by Temperature Induction', *Science*, 227(4682), pp. 23–28.

Ripke, S. *et al.* (2013) 'Genome-wide Association Analysis Identifies 14 New Risk Loci for Schizophrenia', *Nature Genetics*, 45(10), pp. 1150–1159. doi: 10.1038/ng.2742.Genome-wide.

Roberts, R. J., Murray, K. and Roberts, R. J. (1976) 'CRC Critical Reviews in Biochemistry Restriction Endonuclease', *CRC Critical Reviews in Biochemistry*, 4(2), pp. 123–164. doi: 10.3109/10409237609105456.

Robinson, J. T. *et al.* (2011) 'Integrative Genomics Viewer', *Nature Biotechnology*, 29(1), pp. 24–26. doi: 10.1038/nbt.1754.Integrative.

Roda, A. *et al.* (2009) 'Bioluminescence in analytical chemistry and in vivo imaging', *Trends in Analytical Chemistry*. Elsevier Ltd, 28(3), pp. 307–322. doi: 10.1016/j.trac.2008.11.015.

Rotter, A. *et al.* (2012) 'Orexin a expression and promoter methylation in patients with cannabis dependence in comparison to nicotine-dependent cigarette smokers and nonsmokers', *Neuropsychobiology*, 66(2), pp. 126–133. doi: 10.1159/000339457.

Rouhani, F. *et al.* (2014) 'Genetic Background Drives Transcriptional Variation in Human Induced Pluripotent Stem Cells', *PLoS Genetics*, 10(6). doi: 10.1371/journal.pgen.1004432.

Ruan, K. *et al.* (2015) 'Histone H4 acetylation required for chromatin decompaction during DNA replication', *Scientific Reports*. Nature Publishing Group, 5, pp. 1–10. doi: 10.1038/srep12720.

Saavedra, D. *et al.* (2014) 'ITH12410/SC058: A New Neuroprotective Compound with Potential in the Treatment of Alzheimer's Disease', *ACS Chemical Neuroscience*, 5(9), pp. 770–775. doi: 10.1021/cn500131t.

Sadakierska-Chudy, A. and Filip, M. (2015) 'A Comprehensive View of the Epigenetic Landscape. Part II: Histone Post-translational Modification, Nucleosome Level, and Chromatin Regulation by ncRNAs', *Neurotoxicity Research*, 27(2), pp. 172–197. doi: 10.1007/s12640-014-9508-6.

Sakuma, T. *et al.* (2014) 'Multiplex genome engineering in human cells using all-in-one CRISPR/Cas9 vector system', *Scientific Reports*, 4, pp. 4–9. doi: 10.1038/srep05400.

Salazar, M. *et al.* (2009) 'Cannabinoid action induces autophagy-mediated cell death through stimulation of ER stress in human glioma cells', *Journal of Clinical Investigation*, 119(5), pp. 1359–1372. doi: 10.1172/JCI37948.

Salzman, G. C. *et al.* (1975) 'A Flow-System Multiangle Light-Scattering Instrument for Cell Characterization', *Clinical Chemistry*, 21(9), pp. 1297–1304.

Sams-Dodd, F. (1998) 'A test of the predictive validity of animal models of schizophrenia based on phencyclidine and d-amphetamine', *Neuropsychopharmacology*, 18(4), pp. 293–304. doi: 10.1016/S0893-133X(97)00161-9.

Sander, J. D. *et al.* (2011) 'Targeted gene disruption in somatic zebrafish cells using engineered TALENs', *Nature Biotechnology*. Nature Publishing Group, 29(8), pp. 697–698. doi: 10.1038/nbt.1934.

Sanger, F. and Nicklen, S. (1977) 'DNA sequencing with chain-terminating inhibitors', *Proceedings of the National Academy of Sciences of the United States of America*, 74(12), pp. 5463–5467.

Schaub, M. A. *et al.* (2012) 'Linking disease associations with regulatory information in the human genome', *Genome Research*, 22, pp. 1748–1759. doi: 10.1101/gr.136127.111.

Schizophrenia Working Group of the Psychiatric Genomics Consortium (2014) 'Biological Insights From 108 Schizophrenia-Associated Genetic Loci', *Nature*, 511(7510), pp. 421–427. doi: 10.1038/nature13595.Biological.

Schmidt, R., Strähle, U. and Scholpp, S. (2013) 'Neurogenesis in zebrafish - from embryo to adult', *Neural Development*, 8(1), pp. 1–13. doi: 10.1186/1749-8104-8-3.

Schroeder, A. *et al.* (2006) 'The RIN: an RNA integrity number for assigning integrity values to RNA measurements', *BMC Molecular Biology*, 14, pp. 1–14. doi: 10.1186/1471-2199-7-3.

Schrott, R. *et al.* (2019) 'Cannabis use is associated with potentially heritable widespread changes in autism candidate gene *DLGAP2* DNA methylation in sperm', *Epigenetics*. Taylor & Francis, 0(0), pp. 1–13. doi: 10.1080/15592294.2019.1656158.

Schwenk, J. M., Asplund, A. and Edqvist, P. D. (2011) 'The Human Protein Atlas as a proteomic resource for biomarker discovery', *Journal of Internal Medicine*, 270, pp. 428–446. doi: 10.1111/j.1365-2796.2011.02427.x.

Scott, M. S. and Ono, M. (2011) 'From snoRNA to miRNA: Dual function regulatory non-coding RNAs', *Biochimie*. Elsevier Masson SAS, 93(11), pp. 1987–1992. doi: 10.1016/j.biochi.2011.05.026.

Segal-Gavish, H. *et al.* (2017) 'BDNF overexpression prevents cognitive deficit elicited by adolescent cannabis exposure and host susceptibility interaction', *Human Molecular Genetics*, 26(13), pp. 2462–2471. doi: 10.1093/hmg/ddx139.

Sekar, A. *et al.* (2016) 'Schizophrenia risk from complex variation of complement component 4', *Nature*. Nature Publishing Group, 530(7589), pp. 177–183. doi: 10.1038/nature16549.

Shah, B. S., Shimell, J. J. and Bamji, S. X. (2019) 'Regulation of dendrite morphology and excitatory synapse formation by zDHHC15', *Journal of Cell Science*, 132. doi: 10.1242/jcs.230052.

Shah, S. A. *et al.* (2013) 'Protospacer recognition motifs', *RNA Biology*, 10(5), pp. 891–899.

Sharma, S., Kelly, T. K. and Jones, P. A. (2010) 'Epigenetics in cancer', *Carcinogenesis*, 31(1), pp. 27–36. doi: 10.1093/carcin/bgp220.

Sharpley, M. *et al.* (2001) 'Understanding the excess of psychosis among the African-Caribbean population in England', *British Journal of Psychiatry*, 178(SUPPL. 40).

Shelton, M. A. *et al.* (2016) 'Loss of Microtuble-Associated Protein 2 Immunoreactivity Linked to Dendritic Spine Loss in Schizophrenia', *Biological Psychiatry*, 78(6), pp. 374–385. doi: 10.1016/j.biopsych.2014.12.029.Loss.

Shenton, M. E. *et al.* (2001) 'A review of MRI findings in schizophrenia', *Schizophrenia Research*, 49(1–2), pp. 1–52. Available at: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2812015/.

Shi, D. Q. *et al.* (2017) 'New insights into 5hmC DNA modification: Generation, distribution and function', *Frontiers in Genetics*, 8(JUL), pp. 1–11. doi: 10.3389/fgene.2017.00100.

Shipley, M. M., Mangold, C. A. and Szpara, M. L. (2017) 'Differentiation of the SH-SY5Y Human Neuroblastoma Cell Line', *Journal of visualized experiments : JoVE*, (108), pp. 3047–3054. doi: 10.1016/j.bbamem.2015.02.010.Cationic.

Shorter, K. R. and Miller, B. H. (2015) 'Epigenetic Mechanisms in Schizophrenia', *Progress in Biophysics & Molecular Biology*, 118, pp. 1–7. doi: 10.1038/nrn1604.

Siddiqui, I. J., Pervaiz, N. and Abbasi, A. A. (2016) 'The Parkinson Disease gene SNCA: Evolutionary and structural insights with pathological implication', *Scientific Reports*. Nature Publishing Group, 6(24475), pp. 1–11. doi: 10.1038/srep24475.

Sido, J. M. *et al.* (2015) ' Δ 9 -Tetrahydrocannabinol-mediated epigenetic modifications elicit myeloid-derived suppressor cell activation via STAT3/S100A8 ', *Journal of Leukocyte Biology*, 97(4), pp. 677–688. doi: 10.1189/jlb.1a1014-479r.

Siggens, L. and Ekwall, K. (2014) 'Epigenetics , chromatin and genome organization : recent advances from the ENCODE project', *Journal of Internal Medicine*, 276(3), pp. 201–214. doi: 10.1111/joim.12231.

Simon, L. *et al.* (2016) 'Δ9-Tetrahydrocannabinol (Δ9-THC) Promotes Neuroimmune-Modulatory MicroRNA Profile in Striatum of Simian Immunodeficiency Virus (SIV)-Infected Macaques', *Journal of Neuroimmune Pharmacology*, 11(1), pp. 192–213. doi: 10.1007/s11481-015-9645-6.

Slatkin, M. (2008) 'Linkage disequilibrium - Understanding the evolutionary past and mapping the medical future', *Nature Reviews Genetics*, 9(6), pp. 477–485. doi: 10.1038/nrg2361.

Sloan, S. A. *et al.* (2017) 'Human Astrocyte Maturation Captured in 3D Cerebral Cortical Spheroids Derived from Pluripotent Stem Cells', *Neuron*, 95(4), pp. 779-790.e6. doi: 10.1016/j.neuron.2017.07.035.

Sloan, S. A. *et al.* (2018) 'Generation and assembly of human brain region–specific three-dimensional cultures', *Nature Protocols*, 13(September), pp. 2062–2085. doi: 10.1038/s41596-018-0032-7.

Smale, S. T. and Kadonaga, J. T. (2003) 'The RNA Polymerase II Core Promoter', *Annual Review of Biochemistry*, 72, pp. 449–479. doi: 10.1146/annurev.biochem.72.121801.161520.

Smith, A. R. *et al.* (2019) 'Parallel profiling of DNA methylation and hydroxymethylation highlights neuropathology-associated epigenetic variation in Alzheimer's disease', *Clinical Epigenetics*. Clinical Epigenetics, 11(1), pp. 1–13. doi: 10.1186/s13148-019-0636-y.

Spencer, V. A. and Davie, J. R. (1999) 'Role of covalent modifications of histones in regulating gene expression', *Gene*, 240(1), pp. 1–12. doi: 10.1016/S0378-1119(99)00405-9.

Spengler, B. A., Biedler, J. L. and Ross, R. A. (2002) 'A corrected karyotype for the SH-SY5Y human neuroblastoma cell line', *Cancer Genetics and Cytogenetics*, 138, pp. 177–178.

Spiers, H. *et al.* (2015) 'Methylomic trajectories across human fetal brain development', *Genome Research*, 25(3), pp. 338–352. doi: 10.1101/gr.180273.114.

Spiers, H. *et al.* (2017) '5-Hydroxymethylcytosine Is Highly Dynamic Across Human Fetal Brain Development', *BMC Genomics*. BMC Genomics, 18(1), pp. 1–14. doi: 10.1186/s12864-017-4091-x.

Srikanth, P. *et al.* (2015) 'Genomic DISC1 Disruption in hiPSCs Alters Wnt Signaling and Neural Cell Fate', *Cell Reports*. The Authors, 12(9), pp. 1414–1429. doi: 10.1016/j.celrep.2015.07.061.

Srikanth, P. *et al.* (2018) 'Convergence of independent DISC1 mutations on impaired neurite growth via decreased UNC5D expression', *Translational Psychiatry*. Springer US, 8(245), pp. 1–16. doi: 10.1038/s41398-018-0281-9.

Stables, J. *et al.* (2009) 'Development of a Dual Glow-Signal Firefly and Renilla Luciferase Assay Reagent for the Analysis of G-Protein Coupled Receptor Signalling', *Journal of Receptors and Signal Transduction*, 9893. doi: 10.3109/10799899909036660.

Stanslowsky, N. *et al.* (2017) 'Functional effects of cannabinoids during dopaminergic specification of human neural precursors derived from induced pluripotent stem cells', *Addiction Biology*, 22(5), pp. 1329–1342. doi: 10.1111/adb.12394.

Stark, Z. *et al.* (2010) 'De novo 325 kb microdeletion in chromosome band 10q25.3 including ATRNL1 in a boy with cognitive impairment, autism and dysmorphic features', *European Journal of Medical Genetics*. Elsevier Masson SAS, 53(5), pp. 337–339. doi: 10.1016/j.ejmg.2010.07.009.

Stefansson, H. *et al.* (2002) 'Neuregulin 1 and Susceptibility to Schizophrenia', *American Journal of Human Genetics*, 71(4), pp. 877–892. doi: 10.1086/342734.

Stein, J. L. *et al.* (2014) 'A quantitative framework to evaluate modeling of cortical development by neural stem cells', *Neuron*, 83(1), pp. 69–86. doi: 10.1016/j.neuron.2014.05.035.

Stepanenko, A. A. and Dmitrenko, V. V (2015) 'HEK293 in cell biology and cancer research: phenotype, karyotype, tumorigenicity, and stress-induced genome-phenotype evolution', *Gene*. Elsevier B.V., 569(2), pp. 182–190. doi: 10.1016/j.gene.2015.05.065.

Stiles, J. and Jernigan, T. L. (2010) 'The basics of brain development', *Neuropsychology Review*, 20(4), pp. 327–348. doi: 10.1007/s11065-010-9148-4.

Strous, R. D. and Shoenfeld, Y. (2006) 'Schizophrenia, autoimmunity and immune system dysregulation: A comprehensive model updated and revisited', *Journal of Autoimmunity*, 27(2), pp. 71–80. doi: 10.1016/j.jaut.2006.07.006.

Suderman, M., French, R. and Arathimos, R. (2018) 'dmrff: identifying differentially methylated regions efficiently with power and control', *BioRxiv*, pp. 1–26.

Sugano, S. S. *et al.* (2018) 'Efficient CRISPR / Cas9-based genome editing and its application to conditional genetic analysis in Marchantia polymorpha', *PLoS ONE*, 13(10), pp. 1–22.

Suliman, N. A. *et al.* (2018) 'Delta-9-Tetrahydrocannabinol (∆9-THC) Induce Neurogenesis and Improve Cognitive Performances of Male Sprague Dawley Rats', *Neurotoxicity Research*. Neurotoxicity Research, 33(2), pp. 402–411. doi: 10.1007/s12640-017-9806-x.

Sullivan, P. F. (2010) 'The Psychiatric GWAS Consortium: Big science comes to psychiatry', *Neuron*. Elsevier Inc., 68(2), pp. 182–186. doi: 10.1016/j.neuron.2010.10.003.

Sullivan, P. F., Kendler, K. S. and Neale, M. C. (2003) 'Schizophrenia as a Complex Trait Evidence From a Meta-analysis of Twin Studies', *Archives of General Psychiatry*, 60(12), p. 1187. doi: 10.1001/archpsyc.60.12.1187.

Suzuki, M. M. and Bird, A. (2008) 'DNA methylation landscapes: provocative insights from epigenomics', *Nature Reviews Genetics*, 9, pp. 465–476. doi: 10.1038/nrg2341.

Symington, L. S. and Gautier, J. (2011) 'Double-Strand Break End Resection and Repair Pathway Choice', *Annual Review of Genetics*, 45(1), pp. 247–271. doi: 10.1146/annurev-genet-110410-132435.

Szutorisz, H. and Hurd, Y. L. (2016) 'Epigenetic Effects of Cannabis Exposure', *Biological Psychiatry*, 79(7), pp. 586–594. doi: 10.1016/j.biopsych.2015.09.014.Epigenetic.

Szutorisz, H. and Hurd, Y. L. (2018) 'High times for cannabis: Epigenetic imprint and its legacy on brain and behavior', *Neuroscience and Biobehavioral Reviews*. Elsevier, 85, pp. 93–101. doi: 10.1016/j.neubiorev.2017.05.011.

Tadi, V. *et al.* (2016) 'Repurposing the CRISPR-Cas9 system for targeted DNA methylation', pp. 1–14. doi: 10.1093/nar/gkw159.

Taghbalout, A. *et al.* (2019) 'Enhanced CRISPR-based DNA demethylation by Casilio-ME-mediated RNA coupling of methylcytosine oxidation and DNA repair pathways', *Nature Communications*. Springer US, (2019), pp. 1–12. doi: 10.1038/s41467-019-12339-7.

Tahiliani, M. *et al.* (2009) 'Conversion of 5-methylcytosine to 5-hydroxymethylcytosine in mammalian DNA by MLL partner TET1', *Science*, 324(5929), pp. 930–935. doi: 10.1126/science.1170116.

Takahashi, K. *et al.* (2007) 'Induction of Pluripotent Stem Cells from Adult Human Fibroblasts by Defined Factors', *Cell*, 131(5), pp. 861–872. doi: 10.1016/j.cell.2007.11.019.

Tamminga, C. A. and Buchsbaum, M. S. (2004) 'Frontal Cortex Function', *American Journal of Psychiatry*, 16112(December), p. 75390. Available at: http://ajp.psychiatryonline.org.

Tandon, R. *et al.* (2013) 'Definition and description of schizophrenia in the DSM-5', *Schizophrenia Research*, 150(1), pp. 3–10. doi: 10.1016/j.schres.2013.05.028.

Tarjan, D. R., Flavahan, W. A. and Bernstein, B. E. (2019) 'Epigenome editing strategies for the functional annotation of CTCF insulators.', *Nature communications*. Springer US, 10(1), p. 4258. doi: 10.1038/s41467-019-12166-w.

Taylor, P. *et al.* (2014) 'Functional DNA methylation in a transcript specific 3′UTR region of TrkB associates with suicide', *Epigenetics*, 9(8), pp. 37–41. doi: 10.4161/epi.29068.

Teppola, H. *et al.* (2016) 'Morphological Differentiation Towards Neuronal Phenotype of SH-SY5Y Neuroblastoma Cells by Estradiol, Retinoic Acid and Cholesterol', *Neurochemical Research*. Springer US, 41(4), pp. 731–747. doi: 10.1007/s11064-015-1743-6.

Thakore, P. I. *et al.* (2015) 'Highly specific epigenome editing by CRISPR-Cas9 repressors for silencing of distal regulatory elements', *Nature Methods*, 12(12), pp. 1143–1149. doi: 10.1038/nmeth.3630.

The Autism Spectrum Disorders Working Group of The Psychiatric Genomics Consortium (2017) 'Meta-analysis of GWAS of over 16,000 individuals with autism spectrum disorder highlights a novel locus at 10q24.32 and a significant overlap with schizophrenia', *Molecular Autism*, 8(21), pp. 1–17. doi: 10.1186/s13229-017-0137-9.

The ENCODE Project Consortium (2007) ‘Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project’, *Nature*, 447(7146), pp. 799–816. doi: 10.1038/nature05874.

The Genotype Tissue Expression Consortium (2019a) ‘The GTEx Consortium atlas of genetic regulatory effects across human tissues’, *BioRxiv*. doi: 10.1101/787903.

The Genotype Tissue Expression Consortium (2019b) ‘The GTEx Consortium atlas of genetic regulatory effects across human tissues’, *BioRxiv*.

The Schizophrenia Psychiatric Genome-Wide Association Study Consortium (2012) ‘Genome-wide association study identifies five new schizophrenia loci’, *Nature Genetics*, 43(10), pp. 969–976. doi: 10.1038/ng.940.Genome-wide.

Thomas, D. J. *et al.* (2007) ‘Arsenic (+3 Oxidation State) Methyltransferase and the Methylation of Arsenicals’, *Experimental Biology and Medicine (Maywood)*, 232(1), pp. 3–13.

Thomas, M. P. and Potter, B. V. L. (2014) ‘The enzymes of human diphosphoinositol polyphosphate metabolism’, *FEBS Journal*, 281(May), pp. 14–33. doi: 10.1111/febs.12575.

Thomas, P. and Smart, T. G. (2005) ‘HEK293 cell line: A vehicle for the expression of recombinant proteins’, *Journal of Pharmacological and Toxicological Methods*, 51, pp. 187–200. doi: 10.1016/j.vascn.2004.08.014.

Thorvaldsdóttir, H., Robinson, J. T. and Mesirov, J. P. (2013) ‘Integrative Genomics Viewer (IGV): High-performance genomics data visualization and exploration’, *Briefings in Bioinformatics*, 14(2), pp. 178–192. doi: 10.1093/bib/bbs017.

Tiethof, A. K., Richardson, J. R. and Hart, R. P. (2018) ‘Knockdown of butyrylcholinesterase but not inhibition by chlorpyrifos alters early differentiation mechanisms in human neural stem cells’, *bioRxiv*.

Tobi, E. W. *et al.* (2014) ‘DNA methylation signatures link prenatal famine exposure to growth and metabolism’, *Nature Communications*, 5(5592). doi: 10.1038/ncomms6592.

Torrey, E. F. *et al.* (1997) ‘Seasonality of births in schizophrenia and bipolar disorder: A review of the literature’, *Schizophrenia Research*, 28(1), pp. 1–38. doi: 10.1016/S0920-9964(97)00092-3.

Torrey, E. F. *et al.* (2009) ‘Paternal age as a risk factor for schizophrenia: How important is it?’, *Schizophrenia Research*. Elsevier B.V., 114(1–3), pp. 1–5. doi: 10.1016/j.schres.2009.06.017.

Trost, S. *et al.* (2016) ‘Investigating the impact of a genome-wide supported bipolar risk variant of MAD1L1 on the human reward system’, *Neuropsychopharmacology*, 41(11), pp. 2679–2687. doi: 10.1038/npp.2016.70.

Tsai, S. Q. *et al.* (2015) ‘GUIDE-seq enables genome-wide profiling of off-target cleavage by CRISPR-Cas nucleases’, *Nature Biotechnology*, 33(2), pp. 187–198. doi: 10.1038/nbt.3117.

Tunbridge, E. M., Harrison, P. J. and Weinberger, D. R. (2006) ‘Catechol-o-Methyltransferase, Cognition, and Psychosis: Val158Met and Beyond’, *Biological*

Psychiatry, 60(2), pp. 141–151. doi: 10.1016/j.biopsych.2005.10.024.

Tycko, J. *et al.* (2019) 'Identification and mitigation of pervasive off-target activity in CRISPR-Cas9 screens for essential non-coding elements', *BioRxiv*, pp. 1–60.

United Nations Office on Drugs and Crime (2018) *World Drug Report 2018*, *United Nations*. doi: 10.1080/00909887909365203.

Valton, J. *et al.* (2012) 'Overcoming transcription activator-like effector (TALE) DNA binding domain sensitivity to cytosine methylation', *Journal of Biological Chemistry*, 287(46), pp. 38427–38432. doi: 10.1074/jbc.C112.408864.

Vettese-Dadey, M. *et al.* (1996) 'Acetylation of histone H4 plays a primary role in enhancing transcription factor binding to nucleosomal DNA in vitro.', *The EMBO Journal*, 15(10), pp. 2508–2518. doi: 10.1002/j.1460-2075.1996.tb00608.x.

Viana, J. *et al.* (2017) 'Schizophrenia-associated methylomic variation: molecular signatures of disease and polygenic risk burden across multiple brain regions', *Human molecular genetics*, 26(1), pp. 210–225. doi: 10.1093/hmg/ddw373.

Visscher, P. M. *et al.* (2017) '10 Years of GWAS Discovery: Biology, Function, and Translation', *American Journal of Human Genetics*. ElsevierCompany., 101(1), pp. 5–22. doi: 10.1016/j.ajhg.2017.06.005.

Vojta, A. *et al.* (2016) 'Repurposing the CRISPR-Cas9 system for targeted DNA methylation', *Nucleic Acids Research*, 44(12), pp. 5615–5628. doi: 10.1093/nar/gkw159.

Waddington, C. H. (2012) 'The epigenotype', *International journal of epidemiology*, 41(1), pp. 10–13. doi: 10.1093/ije/dyr184.

Wagner, J. R. *et al.* (2014) 'The relationship between DNA methylation, genetic and expression inter-individual variation in untransformed human fibroblasts', *Genome Biology*, 15(2), pp. 1–17. doi: 10.1186/gb-2014-15-2-r37.

Waldo Zuardi, A. *et al.* (2012) 'A Critical Review of the Antipsychotic Effects of Cannabidiol: 30 Years of a Translational Investigation', *Current Pharmaceutical Design*, 18(32), pp. 5131–5140. doi: 10.2174/138161212802884681.

Walker, W. P. *et al.* (2006) 'Genetic Analysis of Attractin Homologs', *Genesis*, 224(September), pp. 219–224. doi: 10.1002/dvg.

Wang, C. and Zhang, Y. (2017) 'Schizophrenia in mid-adulthood after prenatal exposure to the Chinese Famine of 1959 – 1961', *Schizophrenia Research*. Elsevier B.V., 184, pp. 21–25. doi: 10.1016/j.schres.2016.11.030.

Wang, F. *et al.* (2015) 'Zdhhc15b Regulates Differentiation of Diencephalic Dopaminergic Neurons in Zebrafish', *Journal of Cellular Biochemistry*, 116(44), pp. 2980–2991. doi: 10.1002/jcb.25256.

Wang, H. *et al.* (2013) 'One-step generation of mice carrying mutations in multiple genes by CRISPR/cas-mediated genome engineering', *Cell*. Elsevier Inc., 153(4), pp. 910–918. doi: 10.1016/j.cell.2013.04.025.

Watson, C. T. *et al.* (2015) 'Genome-Wide DNA Methylation Profiling Reveals Epigenetic Changes in the Rat Nucleus Accumbens Associated With Cross-

Generational Effects of Adolescent THC Exposure', *Neuropsychopharmacology*. Nature Publishing Group, 40(13), pp. 2993–3005. doi: 10.1038/npp.2015.155.

Weber, A. R. *et al.* (2016) 'Biochemical reconstitution of TET1-TDG-BER-dependent active DNA demethylation reveals a highly coordinated mechanism', *Nature Communications*, 7(0372). doi: 10.1038/ncomms10806.

Wen, Z. *et al.* (2014) 'Synaptic dysregulation in a human iPS cell model of mental disorders', *Nature*. Nature Publishing Group, 515(7527), pp. 414–418. doi: 10.1038/nature13716.

Westra, H. J. *et al.* (2013) 'Systematic identification of trans eQTLs as putative drivers of known disease associations', *Nature Genetics*. Nature Publishing Group, 45(10), pp. 1238–1243. doi: 10.1038/ng.2756.

Wet, J. R. D. E. *et al.* (1985) 'Cloning of firefly luciferase cDNA and the expression of active luciferase in Escherichia coli', *Proceedings of the National Academy of Sciences of the Univted States of America*, 82(December), pp. 7870–7873.

Wheeler, A. L. and Voineskos, A. N. (2014) 'A review of structural neuroimaging in schizophrenia: From connectivity to connectomics', *Frontiers in Human Neuroscience*, 8(AUG). doi: 10.3389/fnhum.2014.00653.

Whiteford, H. A. *et al.* (2013) 'Global burden of disease attributable to mental and substance use disorders: Findings from the Global Burden of Disease Study 2010', *The Lancet*, 382(9904), pp. 1575–1586. doi: 10.1016/S0140-6736(13)61611-6.

Wijsman, E. M. *et al.* (2011) 'Genome-wide association of familial late-onset alzheimer's disease replicates BIN1 and CLU and nominates CUGBP2 in interaction with APOE', *PLoS Genetics*, 7(2). doi: 10.1371/journal.pgen.1001308.

Williams, H. J., Owen, M. J. and O'Donovan, M. C. (2007) 'Is COMT a susceptibility gene for schizophrenia?', *Schizophrenia Bulletin*, 33(3), pp. 635–641. doi: 10.1093/schbul/sbm019.

Wilson, W. *et al.* (2000) 'Brain morphological changes and early marijuana use: A magnetic resonance and positron emission tomography study', *Journal of Addictive Diseases*, 19(1), pp. 1–22. doi: 10.1300/J069v19n01_01.

Wockner, L. *et al.* (2014) 'Genome-wide DNA methylation analysis of human brain tissue from schizophrenia patients', *Translational Psychiatry*, 4(September 2013), pp. 1–8. doi: 10.1038/tp.2013.111.

World Health Organization (2014) *Schizophrenia*, *WHO*. World Health Organization. Available at: https://www.who.int/mental_health/management/schizophrenia/en/#.XWU-Ik3SKt4.mendeley (Accessed: 27 August 2019).

Wu, X. and Zhang, Y. (2017) 'TET-mediated active DNA demethylation: Mechanism, function and beyond', *Nature Reviews Genetics*. Nature Publishing Group, 18(9), pp. 517–534. doi: 10.1038/nrg.2017.33.

Xia, S. *et al.* (2015) 'Experimental validation of candidate schizophrenia gene CALN1 as a target for microRNA-137', *Neuroscience Letters*. Elsevier Ireland Ltd, 602, pp. 110–114. doi: 10.1016/j.neulet.2015.07.001.

Xicoy, H., Wieringa, B. and Martens, G. J. M. (2017) 'The SH-SY5Y cell line in Parkinson's disease research: a systematic review', *Molecular Neurodegeneration*. Molecular Neurodegeneration, 12(10), pp. 1–11. doi: 10.1186/s13024-017-0149-0.

Xie, F. *et al.* (2014) 'Seamless gene correction of β-thalassemia mutations in patient-specific iPSCs using CRISPR/Cas9 and piggyBac', *Genome Research*, 24(9), pp. 1526–1533. doi: 10.1101/gr.173427.114.

Yang, X. *et al.* (2014) 'Histone modifications are associated with 9-tetrahydrocannabinol-mediated alterations in antigenspecific t cell responses', *Journal of Biological Chemistry*, 289(27), pp. 18707–18718. doi: 10.1074/jbc.M113.545210.

Young, J. W. *et al.* (2009) 'Using the MATRICS to guide development of a preclinical cognitive test battery for research in schizophrenia', *Pharmacology and Therapeutics*. Elsevier Inc., 122(2), pp. 150–202. doi: 10.1016/j.pharmthera.2009.02.004.

Young, M. D. *et al.* (2010) 'Gene ontology analysis for RNA-seq: accounting for selection bias', *Genome Biology*, 11(2), pp. 1–12.

Zeissler, M. L. *et al.* (2016) 'Delta-9-tetrahydrocannabinol protects against MPP+ toxicity in SH-SY5Y cells by restoring proteins involved in mitochondrial biogenesis', *Oncotarget*, 7(29), pp. 46603–46614. doi: 10.18632/oncotarget.10314.

Zentner, G. E. and Henikoff, S. (2013) 'Regulation of nucleosome dynamics by histone modifications', *Nature Structural and Molecular Biology*. Nature Publishing Group, 20(3), pp. 259–266. doi: 10.1038/nsmb.2470.

Zhang, T. *et al.* (2018) 'Voltage-gated calcium channel activity and complex related genes and schizophrenia: A systematic investigation based on Han Chinese population', *Journal of Psychiatric Research*. Elsevier, 106(June), pp. 99–105. doi: 10.1016/j.jpsychires.2018.09.020.

Zhang, X. *et al.* (2007) 'HOXC6 and HOXC11 increase transcription of S100B gene in BrdU-induced in vitro differentiation of GOTO neuroblastoma cells into Schwannian cells', *Journal of Cellular and Molecular Medicine*, 11(2), pp. 299–306. doi: 10.1111/j.1582-4934.2007.00020.x.

Zhang, X. *et al.* (2019) 'Genome-wide analysis of epigenetic dynamics across human developmental stages and tissues', *BMC Genomics*, 20(Suppl 2). doi: 10.1186/s12864-019-5472-0.

Zhou, Y. *et al.* (2015) 'Prefrontal cortex and the dysconnectivity hypothesis of schizophrenia', *Neuroscience Bulletin*, 31(2), pp. 207–219. doi: 10.1007/s12264-014-1502-8.

Zhu, Z. *et al.* (2016) 'Integration of summary data from GWAS and eQTL studies predicts complex trait gene targets', *Nature Genetics*, 48(5), pp. 481–487. doi: 10.1038/ng.3538.

Zufferey, R. *et al.* (1998) 'Self-Inactivating Lentivirus Vector for Safe and Efficient In Vivo Gene Delivery', *Journal of Virology*, 72(12), pp. 9873–9880.

Zugno, A. I. *et al.* (2013) 'Chronic exposure to cigarette smoke during gestation results in altered cholinesterase enzyme activity and behavioral deficits in adult rat

offspring: Potential relevance to schizophrenia', *Journal of Psychiatric Research*. Elsevier Ltd, 47(6), pp. 740–746. doi: 10.1016/j.jpsychires.2013.02.001.