UNIVERSITY OF LIVERPOOL

Regulation at the schizophrenia-associated MIR137 locus and repetitive DNA in the regulation and evolution of brain-related pathways.

Thesis submitted in accordance with the requirements of the University of Liverpool for the degree of Doctor in Philosophy by

Olympia Gianfrancesco

September 2017

Acknowledgements

First and foremost, I would like to thank my PhD supervisors, Prof. John Quinn and Dr. Jill Bubb at the University of Liverpool, and Prof. David Collier at Eli Lilly and Company. Your support, encouragement, and guidance during my PhD has been invaluable. To David, thank you for your helpful feedback on my work, which has no doubt made me a better writer and a better scientist, and thank you for making me feel welcome during my time at Lilly. I also extend my thanks to Dr. Karim Malki and Dr. Nathan Lawless at Lilly for their work on RNA-seq which has contributed to this thesis.

To John and Jill, thank you for your genuine care for each member of the group, and for sharing both your wisdom and knowledge, as well as your great sense of humour. You have always encouraged us to work together not just as colleagues, but as a family, and it is for this reason that I have made friends for life in the Quinn Lab. Thank you to those that made me feel so welcome when I joined – Veridiana Pessoa, Alix Warburton, Maurizio Manca, Abigail Savage, Paul Myers, Kejhal Khursheed, and Kate Haddley. Thank you also to the newer members that have rapidly become great friends, and even housemates – Kim Billingsley, Emma Price, Ana Illera, and the Quen Lab Witchers, Jack Marshall and Ben Middlehurst. Whether with lab members past or present, I have been lucky enough to spend every day of the last four years with my favourite people, who kept me smiling and laughing even on the days when nothing was working!

Outside of the lab, I would like to thank my best friend, Alex Vaughn, as well as the rest of my wonderful friends and housemates in Liverpool. My time in this city has been the best of my life, and you have all been a huge part of that.

Big thanks go to my girlfriend, Beth Geary, for your unfailing belief in me, especially on the days that I couldn't see any way through this thesis. You have been there with hugs and kind words of reassurance every time. Thank you not only for giving me a calm space in which to work, but also for keeping me sane through the thesis write-up, with plenty of cuddles, board games, and marathons of Buffy the Vampire Slayer. I don't know how I would've done it without you.

Finally, I would like to thank my family, and especially my mum, both for your invaluable support, and for instilling in me the belief that the natural world is an unending source of wonder and beauty. These are the things that have kept me going throughout my studies and allowed me to get where I am today. Nine years ago, you were advised that I shouldn't apply for University. Thank you for trusting me when I decided to do it anyway.

That this volume exists at all is a testament to the love and support that I have gratefully received from all of you.

Contents

Publications

*Joint first author § Co-editor

Journal articles

1. **Gianfrancesco, O.**, Griffiths, D., Myers, P., Collier, DA., Bubb, VJ., Quinn, JP. Identification and potential regulatory properties of evolutionary conserved regions (ECRs) at the schizophrenia-associated MIR137 locus. J Mol Neurosci. 2016. doi: 10.1007/s12031-016-0812-x. PMID:27525637
2. **Gianfrancesco, O.**, Bubb, VJ., Quinn, JP. SVA retrotransposons as potential modulators of neuropeptide gene expression. Neuropeptides. 2016. doi: 10.1016/j.npep.2016.09.006. PMID:27743609
3. **Gianfrancesco, O.**, Warburton, A., Collier, DA., Bubb, VJ., Quinn, JP. Novel brain expressed RNA identified at the MIR137 schizophrenia-associated locus. Schizophr Res. 2016 doi: 10.1016/j.schres.2016.11.034. PMID:27913161
4. *Billingsley, K., Manca, M., **Gianfrancesco, O.**, Sharp, H., Bubb, VJ., Quinn, JP. Regulatory characterisation of the schizophrenia-associated CACNA1C proximal promoter and the potential role for the transcription factor EZH2 in schizophrenia aetiology. Schizophr Res. 2018 doi: 10.1016/j.schres.2018.02.036. PMID:29501388

Books

1. §Taylor, PJ., **Gianfrancesco, O.** Personal Experiences of Psychological Therapy for Psychosis. Oxford, Routledge (Taylor & Francis Group Ltd.).

Accepted for publication through Routledge as part of the International Society for Psychological and Social Approaches to Psychosis (ISPS) book series.

Abbreviations

5-HTTLPR	Serotonin transporter linked polymorphic region
aa	Amino acid
ADHD	Attention deficit hyperactivity disorder
AGO2	Argonaute 2
ALS	Amyotrophic lateral sclerosis
ASD	Autism spectrum disorder
BED	Browser extensible data
BLAST	Basic local alignment search tool
BLAT	BLAST-like alignment tool
bp	Base pair
BSA	Bovine serum albumin
C10orf26	Chromosome 10 open reading frame 26
C_2H_2	Cysteine 2 histidine 2 motif
C9orf72	Chromosome 9 open reading frame 72
CACNA	Calcium voltage-gated channel subunit alpha family
CACNA1C	Calcium voltage-gated channel subunit alpha 1 C
CACNB	Calcium voltage-gated channel auxiliary subunit beta family
CACNG	Calcium voltage-gated channel auxiliary subunit gamma family
CADPS2	Calcium-dependent secretion activator 2
cAMP	Cyclic adenosine monophosphate
cDNA	Complementary DNA
CENP-A	Centromere protein A
CEPH	Centre d'Etude du Polymorphisme Humain
CEU	Utah residents with Northern and Western European ancestry
ChIP	Chromatin immunoprecipitation
ChIP-seq	ChIP sequencing
CNS	Central nervous system
CO_2	Carbon dioxide
CpG	CG dinucleotide

CSMD1	CUB and sushi multiple domains 1
CTA	Cancer/testis antigen
CTCF	CCCTC-binding factor
DAT1	Dopamine transporter 1
DGCR8	DiGeorge critical region gene 8
DGKQ	Diacylglycerol kinase theta
DISC1	Disrupted in schizophrenia 1
DL-PFC	Dorsolateral prefrontal cortex
DMSO	Dimethyl sulfoxide
DNA	Deoxyribonuleic acid
DNAJC5	DnaJ heat shock protein family (Hsp40) member 5
dNTP	Deoxynucleotide triphosphate
DPYD	Dihydropyrimidine dehydrogenase
DRD	Dopamine receptor family
DRD4	Dopamine receptor 4
DSM	Diagnostic and statistical manual of mental disorders
EBI	European Bioinformatics Institute
ECR	Evolutionary conserved region
EDTA	Ethylenediaminetetraacetic acid
EED	Embryonic ectoderm development
EMBL	European Molecular Biology Laboratory
ENCODE	Encyclopaedia of DNA elements
ERV	Endogenous retrovirus
EST	Expressed sequence tag
EZH2	Enhancer of zeste homologue 2
FOS	Fos proto-oncogene, AP-1 transcription factor subunit
FTLD	Fronto-temporal lobe degeneration
FUS	Fused in sarcoma
G x E	Gene x environment
G4	G quadruplex

GABA	Gamma-aminobutyric acid
GABBR2	Gamma-aminobutyric acid type B receptor subunit 2
GABRA2	Gamma-aminobutyric acid type A receptor alpha 2 subunit
GABRG3	Gamma-aminobutyric acid type A receptor gamma 3 subunit
GAGE	G antigen family
GAK	Cyclin G-associated kinase
GENCODE	Encyclopaedia of genes and gene variants
GO	Gene ontology
GRIA	Glutamate ionotropic receptor AMPA type subunit family
GRID1	Glutamate ionotropic receptor delta type subunit 1
GRID2	Glutamate ionotropic receptor delta type subunit 2
GRIN	Glutamate ionotropic receptor NMDA type subunit family
GTEx	Genotype-tissue expression project
GWAS	Genome Wide Association Studies
H3K27ac	Acetylation of lysine 27 in histone H3
H3K4me1	Monomethylation of lysine 4 in histone H3
H3K4me3	Trimethylation of lysine 4 in histone H3
H3K9ac	Acetylation of lysine 9 in histone H3
HERV	Human endogenous retrovirus
hg	Human genome
HOTAIR	HOX transcript antisense RNA
HSP40	Heat shock protein 40 family
HSP70	Heat shock protein 70 family
HTT	Huntingtin
ICD	International classification of disease
Imir137	Internal MIR137 promoter
iPSC	Induced pluripotent stem cell
JUN	Jun proto-oncogene, AP-1 transcription factor subunit
kb	Kilobase
kDa	Kilodalton

KEGG	Kyoto encyclopaedia of genes and genomes
KRAB	Krüppel-associated box
L1HS	LINE-1 subfamily human specific
L1PA2	LINE-1 subfamily PA2
L1PA3	LINE-1 subfamily PA3
LARII	Luciferase assay reagent II
LD	Linkage disequilibrium
lincRNA	Long intergenic non-coding RNA
LINE	Long interspersed nuclear element
lncRNA	Long non-coding RNA
LTR	Long terminal repeat
MAGE	Melanoma antigen family
MAOA	Monoamine oxidase A
MAST2	Microtubule-associated serine/threonine-protein kinase 2
Mb	Megabase
MET	cMet; tyrosine-protein kinase Met
$MgCl_2$	Magnesium chloride
MGI	Mouse genome informatics
MIAT	Myocardial infarction associated transcript
MIR137	MicroRNA-137
MIR137HG	MicroRNA-137 host gene
MIR941	MicroRNA-941
miRNA	MicroRNA
mRNA	Messenger RNA
Mya	Million years ago
Myrs	Million years
NCBI	National centre for biotechnology information
NCL	Neuronal ceroid lipofuscinosis
ncRNA	Non-coding RNA
NHGRI	National Human Genome Research Institute

NHPRTR	Non-human primate reference transcriptome resource
nt	Nucleotide
OD	Optical density
OR	Odds ratio
ORF	Open reading frame
PARK7	Parkinson disease protein 7
PBMC	Peripheral blood mononuclear cells
PCR	Polymerase chain reaction
PD	Parkinson’s disease
PGC	Psychiatric genomics consortium
PIC	Pre-initiation complex
PLB	Passive lysis buffer
PPT1	Palmitoyl-protein thioesterase 1
PRAME	Preferentially expressed antigen in melanoma
PRC2	Polycomb repressive complex 2
Pre-miRNA	Precursor RNA
Pri-miRNA	Primary microRNA
PTSD	Post-traumatic stress disorder
REST	Repressor element-1 silencing transcription factor
RIN	RNA integrity
RIP	Retrotransposon insertion polymorphism
RISC	RNA induced silencing complex
RNA Pol II	RNA Polymerase II
RNA	Ribonucleic acid
RNA-seq	RNA sequencing
RPKM	Reads per kilobase of transcript per million mapped reads
rpm	Revolutions per minute
SH-SY5Y	Human-derived neuroblastoma cells
SINE	Short interspersed nuclear element
SLC6A4	Serotonin transporter (5-HTT) solute carrier family 6 member 4

snoRNA	Short nucleolar RNA
SNP	Single nucleotide polymorphism
snRNA	Short nuclear RNA
SUZ12	Suppressor of zeste 12 homologue
SV40	Simian virus 40
SVA	SINE-VNTR-Alu
SYNE2	Spectrin repeat containing nuclear envelope protein 2
TBP	TATA-box binding protein
TBPL2	TATA box-binding protein-like protein 2
TCF4	Transcription factor 4
TE	Transposable element
TFBS	Transcription factor binding site
TR	Tandem repeat
TRBP	TAR binding protein
TRIM28	Tripartite motif containing 28
tRNA	Transfer RNA
TRPV	Transient receptor potential cation channel subfamily V
UCSC	University of California, Santa Cruz
UTR	Untranslated region
UV	Ultraviolet
VNTR	Variable number tandem repeat
XPO5	Exportin 5
ZNF	Zinc finger (C_2H_2-type)
ZNF804A	Zinc finger 804A

List of figures

List of tables

Abstract

Maintaining the appropriate transcriptional balance in the cell is a complex process involving numerous mechanisms, including the action of transcription factors and non-coding regulatory elements. Such processes are key to maintaining healthy central nervous system (CNS) functioning, and can be modulated through the interaction of both genes and environment in a 'G x E' mechanism. If the regulation of a certain gene or gene set is altered inappropriately in the brain, this can result in neuronal dysfunction which may contribute to psychiatric or CNS conditions. This thesis primarily aimed to extend our understanding of transcriptional regulation at the MIR137 schizophrenia-associated locus, and to add to our understanding of the role of repetitive DNA and retrotransposons in the regulation and evolution of genes involved in wider CNS pathways.

The chromosome 1p21.3 locus encompassing the microRNA, MIR137, has been repeatedly highlighted by GWAS as one of the most robust loci for association with schizophrenia. The evidence presented in this thesis identified multiple evolutionary conserved regions (ECRs) which act as transcriptional regulators at this locus, as well as a regulatory gene network comprising MIR137 and the transcriptional regulators REST and EZH2, which are likely to modulate the expression of multiple CNS- and schizophrenia-associated gene sets. Extending our view of the MIR137 locus identified a brain-expressed RNA, EU358092, which shared near identical expression and regulatory profiles to MIR137, suggesting potential co-expression and -regulation of RNAs across this locus.

The second half of this thesis explored repetitive DNA, including variable number tandem repeats (VNTRs) and the retrotransposon subfamilies, Long Interspersed Nuclear Elements (LINEs) and SINE-VNTR-Alus (SVAs), which have been shown to

act as modulators of gene expression. Common polymorphisms in the VNTR containing MIR941, a human-specific, brain expressed microRNA at chromosome 20q13.3, resulted in altered copy number of MIR941, with two genotypes being specific to a schizophrenia cohort. SVAs were implicated in the recent evolution of multiple zinc finger loci, which may have had the potential to alter the regulation of large transcriptional networks in a species-specific manner, while LINE elements were likely to have been involved in recent genomic remodelling around GABA and glutamate signalling genes.

Taken together, the work contained in this thesis considered the roles of a wide range of DNA elements with relevance to CNS-expressed genes, from the oldest and most highly conserved regions of the genome, to the most recent retrotransposon insertions. This work identified roles for these elements in the evolution and regulation of genes involved in schizophrenia risk and neuroprotection, and further identified gene networks and additional transcripts which may contribute to the maintenance of healthy brain functioning. The impact of genetic variation at these regions - in the form of single nucleotide polymorphisms (SNPs), altered VNTR copy number, or polymorphic retrotransposon insertions - and their effect on CNS functioning, has been a key theme throughout this work. In conclusion, this would provide evidence to suggest that genetic polymorphisms which alter the function, size, or location of such elements at loci involved in brain-related processes could contribute to schizophrenia risk in a way that would likely be modulated through an interaction between environmental stimuli and genotype.

Chapter 1

General Introduction

1.1 An overview of non-coding RNAs

Since the first draft of the human genome was published in 2001, demonstrating that 2% of the genome comprises coding DNA, it has become apparent that much of the remaining 98% is also likely to be functional (The ENCODE Project Consortium 2012). This has been supported by advances in transcriptomics and genome-wide association studies (GWAS), which have demonstrated that approximately 80% of the genome is capable of being transcribed (Djebali et al. 2012, Hangauer, Vaughn and McManus 2013), and that 88% of disease- or trait-associated single nucleotide polymorphisms (SNPs) reside in non-coding regions (Hindorff et al. 2009). Such results are likely to be highlighting not only the existence of non-coding RNAs (ncRNAs) and regulatory elements that can alter transcription, translation, and local chromatin structure, but also their importance in maintaining the regulatory balance within the cell. This introductory chapter first presents an overview of non-coding RNAs, and further introduces a range of non-coding regulatory elements and the mechanisms through which they can influence chromatin structure and gene expression.

Non-coding RNAs refer to RNA transcripts that are not translated into protein, and can be subdivided into multiple categories based on characteristics such as their function or size. NcRNAs can be divided into the broad categories of housekeeping ncRNAs and regulatory ncRNAs. The former group are constitutively expressed and include transfer RNAs (tRNAs), ribosomal RNAs (rRNAs), and small nuclear/nucleolar RNAs (snRNAs/snoRNAs). Regulatory ncRNAs typically cover microRNAs (miRNAs) and long non-coding RNAs (lncRNA) (Esteller 2011), with the latter further subdivided into long intergenic non-coding RNAs (lincRNAs) and anti-sense RNAs (AS-RNAs)

(Derrien et al. 2012, Harrow et al. 2012). This section focuses on miRNAs and lncRNAs, which are key themes within this thesis.

1.1.1 MicroRNAs

1.1.1.1 A brief overview of microRNA biogenesis

Mature miRNAs are typically 22 nucleotides (nt) in length and are known to play key roles in the post-transcriptional regulation of gene expression. Such regulation predominantly occurs through binding of the miRNA seed sequence, a region of 2-8 nt at the 5' end of the miRNA, to complementary sequences in the 3' untranslated region (3' UTR) of their target mRNAs. With 30-80% of human genes predicted to be regulated by miRNAs (Lewis, Burge and Bartel 2005, Friedman et al. 2009, Lu and Clark 2012), this makes them a key class of regulators within the cell and therefore of significant interest in further understanding both health and disease states.

In order to be processed into mature ~22 nt sequences, miRNAs are first transcribed by RNA Polymerase II or RNA Polymerase III (RNA Pol II or III) as a primary miRNA (pri-miRNA) (Borchert, Lanier and Davidson 2006), which is typically at least 1 kilobase (kb) in length and folds to form a double stranded hairpin-loop structure (Lee et al. 2004) (Figure 1.1). This hairpin structure is recognised by the DGCR8 protein (Han et al. 2006), which is in complex with the Drosha protein. Together, these make up the minimal microprocessor complex, which cleaves the hairpin from the ≥ 1 kb pri-miRNA to form a precursor miRNA or pre-miRNA of between 60-90 nt (Krol et al. 2004). This processing by the microprocessor complex leaves single stranded overhangs of 2-3 nt, which are recognised by the transport protein, Exportin-5 (XPO5), which transports the pre-miRNAs into the cytoplasm for further processing by Dicer (Lund et al. 2004).

Figure 1.1 An overview of miRNA biogenesis.

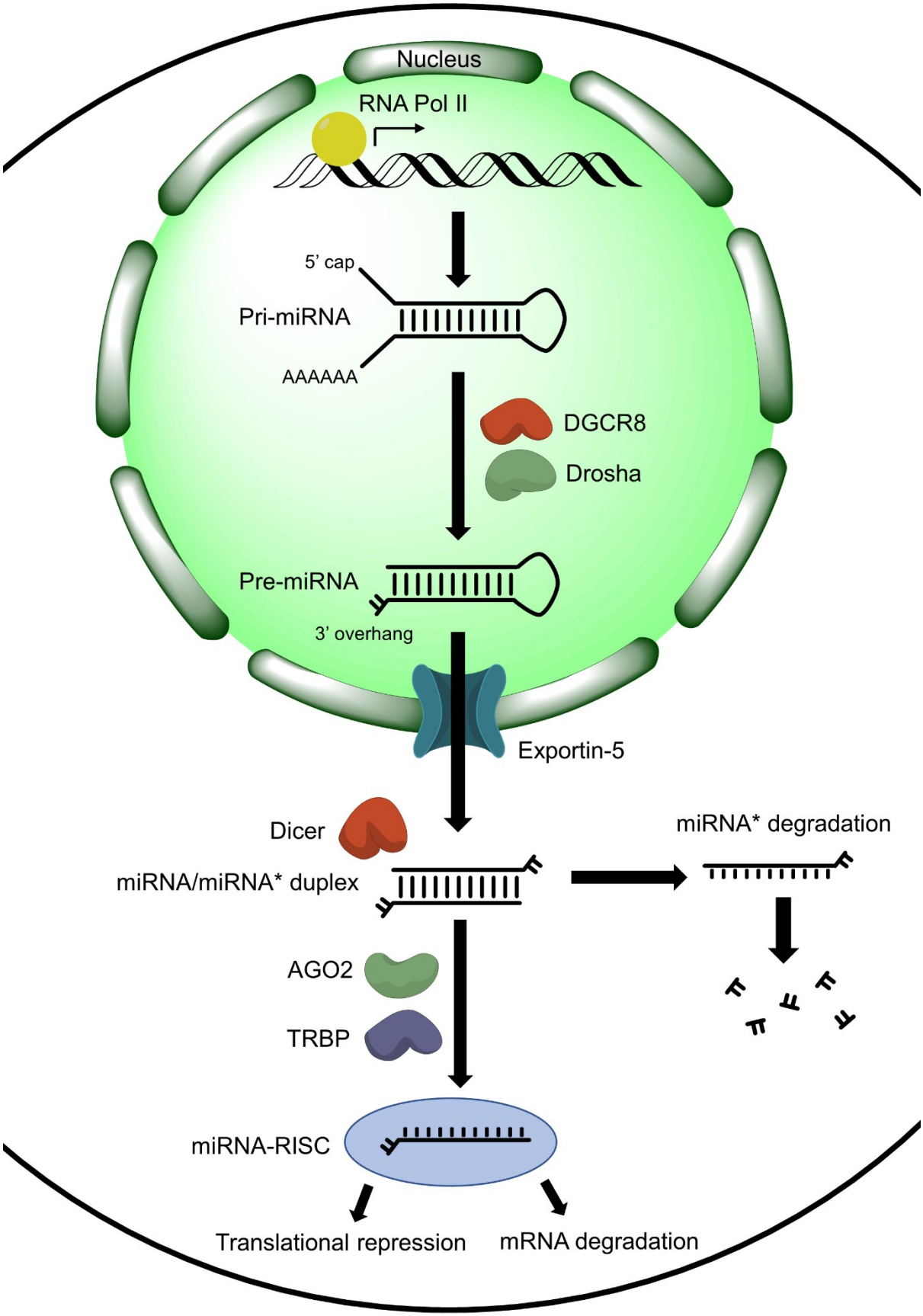


Figure 1.1 An overview of miRNA biogenesis.

Canonical miRNAs are first transcribed as a primary transcript (pri-miRNA) by RNA Pol II. The pri-miRNA is cleaved through the action of Drosha, which is complexed with DCGR8 within the microprocessor unit, resulting in a pre-miRNA which comprises a stem-loop structure of approximately 70 nt. The pre-miRNA is exported to the nucleus by Exportin-5, where it undergoes further processing by Dicer to recognise and cleave the loop structure, leaving a miRNA/miRNA duplex. One strand of the duplex is preferentially loaded into the miRNA-RISC, which is typically made up of Dicer, TRBP, and AGO2. Once in the RISC, the miRNA-RISC is a functional unit that can go on to target mRNA for repression or degradation. The second strand of the duplex (miRNA*) is typically degraded.*

Dicer cleaves the loop of the hairpin-loop structure to leave a ~22 base pair (bp) duplex with one copy of the mature miRNA in a double strand complex with its complementary sequence (referred to as miRNA*). One strand of this duplex is then incorporated into the RNA-induced silencing complex (RISC), while the second strand (usually the miRNA*) is typically degraded. Along with Dicer and the mature miRNA, Argonaute proteins (AGO2) and the TAR binding protein (TRBP) are key components of the miRNA-RISC complex (Chendrimada et al. 2005). After recognising their target mRNA through complementary sites contained within the miRNA, Argonaute proteins cleave the mRNA to prevent translation into proteins, thereby post-transcriptionally regulating expression.

While this process is typical for intergenic miRNAs, work by Rodriguez et al. demonstrated that approximately 40% of miRNAs are found within the introns of protein-coding genes, with a further 10% residing within lncRNA transcripts (Rodriguez et al. 2004). Rodriguez et al. and Baskerville et al. further showed that the expression of intronic miRNAs often displays strong correlation with expression of their host gene, which would suggest co-expression of the transcripts, potentially with the miRNA originating from the same transcript as the host gene (Rodriguez et al. 2004, Baskerville and Bartel 2005). Indeed, it is known that pre-miRNA structures can be processed from the intron of host gene mRNA through microprocessor interaction with the spliceosome, either before or after splicing the introns from the mRNA (Kataoka, Fujita and Ohno 2009, Kim and Kim 2007), which would support the research suggesting co-expression of host genes and their intronic miRNAs.

1.1.1.2 MicroRNAs in the brain and in schizophrenia risk

Work by Ludwig et al. to define the expression patterns of 1997 miRNAs across 24 regions of the human body demonstrated that 11-16% of miRNAs were tissue specific,

with 82.9% having ‘intermediate’ specificity, displaying neither ubiquitous housekeeping expression nor tissue specificity (Ludwig et al. 2016). This work identified three clusters in which miRNA expression was concordant across tissues in the two bodies tested. The first cluster comprised the kidney, liver, stomach, and small intestine, while a second cluster involved the thyroid, nerve, muscle, myocardium, and colon. The third cluster comprised all brain and CNS tissues tested, with four miRNAs found to be expressed exclusively in the brain (Ludwig et al. 2016). Earlier studies of 345 miRNAs across 40 normal human tissues identified miRNAs such as miR-129, miR-219, and miR-330, that were largely brain-specific, with very little or no expression in other tissues tested, with further miRNAs that were specific to the brain and peripheral blood mononuclear cells (PBMCs) (Liang et al. 2007). Liang et al. also noted two miRNAs, miR-199 and miR-214, that were decreased in the adult and foetal brain and adult PBMCs compared to all other tissues tested. It was demonstrated that predicted target genes of miR-199 and miR-214 were required for healthy development and function of the nervous system and haematopoietic system, suggesting that expression of these target genes would be restricted in all tissues except the brain and blood (Liang et al. 2007).

The importance of miRNAs in the brain has been demonstrated in part through studies in mouse models, which show that core components of miRNA biogenesis and the RISC complex, such as Dicer and Argonaute proteins, are localised and enriched at the post-synaptic density, and are modified by neuronal activation (Lugli et al. 2005). Similarly, pri-miRNAs are found in the synaptic fraction in mouse neurons, and are also found to be enriched in the post-synaptic density, along with microprocessor components, Drosha and DGCR8 (Lugli et al. 2008, Lugli et al. 2012). This would

suggest localised miRNA production at synapses to allow rapid, spatio-temporal gene regulation in response to neuronal activation.

Work by Paschou et al. used computational methods to analyse miRNA target sites in human genes encoding 242 pre-synaptic proteins and 304 post-synaptic proteins, demonstrating that 91% of the genes tested were predicted to be miRNA targets (Paschou et al. 2012). In this analysis, it was demonstrated that 77% of pre-synaptic transcripts and 80% of post-synaptic transcripts were regulated by 10 miRNA families. Of these 10 families, five were thought to be either completely or partially primate-specific due to the lack of currently known homologues in other mammalian species.

Comparison of gene expression in human and primate brains has demonstrated that the evolutionary divergence of transcriptomic profiles during development in the human brain was 3-5 times quicker than in the chimpanzee brain (Somel et al. 2011). Here, evolutionary divergence refers to variation in gene expression between different species at different points in development and across the lifetime. Somel et al. demonstrated that this accelerated transcriptomic variation in humans was likely due in part to miRNAs, the evolution of which would allow new regulatory patterns of large networks of genes, rather than a single gene. Somel et al. demonstrated a subset of developmental miRNAs in the human prefrontal cortex that had evolved more rapidly than many other classes of genes, including transcription factors (Somel et al. 2011). This would suggest that evolution of the human brain was directed in part by miRNA-mediated regulatory changes to large networks of brain-expressed genes.

Evidence to support the key roles of miRNAs in brain development and function come from genetic conditions which disrupt components involved in miRNA biogenesis and function, such as DiGeorge syndrome (22q11.2 deletion syndrome) and Fragile X

syndrome. Individuals with DiGeorge syndrome typically have a ~3 megabase deletion at the chromosome 22q11.2 region which results in haploinsufficiency of DGCR8, a key component of the microprocessor complex, as well as approximately 35 other genes, many of which are poorly characterised (Hacıhamdioğlu, Hacıhamdioğlu and Delil 2015). This results in intellectual disability and a significantly increased risk of psychiatric conditions, including a 20-40% incidence of schizophrenia or psychosis (Monks et al. 2014, Jonas, Montojo and Bearden 2014, Schneider et al. 2014). Similarly, Fragile X syndrome results from a CGG trinucleotide repeat expansion in the FMR1 gene encoding the FMRP protein. FMRP is involved in miRNA biogenesis and function through its role in the translation of Drosha and its interaction with AGO2 and the miRNA-RISC complex (Kenny and Ceman 2016, Li et al. 2014, Wan et al. 2017). Alteration of FMRP in individuals with Fragile X syndrome can lead to intellectual disability and seizures, as well as behavioural and psychiatric conditions (Fatemi and Folsom 2011, Heard et al. 2014, Kidd et al. 2014, Schneider et al. 2016). In particular, 22q11.2 deletion has become a widely used model in schizophrenia research (Ellegood et al. 2014, Gur et al. 2017, Meechan et al. 2015).

Returning to work by Paschou et al., of the miRNAs predicted to regulate synaptic transcripts in their computational analysis, the authors report that 25 such miRNAs had previously been associated with schizophrenia through analyses which demonstrated their altered expression in the brains of individuals with this diagnosis (Paschou et al. 2012, Beveridge and Cairns 2012). Further changes in miRNA expression levels have been demonstrated in the brains of individuals with a diagnosis of schizophrenia. For example, Smalheiser et al. found 73 miRNAs with significantly reduced expression in the dorsolateral prefrontal cortex (DL-PFC) of individuals with schizophrenia (Smalheiser et al. 2014). Typically, the more highly enriched a miRNA

was in the synapses of control brains, the more significantly downregulated the miRNA was in the brains of individuals with schizophrenia. This would suggest that alterations in the biogenesis and processing of synaptic miRNAs could be an underlying mechanism in schizophrenia biology (Smalheiser et al. 2014). MiRNAs have also been demonstrated to be altered in the peripheral blood of individuals with schizophrenia and psychosis, and have been proposed as potential biomarkers (Wei et al. 2015, Sun et al. 2015b, Fan et al. 2015). For example, the human-specific MIR941 was identified as being lost from the peripheral blood miRNA networks of high risk individuals who progressed to psychosis at two-year follow up, with no such change in those that did not progress to psychosis (Jeffries et al. 2016). This may suggest that changes in miRNA networks including MIR941 could correlate with the transition from high risk to psychosis. Further understanding of the MIR941 locus in schizophrenia is the focus of Chapter 5.

Genome-wide association studies have also highlighted a number of miRNAs which are likely to play roles in schizophrenia biology, with work by multiple groups demonstrating that miRNAs are significantly enriched for schizophrenia GWAS SNPs (Williamson et al. 2015, Goulart et al. 2015). Most notably, the locus encompassing the miRNA, MIR137, has been repeatedly demonstrated by GWAS to be associated with schizophrenia (Ripke et al. 2013, Schizophrenia Psychiatric Genome-Wide Association Study (GWAS) Consortium 2011). Risk variants at this locus have been associated with altered connectivity between the prefrontal cortex and regions such as the amygdala and hippocampus in healthy controls (Liu et al. 2014a, Mothersill et al. 2014), as well as altered cognitive function and more prominent negative symptoms in individuals with a diagnosis of schizophrenia (Kuswanto et al. 2015, Green et al.

2013, Cummings et al. 2013). Regulation and variation around the MIR137 locus is investigated in Chapters 3 and 4.

1.1.2 Long non-coding RNAs (lncRNAs)

The second class of ncRNAs which will be considered in this thesis are lncRNAs. LncRNAs are defined as transcripts over 200 bp in length which lack protein coding capacity, with the most recent GENCODE release documenting 14,880 such transcripts (Derrien et al. 2012). Of these transcripts, Derrien et al. demonstrated that 64% were intergenic (sometimes referred to as lincRNAs), while the remaining 36% were genic, overlapping introns or exons of protein-coding genes, or with introns spanning across such genes. The majority of lncRNAs are expressed in the same way as mRNAs, being transcribed by RNA Polymerase II from regions that share similar chromatin states and histone modifications. They typically have a 5' cap and a poly(A) tail, with 98% undergoing splicing. For the most part, lncRNAs and mRNAs are indistinguishable, except for the lack of open reading frames in the former. The length and lower abundance of lncRNAs may also be considered distinguishing factors. LncRNAs are typically shorter in length than mRNAs, and the lncRNA class is significantly enriched for two-exon transcripts (Derrien et al. 2012). Annotated lncRNAs in the GENCODE v7 database show that 42% of lncRNAs have only two exons, which is in contrast to protein-coding genes, of which 6% have two exons (Derrien et al. 2012). Further, lncRNAs are expressed at lower levels than protein-coding genes, and often display tissue-specific expression patterns. Cabili et al. characterised the expression patterns of the lincRNA subclass of lncRNAs across 24 tissues and cell types through RNA-seq, which demonstrated that lincRNAs are on average expressed at approximately 10-fold lower levels than protein-coding genes,

with 78% of lincRNAs displaying tissue specificity compared to just 19% of mRNAs (Cabili et al. 2011).

Derrien et al. demonstrated that approximately 30% of lncRNAs in the human genome are primate-specific, with 0.7% being human-specific (Derrien et al. 2012). Work by Cabili et al. showed that 28% of intergenic lncRNAs (lincRNAs) are located within 10 kb of a protein-coding gene, with the protein-coding genes within 10 kb of a lincRNA showing enrichment for roles in development and the regulation of gene expression (Cabili et al. 2011). This study further demonstrated that lincRNAs which are adjacent to protein-coding genes display significant correlation in terms of expression levels to their neighbouring protein-coding gene, likely indicating co-expression. However, the same correlation was demonstrated for neighbouring protein-coding genes, suggesting that this is an effect of proximity to open chromatin, rather than identifying *cis*-regulatory activity of particular lincRNAs (Cabili et al. 2011).

LncRNAs are known to function as regulators of expression through numerous mechanisms. For example, expression of antisense lncRNAs can directly interfere with sense-oriented host gene transcription through steric hindrance. Similarly, expression of lncRNAs across or near to host gene enhancer elements can block or activate the action of such elements, either by preventing DNA or protein interactions due to the movement of transcriptional machinery over the element, or by creating an area of open chromatin around the enhancer (Kornienko et al. 2013). On the other hand, lncRNA transcripts may have the capacity to recruit or block the action of epigenetic modifiers, transcription factors, or polymerases (Kornienko et al. 2013). They can also act as molecular scaffolds to bring together multiple protein units, or act as molecular decoys to sequester specific proteins, among other gene-regulatory mechanisms (Ulitsky and Bartel 2013).

Analysis of the most differentially expressed lncRNAs in the GENCODE v7 database revealed that approximately 40% were accounted for by a brain-specific lncRNA cluster (Derrien et al. 2012). This would highlight their importance as regulatory transcripts in the brain, with lncRNAs increasingly being found to play roles in neurodevelopmental, neurodegenerative, and psychiatric conditions. Work by Ziats et al. identified over 200 differentially expressed lncRNAs in the brains of individuals with autism spectrum disorder (ASD), which were enriched at genomic loci containing genes that have been shown to be involved in cerebral cortex cell migration, and targets of MIR103 and MIR107. These miRNAs have been shown to regulate CDK5R1 expression affecting neuronal migration (Moncini et al. 2011), and altered expression of MIR107 has been demonstrated in psychiatric and neurodegenerative conditions (Santarelli et al. 2011, Nelson and Wang 2010). This may suggest that altered expression of lncRNAs within loci with known CNS function may be an underlying mechanism in ASD (Ziats and Rennert 2013). In peripheral blood, almost 4000 lncRNAs were found to be deregulated in individuals with autism, and over 2000 deregulated in the blood of individuals with major depression (Liu et al. 2014b, Wang et al. 2015).

There is now significant evidence to suggest that altered expression or regulation of lncRNAs is likely to play a role in schizophrenia biology, and as such, their transcriptional regulation is of significant interest in understanding the mechanisms which contribute to this condition. Liao et al. demonstrated methylation differences over lncRNAs in DNA from the peripheral blood of both men and women with schizophrenia, suggesting that deregulation of lncRNAs may contribute to schizophrenia pathology (Liao et al. 2015a, Liao et al. 2015b). This is in agreement with research by Ren et al. which showed modulation of lncRNA networks in the

peripheral blood of individuals with early onset schizophrenia (Ren et al. 2015c). In particular, SNPs across the lncRNA, MIAT, which is known to play a role in brain development, have been associated with risk for paranoid schizophrenia in the Han Chinese population (Rao et al. 2015). Further, MIAT has been shown to be downregulated in the brains of individuals with schizophrenia, with over-expression and knockdown of the lncRNA in human induced pluripotent stem cell (iPSC)-derived neurons shown to result in altered splicing of the schizophrenia-associated DISC1 gene (Barry et al. 2014).

In this thesis, Chapter 3 Part II focuses on lncRNAs in schizophrenia, identifying and characterising a novel, brain-expressed lncRNA, EU358092, at the schizophrenia-associated MIR137 locus.

While it is clear from this section that non-coding DNA can give rise to vast numbers of regulatory transcripts with important functions in modifying CNS-expressed genes, a significant portion of non-coding DNA also functions to regulate expression without being transcribed. The remainder of this introduction provides an overview of non-coding regulatory elements, and how variation in such elements could act as risk factors for psychiatric conditions, either alone, or particularly when paired with stressful environmental stimuli.

1.2 Transcriptional regulatory elements

The previous section gave a brief overview of two classes of ncRNA that have been shown to regulate gene expression, with likely roles in the brain in both health and disease states. In addition, non-coding DNA in the human genome also contains a wide range of elements that are vital in maintaining the proper regulation of gene expression. Such elements can include the promoter regions, typically found around

the transcriptional start sites of a gene, as well as numerous other regulatory elements, ranging widely in their evolutionary age and sequence identity between species, from highly conserved sequences known as Evolutionary Conserved Regions (ECRs), to highly variable repeated sequences which have changed in size or copied to other regions of the genome through species divergence and evolution, in the case of Variable Number Tandem Repeats (VNTRs) and mobile sequences known as Transposable Elements (TEs).

1.2.1 Evolutionary Conserved Regions (ECRs)

Evolutionary conserved regions in the human genome were first described by multiple independent groups in the early 2000s (Sandelin et al. 2004, Bejerano et al. 2004). While the minimum required length and percentage sequence identity vary between studies, ECRs are typically defined as regions of sequence that retain 70% or more sequence identity when compared to the corresponding region of sequence in other species, such as when comparing the human genome to the mouse genome.

Some coding exons are highly conserved between species in order to preserve important aspects of protein function and such regions of DNA would therefore be considered to be ECRs. For example, Bejerano et al. identified 481 ‘ultra-conserved regions’ over 200 bp in size that were conserved with 95-100% sequence similarity between human and mouse genomes. However, only one quarter of these were accounted for by exons, revealing the presence of numerous, as yet uncharacterised regions which were suspected to have important functions (Bejerano et al. 2004). While some exons would be considered ECRs, in this thesis, the use of the term ECR will refer to non-coding regions of high species conservation unless otherwise stated.

Strong evolutionary conservation is a useful indicator of likely function, as any variation at these regions has been restricted or selected out through evolution, presumably to protect and maintain the function associated with the region of conservation. Indeed, work by Prabhakar et al. demonstrated that comparative sequence analysis of multiple primate and mammalian species against the human genome was sufficient to identify conserved *cis*-regulatory elements in the human genome (Prabhakar et al. 2006). However, the reliability of this method was much reduced and struggled to identify true conserved regulators when comparing the human genome against more distant evolutionary relatives such as birds, amphibians, or fish. MacKenzie and Quinn have also demonstrated that comparative sequence analysis across multiple species is a useful method in identifying regulatory elements around CNS-expressed genes (MacKenzie and Quinn 2004).

Evolutionary conserved regions are not distributed randomly in the vertebrate genome, and have been shown to cluster in arrays around their likely target genes, further suggesting a role for these elements in gene regulation (Sandelin et al. 2004). Studies comparing the human genome against multiple vertebrate species, including the mouse and puffer fish, have demonstrated that ECRs are preferentially found around genes involved in development (Sandelin et al. 2004, Woolfe and Elgar 2008, Woolfe et al. 2005, Visel et al. 2008). This would suggest that regions of extreme conservation across the genome are likely to be common to most vertebrates, as many function to regulate crucial stages in early vertebrate development.

Some distant species, such as the mouse and zebrafish, have low levels of overlap in terms of ECRs, with only 10.5% of evolutionary conserved enhancers in the mouse genome having homologues in the zebrafish. However, the 11 homologous ECRs in the two species were found to reside at key developmental transcription factors

including homeobox (Hox) and paired box (Pax) genes (Plessy et al. 2005), which are known to play crucial roles in vertebrate body patterning and tissue specification during early development (Wellik 2007, Mallo, Wellik and Deschamps 2010, Blake and Ziman 2014, Wang et al. 2008). It has further been demonstrated that, even in species with little overlap in conserved sequences, such as when comparing vertebrates with invertebrates, ECRs in each species are still enriched at loci including genes involved in development. For example, Vavouri et al. describe ECRs as being both highly conserved within specific lineages, and yet highly divergent between lineages, demonstrating that while there is little evidence of homologous vertebrate ECRs in invertebrate genomes, ECRs in *C. elegans* and *D. melanogaster* share similar distribution patterns to vertebrate ECRs, demonstrating over-representation at genes involved in development (Vavouri et al. 2007). This would suggest parallel evolution of ECRs in multiple species as a mechanism of regulating important processes in early development.

ECRs have been noted as being active in brain development, with work by Matsunami and Saitou demonstrating that up to a third of ECRs in paralogous regions showed regulatory activity within the brain at multiple stages of development, particularly around genes encoding CNS-expressed transcription factors (Matsunami and Saitou 2013). Similarly, analysing conserved elements in hominids through the comparison of human, chimpanzee, gorilla, orangutan, and gibbon genomes demonstrated that hominid ECRs were preferentially located around transcriptional start sites of genes involved in gene regulation, development, and nervous system processes (Mahmoudi Saber and Saitou 2017). The role of ECRs in regulating brain-expressed genes both *in vitro* and *in vivo* is further supported by work from Paredes et al. and Khursheed et al. Previous work has identified and characterised an ECR within the human dopamine

receptor D4 gene (DRD4), which was shown to be a positive regulator of reporter gene expression in studies using primary cultured cells from the rat frontal cortex (Paredes et al. 2011). Others in the lab characterised an ECR adjacent to the FUS (Fused in Sarcoma) promoter, a candidate gene for motor neuron disease (Kwiatkowski et al. 2009, Vance et al. 2009), demonstrating positive regulatory activity in reporter gene assays and showing activity in developing motor neurons in a chick embryo model (Khursheed et al. 2015). Finally, we have described eight ECRs at the schizophrenia-associated MIR137/DPYD locus, six of which were shown to be positive transcriptional regulators, and two negative transcriptional regulators in a human cell line model (Gianfrancesco et al. 2016b, Gianfrancesco et al. 2017). This work is discussed in Chapter 3.

ECRs are theorised to have retained such high conservation due to their sequences including important transcription factor binding sites (TFBSs) which would allow them to influence the expression of nearby genes. ECRs have been shown to be over-represented for specific transcription factor binding sites, including sites for REST (RE-1 silencing transcription factor) and CTCF (CCCTC-binding factor) (Viturawong et al. 2013, Xie et al. 2007). REST is a well-known regulator of neuronal genes (Qureshi and Mehler 2009, Johnson et al. 2008, Ballas et al. 2005, Gao et al. 2011, Mukherjee et al. 2016, Warburton et al. 2015b), which would be consistent with ECRs being noted around gene loci involved in CNS function, while CTCF is a well characterised regulator of gene expression and genome organisation (Kim, Yu and Kaang 2015), which would support the function of ECRs in gene regulation. Further work has demonstrated the ability of ECRs to respond *in vitro* to over-expression of transcription factors for which their sequence contains conserved binding sites. For example, Paredes et al. demonstrated that overexpression of SP1 significantly repressed

enhancer activity of the DRD4 ECR in primary cultured rat cortex neurons (Paredes et al. 2011).

1.2.2 Repetitive DNA

Repetitive DNA can refer either to sections of DNA that are tandemly repeated, or to transposable elements that have been copied and inserted across the genome giving rise to many duplications of the same sequence in distinct locations. This section will review Variable Number Tandem Repeats (VNTRs), which fall into the former classification of typically short sequences which are repeated in tandem, giving rise to repetitive structure in the genome. Following this, an overview of transposable elements will give an introduction to the subclasses that remain active in the human genome, Long Interspersed Nuclear Elements (LINEs), Short Interspersed Nuclear Elements (SINEs), and SINE-VNTR-Alus (SVAs). LINEs and SINEs are not typically repetitive in their structure of individual elements, but their duplication across the genome has led to them being characterised as repetitive DNA. On the other hand, SVAs, which are primate-specific composite elements, both contain repetitive structures within their sequence and have the capacity to spread across the genome through repeated rounds of copying and insertion.

1.2.2.1 Variable Number Tandem Repeats (VNTRs)

Tandem Repeats (TRs) are short sequences that are repeated in tandem and can be variable in copy number in the population. Such variation can be generated through recombination events, or by strand slippage during replication, causing a mispairing between the template strand and the nascent strand which would result in a deletion or duplication (Gemayel et al. 2012).

TRs can be referred to using a number of terms, including variable number tandem repeats (VNTRs), simple sequence repeats (SSRs), or micro- or mini-satellites depending on their repeat size. Microsatellites are typically defined as a repeating unit of 1-6 bp, whereas minisatellites refer to repeated sequences typically in the range of 6-500 bp (Breen et al. 2008). However, in this thesis, the term 'TR' will be used to describe tandem repeats of multiple sizes. Where these elements are polymorphic for repeat number in the population, they will be referred to as 'VNTRs'.

TRs are found in both coding and non-coding regions, with approximately 17% of human genes containing TR polymorphism in their coding sequence (Gemayel et al. 2012). While primate non-coding regions are enriched for all sizes of microsatellite with repeating units between 1-6 bp, Metzgar et al. demonstrated that tri- and hexanucleotide repeats were consistently over-represented in coding regions (Metzgar, Bytof and Wills 2000). This would suggest that TRs in coding regions are under strong evolutionary selection to prevent repeat units that would result in frame shift mutations. Interestingly, coding TRs in the human genome have been shown to be enriched in genes with particular functions, including genes involved in transcriptional regulation, embryonic morphogenesis, and brain development (Legendre et al. 2007).

TRs are known to have contributed to evolution, both through their ability to alter gene structure, and through their ability to alter gene regulation, often in a copy number-dependent manner (Bilgin Sonay et al. 2015, Gemayel et al. 2010, Jansen, Gemayel and Verstrepen 2012). This could lead to species-specific gene regulation, or variation in gene expression across populations, which would be based on genotype.

Quilez et al. demonstrated that approximately 31% of human genes had a TR within 1 kb either side of their transcriptional start site, with this proximity suggesting the ability of the repeated element to regulate expression from the promoter region (Quilez et al. 2016). In this study, genes with highly variable promoter VNTRs were shown to display significantly higher variability in expression and methylation between individuals, likely directed by differences in VNTR genotype. Over 90% of tandem repeats in this study that were found to be expressed quantitative trait loci (eQTLs) were also found to significantly correlate with changes in local methylation levels. This would be in agreement with work by Sawaya et al., which demonstrated that microsatellites are statistically associated with promoters. While AT-rich repeats make up 36.4% of microsatellites motifs in the human genome, such elements display a strong negative correlation with promoters. Instead, promoter-associated microsatellites typically have a high GC content and are associated with CpG islands (Sawaya et al. 2013).

TRs therefore have the capacity to influence gene expression through a number of mechanisms, including transcription factor binding, DNA methylation, or the formation of secondary structure (Figure 1.2). Simplistically, if a VNTR contains a transcription factor binding site or many CpG dinucleotides, then it would follow that increased copy number of the VNTR would result in either (a) additional binding sites and an increased likelihood of binding transcription factors that could regulate nearby transcripts, or (b) a larger region for activating or repressive methylation and histone modifications which would more strongly influence local chromatin structure. In addition to this, repetitive DNA has the capacity to form alternative DNA structures, which may influence gene expression when situated near to transcriptional start sites.

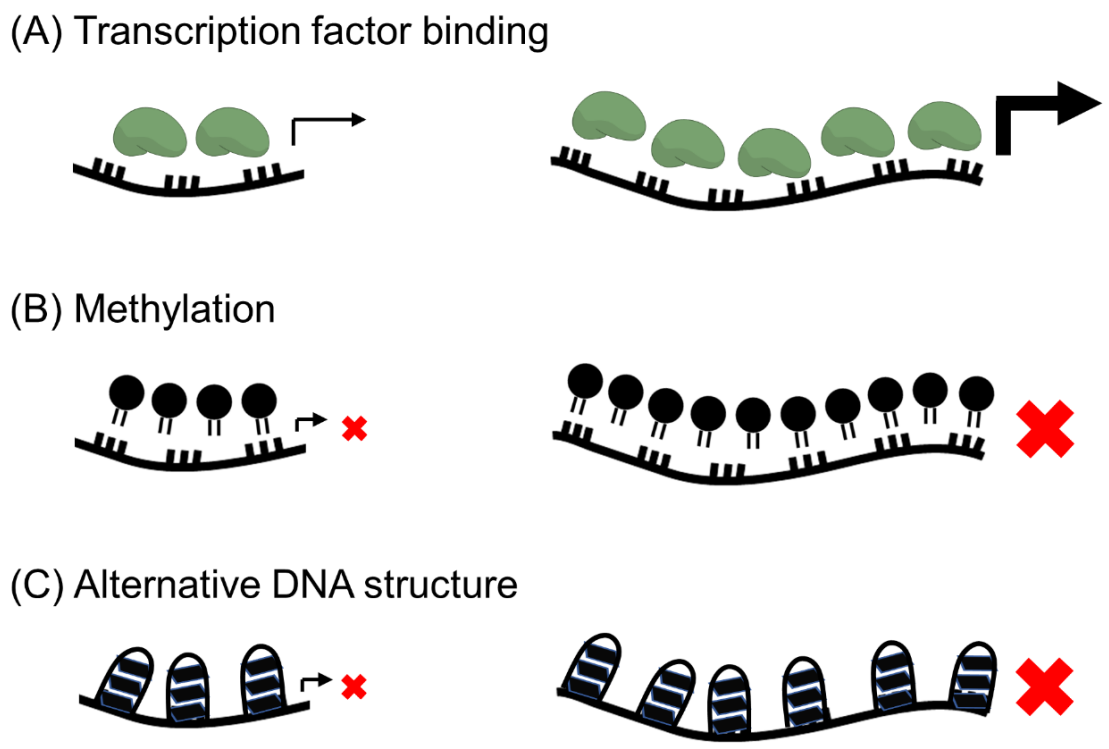


Figure 1.2 Mechanisms of VNTR-mediated gene regulation.

Repetitive regions of DNA such as VNTRs are often found around promoters, with many such promoter VNTRs containing high GC content and transcription factor binding sites. At this location, they may regulate expression through binding transcription factors, or through their methylation or histone modification status, which could alter local chromatin structure making the region more or less accessible for interactions with DNA or proteins. Repetitive regions with a high G content are also thought to be capable of forming alternative DNA structures such as G-quadruplex, which has been demonstrated to alter the regulation of nearby genes. The left-hand column shows the hypothetical regulatory ability of a three-copy repeat VNTR versus a six-copy repeat in the right-hand column. Typically, the more copies of a repeat, the more capacity this region would have to bind transcription factors, form a greater number of copies of alternative DNA structure, or become more heavily epigenetically modified. For this reason, different alleles of VNTR copy number can result in allele-specific gene regulation.

Of particular interest to GC-rich promoter repeats is the G-quadruplex (G4) structure (Kejnovsky, Tokan and Lexa 2015). G4 structures form in regions containing multiple short runs of guanine (G) bases, which form planar tetrad structures through Hoogsteen hydrogen bonding and stack together to form helices (Rhodes and Lipps 2015). Over 40% of human genes contain one or more potential G4 sequences in their promoter region (Huppert and Balasubramanian 2007), and the ability of such sequences to influence gene expression has been demonstrated in numerous studies, showing that mutation or stabilisation of sequences with the potential to form G4 structure results in modified gene expression both *in vitro* and *in vivo* (David et al. 2016, Gu et al. 2012, Shin et al. 2015, Wang et al. 2010).

Many of the most well-known coding VNTRs reside within genes involved in central nervous system (CNS) processes and are known to be implicated in neurodegenerative conditions, such as the $(CAG)_n$ trinucleotide repeat expansion in the HTT gene associated with Huntington's Disease (Reiner, Dragatsis and Dietrich 2011), and the $(GGGGCC)_n$ hexanucleotide repeat expansion in C9orf72 which is implicated in frontotemporal dementia (FTD) and amyotrophic lateral sclerosis (ALS) (Rohrer et al. 2015). This would be in line with work by Legendre et al., which demonstrated that coding TRs were over-represented in genes involved in brain development and neurogenesis (Legendre et al. 2007). Similarly Sawaya et al. demonstrated that non-coding, promoter microsatellites were over-represented at genes involved in neurotransmitter secretion, synaptic transmission, and synaptic plasticity (Sawaya et al. 2013).

A number of non-coding VNTRs have been identified at key behavioural and mental health-related genes, such as the schizophrenia-associated miRNA, MIR137 (Warburton et al. 2015a), monoamine oxidase A (MAOA) (Sabol, Hu and Hamer

1998), dopamine receptor D4 (DRD4) (Paredes, Quinn and D'Souza 2013), dopamine transporter 1 (DAT1) (Guindalini et al. 2006), and the serotonin transporter gene, solute carrier family 6 member 4 (SLC6A4) (Vasiliou et al. 2012). Research has demonstrated the ability of such VNTRs to drive allele-specific expression of CNS genes based on copy number of the repeat both *in vitro* and *in vivo*, with likely implications in the risk for psychiatric conditions (Fiskerstrand, Lovejoy and Quinn 1999, MacKenzie and Quinn 1999, Paredes et al. 2013). For example, Warburton et al. demonstrated that an internal promoter at the schizophrenia-associated MIR137 locus contained a VNTR which supported reporter gene expression in a cell line model (Warburton et al. 2015b). The promoter containing 12 copies of the VNTR supported significantly higher expression than the promoter region containing a 4-copy variation. Similarly, regulation from the MIR137 internal promoter (imir137) in response to over-expression of the transcription factor, REST, was dependent on VNTR copy number. This would provide evidence to suggest that altered copy number of a VNTR can result in differential regulation as directed by transcription factor binding over the repeating unit, with this genotype-dependent effect potentially playing a role in MIR137 regulation in schizophrenia.

1.2.2.2 An overview of transposable elements (TEs)

Transposable elements have been identified in all eukaryotic genomes sequenced to date, and comprise up to half of the human genome (Lander et al. 2001). Although transposable elements were believed for many years to be “selfish” or “junk” DNA, having frequently been referred to as “parasites” in the genome, research supports a role for transposable elements in the process of genomic diversity and evolution (Goodier and Kazazian 2008, Cordaux and Batzer 2009, Erwin, Marchetto and Gage 2014, Vasieva et al. 2016a), and they are thought to play important roles in the

regulation of chromatin structure and gene expression (Elbarbary, Lucas and Maquat 2016).

Transposable elements are divided into a number of sub-categories, with the first division based on their method of movement across the genome (Wicker et al. 2007). Class I transposable elements, known as retrotransposons, move through a “copy-and-paste” mechanism. This involves transcription of the element into an RNA intermediate, which is then reverse transcribed back into DNA and inserted at a new location in the genome (Figure 1.3).

This method of movement through an RNA intermediate allows the creation and insertion of additional copies of the element from the original element, allowing retrotransposons to spread across the genome. On the other hand, class II elements, known as DNA transposons, move through a “cut-and-paste” mechanism. This involves specific transposase enzymes which excise the transposon from the DNA. DNA transposons can be further sub-divided based on the enzymes used to mobilise the element (Wicker et al. 2007).

This section will focus on non-Long Terminal Repeat (LTR) retrotransposons, which include LINE and SVA sequences. This subclass of retrotransposons also includes SINE and Alu elements, however, these elements are not discussed in detail in this thesis. Non-LTR retrotransposons are known to be the most abundant class of mobile element in the human genome.

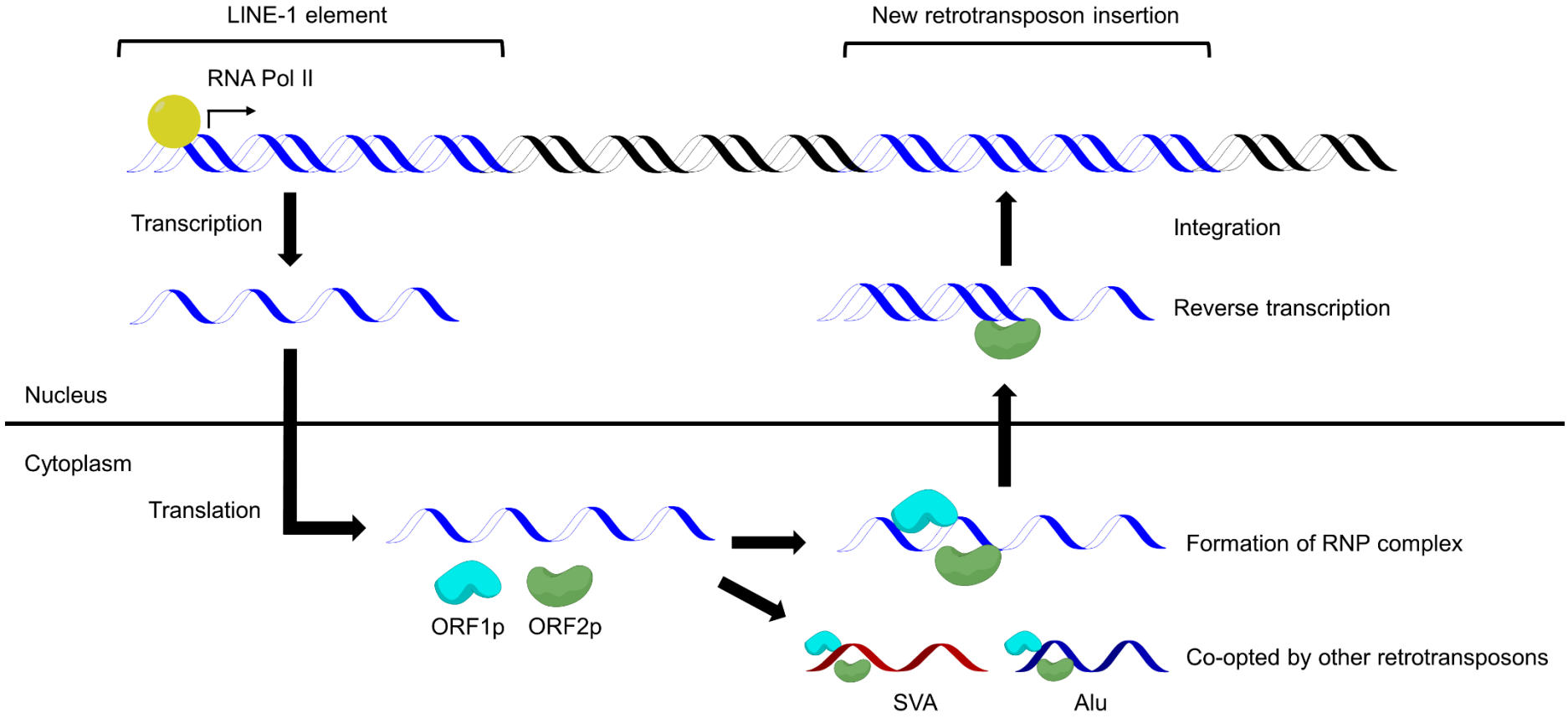


Figure 1.3 Retrotransposon mobilisation and integration.

In order for retrotransposons to mobilise, LINE-1 elements must first be transcribed by RNA Polymerase II. From this transcript, ORF1 and ORF2 can be translated into proteins. ORF1p is an RNA binding protein with chaperone activity, and ORF2p is a reverse transcriptase and endonuclease. Both proteins are required for mobilisation through the formation of a ribonucleoprotein complex (RNP). While LINE-1 encoded proteins preferentially form an RNP with LINE-1 transcripts, they can also be co-opted by other retrotransposons such as SVAs and Alus, which are non-autonomous and therefore dependent one LINE-1 machinery to mobilise. After formation of the RNP, ORF2p nicks one strand of the host DNA and uses this to prime the reverse transcription of the LINE-1 RNA to form a double stranded element which is integrated back into the host genome at the nicked site.

While the majority of transposable elements have become inactive throughout the course of evolution, non-LTR retrotransposons are the only group that are known to remain mobile and active within the human lineage (Mills et al. 2007), with their movement across the genome estimated to result in 0.27% of human diseases (Callinan and Batzer 2006, Hancks and Kazazian 2016), as well as contributing to inter-individual genetic variation in the case of new germline or somatic insertions (Richardson, Morell and Faulkner 2014, Stewart et al. 2011). However, regardless of their mobilisation ability, retrotransposons are able to function as transcriptional regulatory elements (Savage et al. 2013b, Savage et al. 2014, Payton et al. 2016), and it is in this capacity that they are predominantly considered in this thesis.

1.2.2.2.1 Long Interspersed Nuclear Elements (LINEs):

Multiple subfamilies of LINEs have evolved and co-existed over time, with most being distinguished by differing characteristics of their 5' UTR, thought to be activated by different proteins so as not to outcompete each other. Although there are three recognisable LINE families in the human genome (LINE-1, 2, and 3), over the last ~40 million years (Myrs), a single class of LINEs has dominated, and LINE-1 elements are now the only remaining family of LINEs that are active in the human genome (Lander et al. 2001, Khan, Smit and Boissinot 2006). As such, LINE-1s are the only LINE elements considered in this thesis. LINE-1 elements account for approximately 17% of the human genome, comprising an estimated ~500,000 copies (Lander et al. 2001). The canonical LINE-1 element is roughly 6 kb in length and comprises a 5' UTR with internal RNA Pol II site (Swergold 1990), followed by 2 open reading frames, ORF1 and ORF2, a 3' UTR and a poly(A) tail (Figure 1.4a).

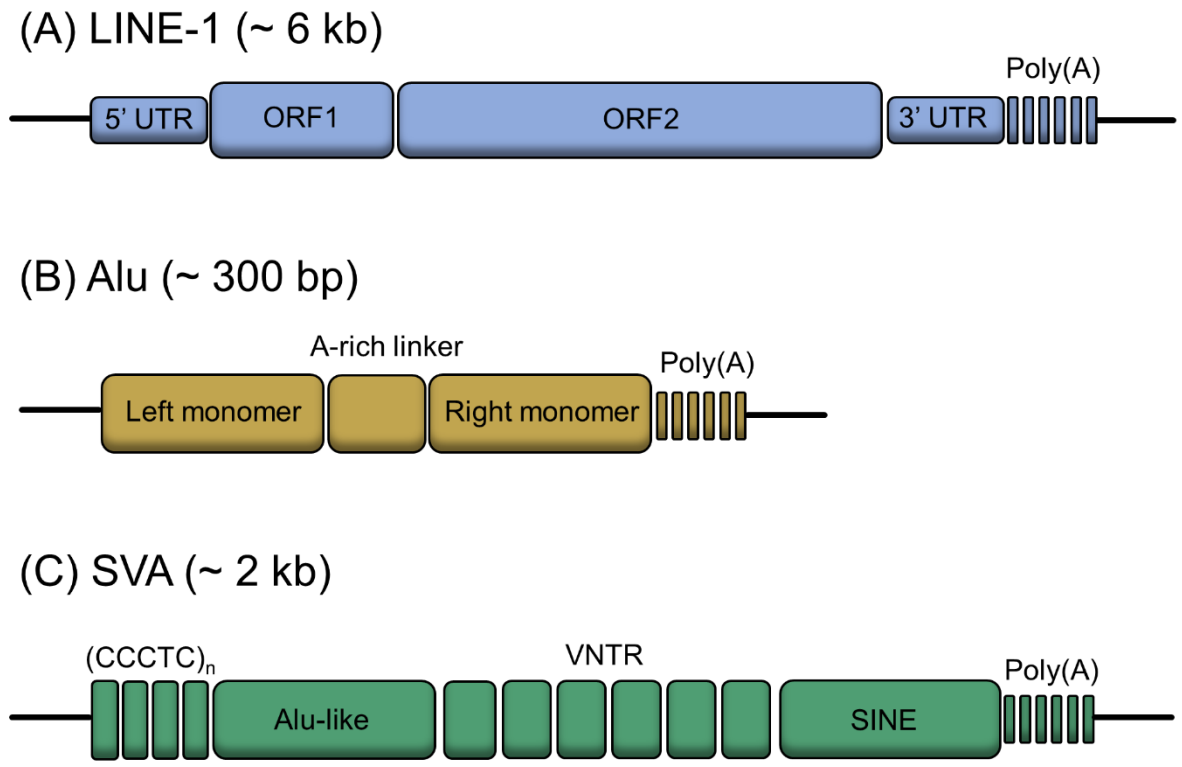


Figure 1.4 Canonical structures of LINE-1, Alu, and SVA elements.

(a) A canonical LINE-1 element is approximately 6 kb in length, comprising a 5' UTR, two open reading frames (ORF1 and ORF2), and a 3' UTR, followed by a poly(A) tail. ORF1 encodes an RNA binding protein with chaperone activity, while ORF2 encodes a reverse transcriptase and endonuclease. Both are required for mobilisation and can be co-opted by non-autonomous retrotransposons.

(b) Alu elements are approximately 300 bp in size, consisting primarily of two monomers which are derived from the signal recognition particle RNA (7SL RNA) sequence. The left and right monomers are separated by an A-rich linker with the sequence $(A)_5TAC(A)_6$. Alu elements also contain a poly(A) signal at their 3' end.

(c) Canonical SVAs are approximately 2 kb in length, though may be polymorphic in size due to variation within their repetitive domains. Typically, SVAs have a $(CCCTCT)_n$ hexamer repeat at the 5' end, except for the F1 subfamily which lacks this region. Following the hexamer repeat is an Alu-like domain, a VNTR, a SINE-derived region, and a poly(A) tail.

ORF1 encodes a ~40 kDa RNA binding protein with nucleic acid chaperone activity (Martin and Bushman 2001), and ORF2 encodes a ~150 kDa protein with both endonuclease and reverse transcriptase activity (Feng et al. 1996, Mathias et al. 1991). Due to their ability to express their own mobilisation machinery, LINE-1s are the only mobile, autonomous retrotransposon in modern humans. However, the large majority of LINE-1 elements have been truncated, therefore having lost the ability to mobilise further, with only ~90 'hot L1' elements remaining actively mobile in the human genome (Brouha et al. 2003). While the ORF1 and 2 proteins (ORF1p and ORF2p) have been shown to preferentially bind their own LINE-1 mRNA sequence (Wei et al. 2001), they are also co-opted by non-autonomous transposable elements such as SVAs in order to mobilise in *trans* (Raiz et al. 2012).

LINE-1 elements are not distributed randomly across the genome are known to preferentially cluster on the X chromosome, accounting for approximately one third of the chromosome's size (Ross et al. 2005). LINE-1 elements are believed to play a role in the propagation and patterning of X chromosome inactivation (Chow et al. 2010). This was supported by work elucidating the regulatory consequences of X;autosome translocations, in which autosomal genes were shown to be inactivated when a translocation event moved them to an inactive X chromosome (Bala Tannan et al. 2014). LINE-1 elements were found to be enriched around these inactivated autosomal genes, as well as enriched around inactive X chromosome genes compared to those which escape X inactivation (Bala Tannan et al. 2014). This would be suggestive of a role for LINE-1 elements in the spreading of X chromosome inactivation.

Many other groups have demonstrated the ability of LINE-1 elements to modulate gene expression through a number of mechanisms, which can contribute to both

evolution and to disease states or disease risk. For example, *in vitro* studies into somatic retrotransposon insertions in iPSCs demonstrated that a LINE-1 insertion into the intron of the CADPS2 gene resulted in significantly altered CADPS2 expression (Klawitter et al. 2015). Further, studies in cancer have demonstrated that hypomethylation of LINE-1 elements in the introns of some oncogenes, particularly the MET gene, activate gene expression and the expression of alternative transcripts in cancerous cells (Wolff et al. 2010, Hur et al. 2014, Zhu et al. 2014, Zhu et al. 2015). Similarly, LINE-1 elements are able to splice into other genes, with some elements further having the ability to act as promoters. This could result in disrupted gene expression or the creation of alternative transcripts, with LINE-1 promoter activity driving the expression of novel and potentially species-specific transcripts from its bi-directional promoter (Belancio, Hedges and Deininger 2006, Criscione et al. 2016). Increased LINE-1 expression has also been demonstrated in the brains of individuals with schizophrenia and autism (Bundo et al. 2014, Shpyleva et al. 2017). Finally, recently evolved subfamilies of LINE-1 have G-rich sequences at their 3' end with the potential to form G4 structure. This may further contribute to the ability of LINE-1 elements to regulate the expression of nearby genes (Sahakyan et al. 2017).

1.2.2.2.2 Alu elements:

Alu elements are a primate-specific member of the SINE subfamily of retrotransposons. After expanding rapidly in the primate lineage, there are now approximately 1.1 million copies in the human genome, comprising around 10% of the total mass of the genome, and making them the most abundant of all transposable elements (Wang, 2014). Their vast number makes them difficult to study, and thus the Alu class of elements is not considered in this thesis, except as a component in the composite SVA retrotransposon subfamily.

Individual Alu elements are approximately 300 bp in length and primarily comprised of two dimers which resulted from a fusion between two ancient derivatives of the 7SL RNA (Ullu and Tschudi 1984, Mighell, Markham and Robinson 1997). The two monomers are separated by a short A-rich sequence, with another A-rich sequence at the 3' end, followed by flanking direct repeat sequences and a poly(A) tail (Figure 1.4b). Despite their large number in the genome, the majority of Alu elements have been rendered immobile due to mutation, leaving approximately 1836 highly active 'hot' Alu elements (Bennett et al. 2008). As with other retrotransposons, Alu elements are known to act as regulators of gene expression (Payton et al. 2016).

1.2.2.2.3 SINE-VNTR-Alus (SVAs):

SVAs are the most recently evolved family of active non-LTR retrotransposable elements, being present only in hominid genomes. There are approximately 2700 annotated SVA copies in the human reference genome, which can be divided into seven subfamilies (SVA A-F1; Figure 1.5) based on their SINE region and evolutionary age (Lander et al. 2001, Savage et al. 2013b). The older SVA subfamilies A-C are present in primate species between and including chimpanzees and humans, while more recent subfamilies E-F1 are present only in humans. The SVA D subfamily is approximately 9.6 Myrs old and is the largest SVA subfamily, accounting for over 40% of the total number of SVA elements. SVA Ds are also the only SVA subfamily that has some copies present only in humans, while other copies may be present in species from chimpanzees to humans (Wang et al. 2005). While SVAs remain mobile in the human lineage, they are dependent on LINE-encoded proteins in order to mobilise (Raiz et al. 2012).

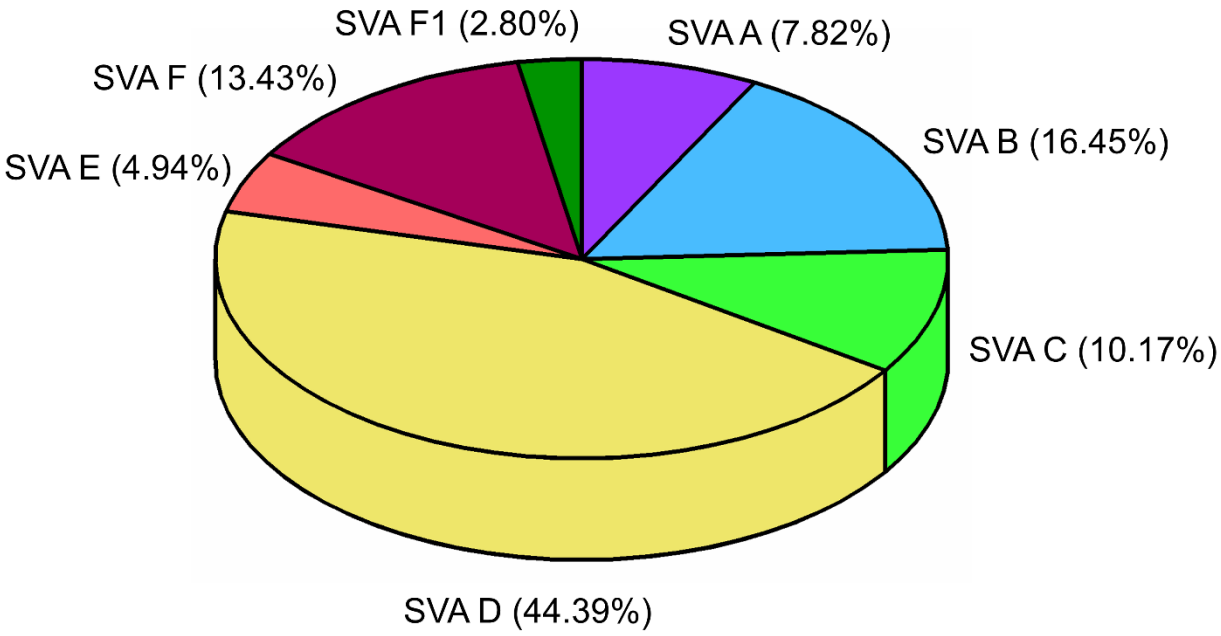


Figure 1.5 Breakdown of SVA subtypes present in the human genome.

Of the 2676 SVAs annotated on the hg19 human genome build, the largest subtype is SVA D, which comprises 44% of the total number of SVAs. SVA B, C, and F each comprise over 10% of the total, while SVA As make up 8% and SVAs E and F1 comprise 5 and 3%, respectively.

A canonical SVA is comprised of 5 main components, beginning with (1) a simple hexamer repeat of $(CCCTCT)_n$ at the 5' end, followed by (2) an Alu- like region made up of 2 antisense Alu fragments separated by a region of intervening sequence, (3) a variable number tandem repeat (VNTR), (4) a SINE region derived from the 3' LTR of the retroviral HERV-K10 element, and finally (5) a 3' poly(A) signal (Wang et al. 2005) (Figure 1.4c). A typical full length SVA element is approximately 2 kb in size, however, due to the young evolutionary age and the repetitive nature of the SVA structure, many SVA elements are polymorphic in the human population. This polymorphism typically occurs within the CCCTCT hexamer repeat, the VNTR, and sometimes in the poly(A) region (Savage et al. 2013b). In the case of the seventh SVA family, known as F1, this region lacks the 5′ CCCTCT hexamer repeat, instead containing a 5′ transduction of exon 1 of the MAST2 gene (Bantysh and Buzdin 2009).

SVA elements are preferentially found in regions of medium to high GC content, with evolutionarily younger SVA species preferentially inserting into higher GC regions than older subfamilies, with families E and F peaking in regions of 48-52% GC compared to older subfamilies that were found to be enriched in regions of 40-42% GC (Lander et al. 2001, Wang et al. 2005). In addition to this, SVAs themselves are generally around 60% GC, with this potentially even exceeding 70% within the VNTR region (Wang et al. 2005). Since the definition of a CpG island is a sequence over 200 bp with a GC content exceeding 50% and an observed-to-expected CpG ratio over 60%, SVAs may be simplistically thought of as mobile CpG islands (Strichman-Almashanu et al. 2002). For this reason, the movement of such elements may have the potential to create novel CpG islands which could impact architecture and function at the surrounding genomic region.

SVAs are non-randomly distributed throughout the genome, with chromosomes 1, 17, 19 and 22 containing a higher proportion of SVA sequences than would be predicted based on their size (Wang et al. 2005). Work by others has demonstrated that SVAs are over-represented at genic regions (Savage et al. 2013b), particularly around genes involved in CNS-related processes, and those that have been linked to Parkinson's Disease (Vasieva et al. 2016a, Savage 2013). Given their high GC content and general proximity to genes, SVAs are likely to have the potential to positively or negatively regulate expression through mechanisms such as the recruiting transcription factors or altering local chromatin structure based on epigenetic marks across the element. Indeed, SVA elements at the PARK7 and FUS gene loci have been shown to modulate expression of a reporter gene both *in vitro* in cell line models, and *in vivo* in chick embryo models (Savage et al. 2013b, Savage et al. 2014).

One mechanisms through which SVAs may regulate gene expression based on their GC content is through the formation of alternative DNA structures which have the potential to affect transcription, such as G4 structures. Previous work has demonstrated that, while SVAs account for only 0.13% of the total genome, they make up nearly 2% of DNA that is predicted to form G4 structures (Savage et al. 2013b). In particular, the 5′ CCCTCT hexamer repeat in SVAs is likely to be the most amenable to G4 formation. Further, as the evolutionary age of the SVA subtypes decreases, the percentage of sequence capable of quadruplex formation was found to increase, giving the human-specific SVA E-F1 subtypes the most potential to form transcriptional modulatory G4 structures. This is due in part to the increased copy number of the CCCTCT hexamer repeat in younger SVA subtypes, and also to the increase in VNTR potential to form quadruplex, as younger subtypes contain two VNTRs with a higher

GC content, as opposed to one with lower GC content in the older SVAs (Savage et al. 2013b).

As described for LINE-1 elements, Kim et al. have further demonstrated the capacity of SVAs to act as novel promoters (Kim and Hahn 2010, Kim and Hahn 2011). Using bioinformatic analysis to identify human-specific insertions that were expressed as exons, this work identified 12 cases in which human-specific SVA insertions led to the formation of novel promoters, driving expression of human-specific transcripts originating from within the SVA. For example, an SVA insertion 5′ to the TBPL2 gene has been shown to act as a novel promoter for a human-specific transcript of TBPL2, in which SVA derived sequence is included as a novel first exon (Kim and Hahn 2011). This work also demonstrated the ability of SVA insertions to drive expression of novel antisense transcripts, including a human-specific SVA insertion at intron 45 of the SYNE2 gene, which acts as a novel promoter for the expression of six antisense ESTs (Kim and Hahn 2010).

This section has provided an overview of genomic elements that have the capacity to modulate gene expression and epigenetic regulation, ranging from the most highly conserved elements in the genome, to recent and polymorphic repetitive sequences. These elements have been demonstrated to support allele-specific regulatory activity, and their transcriptional effects have been shown to be stimulus-inducible. It is therefore likely that such elements play a role in controlling gene expression in response to external stimuli such as stress. The following section will provide a basic overview of this type of gene-environment interaction and a brief introduction the literature on stress as a risk factor for schizophrenia.

1.3 Gene x Environment interactions and stress as a risk factor for schizophrenia

The above sections reviewed different types of transcriptional regulatory elements, from the most highly conserved regions of the genome, ECRs, to the most recent and variable evolutionary changes such as polymorphic VNTRs or new retrotransposon insertions. In demonstrating how these elements can regulate gene expression, it would also follow that genetic variation within these elements could alter their regulatory potential.

Transcriptional regulation can be controlled through DNA methylation and histone modifications that change chromatin conformation, and through the action of transcription factors, which bind to regulatory elements and function either to block or recruit the necessary factors for gene expression. DNA is typically found in a condensed complex of DNA, RNA, and protein known as chromatin. Within this structure, DNA is packaged into basic units known as nucleosomes, which consist of approximately 147 bp of DNA wound around a histone octomer core (Figure 1.6). Each histone octomer is made up of two of each of the four histone proteins, H2A, H2B, H3, and H4. This condensed structure allows tight transcriptional control within the cell, with regions of more loosely packed chromatin (euchromatin) at loci undergoing active expression, and regions of tightly packed chromatin (heterochromatin) at regions encompassing silenced genes.

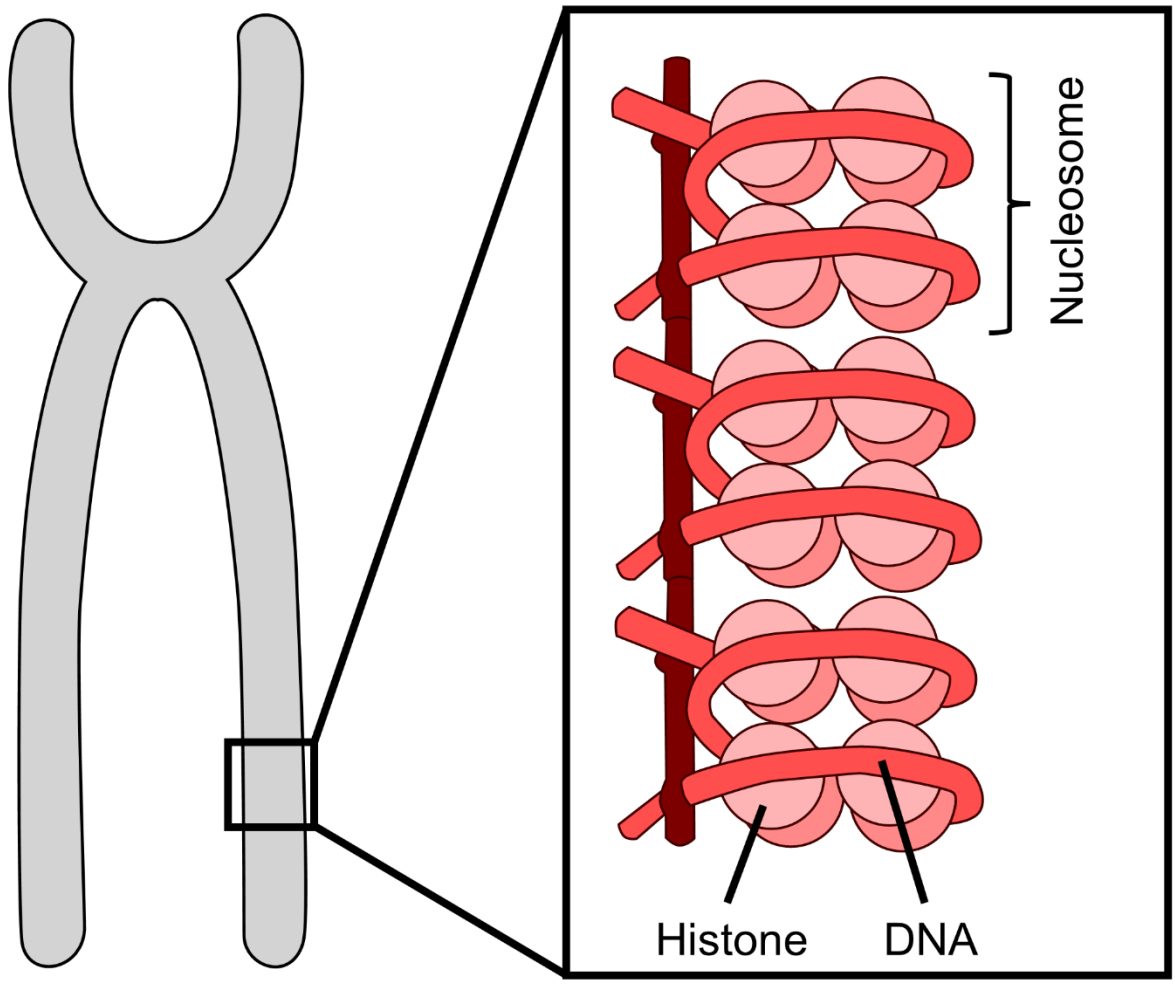


Figure 1.6 Chromatin organisation and nucleosome structure

DNA is bound and packaged into chromatin, which can be tightly packed (heterochromatin) around silenced genes, or open (euchromatin) around active genes. Chromatin is comprised of repeating units called nucleosomes. One nucleosome contains a histone octamer core, around which 147 bp of DNA is wrapped. Each histone octamer contains two copies of the H2A, H2B, H3, and H4 histone proteins.

In order to alter this structure to regulate gene expression, epigenetic modifications in the form of methylation or acetylation are added to histone proteins within each nucleosome core, and the chromatin is then remodelled. Simplistically, this loosening or tightening of chromatin structure via histone modification and chromatin remodelling can regulate expression by making genes more or less accessible to transcription factors and polymerases.

Processes such as neuronal development and adult neurogenesis rely on epigenetic interactions, and disruption of these processes has been implicated in CNS conditions. Epigenetic mechanisms such as DNA methylation and histone modifications are dynamic processes that allow the cell to respond in a transient manner to its environment, and thus they are typically at the interface of so called 'gene x environment' interactions (G x E).

G x E typically refers to a process in which an environmental stimulus alters the expression of a gene, and is often used to describe a situation in which the same environmental stimulus modifies gene expression differently based on variation in the genetic element through which the transcriptional response is mediated. Such a process may result in distinctly different phenotypic outcomes in response to the same environmental change, which would be influenced by an individual's genotype. This is the basic principle underlying the stress-vulnerability model of mental illness (Nuechterlein and Dawson 1984, Zubin and Spring 1977), when considering genetic variation as the 'vulnerability' aspect (Figure 1.7).

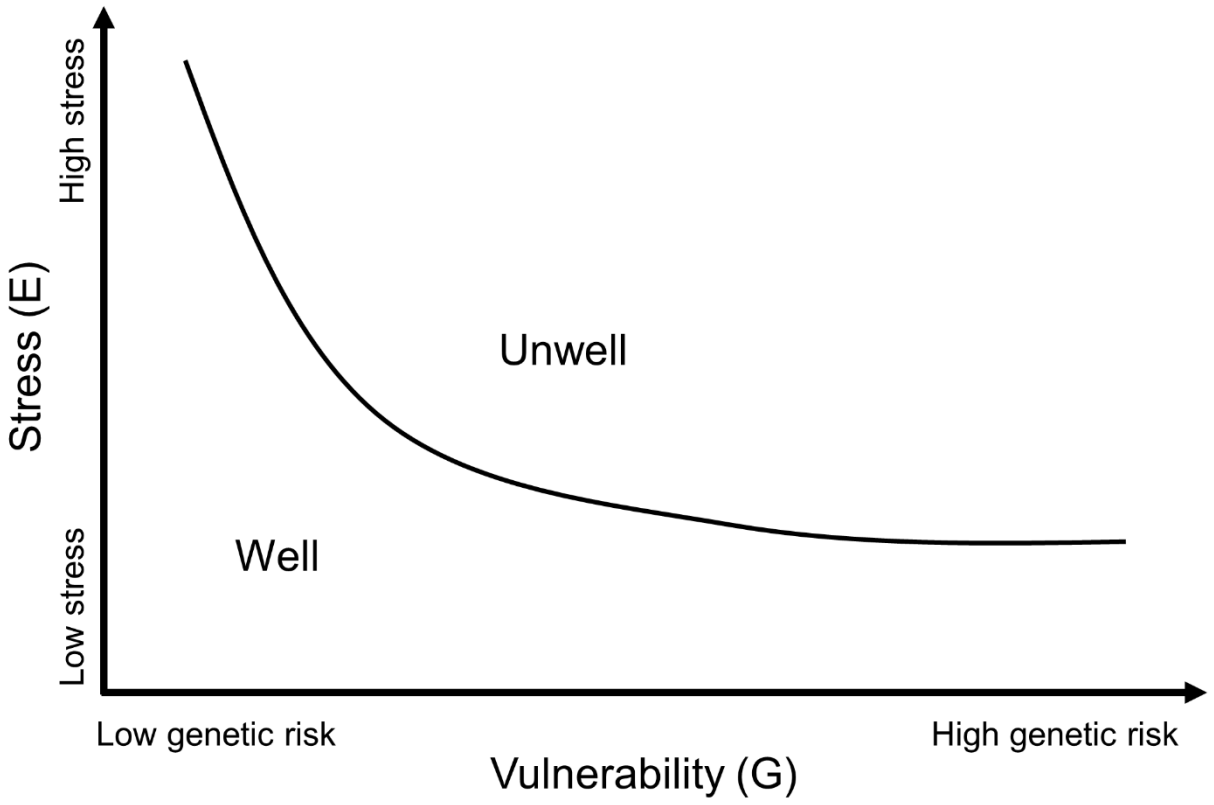


Figure 1.7 The stress-vulnerability model of schizophrenia.

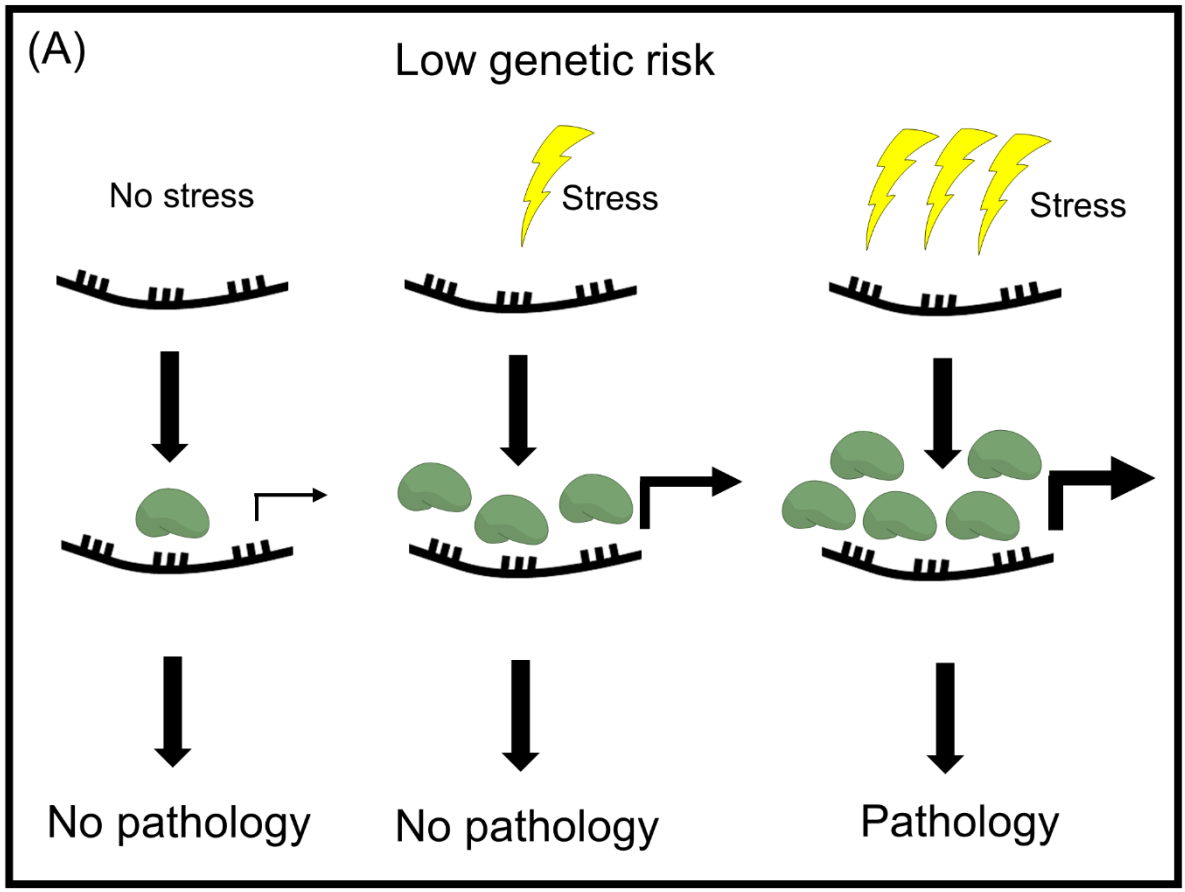
The stress-vulnerability model posits a gene x environment interaction in the risk for schizophrenia, with individuals at a genetic high risk for the condition being more likely to develop schizophrenia after experiencing a moderate amount of stress. Individuals with low genetic risk may still experience schizophrenia, though likely in response to significantly higher levels of stress. Overall, the stress-vulnerability model suggests that an individual's ability to remain psychologically well during or following periods of stress is in part influenced by their genotype. 'Stress' on the Y axis can refer to numerous experiences at multiple developmental time points, including maternal stress or infection, childhood adversity, neglect, or trauma, and experiences of stress or adversity in adulthood including poverty, abuse, the loss of a loved one, the loss of a job, and numerous other factors that may be classed as traumatic or resulting in chronic stress.

‘Schizophrenia’ is a diagnostic term used to describe diverse combinations of emotional, psychological, and physical experiences, ranging from differences in perceptual experience, such as hearing voices, to changes in social functioning and cognitive processes. These experiences exist on a continuum in the general population, occurring both in individuals that would typically be considered psychologically well, and those that would be considered psychologically unwell (Johns et al. 2014, Nuevo et al. 2012, de Leede-Smith and Barkus 2013).

The present criteria for a diagnosis of schizophrenia is defined in the Diagnostic and Statistical Manual of Mental Disorders IV, and includes two or more experiences such as hallucinations, delusions, disorganised or catatonic behaviour, and/or negative symptoms such as a lack of emotion or motivation. In addition, the diagnostic criteria specifies a decline in the individual’s ability to function socially or in their work environment, with these experiences having affected the individual for a duration of at least six months (American Psychiatric Association 2013).

The stress-vulnerability model for schizophrenia posits that every individual has some amount of genetic risk for experiencing schizophrenia, but that only people who experience specific amounts of stress, combined with genetic risk, are likely to become unwell. In this model, individuals with low genetic risk who experience extreme amounts of stress are at risk of developing schizophrenia, as are individuals with high genetic risk who experience more moderate amounts of stress (Figure 1.8). The model would suggest that the lower an individual’s genetic risk for schizophrenia, the higher amount of stress could be tolerated before they experienced distressing symptoms.

Figure 1.8 The influence of genetic variation on molecular response to stress.



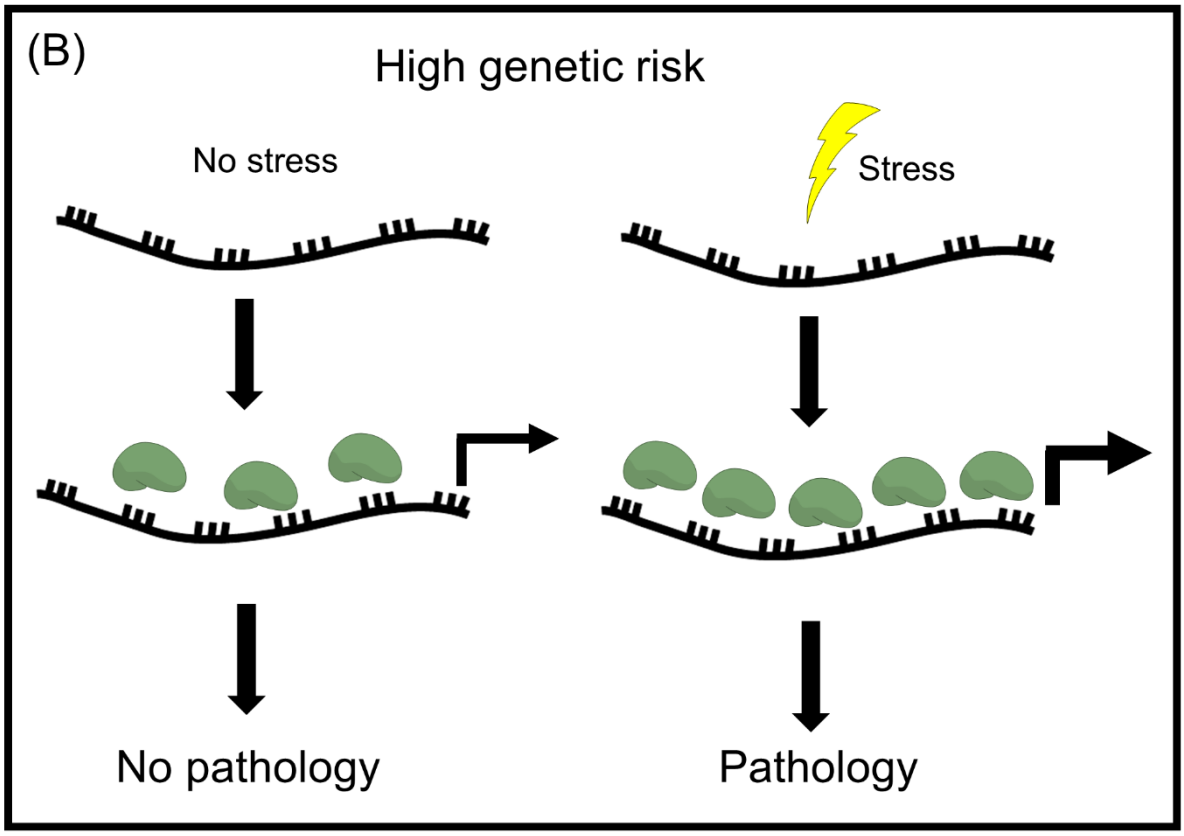


Figure 1.8 The influence of genetic variation on molecular response to stress.

This hypothetical example uses a VNTR as a genetic variant which may differentially influence two individuals' molecular responses to the same stress based on their genotype. In (a) the low genetic risk individual has a three-copy repeat of the VNTR which is adjacent to a gene promoter and typically binds one unit of a transcription factor under normal conditions to maintain the correct expression of the adjacent gene. Under moderate stress conditions, increased expression of the transcription factor results in higher expression of the gene in question through the increased activity of the adjacent VNTR, though this is not enough to result in pathology. However, under conditions of extreme stress, a significant overabundance of the transcription factor may lead to constant activity at the VNTR, which could result in high expression of the adjacent gene and contribute to a state of illness. On the other hand, the high risk individual in (b) has a six-copy VNTR which is capable of binding more copies of the transcription factor. They therefore have increased expression of the adjacent gene under normal conditions compared to the low risk individual, though this is not problematic until person (b) experiences a moderate amount of stress. Under stress conditions, the six-copy VNTR can bind a larger number of transcription factor units than the three-copy, resulting in significantly increased gene expression which may contribute to schizophrenia-related pathology.

Work by the Schizophrenia Working Group of the Psychiatric Genetics Consortium compiled schizophrenia GWAS data from multiple studies to correlate schizophrenia risk decile based on genetics with an odds ratio (OR) for schizophrenia diagnosis. An OR represents the association between an exposure and an outcome, and is a measure of the likelihood that the outcome will occur following the exposure compared to the likelihood that the outcome would occur without exposure. In this study, the tenth decile represents the highest genetic risk profile, and was found to be associated with an OR of 8-20 for schizophrenia, varying based on the sample set used for the GWAS analysis (Schizophrenia Working Group of the Psychiatric Genomics Consortium 2014). This would provide clear evidence that individuals with higher numbers of schizophrenia-associated SNPs have increased odds of developing the condition.

Stress reactivity is known to be correlated with likelihood of experiencing psychosis, which would classically be associated with schizophrenia (DeVylder et al. 2016). An individuals' stress response may be genetically influenced, but is also known to be modified by previous experience in which early life stress can sensitise an individual to future stress (Cristobal-Narvaez et al. 2016, Gorka et al. 2014).

In general, psychologically healthy individuals who report psychosis-like experiences have been found to have higher incidences of childhood trauma (Krakvik et al. 2015, Daalman et al. 2012, Shevlin et al. 2011).

Early experiences of trauma and adversity are well known risk factors for mental illness, with a World Health Organisation (WHO) study of almost 52,000 individuals across 21 countries demonstrating that childhood adversity accounted for 30% of all adult mental health conditions (Kessler et al. 2010). Similarly, a meta-analysis of studies assessing the link between childhood adversity and psychosis demonstrated

that 78% of studies tested showed a positive association between childhood adversity and psychosis, and that 33% of the population risk for psychosis could be attributed to experiences of childhood adversity (Varese et al. 2012). Psychosis is often associated with a diagnosis of schizophrenia, however, psychosis as a symptom can also fall within additional diagnostic categories such as other schizophrenia-spectrum or delusional disorders, as well as bipolar disorder and depression with psychotic features. “Mental illness” as described in the WHO study by Kessler et al. refers to 20 different diagnosable conditions according to the DSM IV criteria, covering multiple diagnoses in the broad categories of mood, anxiety, behavioural, and substance abuse disorders. A full list of the conditions included in this study can be found in Kessler et al. 2010 (Kessler et al. 2010).

While specific types of adverse experience were all shown increase schizophrenia risk to a similar degree in this study (Varese et al. 2012), it has recently been demonstrated that specific types of adversity, while not differentially correlated with schizophrenia diagnosis, do show a correlation with specific psychotic symptoms (Bentall et al. 2012, Wickham, Sitko and Bentall 2015). For example, Bentall et al. demonstrated that childhood rape was associated with a significantly increased risk of hearing voices (OR 8.90, $p < 0.05$), but did not significantly alter risk of paranoia (OR 2.78) (Bentall et al. 2012). On the other hand, growing up in institutional care was significantly associated with increased odds of experiencing paranoia (OR 11.08, $p < 0.05$), but did not significantly alter risk of voice hearing (OR 3.45). Some incidences of trauma were also demonstrated to significantly increase the risk of both voice hearing and paranoia, such as in the case of physical abuse (OR 8.52 for paranoia, OR 4.79 for voice hearing; $p < 0.05$) (Bentall et al. 2012).

Overall, and similarly to the cumulative effect of increasing numbers of schizophrenia GWAS SNPs, Shevlin et al. demonstrated that experiencing two or more traumas was predictive of psychosis, with increased number of traumas also demonstrating an additive effect on risk. In this study, individuals who had experienced five traumas had an OR for schizophrenia between 30-193 depending on the cohort studied (Shevlin et al. 2008). Similarly, the 2012 meta-analysis by Varese et al. highlighted that nine of the 10 studies that tested for dose response in their meta-analysis demonstrated a positive correlation between increased number of adverse events and increased risk of psychosis (Varese et al. 2012). In this instance, 'psychosis' refers to individuals with a DSM diagnosis of psychotic disorder, schizophrenia, or schizoaffective disorder based on the criteria detailed in DSM versions III to IV-TR or ICD (International Classification of Diseases) versions 9 or 10.

This pattern also remained when considering symptoms individually rather than diagnosis. Bentall et al. demonstrated the dose response effect of trauma for risk of paranoia and hearing voices, with a single adverse experience significantly increasing the risk for both, with additional traumatic experiences having an additive effect on risk. Individuals with one to four experiences of trauma displayed an OR of 3.33-17.54 for paranoia, and an OR of 2.31-27.42 for voice hearing, which predominantly increased with number of adverse experiences (Bentall et al. 2012). While the genetic study did not assess trauma, and the trauma studies did not account for genetics, the latter studies demonstrate that the effect of trauma is at least as strong, if not significantly stronger, in influencing an individual's risk for experiencing schizophrenia or psychosis.

High stress sensitivity is a known risk factor for schizophrenia, with research by DeVylder et al. across almost 177,000 people from 39 countries demonstrating a 4.6%

prevalence of psychotic symptoms in the lowest stress sensitivity group compared to a prevalence of 22.4% in individuals with the highest stress sensitivity scores in this study (DeVylder et al. 2016). Childhood adversity has been shown to be associated with changes in brain structure and connectivity in psychologically healthy individuals (McCarthy-Jones et al. 2017, Herringa et al. 2013), with these physical changes being shown to correlate with mental health-related traits that are likely to sensitise the individual to future life stress (Herringa et al. 2013). For example, differences in hippocampal and medial prefrontal cortex grey matter volume as a result of childhood adversity have been shown to be associated with increased anxiety in response to recent life stress (Gorka et al. 2014), suggesting that differences in brain structure seen in those with experience of childhood trauma could contribute to schizophrenia risk through modulating sensitivity to future stress. This would support a second model of schizophrenia known as the 'traumagenic neurodevelopmental model', which suggests that experiences of early life stress, combined with genetic risk, have the potential to alter the trajectory of brain development in ways that would put an individual at higher risk of experiencing psychosis in response to future adversity (Figure 1.9) (Read et al. 2001, Read et al. 2014).

In line with the traumagenic neurodevelopmental model, and adding another layer of complexity, an individual's genotype has been shown to modulate the effects of childhood trauma on changes in brain structure.

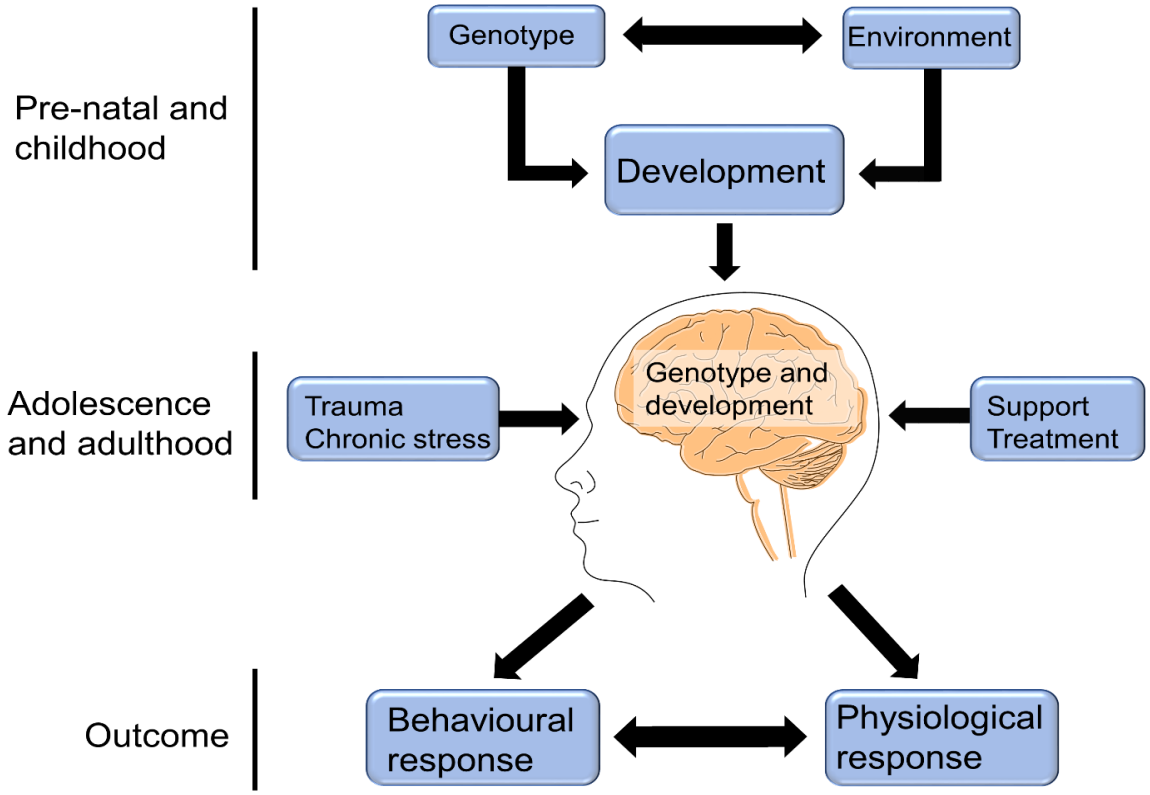


Figure 1.9 Traumagenic neurodevelopmental model of psychiatric conditions.

Based on a figure by Heim et al. (Heim et al. 2008). The traumagenic neurodevelopmental model of schizophrenia suggests that the interaction between an individual's genotype and environment will influence their brain development in ways which may put them at increased risk of experiencing schizophrenia in response to future life stress. For example, maternal stress or infection, coupled with childhood adversity may negatively impact a child's brain development, resulting in molecular or structural changes that affect their response to adversity later in life. Experiences in adolescence and adulthood can either result in further increased risk, such as in the case of trauma or chronic stress, or can reduce the individual's risk, in the case of receiving social support or treatment. The combination of genotype, brain development, and adolescent/adult life experiences can combine to influence the individual's molecular/physiological response and their behavioural response to future stress, which can in turn interact and impact their risk of schizophrenia.

For example, Dannlowski et al. have demonstrated the effect of the oxytocin receptor SNP, rs53576, on ventral striatum grey matter volume in individuals with varying levels of childhood trauma. Individuals who are homozygous for the G allele of rs53576 show marked decreases in ventral striatum grey matter volume in response to childhood trauma, compared to A allele carriers, who show a small increase in grey matter volume at this region in individuals with the highest trauma scores (Dannlowski et al. 2016).

Stress and trauma are clear and reproducible risk factors for schizophrenia and schizophrenia-like experiences, as well as being known modulators of epigenetic modifications and of gene expression (Klengel and Binder 2015, Vaiserman 2015). It is therefore clear that our life experiences, genes, and brains are in complex and constant interaction which can alter our gene expression, development, and behavioural response to future events.

For this reason, the work presented in this thesis aimed to characterise and further understand regulatory elements and transcriptional networks that may be involved in mental health and brain function, with the understanding that these pathways are likely to be modulated by stress and trauma in a way that would be influenced by an individual's genotype, with implications for their development and mental health. Such work built on previous studies assessing regulatory elements at CNS regions, in which variation had been linked to allele-specific expression or risk for CNS conditions (Davidson et al. 2011, D'Souza et al. 2013, Fiskerstrand et al. 1999, Hing et al. 2012, MacKenzie and Quinn 1999, Paredes et al. 2013, Warburton et al. 2015a). This included the characterisation of both highly conserved regulatory elements (ECRs), as well as evolutionary new and polymorphic elements (VNTRs, retrotransposons), with work to identify and describe their regulatory function within neuronal cell lines or in

ex vivo brain samples. Throughout this work, we were mindful of common variation in such elements, and the potential of variants to introduce allele-specific expression of genes of interest, and the potential for allele-specific molecular response to environmental challenge, which could have the potential to alter developmental trajectories or risk for disease in an epigenetic manner.

Chapter 2

Materials and Methods

2.1 Materials

2.1.1 Commonly used solutions

TBE buffer (5x): 108g Tris base (Sigma), 55g Boric acid (Sigma), 5.84g EDTA (Sigma), made up to 2L with distilled water.

LB agar: 40g/L in distilled water (Fluka Analytical).

LB broth: 25g/L in distilled water (Fluka Analytical).

2.1.2 Human DNA samples for genotype analysis

Genomic DNA from 823 individuals with a diagnosis of schizophrenia (average age 37.66 years) and 762 healthy controls (average age 46.27 years) were provided by Prof. Dan Rujescu from the Department of Psychiatry at the Martin Luther University Halle-Wittenberg, Germany. All individuals were of German or Central European descent, with schizophrenia status confirmed by diagnosis according to the Diagnostic and Statistical Manual of Mental Disorders (DSM-IV) and the International Classification of Disease (ICD-10). Healthy controls were selected from the general population in Munich. In order to rule out any control individuals that had experience, or who may have been at familial high risk, of a psychiatric disorder, initial screening and detailed medical and psychiatric histories of both the individual and their first-degree relatives were assessed through the structured clinical interviews, SCID I and SCID II. All individuals were provided with detailed information on the study, and provided written informed consent. The study was approved by the ethics committee of Ludwig Maximilians University in Munich, Germany, and carried out in accordance with ethical standards as detailed in the Declarations of Helsinki.

2.1.3 Human cell line used for in vitro models

SH-SY5Y (CRL-2266) is a human neuroblastoma cell line from the American Type Culture Collection (ATCC).

2.1.4 Cell culture media for SH-SY5Y cells

2.1.4.1 Complete media for culturing cells

Complete SH-SY5Y media was made up as follows: 1:1 mix of Minimal Essential Medium Eagle (Sigma) and Nutrient Mixture F-12 Ham (Sigma), supplemented with 10% foetal bovine serum (Sigma), 1% penicillin/streptomycin (100 U/ml, 100 mg/ml; Sigma), 1% (v/v) 200 mM L-glutamine (Sigma), and 1% (v/v) 100 mM sodium pyruvate (Sigma).

2.1.4.2 Freezing media for long term storage in liquid nitrogen

Freezing media for long term storage of SH-SY5Y cells in liquid nitrogen was comprised of 90% foetal bovine serum (Sigma) and 10% DMSO (Sigma).

2.1.4.3 Drug treatments for SH-SY5Y cell line model

1 mM amphetamine stock: Amphetamine (Sigma) was dissolved and made up to a stock concentration of 1 mM in filter sterilised distilled water. For use in cell treatments, stock solution was diluted to a final concentration of 10 μM in complete SH-SY5Y culture media (Jones and Kauer 1999, Shyu et al. 2004).

1 mM cocaine hydrochloride stock: Cocaine hydrochloride (Sigma) was dissolved and made up to a stock concentration of 1 mM in filter sterilised distilled water. For use in cell treatments, stock solution was diluted to a final concentration of 10 μM in complete SH-SY5Y culture media (Vasiliou et al. 2012).

1 M lithium chloride stock: Lithium chloride (Sigma) was dissolved and made up to a stock concentration of 1 M in filter sterilised distilled water. For use in cell treatments,

stock solution was diluted to a final concentration of 1 mM in complete SH-SY5Y culture media (Hing et al. 2012, Roberts et al. 2007).

1 M sodium valproate stock: was dissolved and made up to a stock concentration of 1 M in filter sterilised distilled water. For use in cell treatments, stock solution was diluted to a final concentration of 5 mM in complete SH-SY5Y culture media (Phiel et al. 2001, Zhang et al. 2003, Pan et al. 2005).

All stock solutions were stored at – 20 ºC.

2.2 Methods

2.2.1 Primer design for PCR

Primers for PCR were designed by first downloading the sequence for the region of interest from the UCSC Genome Browser (https://genome.ucsc.edu), with additional flanking sequence to allow for the design of primers. Sequences of over 18 bp were selected by eye, and the suitability was tested using OligoAnalyzer, hosted by Integrated DNA Technologies (https://eu.idtdna.com/calc/analyzer). This allowed selection of primers with a melting temperature between 55 – 65 ºC and a GC content of 40 – 60%. OligoAnalyzer was also used to identify the potential formation of hairpin structures, as well as the likelihood of homo- and hetero-dimers, allowing refinement of the selected primers to minimise the formation of these structures. After selection, primer specificity was tested using the UCSC Genome Browser's In Silico PCR tool. Primers were purchased from Eurofins Genomics.

2.2.2 Cloning methods using Gibson Isothermal Assembly

2.2.2.1 Primer design for cloning

PCR primers for cloning were designed as outlined in Section 2.2.1, with the addition of a ≥ 16 bp tag sequence which was complementary to the sequence flanking the cut site in the backbone vector. In this case, cutting pGL3P with SmaI resulted in the following flanking sequences: AGCTCTTACGCGTGCTAG added to the 5' end of all forward primers, and AGATCGCAGATCTCGAG added to the 5' end of all reverse primers.

2.2.2.2 PCR amplification using a proof-reading enzyme

PCR amplification of regions for cloning was carried out using the proof-reading Phusion High-Fidelity DNA Polymerase (NEB), with the master mix as below.

Component	**Volume**
Phusion HF Buffer (5x)	10 µl
dNTPs (10 mM each nucleotide)	1 µl
Forward primer (20 pmol/µl)	2.5 µl
Reverse primer (20 pmol/µl)	2.5 µl
Betaine	2.5 µl
Phusion DNA polymerase (2 U/µl)	0.5 µl
DNA template	2.5 µl (10 ng/µl)
Nuclease free water	21 µl
Total volume	50 µl

2.2.2.3 Agarose gel electrophoresis

Both PCR reactions and diagnostic restriction digests were analysed by gel electrophoresis, typically using 1.5% agarose depending on the expected band sizes. DNA bands were visualised through the addition of 0.5 µl of ethidium bromide (Sigma 10 mg/ml) per 10 ml of gel, and band size estimated through comparison to either 100

bp or 1 kb ladders (Promega), or a 2-log ladder (NEB). Gels were typically run at 100 volts for 1.5 hours, varying dependent on the gel percentage and size of the DNA fragments. Gels were visualised and imaged using the BioDoc-it Imagine System UV transilluminator.

2.2.2.4 Purification of DNA from agarose gels

After separation by gel electrophoresis, the appropriate bands were excised from the gel and column purified using the Wizard SV Gel and PCR Clean-Up System (Promega) following the manufacturer's protocol. DNA was eluted in 30 µl of nuclease free water.

2.2.2.5 Restriction enzyme digests

Restriction enzyme digests were used both to cut a vector to create specific nucleotide overhangs in preparation for cloning, and later diagnostically to determine successful cloning through visualising the presence and/or orientation of an insert. Restriction enzymes were purchased either from Promega or NEB, and used in the reaction as follows:

Component	**Volume**
Buffer (10x)	2 µl
BSA (10 mg/ml)	0.2 µl
Enzyme (10 U/µl)	0.5 µl
DNA	X µl
Nuclease free water	Y µl
Total volume	20 µl

Restriction enzyme digests were incubated at the appropriate temperature for the enzyme of use, typically for 1-2 hours. After digestion, the reaction mix was run on an

agarose gel (Section 2.2.2.3) to confirm restriction and/or to confirm the presence and orientation of the insert.

2.2.2.6 Ligation using Gibson isothermal assembly

Before cloning, the amount of vector and insert for ligation using Gibson isothermal assembly was calculated. NEB guidelines recommended the use of 100 ng of cut vector, and an insert:vector ratio of 3:1 for the assembly of 2-3 fragments. The amount of insert for each fragment was calculated using the following equation:

$$Insert\ (ng) = \frac{vector\ (ng)\ x\ size\ of\ insert\ (kb)}{size\ of\ vector\ (kb)}\ x\ ratio\ \left(\frac{insert}{vector}\right)$$

The ligation reaction was set up as follows:

Component	**Volume**
Insert DNA	X µl
Vector (50 ng/µl)	2 µl
Gibson assembly master mix (2x)	10 µl
Nuclease free water	Y µl
Total volume	20 µl

The Gibson assembly master mix contains a 5’ exonuclease, a DNA polymerase, and a DNA ligase. All reactions were incubated at 50 ºC for 15 minutes. During the time taken for the thermocycler to reach 50 ºC, the exonuclease was able to chew back the 5’ ends of the cut vector and insert, producing complementary overhangs. Upon reaching 50 ºC, the exonuclease was inactivated, leaving the remaining time for the insert and vector to anneal at the complementary overhangs, the DNA polymerase to extend the 3’ ends, and the DNA ligase to repair the backbone (Figure 2.1).

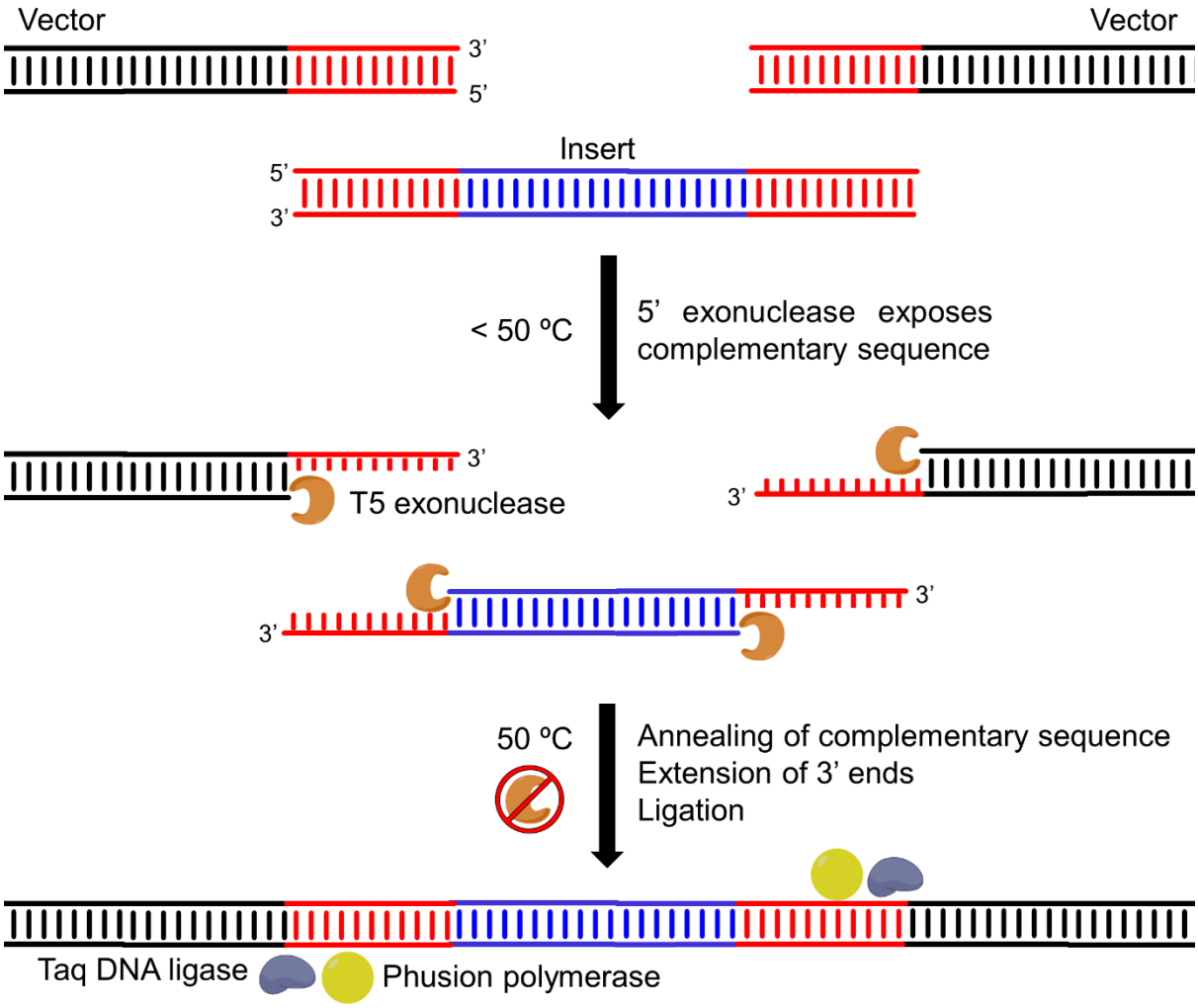


Figure 2.1 Gibson isothermal assembly reaction

Before amplifying the desired region of DNA which will form the insert for the construct, PCR primers are designed to add ≥ 16 bp overhangs to the insert which will be complementary to the sequence on either side of the restriction enzyme site used to linearise the vector. Once amplified, a 3:1 ratio of insert to vector is added to the Gibson isothermal assembly mix, which contains a 5' exonuclease, a DNA ligase, and a DNA polymerase. The reaction mixture is heated to 50 ºC. During the time it takes for the reaction to reach 50 ºC, the 5' exonuclease chews back the 5' strands of the insert and vector revealing complementary overhangs. Once the reaction reaches 50 ºC, the exonuclease is inactivated. The complementary overhangs anneal, and the polymerase fills in any gaps using the complementary strand as a template, with the ligase ligating the backbone of the vector and insert to form a complete vector with insert.

2.2.2.7 Transformation of ligation mix into competent E. coli cells

Gibson assembly ligation mix was transformed into XL Gold ultracompetent cells (Agilent) as detailed in the manufacturer's protocol. Briefly, 100 µl aliquots of cells were thawed on ice and placed in pre-chilled 14 ml falcon tubes. 4 µl β-mercaptoethanol was added to each aliquot and incubated on ice for 10 minutes, swirling gently every two minutes. 2 µl of ligation mixture was added to an aliquot of cells, swirled gently to mix, and then incubated on ice for 30 minutes.

After incubation, cells were heat shocked in a water bath at 42 ºC for 30 seconds and incubated on ice for 2 minutes. Next, 0.9 ml NZY^{+} broth (pre-heated to 42 ºC) was added to each, and tubes were incubated at 37 ºC for one hour, shaking at 225 rpm.

Following one hour incubation, 200 µl of each transformation mix was plated on LB agar plates supplemented with 100 µg/ml ampicillin and incubated overnight at 37 ºC.

2.2.2.8 Purification of plasmid DNA from transformed E.coli

2.2.2.8.1 Mini-prep of plasmid DNA

Individual colonies were picked from an overnight incubation of transformed *E.coli* culture on LB agar plates (100 µg/ml ampicillin) (Section 2.2.2.7) and grown overnight in 3 ml LB broth with 100 µg/ml ampicillin at 37 ºC, shaking at 225 rpm. Plasmid DNA was extracted and purified from 1 ml of the 3 ml overnight culture using the QIAprep spin miniprep kit (Qiagen) or the Wizard plus SV miniprep DNA purification system (Promega) according to the manufacturer's protocols. DNA was eluted in 30 µl nuclease free water and either stored at – 20 ºC or directly used for restriction enzyme digest (Section 2.2.2.5) to determine the presence or absence of an insert.

2.2.2.8.2 Maxi-prep of plasmid DNA

In order to extract plasmids with greater purity and yield for use in reporter gene assays, 100 ml bacterial cultures were grown overnight by adding 100 µl of 3 ml mini-prep culture (Section 2.2.2.8.1) to 100 ml LB broth with 100 µg/ml ampicillin, grown overnight at 37 ºC and shaking at 225 rpm. Plasmid DNA was extracted and purified using the Qiagen plasmid maxi kit according to manufacturer's protocols for purifying high copy plasmids. DNA pellets were resuspended in 150 µl nuclease free water and the concentration determined using a Nanodrop 8000 before storing for later use at –20 ºC.

2.2.2.9 Sequencing

Sequencing verification of plasmids with successfully cloned inserts was carried out externally by Source Bioscience, with samples of 5 µl at 100 ng/µl, and 5 µl of each sequencing primer at 3.2 pmol/µl. Sequencing primers for the pGL3P vector are outlined in Section 8.

2.2.3 Generation of reporter gene constructs

Evolutionary conserved regions (ECRs) at the MIR137 and EU358092 locus were amplified using Phusion High-Fidelity Polymerase (Section 2.2.2.2; Section 8) and ligated into the multiple cloning site of the pGL3P luciferase reporter vector through Gibson isothermal assembly (Section 2.2.2.6). Ligation reactions were transformed into chemically competent XL-Gold *E.coli* (Section 2.2.2.7) for selection. Successful clones were verified by restriction enzyme digest and sequencing (Sections 2.2.2.5 and 2.2.2.9), and maxi-preps (Section 2.2.2.8.2) were used to generate a high yield of each plasmid for luciferase reporter gene assays.

2.2.4 Cell culture methods

2.2.4.1 Culturing SH-SY5Y cells

Human SH-SY5Y neuroblastoma cells were maintained in SH-SY5Y media (Section 2.1.4.1) at 37 ºC and 5% CO_2, typically grown and maintained in T175 flasks and passaged to new flasks when approaching 70-80% confluency. To passage cells, media was removed and the cells washed with 10 ml sterile phosphate buffered saline (PBS) (Sigma) which had been pre-warmed to 37 ºC. After washing and removing the PBS, 5 ml of pre-warmed, 1x trypsin (Sigma) was added and the flasks incubated at 37 ºC for 3 minutes in order to detach cells from the bottom of the flask. To neutralise the trypsin, cells were washed and resuspended in 10 ml pre-warmed SH-SY5Y media by repeated pipetting. 1 ml of cell suspension was then transferred to a new T175 flask containing 30 ml pre-warmed SH-SY5Y media, and incubated at 37 ºC with 5% CO_2. Cell lines were tested for mycoplasm every six months using the MycoAlert Mycoplasma Detection kit (Lonza), and underwent short tandem repeat (STR) analysis to confirm cell line identity.

2.2.4.2 Counting cells using a haemocytometer

A haemocytometer was used to count the number of cells per ml of media. The central counting square of the haemocytometer contains a 5 x 5 grid of 25 squares bound by three parallel lines, each containing 25 smaller squares in the same 5 x 5 configuration. Both the coverslip and haemocytometer were washed with 70% ethanol before and after use. During the passaging of cells (Section 2.2.4.1), after washing and resuspending in 10 ml media, 20 µl of cell suspension was transferred to a 1.5 ml Eppendorf where it was mixed 1:1 with 20 µl of trypan blue (Sigma) to stain dead cells. 20 µl of the cell suspension stained with trypan blue was introduced under the coverslip and onto the counting surface of the haemocytometer. The counting surface

was visualised using a light microscope and 10 x objective lens. The number of cells in at least three of the larger 5 x 5 squares were counted, including cells which touched the top or left boarders, and excluding those touching the bottom and right-hand boarders. An average was taken of the number of cells per large 5 x 5 square, and then multiplied by two to correct for the dilution caused by the addition of trypan blue. As a large 5 x 5 region corresponds to 0.1 mm^3, the average number of cells multiplied by two was then multiplied by 10,000 to give the number of cells per 1 cm^3 or 1 ml of media. This number was then used to calculate the seeding density for cells.

2.2.4.3 Freezing cells in liquid nitrogen for long term storage

Cells were frozen in freezing media (Section 2.1.4.2) and placed in liquid nitrogen for long term storage. Cells were grown in T175 flasks until 70 – 80% confluent, and passaged as in Section 2.2.4.1, with the exception of neutralising trypsin and resuspending cells using 10 ml freezing media in place of the usual cell culture media. The cells suspended in freezing media were then aliquoted into 1.8 ml cryovials and placed into an isopropanol-containing Mr Frosty freezing container which was stored at – 80 ºC for 24 hours, before transferring the cryovials to liquid nitrogen.

2.2.4.4 Drug treatments

Drug treatments were carried out using concentrations for cell culture that had previously been optimised in the lab or reported in the literature. Stock solutions of each drug were diluted in cell culture media to the appropriate working concentration as outlined in Section 2.1.4.3. Cells were cultured for 24 hours, before undergoing one hour incubation with either amphetamine, cocaine, lithium, or sodium valproate, with cells harvested directly after the one hour drug treatment for RNA extraction (n = 4). Basal untreated cells were also harvested for use as controls.

2.2.4.5 Transient transfections for reporter gene assays

SH-SY5Y cells were seeded at approximately 100,000 cells per well across 24-well plates and incubated for 24 hours at 37 ºC and 5% CO_2. After 24 hours, each well was transfected with 1 µg plasmid DNA and 20 ng TK Renillia luciferase plasmid as an internal control (n = 4) using the TurboFect transfection reagent (Thermo Scientific) in accordance with the manufacturer's guidelines. The empty pGL3P backbone was used as a control to determine baseline luciferase expression. Cells were processed 48 hours post-transfection using the Dual Luciferase Reporter Assay system (Promega).

2.2.5 Luciferase reporter gene assays

2.2.5.1 Cell lysis

At 48 hours post-transfection, cell culture media was removed from each well and the cells washed with 1 x PBS. 5 x Passive Lysis Buffer (PLB) was diluted to 1 x PLB with nuclease free water and 100 µl added to each well. After the addition of 1 x PLB, plates were incubated at room temperature on a rocking platform for 15 minutes. 20 µl of each cell lysate was transferred to a white 96-well plate for analysis using the Glomax 96 Microplate Luminometer (Promega).

2.2.5.2 Measuring reporter gene levels by dual luciferase assay

Using the Dual-Luciferase Reporter Assay system (Promega), 100 µl of luciferase assay reagent II (LARII) and 100 µl Stop and Glo reagent were prepared per sample according to the manufacturer's protocol and allowed to reach room temperature. The Glomax 96 Microplate Luminometer (Promega) was set up using default settings for a dual luciferase reporter gene assay with two injectors, and both injectors were flushed three times with distilled water, 70% ethanol, and again with distilled water. A fourth flush with air was used to clear residual ethanol or water from the injectors, which were

then primed with the LARII in injector 1 and the Stop and Glo reagent in injector 2. The white 96-well plate containing cell lysates was placed into the luminometer and the dual luciferase program was run. The LARII reagent is injected first and a measurement of the bioluminescence produced by firefly luciferase catalysis is recorded. Following this, the Stop and Glo reagent quenches this reaction, allowing measurement of the bioluminescence produced by catalysis of the *Renilla* luciferase protein as an internal control.

2.2.5.3 Statistical analysis

Fold changes in firefly luciferase activity supported by the plasmids were normalised to the *Renilla* luciferase measurement for their corresponding well. Each transfection was carried out across four wells, which were then averaged to minimise variation due to transfection efficiency or pipetting errors. The bioluminescence measurements supported by the empty pGL3P control transfection were used as a baseline of 1, to which all fold-changes in expression were normalised. Significance was calculated using a two-tailed t-test and scored as follows: *$p < 0.05$, **$p < 0.01$, ***$p < 0.001$. For each transfection, $n = 4$. The p-values calculated using the two-tailed t-test were not corrected for multiple testing. However, luciferase data for nine ECRs is outlined in this thesis, which would give a Bonferroni-corrected p-value of 0.006 for significance.

2.2.6 Analysis of endogenous gene expression

2.2.6.1 RNA extraction

Total RNA was extracted using TRIzol reagent (Invitrogen) according to manufacturer's protocols. Briefly, around 400,000 SH-SY5Y cells were seeded into 6-well plates and incubated for 24 hours. After 24 hours, or after 24 hours and a one hour incubation with treatment, cell culture media was removed from each well and 1 ml TRIzol was added per 10 cm^2. The cells were pipetted repeatedly in TRIzol to lyse,

before transferring the lysate to a 1.5 ml Eppendorf and incubating at room temperature for 5 minutes. Following incubation, 0.2 ml chloroform per 1 ml of TRIzol reagent was added to each sample and shaken vigorously by hand for 15 seconds. Samples were then incubated at room temperature for a further 2 – 3 minutes and centrifuged at 12,000 x g for 15 minutes at 4 ºC. Centrifugation allows phase separation of three layers, including the colourless upper layer containing RNA, a thin white middle layer containing DNA, and a lower pink organic layer containing protein. For RNA extractions, approximately 500 µl of the upper colourless layer was removed and transferred into a fresh 1.5 ml Eppendorf. 500 µl of 100% molecular grade isopropanol was added to each sample per 1 ml of TRIzol reagent and incubated at room temperature for 10 minutes. All samples were then centrifuged for 10 minutes at 12,000 x g and 4 ºC, after which the supernatant was removed to leave an RNA pellet. To wash the RNA pellet, 1 ml of 75% molecular grade ethanol per 1 ml of TRIzol used is added to each pellet. Samples were then vortexed and centrifuged at 7,500 x g for 5 minutes at 4 ºC. After re-pelleting the RNA and removing the supernatant, the pellets were left to air dry for 5 – 10 minutes, before being resuspended in 20 µl of nuclease free water. Samples were then incubated at 55 ºC for 10 – 15 minutes to aid solubilisation. The concentration of RNA was determined using a Nanodrop 8000 (Section 2.2.6.2) and then used either directly for first strand cDNA synthesis (Section 2.2.6.3) or stored at – 80 ºC for later use.

2.2.6.2 Quantifying RNA concentration by spectrophotometry

RNA concentration was measured using a Nanodrop 8000. The nanodrop was set to the RNA setting and calibrated with 1 µl of nuclease free water (which was used to dissolve the RNA pellet). After calibrating the nanodrop and cleaning the pedestal, 1 µl of the RNA sample was loaded and the absorbance measured. Depending on the

concentration of the sample, varying amounts of UV light will be absorbed at 260 nm by the nucleic acids. By measuring the optical density (OD) of the RNA, the nanodrop is able to calculate the concentration of the sample. Measurements of RNA quality were also noted through the measurement of the 260/280 ratios, with samples giving a 260/280 reading of 2.0 and above being regarded as high quality RNA.

2.2.6.3 First strand synthesis of cDNA

cDNA was synthesised from extracted RNA (Section 2.2.6.1) using the GoScript Reverse Transcription system (Promega) following the manufacturer's protocol. Typically, up to 5 µg of total RNA was used to synthesise cDNA, with the volume of RNA used from each sample varying with sample concentration. The initial step involved addition of random primers to the RNA followed by denaturing.

Component	**Volume**
RNA (up to 5 µg)	X µl
Random primers (500 ng)	1 µl
Nuclease free water	Y µl
Total volume	5 µl

The RNA and primer mixture was incubated at 70 ºC for 5 minutes to denature and then cooled on ice for 5 minutes. Following cooling, the reverse transcription mix below was made up separately and then added to the RNA and primer mix to give a total reaction volume of 20 µl.

Component	**Volume**
GoScript reaction buffer (5 x)	4 µl
$MgCl_2$ (25 mM)	4 µl
dNTPs (10 mM each nucleotide)	1 µl
Recombinant RNasin ribonuclease inhibitor (40 U/µl)	0.5 µl

GoScript reverse transcriptase	1 µl
Nuclease free water	X µl
Total volume	15 µl

15 µl of the above reverse transcription mix was added to each 5 µl mix of RNA and primers and incubated at 25 ºC for 5 minutes to allow primer annealing. After this, all samples were incubated at 42 ºC for 1 hour to allow extension by the reverse transcriptase enzyme. Samples were then heated to 70 ºC for 15 minutes to inactivate the reverse transcriptase and end the reaction. The resulting cDNA was diluted appropriately using nuclease free water and stored at – 20 ºC for later use.

2.2.6.3 PCR of cDNA for expression analysis

For analysis of gene expression, cDNA (Section 2.2.6.2) was amplified using GoTaq DNA polymerase (Promega) according to the manufacturer's guidelines. The following GoTaq Flexi master mix was used for amplification reactions, using an appropriate dilution of cDNA.

Component	**Volume**
GoTaq Flexi reaction buffer (5 x)	5 µl
$MgCl_2$ (25 mM)	4 µl
dNTPs (10 mM of each nucleotide)	1 µl
Forward primer (20 pmol/µl)	0.25 µl
Reverse primer (20 pmol/µl)	0.25 µl
GoTaq DNA polymerase (5 U/µl)	0.25 µl
cDNA template (appropriately diluted)	1 µl
Nuclease free water	X µl
Total volume	25 µl

PCR was carried out using a PeqSTAR 2X thermocycler (Peqlab). The annealing temperatures and primer sets used are outlined in Section 8. For amplification of the EU358092 transcript from basal and drug treated samples, undiluted cDNA was used. For β-actin, a cDNA dilution of 1:200 was used. After amplification, samples were run on an agarose gel (Section 2.2.2.3) or stored at – 20 ºC for later use.

2.2.7 Genotyping

2.2.7.1 PCR amplification of the MIR941 VNTR

The MIR941 VNTR was genotyped in 342 individuals with a diagnosis of schizophrenia and 340 psychologically healthy controls (Section 2.1.2) by PCR. Primers were designed as outlined in Section 2.2.1 and detailed in Section 8. Amplification of the target region was carried out using 5 ng of DNA and the 2x ReddyMix PCR Master Mix (Thermo Fisher) as below:

Component	**Volume**
ReddyMix PCR master mix (2x)	7.5 µl
Forward primer (20 pmol/µl)	1 µl
Reverse primer (20 pmol/µl)	1 µl
DNA template (5 ng/µl)	1 µl
Nuclease free water	9.5 µl

The PCR reactions were carried out using a PeqSTAR 2X thermocycler (Peqlab) set to the programme detailed in Section 8, and all samples were run on an agarose gel (Section 2.2.2.3) to determine polymorphism. For verification of the polymorphic sequence, at least one of each allele was sequenced by Source Bioscience (Section 2.2.2.9) after being purified from an agarose gel (Section 2.2.2.4).

2.2.7.2 Statistical analysis

Clump 24 analysis was used to test the significance of MIR941 VNTR genotype frequency between schizophrenia cases and controls (Section 2.1.2). The Clump 24 analysis software (accessed through: http://www.davecurtis.net/dcurtis/software.html) is designed to assess the significance of the departure of observed values from the expected values, utilising a Monte Carlo-based method. This is achieved by carrying out repeated simulations (typically 10,000) to generate 2 x N contingency tables. Each simulated table has the same marginal totals as the input data, and randomly simulated data is used to count the number of times that the chi squared value for the input table is achieved through random simulation. Clump also provides a fourth chi squared value (T4) which is generated through 'clumping' columns into a new 2 x 2 table. This is designed to maximise the chi squared value, and is able to directly test the hypothesis that certain genotypes are more common in cases than controls (Sham and Curtis 1995).

Clump analysis uses Monte Carlo methods to test the significance of chi squared values by assessing how many times the observed value is reproduced by chance from randomly simulated data sets. The programme then outputs four test statistics based on different methods, as described below:

T1 is based on the raw 2 x N table of original input data.

T2 is based on the original input data, with columns containing data on low frequency alleles clumped together.

T3 takes the most significant of all the 2 x 2 tables created by comparing each column of the original input table (excluding columns with data on rare alleles) against the total of all other columns.

T4 is based on a 2 x 2 table which is generated through clumping the columns of the original input table to maximise the chi squared value.

2.2.8 RNA-seq protocol and statistical analysis

Access to RNA-seq data was restricted to Eli Lilly employees working on Eli Lilly property only. Therefore, all RNA-seq analysis detailed in this thesis was carried out by Dr. Karim Malki and Dr. Nathan Lawless at Eli Lilly and Company Limited, with support from Dr. Andrew Jaffe at the Lieber Institute for Brain Development. RNA-seq data was obtained from the dorsolateral pre-frontal cortex of 155 individuals with schizophrenia and 196 healthy controls, as provided and in accordance with guidelines set out by the Lieber Institute. The following methods were provided by Dr. Karim Malki of Eli Lilly and Company.

2.2.8.1 Post-mortem brain samples

Human post-mortem brain tissues were obtained by autopsy and donated to the Lieber Institute for Brain Development from the Offices of the Chief Medical Examiner of the District of Columbia and of the Commonwealth of Virginia, Northern District. Informed consent was obtained from the next of kin following protocol 90-M-0142 approved by the NIMH/NIH Board. Sample clinical characteristics, including diagnosis were obtained using standard protocols described by Lipska et al. (Lipska et al. 2006). Subjects with evidence of micro- or macroscopic neuropathology following examination were excluded from this study. Toxicological analysis of biological samples to measure antipsychotic use was performed at time of death.

2.2.8.2 RNA extraction and sequencing

RNA extraction and sequencing was performed using previously published protocols described by Jaffe et al. (Jaffe et al. 2017a). Briefly, RNA was extract from ~100 mg

of post-mortem tissue homogenates of DL-PFC grey matter using the RNeasy kit (Qiagen) following standard protocols. All poly(A)-containing RNA molecules were DNAse treated from 1 µg of total RNA. Sequencing libraries were constructed using the Illumina TruSeq © RNA sample preparation kit, and sequencing was performed using the Illumina HiSeq 2000.

FASTQ files containing raw reads were aligned to the human genome (UCSC Genome Browser, hg19) using Top Hat v2.0.4. Transcriptomic profiles for all samples were determined using several convergent expression methods based on known gene annotation (gene, exon count, and transcript-level quantification) and methods that rely solely on read alignments (for example exon-exon splice junctions). Gene counts and exon counts were generated using the featureCounts tool based on Ensembl v75 (Liao, Smyth and Shi 2014).

Counts were converted to reads per kilobase of transcript per million mapped reads (RPKM) using the total number of aligned reads and used as dependent variables in statistical models. The DER Finder approach, implemented in the derfinder package for R (cran –R) was used to identify expression regions (ER).

2.2.8.3 Statistical analysis

For the purpose of this analysis, only subjects between the ages of 17-80, with a gene assignment rate >0.5, mapping rate >0.7, RNA integrity (RIN) >6, self-reported Caucasian or African American ancestry, and falling within the normal range of the distribution on the 2nd ancestry principle component (PC) were included. A total of 155 schizophrenia cases and 196 controls were included. A weighted, multiple linear model was fitted using the lmFit function and genes were ranked in order of evidence for differential expression using the ebayes function, both implemented in the limma

package for R, available on Bioconductor (https://bioconductor.org/packages/release/bioc/html/limma.html) (Ritchie et al. 2015).

Log_2 adjusted RPKM measures with an offset of 1 for each of the candidate genes were used as depended measure, and schizophrenia diagnosis modelled as fixed effects, adjusting for age, sex, and ancestry in the form of 5 SNP-derived principle components (PCs 1, 5, 6, 9, 10), mitochondrial mapping rate, RIN, and rate of gene assignment. Additionally, we modelled the first 12 PCs derived from the degradation matrix based on polyA+ libraries using the Quality Surrogate Variable method described by Jaffe et al. (Jaffe et al. 2017a, Jaffe et al. 2017b).

The moderated F-statistic (F) summarises the t-statistics for each gene into an overall test of significance based on a model which includes the context of all covariates in the analysis. The q-value is based on the distribution of p-values for each gene and is corrected for multiple testing, highlighting which genes are still significant in this model after correction. The q-value is a measure of the goodness of fit of the model when all covariates are considered.

2.2.9 Bioinformatic analysis

2.2.9.1 Using ECR Browser to view evolutionary conservation

The Evolutionary Conserved Region Browser (ECR Browser; https://ecrbrowser.dcode.org/) was used to identify regions of high conservation around the MIR137 and EU358092 locus, using an alignment of seven vertebrate species against the human genome, including puffer fish (*Fugu rubripes*), frog *(Xenopus tropicalis*), chicken (*Gallus gallus*), opossum (*Monodelphis domestica*), mouse (*Mus musculus*), dog (*Canis familiaris*), and rhesus macaque (*Macaca mulatta*). Regions with over 70% conservation back to either opossum, chicken, frog,

or puffer fish were considered to be highly conserved through evolution. This informed the selection of regions to study when testing for evolutionary conserved modulators of gene expression around the MIR137 locus.

2.2.9.2 Using UCSC Genome Browser and Galaxy for bioinformatic analysis

The UCSC Genome Browser (https://genome.ucsc.edu) was used to carry out bioinformatic analyses across the human genome using the 2009 GRCh37/hg19 genome build. In particular, the UCSC table browser was accessed through the web-based platform Galaxy (https://usegalaxy.org/). This allowed the download, upload, and intersection of data sets available on the UCSC Genome Browser, as well as custom data sets. Specifically, the UCSC Genome Browser and Galaxy were used to download and overlay ENCODE ChIP-seq data for transcription factors EZH2, SUZ12, and REST, with the transcriptional start sites of all genes annotated on the UCSC 'known gene' track. This allowed identification of gene sets with EZH2, SUZ12, or REST binding within 500 bp of the transcriptional start site. Similarly, these tools were used to identify and download co-ordinates of retrotransposable element subfamilies SVA, L1HS, L1PA2, and L1PA3 from the human 'repeat masker' data set, and identify their proximity to genes on a genome-wide scale by overlaying their co-ordinates with the co-ordinates of all known transcripts in the 'known gene' data set, adding 5 kb upstream to capture retrotransposable elements in the promoter region of genes.

Data sets could be specified and downloaded through Galaxy or the UCSC Table Browser and saved as browser extensible data (.BED) files. All .BED files used in this thesis are available on the accompanying disk in the Supplementary Files folder. These data sets could then be loaded into UCSC separately and intersected using the UCSC Table Browser. For example, after downloading and saving the co-ordinates of all EZH2 binding signals across the genome from the ENCODE ChIP-seq data set,

and all the transcriptional start sites in the genome from the 'known gene' data set, 500 bp was added and subtracted from each transcriptional start site co-ordinate in Excel to give a 1 kb minimal promoter region. These co-ordinates were then saved as a new .BED file and uploaded through the UCSC Table Browser. To identify all transcriptional start sites with EZH2 binding within 500 bp, the Table Browser tool was instructed to intersect data from the two files and return a list of transcriptional start site co-ordinates and corresponding gene names only for those regions which overlapped the co-ordinates of EZH2 binding in the second file.

Further, the UCSC Table Browser can be used to upload data from external sources for viewing alongside other data sets available through the genome browser. For example, schizophrenia GWAS data was downloaded as a .BED file from the PGC website and uploaded to the UCSC Genome Browser to overlay with conservation data.

2.2.9.3 Using Ricopili to visualise schizophrenia GWAS data at the MIR137 locus

Ricopili (https://data.broadinstitute.org/mpg/ricopili/), a web-based GWAS visualisaition tool hosted by the Broad Institute, was used to view the distribution of schizophrenia GWAS SNPs from the Psychiatric Genomics Consortium's 2013 schizophrenia GWAS data set ('PGC_SCZ52_may13') around the MIR137 locus.

2.2.9.4 Using HapMap Genome Browser and HaploView for linkage disequilibrium analysis

SNP genotype data for the CEU/CEPH European cohort across the MIR137 locus (chr1:98,498,912–98,595,043 and chr1:98,105,779–98,855,147) was downloaded from the now retired HapMap Genome Browser, release #28 (August 2010), and uploaded into Haploview 4.2 for analysis

(https://www.broadinstitute.org/haploview/haploview). Haploview is freely available software supported by the Broad Institute which enables analysis of linkage disequilibrium (LD) and the definition of haplotype blocks (Barrett et al. 2005). LD analysis was performed using default parameters and outputting the D prime (D') statistic. The D' statistic is derived from D, the coefficient of linkage disequilibrium, which measures the difference between the observed frequency of alleles at adjacent loci on a single chromosome, and the expected frequency if the alleles were segregating randomly. Because the D statistic value is reliant on the frequency of the alleles in question, this can be a problematic measurement when considering different groups of alleles. Therefore, the D' statistic is a normalised version of the D statistic, which is generated by dividing D by the theoretical maximum difference between the observed and expected allele frequencies. Alleles are said to be in complete LD when D' is equal to one, or high LD when D' is equal to or higher than 0.8. Haplotype blocks were determined using default parameters, as defined by Gabriel et al (Gabriel et al. 2002).

2.2.9.5 Using HaploReg v4.1 to access chromatin state and histone modification data in a range of human tissues and cell lines

SNPs of interest were input into HaploReg v4.1 (http://archive.broadinstitute.org/mammals/haploreg/haploreg.php), a web-based tool for exploring data on chromatin states and histone marks around the input SNPs based on ChIP-seq data in 130 human tissues and cell lines (Ward and Kellis 2012, Ward and Kellis 2016). This allowed us to determine whether the regions including the selected SNPs had potential transcriptional regulatory properties, and if so, which tissues these regulatory regions may be active in.

2.2.9.6 Using Enrichr to perform enrichment analysis on gene lists

Gene lists generated through the overlaying of transcription factor binding or retrotransposon co-ordinates with the co-ordinates of annotated genes (Section 2.2.9.2) were analysed using the Ma'ayan Lab Enrichr tool (http://amp.pharm.mssm.edu/Enrichr/) (Chen et al. 2013, Kuleshov et al. 2016). Enrichr performs enrichment analysis of an input gene list against previously annotated gene lists known to be involved in specific pathways or functions by drawing data from numerous sources including those providing data on regulation, pathways, ontologies, tissue distribution, and disease states. Enrichr uses a pre-computed look up table with expected values for each enrichment term based on a large number of random test gene lists. The standard deviation of the original input list from the expected value based on random gene lists is used to determine the significance of the input gene list for each particular enrichment term. We predominantly made use of enrichment data for our gene lists based on information from the Gene Ontology Consortium (GO), and from the Mouse Genome Informatics (MGI) mouse phenotype data set, both of which are available through Enrichr.

2.2.8.7 R programming language for analysis of retrotransposon distribution data

R is a freely available software package, language, and environment that is designed for handling large data sets and performing statistical computing. In this instance, we used R to handle large data, with a short script written by Dr. Bethany Geary to count the number of retrotransposons per Mb across the human genome based on input data downloaded from the UCSC Genome Browser's 'repeat masker' data set (Section 2.2.9.2). Further, we used a publicly available script provided by Dr Giovanni M Dall'Olio through the BioStars website to count the number of transcripts per

megabase across the genome (https://www.biostars.org/p/169171/#169211). Briefly, this accessed the human genome build 19 ‘known gene’ data set from UCSC through the ‘human.genes’ object using the ‘Homo.sapiens’ Bioconductor package, and counted the number of transcripts present in a specific genomic range (in this case, windows of 1 Mb).

Chapter 3

Identification and characterisation of regulatory domains and a non-coding RNA at the MIR137 locus and their potential involvement in brain development and schizophrenia.

Part I: Evolutionary conserved regions at the MIR137 locus

Part II: A novel brain-expressed RNA, EU358092, at the MIR137 locus

The work in the following sections has been published in the Journal of Molecular Neuroscience (Part I, doi: 10.1007/s12031-016-0812-x) and in Schizophrenia Research (Part II, doi: 10.1016/j.schres.2016.11.034).

Part I: Evolutionary conserved regions at the MIR137 locus

3.1 Introduction

Meta-analyses of genome-wide association studies in schizophrenia have identified the MIR137 locus on chromosome 1p21.3 (chr1:98,298,371–98,581,337, GRCh37/hg19) as one of the most highly associated regions for schizophrenia based on significant p-values for risk (Ripke et al. 2013, Schizophrenia Psychiatric Genome-Wide Association Study 2011). The miRNA, MIR137, is highly expressed in the brain and has known roles in a range of CNS-related pathways and processes, including in neural development, adult neurogenesis, and the regulation of synaptic plasticity (Smrt et al. 2010, Szulwach et al. 2010, Siegert et al. 2015). Further to this, MIR137 is known to target multiple other schizophrenia-associated GWAS genes, including five of the top 14 GWAS hits, CSMD1, C10orf26, CACNA1C, TCF4, and ZNF804A (Kim et al. 2012, Kwon, Wang and Tsai 2013, Collins et al. 2014). This would suggest that MIR137 may sit at the top of a larger schizophrenia-associated regulatory network, with even subtle changes in the levels of MIR137 having the potential for far-reaching downstream effects in terms of the regulation of schizophrenia-associated gene networks.

For this reason, it is important to determine regulatory elements around the MIR137 locus that may contribute to the modulation of MIR137 levels in a tissue-specific and stimulus inducible manner. Sequencing across the MIR137 locus by Duan et al. identified 133 rare SNP variants (minor allele frequency < 0.5), which were found to be over-represented in regulatory non-coding regions such as promoters and enhancers when compared to insulator regions (as defined by histone methylation patterns). The frequency of these rare variants was determined through Sanger sequencing in 2,610 individuals with a diagnosis of schizophrenia and 2,611 controls,

and significance calculated using a two-sided Fisher's exact test on the minor allele count of both cohorts. In particular, this work identified a rare SNP, 1:g.98515539 A > T, in an enhancer element, which was shown to be associated with schizophrenia and reduced the enhancer activity predicting lower MIR137 expression (Duan et al. 2014).

Previous work by the group has shown that overlaying GWAS data onto comparative genomics can identify regulatory regions in non-coding DNA that are likely to be involved in a range of conditions, including depression, alcoholism, and obesity (Davidson et al. 2011, Davidson et al. 2016, Hing et al. 2012, Paredes et al. 2011). Using a similar method to identify and characterise conserved regulatory domains around the MIR137 locus, we made use of the ECR Browser and the UCSC Genome Browser to first align and compare multiple vertebrate genomes to identify regions of high conservation, and then to overlay publicly available schizophrenia GWAS data to identify potential regulatory regions with relevance to schizophrenia biology. Each ECR was tested in reporter gene constructs to assess regulatory potential in the SH-SY5Y neuroblastoma cell line, and the interrogation of chromatin state and histone modification data over each ECR through the HaploReg v4.1 tool allowed us to extend our hypotheses as to their function *in vivo*.

3.2 Aims

- Comparative genomic analysis of the MIR137 locus to identify regions of high conservation across vertebrate species.
- Use of publicly available schizophrenia GWAS data from the Psychiatric Genomics Consortium to identify schizophrenia-associated SNPs within or nearby regions of high conservation at the MIR137 locus.
- Use HaploReg to identify tissues in which selected regions of conservation may be functionally active based on chromatin state and histone modification data.
- Validate the regulatory potential of selected conserved regions using reporter gene assays in an SH-SY5Y neuroblastoma cell line model.

3.3 Results

3.3.1 Bioinformatic analysis using Ricopili and the ECR browser identified seven regions of high evolutionary conservation at the MIR137 schizophrenia GWAS locus.

Ricopili, a web based tool hosted by the Broad Institute, allows visualisation of GWAS data overlaid onto the genome, enabling rapid identification and prioritisation of regions of interest for further study. We used the Ricopili tool to identify a select region around MIR137 showing the highest association with schizophrenia through GWAS for the further study of this locus.

Viewing the 'PGC_SCZ52_may13' data set, or the Psychiatric Genomics Consortium's second schizophrenia GWAS data set published in May 2013 (PGC2; available at https://www.med.unc.edu/pgc/results-and-downloads), on Ricopili defined the MIR137 GWAS locus as chr1:98,298,371-98,595,000 (GRCh37/hg19), which includes the long precursor transcript of MIR137, known as MIR137HG, and the neighbouring gene, DPYD (Figure 3.1a). As the strongest signal for schizophrenia-association was directly over the MIR137HG transcript and extended upstream of the gene, we focused in on a smaller region encompassing MIR137HG and an area of 80 kb upstream (chr1:98,448,000-98,595,000; GRCh37/hg19) (Figure 3.1b).

Having defined the region of interest based on schizophrenia GWAS data, we next set about identifying ECRs as a method of identifying potential regulatory elements that may modulate expression of MIR137.

Figure 3.1 Distribution of schizophrenia GWAS SNPs across the MIR137/DPYD locus and ECRs.

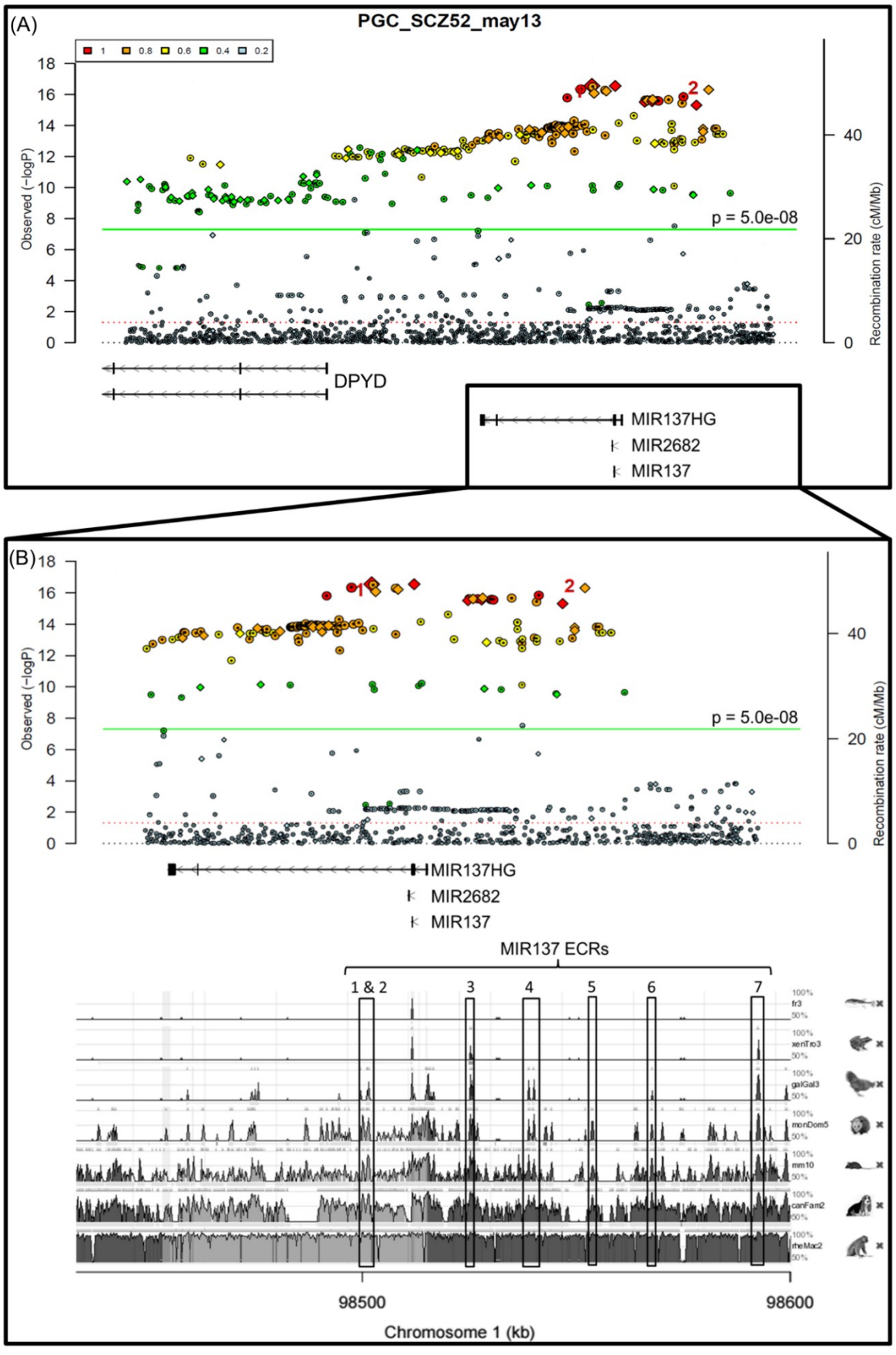


Figure 3.1 Distribution of schizophrenia GWAS SNPs across the MIR137/DPYD locus and ECRs.

(a) Visualisation of the Psychiatric Genomics Consortium's May 2013 schizophrenia GWAS data across the full schizophrenia-associated locus (chr1:98,298,371-98,595,000, GRCh37/hg19), using the Broad Institute's Ricopili tool. The boxed region including MIR137 and upstream regions (chr1:98,448,000-98,595,000; GRCh37/hg19) displays the highest signal at this locus in terms of association for schizophrenia, suggesting that MIR137 and potential upstream regulatory regions are likely to be the causal association at this locus. The boxed region is expanded below, and schizophrenia GWAS SNP data overlaid onto species conservation data from the ECR browser.

(b) Statistically significant schizophrenia GWAS SNPs are represented by coloured circles or diamonds above the green line ($p = 5.0 \times 10^{-8}$). Red numbers 1 and 2 highlight the schizophrenia GWAS SNPs rs1625579 and rs1198588, respectively, which were identified as the most significant SNPs for schizophrenia at this locus in the previous PGC GWAS analysis. ECRs selected for study are boxed and numbered. The conserved peaks between boxes 2 and 3 represent the miRNA and its proximal promoter, which were not included in this analysis, but highlight conservation as an indicator of functionality.

ECR browser is a web based tool which enables interactive visual browsing of evolutionary conserved regions through alignment and comparison of multiple vertebrate genomes. This enables fast identification of potentially important regulatory elements or ncRNAs, as high conservation of sequence through species suggests important functions at these loci.

We used the ECR browser to examine the region at chr1:98,448,000-98,595,000, using an alignment of seven species against the human genome; from the puffer fish (*Fugu rubripes*), frog *(Xenopus tropicalis*), and chicken (*Gallus gallus*), through to the opossum (*Monodelphis domestica*), mouse (*Mus musculus*), dog (*Canis familiaris*), and rhesus macaque (*Macaca mulatta*).

Through multiple species comparison, we identified seven regions of interest, based on their strong evolutionary conservation, which is often sufficient to identify regulatory elements (Prabhakar et al. 2006). All regions were selected based on their high evolutionary conservation, with two such highly conserved regions also noted as being in close proximity to the most significant GWAS SNPs for schizophrenia at this locus (rs1625579 and rs1198588) (Figure 3.1b). The ECRs were labelled 1 to 7, numbered from within the MIR137 transcript and increasing in number with increasing distance upstream of MIR137. Of the seven selected ECRs, ECR 3 and 7 were the most highly preserved, with conservation back to the frog genome. ECRs 1, 2, 4, and 6 were conserved back to the chicken genome, and ECR 5 was conserved to the opossum.

All ECRs studied in this chapter were defined through bioinformatic analysis of conservation data prior to any further study. All ECRs were selected on the basis of high evolutionary conservation, and no further elements were selected or discarded as the study progressed. All ECRs selected at the beginning of this study were

analysed in the same way, and no elements were specifically selected for further study over others. Where data may appear to be missing for certain elements, it will be noted that this is because no data was available, or no results were found.

3.3.2 Schizophrenia GWAS SNPs are within or adjacent to five of the seven ECRs, and ECRs 1 to 5 form a haplotype block.

After identification of our ECRs of interest, we used the UCSC Genome Browser to overlay the PGC2 schizophrenia GWAS data onto the human genome to allow for finer and more flexible observation of the data set compared to the basic image viewing provided by the Ricopili tool.

The UCSC Genome Browser is a web based tool hosted by the University of California, Santa Cruz, which enables access to, and interactive visual browsing of, a vast number of data sets aligned across the genomes of many different species. In particular, for this work, we made use of the browser's ability to enable viewing of gene loci in the human genome, overlaid with data on evolutionary conservation, as well as the locations of CpG islands, DNase I hypersensitivity sites, active histone modifications, and transcription factor binding; all of which may indicate active regulatory regions. In addition to the many data sets available to view on the UCSC Genome Browser, a key advantage is the ability to upload external data sets in a manner that allows visualisation of this data alongside other data tracks provided by the tool. This enabled us to directly view and compare the PGC2 schizophrenia GWAS data with evolutionary conservation across the region, which was not possible with the ECR Browser or Ricopili tools (see Section 2.2.9.2).

Our selected ECRs span a region of 96 kb, beginning in the third intron of MIR137HG and spanning 80 kb upstream. We find 80 statistically significant schizophrenia GWAS

SNPs across this 96 kb area in the PGC2 data set. It is likely that many of the statistically significant SNPs across this region are in LD with the true causally associated SNPs and therefore may not contribute to risk individually. One method to try to identify causally associated risk SNPs is to look for variants with the potential to alter processes such as regulatory activity or transcription factor binding. For this reason, we narrowed down this analysis to focus on SNPs that fell within regions of high evolutionary conservation, as variation in such regions could have the potential to alter conserved regulatory function. Of the 80 GWAS SNPs across the 96 kb region at the MIR137 locus, we find three within the ECR 3 sequence, one within the ECR 7 sequence, and an additional seven GWAS SNPs adjacent to ECRs 1, 2, and 5 (Figure 3.2).

In order to assess linkage disequilibrium (LD) across the MIR137 locus ECRs, we utilised the now retired International HapMap Project resource to access the most recent Phase 3 data on SNP genotype frequency in the CEU cohort, which comprised genotype data from 165 Utah residents with Northern and Western European ancestry. The genotype data for all SNPs across the MIR137 region of interest (chr1:98,498,912–98,595,043) was downloaded from the International HapMap website, which was hosted by the National Centre for Biotechnology Information (NCBI). The data was then loaded into the HaploView software (Section 2.2.9.4), which was developed at, and is hosted by, the Broad Institute.

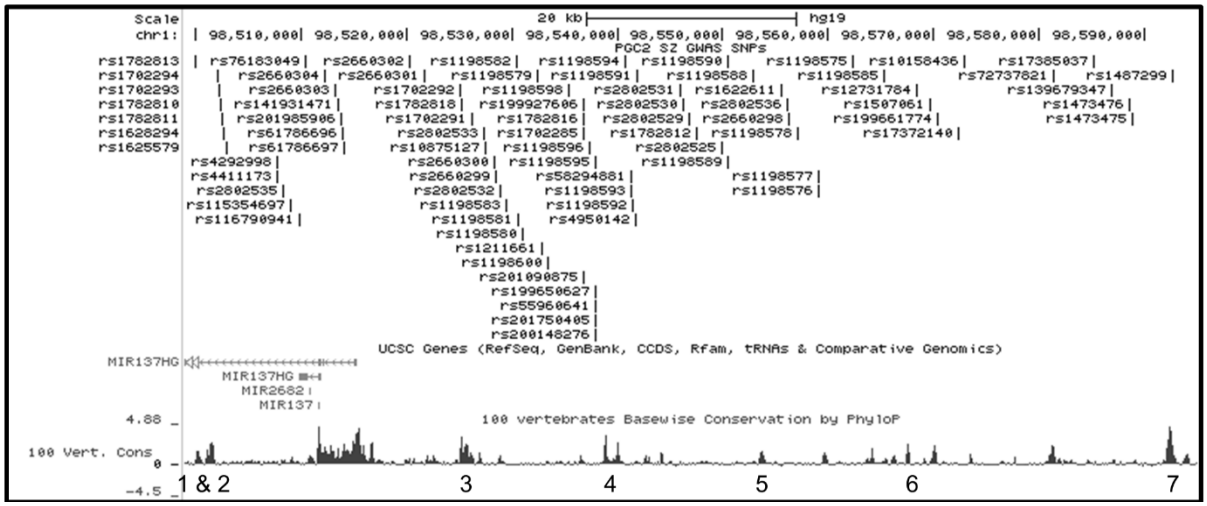


Figure 3.2 Visualisation of schizophrenia GWAS SNPs across MIR137 and the upstream region containing selected ECRs.

Schizophrenia GWAS data from the PGC2 was uploaded onto the UCSC Genome Browser and overlaid onto annotated genes. A schematic of MIR137 and its upstream region containing the ECRs of interest demonstrates 80 schizophrenia GWAS SNPs across this region. ECRs are numbered 1–7 and visualised as peaks on the 100 vertebrates basewise conservation track from the UCSC Genome Browser which compares the genomes of 100 vertebrate species to identify regions of high conservation.

The HaploView software provides a simple interface to run LD analysis on large data sets downloaded from the HapMap project and provides a graphical output in the form of LD maps, which give a visual representation of the SNPs and regions that are in LD with each other across the genomic input area.

Upon loading the MIR137 locus SNP data into HaploView, we selected 21 SNPs, including nine schizophrenia GWAS SNPs, within and adjacent to the seven ECRs of interest for analysis. Visualising the D' value for LD across this region using the default software parameters and the default haplotype block definition (Gabriel et al. 2002) showed that seven of the nine schizophrenia GWAS SNPs used in this analysis were in a haplotype block linking the MIR137 ECRs 1, 2, 3, and 5 over a region of ~54.9 kb (chr1:98,499,795–98,554,659) (Figure 3.3). The D' value is a normalised version of the D statistic (coefficient of LD), both of which are described in further detail in Section 2.2.9.4. A D' value equal to or greater than 0.8 signifies strong LD, with a D' of 1 indicating complete LD.

From this analysis, we also find that the schizophrenia GWAS SNPs within the MIR137 locus ECRs (excluding ECR 7) are preferentially in complete or high LD with each other. In contrast to this, the non-GWAS SNPs in the same regions show significantly lower LD with each other. This may suggest that the schizophrenia-associated SNPs across the MIR137 locus ECRs could function together, influencing risk through a shared or combined mechanism. However, this may also suggest that one SNP in this set is a functional risk allele and the remainder are statistically significant due to being in strong LD.

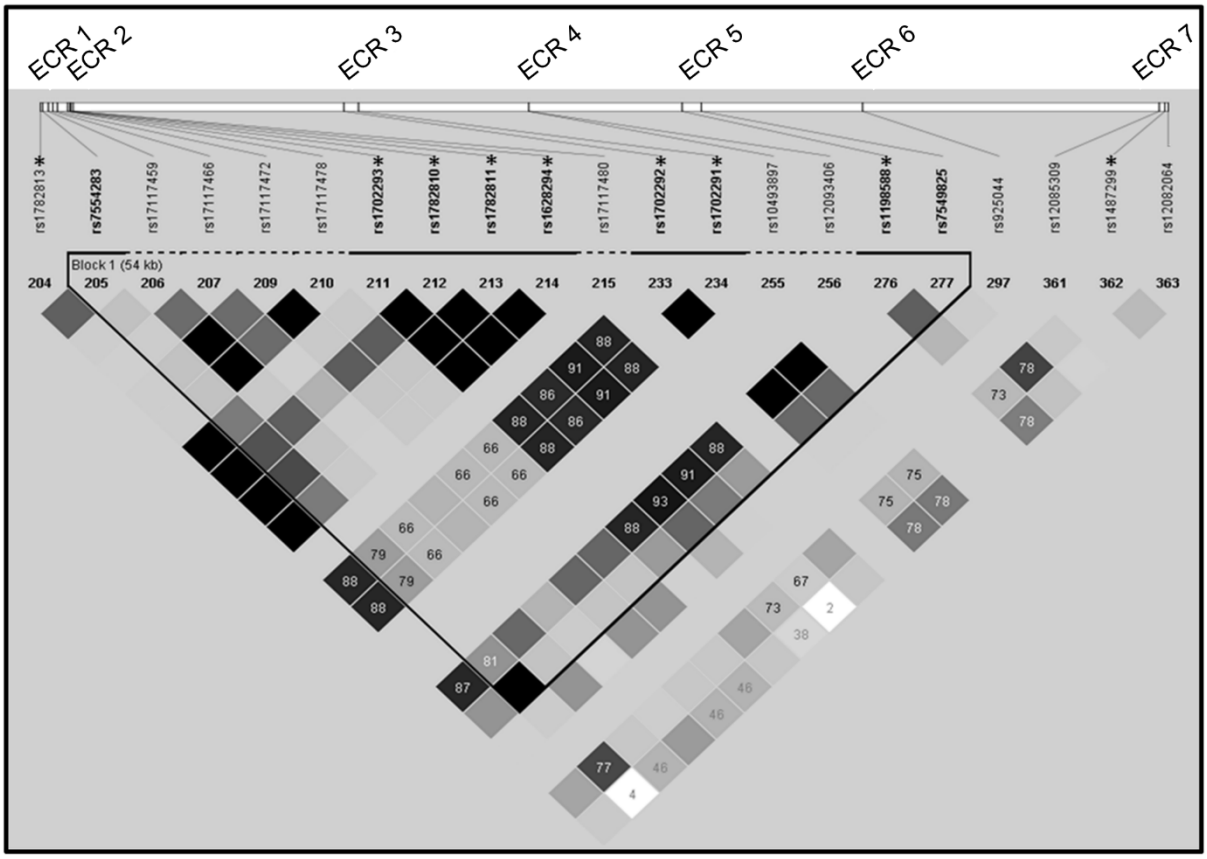


Figure 3.3 Linkage disequilibrium map of SNPs within or around ECRs at the MIR137 locus.

LD analysis of SNPs at the MIR137 ECRs was carried out using SNP genotype data from the HapMap CEU cohort, with D' values shown. Data for 9 of the schizophrenia GWAS SNPs (starred) was available in this cohort, and analysed alongside data for non-GWAS HapMap SNPs within the same ECRs. In this analysis, a haplotype block was found to link SNPs from ECRs 1–5, with schizophrenia GWAS SNPs at ECR 1, 2, 3 and 5 in strong linkage disequilibrium with each other.*

3.3.3 Bioinformatic analysis of MIR137 ECR function in *ex vivo* samples

In order to begin characterising the potential functions of each ECR, we used HaploReg v4.1. HaploReg is a tool from the Broad Institute which allows the viewing of chromatin state and histone modification data from a wide range of human tissues and cell lines across the region of an input SNP, thereby facilitating the drawing of conclusions as to the function of such regions in tissue or in cell lines.

By inputting SNPs within or adjacent to the MIR137 ECRs, we were able to view such data across our ECRs of interest. H3K4me1 and H3K27ac histone marks are indicative of active enhancers, whereas H3K4me3 and H3K9ac are indicative of transcriptionally active promoters. This data showed H3K4me1, H3K4me3, and H3K27ac histone modifications across MIR137 ECRs 1 and 2 (input SNPs rs7554283 and rs17117472, respectively) in all 10 brain regions tested (Figure 3.4a, b, and Supplementary Data 3.1), including the strongest data suggesting enhancer and potential promoter activity around ECRs 1 and 2 in the hippocampus, inferior temporal lobe, angular gyrus, and the foetal brain.

We also identify histone marks suggestive of enhancer and promoter activity in the NHA (Normal Human Astrocyte) primary cell line, as well as H3K4me1 modifications around ECR 1 and 2 in primary cultured neurospheres derived from the cortex and ganglion eminence.

Figure 3.4 Chromatin state and histone modifications at MIR137 ECRs 1, 2, and 6.

(A) ECR 1

Description	Chromatin states (15-state model)	Chromatin states (25-state model)	H3K4me1	H3K4me3	H3K27ac	H3K9ac
HUES64 Cells			H3K4me1_Enh			
H1 Derived Neuronal Progenitor Cultured Cells			H3K4me1_Enh			
hESC Derived CD56+ Mesoderm Cultured Cells		18_EnhAc				
hESC Derived CD56+ Ectoderm Cultured Cells			H3K4me1_Enh			
Brain Hippocampus Middle	7_Enh	16_EnhW1	H3K4me1_Enh	H3K4me3_Pro	H3K27ac_Enh	
Brain Substantia Nigra		16_EnhW1	H3K4me1_Enh			
Brain Anterior Caudate		16_EnhW1	H3K4me1_Enh			
Brain Cingulate Gyrus	7_Enh	16_EnhW1	H3K4me1_Enh	H3K4me3_Pro		
Brain Inferior Temporal Lobe		16_EnhW1		H3K4me3_Pro	H3K27ac_Enh	
Brain Angular Gyrus	7_Enh	16_EnhW1	H3K4me1_Enh	H3K4me3_Pro	H3K27ac_Enh	H3K9ac_Pro
Brain_Dorsolateral_Prefrontal_Cortex	7_Enh	16_EnhW1	H3K4me1_Enh	H3K4me3_Pro	H3K27ac_Enh	
Brain Germinal Matrix		17_EnhW2	H3K4me1_Enh			
Fetal Brain Female	2_TssAFlnk	13_EnhA1	H3K4me1_Enh	H3K4me3_Pro		
Fetal Brain Male	7_Enh	2_PromU	H3K4me1_Enh			

(B) ECR 2

Description	Chromatin states (15-state model)	Chromatin states (25-state model)	H3K4me1	H3K4me3	H3K27ac	H3K9ac
Brain Hippocampus Middle		22_PromP	H3K4me1_Enh	H3K4me3_Pro		
Brain Substantia Nigra			H3K4me1_Enh			
Brain Anterior Caudate		17_EnhW2	H3K4me1_Enh			
Brain Cingulate Gyrus			H3K4me1_Enh			
Brain Inferior Temporal Lobe		17_EnhW2		H3K4me3_Pro	H3K27ac_Enh	
Brain Angular Gyrus		17_EnhW2	H3K4me1_Enh	H3K4me3_Pro	H3K27ac_Enh	H3K9ac_Pro
Brain_Dorsolateral_Prefrontal_Cortex	7_Enh				H3K27ac_Enh	
Brain Germinal Matrix	7_Enh		H3K4me1_Enh			
Fetal Brain Female	2_TssAFlnk	14_EnhA2	H3K4me1_Enh	H3K4me3_Pro		
Fetal Brain Male	7_Enh	22_PromP	H3K4me1_Enh	H3K4me3_Pro		

(C) ECR 6

Description	Chromatin states (15-state model)	Chromatin states (25-state model)	H3K4me1	H3K4me3	H3K27ac	H3K9ac
Brain Hippocampus Middle			H3K4me1_Enh			
Brain Cingulate Gyrus			H3K4me1_Enh			
Brain Germinal Matrix		17_EnhW2	H3K4me1_Enh			
Fetal Brain Female	7_Enh		H3K4me1_Enh			
Fetal Brain Male			H3K4me1_Enh			

Key for chromatin state and histone modification data:

Chromatin states: Enh = enhancer, EnhA1/2 = active enhancer 1 or 2, EnhAc = enhancer acetylation only, EnhAF = active enhancer flank, EnhW1/2 = weak enhancer 1 or 2, PromP = poised promoter, PromU = promoter upstream transcriptional start site, TssA = active transcription start site, TssAFlank = flanking active transcriptional start site.

Histone modifications: Enh = enhancer, Pro = promoter, Black = no available data.

Yellow = Enhancer (Enh), weak enhancer (EnhW1, EnhW2). Orange = Active enhancer (Enh), flanking active enhancer (EnhAF). Red = Active transcriptional start site (TssA), flanking active transcriptional start site (TssAFlnk), promoter (Pro), promoter upstream transcriptional start site (PromU). Pink = Poised promoter (PromP).

Figure 3.4 Chromatin state and histone modifications at MIR137 ECRs 1, 2, and 6.

Chromatin state and histone modification data from the Roadmap Epigenomics Consortium shows that ECRs 1, 2, and 6 are active regulators of transcription in many areas of the human brain. In particular, as well as being regulatory domains across the brain, data on H3K4me3 histone modifications for ECRs 1 and 2 suggests promoter activity around this area in numerous brain regions, including the hippocampus, inferior temporal lobe, angular gyrus, and in the foetal brain. Data for ECR 6 does not suggest promoter activity, though this region displays the hallmarks of transcriptional regulatory activity in a range of brain areas including the hippocampus, cingulate gyrus, and foetal brain.

We see no evidence of regulatory activity at these ECRs in embryonic or neural progenitor cells. Such data therefore provides evidence supporting MIR137 ECRs 1 and 2 as active enhancers and/or promoters in multiple regions of the adult brain.

Similarly, MIR137 ECR 6 (input SNP rs925044) shows H3K4me1 histone marks at a number of brain regions including the hippocampus, cingulate gyrus, germinal matrix, and across the foetal brain. This pattern of histone methylation is consistent with MIR137 ECR 6 acting similarly to ECRs 1 and 2, being an active transcriptional regulator in multiple areas of the brain (Figure 3.4c). Conversely, analysis of data across MIR137 ECRs 3, 4, and 5 (input SNPs rs1702292, rs72969673, and rs114522393, respectively) shows little to no data within the brain, with only ECR 4 displaying regulatory activity (H3K27ac) in the anterior caudate. Instead, data returned for these ECRs shows activity in multiple human embryonic stem cell lines (ESCs), human iPSCs, and stem cell derived neuronal progenitor cells (Figure 3.5). The predominant histone modification across ECRs 3 and 4 is H3K4me1, suggesting that these ECRs are enhancers in embryonic cells and in neuronal progenitor cells (Figure 3.5a, b). In addition to H3K4me1 histone modifications, H3K27ac and H3K9ac marks are also seen across ECR 5 in a greater number of embryonic and iPSC lines compared to ECRs 3 and 4, suggesting a role for ECR 5 not only in transcriptional regulation, but also suggesting active transcription from a nearby promoter in embryonic cells and neuronal progenitors during development (Figure 3.5c). No relevant data was seen over the MIR137 ECR 7 locus.

Figure 3.5 Chromatin state and histone modifications at MIR137 ECRs 3, 4, and 5.

(A) ECR 3

Description	Chromatin states (15-state model)	Chromatin states (25-state model)	H3K4me1	H3K4me3	H3K27ac	H3K9ac
ES-WA7 Cells	7_Enh					
ES-I3 Cells	7_Enh		H3K4me1_Enh			
HUES6 Cells			H3K4me1_Enh			
HUES48 Cells	7_Enh		H3K4me1_Enh			
HUES64 Cells			H3K4me1_Enh			
iPS-15b Cells			H3K4me1_Enh			
H1 Derived Neuronal Progenitor Cultured Cells			H3K4me1_Enh			
H9 Derived Neuronal Progenitor Cultured Cells	7_Enh		H3K4me1_Enh			
H9 Derived Neuron Cultured Cells			H3K4me1_Enh			
hESC Derived CD56+ Mesoderm Cultured Cells	7_Enh		H3K4me1_Enh			

(B) ECR 4

Description	Chromatin states (15-state model)	Chromatin states (25-state model)	H3K4me1	H3K4me3	H3K27ac	H3K9ac
ES-WA7 Cells			H3K4me1_Enh			
ES-I3 Cells	7_Enh		H3K4me1_Enh			
HUES48 Cells			H3K4me1_Enh			
HUES64 Cells			H3K4me1_Enh			
iPS-15b Cells			H3K4me1_Enh			
Brain Anterior Caudate					H3K27ac_Enh	

(C) ECR 5

Description	Chromatin states (15-state model)	Chromatin states (25-state model)	H3K4me1	H3K4me3	H3K27ac	H3K9ac
ES-I3 Cells	7_Enh		H3K4me1_Enh			
HUES6 Cells		15_EnhAF			H3K27ac_Enh	
HUES48 Cells		15_EnhAF	H3K4me1_Enh		H3K27ac_Enh	H3K9ac_Pro
HUES64 Cells	7_Enh	15_EnhAF	H3K4me1_Enh		H3K27ac_Enh	H3K9ac_Pro
H1 Cells					H3K27ac_Enh	
ES-UCSF4 Cells			H3K4me1_Enh			
iPS-20b Cells			H3K4me1_Enh	H3K4me3_Pro		H3K9ac_Pro
iPS-18 Cells	7_Enh	18_EnhAc	H3K4me1_Enh			H3K9ac_Pro
iPS-15b Cells			H3K4me1_Enh			
H9 Derived Neuron Cultured Cells			H3K4me1_Enh			
hESC Derived CD56+ Mesoderm Cultured Cells			H3K4me1_Enh		H3K27ac_Enh	
hESC Derived CD56+ Ectoderm Cultured Cells					H3K27ac_Enh	

Key for chromatin state and histone modification data:

Chromatin states: Enh = enhancer, EnhA1/2 = active enhancer 1 or 2, EnhAc = enhancer acetylation only, EnhAF = active. Histone modifications: Enh = enhancer, Pro = promoter, Black = no available data.

Yellow = Enhancer (Enh), weak enhancer (EnhW1, EnhW2). Orange = Active enhancer (Enh), flanking active enhancer (EnhAF). Red = Active transcriptional start site (TssA), flanking active transcriptional start site (TssAFlnk), promoter (Pro), promoter upstream transcriptional start site (PromU). Pink = Poised promoter (PromP).

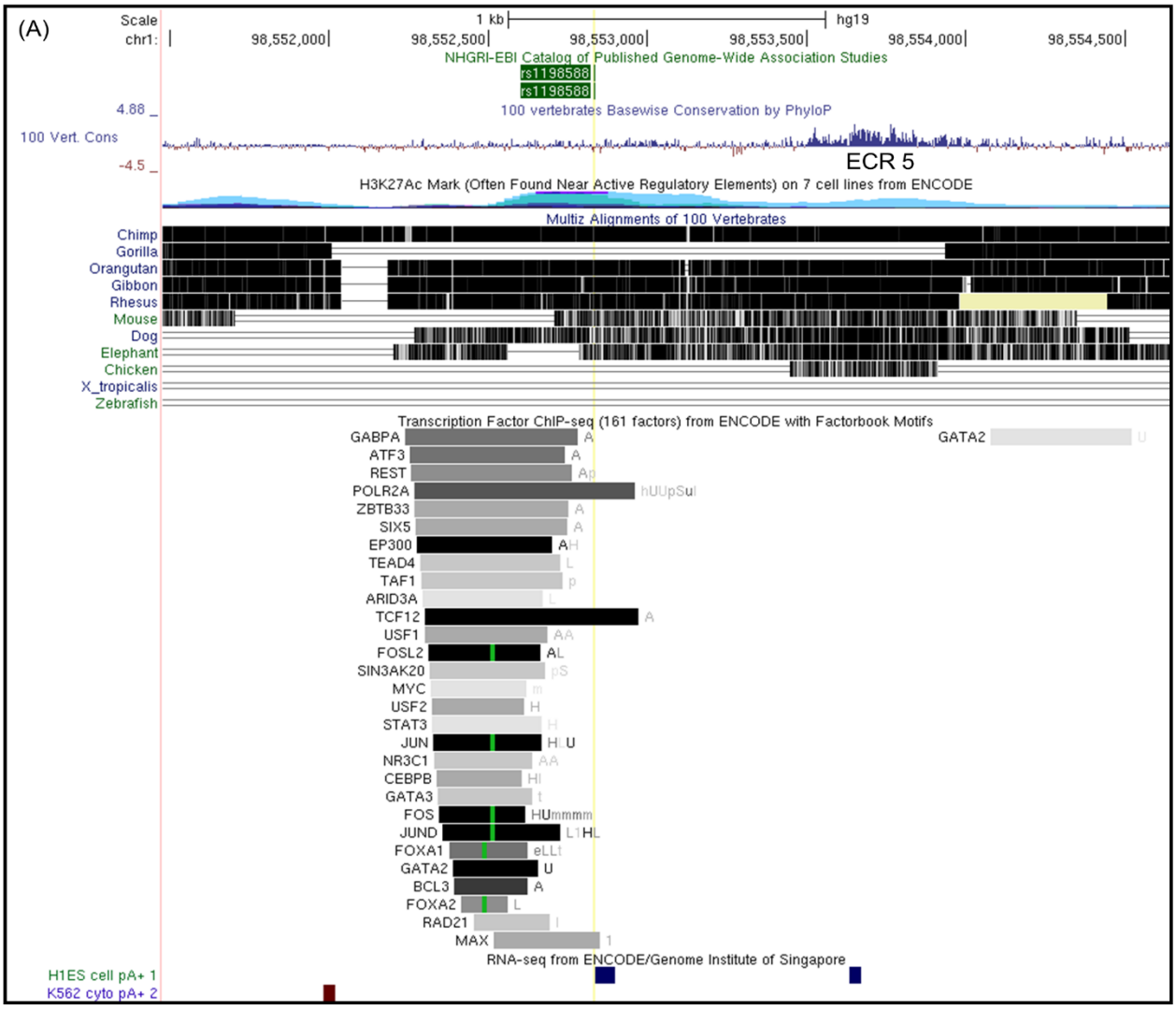
Figure 3.5 Chromatin state and histone modifications at MIR137 ECRs 3, 4, and 5.

Chromatin state and histone modification data from the Roadmap Epigenomics Consortium shows that ECRs 3 and 4 are active regulators of transcription in multiple ESC lines, with evidence for ECR 3 (a) activity in neuronal progenitor cells and stem cell derived neurons, as well as six other stem cell lines, and ECR 4 (b) activity in five stem cell lines and in the anterior caudate in the adult brain. ECR 5 (c) shows clear chromatin state and histone modification data to suggest regulatory activity in numerous stem cell lines, including stem cell derived neurons, with evidence of promoter activity at this ECR in four of the 13 stem cell lines with available data.

While the main active histone signals are centred around activity in the brain or during embryonic development, we also find histone modification data for multiple ECRs as active enhancers or promoters in mesenchymal stem cells derived from multiple tissues, and data suggesting enhancer activity in multiple muscle cell types, breast mammary epithelial cells, as well as some fibroblasts and keratinocytes (Supplementary Data 3.1). Regulatory activity at this locus in mammary cells may be of interest due to the known role of MIR137 in breast cancer (Zhao et al. 2012, Ying, Sun and He 2017, Lee et al. 2015).

As ECR 5 was adjacent to one of the most significant GWAS SNPs for schizophrenia at this locus, rs1198588, we decided to utilise HaploReg in an attempt to gain further information on the functional significance of the highly associated SNP at this region. It was found that rs1198588 overlaps a region of active histones and the binding of 29 transcription factors according to ENCODE ChIP-seq data on the UCSC Genome Browser (Figure 3.6a). This included factors such as RNA Polymerase II, which was found to bind at this region in seven different cell lines including the brain derived SK-N-SH and U87 lines (data not shown), suggesting potential transcription from this locus in the brain. Further interrogation of data available on the UCSC Genome Browser demonstrated that RNA-seq data from the embryonic H1ES cell line and the K562 leukaemia cell line identified expression at this locus. In addition to this, chromatin state and histone modification data across the GWAS SNP from HaploReg demonstrated strong evidence of regulatory activity across a wide range of stem cell lines, including likely promoter activity in six stem cell lines. We further note chromatin state and histone modification data suggesting regulatory activity at this region in the hippocampus, and poised promoter status in the anterior caudate (Figure 3.6b).

Figure 3.6 Evidence of potential transcriptional activity around the schizophrenia GWAS SNP, rs1198588.



(B)

Description	Chromatin states (15-state model)	Chromatin states (25-state model)	H3K4me1	H3K4me3	H3K27ac	H3K9ac
ES-WA7 Cells		14_EnhA2	H3K4me1_Enh	H3K4me3_Pro		
H9 Cells		18_EnhAc			H3K27ac_Enh	
ES-I3 Cells	7_Enh	14_EnhA2	H3K4me1_Enh			
HUES6 Cells	7_Enh	15_EnhAF	H3K4me1_Enh		H3K27ac_Enh	H3K9ac_Pro
HUES48 Cells	7_Enh	13_EnhA1	H3K4me1_Enh		H3K27ac_Enh	H3K9ac_Pro
HUES64 Cells	7_Enh	13_EnhA1	H3K4me1_Enh		H3K27ac_Enh	H3K9ac_Pro
H1 Cells		15_EnhAF	H3K4me1_Enh		H3K27ac_Enh	
ES-UCSF4 Cells	7_Enh	18_EnhAc	H3K4me1_Enh			
iPS-20b Cells	7_Enh	14_EnhA2	H3K4me1_Enh	H3K4me3_Pro		H3K9ac_Pro
iPS-18 Cells	7_Enh	14_EnhA2	H3K4me1_Enh		H3K27ac_Enh	H3K9ac_Pro
iPS-15b Cells	7_Enh	18_EnhAc	H3K4me1_Enh			
iPS DF 6.9 Cells		18_EnhAc	H3K4me1_Enh			
iPS DF 19.11 Cells		18_EnhAc			H3K27ac_Enh	
H1 Derived Neuronal Progenitor Cultured Cells		17_EnhW2	H3K4me1_Enh		H3K27ac_Enh	
H9 Derived Neuronal Progenitor Cultured Cells	7_Enh	18_EnhAc	H3K4me1_Enh			
H9 Derived Neuron Cultured Cells	7_Enh	14_EnhA2	H3K4me1_Enh			
hESC Derived CD56+ Mesoderm Cultured Cells	7_Enh	13_EnhA1	H3K4me1_Enh		H3K27ac_Enh	
hESC Derived CD56+ Ectoderm Cultured Cells		15_EnhAF			H3K27ac_Enh	
hESC Derived CD184+ Endoderm Cultured Cells		15_EnhAF			H3K27ac_Enh	
H1 BMP4 Derived Mesendoderm Cultured Cells	7_Enh	15_EnhAF			H3K27ac_Enh	
H1 BMP4 Derived Trophoblast Cultured Cells					H3K27ac_Enh	
H1 Derived Mesenchymal Stem Cells		15_EnhAF			H3K27ac_Enh	
Brain Hippocampus Middle		16_EnhW1	H3K4me1_Enh			
Brain Anterior Caudate		22_PromP	H3K4me1_Enh			

Figure 3.6 Evidence of potential transcriptional activity around the schizophrenia GWAS SNP, rs1198588.

(a) Visualisation of the region encompassing rs1198588 and ECR 5 demonstrates active histones and the binding of 29 transcription factors, including RNA Polymerase II, according to ENCODE ChIP-seq data. This would be suggestive of transcription from this region, with ENCODE RNA-seq data further suggesting expression around this region in the H1ES human embryonic stem cell line and the K562 human leukaemia cell line. The schizophrenia GWAS SNP at this locus lies within the peak of active histone signal and overlaps the ENCODE ChIP-seq data for a number of transcription factors, suggesting that this SNP may have the potential to alter regulation at this region.

(b) Chromatin state and histone modification data from HaploReg across rs1198588 provides strong evidence for regulatory activity around this SNP in 22 embryonic stem cell lines and stem cell-derived lines, as well as in the hippocampus and anterior caudate. This data further suggests promoter activity around the region of the SNP in six stem cell lines, which would support the RNA-seq and transcription factor binding data in (a).

Key for chromatin state and histone modification data:

Chromatin states: Enh = enhancer, EnhA1/2 = active enhancer 1 or 2, EnhAc = enhancer acetylation only, EnhAF = active enhancer flank, EnhW1/2 = weak enhancer 1 or 2, PromP = poised promoter.

Histone modifications: Enh = enhancer, Pro = promoter, Black = no available data.

Yellow = Enhancer (Enh), weak enhancer (EnhW1, EnhW2). Orange = Active enhancer (Enh), flanking active enhancer (EnhAF). Red = Active transcriptional start site (TssA), flanking active transcriptional start site (TssAFlnk), promoter (Pro), promoter upstream transcriptional start site (PromU). Pink = Poised promoter (PromP).

This would support a role for the region encompassing ECR 5 and the schizophrenia GWAS SNP, rs1198588, in driving the expression or regulation of transcripts at this locus in embryonic development and neural progenitor cells, as well as regions of the adult brain including the hippocampus and anterior caudate. Such data may suggest that rs1198588 could be contributing to risk through the alteration of potential expression from this locus, in a similar manner to that which has been demonstrated for the rs2660403 schizophrenia GWAS SNP at the MIR137 internal promoter (Warburton et al. 2015a). One way to test this hypothesis would be to synthetically create multiple reporter gene constructs with identical sequence except for alleles of the rs1198588 SNP and assay these for any change in reporter gene levels. This experiment could be carried out in both the pGL3-P and pGL3-B vector backbones in order to also test for promoter activity within this sequence.

Taken together, these data would suggest that we can split our model for ECR function at the MIR137 locus into two timepoints, with ECRs 3, 4, and 5 likely playing a role in transcriptional regulation during development in the embryo and early brain formation, with ECR 5 and the schizophrenia GWAS SNP, rs1198588, marking a potentially important region which could be active in early development and may have implications in schizophrenia. Alternatively, ECRs 1, 2, and 6 are likely to be functional in multiple regions of the adult brain. This data may therefore implicate regulation at the MIR137 locus in both the developmental and adult risk factors for schizophrenia, with the function of different ECRs perhaps being important in the CNS at different time points.

3.3.4 Transcriptional regulatory activity of MIR137 ECRs

We next sought to validate the potential regulatory activity of MIR137 ECRs 1 to 7 *in vitro* using the SH-SY5Y neuroblastoma cell line as a model. The seven selected

ECRs were cloned into the pGL3P luciferase reporter vector upstream of the minimal SV40 promoter, and potential transcriptional regulatory function was assessed through dual luciferase reporter assays using the GloMax luciferase luminometer to measure fluorescence. Expression from the pGL3P vector results in the production of firefly luciferase, and the pRL-TK Renillia luciferase vector was used as an internal control.

When compared to the baseline expression of luciferase from the unmodified pGL3P backbone, five of the seven ECRs (MIR137 ECR 1, 3, 4, 6, and 7) were found to have positive regulatory effects, with a modest but significant increase of between 1.25- to 1.75-fold luciferase expression. MIR137 ECRs 1 and 6 were found to be the strongest positive regulators in this analysis. Conversely, ECRs 2 and 5 were found to be negative regulators of reporter gene expression in the SH-SY5Y cell line, decreasing luciferase signal by approximately 0.5- and 0.8-fold, respectively (Figure 3.7). We note that chromatin state and histone modification data predicted enhancer activity of ECRs 2 and 5, while these regions were found to be negative regulators in the SH-SY5Y cell line model. This is likely due to cell line specific effects and as such, these regions may display different functional activity in different cell lines or in *ex vivo* samples such as in the data from HaploReg. Further, the regulatory activity observed in a cancer-derived cell line may not accurately reflect activity in a healthy cell.

No data was available over ECR 7 in the HaploReg data set, however reporter gene assays in this cell line model suggested that this ECR can act as a positive regulator of gene expression.

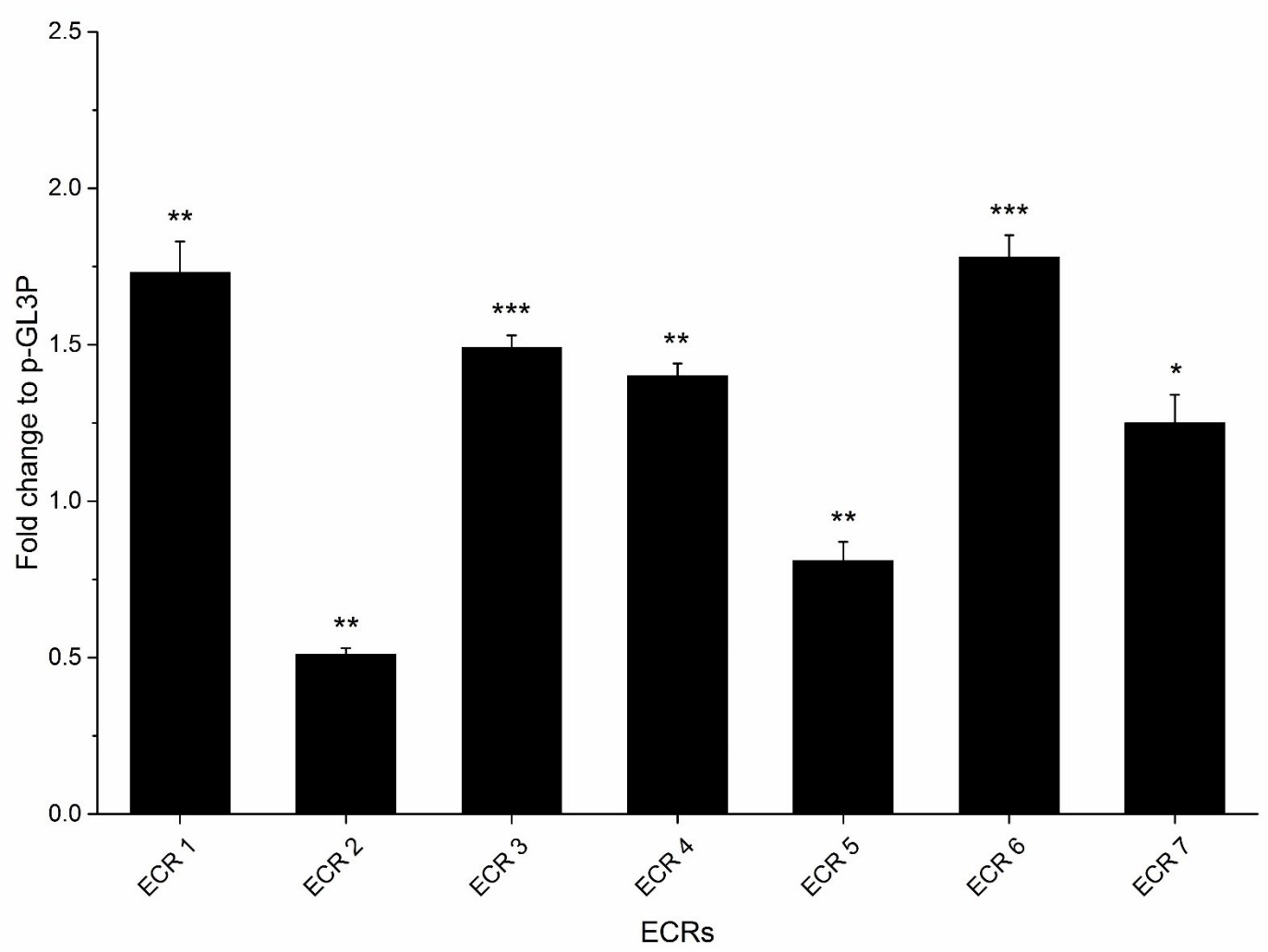


Figure 3.7 Transcriptional regulatory activity of MIR137 locus ECRs in a neuroblastoma cell line model.

*Regulatory function of ECR sequences was assessed by dual luciferase assay in the pGL3P reporter vector in SH-SY5Y neuroblastoma cells under basal conditions. All ECRs tested displayed regulatory properties in vitro, with five showing positive regulatory function and the remaining two displaying negative regulatory effects in this cell line. Error bars display the standard error for each ECR, calculated as the standard divided by number of luciferase assays (n = 4). *P < 0.1, **P < 0.01, ***P < 0.001..*

As the MIR137 locus is shown to be highly associated with schizophrenia through GWAS, new expressed sequence tags (ESTs) and RNAs from within this region warrant further study for their potential involvement in brain development and function. In this regard, further bioinformatic investigation of the MIR137 ECRs was conducted using GenBank data on human ESTs, histone modification data from HaploReg, and ENCODE ChIP-seq data on transcription factor binding.

Starting with the MIR137 ECR 1, data on ESTs from GenBank showed an uncharacterised transcript, AW901379, expressed approximately 525 bp adjacent to this ECR. Further data from GenBank characterised AW901379 as a 299 bp mRNA that was identified as expressed in normal nervous tissue (Figure 3.8a).

Further analysis of the HaploReg data over the MIR137 ECR 1 (Figure 3.4a), demonstrated H3K4me1 modifications at this region in human embryonic stem cells and derived neuronal progenitor cells, suggesting transcriptional regulatory activity at this ECR in both embryonic and brain development. We also see H3K4me1 histone marks at ECR 1 in eight of the ten brain regions with data available in HaploReg, suggesting that this site is an active regulator across many brain regions, with additional H3K4me3 data also suggesting transcription from an active promoter at this ECR in the hippocampus, cingulate gyrus, inferior temporal lobe, angular gyrus, dorsolateral prefrontal cortex, and foetal brain. Taken together, these data suggest that the MIR137 ECR 1 may act as both a regulator of MIR137, and potentially a promoter region or regulatory element for an as yet uncharacterised CNS-expressed transcript, AW901379, which sits within the third intron of MIR137HG.

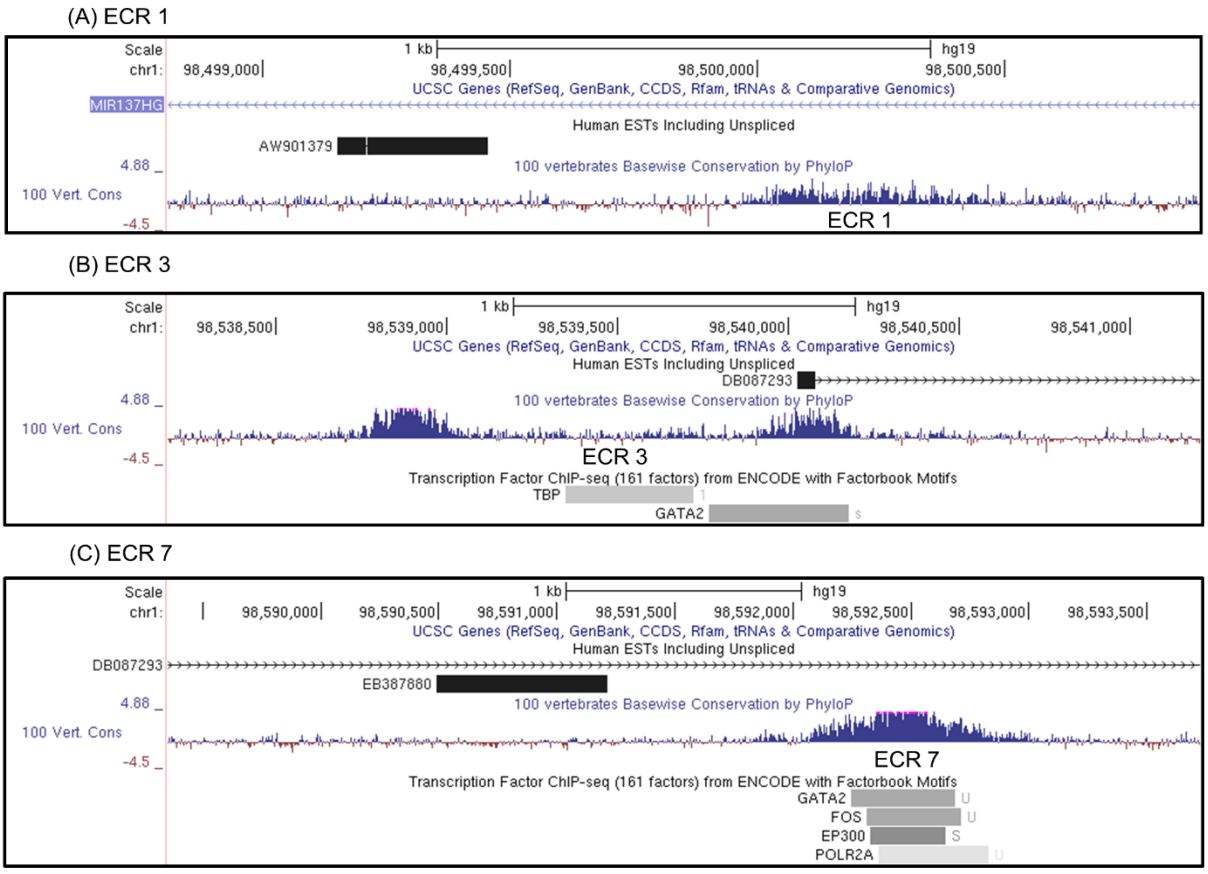


Figure 3.8 Uncharacterised transcripts at the MIR137 locus that may be regulated by ECRs.

(a) Schematic of ECR 1 and its flanking region, showing the human EST, AW901379. GenBank lists this EST as being identified in nervous tissue, and HaploReg data from Figure 3.4a showed histone modifications across this ECR that would be indicative of active transcription from this locus in the human brain, suggesting a potential CNS-expressed transcript originating adjacent to, and potentially regulated by, ECR 1.

(b) As well as acting as a transcriptional regulator, particularly in embryonic cell lines, the second peak of conservation at ECR 3 was found to potentially code for an exon in an uncharcterised testes-expressed EST, DB087293. We also note ENCODE ChIP-seq data demonstrating TATA box binding protein at this region, which may suggest expression from this ECR.

(c) Finally, binding of multiple transcription factors including RNA Polymerase II at ECR 7 suggested transcription from this region. We identify EB307880, an optic nerve-expressed EST beginning approximately 1 kb from ECR 7.

We find further evidence of potential transcripts around ECRs 3 and 7, with a testis expressed EST, DB087293, beginning within MIR137 ECR 3 and using part of this ECR as the first exon (Figure 3.8b). Accompanying data around ECR 3 lends support to transcription from this region, with ENCODE ChIP-seq data showing binding of the transcription factor TBP (TATA-box binding protein), the binding of which is the initial step in the formation of the RNA Polymerase II pre-initiation complex (PIC) which is necessary for transcription. This may suggest an alternative hypothesis that part of ECR 3 is conserved due to coding for an exon in the DB087293 transcript, however no evidence of an equivalent exon at this region was found in other species with transcript data available on the UCSC Genome Browser.

We also note EB307880, an EST which has been identified as expressed in the optic nerve, beginning approximately 1 kb from ECR 7, with ENCODE ChIP-seq data showing transcription factors including RNA Polymerase II binding across the ECR, suggesting transcription from this region (Figure 3.8c). The identification of an optic nerve-expressed EST at this locus may be of interest due to the known role of the MIR137 locus in uveal melanoma (Chen et al. 2011a). Regions around ECRs 2, 4, 5, and 6 were also examined bioinformatically for evidence of additional mRNA transcripts, however, no data was found to suggest the presence of additional mRNAs.

With the exception of DB087293, which overlaps part of the ECR 3 sequence, the exons of the other ESTs noted in this section are significantly less conserved than their adjacent ECRs. This could suggest that, if such ESTs were regulated by their adjacent ECRs, it is likely that the ECRs were already functional regulators before the evolution of nearby ESTs, which may have co-opted the ECRs as regulators at a later time.

Further work such as RT-PCR or RNA-seq in the relevant tissues may be able to confirm the presence of transcripts expressed from or adjacent to the ECRs in this chapter. In order to test for potential promoter function, the ECRs could be cloned into the pGL3B basic vector and their effect on a reporter gene assayed. In addition, if expression of these RNAs was able to be reliably identified, it may be of interest to use knock-down or knock-out techniques to begin to identify potential RNA function, and to characterise the effect of loss or mutation of these ECRs on the expression of RNAs across this locus.

3.4 Discussion

Data from genome-wide association studies has demonstrated that approximately 88% of common single nucleotide risk factors for complex disease are found in non-coding regions of the genome (Hindorff et al. 2009). The same is true for mental health GWAS, with numerous regions of non-coding DNA across the genome having been linked to genetic susceptibility to schizophrenia (Ripke et al. 2013). As GWAS continues to identify non-coding variants for disease risk, it has been proposed that the next major challenge in the post-GWAS era lies in deciphering the mechanisms which link non-coding risk variants to disease (Tak and Farnham 2015, Sadee et al. 2014, Huang 2015, Zhang and Lupski 2015). It is likely that many of the associated variants identified through GWAS are highlighting non-coding regulatory elements or mechanisms that modulate the tissue-specific and stimulus inducible regulation of gene expression or RNA processing (Quinn et al. 2013), rather than changes to coding regions which would alter protein conformation or function. This would be consistent with the fact that many psychiatric conditions, including schizophrenia, are episodic in nature, thereby suggesting that the molecular underpinnings of such conditions are perhaps more likely to involve changes in the way an individual responds at the cellular

level to environmental challenge, rather than the more permanent effects of, for example, physically altered genes or proteins. In support of this, Hoffmann et al. demonstrate that a number of non-coding regions identified in schizophrenia GWAS are in fact regulatory domains which are foetal methylation quantitative trait loci (meQTLs), meaning that epigenetic changes across these loci in response to the environment may contribute to the neurodevelopmental risk for schizophrenia (Hoffmann, Ziller and Spengler 2016).

Meta-analyses of schizophrenia GWAS data have repeatedly highlighted MIR137 as one of the most highly associated genes for schizophrenia risk, with the GWAS signal stretching across MIR137 and far into the flanking up- and downstream regions (chr1:98,298,371-98,595,000, GRCh37/hg19) (Ripke et al. 2013, Schizophrenia Psychiatric Genome-Wide Association Study 2011). With the most highly significant schizophrenia GWAS SNPs residing either within MIR137 introns (eg. rs1625579) or up to 38 kb upstream (eg. rs1198588) (Figure 3.1a), this is may suggest the involvement of non-coding regulatory elements both within and upstream of the MIR137 transcripts. Such regions may be experimentally identified in a similar manner to work by Duan et al., who sequenced the MIR137 locus to identify rare SNPs in non-coding functional regions which may contribute to schizophrenia risk. After identifying rare SNPs in schizophrenia samples, Duan's methods then involved cross-referencing these SNPs with databases such as ENCODE to identify nearby regulatory elements or predicted binding sites which may be altered by the SNP. Regulatory potential of the selected regions was carried out by luciferase assay for each allele, and allele-specific transcription factor affinity was validated by electrophoresis mobility shift assays (EMSAs). Finally, Chromatin Conformation Capture (3C) was performed to

identify whether the selected non-coding regions physically interacted with the MIR137 core promoter or with other known regulatory regions (Duan et al. 2014).

Others in the group have previously used comparative genetics overlaid with GWAS data to begin elucidating the mechanism linking GWAS variants and disease risk, identifying regions of high evolutionary conservation which act as transcriptional regulators both *in vitro* and *in vivo* with relevance to psychiatric and CNS conditions (Davidson et al. 2011, Davidson et al. 2006, Hing et al. 2012, Khursheed et al. 2015). Here we use a similar approach, comparing seven vertebrate genomes against the human genome to identify regions of high conservation across MIR137 and its upstream region (chr1:98,448,000-98,595,000). Schizophrenia GWAS data was overlaid onto evolutionary conservation data to identify and prioritise the study of conserved regions with potential relevance to schizophrenia.

This analysis identified seven ECRs at the MIR137 locus, with two residing within the third intron of MIR137HG, and five upstream, with over 70% conservation back to the opossum, chicken, or frog genomes (Figure 3.1b). Overlaying schizophrenia GWAS data from the Psychiatric Genomics Consortium's PGC2 study revealed that the regions encompassing ECRs 1, 2, 3, 5, and 7 contained schizophrenia GWAS SNPs, which may suggest genotype-dependent modulation of the potential activity at these regions, similar to which has been previously shown at the MIR137 internal promoter (Warburton et al. 2015a).

Initial bioinformatic analysis of the chromatin state and histone modifications across each ECR in a range of tissues demonstrated a clear split in terms of the timepoint at which each ECR may be active. ECRs 1, 2, and 6 show chromatin state and histone modifications indicative of regulatory activity in numerous brain regions, including the

hippocampus, cingulate gyrus, foetal brain, and others (Figure 3.4). On the other hand, ECRs 3, 4, and 5 demonstrate regulatory activity primarily in embryonic stem cell lines (Figure 3.5), suggesting that they may be functional in regulating expression from this locus during early development, perhaps contributing to the neurodevelopmental risk for schizophrenia if environmental factors or schizophrenia GWAS SNPs alter regulation by these elements. Such data is useful in further delineating mechanisms associated with regulation across the MIR137 locus in the development of schizophrenia, whether through regulation by ECRs 3, 4, and 5 in early foetal development, or by ECRs 1, 2, and 6 in the adult brain.

Taking this further, we confirmed the regulatory activity of all seven ECRs through reporter gene assays in the SH-SY5Y neuroblastoma cell line. This demonstrated that five of the seven ECRs acted as positive regulators of expression, and two as negative regulators in this model (Figure 3.7), though this is likely to be cell- and environment-specific. This is in line with existing studies, which have demonstrated that similarly conserved regions which function *in vitro* in cell line models can also support tissue-specific reporter gene expression *in vivo* in both mouse models and in the chick embryo (Davidson et al. 2006, Davidson et al. 2011, Khursheed et al. 2015).

Finally, we note the presence of novel and uncharacterised transcripts around three of the ECRs characterised in this study (Figure 3.8). This may suggest that the transcriptional landscape around the MIR137 locus is more complex than is currently understood, and could point to additional roles for the ECRs characterised in this section as regulators or promoters for transcripts other than MIR137, which may act in a concerted manner to regulate multiple RNAs at this locus. Deeper analysis of uncharacterised transcripts at the MIR137 locus is of interest, as it may allow the

identification of new RNAs which could contribute to the schizophrenia risk across this region.

3.5 Summary

In summary, bioinformatic analysis of the schizophrenia-associated MIR137 locus identified seven highly conserved ECRs. The transcriptional regulatory activity of each ECR was confirmed through reporter gene assays in a neuroblastoma cell line, and chromatin state and histone modification data from multiple tissues was used to build up a picture of the activity of these regions in *ex vivo* samples. Data from the PGC schizophrenia GWAS studies shows schizophrenia-associated SNPs within or adjacent to multiple ECRs, which suggests their potential involvement in 'gene x environment' mechanisms in regulating expression at the MIR137 locus with relevance to schizophrenia biology. We noted additional transcriptional complexity at this locus, with a number of uncharacterised transcripts emanating from three of the ECRs in this study.

We conclude that seven highly conserved regions at the MIR137 locus display transcriptional regulatory activity in the SH-SY5Y neuroblastoma cell line which may have the potential to affect gene expression during development or in the adult brain. Luciferase assays for all ECRs that contained schizophrenia GWAS SNPs in this study were on genomic DNA containing the common, non-risk SNP. For each ECR either containing or adjacent to a schizophrenia GWAS SNP, this work should be extended to compare the effects of both SNP alleles on the regulatory potential of the ECR to determine any potential functional risk SNPs and to differentiate these from SNPs that may be statistically significant due to being in LD with true causal variants across the locus. Publicly available data sets should also be utilised to identify any potential eQTL SNPs within these ECRs, and to investigate potential tissue-specific eQTL effects. We

also note the presence of additional non-coding RNAs at this region, which require further study to validate, characterise, and to begin to describe their potential function.

Part II of this chapter extends our view of the MIR137 locus to address and characterise one such novel RNA, EU358092, which lies downstream of the MIR137HG transcript.

Part II: A novel brain-expressed RNA, EU358092, at the MIR137 locus.

3.6 Introduction

As discussed in the previous section, the majority of schizophrenia-associated SNPs at the MIR137 locus lie in non-coding regions (Ripke et al. 2013, Schizophrenia Psychiatric Genome-Wide Association Study 2011). In Part I of Chapter 3, we explored potential mechanisms linking risk SNPs to disease through focussing on highly conserved, non-coding regulatory elements as potential modulators of expression, which may contribute to schizophrenia biology when regulation through these elements is disrupted, for example by environmental challenge or by genetic variation. However, the identification of the miRNA, MIR137, as a major GWAS hit for schizophrenia also demonstrates the potential role for risk SNPs in highlighting the importance of ncRNAs.

Many ncRNAs, including miRNAs and lncRNAs (Section 1.1), are highly expressed in the brain and are known to play crucial roles in brain development, function, and in a wide range of CNS conditions (Cao, Li and Chan 2016, Hu and Li 2017, Briggs et al. 2015, Shi, Zhang and Qin 2017).

Key evidence for the general role of ncRNAs in neuropsychiatric conditions comes from the significantly increased risk of schizophrenia and other neurological and cognitive conditions in individuals with DiGeorge Syndrome, or 22q11.2 deletion syndrome (Hoeffding et al. 2017, Forstner et al. 2013, Tang et al. 2015). This condition is caused by a deletion at the 22q11.2 locus which includes the DGCR8 gene, a microprocessor complex subunit which is required for miRNA biogenesis, and is often used as a model for schizophrenia (Meechan et al. 2015).

Variants in the lncRNA, MIAT, are also known to add to the risk of paranoid schizophrenia in the Han Chinese population (Rao et al. 2015), and a number of studies using peripheral blood mononuclear cells (PBMCs) from individuals with a diagnosis of schizophrenia and controls have highlighted deregulation of multiple miRNAs (Camkurt et al. 2016, Jeffries et al. 2016) and up to 125 lncRNAs (Ren et al. 2015c, Chen et al. 2016b) in individuals with schizophrenia.

Here we focus on the two exon EU358092 transcript, which lies between MIR137 and its neighbouring gene, DPYD, at chr1:98,399,030-98,407,302. Initial observation of this region identified two ECRs with histone marks suggestive of active expression and regulation. Upon further investigation, we identified the EU358092 transcript beginning within the first ECR using the UCSC Genome Browser's 'human mRNA' data set, which stated that this mRNA had been identified in the foetal brain.

Characterisation of the evolutionary conservation of EU358092, the schizophrenia GWAS SNPs across the locus, and its expression both *in vitro* and *in vivo*, demonstrated that EU358092 had much in common with MIR137 and should therefore be further studied as a potential contributor to schizophrenia risk at this locus.

3.7 Aims

- Further analyse the extended MIR137 locus, as defined by the GWAS signal over this region, to identify potential novel transcripts that may contribute to schizophrenia risk at this locus.
- Overlay schizophrenia GWAS data across the MIR137 locus RNA, EU358092, and further understand linkage disequilibrium across this region.
- Characterise the expression patterns of EU358092 in humans and primates using publicly available transcriptomic data.
- Test for EU358092 expression in a neuroblastoma cell line model, and demonstrate stimulus-inducible expression upon exposure to psychoactive compounds.
- Identify and characterise the regulatory potential of two evolutionary conserved regions at the EU358092 promoter.

3.8 Results

3.8.1 Bioinformatic analysis of the MIR137 locus identifies regions of conservation and transcriptional activity

During the analysis of ECRs at the MIR137 locus, we extended our view of the locus downstream of MIR137 to also encompass the start of the neighbouring gene, DPYD, as schizophrenia GWAS data extends from MIR137 over this flanking region (Figure 3.9a). We identified two further ECRs of potential interest in the region upstream of DYPD, initially drawn to these regions due to their high evolutionary conservation back to the mouse (Figure 3.9b). We initially postulated that these ECRs may be regulators of the adjacent gene, DPYD, or perhaps a nearby, unidentified transcript, as an area of H3K27ac histone marks approximately 3 kb downstream of one of the ECRs suggested that they may be directing regulation of an unknown transcript starting around this area.

Indeed, further analysis of the locus using the UCSC Genome Browser to view human mRNAs revealed a brain expressed ncRNA, EU358092, originating from within one of the ECRs (Figure 3.9b). The two ECRs were consequently named EU 1, which partially encodes the first exon of EU358092, and EU 2, which is 2.6 kb upstream of the transcriptional start site of EU358092.

Observation of data from both the ECR Browser and UCSC Genome Browser showed that EU 1 displayed 84.3% conservation back to the mouse genome, while EU 2 was conserved 66.7 and 79.1% to the mouse genome over the two sections of high conservation making up EU 2 (Figure 3.10).

Figure 3.9 Schizophrenia GWAS data across the chromosome 1p21.3 MIR137 locus, expanded to show the position of an uncharacterised RNA, EU358092.

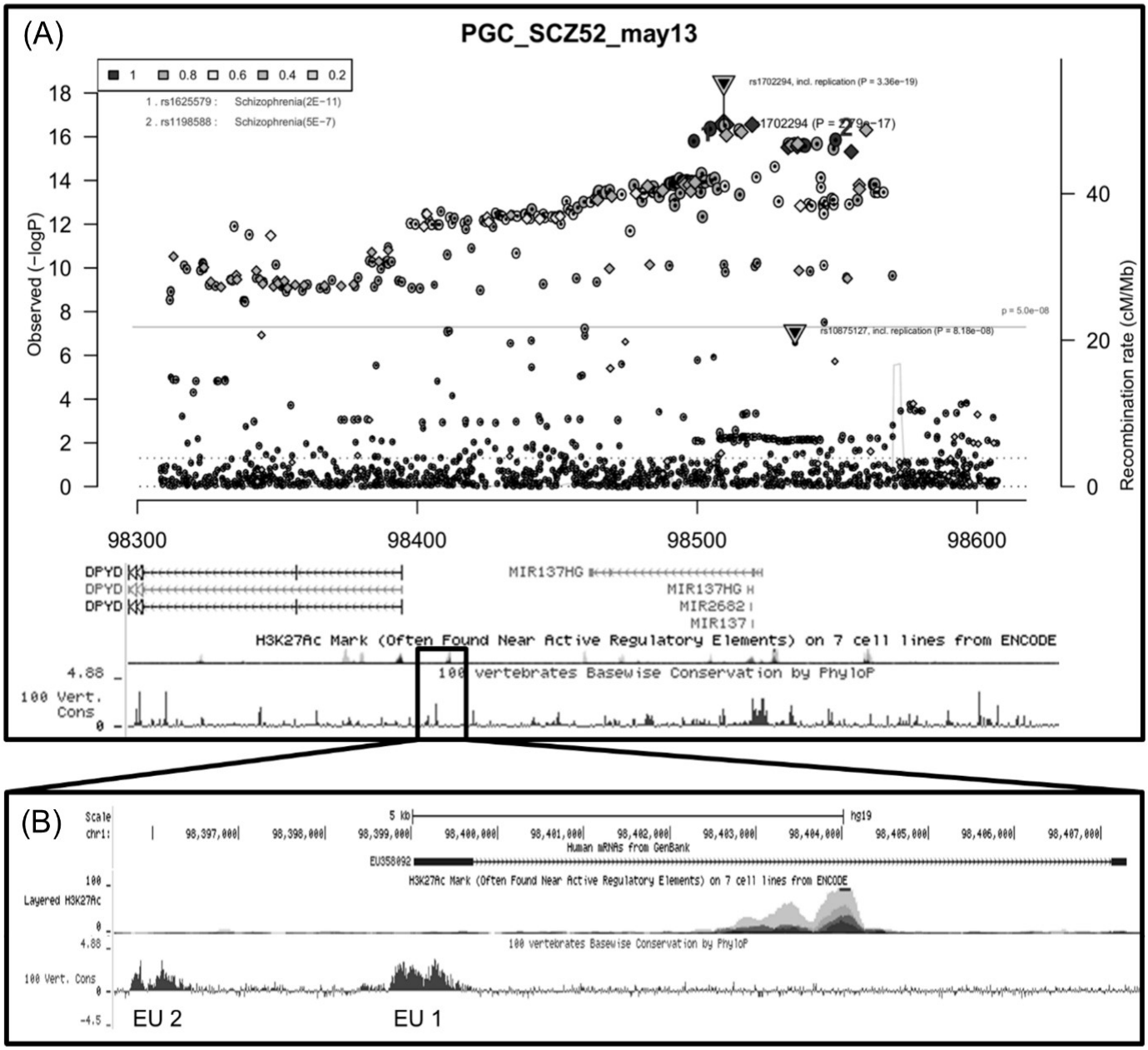


Figure 3.9 Schizophrenia GWAS data across the chromosome 1p21.3 MIR137 locus, expanded to show the position of an uncharacterised RNA, EU358092.

(a) A composite figure overlaying data across the MIR137 and DPYD genes at chr1:98,300,000–98,600,000. The Psychiatric Genomics Consortium's May 2013 schizophrenia GWAS data (PGC SCZ52), accessed and viewed through Ricopili, was overlaid onto annotated genes, histone modifications, and evolutionary conservation data from the UCSC Genome Browser across the same genomic region. Points above the horizontal line in the GWAS data represent GWAS SNPs with a statistically significant p*-value of* 5.0×10^{-8} *or lower, demonstrating a large number of genome-wide SNPs for association with schizophrenia across the locus, which extend from MIR137 into EU358092 and the neighbouring gene, DPYD. Evolutionary conservation and H3K27ac histone modifications over this locus are shown below, and are markers of potential conserved function and active transcriptional regulation, respectively. The region encompassing EU358092 is boxed and expanded in (b).*

(b) The EU358092 boxed region in (a) is expanded, showing more clearly the peak of H3K27Ac histone modifications, and presenting sequence comparison between multiple species. This identified two evolutionary conserved regions which are displayed as peaks named EU 1 and EU 2, numbered from the start of the transcript and outwards. We find that EU 1 overlaps the first exon of EU358092, and may therefore display conservation due to being an exon. EU 2 on the other hand is approximately 2.6 kb upstream of the predicted transcriptional start site of EU358092. The location of EU 2 upstream of the EU358092 transcriptional start site may suggest that it could exert a regulatory effect on the EU358092 promoter.

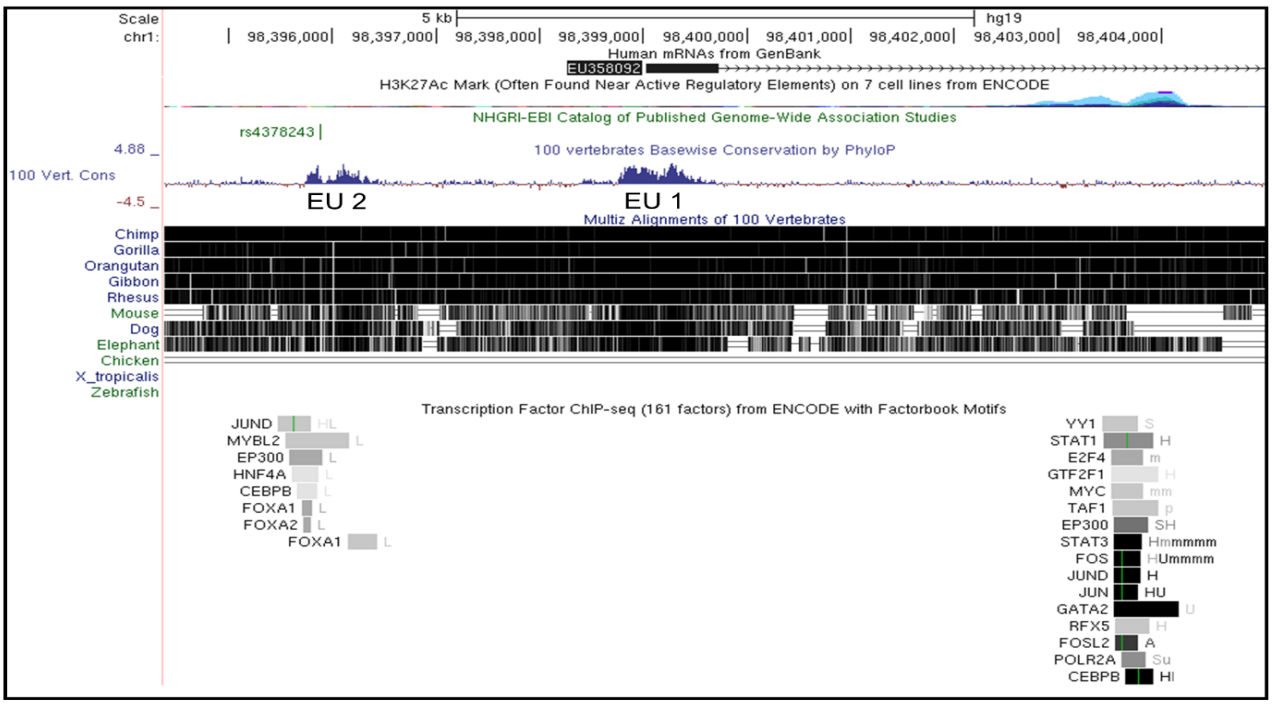


Figure 3.10 Evolutionary conservation and transcription factor binding at the EU358092 locus.

The region around the EU358092 transcriptional start site and ECRs demonstrates high conservation across the majority of this region back to species including the elephant, dog, and mouse. ECRs EU 1 and EU 2 are represented as peaks on the 100 vertebrates basewise conservation track. EU 2 contains a GWAS SNP with significance for both schizophrenia and educational attainment. Multiple transcription factors are likely to bind across this ECR and SNP according to ENCODE ChIP-seq data, which may modulate EU358092 expression, potentially in a manner dependent on the genotype of the GWAS SNP. EU 1 partially encodes the first exon of EU358092, and does not have any ENCODE ChIP-seq data over this region. A region of active histones within the EU358092 intron overlaps a region of partial primate-specific sequence. ENCODE ChIP-seq data at this region demonstrates the binding of 16 transcription factors including RNA Polymerase II, suggesting active regulation or transcription from this area, which may be altered in, or specific to, primates based on evolutionary conservation of the DNA sequence at this region. The presence of FOS and JUN binding at this region would also suggest activity in the brain, as these proteins are classical immediate-early response transcription factors which modulate expression in neurons.

Further interrogation of data available around the ECRs through the UCSC Genome Browser displayed transcription factor binding across EU 2 according to ENCODE ChIP-seq data, with signals for JUND, MYBL2, EP300, HNF4A, CEBPB, FOXA1, and FOXA2 overlapping the conserved region, predominantly in the HepG2 liver cell line (Figure 3.10). Further, EU 2 was shown to contain the G/T SNP, rs4378243, at the position chr1:98395881 (hg19) which has been identified as one of a number of schizophrenia risk SNPs across this region in the PGC2 schizophrenia GWAS data.

The rs4378243 SNP has also been highlighted by Okbay et al. as being associated with educational attainment, as measured by years in education (Okbay et al. 2016), which showed a small but statistically significant positive correlation with increased schizophrenia risk in this study. While schizophrenia diagnosis or risk has typically been associated with lower educational attainment or IQ (Lencz et al. 2014, Mistry et al. 2017, Kendler et al. 2015, Tempelaar et al. 2017), other studies based on polygenic scores for schizophrenia and educational attainment have demonstrated effects in both directions. For example, work by Le Hellard et al. identified 30 SNPs across 26 genomic intervals which were associated with both schizophrenia and educational attainment. 46.7 % (14/30) of the SNPs were associated with both increased schizophrenia risk and increased educational attainment, while 23.3% (7/30) of SNPs were associated with lower educational attainment and increased schizophrenia risk (Le Hellard et al. 2017). The literature for the direction of correlation between schizophrenia risk and educational attainment at the MIR137 locus is currently unclear. For example, Cosgrove et al. demonstrated that increased polygenic risk for schizophrenia at this locus was associated with lower IQ (Cosgrove et al. 2017), while Le Hellard et al. demonstrated that SNPs at the MIR137 locus that are significantly

associated with both schizophrenia risk and educational attainment show a positive correlation for these traits (Le Hellard et al. 2017).

While measures such as educational attainment are likely to have a strong social component, this may suggest a role for the EU358092/MIR137 locus in influencing broader aspects of cognition.

The aforementioned region of H3K27ac histone modifications identified downstream of the ECRs on the UCSC Genome Browser was found to lie within the intron of EU358092 and overlapped multiple transcription factor binding signals according to ENCODE ChIP-seq data. This region spans approximately 2.3 kb, around half of which is made up of primate-specific sequence, being conserved only back to the Rhesus macaque based on the Multiz alignment of 100 vertebrate species available through the UCSC Genome Browser. Among the 16 transcription factors identified as binding at this region, we see RNA Polymerase II binding in two cell lines, the brain-derived SK-N-SH and U87, suggesting the potential for expression from this region in brain tissue (Figure 3.10). We also note FOS and JUN binding at this region according to ENCODE ChIP-seq data. FOS and JUN are immediate-early transcription factors and classical markers of neuronal activation which can be induced by stress (Hoffman, Smith and Verbalis 1993, Sagar, Sharp and Curran 1988, Cullinan et al. 1995, Honkaniemi et al. 1992). This may further suggest activity around this region in the CNS, with potential regulation in response to environmental stimuli through FOS and JUN.

While much research has focussed on MIR137 as the gene responsible for schizophrenia association across this locus, we find that EU358092 shares a number of equally important attributes with MIR137 and is therefore worth further study to

assess the potential of this RNA to function within tissues and pathways that may indicate an additive affect to the schizophrenia risk across the MIR137 locus.

Continuing our use of the Multiz 100 vertebrate genome alignment, we find that the exons of EU358092 are as conserved as the second exon of the MIR137 short transcript, AK311400, as well as exons three and four of MIR137HG. Using BLAT to search for the complete EU358092 sequence (8273 bp) in other species, we find that it is conserved 94.1% to the Rhesus macaque over a genomic distance of 8463 bp (BCM Mmul_8.0.1/rheMac8, November 2015), and 90.2% back to the mouse genome over a significantly smaller region of 380 bp (GRCm38/mm10, December 2011) (data not shown).

3.8.2 Bioinformatic data supports long non-coding RNA status of EU358092

Collecting data from multiple bioinformatic sources provided support for EU358092 as a lncRNA. Both GenBank and AceView databases of annotated genes listed the RNA as a human, brain-expressed, 869 bp mRNA, with data from GenBank identifying this transcript in the foetal brain, and AceView listing the transcript as identified in the amygdala. Data from multiple sources suggested that EU358092 was comprised of two exons, spanning a genomic distance of 8.27 kb (chr1:98,399,030–98,407,302). In order to assess the protein coding potential for EU358092, we utilised the ExPASy translation tool to predict protein sequence. This identified multiple stop codons in each of the three possible reading frames (data not shown), with the longest potential protein coding open reading frame being 85 amino acids in length. While this is not definitive, such observations would support the likelihood of EU358092 functioning as an untranslated RNA.

3.8.3 GWAS and LD support a role for EU358092 in schizophrenia

Using the UCSC Genome Browser to overlay the Psychiatric Genomics Consortium's 2013 schizophrenia GWAS data alongside published GWAS data for multiple traits and conditions from the National Human Genome Research Institute (NHGRI) and the European Bioinformatics Institute (EMBL-EBI), we identify 16 schizophrenia risk SNPs with a p-value of < 0.001 across the area encompassing the EU358092 transcript and its upstream ECRs (Figure 3.11a). Of these 16 schizophrenia-associated SNPs, one of these, the aforementioned rs4378243, resides within the EU 2 ECR, and is also listed as being associated with educational attainment, with a p-value of 1 $x10^{-9}$. Published in Nature in 2016, this paper also identified two additional SNPs around the MIR137 locus, rs1198575 and rs17372140, with genome wide significance for educational attainment (p-value = 2 $x10^{-6}$ and 9 $x10^{-9}$, respectively) (Okbay et al. 2016). Okbay et al. demonstrate that educational attainment, measured in this instance by years in education, is positively correlated with diagnoses of schizophrenia and bipolar disorder, as well as measures of cognitive performance.

In addition to the schizophrenia and educational attainment GWAS SNP, rs4378243, we find a second schizophrenia GWAS SNP in the PGC2 data set, rs4294451, within 20 bp of the EU 2 ECR, suggesting that variation at this highly conserved potential regulatory element may modulate educational attainment and/or schizophrenia risk (Figure 3.11a).

Figure 3.11 Schizophrenia risk SNPs and linkage disequilibrium across EU358092.

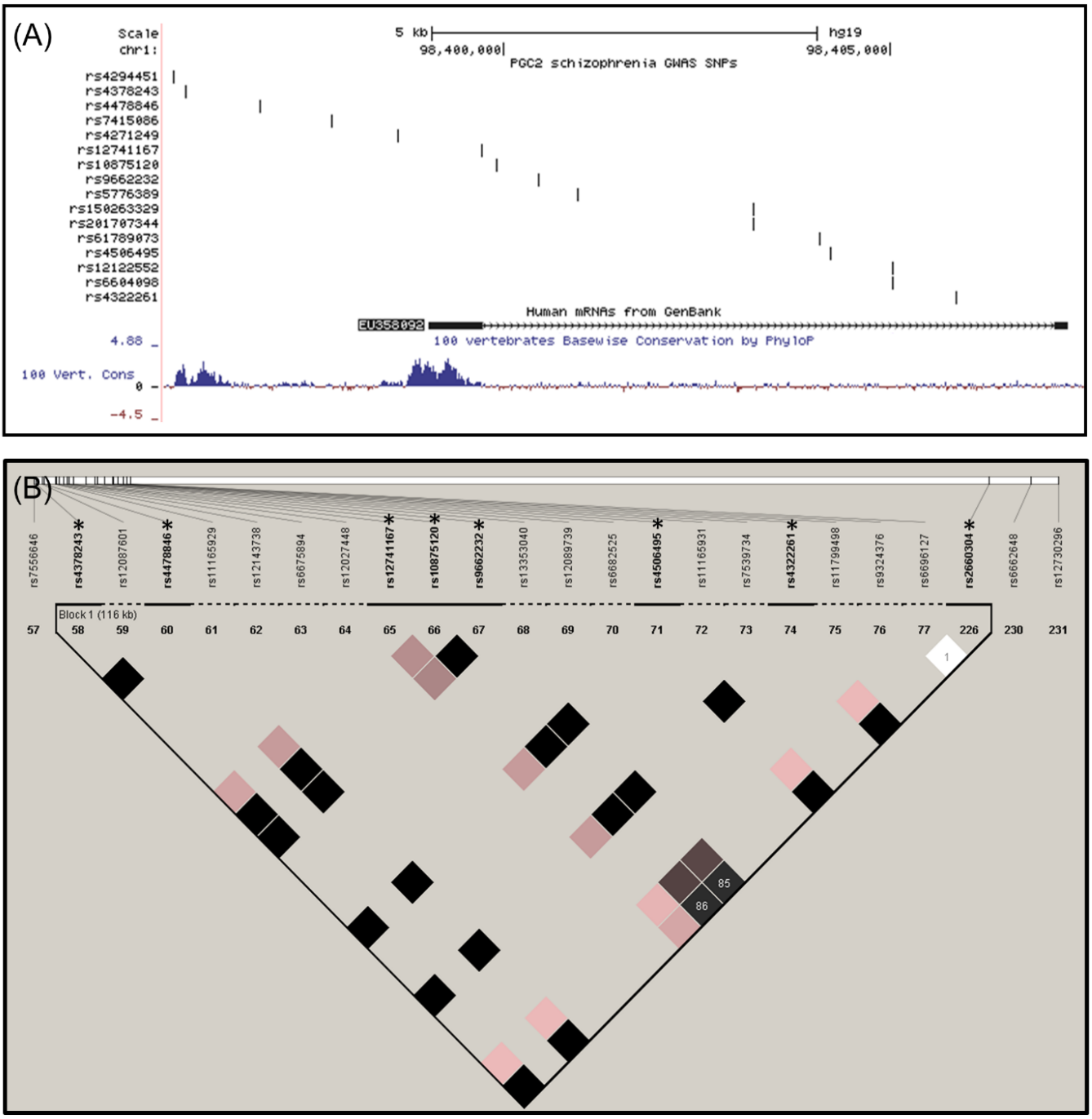


Figure 3.11 Schizophrenia risk SNPs and linkage disequilibrium across EU358092.

(a) *Visualisation of the PCG2 schizophrenia GWAS data across the EU358092 locus and identified ECRs demonstrated 16 schizophrenia GWAS SNPs across this region. Two schizophrenia GWAS SNPs (rs4294451 and rs4378243) were within EU358092 ECR 2 (EU 2), and may modify the function of this ECR in a genotype dependent manner, or highlight this as a region of functional interest.*

(b) *Linkage disequilibrium analysis of SNPs at EU358092 was carried out using SNP genotype data from the HapMap CEU cohort, in which data for seven of the 16 schizophrenia GWAS SNPs (starred*) was available. Non-GWAS HapMap SNPs around the selected ECRs were also included in this analysis, however, the minor allele frequency for these SNPs was not sufficient to assess LD. D' values of LD are shown in this figure. All schizophrenia GWAS SNPs at this locus, with the exception of the EU358092 splice site SNP rs12741167, were in complete LD and form a haplotype block as defined by Gabriel et al. (Gabriel et al. 2002). Further, we note that the schizophrenia GWAS SNPs at EU358092 were in LD with the MIR137 internal promoter GWAS SNP, rs2660304. LD between EU358092 and MIR137 therefore makes it difficult to distinguish whether the GWAS signal across this locus is highlighting EU358092 or MIR137, which may be acting separately or synergistically in adding to the schizophrenia association at this locus.*

Of the remaining 14 schizophrenia GWAS SNPs over the EU358092 region, three were found to be within an area of approximately 2.2 kb upstream of the transcriptional start site, one directly on the annotated splice site at the end of the first exon (rs12741167), and 10 across the EU358092 intron.

In order to assess linkage disequilibrium (LD) across EU358092 and the wider MIR137 region, we again utilised SNP genotype data from the HapMap CEU cohort. Data for seven of the 16 schizophrenia associated SNPs across EU358092 was available in the HapMap dataset. Other non-GWAS SNPs available in the HapMap were also included in this analysis, but lacked sufficient data to assess LD. In addition to SNPs across EU358092, three SNPs at the MIR137 major and internal promoters, including the schizophrenia associated rs2660304, were also included in this analysis. A total of 24 SNPs were included, seven of which were schizophrenia GWAS SNPs across EU358092, and one of which was a schizophrenia GWAS SNP at the MIR137 internal promoter.

LD analysis across the EU358092 and MIR137 locus showed that six of the seven schizophrenia GWAS SNPs tested across EU358092 were in complete LD with each other, and in complete (D' = 1) or high (D′ ≥ 0.85) LD with the MIR137 internal promoter GWAS SNP, rs2660304 (Figure 3.11b). The schizophrenia GWAS SNP, rs12741167, which sits directly at the annotated splice site at the end of the first exon of EU358092, was the only exception, and displayed partial LD with other GWAS SNPs in the locus. The seven schizophrenia GWAS SNPs found to be in LD in this analysis formed a haplotype block which spanned both EU358092 and MIR137. SNPs at the MIR137HG promoter, rs6662648 and rs12730296, unfortunately lacked sufficient data to assess whether the LD across EU358092 and the MIR137 internal promoter schizophrenia GWAS SNPs extended further upstream to the MIR137HG promoter. Overall, this data

may suggest that schizophrenia risk at this locus is mediated through related mechanisms affecting both EU358092 and MIR137 together.

3.8.4 Expression of EU358092 and activity of EU ECRs in *ex vivo* samples

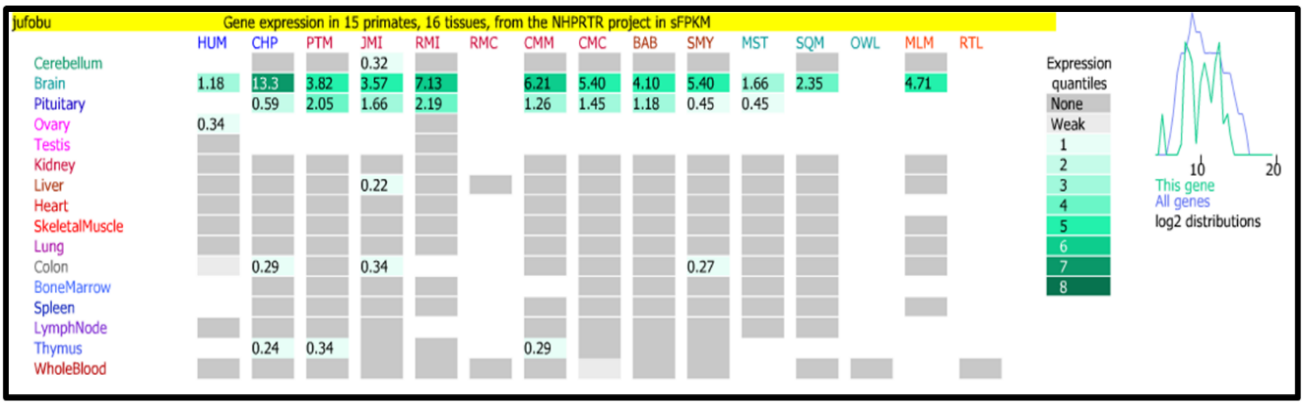
In order to further understand the potential functions of the EU358092 RNA, we used multiple bioinformatic resources to look at its expression in human and primate tissues. We first utilised the Non-Human Primate Reference Transcriptome Resource (NHPRTR), which is accessible through AceView. The NHPRTR allows simple viewing and comparison of RNA-seq data from multiple tissues across 15 different primate species and subspecies. This data set showed that expression of EU358092 (termed 'jufobu' in this database) is restricted to the brain and pituitary gland and that expression of the RNA is conserved throughout all primate species in this data set (Figure 3.12a).

To compare this expression pattern to other genes around this locus, we also viewed data on the expression of neighbouring genes MIR137 and DPYD in this resource. MIR137 displays a near identical expression pattern to EU358092 across primate species, with expression being restricted primarily to the brain and pituitary gland in this data set (Figure 3.12b). This related pattern of expression in humans and primates may suggest that EU358092 and MIR137 are co-expressed *in vivo*. Conversely, DPYD was found to be expressed at much higher levels and across the majority of tissues in all primate species with data available in this resource (Figure 3.12c).

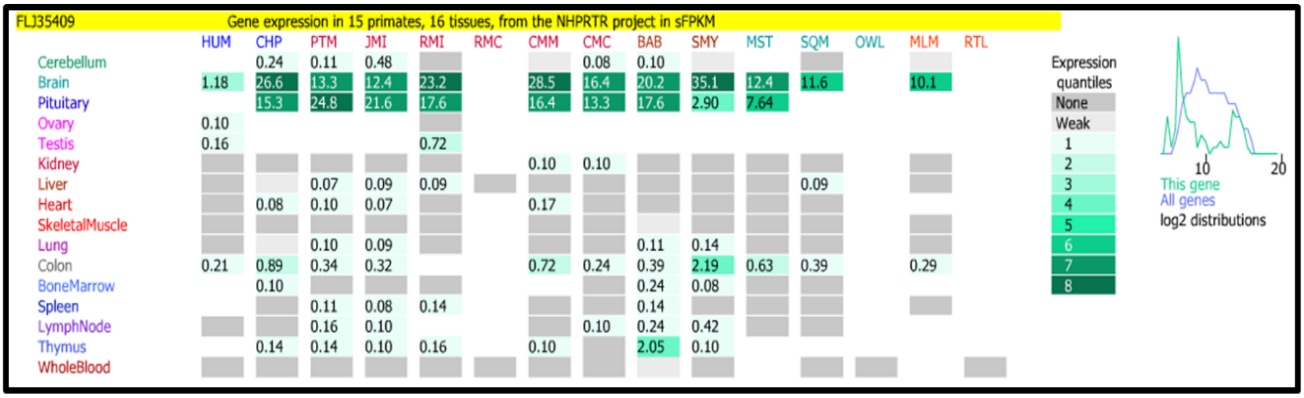
As EU358092 expression is primarily restricted to the brain, we made use of the Sestan Lab microarray expression data from four late-mid stage foetal brains in order to gain a more precise understanding of how EU358092 is expressed across different brain regions in humans.

Figure 3.12　Tissue expression profiles of EU358092, MIR137, and DPYD in humans and primates.

(A) EU358092



(B) MIR137



(C) DPYD

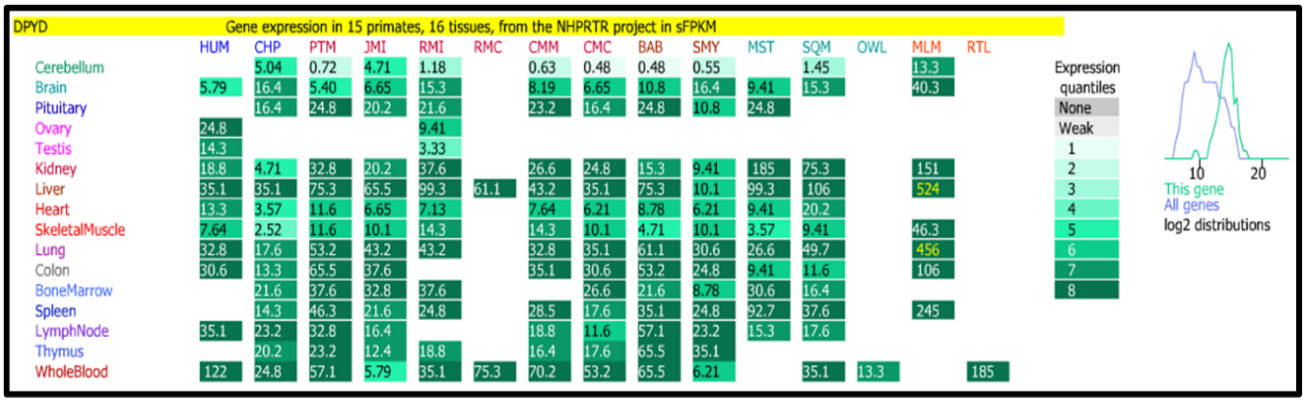


Figure 3.12 Tissue expression profiles of EU358092, MIR137, and DPYD in humans and primates.

RNA-seq data from the Non-Human Primate Reference Transcriptome Resource (NHPRTR) comparing EU358092, MIR137, and DPYD expression across 16 tissues in humans and primates. MIR137 and EU358092 display almost indistinguishable expression profiles, with their expression being limited to the brain and pituitary in this data set. EU358092 expression was found to be consistently lower than MIR137 across all primate species except humans, where MIR137 and EU358092 expression levels were identical in the human brain in this data set. This data may suggest co-ordinated expression of MIR137 and EU358092. On the other hand, DPYD, MIR137's neighbouring gene which also resides within the area covered by schizophrenia GWAS association, is expressed almost ubiquitously across all tissues with available data in this resource.

HUM = human; CHP = chimp; PTM = pig-tailed macaque; JMI = Japanese macaque; RM(I/C) = rhesus macaque Indian/Chinese; CM(M/C) = cynomolgus macaque Mauritian/Chinese; BAB = olive baboon; SMY = sooty mangabey; MST = marmoset; SQM = squirrel monkey; OWL = owl monkey; MLM = mouse lemur; RTL = ring-tailed lemur.

This data set is available to view through the UCSC Genome Browser. At the late-mid foetal time point, we observed EU358092 expression in the orbital prefrontal, medial prefrontal, temporal association, and temporal auditory cortices, with low or no expression in the hippocampus, striatum, thalamus, or cerebellum (Figure 3.13a). Similarly, expression of MIR137 in the foetal brain displayed a broadly similar pattern to that of EU358092, with the miRNA being most highly expressed in the temporal association and temporal auditory cortices, and expression also seen across multiple regions of the prefrontal cortex, including the medial, dorsolateral, and ventrolateral areas (Figure 3.13b). As with EU358092, low or no expression of MIR137 was seen in the hippocampus, thalamus, or cerebellum. This similar expression pattern seen for EU358092 and MIR137 both across tissues in humans and other primates, and across multiple foetal brain regions in humans, provides further evidence to suggest co-expression of the RNAs at this locus within the human brain, and may point to an important but as yet undetermined brain-related function for EU358092.

After determining the active expression of EU358092 in the human brain, we made further use of the HaploReg v4.1 tool. This allowed interrogation of histone modification and chromatin state data around the EU358092 ECRs across multiple human tissues and cell lines. The EU 1 ECR partially encodes the first exon of EU358092. The ECR does not contain any common SNPs, but has three flanking SNPs, rs4271249, rs12741167, and rs79895962. The closest SNPs on either side of the ECR, rs4271249 and rs12741167, were used in HaploReg to view histone and chromatin data over the EU 1 ECR. Data over both SNPs is broadly in agreement, showing H3K4me1 histone modifications in a number of embryonic and iPSC lines as well as H9 derived neuron cultured cells (Figure 3.14a).

Figure 3.13 Comparison of MIR137 and EU358092 expression in the human foetal brain.

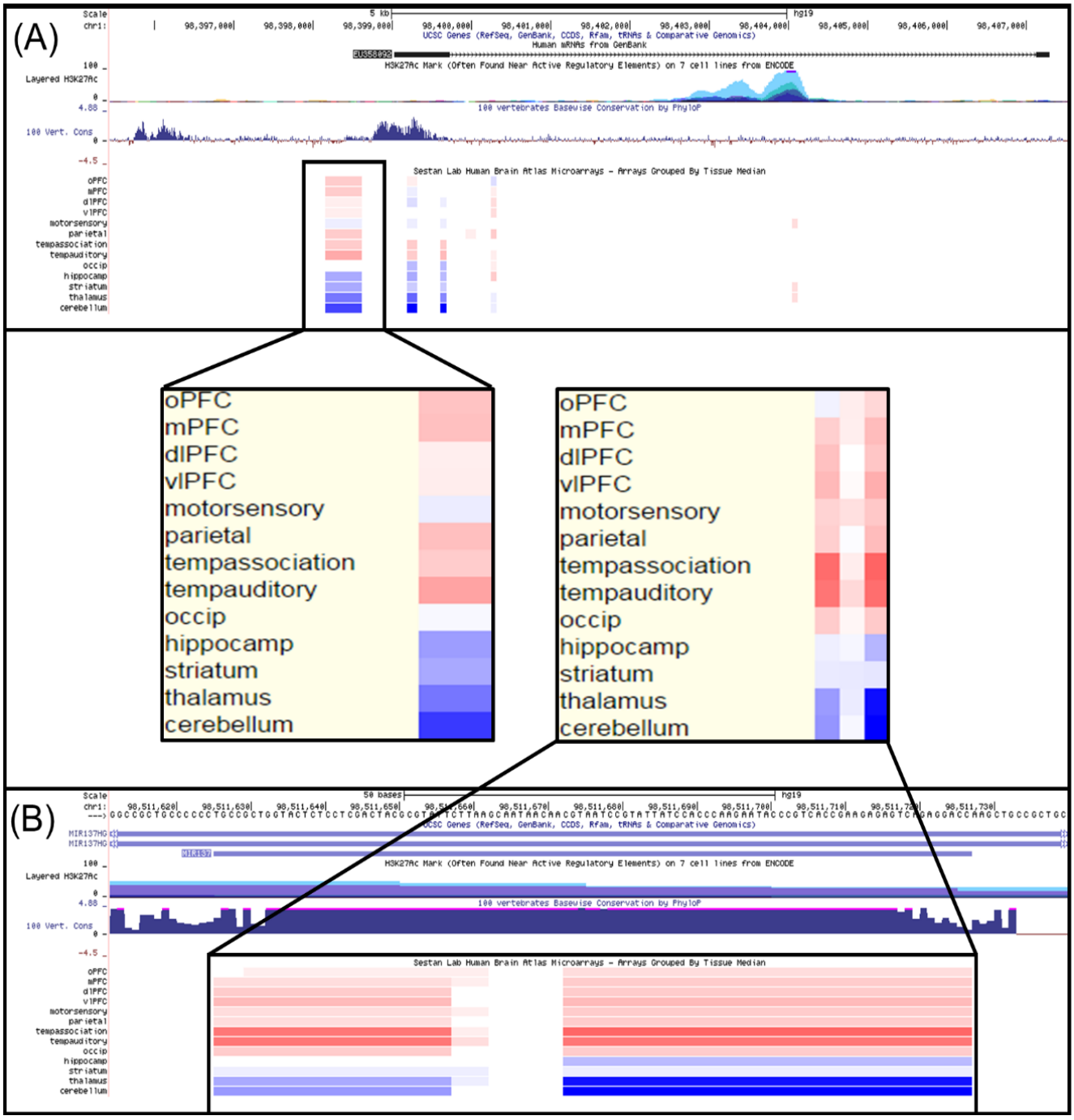


Figure 3.13 Comparison of MIR137 and EU358092 expression in the human foetal brain.

(a) Microarray expression data from four mid-late foetal stage brains (accessible through the Sestan Lab Microarray track on the UCSC Genome Browser) showed EU358092 expression (red) in the orbital and medial prefrontal cortices, as well as the parietal, temporal association and temporal auditory cortices. No expression (blue) of EU358092 was seen in the hippocampus, striatum, thalamus, or cerebellum at this time point.

(b) Expression data from the same resource showed MIR137 expression across regions of the prefrontal cortex, including the medial, dorsolateral, and ventrolateral regions, as well as being most highly expressed in the temporal association and temporal auditory cortices during this stage of brain development in the foetus. MIR137 was not found to be expressed in the thalamus or cerebellum in this data set.

The similar expression patterns of EU358092 and MIR137 across multiple regions of the human foetal brain may suggest co-ordinated expression of genes at this locus, which would provide further support for a potentially important brain-related function for EU358092.

(A) ECR EU 1

Description	Chromatin states (15-state model)	Chromatin states (25-state model)	H3K4me1	H3K4me3	H3K27ac	H3K9ac
ES-I3 Cells			H3K4me1_Enh			
HUES48 Cells			H3K4me1_Enh			
iPS-20b Cells			H3K4me1_Enh			
iPS-15b Cells			H3K4me1_Enh			
H9 Derived Neuron Cultured Cells	7_Enh		H3K4me1_Enh			
Brain Hippocampus Middle	7_Enh		H3K4me1_Enh		H3K27ac_Enh	
Brain Substantia Nigra		17_EnhW2	H3K4me1_Enh	H3K4me3_Pro	H3K27ac_Enh	
Brain Anterior Caudate					H3K27ac_Enh	
Brain Cingulate Gyrus		17_EnhW2	H3K4me1_Enh		H3K27ac_Enh	
Brain Inferior Temporal Lobe	7_Enh	17_EnhW2	H3K4me1_Enh		H3K27ac_Enh	
Brain Angular Gyrus		18_EnhAc	H3K4me1_Enh	H3K4me3_Pro	H3K27ac_Enh	H3K9ac_Pro
Brain_Dorsolateral_Prefrontal_Cortex					H3K27ac_Enh	
Fetal Brain Male			H3K4me1_Enh			

(B) ECR EU 2

Description	Chromatin states (15-state model)	Chromatin states (25-state model)	H3K4me1	H3K4me3	H3K27ac	H3K9ac
Brain Substantia Nigra				H3K4me3_Pro		
Brain Cingulate Gyrus			H3K4me1_Enh		H3K27ac_Enh	
Brain Angular Gyrus					H3K27ac_Enh	
Brain Germinal Matrix		19_DNase				
Fetal Brain Female		19_DNase				
Fetal Brain Male	7_Enh	19_DNase	H3K4me1_Enh			

Figure 3.14 Chromatin state and histone modification data at EU358092 ECRs.

(a) Chromatin state and histone modification data from HaploReg v4.1 across the EU358092 ECRs demonstrates clear evidence of EU 1 acting as a transcriptional regulator in vitro in multiple embryonic stem cell lines, as well as in multiple human brain regions. H3K4me3 and H3K9ac histone modifications also suggest that this ECR may act as a promoter in the substantia nigra and angular gyrus.

(b) Chromatin state and histone modification data suggests that the EU 2 ECR may act as a transcriptional regulator in multiple human brain regions, including the cingulate gyrus, angular gyrus, and foetal brain. Further, H3K4me3 histone modifications at this ECR suggest that it may have the capacity to act as a promoter region in the substantia nigra.

Chromatin states: Enh = enhancer, EnhAc = enhancer acetylation only, EnhW1/2 = weak enhancer 1 or 2.

Histone modifications: Enh = enhancer, Pro = promoter, Black = no available data.

Yellow = Enhancer (Enh), weak enhancer (EnhW1, EnhW2). Orange = Active enhancer (Enh), flanking active enhancer (EnhAF). Red = Active transcriptional start site (TssA), flanking active transcriptional start site (TssAFlnk), promoter (Pro), promoter upstream transcriptional start site (PromU). Pink = Poised promoter (PromP).

Further H3K4me1 modifications at this ECR are seen in a range of primary T helper and regulatory cells from peripheral blood (Supplementary Data 3.2). However, by far the clearest signal is found in the brain, with H3K4me1 and H3K27ac histone modifications and chromatin structure indicative of enhancer activity seen across multiple brain regions, including the hippocampus, substantia nigra, cingulate gyrus, and inferior temporal lobe. We also observe H3K4me3 and H3K9ac marks, indicative of nearby promoter activity, in the substantia nigra and angular gyrus, as well as some additional evidence of enhancer activity in the anterior caudate and foetal brain (Figure 3.14a). Of the ten brain regions with available data in the HaploReg database, we find evidence to support activity around the EU 1 ECR in eight brain regions. Taken together, these data would imply transcription and regulation around this ECR predominantly in the brain, which would be in line with the brain-expressed EU358092 RNA originating from within this ECR. These data may also imply a potential regulatory role for this region in embryonic development and in primary T cells.

The GWAS SNP for both schizophrenia and educational attainment, rs4378243, lies within the EU358092 ECR 2 (EU 2) and was used as the input SNP for gathering data across this ECR. Data from HaploReg shows a combination of H3K4me1, H3K4me3, and H3K27ac histone modifications, as well as DNase I sensitivity across the EU 2 ECR in multiple brain regions, including the substantia nigra, cingulate gyrus, angular gyrus, and the foetal brain, as well as DNase I hypersensitivity in multiple lines of primary cultured neurospheres (Figure 3.14b). Outside of the brain, there is strong histone and chromatin state evidence for functional promoter and/or enhancer activity at this ECR in the liver, with H3K4me1, H3K4me3, H3K27ac, H3K9ac, and chromatin conformation suggesting an enhancer within close proximity of a poised promoter and flanking transcriptional start site (Supplementary Data 3.2). Poised chromatin is

defined as an area which has both active and repressive histone modifications. Such patterns of histone modification are typically found at promoters, indicating that the region is silenced but ready to be activated. Such data would correlate with transcription factor binding data from ENCODE ChIP-seq across this region, signals from which were primarily identified in the HepG2 liver derived hepatocellular carcinoma cell line. We also see histone modifications suggestive of regulatory activity in a number of cell lines, such as HepG2 hepatocellular carcinoma, A549 lung carcinoma, and human skeletal muscle and myoblast cells (HSMM).

The above data supports a potential role for the EU 2 ECR in regulating expression at this region in multiple brain regions in humans, and may also suggest a previously unexpected role in the liver. Due to its proximity to the DPYD transcriptional start site approximately 9.2 kb away, this ECR may also have the capacity to influence DPYD expression. Further work to clarify the role of the EU 2 ECR could employ chromatin capture technology in a range of cell lines or tissues, including the HepG2 liver cell line or ex vivo liver tissue, to test for potential interactions between the EU 2 ECR and the DPYD major promoter, and other potential interactions that may reveal any tissue- or cell line-specific roles for this conserved region.

3.8.5 Activity of the EU358092 locus in vitro

After using publicly available data sets to gather evidence for the expression of EU358092 across the human brain, and to suggest transcriptional activity at the EU 1 and EU 2 ECRs in the brain, we next aimed to identify expression of this RNA and activity of its ECRs in the SH-SY5Y neuroblastoma cell line.

Warburton et al. has previously shown that MIR137 expression is modulated in response to challenge with drugs such as cocaine in this cell line (Warburton et al.

2015b). To further address the similarities between EU358092 and MIR137, and to investigate the possibility of co-expression and co-regulation, we conducted similar experiments to assess the response of EU358092 to challenge with lithium chloride, sodium valproate, cocaine, and amphetamine, *in vitro*. These drugs were selected for a number of reasons. Firstly, all four are well known for their CNS effects, with lithium and sodium valproate having been used in the treatment of brain-related conditions such as mood disorders and epilepsy for many decades (Won and Kim 2017, Loscher 2002). In particular, sodium valproate is a known histone deacetylase (HDAC) inhibitor (HDI). HDI's block the action of HDACs, which typically remove acetyl groups from lysine residues in histones, resulting in chromatin condensation and reduced gene expression. The inhibition of HDACs can therefore result in increased open chromatin and increased gene expression, and HDIs are known to modulate expression of a number of genes associated with CNS conditions such as BDNF (Koppel and Timmusk 2013, Wu et al. 2008). For this reason, sodium valproate was also selected due to its known induction of a number of brain-expressed genes.

Similarly, cocaine and amphetamine are known for their psychoactive effects, and may precipitate relapse in individuals with schizophrenia, or mimic the effects of schizophrenia in drug users, resulting in side effects such as paranoia and hallucinations (Roncero et al. 2014, San et al. 2013, Bramness and Rognli 2016, McKetin et al. 2016). Secondly, all four drugs have been used routinely in the literature as an established challenge applied to neuronal cell lines (Billingsley et al. 2018, Hing et al. 2012, Roberts et al. 2007, Vasiliou et al. 2012, Warburton et al. 2015b, Brotons et al. 2011, Warburton et al. 2015c). Therefore, using these drugs in this particular study allowed comparison of results to a collection of previous work.

PCR for EU358092 was carried out using cDNA from SH-SY5Y cells under basal conditions, and from cells that had undergone 1 hour drug treatment with either 1 mM lithium chloride, 5 mM sodium valproate, 10 μM cocaine, or 10 μM amphetamine. We found that EU358092 is expressed in the SH-SY5Y cell line under basal conditions, and responds to drug treatments in a stimulus-inducible manner. EU358092 expression was found to be increased after 1 hour treatment with 5 mM sodium valproate, whereas treatment with 10 μM cocaine resulted in a strong decrease in expression to the point that no band was visible in this sample after PCR (Figure 3.15). This reflects previous data relating to the expression of MIR137, which shows that some transcripts of MIR137 are strongly repressed in the same cell line after 1 hour exposure to 10 μM cocaine (Warburton et al. 2015b).

To assess the transcriptional regulatory potential of the EU358092 ECRs, EU 1 and EU 2, their sequences were cloned into the pGL3P luciferase reporter vector and assayed for regulatory function 48 hours post-transfection in the SH-SY5Y cell line.

EU 2, which resides approximately 2.6 kb upstream of the annotated transcriptional start site of EU358092, and shows histone marks indicative of regulatory activity in *ex vivo* brain samples, was also confirmed to be a positive transcriptional regulator *in vitro*, supporting a 2.12-fold (t-test p-value = 7.97×10^{-4}) increase in reporter gene expression (Figure 3.16). Conversely, the EU 1 ECR, which partially encodes the first exon of EU358092 and had histone marks suggestive of active transcription from this region in the brain, was found to have no effect on the baseline expression of the reporter gene (t-test p-value = 0.9). This finding is likely to be in line with EU 1 partially encoding the first exon of EU358092, and therefore being conserved due to its function as part of the RNA, rather than due to any conserved regulatory function.

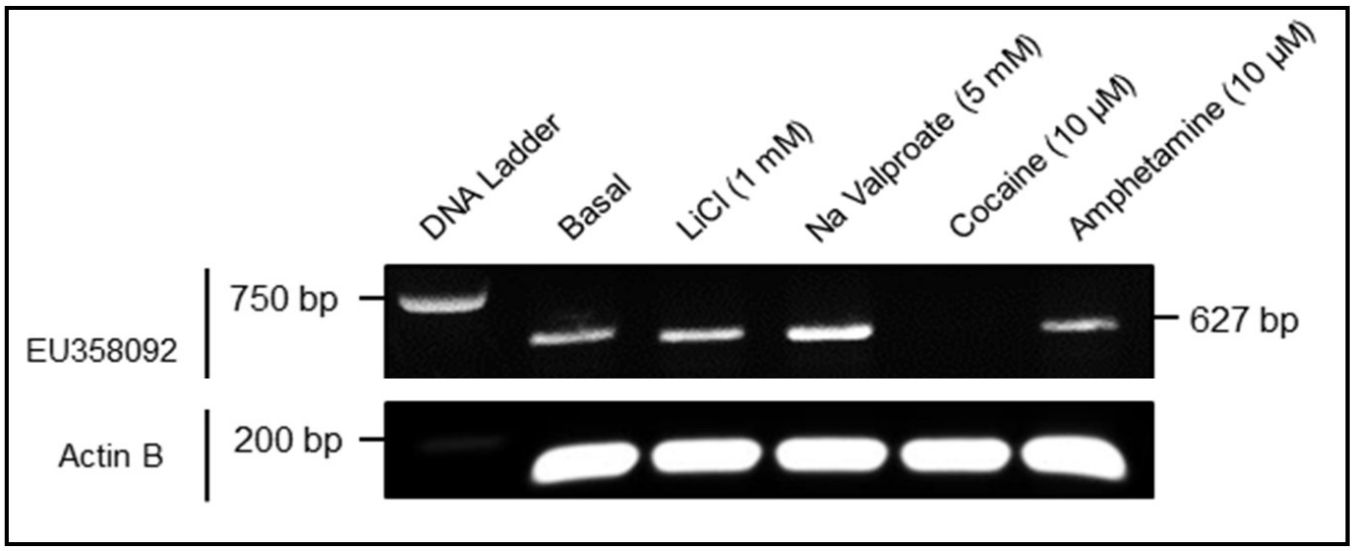


Figure 3.15 Expression of EU358092 in an SH-SY5Y neuroblastoma cell line model.

Expression of EU358092 was addressed in vitro in the SH-SY5Y neuroblastoma cell line. RT-PCR was carried out using cDNA from SH-SY5Y cells under basal conditions, and after the following treatments: lithium chloride (1 mM), sodium valproate (5 mM), cocaine (10 μM), and amphetamine (10 μM). PCR for the housekeeping gene, β-actin, was used as an internal control. EU358092 was found to be expressed in the SH-SY5Y cell line under basal conditions, with increased EU358092 expression following sodium valproate treatment, and strongly downregulated expression following cocaine treatment. A similar pattern of expression in response to cocaine challenge was also seen for MIR137 in previous studies in the lab (Warburton et al. 2015b). This may suggest co-ordinated regulatory responses of transcripts at this locus in response to challenge.

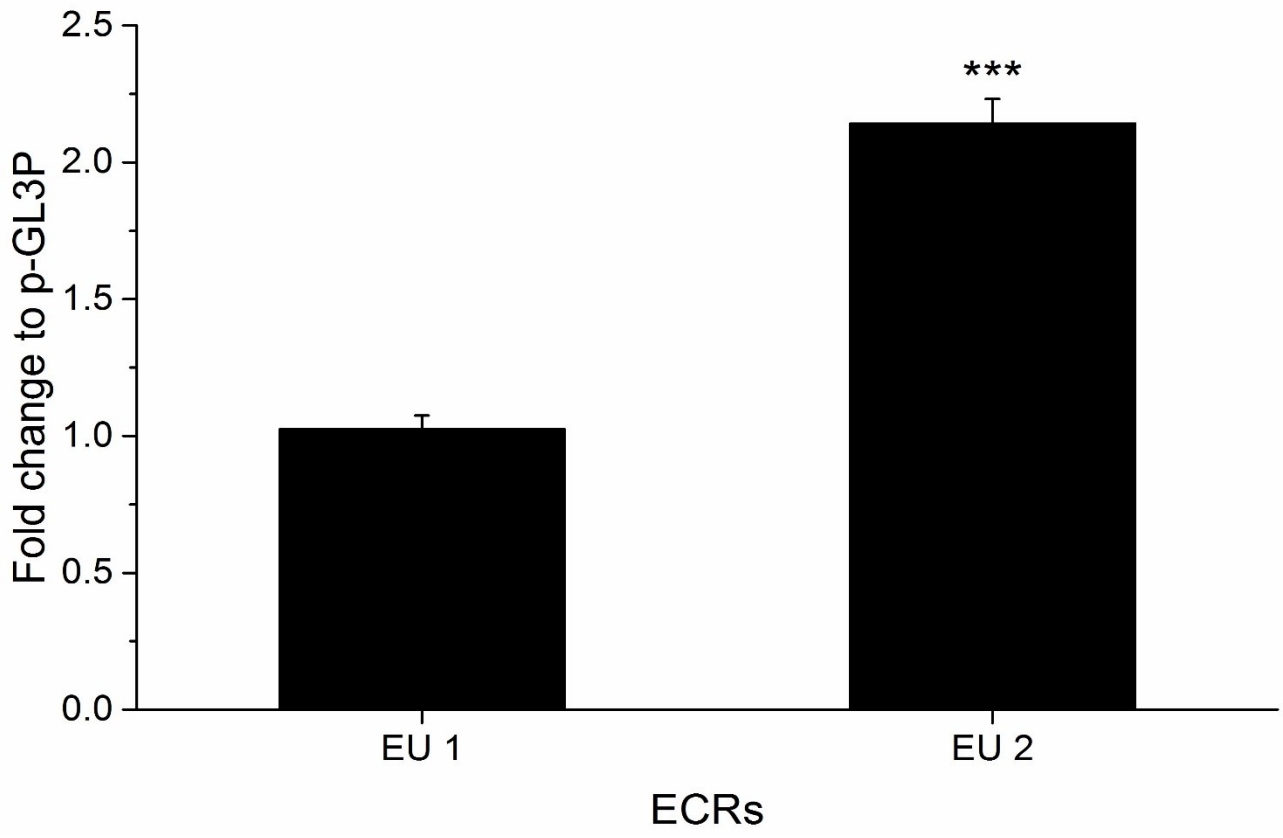


Figure 3.16 Transcriptional regulatory activity of EU358092 ECRs *in vitro* in an SH-SY5Y neuroblastoma cell line model.

*The regulatory potential of EU358092 ECRs was tested by cloning ECR sequences into the pGL3-Promoter (pGL3P) vector and performing luciferase reporter gene assays in the SH-SY5Y neuroblastoma cell line under basal conditions. EU358092 ECR 2 (EU 2) supported a 2.12-fold increase in reporter gene expression compared to baseline expression from the empty pGL3-P vector, supporting a regulatory role for this sequence which may modulate EU358092 expression. On the other hand, EU 1 displayed no regulatory function in this assay. This contrasts with chromatin state and histone modification data from HaploReg, but would be consistent with this region displaying evolutionary conservation because it encodes the first exon of the EU358092 transcript, rather than being a conserved regulatory element. Error bars display the standard error for each ECR, calculated as the standard divided by number of luciferase assays (n = 4), ***p < 0.001.*

The EU 1 data from HaploReg suggesting regulatory potential and promoter function around this ECR in the brain may have been in relation to a nearby promoter region for EU358092.

3.8.6 Evidence for an antisense transcript of EU358092 in the brain and in schizophrenia biology

Since carrying out the above work, the updated genome build (GRCh38/hg38) has listed the transcript RP11-272L13.3, also named LINC01930 in the RefSeq database, which is identical to the EU358092 sequence with the exception of being annotated as transcribed in the antisense direction. We tested the RP11-272L13.3 sequence for potential coding capacity using the ExPASy translation tool, and similarly found no evidence for protein coding capacity, with the longest potential protein in any reading frame being 73 amino acids (data not shown).

Expression data for RP11-272L13.3 is available through GTEx, the Geneotype-Tissue Expression project hosted by the Broad Institute, which allows searching of genes and transcripts, and outputs RNA-seq data from a total of 855 different samples, from 544 individuals and across 53 different tissue types. This data set includes RNA-seq across 14 CNS regions including the spinal cord and pituitary gland, and supports our previous findings that EU358092/RP11-272L13.3 is expressed almost exclusively within the brain, spinal cord, and pituitary gland, with some evidence for very low expression in transformed fibroblasts, the sigmoid colon, and the ovary (Figure 3.17).

Data from GTEx is from adult donors, and we therefore note some differences in EU358092/RP11-272L13.3 expression compared to previous work using the Sestan Lab microarray data from the foetal brain.

Figure 3.17 Expression of EU358092/RP11-272L13.3 and MIR137HG in human tissues from the GTEx database.

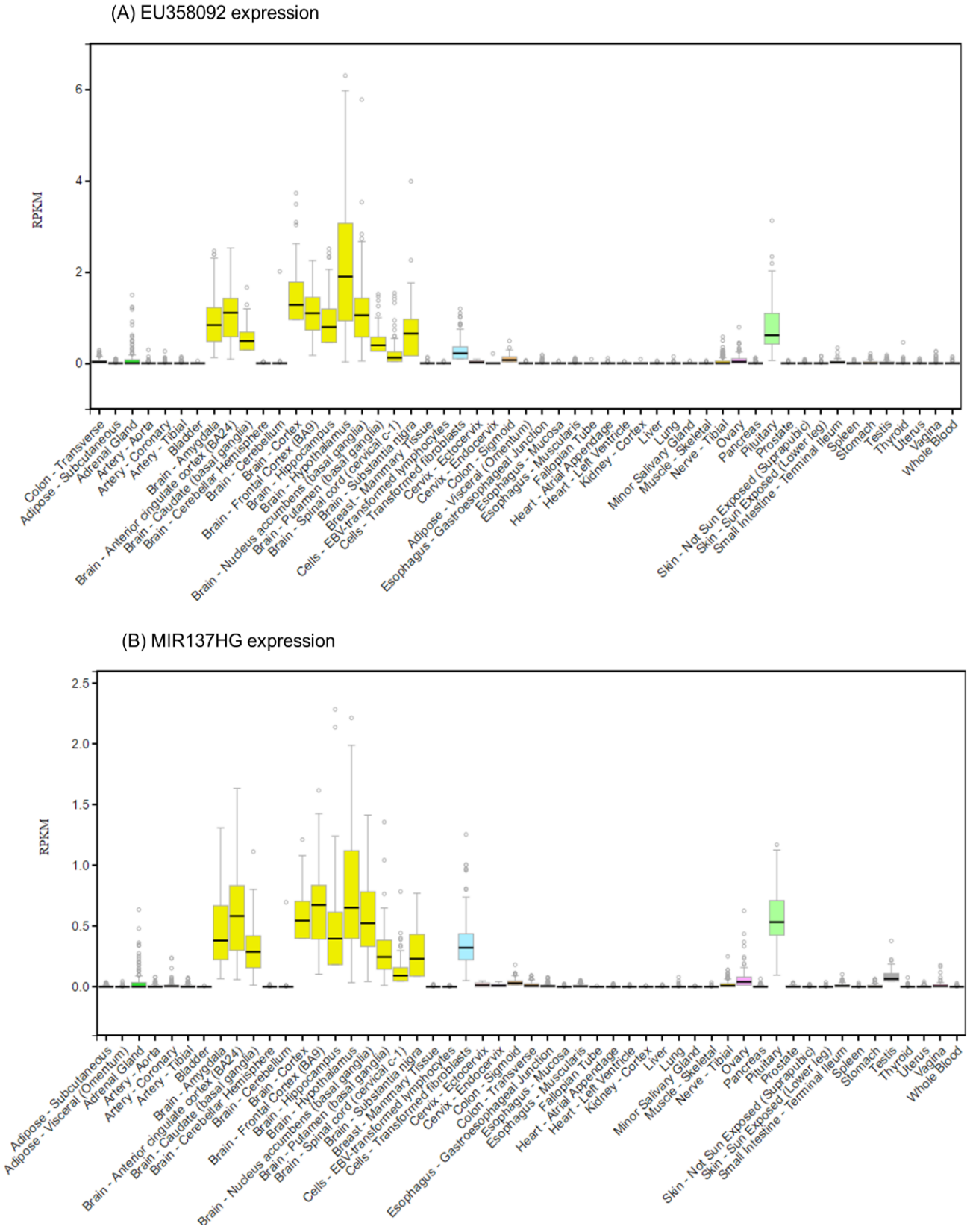


Figure 3.17 Expression of EU358092/RP11-272L13.3 and MIR137HG in human tissues from the GTEx database.

(a) RNA-seq data for EU358092/RP11-272L13.3 expression across 53 adult tissues demonstrated that expression of this transcript was primarily restricted to the CNS, including expression in 10 brain regions, the spinal cord, and the pituitary gland, as well as low expression identified in transformed fibroblasts, the sigmoid colon, and the ovary. The highest expression of EU358092/ RP11-272L13.3 was found across the brain, and particularly in the hypothalamus.

(b) The expression pattern of the precursor MIR137HG in this data set was almost identical to that of EU358092/RP11-272L13.3, with highest expression across the brain and CNS tissues, moderate expression in transformed fibroblasts, and low expression in the sigmoid colon, ovary, and testis. In general, MIR137HG was found to be expressed at higher levels than EU358092/RP11-272L13.3 and displayed almost identical changing expression patterns across the 12 CNS tissues, except for in the frontal cortex (BA9) where MIR137 levels were increased relative to other brain regions, while EU358092/RP11-272L13.3 expression was decreased in this area.

No expression of MIR137 or EU358092/RP11-272L13.3 was identified in the cerebellum. Overall, this would provide further evidence to suggest that MIR137 and the EU358092/RP11-272L13.3 transcript are likely to be co-expressed in vivo in the human brain.

While the two data sets agree on low or no expression of this RNA in the cerebellum, GTEx data from the adult brain shows the highest expression in the hypothalamus, while the area constituting the thalamus was shown to have minimal expression of EU358092 in the foetal brain. Similarly, the adult brain shows expression of the RNA in the hippocampus, which was not observed in the foetal brain. Further, the data from GTEx confirms expression of EU358092/RP11-272L13.3 in 12 of the 14 CNS regions tested, including the amygdala, anterior cingulate cortex (BA24), the caudate, nucleus accumbens, putamen (basal ganglia), both the cortex and frontal cortex (BA9), substantia nigra, cervical C1 spinal cord, and the pituitary gland, as well as regions mentioned above.

As with the NHPRTR and Sestan Lab data sets, we next compared data on EU358092/RP11-272L13.3 with MIR137HG data from GTEx. Most notably, visualisation of RNA-seq data on MIR137 from the GTEx database demonstrates an almost identical expression pattern to EU358092/RP11-272L13.3 across the 53 tissues tested, with near identical expression profiles across different brain regions, with the exception of the frontal cortex in which MIR137 expression is increased relative to other brain regions, whereas EU358092/RP11-272L13.3 is decreased. This may suggest that the two ncRNAs at this locus would continue to be co-regulated both during development and also in the adult brain.

As we have previously identified EU358092/RP11-272L13.3 as residing within a region of high GWAS association for schizophrenia, with multiple GWAS SNPs across its sequence and likely regulatory elements, we next sought to identify whether this RNA may be deregulated in the brains of individuals with a diagnosis of schizophrenia. RNA-seq data from the dorsolateral prefrontal cortex (DL-PFC) of 155 individuals with schizophrenia and 196 controls was generated and analysed by groups at the Lieber

Institute for Brain Development and from Eli Lilly and Company, with data presented in this thesis generated by Dr. Karim Malki, Dr. Nathan Lawless, and Dr. Andrew Jaffe (Section 2.2.8).

RNA-seq data confirmed expression of EU358092/RP11-272L13.3 (referred to as LINC01930 in this data set) in the DL-PFC of both case and control individuals, and statistical analysis performed by Dr. Karim Malki showed that expression of this RNA was significantly different in the DL-PFC of individuals with schizophrenia when compared to controls, with an F statistic of 22.05, and a q-value of 1.67×10^{-52}. The moderated F-statistic (F) is an overall test for significance which summarises multiple t-statistics reported for each individual gene in a model considering all covariates. The q-value is based on the distribution of p-values and is corrected for multiple testing to highlight which genes remain significant after multiple testing and their goodness of fit to the model. However, despite the apparent difference in EU358092 expression between cohorts in this model, we note that the strength of the association is low and that medication effects were unable to be fully accounted for. For this reason, such results should be interpreted with caution. In order to gain a clearer understanding of EU358092 expression in schizophrenia, this would require independent replication and validation through additional studies using different cohorts.

GTEx data confirms and extends our findings from the NHPRTR, which demonstrated that EU358092/RP11-272L13.3 was expressed at a lower level than MIR137HG in general in data from whole brains of humans and multiple primate species. This finding is extended in the GTEx data for both transcripts, showing that EU358092/RP11-272L13.3 is expressed at lower levels than MIR137HG across all specific brain regions and tissues which show expression of both. GTEx data also provided a window into the differences in MIR137HG and EU358092/RP11-272L13.3 expression across the

adult brain in comparison to microarray data from the foetal brain, with the clearest changes being seen in the increased expression of both transcripts in the hippocampus and hypothalamus in adult brains.

Taken together, this data provides further support to the idea that both EU358092/RP11-272L13.3 and MIR137 are likely to be co-expressed and co-regulated across multiple brain regions *in vivo*, with similar patterns of changing expression from foetal brain development and into adulthood. As EU358092 expression was also demonstrated to be altered in the schizophrenia DL-PFC, this transcript may further add to the GWAS association at this locus.

3.9 Discussion

Genome-wide association studies in schizophrenia have repeatedly highlighted the MIR137 locus on chromosome 1p21.3 as being strongly associated with schizophrenia. The location of genome-wide significant SNPs across this locus stretches into the regions flanking MIR137 and overlaps the neighbouring gene, DPYD. The presence of schizophrenia GWAS SNPs in the region between the MIR137 sequence and protein coding regions of DPYD may suggest a regulatory mechanism at this locus that could result in modulation of expression of these genes, with potential additive effects to the association of schizophrenia risk at this locus. As a miRNA, MIR137 will modulate multiple mRNA targets, with consequences at the protein level. These targets include transcripts from many schizophrenia-associated genes, such as CACNA1C (Collins et al. 2014, Kim et al. 2012, Kwon et al. 2013), and thus MIR137 has been highlighted as a potential modulator of CNS function that could reveal underlying schizophrenia biology. Others in the group have previously used bioinformatics and functional genetics to address differential regulation of MIR137

(Warburton et al. 2015a, Warburton et al. 2015b), and we have used similar methods here to identify the potential involvement of the RNA EU358092 in schizophrenia.

EU358092 was found to be as conserved as MIR137 (Figure 3.9) and contained multiple schizophrenia GWAS SNPs across both its sequence and upstream conserved regions (Figure 3.11). Analysis of EU358092 expression *ex vivo* in CNS tissue (Figure 3.12, 3.13, and 3.14) suggested that EU358092 and MIR137 were co-expressed in multiple brain regions in both the foetal and adult human brain. Using the SH-SY5Y neuroblastoma cell line as a model, we demonstrated the stimulus inducible expression of EU358092 in response to drug challenge (Figure 3.15), which was found to be similar to the expression patterns previously published on for MIR137 (Warburton et al. 2015b). These data suggested that both EU358092 and MIR137 were likely to respond to related transcriptional cues both from within the cell and from the environment.

Interrogation of the EU358092 sequence demonstrated that this RNA had many of the characteristics of a lncRNA. Although only a small fraction (less than 2%) of the human genome encodes proteins, the majority of the human genome is capable of being transcribed (Djebali et al. 2012), with lncRNAs known to be the largest class of non-coding transcripts, defined as RNAs of over 200 nucleotides in length which lack protein coding potential. Indeed, the 2012 GENCODE v7 release collected 14,880 annotated human lncRNAs (Derrien et al. 2012). Despite their abundance in the cell, the function of many lncRNAs remains unknown, however, they often display tissue-specific expression patterns and are thought to function as regulators of gene expression through a broad array of mechanisms (Spadaro et al. 2015, Ulitsky and Bartel 2013). Further, lncRNAs are increasingly being found to play roles in neurodevelopmental, neurodegenerative, and psychiatric disorders, as outlined in

Section 1.1.2. (Ziats and Rennert 2013, Wang et al. 2015, Riva, Ratti and Venturin 2016, Liu et al. 2014b).

For example, Liao et al. have demonstrated differences in methylation over lncRNAs in the peripheral blood of both men and women with a diagnosis of schizophrenia (Liao et al. 2015b, Liao et al. 2015a), and work by Ren et al. has shown modulation of lncRNA networks in the peripheral blood of individuals with early onset schizophrenia (Ren et al. 2015c). More specifically, genetic variants of the lncRNA MIAT (also termed Gomafu) have been associated with paranoid schizophrenia in the Han Chinese population (Rao et al. 2015). More global studies by Chen et al. found that 125 lncRNAs were deregulated in peripheral blood mononuclear cells (PBMCs) of individuals with schizophrenia, and that changes in the expression of specific lncRNAs correlated with improved symptoms after treatment with antipsychotic medication (Chen et al. 2016b). With clear differences in lncRNA expression in the blood of individuals experiencing schizophrenia, such differences have been suggested as potential biomarkers for distinguishing schizophrenia from depression and anxiety conditions (Cui et al. 2017a, Cui et al. 2017b).

Analysis of RNA-seq data from the DL-PFC of 155 individuals with schizophrenia and 196 controls confirmed the expression of EU358092 in the adult DL-PFC, with evidence suggesting that the lncRNA was deregulated in the DL-PFC of individuals with schizophrenia. Such data may implicate this lncRNA in schizophrenia biology, and in adding to the GWAS association at this locus. The studies outlined above did not contain data on EU358092, and thus the work presented here would add EU358092 as a lncRNA of interest for further research considering lncRNA deregulation as a mechanism in schizophrenia biology.

Further work to begin to characterise the expression and function of EU358092 would be key to understanding its potential role in the brain and in schizophrenia risk. Methods such as RNA Fluorescent In Situ Hybridisation (RNA-FISH) could be utilised to determine the subcellular localisation of EU358092, which would aid in narrowing down potential functional roles, such as the regulation of gene expression in the nucleus, or sequestering proteins or miRNAs in the cytoplasm (Chen 2016). If results from localisation studies suggested a role in the nucleus, this could be followed up by higher resolution localisation techniques such as ChIRP (Chromatin Isolation by RNA Purification) or RAP (RNA Antisense Purification) to identify areas of chromatin that interact with EU358092. Similarly, RNA-RAP could be utilised to identify RNA-RNA interactions involving the lncRNA (Kashi et al. 2016). If evidence from the above studies indicated a potential role for EU358092 in gene regulation, then over-expression and knock-down studies in a cell line model followed by expression studies such as microarray or RNA-seq analysis may identify potential gene networks that may be regulated by the expression of this lncRNA. Finally, further characterisation of EU358092 expression at different developmental time points in the brain or neural precursors may be of interest to begin to unpick the potential role of this lncRNA in brain function and/or development, and to elucidate time points (if any) at which deregulation of EU358092 could affect development or schizophrenia risk.

3.10 Summary

In this section, bioinformatic analysis of the schizophrenia genome-wide associated region around MIR137 and DPYD using the UCSC Genome Browser identified an uncharacterised, brain-expressed RNA at the chr1:98,399,030-98,407,302 locus, which contained multiple schizophrenia GWAS SNPs both across its sequence and spanning into upstream regulatory regions. We demonstrated that this RNA was as

conserved as many of the neighbouring MIR137 transcripts, with both EU358092 and MIR137 displaying similar characteristics in terms of their expression profiles across primate tissues and across the human brain, as well as in response to drug challenge in a neuroblastoma cell line model. Further, we characterised an upstream evolutionary conserved region, which was shown to be a positive regulator of expression in a reporter gene model. This regulatory region contained a GWAS SNP for schizophrenia and educational attainment, with evidence of active regulatory activity in multiple human brain regions, according to publicly available histone modification data. We further demonstrated that EU358092 is expressed in the adult DL-PFC, where it is deregulated in individuals with schizophrenia. Overall, the work in this section highlights EU358092 as an additional lncRNA for consideration as a transcript of interest in schizophrenia biology.

Chapter 4

The MIR137-REST-EZH2 gene network is altered in schizophrenia.

The work in this section is contained in two manuscripts entitled:

1. The MIR137-REST-EZH2 gene network regulates expression of a large schizophrenia-associated gene set and is altered in the schizophrenia dorsolateral pre-frontal cortex.
2. Regulatory characterisation of the schizophrenia-associated CACNA1C proximal promoter and the potential role for the transcription factor EZH2 in schizophrenia aetiology.

4.1 Introduction

In the previous chapter, we considered non-coding evolutionary conserved regions around the MIR137 locus that could modulate expression of MIR137 and potentially influence schizophrenia risk based on SNPs or regulation of expression in response to the environment. This chapter continues the theme of MIR137 regulation, however, rather than considering genetic elements, we here consider regulatory pathways and transcription factors that may modulate MIR137 expression and downstream networks by binding at the two MIR137 promoters, MIR137HG and the internal MIR137 imir promoter.

In previous investigations of MIR137 regulation, others in the group have used ENCODE ChIP-seq data across the miRNA to identify the transcription factor REST, a well-known regulator of neuronal genes (Qureshi and Mehler 2009, Johnson et al. 2008, Ballas et al. 2005, Gao et al. 2011, Mukherjee et al. 2016, Abrajano et al. 2009), binding at the MIR137 imir promoter. Using luciferase reporter gene assays, it was shown that overexpression of REST modulated the expression of a luciferase reporter gene from the MIR137 imir promoter, and that a schizophrenia GWAS SNP (rs2660304) at this promoter supported differential allele specific expression under basal conditions (Warburton et al. 2015b, Warburton et al. 2016).

In this chapter, we made further use of ENCODE ChIP-seq data to identify EZH2 binding at the promoter of the long MIR137 precursor transcript, MIR137HG, and also binding across the schizophrenia-associated GWAS SNP, rs2660304, adjacent to the MIR137 imir promoter.

EZH2 is a histone-lysine N-methyltransferase enzyme and a known target of MIR137 (Sun et al. 2015a, Ren et al. 2015b, Szulwach et al. 2010), which primarily functions

as the catalytic subunit of the Polycomb Repressive Complex 2 (PRC2) and is known to interact with REST. Tsai et al. demonstrated that the PRC2 can interact with REST via the lncRNA, HOTAIR, with the 5' region of the lncRNA binding to the PRC2 and the 3' region binding to LSD1, which is frequently found in a complex with REST and CoREST (Tsai et al. 2010). However, others have shown that REST can still interact with PRC2 members, EZH2 and SUZ12, even after RNase treatment to degrade HOTAIR, and demonstrate the direct interaction between REST and the PRC2 by co-immunoprecipitation (Dietrich et al. 2012). Within the PRC2, EZH2 works alongside SUZ12 and EED to repress gene expression through tri-methylation of histone 3 at the lysine 27 position (H3K27me3). However, in addition to its function within the PRC2 as a transcriptional repressor, studies in cancer are beginning to highlight a role for EZH2 functioning independently of the PRC2 complex as an activator of gene expression (Lee et al. 2011, Deb, Thakur and Gupta 2013, Lawrence and Baldwin 2016).

When functioning within the PRC2, a key role for EZH2 has been proposed in regulating gene expression during development, and its function in the brain includes roles in regulating neural patterning in the early embryo (Qi et al. 2013). Deletion of EZH2 in mouse models has also demonstrated multiple roles for EZH2/PRC2 in brain development, such as the regulation of neurogenesis (Ronan, Wu and Crabtree 2013), and regulation of the neuronal precursor cell shift from the production of neurons to the production of astrocytes (Hirabayashi et al. 2009). Further, studies in humans have suggested the involvement of EZH2 in multiple neurological conditions. For example, Li et al. have demonstrated that SNPs in EZH2 have been associated with autism risk through family based association studies in the Han Chinese population (Li et al. 2016). EZH2 also lies within the chromosome 7q35-36 locus, which is a known linkage

region for autism and language traits in autistic individuals (Alarcón et al. 2002, Alarcon et al. 2008). Additionally, von Schimmelmann et al. have identified a role for EZH2 in regulating genes involved in neurodegeneration (von Schimmelmann et al. 2016). In the case of schizophrenia, methylation studies in whole blood have identified EZH2 binding at all three of the top methylome-wide associated genes for this condition, SDCCAG8, CREB1, and ATXN7 (Kumar et al. 2015).

In this chapter, we used a combination of bioinformatic analysis and RNA-seq to describe a model in which MIR137, REST, and EZH2/PRC2 form a core regulatory group that is likely to influence the regulation of a much larger downstream network of CNS and schizophrenia-associated genes. This was achieved firstly through the use of ENCODE ChIP-seq data, which identified EZH2 binding across both the MIR137HG and MIR137 imir promoters, suggesting a regulatory feedback loop between these two genes. Using the UCSC Genome Browser and Galaxy, we overlaid ENCODE ChIP-seq data onto all annotated transcriptional start sites for the 78,807 transcripts in the hg19 ‘known gene’ data set, which generated gene lists with EZH2 alone, EZH2 and SUZ12 (as a proxy for EZH2 binding within the PRC2 complex), or REST binding within 500 bp of their transcriptional start sites. Running these gene lists through an enrichment analysis tool allowed us to identify pathways in which these REST or EZH2/PRC2 target gene sets may be involved, and allowed us to build a model describing the potential roles for these genes in schizophrenia.

Finally, this model was tested in clinical samples using RNA-seq data from the DL-PFC of 155 individuals with schizophrenia and 196 controls, which was analysed by groups at Eli Lilly and the Lieber Institute for Brain Development.

4.2 Aims

- Interrogate ENCODE ChIP-seq data to identify transcription factors that bind across MIR137 promoters.
- Bioinformatic analysis using ENCODE ChIP-seq data and Galaxy to identify REST and EZH2 target gene networks.
- Perform enrichment analysis on the above gene sets in order to identify pathways and processes that may be affected by changes in the MIR137-REST-EZH2 network.
- Use RNA-seq data from the DL-PFC of individuals with schizophrenia and controls in order to analyse the MIR137-REST-EZH2 model as a potential mechanism in schizophrenia biology.

4.3 Results

4.3.1 Bioinformatic analysis shows REST and EZH2 binding at the MIR137 promoters

In order to investigate potential pathways and transcription factors that regulate expression of MIR137, we addressed ENCODE ChIP-seq data, which is available to view across the genome through the UCSC Genome Browser. Work by Warburton et al. has previously used such data to observe REST binding over the MIR137 imir promoter, and validated this by overexpression of REST in cell line models, which was shown to modulate endogenous MIR137 expression (Warburton et al. 2015b). Here we use a similar strategy, but extend our study using bioinformatic approaches to build up regulatory networks that may be important in the brain and affected in schizophrenia.

ENCODE ChIP-seq data shows four transcription factors – REST, EZH2, RNA Polymerase II, and CEBP – binding at the MIR137 imir promoter, and eight transcription factors binding directly over the MIR137HG promoter – REST, EZH2, CTCF, GATA2, GATA3, RAD21, SIN3AK20, and YY1 (Figure 4.1). We also note 25 further instances of transcription factor binding within the first intron of MIR137HG, which may have the potential to impact regulation from one or both promoters.

As REST and EZH2 are the only two transcription factors in this data set which are present at both the MIR137 imir and the MIR137HG promoters, we carried these forward for further bioinformatic analysis. According to ENCODE ChIP-seq, REST binding at the MIR137 imir promoter was seen in the PFSK-1 neuroectodermal cell line, and EZH2 binding across this region was identified in the H1 human embryonic stem cell line (H1-hESC), NHEK epidermal keratinocytes, and the Dnd41 T cell leukaemia cell line.

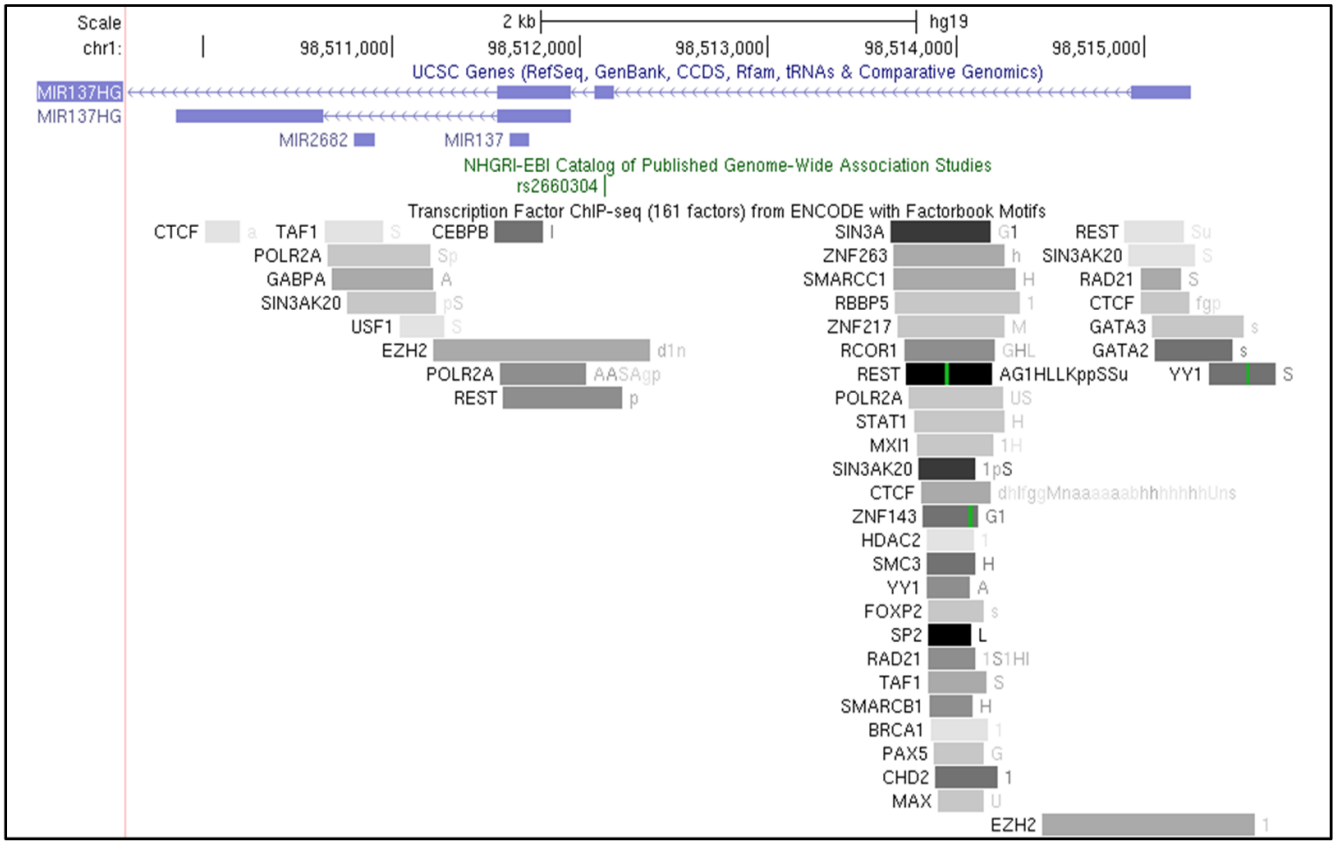


Figure 4.1 ENCODE ChIP-seq data on transcription factor binding over the MIR137 promoters.

The UCSC Genome Browser was used to visualise ENCODE ChIP-seq and NHGRI-EBI GWAS data over the region encompassing the MIR137 promoters (GRCh37/hg19). ENCODE ChIP-seq data shows eight transcription factors binding directly over the transcriptional start site or first exon of the long MIR137HG precursor transcript, and four transcription factors binding at the MIR137 imir promoter adjacent to the schizophrenia GWAS SNP, rs2660304. Numerous other transcription factors were identified as binding around the MIR137 introns and may also have the ability to modulate gene expression across this region. EZH2 and REST were the only two transcription factors identified as binding at the transcriptional start sites of both the MIR137HG and the MIR137 imir promoter. Further, the EZH2 and REST ChIP-seq signals are the only two which overlap the schizophrenia-associated GWAS SNP, rs2660304.

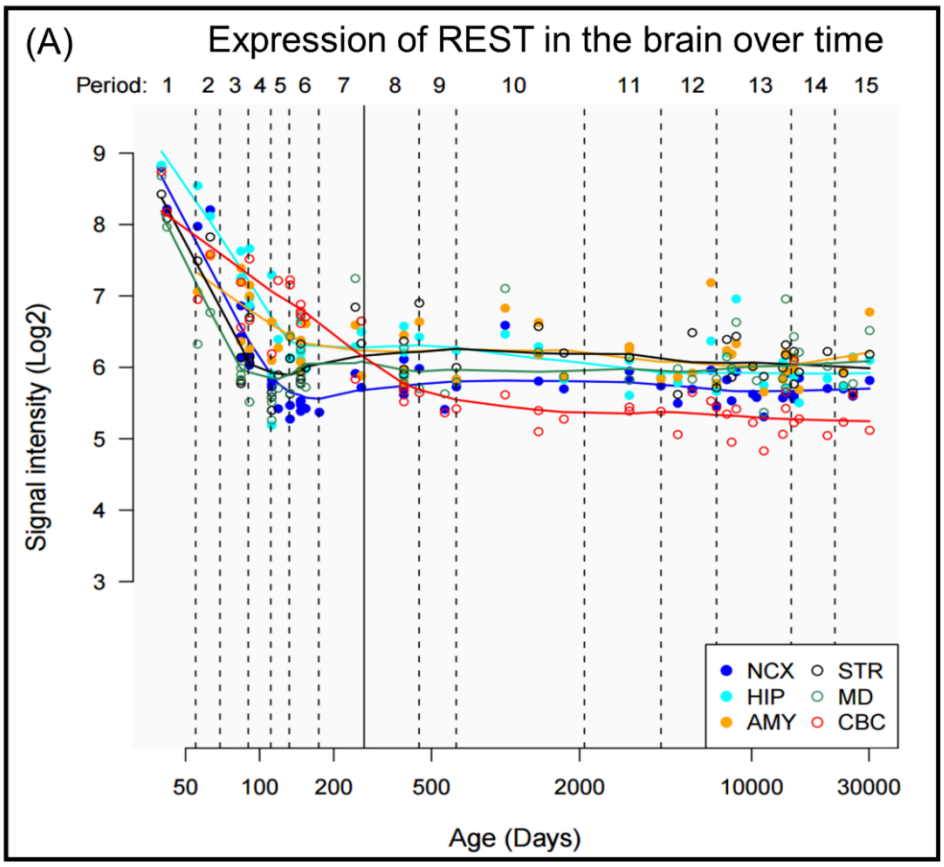
Of particular interest, we find that the ENCODE ChIP-seq signals for REST and EZH2 at this promoter are the only two that overlap the nearby schizophrenia GWAS SNP, rs2660304 (Figure 4.1), which is known to support allele-specific expression from the MIR137 imir promoter based on SNP genotype (Warburton et al. 2016). At the MIR137HG promoter, REST was identified as binding at this region in the brain-derived SK-N-SH neuroblastoma cell line, and the U87 glioblastoma cell line, whereas EZH2 was again identified as binding at the MIR137HG promoter in H1-hESCs.

4.3.2 REST and EZH2 are highly expressed across the developing brain, with expression plateauing around birth

In order to gather information on REST and EZH2 expression in the brain at different time points, we made use of data collected in the Human Brain Transcriptome Resource (http://hbatlas.org/) to observe expression of REST and EZH2 in multiple brain regions across the lifetime, from the embryonic stages to late adulthood (60 years and over). The Human Brain Transcriptome includes data across six brain regions – the neocortex (NCX), hippocampus (HIP), amygdala (AMY), striatum (STR), mediodorsal nucleus of the thalamus (MD), and cerebellar cortices (CBC) – from a total of 57 clinically and genetically healthy individuals (31 males and 26 females), ranging from 5.7 weeks' post-conception to 82 years of age.

Using this data set, we find that both REST and EZH2 show broadly similar expression patterns across the brain, with the highest levels of both genes seen in all brain areas tested at the earliest time point during the embryonic stages at 4-8 weeks' post-conception (Figure 4.2).

Figure 4.2 REST and EZH2 expression profiles in the brain through the lifetime.



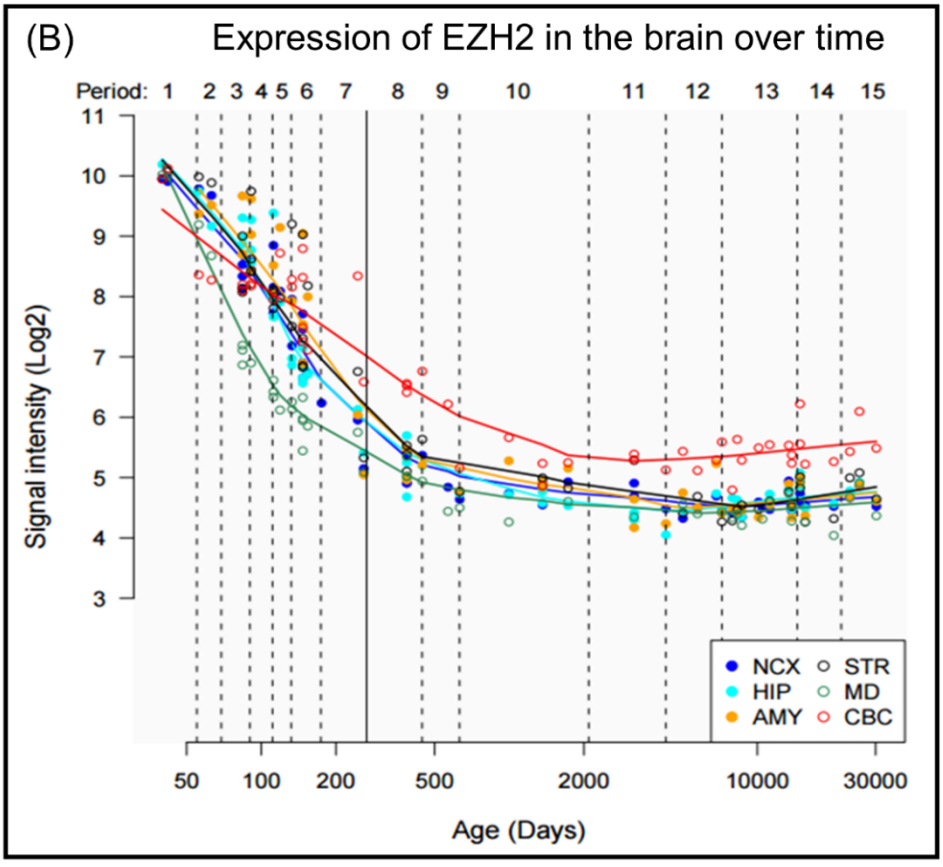


Figure 4.2 REST and EZH2 expression profiles in the brain through the lifetime.

Data for REST and EZH2 expression across the brain, from embryonic development through to 82 years of age, was accessed through the Human Brain Transcriptome resource.

(a) We find the highest expression of REST across the brain in the earliest embryonic and foetal stages at 4-8 weeks' post-conception (Period 1), with expression decreasing across the brain through development and plateauing in the majority of brain regions between 19-24 weeks' post-conception (Period 6), with the exception of REST expression in the cerebellar cortices, which plateaus at around 6-12 months after birth.

(b) Similarly, EZH2 expression is highest in the brain during the embryonic stages of development (4-8 weeks' post-conception, Period 1). EZH2 expression then decreases globally across the brain throughout foetal development and plateaus shortly after birth at around 6-12 months of age (Period 9). This would suggest that REST and EZH2 in the brain are predominantly involved in development, after which expression is restricted to lower levels throughout the rest of the lifetime.

The Human Brain Transcriptome resource contains gene expression data from 57 individuals (31 males and 26 females), testing 16 brain regions including 11 regions of the neocortex (NCX), and one region from the hippocampus (HIP), amygdala (AMY), striatum (STR), mediodorsal nucleus of the thalamus (MD), and the cerebellar cortex (CBC). Periods 1-15 represent different stages of development and age, as detailed in Kang et al. (Kang et al. 2011). Periods 1-7 encompass the embryonic and foetal development stages, whereas periods 8-15 refer to the ages of birth to 60+ years of age. The solid vertical between period 7 and 8 indicates birth.

Expression of both REST and EZH2 then rapidly declines during foetal development, with REST reaching its lowest point in the striatum and the mediodorsal nucleus of the thalamus around the early-mid foetal stage at 13-16 weeks' post-conception, and REST expression in the neocortex drops to its lowest across the lifetime at the late-mid foetal stage (19-24 weeks' post-conception) (Figure 4.2a). There is a small resurgence of REST expression in these regions before birth, which then plateaus over the lifetime.

The lowest starting expression of REST in the embryonic brain is seen in the amygdala, but as expression decreases across all brain regions until birth, REST expression in the amygdala then becomes one of the highest across the six brain regions tested (Figure 4.2a). The cerebellar cortices are the main outlier, with REST expression continuing to decrease in this brain region even after birth, only beginning to flatten out at around mid to late childhood (6-12 years old), and then remaining expressed at lower levels than in other brain regions (Figure 4.2a).

Expression of EZH2, similar to REST, is at its highest levels across the brain regions tested during early embryonic development, and decreases steadily throughout foetal development (Figure 4.2b). The decrease in EZH2 expression during foetal development is most pronounced in the mediodorsal nucleus of the thalamus, whereas expression in the cerebellar cortices starts lower than the other brain regions and sees a shallower drop over the course of development. EZH2 expression in the remaining brain areas tested is comparable between regions, with very similar expression patterns in the neocortex, hippocampus, amygdala, and striatum. While diverging from expression in other areas during development *in utero*, EZH2 expression in the mediodorsal nucleus of the thalamus comes to more closely mirror that of the

neocortex, hippocampus, amygdala, and striatum after birth, while narrowly remaining at the lowest level of all areas tested.

EZH2 expression in the cerebellar cortices stands out as beginning at the lowest for all regions tested in the early embryonic stages, but then has the highest levels of expression from the early mid-foetal stage (16 to 19 weeks' post conception) and throughout the lifetime. Expression of EZH2 in all brain regions tested decreases steadily from the highest levels at 4-8 weeks' post-conception, plateauing around early childhood at around 1-6 years of age.

With the expression patterns of both REST and EZH2 now identified across the brain during both healthy development and over the lifetime, we can extend our hypothesis as to the potential for REST and EZH2 deregulation as a mechanism in schizophrenia. Both REST and EZH2 are highly expressed in the developing brain during the embryonic and foetal stages. It is therefore possible that altered REST or EZH2 expression in the brain during this time period could contribute to the neurodevelopmental aspects of schizophrenia, however, this is not a hypothesis that we were able to test. Alternatively, the expression of REST and EZH2 is maintained at low levels across the brain throughout the lifetime after birth. A second hypothesis, and one that can be tested with the samples available to us, would be that aberrant reactivation of REST or EZH2 in the adult brain may modulate the expression of MIR137 and other schizophrenia-associated targets, thereby contributing to schizophrenia risk.

Comparing RNA-seq data from individuals with schizophrenia who had been exposed to known neurodevelopmental risk factors, such as early life stress or maternal infection, would be of interest in attempting to elucidate potential long term

transcriptional changes in individuals with exposure to known neurodevelopmental risk factors. However, no information was available to us regarding exposure to neurodevelopmental risk factors for the RNA-seq sample set available, and no RNA-seq data was available to us from, for example, iPSCs from individuals with schizophrenia that had undergone treatment to mimic exposure to such risk factors.

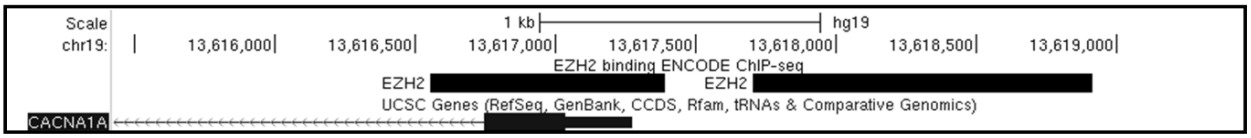
4.3.3 Identifying REST and EZH2 target genes using ENCODE ChIP-seq data

Having extended our hypothesis through the identification of REST and EZH2 expression patterns over the lifetime, we next compiled lists of all genes that were expected to be modulated by these transcriptional regulators (Section 2.2.9.2). Through simple observation of genes of interest, we determined that EZH2 was likely to regulate many gene sets with members which are associated with schizophrenia, such as calcium channels (CACNA1A, CACNA1B, CACNA1C, CACNA1D, CACNA1E, CACNA1G, CACNA1H, CACNA2D3, CACNB2, CACNG4, CACNG6, CACNG8), glutamate receptors (GRIN1, GRIN2A, GRIN2B, GRIN2C, GRIN2D, GRIN3A, GRIN3B, GRIA2, GRIA4), and dopamine receptors (DRD1, DRD2, DRD4, DRD5) (Figure 4.3).

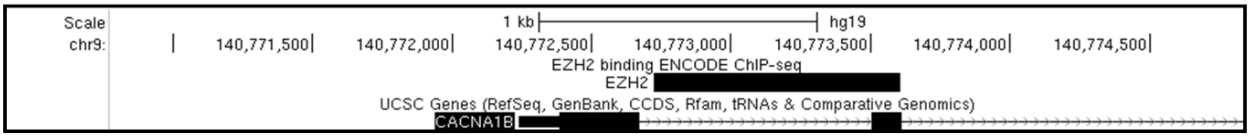
To address global REST and EZH2 targets, we used Galaxy to access and overlay ENCODE ChIP-seq data and annotated transcriptional start site data available through the UCSC Genome Browser. Firstly, Galaxy was used to download the co-ordinates of all transcriptional start sites listed in the UCSC Genome Browser's 'known gene' data set for the hg19 genome build. We added 500 bp either side of each transcriptional start site to represent a stringent 1 kb minimal promoter region. Galaxy was further used to download the co-ordinates of all ENCODE ChIP-seq signals for EZH2, REST, and SUZ12.

Figure 4.3 ENCODE ChIP-seq data for EZH2 at the promoters of multiple schizophrenia-associated genes.

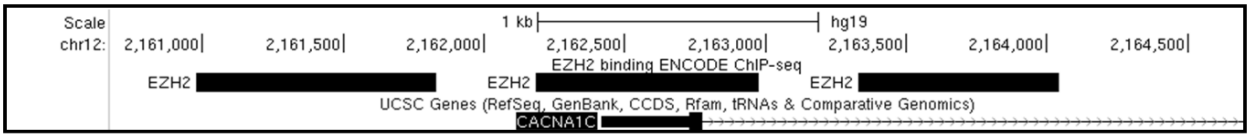
CACNA1A



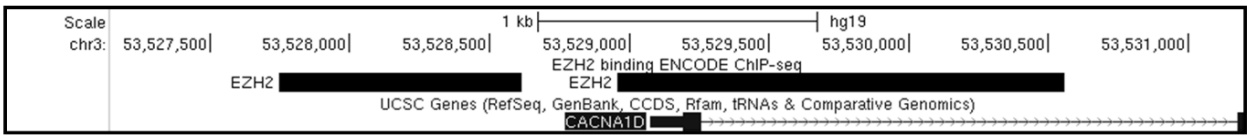
CACNA1B



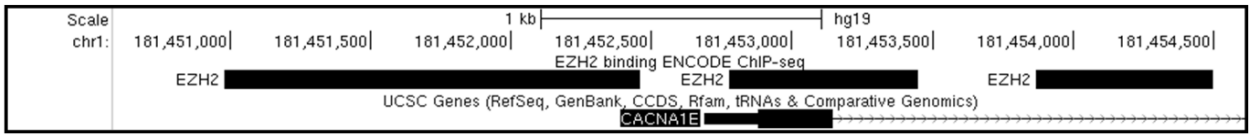
CACNA1C



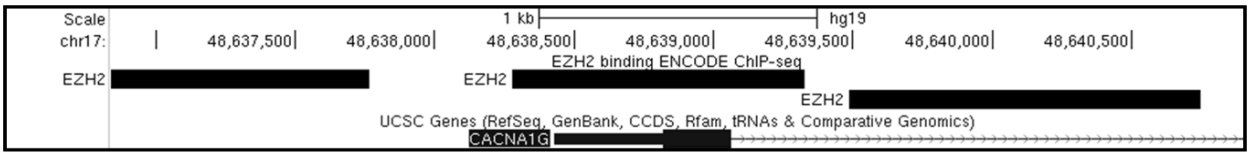
CACNA1D



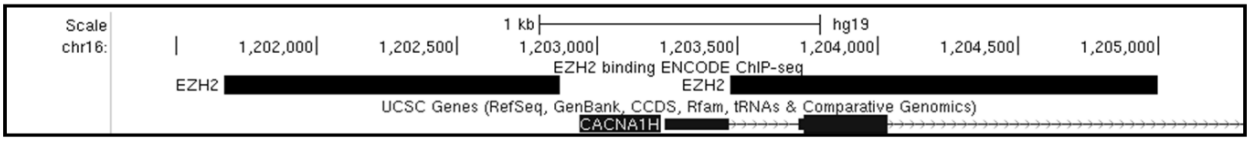
CACNA1E



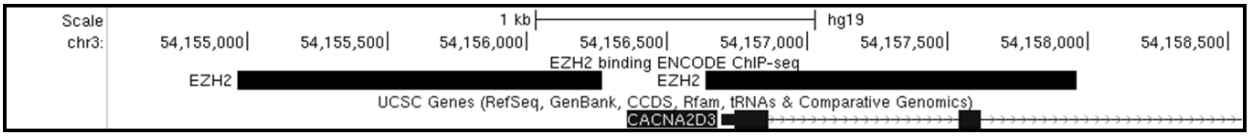
CACNA1G



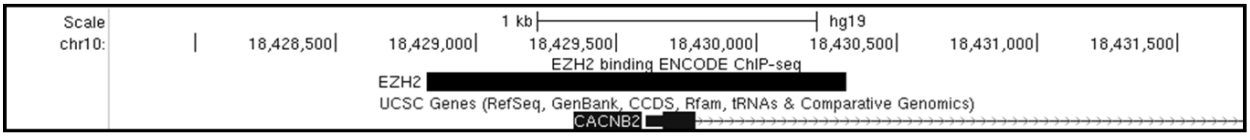
CACNA1H



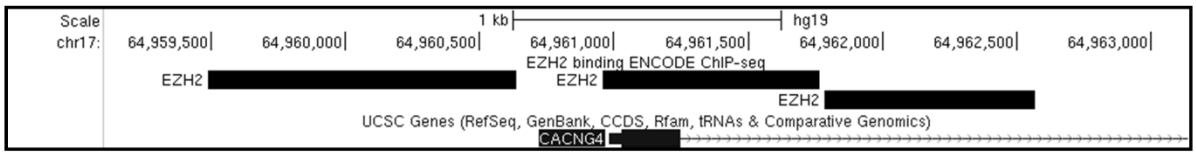
CACNA2D3



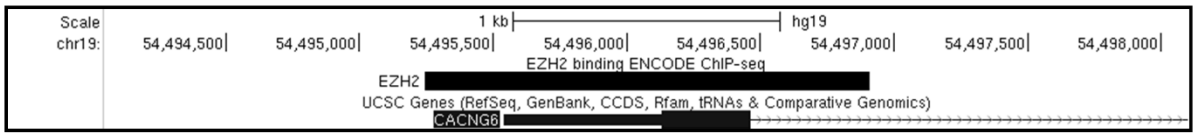
CACNB2



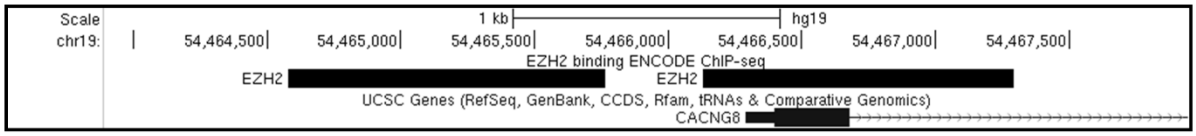
CACNG4



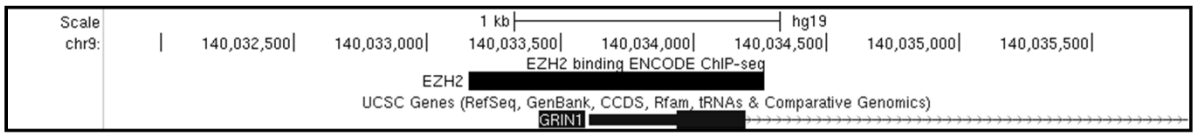
CACNG6



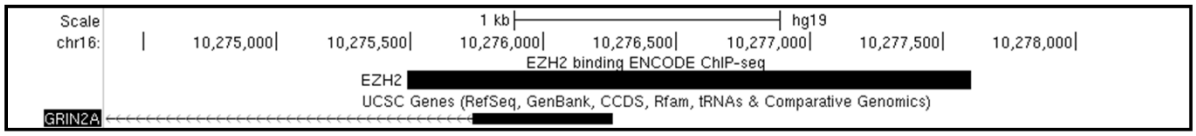
CACNG8



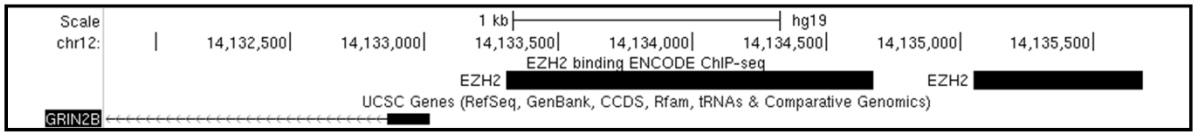
GRIN1



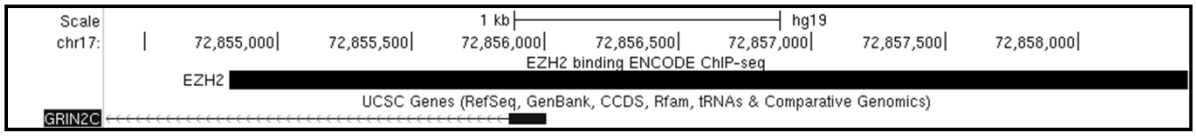
GRIN2A



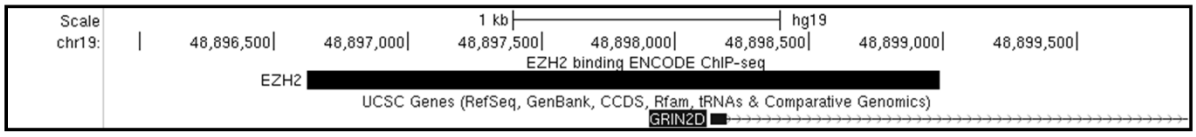
GRIN2B



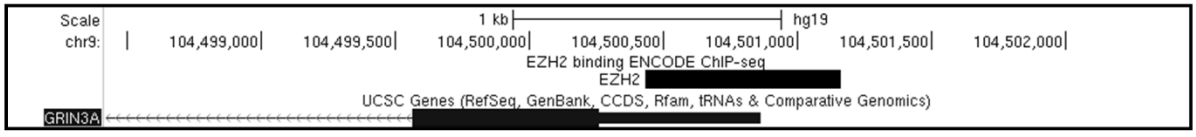
GRIN2C



GRIN2D



GRIN3A



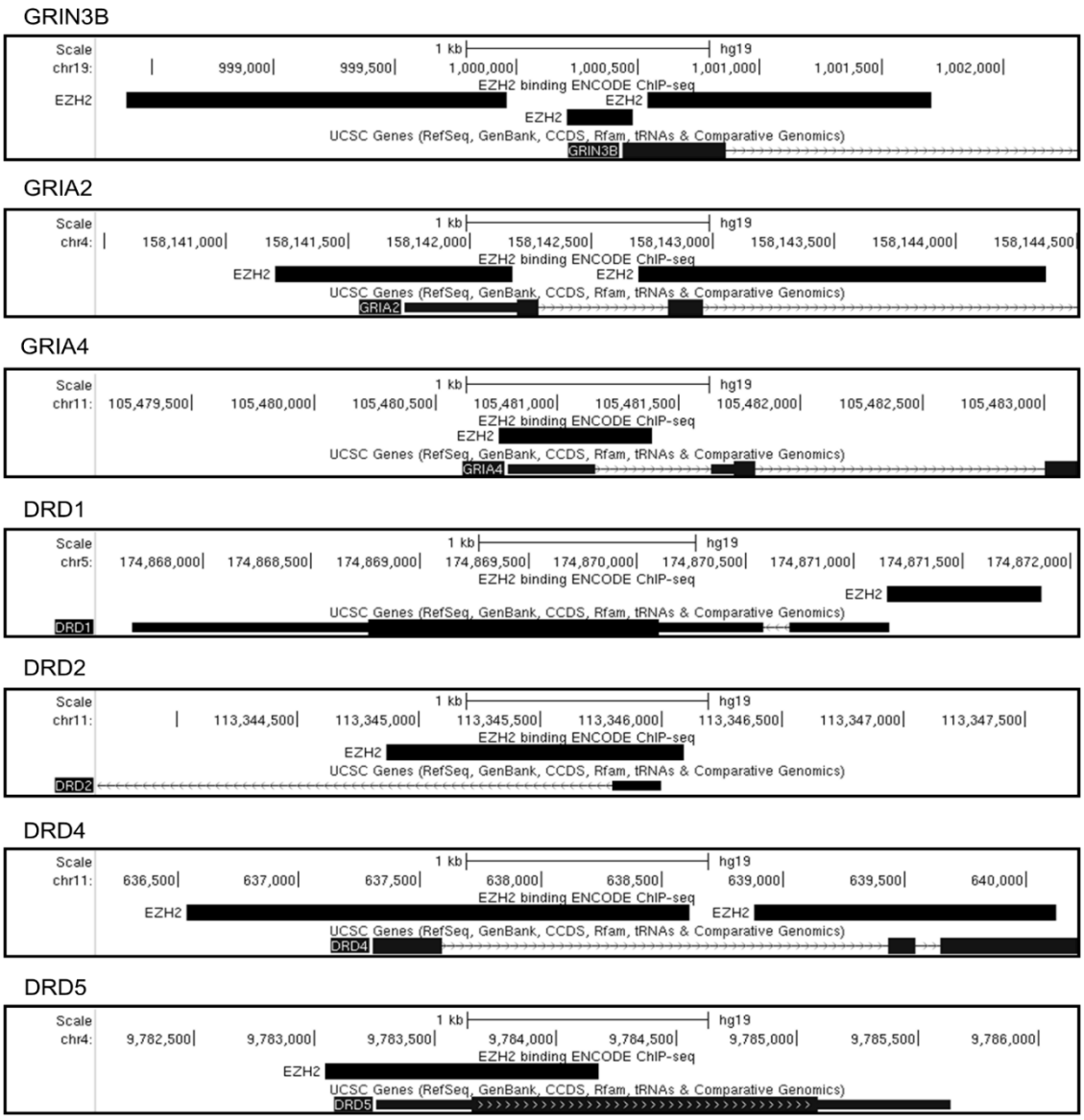


Figure 4.3 ENCODE ChIP-seq data for EZH2 at the promoters of multiple schizophrenia-associated genes.

Interrogation of ENCODE ChIP-seq data for EZH2 using a candidate gene approach to view genes on the UCSC Genome Browser demonstrated EZH2 binding across or around the transcriptional start sites of numerous genes known to be associated with schizophrenia, including multiple genes from schizophrenia-associated gene families such as those encoding subunits of calcium channels, glutamate receptors, and dopamine receptors. Given the presence of EZH2 at the promoters of numerous hand selected CNS genes, we next extended our study to consider all genes with ChIP-seq evidence of EZH2 binding at one or more of their transcriptional start sites.

These data sets were saved as custom UCSC tracks (which are available as .BED files on the supplementary data disk; Supplementary File 4.1, 4.2, 4.3) and uploaded to the UCSC Genome Browser. This allowed us to overlay the transcriptional start site co-ordinates (plus and minus 500 bp; Supplementary File 4.4) with each data set for EZH2, REST, and SUZ12 to generate lists of genes with binding of these factors at the minimal promoter of one or more of their transcripts, according to ENCODE ChIP-seq data. For the gene list with both EZH2 and SUZ12 binding, representing PRC2 binding, we took all genes with EZH2 binding within 500 bp either side of the transcriptional start site and further narrowed down this list by keeping only those which also displayed SUZ12 binding within an area of 1 kb from the EZH2 signal.

Overall, this generated four gene lists for analysis which had the following factors binding within 500 bp of one or more transcriptional start sites:

1) EZH2 alone – 1780 genes
2) EZH2 and SUZ12 (PRC2) – 900 genes
3) REST – 3691 genes
4) SUZ12 alone – 91 genes

4.3.4 Enrichment analysis of genes with REST and EZH2 ENCODE ChIP-seq data supports a role for REST- and EZH2-mediated modulation of MIR137 and larger schizophrenia-associated gene networks

Using information on REST and EZH2 expression across the brain from the Human Brain Transcriptome, we hypothesised that there may be two timepoints at which deregulation of the MIR137-REST-EZH2 pathway may contribute to schizophrenia risk. Firstly, if this pathway is disrupted *in utero*, this may contribute to the neurodevelopmental risk factors for schizophrenia. Secondly, as we see that REST

and EZH2 remain at lower levels across the brain throughout the lifetime from shortly after birth, we would theorise that the inappropriate re-activation of EZH2 and/or REST in the adult brain could contribute to schizophrenia risk in the adult, either directly by modulating expression of their own downstream targets, or indirectly through modulating MIR137 expression, which would feedback onto the regulation of MIR137's schizophrenia-associated target genes.

Overall, we predicted that MIR137, REST, and EZH2 formed a key regulatory pathway, with the genes functioning to regulate both each other in addition to their own distinct gene sets. In this way, changes in any one of the three genes would likely disrupt the balance of this pathway, resulting in altered regulation of a predicted set of schizophrenia-associated genes (Figure 4.4).

In order to gain a greater understanding of the downstream pathways and processes that may be altered by deregulation of the MIR137-REST-EZH2 network, and how this could influence schizophrenia risk, we performed multiple enrichment analyses on the EZH2, PRC2, and REST target gene lists using the Ma'ayan Lab Enrichr tool (Chen et al. 2013, Kuleshov et al. 2016).

Enrichment analysis for the 1780 EZH2 target genes using the 5192 terms listed in the Gene Ontology's (GO) biological processes set demonstrated enrichment for EZH2 target genes in processes involving:

- Behaviour — adjusted $p = 1.69 \times 10^{-9}$
- Synaptic transmission — adjusted $p = 1.40 \times 10^{-8}$
- Single organism behaviour — adjusted $p = 2.07 \times 10^{-5}$
- Multicellular organismal response to stress — adjusted $p = 1.46 \times 10^{-4}$
- Regulation of synaptic transmission — adjusted $p = 2.55 \times 10^{-3}$

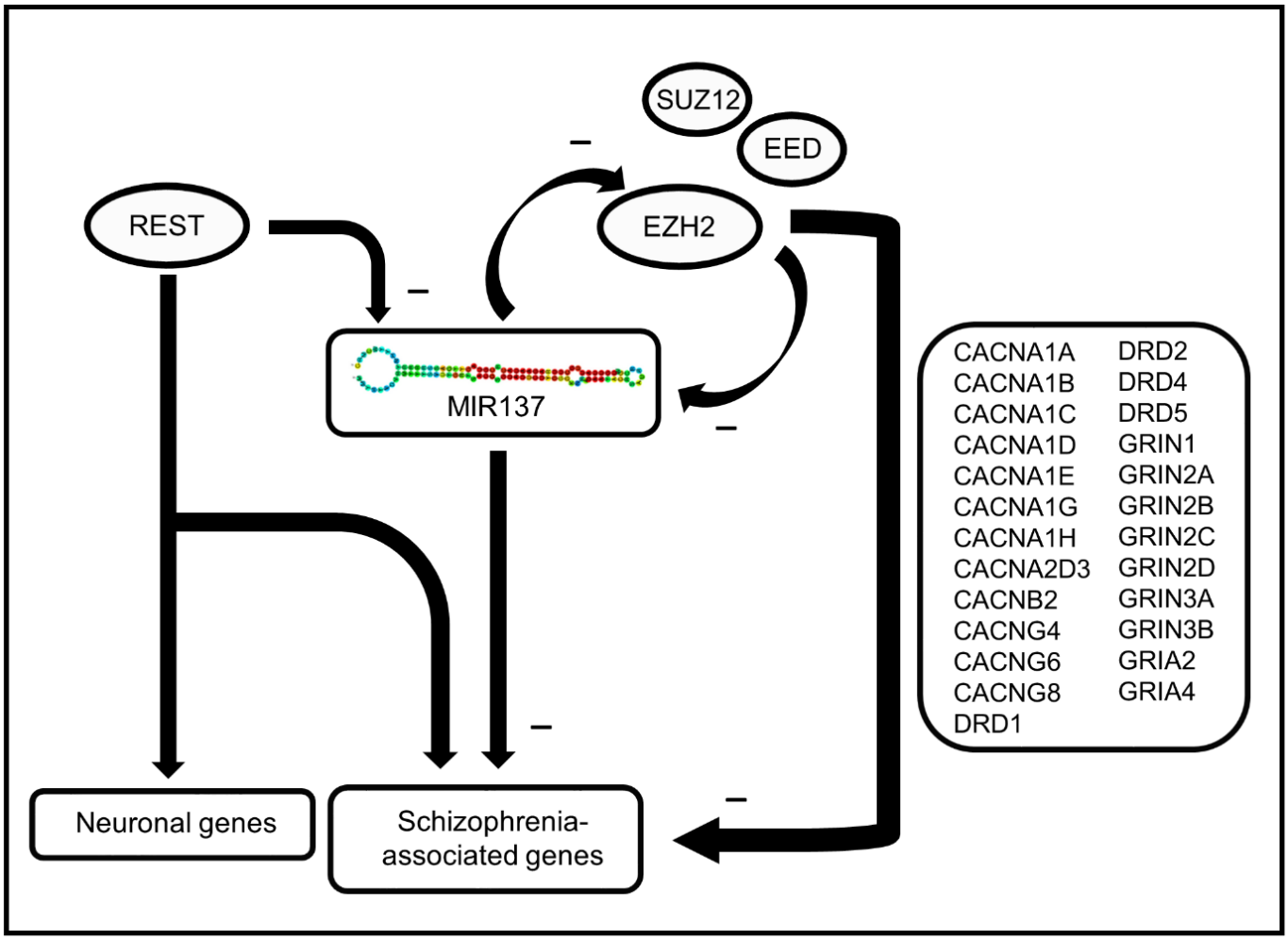


Figure 4.4 Proposed MIR137-REST-EZH2 regulatory network.

MIR137 has been repeatedly shown to be associated with schizophrenia through GWAS, as well as being a known regulator of other schizophrenia-associated genes. REST is a well-established regulator of neuronal genes, and previous work in the lab has validated REST as a regulator of MIR137 (Warburton et al. 2015b). Further, EZH2 is a known target of MIR137, with ENCODE ChIP-seq data demonstrating EZH2 binding at the promoters of MIR137, as well as numerous other genes that have been implicated in schizophrenia biology. In addition to this, REST has been shown to interact with EZH2 within the PRC2 both directly and also through the lncRNA, HOTAIR (Dietrich et al. 2012, Tsai et al. 2010). The former would suggest a regulatory feedback loop between MIR137 and EZH2, and would overall suggest a three-gene network which functions together to regulate the expression of MIR137, REST, and EZH2, and which in turn would influence the expression of each of their distinct CNS- and schizophrenia-related gene sets.

- Learning and memory adjusted p = 2.50 $x10^{-3}$
- Cognition adjusted p = 3.1 $x10^{-3}$

All p-values above are derived from the Fisher's exact test and adjusted using the Benjamini-Hochberg procedure. Many more biological processes with statistical significance of $p < 0.05$ (Benjamini-Hochberg adjusted Fisher's exact p) were also identified, including behavioural defence and behavioural fear responses (Figure 4.5a and Supplementary Data 4.1).

In this context, the definition of the Gene Ontology term 'behaviour' is defined on the Mouse Genome Informatics online resource (http://www.informatics.jax.org/vocab/gene_ontology/GO:0007610) as "the internally coordinated responses (actions or inactions) of animals (individuals or groups) to internal or external stimuli, via a mechanism that involves nervous system activity.

To gain a more complete understanding of the functions associated with each gene set, we correlated biological processes with *in vivo* data, using the Mouse Genome Informatics (MGI) phenotype database to address phenotypes most frequently associated with knockout or mutation of the target genes in each list.

Of the possible 476 terms, mouse phenotypes associated with knockout or mutation of the 1780 EZH2 target genes supported the biological processes data, with highest enrichment for:

- Abnormal synaptic transmission adjusted p = 1.18 $x10^{-8}$
- Abnormal learning and memory adjusted p = 1.99 $x10^{-4}$
- Abnormal emotional and affective behaviour adjusted p = 3.20 $x10^{-3}$
- Abnormal nervous system adjusted p = 3.73 $x10^{-3}$

Figure 4.5 Enrichment analysis of the EZH2 target gene list suggests involvement in behaviour and synaptic transmission.

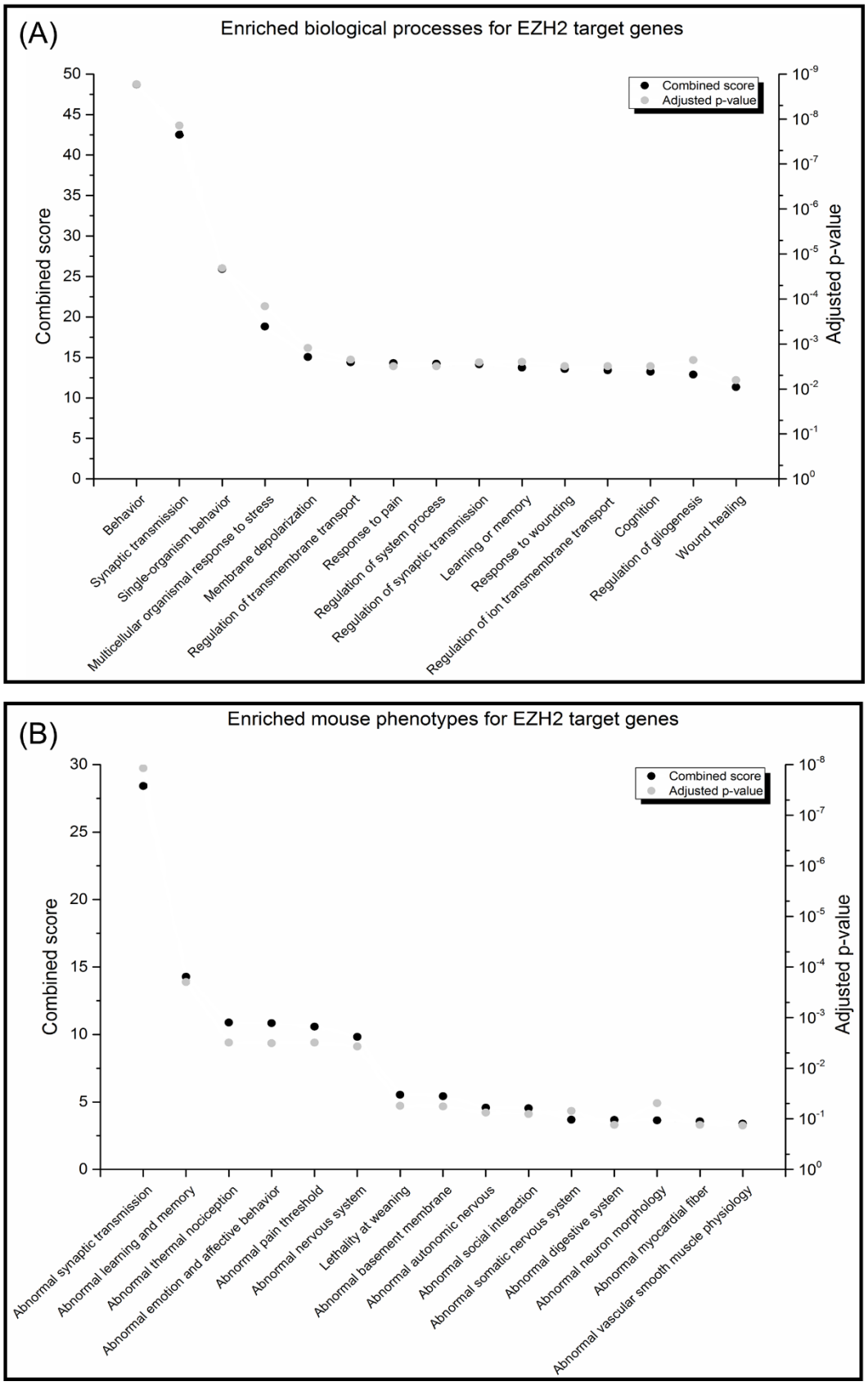


Figure 4.5 Enrichment analysis of the EZH2 target gene list suggests involvement in behaviour and synaptic transmission.

(a) Enrichment for biological process involving target genes with EZH2 binding alone demonstrated highest enrichment for roles in behaviour (Benjamini-Hochberg adjusted Fisher's exact p = 1.69 x10^{-9}) and synaptic transmission (Benjamini-Hochberg adjusted Fisher's exact p = 1.40 x10^{-8}), in addition to further neurological processes including regulation of synaptic transmission (Benjamini-Hochberg adjusted Fisher's exact p = 2.55 x10^{-3}), learning and memory (Benjamini-Hochberg adjusted Fisher's exact p = 2.50 x10^{-3}), and cognition (Benjamini-Hochberg adjusted Fisher's exact p = 3.13 x10^{-3}). We also note enrichment of processes involved in response to stress or other stimuli, such as single organism behaviour (Benjamini-Hochberg adjusted Fisher's exact p = 2.07 x10^{-5}), multicellular organismal response to stress (Benjamini-Hochberg adjusted Fisher's exact p = 1.46 x10^{-4}), and response to pain (Benjamini-Hochberg adjusted Fisher's exact p = 3.13 x10^{-3}). This would suggest that, in healthy individuals, the gene set targeted by EZH2 alone plays a role in regulating behaviour and synaptic transmission, as well as being involved in more general cognitive and stimulus response processes.

(b) Enrichment of the same gene set in the Mouse Genome Informatics database supported the above findings from biological processes data, demonstrating the clearest enrichment for phenotypes involving abnormal synaptic transmission (Benjamini-Hochberg adjusted Fisher's exact p = 1.18 x10^{-8}) and abnormal learning and memory (Benjamini-Hochberg adjusted Fisher's exact p = 1.99 x10^{-4}) in mice with disruption or deletion of genes in this target list. Other enriched phenotypes with relevance to schizophrenia aetiology include abnormal emotional and affective behaviour (Benjamini-Hochberg adjusted Fisher's exact p = 3.20 x10^{-3}), and abnormal nervous system (Benjamini-Hochberg adjusted Fisher's exact p = 3.73 x10^{-3}). We further note enrichment for processes related to altered sensory perception, such as abnormal thermal nociception (Benjamini-Hochberg adjusted Fisher's exact p = 3.13 x10^{-3}) and abnormal pain threshold (Benjamini-Hochberg adjusted Fisher's exact p = 3.13 x10^{-3}), which are in agreement with the biological processes data suggesting that the EZH2 alone gene set is involved in response to pain.

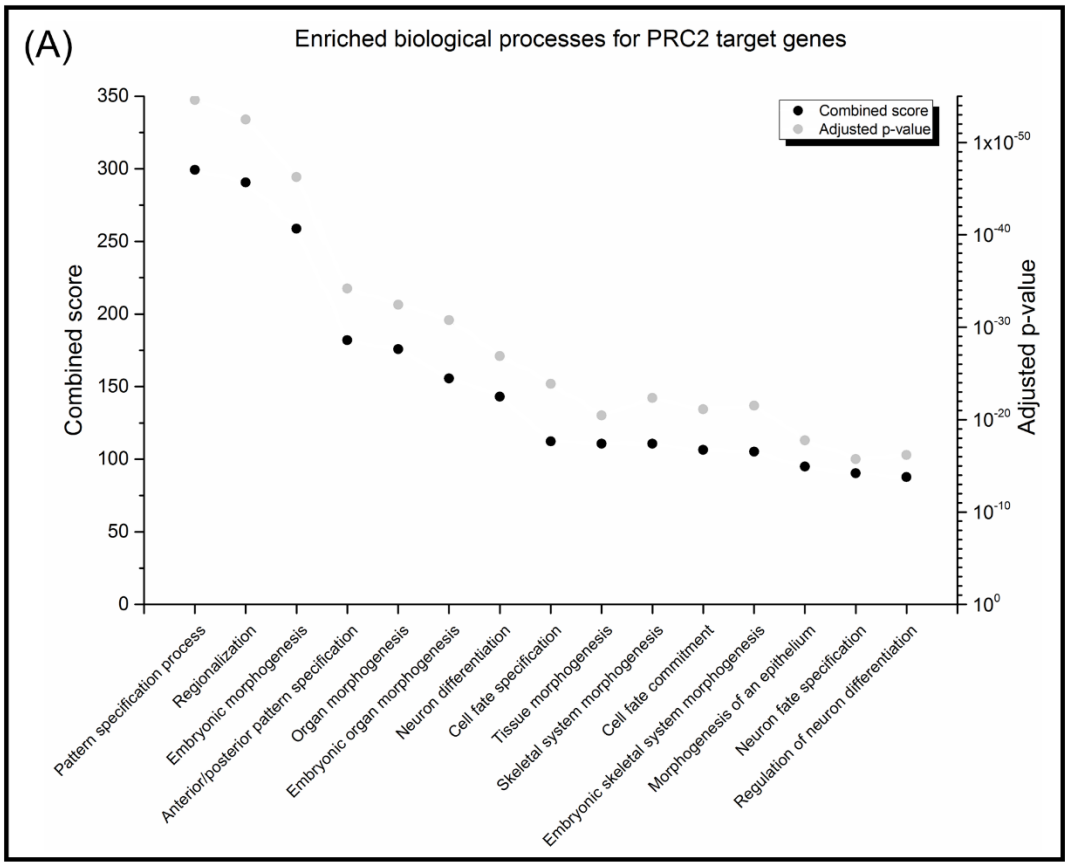
All p-values above are derived from the Fisher's exact test and adjusted using the Benjamini-Hochberg procedure. Our enrichment analyses also suggest a role for EZH2 and its 1780 target genes in the perception and response to pain, with response to pain, abnormal thermal nociception, and abnormal pain threshold all being significantly enriched with a Benjamini-Hochberg adjusted Fisher's exact p-value of 3.13×10^{-3} (Figure 4.5b, and Supplementary Data 4.1). This would be in line with studies reporting altered pain processing and perception in animal models and some individuals with a diagnosis of schizophrenia (Walsh et al. 2010, Boettger, Grossmann and Bar 2013, Minichino et al. 2016), as well as a newly discovered role for EZH2 in regulating stress response to heat in mice (Zovoilis et al. 2016).

Enrichment for biological processes associated with the 900 PRC2 target genes strongly supported the literature in suggesting a role for PRC2-regulated genes in embryonic development, and included processes such as pattern specification and regionalisation in the embryo, as well as pathways regulating embryonic and tissue morphogenesis. Beyond embryonic developmental processes, however, we also found this gene set to be highly enriched for roles in:

- Neuron differentiation — adjusted $p = 1.32 \times 10^{-27}$
- Neuron fate specification — adjusted $p = 1.82 \times 10^{-16}$
- Regulation of neuron differentiation — adjusted $p = 6.80 \times 10^{-17}$
- Positive regulation of nervous system development — adjusted $p = 6.30 \times 10^{-13}$

All p-values above are derived from the Fisher's exact test and adjusted using the Benjamini-Hochberg procedure. Multiple other highly significant, brain related processes were also identified in this analysis and are detailed in the supplementary data (Figure 4.6a and Supplementary Data 4.1).

Figure 4.6 Enrichment analysis of the PRC2 gene set implicates this network in nervous system development.



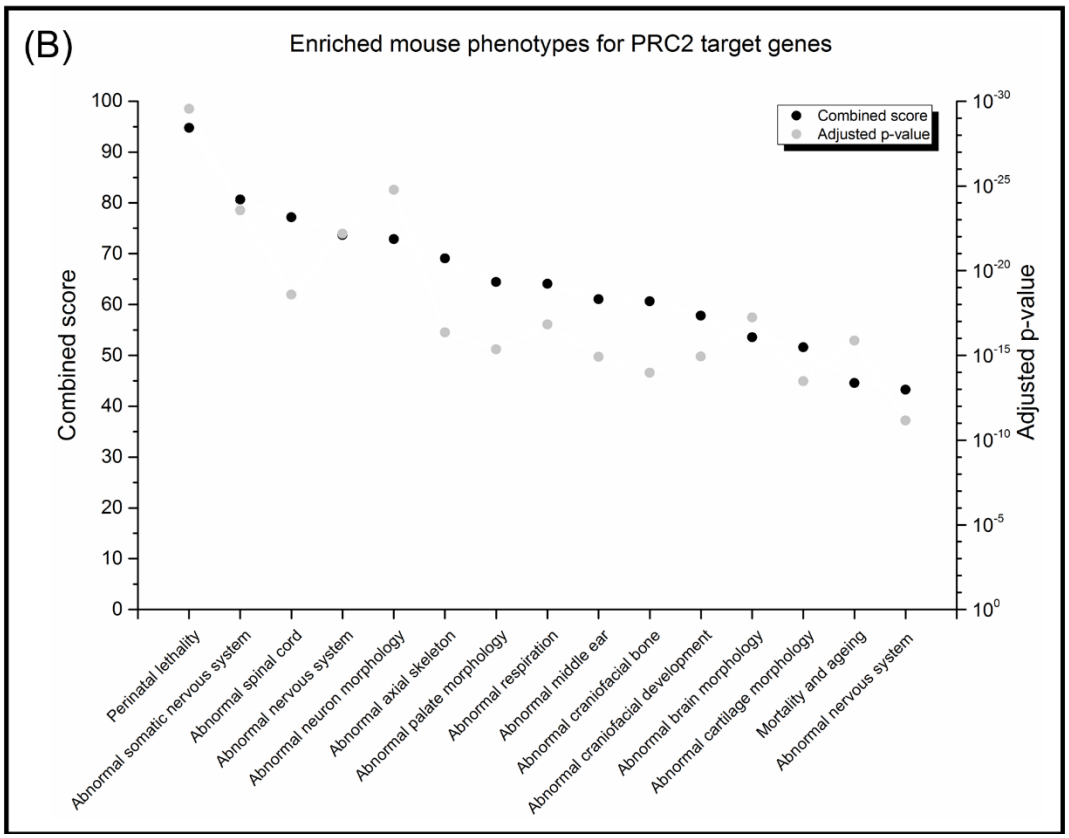


Figure 4.6 Enrichment analysis of the PRC2 gene set implicates this network in nervous system development.

(a) Enrichment analysis using the 900 genes with both EZH2 and SUZ12 binding at their promoters suggested numerous roles for EZH2 in the PRC2 complex in development, as has been widely reported in the literature. Specifically, genes targeted by EZH2 within the PRC2 were enriched for roles in patterning and regionalisation processes in the embryo, as well as general embryonic and organ morphogenesis. This gene set was also enriched for roles in neuron differentiation (Benjamini-Hochberg adjusted Fisher's exact p = 1.32 x10^{-27}), regulation of neuron differentiation (Benjamini-Hochberg adjusted Fisher's exact p = 6.80 x10^{-17}), and neuron fate specification (Benjamini-Hochberg adjusted Fisher's exact p = 1.82 x10^{-16}), in addition to many other highly statistically significant developmental and brain related processes (Supplementary Data 4.1). This would suggest a highly significant role for the PRC2 targeted gene set in regulating many aspects of embryonic development, including neuron differentiation in the developing brain.

(b) Mouse phenotypes associated with disruption or deletion of the PRC2 targeted gene list also support a role for PRC2 targeted genes in healthy CNS development, with highly enriched mouse phenotypes including abnormal somatic nervous system (Benjamini-Hochberg adjusted Fisher's exact p = 2.72 x10^{-24}), abnormal spinal cord (Benjamini-Hochberg adjusted Fisher's exact p = 2.56 x10^{-19}), abnormal nervous system (Benjamini-Hochberg adjusted Fisher's exact p 6.62 x10^{-23} and 7.06 x10^{-12}), abnormal neuron morphology (Benjamini-Hochberg adjusted Fisher's exact p = 1.68 x10^{-25}), and abnormal brain morphology (Benjamini-Hochberg adjusted Fisher's exact p = 5.79 x10^{-18}). We also see enrichment for phenotypes involving abnormal development or morphology of a wide range of organs and bodily structures (Supplementary Data 4.1), indicating a more general role for the PRC2 targeted gene set in ensuring healthy development.

Four of the top five most highly enriched mouse phenotypes for the PRC2 gene set in the MGI mouse phenotype database were CNS-related, with animals displaying:

- Abnormal somatic nervous system — adjusted $p = 2.72 \times 10^{-24}$
- Abnormal spinal cord — adjusted $p = 2.56 \times 10^{-19}$
- Abnormal nervous system — adjusted $p = 6.62 \times 10^{-23}$
- Abnormal neuron morphology — adjusted $p = 1.68 \times 10^{-25}$

All p-values above are derived from the Fisher's exact test and adjusted using the Benjamini-Hochberg procedure. The most highly enriched phenotype was perinatal lethality (Benjamini-Hochberg adjusted Fisher's exact $p = 2.86 \times 10^{-30}$), which, along with other highly enriched phenotypes for abnormal development, demonstrates a critical role for the PRC2 complex in mediating healthy development of the organism as a whole (Figure 4.6b and Supplementary Data 4.1).

Further to this, enrichment analysis using the REST target gene set predominantly supported a role for this pathway in:

- ncRNA metabolism — adjusted $p = 7.85 \times 10^{-8}$
- ncRNA processing — adjusted $p = 2.94 \times 10^{-5}$
- Gene expression — adjusted $p = 1.94 \times 10^{-6}$
- Nucleosome assembly — adjusted $p = 4.77 \times 10^{-7}$
- Nucleosome organisation — adjusted $p = 1.86 \times 10^{-6}$

All p-values above are derived from the Fisher's exact test and adjusted using the Benjamini-Hochberg procedure. This would support the literature in suggesting a role for REST in regulating ncRNA networks and gene expression mechanisms in the CNS (Qureshi and Mehler 2009, Rossbach 2011, Ernsberger 2012). No mouse phenotypes were significantly enriched for mutation or knockout of the REST target gene set

(Benjamini-Hochberg adjusted Fisher's exact p = 1) (Figure 4.7 and Supplementary Data 4.1). Repeating these analyses for the 91 genes with SUZ12 binding independently of EZH2 yielded no significant results for either enriched biological processes (Benjamini-Hochberg adjusted Fisher's exact p = 0.45) or mouse phenotypes (Benjamini-Hochberg adjusted Fisher's exact p = 1) (Supplementary Data 4.1).

4.3.5 The MIR137-REST-EZH2 network is altered in the schizophrenia DL-PFC

Having demonstrated through enrichment analysis that, in particular, EZH2 and PRC2 were likely to be key regulators of healthy brain development, function, and behaviour, we next set out to characterise the MIR137-REST-EZH2 pathway in clinical samples, using RNA-seq data from the DL-PFC of 155 individuals with a diagnosis of schizophrenia and 196 controls. This data was processed and analysed by groups at the Lieber Institute for Brain Development, and at Eli Lilly and Company, as described in Section 2.2.8. Data for the mature MIR137 transcript was not available in this data set, therefore we used data for the long precursor transcript, MIR137HG, as a proxy for MIR137 expression. However, it should be noted that the levels of precursor miRNA expression may not be an accurate depiction of mature miRNA levels due to the short-lived nature of these structures. The half-life of the MIR137 precursor is unknown, but work by Marzi et al. in a mouse fibroblast cell line model demonstrated short half-lives of between 1.1-1.9 hours for the four miRNA precursors tested (Marzi et al. 2016).

Figure 4.7 Enrichment analysis of the REST gene set implicates this network in gene regulation and ncRNA processing.

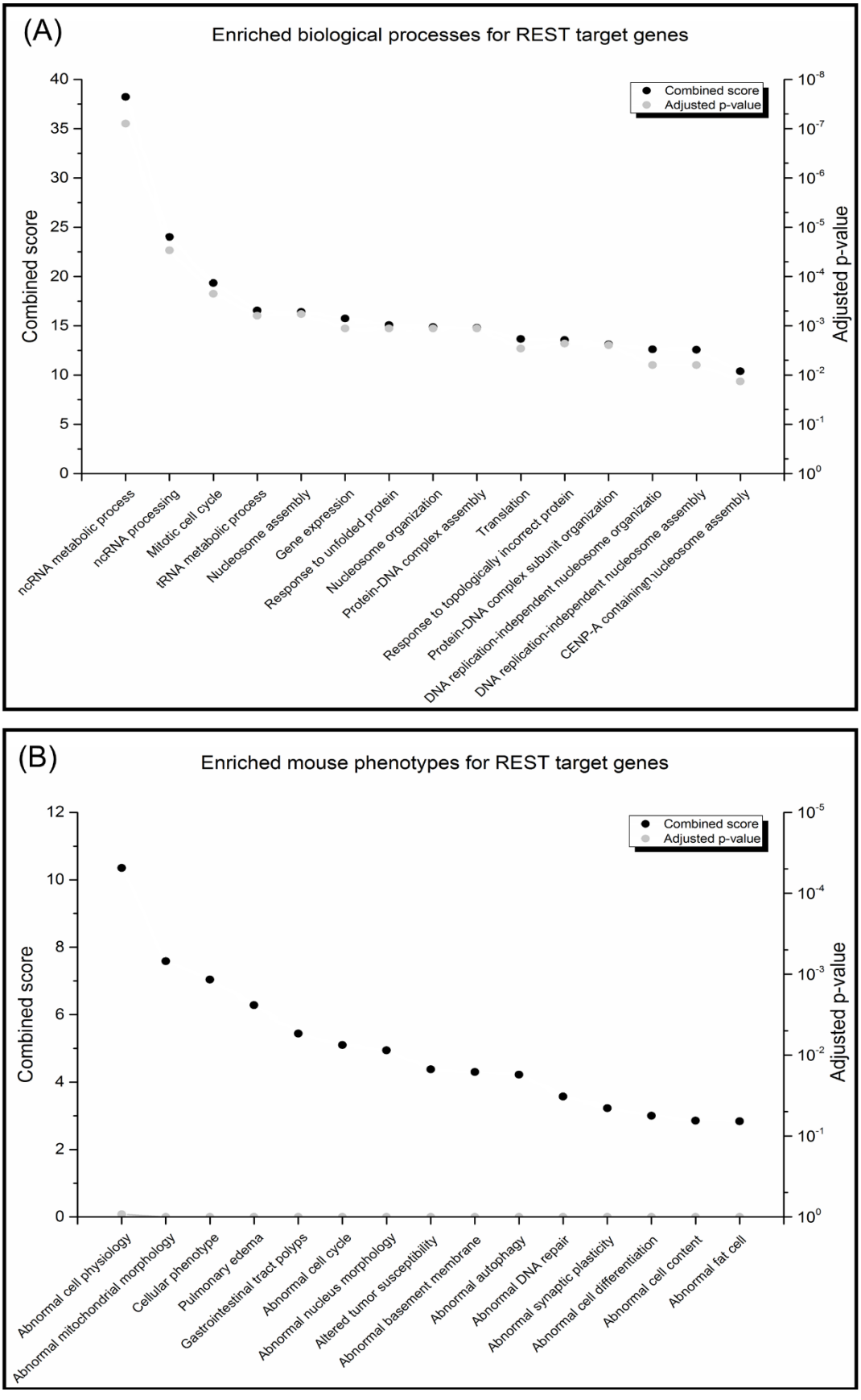


Figure 4.7 Enrichment analysis of the REST gene set implicates this network in gene regulation and ncRNA processing.

a) *Enrichment analysis using the 3691 genes with REST binding at their promoters suggested roles for REST in regulating ncRNA metabolism and processing (Benjamini-Hochberg adjusted Fisher's exact p = 7.85 x10^{-8} and 2.94 x10^{-5}), nucleosome organisation and assembly (Benjamini-Hochberg adjusted Fisher's exact p = 1.86 x10^{-6} and 4.77 x10^{-7}), and gene expression (Benjamini-Hochberg adjusted Fisher's exact p = 1.94 x10^{-6}). This would be in line with the literature describing REST as a transcriptional regulator, as well as the known role of REST in regulating the miRNA, MIR137.*

b) *Mouse phenotypes associated with disruption or deletion of REST target genes did not identify any significantly enriched terms at the Benjamini-Hochberg adjusted Fisher's exact p-value level, however the combined score hinted at potential roles in maintaining normal cell physiology, and morphology of organelles such as the mitochondria and nucleus.*

However, work based on the transcription of almost 100 different miRNAs demonstrated a strong correlation between transcription of primary miRNAs and the levels of mature miRNA, at least for miRNAs that are organised as single transcriptional units (ie. not clustered) such as MIR137 (Marzi et al. 2016). In this study, data for the mature MIR137 miRNA was not available, thus the MIR137HG precursor was used as a proxy. This data should therefore be interpreted with caution, as we cannot confirm that MIR137HG levels are reliably reflective of mature MIR137 levels in the cell. Ideally, this work should be repeated using miRNA specific RNA-seq protocols to replicate and confirm the results presented in this thesis.

RNA-seq analysis revealed differential expression of all members of the proposed MIR137HG-REST-EZH2 regulatory model in the DL-PFC of individuals with schizophrenia when compared to controls. This included deregulation of EZH2's PRC2 partners, SUZ12 and EED. Both REST and EZH2 were found to be deregulated in the DL-PFC of individuals with schizophrenia compared to controls ($F = 132.82$ and 21.00, respectively; $q < 0.001$), which correlated with altered expression of the precursor MIR137HG in the same sample set ($F = 33.77$; $q < 0.001$). Further, we note deregulation of the remaining two PRC2 member genes in the schizophrenia DL-PFC, SUZ12 ($F = 119.68$; $q < 0.001$) and EED ($F = 38.66$; $q < 0.001$) (Table 4.1). This would support our model in which deregulation of MIR137HG, REST, and EZH2 alters the expression not only of each other, but is also likely to result in altered expression of their downstream targets, many of which we have demonstrated through enrichment analysis to be involved in brain-related processes such as synaptic transmission.

Table 4.1 The MIR137-REST-EZH2 network is deregulated in the schizophrenia DL-PFC.

Gene name	F	q
MIR137HG	33.77	1.34E-71
REST	132.82	7.86E-149
EZH2	21.00	1.63E-50
SUZ12	119.68	2.08E-142
EED	38.66	2.81E-78

Analysis of RNA-seq data from the DL-PFC of individuals with schizophrenia and controls demonstrated that all members of the MIR137HG-REST-EZH2 network, including SUZ12 and EED, were deregulated in the schizophrenia brain. Our model would predict that increased expression of REST and EZH2 would result in downregulation of MIR137, all three of which could have regulatory effects on their downstream CNS-expressed target genes.

The F statistic is a summary of multiple t-test statistics and represents an overall test for significance based on a model including all covariates. The q-value is based on the distribution of p-values and accounts for multiple testing.

4.4 Discussion

Significant attention and research has gone into understanding the functions and transcriptional targets of MIR137, which has furthered our knowledge with regard to how this miRNA could contribute to schizophrenia susceptibility. For example, studies show that MIR137 regulates the expression of a large network of schizophrenia-associated genes with roles in nervous system development and function, synaptogenesis, synaptic plasticity, and neuronal transmission (Olde Loohuis et al. 2017, Strazisar et al. 2015, Siegert et al. 2015). Further, variants at the MIR137 locus, in the form of SNPs or altered copy number of repetitive regulatory elements, have been linked to alterations in cognitive function and white matter integrity (Gonzalez-Giraldo, Gonzalez-Reyes and Forero 2016, Cosgrove et al. 2017, Kuswanto et al. 2015). While such work is of value in understanding the mechanisms associated with MIR137 function, it is equally important to understand the expression and regulation of MIR137 in the brain, as even small changes in the levels of this miRNA will likely have far-reaching transcriptional and functional effects. To our knowledge, there is only a single published study assessing MIR137 levels in the brains of individuals with schizophrenia, which demonstrated no significant difference compared to individuals with bipolar disorder or controls (Guella et al. 2013), though this may have lacked power due to the lower sample size. In this chapter, we set out to build a gene network model centred around the regulation of MIR137 expression, and then to test the validity of this model in clinical samples, using RNA-seq data from the DL-PFC of 155 individuals with a diagnosis of schizophrenia and 196 controls.

Others in the lab have previously demonstrated that the transcription factor REST binds at the MIR137 imir promoter and represses transcription from this region (Warburton et al. 2015b, Warburton et al. 2016), while other groups have

demonstrated regulation of MIR137 by the androgen receptor (Nilsson et al. 2015). Further interrogation of ENCODE ChIP-seq data showed the transcription factor EZH2 also binding at both MIR137 promoters in multiple cell lines, and overlapping a schizophrenia GWAS SNP at the MIR137 imir promoter (Figure 4.1) which was known to direct allele specific expression based on the genotype of this SNP (Warburton et al. 2016).

EZH2 is a known target of MIR137 (Szulwach et al. 2010, Sun et al. 2015a, Ren et al. 2015b), and data from ENCODE ChIP-seq would therefore suggest a regulatory feedback loop in which MIR137 and EZH2 regulate each other. REST is also involved in this pathway through its regulation of MIR137, and further for its known ability to recruit EZH2 within the PRC2 to RE-1 sites through the lncRNA, HOTAIR (Dietrich et al. 2012, Tsai et al. 2010). Indeed, the interaction between REST and EZH2 in schizophrenia has been suggested previously by Tamminga et al. Using the mouse hippocampus as a model, Tamminga et al. demonstrated that Ezh2/Prc2 and Rest were enriched at the promoter of grin2b, an ionotropic glutamate receptor subunit gene that has been implicated in schizophrenia (Guo et al. 2016, Zhang et al. 2015, Yang et al. 2015), at the P15 stage in the mouse hippocampus. This group further demonstrated that maternal deprivation in early life resulted in altered activation of Rest in the mouse hippocampus, leading to suppression of Rest-dependent epigenetic remodelling at the promoter of the schizophrenia-associated grin2b gene through co-ordinated action with the Prc. This resulted in deficits in the maturation of glutamatergic synapses in the hippocampus (Tamminga and Zukin 2015). Such work would suggest an environmentally regulated Rest/Ezh2-mediated mechanism with relevance to schizophrenia.

We therefore proposed a three-gene regulatory model based on the interactions between MIR137, REST, and EZH2, outlined in Figure 4.4, which we believe could be modulated in response to an individual's environment to influence risk of psychiatric conditions, particularly in the case of schizophrenia. ENCODE ChIP-seq data supported and extended work by Tamminga et al. by further demonstrating EZH2 binding at the promoters of many gene families that are associated with schizophrenia, such as calcium channels, glutamate receptors, and dopamine receptors (Figure 4.3). This led us to extend the model and to propose a mechanism through which MIR137, REST, and EZH2 regulated not only the expression of each other, but also of a larger group of downstream schizophrenia-associated genes (Figure 4.4).

In order to further characterise the gene sets and functions that may be regulated by EZH2 and REST, we used an unbiased, genome-wide bioinformatic approach, to identify gene sets that displayed EZH2, PRC2, or REST binding at one or more of their transcriptional start sites. Performing enrichment analyses on these target gene lists demonstrated a role for EZH2 acting alone to regulate processes such as behaviour and synaptic transmission (Figure 4.5a), which is supported by mouse phenotype data suggesting that knockout or mutation of genes in this set results in abnormal synaptic transmission, as well as altered learning, memory, and affective behaviour in mouse models (Figure 4.5b). On the other hand, the gene set regulated by EZH2 acting alongside SUZ12 as part of the PRC2 showed strong enrichment for roles in embryonic development (Figure 4.6a), and specifically in processes involved in nervous system development (Figure 4.6b). This would support the existing literature outlining the function of PRC2 in embryonic and foetal development (Coskun, Tsoa and Sun 2012, Aloia, Di Stefano and Di Croce 2013, Blackledge, Rose and Klose

2015, Piunti and Shilatifard 2016), and specifically in brain development (Qi et al. 2013, Ronan et al. 2013, Hirabayashi et al. 2009).

Taken together, these data suggest a role for EZH2, both independently and in the PRC2, in the healthy development and function of the nervous system, with dysregulation of EZH2/PRC2 and their target genes potentially contributing to schizophrenia risk. This could occur at two distinct timepoints. It is clear that PRC2 plays important roles in development, and particularly nervous system development (Figure 4.6). Therefore, altered regulation of EZH2/PRC2 during development may contribute to the neurodevelopmental risk for schizophrenia. Secondly, the EZH2 target gene set was shown to be involved in processes such as behaviour and synaptic transmission (Figure 4.5), and expression of EZH2 was further demonstrated to be maintained at low levels across the brain throughout the lifetime after birth (Figure 4.2). This may suggest that anomalous reactivation of EZH2 expression in the adult brain could result in changes in gene expression which would alter not only MIR137 and its transcriptional targets, but also EZH2's own distinct gene set, which would likely impact pathways involved in synaptic transmission and behaviour.

This hypothesis can be tested through measuring EZH2 expression in RNA-seq samples from the brains of individuals with schizophrenia compared to controls. The results presented in this thesis demonstrate altered expression of EZH2 in the schizophrenia DL-PFC, however, due to the analysis methods used by the group at Eli Lilly, a direction of change was not able to be determined. To extend this work, RNA-seq data from the PD_NGSAtlas (http://bioinfo.hrbmu.edu.cn/pd_ngsatlas/), a reference database containing RNA-seq data from a small number of individuals with schizophrenia (6) and controls (4) was utilised (Zhao et al. 2014). Data from this resource demonstrated a 3-fold increase in EZH2 expression in the anterior cingulate

cortex of individuals with schizophrenia compared to controls, which correlated with a downregulation in the expression of a schizophrenia-associated GWAS gene, CACNA1C. Follow up work by Kimberley Billingsley and Dr. Maurizio Manca demonstrated that EZH2 over-expression resulted in repression of reporter gene expression from the schizophrenia-associated CACNA1C promoter, suggesting that increased levels of EZH2 in the schizophrenia brain could have downstream effects on the regulation of other disease-associated genes (Billingsley et al. 2018).

Further, enrichment analysis using the REST target gene set demonstrated a role for this network in the regulation of ncRNA processing and metabolism (Figure 4.7a), in line with literature reporting REST as a regulator of ncRNA networks in the CNS (Qureshi and Mehler 2009, Rossbach 2011, Ernsberger 2012).

The MIR137-REST-EZH2 model was assessed using RNA-seq data from the DL-PFC of 155 individuals with a diagnosis of schizophrenia and 196 controls. This data was analysed by Dr. Karim Malki, Dr. Nathan Lawless, and Dr. Andrew Jaffe at Eli Lilly and the Lieber Institute for Brain Development. This demonstrated deregulation of all members of the MIR137-REST-EZH2 network, including members of the PRC2 complex, SUZ12 and EED (Table 4.1). No data was available for mature miRNAs in this data set, thus we used data on the precursor transcript, MIR137HG, as a proxy for mature MIR137 levels. To our knowledge, this is the first study to confirm deregulation of MIR137HG in the brains of individuals with schizophrenia, despite extensive research into this miRNA in relation to the condition. In this data set, MIR137HG was significantly deregulated in the DL-PFC of individuals with a diagnosis of schizophrenia when compared to controls ($F = 33.77$; $q < 0.001$), along with REST and EZH2 ($F =$ and 132.82 and 21.00, respectively; $q < 0.001$). While the statistical analysis performed on this data does not provide a direction of change for differential

expression, we would hypothesise from our model that REST and EZH2 may be increased in the schizophrenia DL-PFC, which could then result in downregulation of MIR137HG.

Unfortunately, we were unable to obtain access to the raw RNA-seq data, or to information on the Lilly analysis pipeline, both of which were restricted to commercial partners at Eli Lilly and Company. We were therefore limited only to the use of data that had already been analysed within Lilly. One key question that remains, and would have been particularly of interest, is the question of transcript ratios for the genes discussed in this chapter. The key genes of interest, MIR137, REST, and EZH2, are all known to exist in multiple splice variants. In particular, previous work has demonstrated a REST transcript (known as short form NRSF or sNRSF in humans, and REST4 in rodents) which results in a truncated protein (Coulson et al. 2000, Coulson et al. 2003). Upregulated prefrontal cortex expression of REST4 in response to early life stress has been demonstrated in rat models, and over-expression in this brain region was shown to alter vulnerability to later life stress (Uchida et al. 2010). Altered expression of REST4/sNRSF has also been studied in the context of epilepsy, as well as cancers including glioma and small cell lung cancer (Spencer et al. 2006, Ren et al. 2015a, Coulson et al. 2000, Coulson et al. 2003). This transcript in particular would therefore have been of interest to assess potential differential expression in individuals with schizophrenia. However, the RNA-seq data at our use was a “roll up” data set which combined expression data of multiple transcripts into an overall level of gene expression, and thus no data was available for specific transcripts. Such as with the REST4/sNRSF transcript, it may be likely that only specific transcripts are deregulated in a disease state, which in this case would be obscured by reporting only an overall expression level. Similarly, where one transcript may be downregulated and

another upregulated abnormally in a disease state, combining expression data across genes could risk erroneously reporting no change in expression. Further work to assess potential differential transcript expression for the genes of interest in this chapter would be of great interest. Furthermore, if access to the raw RNA-seq data was available, analysing the data in such a way as to show the direction of change in expression would provide a clearer understanding of the processes which may be altered in the schizophrenia cohort.

In addition to regulating MIR137, and thereby indirectly modulating the downstream schizophrenia-associated MIR137 target gene set, we demonstrated that EZH2 and REST are also likely to each affect their own distinct CNS target gene sets, as characterised by genome-wide transcription factor binding and enrichment analyses (Figures 4.4, 4.5, and 4.6). This would add another layer of complexity to the model, as changes in REST and EZH2 expression in the brains of individuals with schizophrenia may result not only in the altered regulation of MIR137 and its target genes, but also in altered regulation of wider CNS- and schizophrenia-associated pathways directed by REST and EZH2.

4.5 Summary

In conclusion, the data contained in this chapter demonstrates that the miRNA precursor, MIR137HG, is significantly deregulated in the DL-PFC of individuals with schizophrenia compared to controls. Combining data from our model and from RNA-seq data would suggest that this deregulation of MIR137HG is likely due in part to the altered expression of REST and EZH2, which are validated and predicted regulators of MIR137, respectively. As well as binding at and likely regulating expression from the MIR137 promoters, both REST and EZH2 were also found to target their own distinct brain- and schizophrenia-associated gene networks through interrogation of ENCODE ChIP-seq data and enrichment analysis. The MIR137-REST-EZH2 model demonstrated that changes in the level of any one of the three genes may result in altered regulation of both the MIR137 schizophrenia-associated gene set as well as deregulation of CNS- and schizophrenia-associated REST or EZH2 directed pathways. As such, this gene network may play a key role in regulating underlying mechanisms in schizophrenia biology.

Chapter 5

The DNAJ gene family and MIR941 in schizophrenia

5.1 Introduction

In the previous two chapters, we have considered the schizophrenia-associated miRNA, MIR137, attempting to elucidate regulatory mechanisms and pathways through identifying and testing conserved regulatory elements, and through observing transcription factor binding and regulatory gene networks. In this chapter, focus turns to a second miRNA, MIR941. MIR941, is a human-specific, brain-expressed transcript that is located within the DNAJC5 gene at the chromosome 20q13.3 locus (Hu et al. 2012). It has previously been demonstrated that this miRNA resides within a variable number tandem repeat (VNTR) which is polymorphic in copy number across the population (Hu et al. 2012). Further, MIR941 has been shown to be lost from the miRNA interaction networks in the blood of individuals with psychosis, and is altered in response to drug treatment for major depressive disorder in mouse models (Jeffries et al. 2016, Belzeaux et al. 2012).

The MIR941 region was selected for study as it contained multiple characteristics that were of interest and that had been previously studied at other schizophrenia or CNS gene loci. This included primate-specific sequence, a variable number tandem repeat (VNTR), a human-specific, brain-expressed microRNA, and the location of this microRNA within a locus that had previously been associated Kufs disease, a neurodegenerative condition which can include psychiatric symptoms such as psychosis (Wisneiwski et al. 1992, Lewandowska et al. 2009, Sandyk 1981, Tobo et al. 1984). Previous work has demonstrated a VNTR at the MIR137 schizophrenia-associated locus as supporting allele-specific expression and altering RNA structure of transcripts including the repetitive region, as well as correlating with altered performance on the Stroop test in healthy individuals (Warburton et al. 2015a, Warburton et al. 2015b, Mamdani et al. 2013, Gonzalez-Giraldo et al. 2016).

We were therefore interested in the potential role for the MIR941 VNTR and wider MIR941/DNAJC5 locus in the CNS and in CNS conditions, taking a similar approach to that used by Warburton et al. to study this region and to build on further work around VNTRs at brain-expressed microRNA loci (Warburton et al. 2015a, Warburton et al. 2015b). However, such candidate gene approaches have been criticised as being subjective, as well as potentially limiting the view of true causal risk loci, and potentially leading to results that are not reproducible. It has also been argued that the prior knowledge on which candidate gene selection is based is incomplete, and thus unbiased approaches such as GWAS would be preferable over hypothesis-driven gene selection based on incomplete information (Tabor, Risch and Myers 2002, Zhu and Zhao 2007).

Schizophrenia was selected for the condition of interest in which to study this region because of the loci's known role in the CNS, as well as the availability of schizophrenia samples in the lab. This work is being further extended by Ana Illera-Lopez to study MIR941/DNAJC5 in cognitive ageing and in neurodegenerative conditions, which may be more appropriate functionally due to the known role of DNAJC5 in the neurodegenerative condition, Kuf's disease.

VNTRs are known to act as transcriptional regulators (Warburton et al. 2015b, Paredes et al. 2013, Vasiliou et al. 2012), and have been linked to a range of behaviours and psychiatric conditions, including aggression, substance and alcohol abuse, depression, and schizophrenia, often acting in a 'gene x environment' manner (Peitl, Stefanovic and Karlovic 2017, Zhang et al. 2017, Ma, Fan and Li 2016, Bilic et al. 2014, Mallard et al. 2016). For example, Warburton et al. and Mamdani et al. have previously characterised a polymorphic VNTR at the schizophrenia-associated MIR137 locus, which modulates gene expression in an allele dependent manner

(Warburton et al. 2015b), and alters RNA folding when this region is included in the precursor transcript (Mamdani et al. 2013). It is therefore clear that polymorphic repeats that alter the regulation or structure of genes involved in brain-related pathways can contribute to an individual's risk of psychiatric illness. The fact that altered copy number of the MIR941 VNTR results in altered copy number of this brain-expressed miRNA provides an additional level of interest in studying this region in CNS and psychiatric conditions.

MIR941's host gene, DNAJC5 (also known as cysteine string protein; CSP), is a pre-synaptic vesicle protein with known roles in neuroprotection, neurotransmitter release, and synapse maintenance. Mutations in this gene are known to result in adult-onset neuronal ceroid lipofuscinosis (NCL), a neurodegenerative condition resulting from the inappropriate accumulation of lipopigments within multiple organs including the brain (Burgoyne and Morgan 2015, Cadieux-Dion et al. 2013, Benitez et al. 2011, Lopez-Ortega, Ruiz and Tabares 2017). The broader DNAJ family of heat shock proteins (HSP40) are also known to be expressed in the brain, with wide-ranging roles in neuroprotection and involvement in CNS conditions such as amyotrophic lateral sclerosis (ALS), Alzheimer's disease, Parkinson's disease, and schizophrenia (Chen et al. 2016a, Tiwari et al. 2015, Yuan et al. 2016, Vilarino-Guell et al. 2014, Edvardson et al. 2012, Liu et al. 2011).

The DNAJC5/MIR941 region at 20q13.3 is therefore of interest in understanding healthy CNS functioning and this region's potential role in psychiatric conditions. The polymorphic MIR941 VNTR results in altered copy number of the miRNA in the population, but could also act as a regulatory domain for DNAJC5, with the additional possibility of altering the RNA or protein structure in DNAJC5 transcripts which utilise this region as an exon. In this chapter, we use bioinformatic analysis to further

characterise transcripts at this locus, and genotype the VNTR in both schizophrenia and control cohorts. We also make further use of RNA-seq data from the DL-PFC of individuals with schizophrenia and controls to analyse the expression of DNAJC5 and the wider DNAJ gene family in the brain and in schizophrenia.

5.2 Aims

- Characterise transcripts at the DNAJC5 locus using the UCSC Genome Browser.
- Identify transcriptional regulatory activity around transcriptional start sites at this locus in *ex vivo* samples using publicly available histone modification data to build up a model of activity at this locus in humans.
- Identify pathways and phenotypes that may be altered by changes in expression at the DNAJC5 locus.
- Genotype the MIR941 VNTR in control and schizophrenia cohorts to identify any potential schizophrenia risk variants.
- Analyse the expression of DNAJC5 and the wider DNAJ gene family in RNA-seq data from the DL-PFC of individuals with schizophrenia and controls to identify potential roles for these genes in underlying schizophrenia biology.

5.3 Results

5.3.1 Bioinformatic analysis of the DNAJC5/MIR941 locus

Bioinformatic analysis of the DNAJC5/MIR941 locus was carried out using the UCSC Genome Browser (GRCh37/hg19). Visualisation of the Chr20q13.3 locus showed two transcripts of DNAJC5, with a main transcript covering 40.93 kb and a second, shorter transcript of 16.58 kb, referred to as AK128776 in GenBank. The two transcripts differ predominantly in their first exon, while sharing exons 2, 3, 4, and their final exon. The first and final exons of both transcripts were annotated as non-coding 5' and 3' UTRs on the UCSC genome browser, leaving DNAJC5 and AK128776 with similar protein-coding sequences except for an additional ~75 bp coding exon between the fourth exon and the 3' UTR in the AK128776 transcript. Exons 2, 3, and 4 in both transcripts were found to be very highly conserved according to data from the Multiz alignment of vertebrate species (Figure 5.1a), which shows conservation of these regions back to zebrafish. Exon 5 of AK128776 shares some conservation back to chicken, while the shared 3' UTR of both transcripts shows some conservation back to the mouse genome. Where the 5' UTR of the major DNAJC5 transcript is also predominantly conserved back to the mouse (Figure 5.1a), the 5' UTR of AK128776 overlaps a primate-specific CpG island and simple repeat, which contains the miRNA, MIR941 (Figure 5.1b). The bulk of AK128776's 5' UTR is comprised of this tandem repeat sequence, making up 710 bp of the approximately 880 bp exon and extending upstream of the transcriptional start site.

Further interrogation revealed that AK128776 was a 4037 bp coding mRNA resulting in a 167 aa protein. Data from GenBank confirmed that AK128776 had been identified in the human hippocampus.

Figure 5.1 Visualisation of the DNAJC5/MIR941 locus using the UCSC Genome Browser.

Figure 5.1 Visualisation of the DNAJC5/MIR941 locus using the UCSC Genome Browser.

(a) Visualisation of the complete DNAJC5/MIR941 locus (chr20:62,524,092-62,569,394) demonstrated two transcripts of DNAJC5 at this region which primarily differed in their 5' UTR, as well as an additional exon present in the shorter AK128776 transcript. Data on the Multiz alignment of multiple vertebrates demonstrates that the shared exons 2, 3, and 4 are highly conserved back to the zebrafish, with significant conservation of the shared 3' UTR back to the mouse genome. The 5' UTR of the shorter DNAJC5 transcript, AK128776, overlaps a CpG island and the MIR941 miRNA. This region is expanded in (b). SNPs noted in this figure (yellow lines) will be used in bioinformatic analysis in Figure 5.2.

(b) Visualisation of the region encompassing MIR941 and the 5' UTR of AK128776 (chr20:62,546,952-62,554,433) demonstrated that the area containing the miRNA and making up much of the 5' UTR was a high GC simple tandem repeat, resulting in a CpG island at this region. Using the Multiz alignment, the tandem repeat region was found to be primate-specific, with evidence of this sequence in humans, chimpanzees, and the Rhesus macaque. The 'RepeatMasker' data set on the UCSC Genome Browser also identified a primate-specific short interspersed nuclear element (SINE) and long terminal repeat (LTR) retrotransposon insertion of the AluS and ERVK subfamilies, respectively. These elements lie approximately 1.9 kb upstream of the AK128776/MIR941 transcriptional start site and could potentially act as regulatory domains from this location.

In order to gain insight into the transcriptional activity around the different transcripts at the DNAJC5 locus in human tissues, the HaploReg v4.1 tool was utilised to view chromatin state and histone modification data around the transcriptional start sites of the major DNAJC5 transcript and the AK128776 and MIR941 transcripts.

While the MIR941 VNTR can alter the copy number of the miRNA based on genotype, previous work has shown that VNTRs can act as transcriptional regulators with allele specific effects (Warburton et al. 2015b, Paredes et al. 2013, Klenova et al. 2004). In order to assess the regulatory potential around the transcriptional start sites of the major DNAJC5 transcript and the AK128776/MIR941 transcripts, we made use of the HaploReg v4.1 tool to access chromatin state and histone modification data across 130 different human tissues and cell lines.

Using rs115636427 as a proxy SNP for the full length DNAJC5 promoter revealed chromatin state and histone modification data suggestive of active promoter status at this region in all 130 human tissues and cell lines tested (Supplementary Data 5.1). This included chromatin state and H3K4me3 histone modifications suggestive of promoter activity in all 15 brain regions and cell lines, from neuronal progenitor cells and the foetal brain, to multiple regions of the adult brain (Figure 5.2a). Conversely, using rs2427555 and rs4809249 as proxy SNPs for the AK128776/MIR941 transcriptional start site demonstrated evidence of active promoter status in H1-derived neuronal progenitor cells and the hippocampus, as well as highlighting H3K27ac marks demonstrating transcriptional enhancer activity in the hippocampus, substantia nigra, anterior caudate, cingulate gyrus, interior temporal lobe, and angular gyrus (Figure 5.2b).

Figure 5.2 Chromatin state and histone modifications at the DNAJC5 and AK128776/MIR941 promoters.

(A) DNAJC5 full length promoter (rs115636427)

Description	Chromatin states (15-state model)	Chromatin states (25-state model)	H3K4me1	H3K4me3	H3K27ac	H3K9ac
H1 Derived Neuronal Progenitor Cultured Cells	1_TssA	1_TssA		H3K4me3_Pro		H3K9ac_Pro
H9 Derived Neuronal Progenitor Cultured Cells	1_TssA	1_TssA	H3K4me1_Enh	H3K4me3_Pro		
H9 Derived Neuron Cultured Cells	1_TssA	1_TssA		H3K4me3_Pro		
Ganglion Eminence derived primary cultured	1_TssA	2_PromU		H3K4me3_Pro		
Cortex derived primary cultured neurospheres	2_TssAFlnk	1_TssA	H3K4me1_Enh	H3K4me3_Pro		
Brain Hippocampus Middle	1_TssA	1_TssA	H3K4me1_Enh	H3K4me3_Pro	H3K27ac_Enh	
Brain Substantia Nigra	1_TssA	1_TssA		H3K4me3_Pro	H3K27ac_Enh	H3K9ac_Pro
Brain Anterior Caudate	1_TssA	1_TssA		H3K4me3_Pro	H3K27ac_Enh	H3K9ac_Pro
Brain Cingulate Gyrus	1_TssA	1_TssA		H3K4me3_Pro	H3K27ac_Enh	
Brain Inferior Temporal Lobe	1_TssA	1_TssA		H3K4me3_Pro	H3K27ac_Enh	H3K9ac_Pro
Brain Angular Gyrus	1_TssA	1_TssA		H3K4me3_Pro	H3K27ac_Enh	H3K9ac_Pro
Brain_Dorsolateral_Prefrontal_Cortex	1_TssA	1_TssA		H3K4me3_Pro	H3K27ac_Enh	
Brain Germinal Matrix	1_TssA	1_TssA		H3K4me3_Pro		
Fetal Brain Female	1_TssA	1_TssA		H3K4me3_Pro		
Fetal Brain Male		1_TssA	H3K4me1_Enh			

(B) AK128776/MIR941 promoter (rs2427555 and rs4809249)

Description	Chromatin states (15-state model)	Chromatin states (25-state model)	H3K4me1	H3K4me3	H3K27ac	H3K9ac
H1 Derived Neuronal Progenitor Cultured Cells						H3K9ac_Pro
Brain Hippocampus Middle		12_TxEnhW	H3K4me1_Enh		H3K27ac_Enh	
Brain Substantia Nigra					H3K27ac_Enh	
Brain Anterior Caudate			H3K4me1_Enh		H3K27ac_Enh	
Brain Cingulate Gyrus			H3K4me1_Enh		H3K27ac_Enh	
Brain Inferior Temporal Lobe					H3K27ac_Enh	
Brain Angular Gyrus			H3K4me1_Enh		H3K27ac_Enh	
Brain_Dorsolateral_Prefrontal_Cortex					H3K27ac_Enh	

Chromatin states: Enh = enhancer, PromU = promoter upstream transcriptional start site, TssA = active transcription start site, TssAFlank = flanking active transcriptional start site, TXEnhW = weak transcriptional enhancer.

Histone modifications: Enh = enhancer, Pro = promoter, Black = no available data.

Yellow = Enhancer (Enh), weak enhancer (EnhW1, EnhW2). Orange = Active enhancer (Enh), flanking active enhancer (EnhAF). Red = Active transcriptional start site (TssA), flanking active transcriptional start site (TssAFlnk), promoter (Pro), promoter upstream transcriptional start site (PromU). Pink = Poised promoter (PromP).

Figure 5.2 Chromatin state and histone modifications at the DNAJC5 and AK128776/MIR941 promoters.

(a) Chromatin state and histone modification data from HaploReg using the full length DNAJC5 promoter SNP, rs115636427, demonstrated clear promoter activity in all tissues and cell lines tested (Supplementary Data 5.1). This included 15 neuronal cell lines or brain regions, ranging from stem cell derived neural progenitors and the foetal brain to multiple regions of the adult brain, such as the DL-PFC and the hippocampus. Such data would suggest ubiquitous expression of DNAJC5 across numerous organs and tissues.

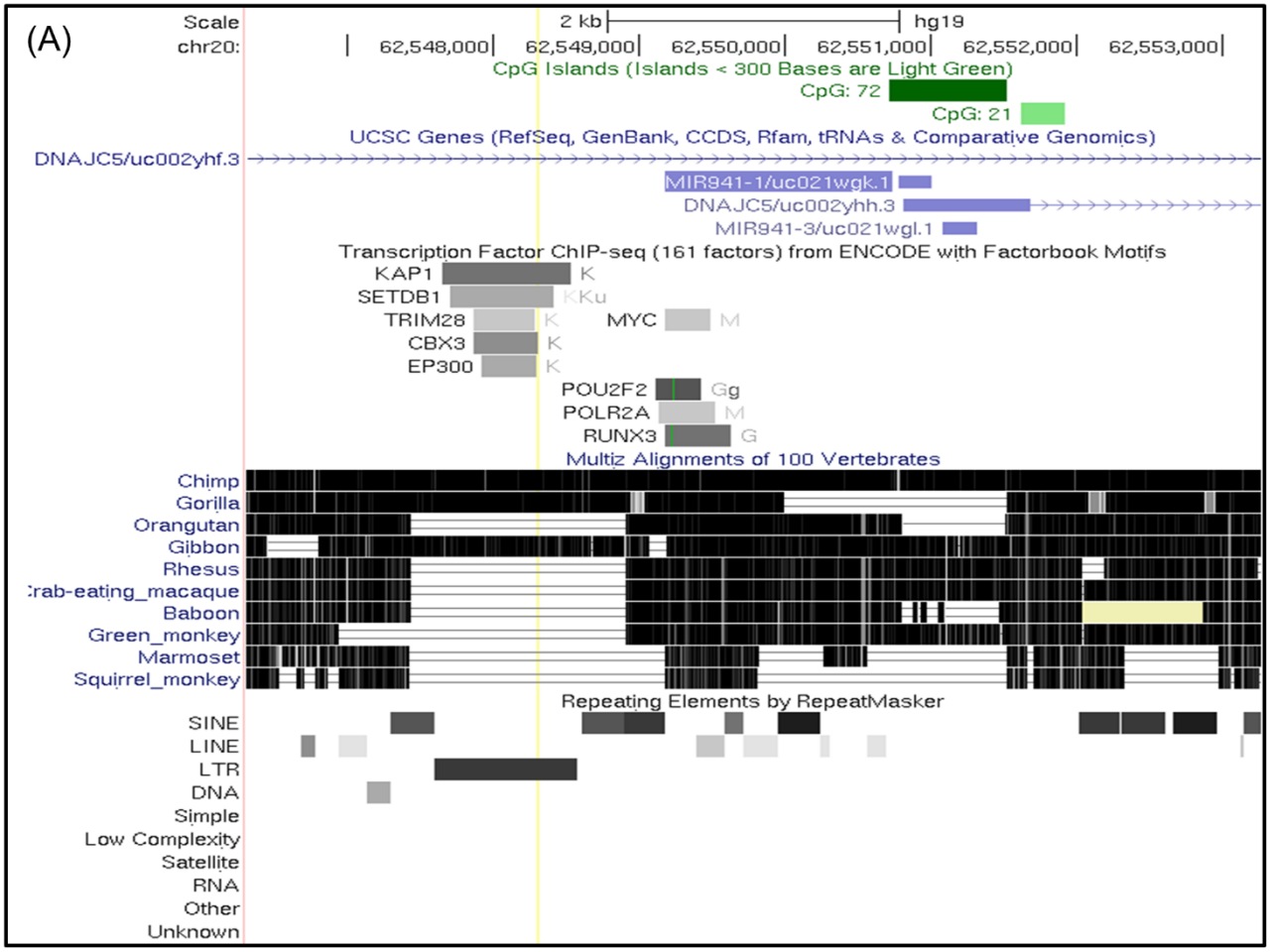
(b) HaploReg data across the AK128776/MIR941 promoter region based on the SNPs, rs2427555 and rs4809249, instead demonstrated limited regulatory activity which was predominantly found in the adult brain. H3K4me1 and H3K27ac histone modifications suggest transcriptional regulatory function in the hippocampus, anterior caudate, cingulate gyrus, and angular gyrus, with H3K27ac histone marks indicating potential regulatory activity in multiple other brain regions. Further, H3K9ac histone modifications were suggestive of promoter activity at this region in H1-derived neuronal progenitor cells, and chromatin state data suggests transcription from this region in the hippocampus, which would support GenBank data on the AK128776 transcript. Such data would suggest that the expression from the AK128776/MIR941 transcriptional start site is likely to be much more restricted than from the full length DNAJC5 promoter, which may be expected due to the recent evolutionary origin of these transcripts.

In addition to transcriptional regulatory activity in the brain, this data set suggests potential activity around the AK128776/MIR941 promoter in other cells including primary B and T cells, and primary natural killer cells, H1-derived mesenchymal stem cells, and others (Supplementary Data 5.1). This may suggest that AK128776 and MIR941 are more selectively expressed, in comparison to the ubiquitous full length DNAJC5 transcript, with chromatin state data suggesting transcription around AK128776/MIR941 in the hippocampus, thereby supporting the GenBank data identifying this mRNA in the same brain region.

Further, interrogation of UCSC data identified a chimpanzee and human-specific ERVK and Alu retrotransposon insertion approximately 1.9 kb upstream of the AK128776/MIR941 transcriptional start site (Figure 5.3a). Multiple transcription factors bind across this region according to ENCODE ChIP-seq data available through the UCSC Genome Browser. In particular, TRIM28 (also known as KAP1) is known to bind and repress endogenous retroviral elements (ERVs). Recent studies have demonstrated that the TRIM28-mediated repression of ERVs can spread to nearby genes resulting in modulation of gene expression directed by the activation status of nearby retroviral elements. It is known that ERV-derived sequences can act as transcriptional regulators in transgenic embryos, with their regulatory function modulated by their TRIM28 sensitivity (Rowe et al. 2013). Further, it has been demonstrated that predominantly primate-specific ERVs, such as the one identified at the MIR941 locus, regulate transcriptional networks in human neural progenitor cells through TRIM28 binding and establishing local heterochromatin, which then extends to regulate the expression of adjacent genes (Brattas et al. 2017, Fasching et al. 2015).

Figure 5.3 Transcriptional regulatory potential of retrotransposon insertions upstream of the AK128776/MIR941 transcriptional start site.

(A)



(B)

Description	Chromatin states (15-state model)	Chromatin states (25-state model)	H3K4me1	H3K4me3	H3K27ac	H3K9ac
Brain Hippocampus Middle		12_TxEnhW	H3K4me1_Enh		H3K27ac_Enh	
Brain Substantia Nigra	7_Enh		H3K4me1_Enh		H3K27ac_Enh	
Brain Anterior Caudate		12_TxEnhW	H3K4me1_Enh		H3K27ac_Enh	H3K9ac_Pro
Brain Cingulate Gyrus	7_Enh		H3K4me1_Enh		H3K27ac_Enh	H3K9ac_Pro
Brain Inferior Temporal Lobe		12_TxEnhW			H3K27ac_Enh	H3K9ac_Pro
Brain Angular Gyrus			H3K4me1_Enh		H3K27ac_Enh	H3K9ac_Pro
Brain_Dorsolateral_Prefrontal_Cortex			H3K4me1_Enh		H3K27ac_Enh	

Figure 5.3 Transcriptional regulatory potential of retrotransposon insertions upstream of the AK128776/MIR941 transcriptional start site.

(a) ENCODE ChIP-seq data demonstrated transcription factor binding over the primate-specific LTR (ERVK) and SINE (Alu) insertions approximately 1.9 kb upstream of the AK128776/MIR941 transcriptional start site. In particular, the binding of TRIM28 (KAP1) across this region is of interest due to its known role in the repression of ERV elements, which has been shown to alter the expression of adjacent genes through the extension of local heterochromatin. The binding of RNA Polymerase II between the retrotransposons and the transcriptional start site also provided further evidence of expression from this region.

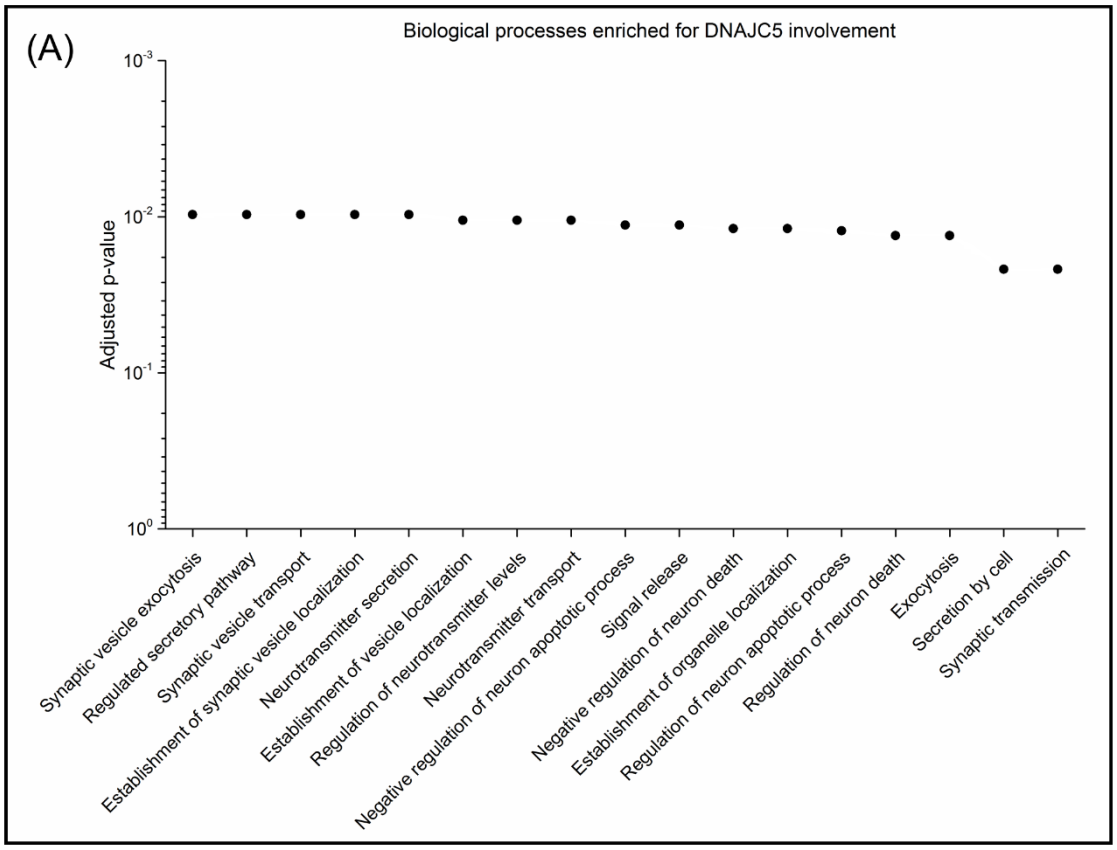
(b) H3K4me1 and H3K27ac histone modification data from HaploReg using the input SNP, rs113385836, suggested transcriptional regulatory activity in multiple regions of the adult brain including the hippocampus, substantia nigra, anterior caudate, cingulate gyrus, inferior temporal lobe, angular gyrus, and DL-PFC. Further, chromatin state data suggested transcription and weak regulatory activity at this region in the hippocampus, anterior caudate, and inferior temporal lobe, with H3K9ac data suggesting promoter activity in the anterior caudate, cingulate gyrus, inferior temporal lobe, and angular gyrus. No relevant data was seen in neuronal progenitors or in the foetal brain, suggesting regulatory activity across the ERVK and Alu elements at this locus may be restricted to the adult brain.

Similarly, multiple groups have demonstrated ability of Alu elements to act as transcriptional regulators (Bouttier et al. 2016, Rajendiran et al. 2016, Payton et al. 2016). Thus, with these elements as known modulators of gene expression, we used the proxy SNP rs113385836, which resides within the ERVK element, to identify potential transcriptional regulatory properties at this region. Data from HaploReg across this SNP demonstrates promoter activity almost exclusively in the brain, with H3K9ac histone modifications indicating promoter status in the anterior caudate, cingulate gyrus, angular gyrus, and inferior temporal lobe, as well as chromatin state data suggesting transcription at this region in the hippocampus, anterior caudate, and inferior temporal lobe (Figure 5.3b; Supplementary Data 5.1).

As the MIR941 VNTR may possess the ability to modulate expression across the DNAJC5 locus, the Enrichr tool was used to gain insight into which processes and phenotypes may be altered by changes in DNAJC5 expression (Figure 5.4). Enrichment analysis for DNAJC5 using the 5192 terms in the 2015 Gene Ontology biological processes data set showed significant enrichment for roles in:

- Synaptic vesicle exocytosis — adjusted p-value = 9.69×10^{-3}
- Synaptic vesicle transport — adjusted p-value = 9.69×10^{-3}
- Synaptic vesicle localisation — adjusted p-value = 9.69×10^{-3}
- Neurotransmitter secretion — adjusted p-value = 9.69×10^{-3}
- Regulation of neurotransmitter levels — adjusted p-value = 1.05×10^{-2}
- Neurotransmitter transport — adjusted p-value = 1.05×10^{-2}
- Regulation of neuron apoptotic processes — adjusted p-value = 1.23×10^{-2}
- Regulation of neuron death — adjusted p-value = 1.32×10^{-2}
- Synaptic transmission — adjusted p-value = 2.17×10^{-2}

Figure 5.4 Enrichment analysis for DNAJC5 demonstrates a role in synaptic vesicles and neurotransmission as well as CNS-related phenotypes.



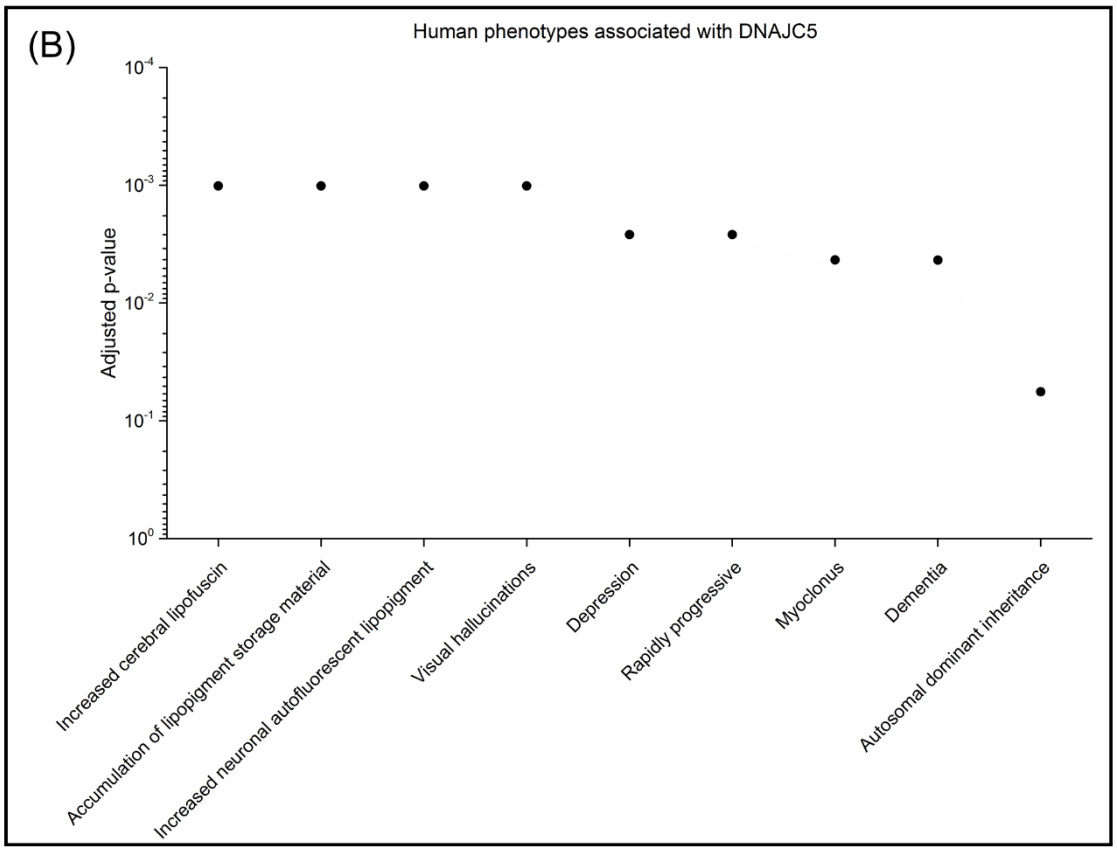


Figure 5.4 Enrichment analysis for DNAJC5 demonstrates a role in synaptic vesicles and neurotransmission as well as CNS-related phenotypes.

(a) *Enrichment analysis for DNAJC5 using the 5192 terms in the Gene Ontology biological processes data set returned 17 significantly associated processes, which were predominantly involved in synaptic vesicle function, neurotransmission, and the regulation of neuron death. This clearly demonstrated a role for DNAJC5 in CNS-related processes.*

(b) *Enrichment analysis for DNAJC5 using the 1779 terms in the Human Phenotype Ontology data set supported the known role of this gene in NCL-related build-up of neuronal lipopigments, as well as demonstrating significant enrichment for visual hallucinations, depression, and dementia. Both dementia and psychosis have been noted as features of adult onset NCL, in addition to impaired voluntary movement and myoclonic epilepsy, which is highlighted in the statistical significance for myoclonus. Finally, enrichment for the term 'rapidly progressive' is likely to refer to the progressive degeneration noted in individuals with NCL.*

All p-values above and below are derived from the Fisher's exact test and adjusted using the Benjamini-Hochberg procedure. In addition, processes involved in secretion, exocytosis, and signal release were also found to be significant at the Benjamini-Hochberg adjusted Fisher's exact p-value level (Figure 5.4a; Supplementary Data 5.2). Similarly, enrichment analysis for the 1779 terms in the Human Phenotype Ontology data set supported the known role for DNAJC5 in phenotypes seen in adult onset NCL, as well identifying additional CNS-related phenotypes, with significance for:

- Increased cerebral lipofuscin — adjusted p-value = 1.01×10^{-3}
- Accumulation of lipopigment storage material — adjusted p-value = 1.01×10^{-3}
- Increased neuronal lipopigment — adjusted p-value = 1.01×10^{-3}
- Visual hallucinations — adjusted p-value = 1.01×10^{-3}
- Depression — adjusted p-value = 2.63×10^{-3}
- Dementia — adjusted p-value = 4.33×10^{-3}

While psychosis and dementia are noted symptoms of NCL (Josephson et al. 2001, Wisneiwski et al. 1992), the enrichment for visual hallucinations may also suggest a role for DNAJC5 in schizophrenia-spectrum experiences. Further, enrichment for depression and dementia may point to a wider role for DNAJC5 in CNS-related conditions (Figure 5.4b; Supplementary Data 5.2). Other enriched terms included myoclonus, which is likely to refer to the impaired voluntary movement and myoclonic epilepsy which is seen in some individuals with NCL, with the term 'rapidly progressive' highlighting the degenerative nature of NCL.

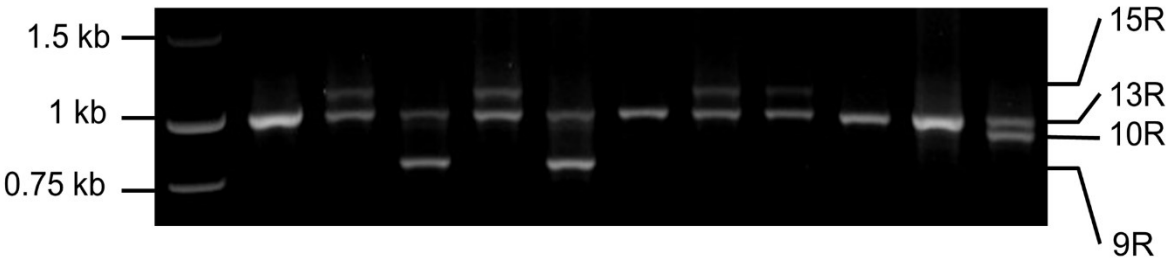
5.3.2 Genotyping of the MIR941 VNTR reveals sex differences and two unique schizophrenia-associated genotypes

The tandem repeat containing MIR941 has previously been shown to be polymorphic in the human population (Hu et al. 2012), with MIR941 itself being implicated in psychosis and depression (Jeffries et al. 2016, Belzeaux et al. 2012). We therefore carried out PCR analysis to genotype this VNTR in 342 individuals with a diagnosis of schizophrenia and 340 controls from a German and Central European cohort (see Section 2.1.2). This analysis identified four common alleles of the MIR941 VNTR in the general population (Figure 5.5a), which arranged to give nine of the possible 10 different genotypes, with no individuals found to be homozygous for the rarest allele, which contained 9 repeats of the VNTR. Sequencing of the four alleles revealed that the main repeating unit of the VNTR was 56 bp in length, with each repeat including a copy of the 22 bp mature MIR941 sequence, according to the miRNA sequence from miRbase (http://www.mirbase.org). Sequencing also confirmed the findings of Hu et al., demonstrating a common C/G SNP of unknown functional significance at base 15 of 22 of the miRNA sequence, which are referred to in their communication as MIR941, for copies with the C SNP, and MIR941*, for copies with the G SNP (Hu et al. 2012).

The shortest and rarest allele was found to contain 9 repeats of the VNTR and mature miRNA sequence (9R) and had an allele frequency in the control population of 5%. This allele contained three copies of MIR941 (blue) and six copies of MIR941* (green) in the following order: 1x MIR941*, 3x MIR941, 5 x MIR941* (Figure 5.5b). No homozygous 9R individuals were identified in this study.

Figure 5.5 Identification and sequence of MIR941 alleles.

(A) MIR941 VNTR alleles



(B) 9R allele

```
CACCCGGCTGTGTGGACATGTGCCCAGGGCCCAGGACAGCGCCACGGAAGAGGACG
CACCCGGCTGTGTGCACATGTGCCCAGGGCCCGGGACAGCGCCACGGAAGAGGACG
CACCCGGCTGTGTGCACATGTGCCCAGGGCCCGGGACAGCGCCACGGAAGAGGACG
CACCCGGCTGTGTGCACATGTGCCCAGGGCCCGGGACAGCGCCACGGAAGAGGACG
CACCCGGCTGTGTGGACATGTGCCCAGGGCCCGGGACAGCGCCACGGAAGAGGACG
CACCCGGCTGTGTGGACATGTGCCCAGGGCCCGGGACAGCGCCACGGAAGAGGACG
CACCCGGCTGTGTGGACATGTGCCCAGGGCCCGGGACAGCGCCACGGAAGAGGACG
CACCCGGCTGTGTGGACATGTGCCCAGGGCCCGGGACAGCGCCACGGAAGAGGACG
CACCCGGCTGTGTGGACATGTGCCCAGGGCCCGGGACAGCGCCACGGAAGAGGACG
```

(C) 10R allele

```
CACCCGGCTGTGTGGACATGTGCCCAGGGCCCAGGACAGCGCCATGGAAGAGGACA
CACCCGGCTGTGTGGACATGTGCCCAGGGCCCAGGACAGCGCCACGGAAGAGGACG
CACCCGGCTGTGTGCACATGTGCCCAGGGCCCGGGACAGCGCCACGGAAGAGGACG
CACCCGGCTGTGTGCACATGTGCCCAGGGCCCGGGACAGCGCCACGGAAGAGGACG
CACCCGGCTGTGTGCACATGTGCCCAGGGCCCGGGACAGCGCCACGGAAGAGGACG
CACCCGGCTGTGTGGACATGTGCCCAGGGCCCGGGACAGCGCCACGGAAGAGGACG
CACCCGGCTGTGTGGACATGTGCCCAGGGCCCGGGACAGCGCCACGGAAGAGGACG
CACCCGGCTGTGTGGACATGTGCCCAGGGCCCGGGACAGCGCCACGGAAGAGGACG
CACCCGGCTGTGTGGACATGTGCCCAGGGCCCGGGACAGCGCCACGGAAGAGGACG
CACCCGGCTGTGTGGACATGTGCCCAGGGCCCGGGACAGCGCCACGGAAGAGGACG
```

(D) 13R allele

```
CACCCGGCTGTGTGGACATGTGCCCAGGGCCCAGGACAGCGCCACGGAAGAGGACA
CACCCGGCTGTGTGGACATGTGCCCAGGGCCCGGGACAGCGCCACGGAAGAGGACG
CACCCGGCTGTGTGCACATGTGCCCAGGGCCCGGGACAGCGCCACGGAAGAGGACG
CACCCGGCTGTGTGCACATGTGCCCAGGGCCCGGGACAGCGCCACGGAAGAGGACG
CACCCGGCTGTGTGCACATGTGCCCAGGGCCCGGGACAGCGCCACGGAAGAGGACG
CACCCGGCTGTGTGGACATGTGCCCAGGGCCCGGGACAGCGCCACGGAAGAGGACGCACAGGACAGCGCCACGGAAGAGGACG
CACCCGGCTGTGTGCACATGTGCCCAGGGCCCGGGACAGCGCCACGGAAGAGGACG
CACCCGGCTGTGTGCACATGTGCCCAGGGCCCGGGACAGCGCCATGGAAGAGGACG
CACCCGGCTGTGTGCACATGTGCCCAGGGCCCGGGACAGCGCCACGGAAGAGGACG
CACCCGGCTGTGTGCACATGTGCCCAGGGCCCGGGACAGCGCCACGGAAGAGGATG
CACCCGGCTGTGTGGACATGTGCCCAGGGCCCGGGACAGCGCCACGGAAGAGGACG
CACCCGGCTGTGTGGACATGTGCCCAGGGCCCGGGACAGCGCCACGGAAGAGGACG
CACCCGGCTGTGTGGACATGTGCCCAGGGCCCGGGACAGCGCCACGGAAGAGGACG
```

Figure 5.5 Identification and sequence of MIR941 alleles.

(a) PCR across the MIR941 VNTR in control samples identified four main alleles, with PCR fragments ranging between approximately 800 bp to over 1 kb in size. Sequencing across these alleles demonstrated that each contained a different number of VNTR repeats which resulted in a different number of MIR941 copies. This ranged from nine repeats in the smallest and rarest allele (9R) to a predicted 15 repeats in the largest allele (15R).

(b) The 9R allele has nine repeats of the VNTR and miRNA, with three MIR941 copies (blue) and six MIR941 copies (green), which are defined by a C/G SNP at base 15 in the mature miRNA sequence.*

(c) The 10R allele has 10 repeats of the VNTR with three copies of MIR941 and seven copies of MIR941. Both the 9R and 10R alleles appear to be the same in terms of the positioning of MIR941/MIR941* copies, with the exception of the 9R allele missing the first copy.*

(d) The 13R allele has 13 repeats of the VNTR with seven copies of MIR941 and six copies of MIR941. Similar to the 10R allele, the 13R allele begins with two copies of MIR941* followed by three copies of MIR941. However, the sixth repeat of the VNTR contains a MIR941* copy and is extended from the usual 56 bp to 83 bp. Following this extended repeat, the 13R allele has a further four MIR941 copies and three MIR941* copies.*

The high GC content and repetitive nature of this region meant that we were unable to reliably sequence the 15R allele, however, due to its approximate 100 bp size difference compared to the 13R allele as visualised on an agarose gel, we predict that this allele may have an additional two copies of the 56 bp repeat.

The second rarest allele, with an allele frequency of 6%, had 10 copies of the VNTR (10R) including three copies of MIR941 and seven copies of the MIR941*, in a pattern of 2x MIR941*, 3x MIR941, 5x MIR941* (Figure 5.5c). We found 1.18% of the control population to be homozygous for the 10R allele.

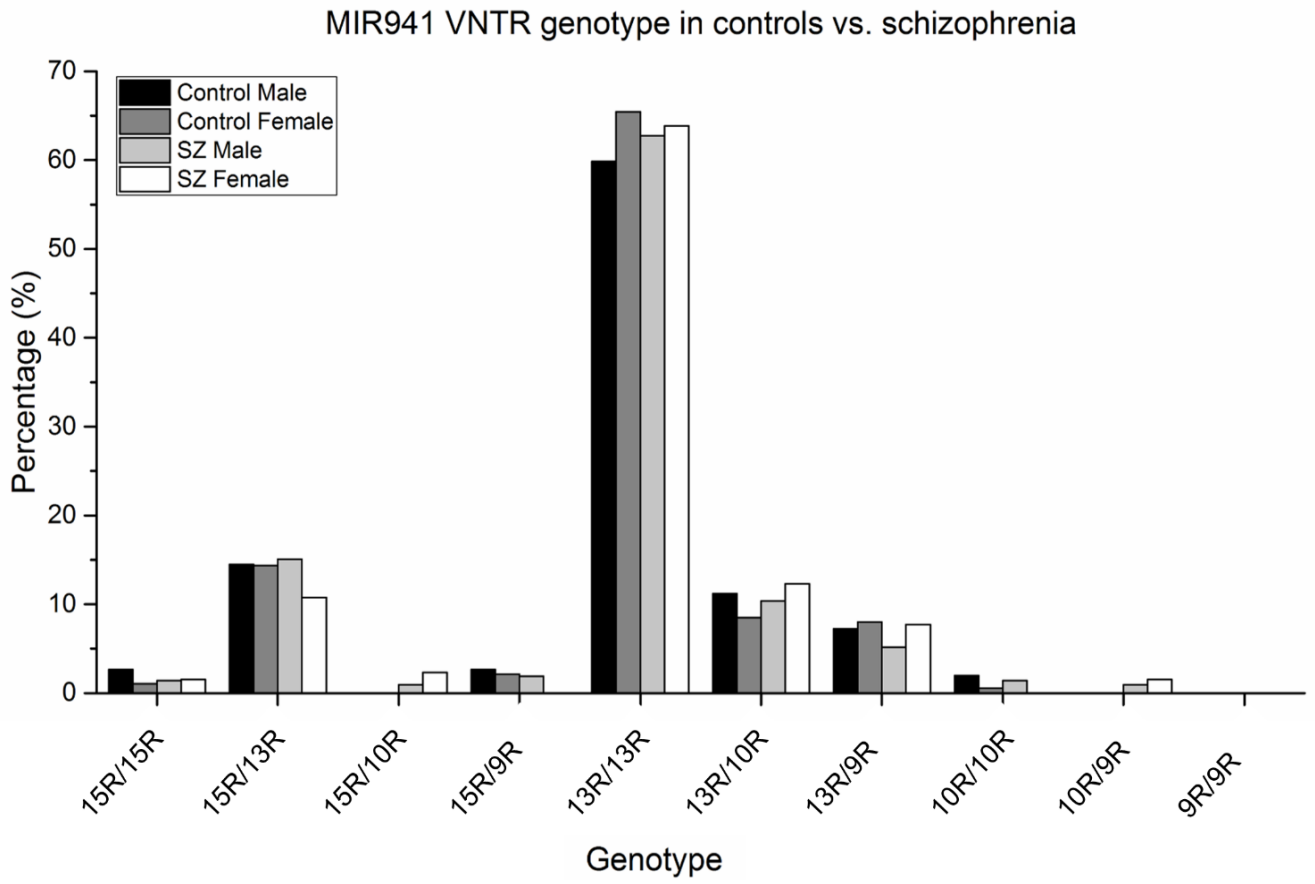
The most common allele was found to contain 13 repeats of the VNTR and mature miRNA sequence (13R), and had an allele frequency of 78.8% in the control population. The 13R allele included six copies of MIR941 and seven copies of MIR941*. The order of the miRNA copies was as follows: 2x MIR941*, 3x MIR941, 1x MIR941*, 4x MIR941, 3x MIR941* (Figure 5.5d). Instead of the usual 56 bp repeat, the seventh repeat in this allele had expanded to 83 bp in length. In the control population, 62.94% of individuals were found to be homozygous for the 13R allele, with 94.71% of individuals having at least one 13R allele.

Due to the size, the repetitive nature of the sequence, and the 71% GC content of this region, the largest allele was unable to be sequenced with confidence. However, given the approximate 100 bp size difference between the largest allele and the 13R allele on a gel (Figure 5.5a), we predict that the largest allele may contain up to two extra copies of the 56 bp repeat, with an estimated 15 copies of MIR941. As such, this allele is referred to as 15R. The 15R allele was found to have a frequency of 10.15% within the control population, with 1.76% of this group being homozygous, and 18.53% containing at least one 15R allele.

Comparing genotype analysis of the MIR941 VNTR in schizophrenia versus control populations notably demonstrated two genotypes that were present exclusively in the schizophrenia population. Five individuals in the schizophrenia cohort were found to have the 15R/10R genotype (1.46%) and four individuals had the 10R/9R genotype

(1.17%). This gave a total of 2.63% of individuals with schizophrenia in this data set who had a MIR941 VNTR genotype that was not found in control individuals (Figure 5.6, Table 5.1). No individuals from either population were found to be homozygous for the 9R allele, which may be expected due to its low allele frequency in both control and schizophrenia populations (5.00% and 4.24%, respectively). It may be of interest to note that the 9R allele was only found in the control cohort in combination with longer alleles (13R or 15R), whereas 1.17% of individuals with schizophrenia had the genotype 10R/9R which is comprised of two short alleles. This could suggest that the shortest allele with the fewest repeats of MIR941 may be detrimental but can be offset by pairing with a longer allele. However, the 15R/10R genotype was also found to be exclusive to the schizophrenia cohort in this study, with 1.46% of individuals with schizophrenia having this genotype. Based on the allele frequencies of the mixed sex control population, we would expect to find one individual in a cohort of 340 controls with the 10R/9R genotype (0.5%. 1.7 individuals), and 4 individuals with the 15R/10R genotype (1.3%, 4.42 individuals). Therefore, at present, the lack of representation of these genotypes in the control population may be due to chance given the low number of samples in this analysis. For this reason, it is not possible to determine risk associated with the schizophrenia-specific genotypes identified in this study, as much larger numbers would be required for genotyping before any association could be reliably identified. Further work is currently being carried out by Ana Illera-Lopez to extend this project and to determine whether these genotypes are found in other cohorts and at what frequency, including in Alzheimer's disease and healthy elderly cohorts.

Figure 5.6 and Table 5.1 Breakdown of MIR941 genotypes by diagnosis and sex.



	Control Male		Control Female		Schizophrenia Male		Schizophrenia Female	
Genotype	**Number**	**Percentage**	**Number**	**Percentage**	**Number**	**Percentage**	**Number**	**Percentage**
15R/15R	4	2.63	2	1.06	3	1.42	2	1.54
15R/13R	22	14.47	27	14.36	32	15.09	14	10.77
15R/10R	0	0.00	0	0.00	2	0.94	3	2.31
15R/9R	4	2.63	4	2.13	4	1.89	0	0.00
13R/13R	91	59.87	123	65.43	133	62.74	83	63.85
13R/10R	17	11.18	16	8.51	22	10.38	16	12.31
13R/9R	11	7.24	15	7.98	11	5.19	10	7.69
10R/10R	3	1.97	1	0.53	3	1.42	0	0.00
10R/9R	0	0.00	0	0.00	2	0.94	2	1.54
9R/9R	0	0.00	0	0.00	0	0.00	0	0.00
Total	152	100.00	188	100.00	212	100.00	130	100.00

Figure 5.6 and Table 5.1 Breakdown of MIR941 genotypes by diagnosis and sex.

Raw numbers and percentages of MIR941 VNTR genotypes in 340 controls (152 men and 188 women) and 342 individuals with schizophrenia (212 men and 130 women). The 13R/13R genotype was by far the most common with approximately 60% or more of each grouping being homozygous for this allele, with control males demonstrating the lowest percentage of 13R homozygotes at 59.87%. 15R/13R was found to be the second most common genotype at around 14% in both male and female control populations and 15% in the male schizophrenia cohort, however women with schizophrenia were underrepresented for this genotype with only 10.77% of this group having the 15R/13R genotype. The 13R/10R genotype ranged from 12.31% in the female schizophrenia group to 8.51% in the control female group, with both male populations displaying 10-11% of individuals with this genotype. On the other hand, the 13R/9R genotype was similar across both control and schizophrenia females as well as control males at 7-8%, while being decreased in men with schizophrenia at 5.19%. Most notably, the 15R/10R and 10R/9R genotypes were identified only in individuals with schizophrenia. Given that around 95% of control individuals have at least one copy of the 13R allele, these genotypes unique to the schizophrenia cohort may suggest that deviation from the 13-repeat allele could have detrimental effects.

Taken together, and with the knowledge that approximately 95% of control individuals have at least one 13R allele, we can speculate only that deviation from the 13-repeat allele in either direction may be detrimental, but further study would be needed to confirm this.

Genotype analysis was further stratified by sex, which revealed notable differences in genotype frequency between men and women in the control population, with the 10R/10R, 15R/15R, and 13R/10R genotypes being decreased in control women when compared to men, by 73.10%, 59.70%, and 23.88%, respectively, and the 13R/13R genotype increased by 9.29% in control women compared to men (Figure 5.6, Table 5.1).

Reanalysis by sex in the schizophrenia cohort showed the clearest differences for the schizophrenia-specific genotypes noted above, with the 15R/10R genotype increased by 145.74% in women with schizophrenia when compared to men with the same diagnosis. Further, the schizophrenia-specific 10R/9R genotype was found to be increased by 63.83% in women compared to men with schizophrenia. Finally, the 15R/13R genotype, which was present in both controls and cases, was found to be decreased by 28.63% in women with schizophrenia compared to men with schizophrenia (Figure 5.6, Table 5.1).

Given the sex differences observed in the control population, we went on to compare control versus schizophrenia women, and control versus schizophrenia men, in order to minimise skewing of the data due to sex-based genotype differences. In men with schizophrenia, we note a lower frequency of the genotypes 15R/15R, 15R/9R, and 13R/9R, with percentage decreases of 46.01%, 28.14%, and 28.31%, respectively, compared to the control male population (Figure 5.6, Table 5.1). In females, we see a

25.00% decrease of the 15R/13R allele and a 44.65% increase of the 13R/10R allele in women with schizophrenia. Further, in the female population, the 15R/10R and 10R/10R genotypes were observed only in control women, however the absence of these genotypes in women with schizophrenia may have been due to the smaller sample size for this cohort (130 women with schizophrenia versus 188 control women).

In order to test the statistical significance of the genotype differences noted between groups, Clump analysis was applied to the genotype data separated both by diagnosis and sex. Clump analysis uses Monte Carlo methods to evaluate the significance of chi squared values which are produced through multiple analyses outlined in Section 2.2.7.2, and tests the significance of these values by assessing how often the observed values can be replicated by randomly simulated data sets (Sham and Curtis 1995).

This analysis did not support a statistically significant difference between any of the five separate groupings of male and female control and/or schizophrenia genotypes (Table 5.2), suggesting that differences in MIR941 genotype are not associated with schizophrenia diagnosis or sex. The mixed sex control vs schizophrenia analysis approached significance for the T4 test (Chi squared p-value = 0.07), however this trend towards significance was not observed after separating the data by sex to account for both the differences in male and female sample number between groups, and the variability in genotype between the sexes, which may suggest that this is not a true result. While Clump analysis is useful for generating more accurate statistics when genotype frequency is low (below 5-10), some of the rarer genotypes had no data in this sample set which may have made assessment of statistical significance difficult. It is therefore likely that this analysis would need to be extended to a larger

cohort before potential statistical significance could be identified. However, variation between sexes in the general population, as well as the 15R/10R and 10R/9R genotypes being specific to individuals with schizophrenia in this analysis, and particularly over-represented in women with schizophrenia, remain of interest in understanding the potential roles of MIR941 VNTR genotype in CNS conditions.

	p-value from chi-squared distribution			
Analysis	**T1**	**T2**	**T3**	**T4**
Control vs schizophrenia (Mixed sex)	0.2402	0.2688	0.2400	0.0729
Control vs schizophrenia (Male)	0.8555	0.8348	0.4187	0.9132
Control vs schizophrenia (Female)	0.1093	0.2930	0.1445	0.2629
Control (Male vs Female)	0.8555	0.8348	0.4187	0.9132
Schizophrenia (Male vs Female)	0.5545	0.4892	0.2552	0.5109

Table 5.2 Clump analysis to determine statistical significance of genotype differences between cohorts.

Clump analysis was performed to compare five different combinations of groups including control vs schizophrenia cohorts (mixed sex), control vs schizophrenia males, control vs schizophrenia females, control males vs control females, and schizophrenia males vs schizophrenia females. Of the four separate Chi squared p-values returned by Clump analysis based on the multiple tests described in Section 2.2.7.2, statistical significance (Chi squared p-value < 0.05) was not achieved for any pairing of groups. The difference in genotype between mixed sex control and schizophrenia cohorts approached significance for the T4 test with a Chi squared p-value of 0.07, however this trend towards significance was not found upon separating the cohorts to control for sex differences, suggesting that the imbalance of male and female samples in the two cohorts may have been impacting the results.

5.3.3 DNAJC5 and the wider DNAJ gene family are deregulated in the DL-PFC of individuals with schizophrenia

MIR941 has been shown to be lost from the peripheral blood miRNA interaction networks of individuals who progress to psychosis at two-year follow up, compared to both controls and high risk individuals who do not progress to psychosis (Jeffries et al. 2016). However, this did not correlate with change in MIR941 expression levels in those that progressed to psychosis. Nonetheless, the loss of MIR941 from the miRNA networks of individuals experiencing psychosis highlighted this as a miRNA of potential interest in schizophrenia biology.

To assess expression around the DNAJC5/MIR941 locus in the brain, we made use of RNA-seq data from the DL-PFC of 155 individuals with schizophrenia and 196 controls from the Lieber Institute for Brain Development. No data was available in this resource for MIR941 or AK128776, thus the full length DNAJC5 transcript was used as a proxy for expression around this locus. As this analysis did not differentiate between DNAJC5 and AK128776, it is possible that the overall levels reported in this data set may include both transcripts. Of the 63,677 genes tested in this analysis, DNAJC5 was the 27th most significantly differentially regulated gene between groups in the DL-PFC ($F = 228.23$; $q = 7.68 \times 10^{-183}$), showing strong deregulation in individuals with schizophrenia when compared to controls (Table 5.3). This put DNAJC5 within the top 0.05% of genes that were most significantly deregulated in the schizophrenia DL-PFC in this analysis. While the data available for this study did not provide information on MIR941 expression, the finding that DNAJC5 is highly deregulated in the schizophrenia DL-PFC could potentially suggest deregulation across the locus encompassing AK128776 and MIR941. Research into miRNA biogenesis has demonstrated that miRNAs which are located within introns of coding genes are often

co-expressed with their host gene (Rodriguez et al. 2004, Baskerville and Bartel 2005), being processed from the intron of the coding RNA either before or after splicing (Kataoka et al. 2009, Kim and Kim 2007). While we cannot equate host gene expression with miRNA expression in this data set, we could hypothesise that strong deregulation of DNAJC5 in the schizophrenia brain would impact expression of the internal MIR941 transcript. This hypothesis should be tested using RNA-seq data that has been processed and analysed using techniques appropriate for miRNA expression analysis.

Due to the known role of DNAJ genes in a broad range of CNS conditions (Chen et al. 2016a, Tiwari et al. 2015, Vilarino-Guell et al. 2014, Edvardson et al. 2012, Liu et al. 2011), this analysis was next extended to encompass the complete DNAJ gene family. Firstly, this demonstrated that all 49 genes of the DNAJ family were identified as expressed in the human DL-PFC. Analysis of differential expression in schizophrenia and control brains further revealed that 46 out of the 49 DNAJ genes were significantly deregulated in the schizophrenia DL-PFC when compared to controls, at the significant q-value level of < 0.001 (Table 5.3).

Taken together, these data would strongly suggest a role for altered expression around the DNAJC5 locus in schizophrenia, with potential to deregulate both DNAJC5-related processes as well as those involving the brain-expressed miRNA, MIR941. In addition to associating the 20q13.3 locus, these data also suggest a wider role for the DNAJ gene family and the HSP40 heat shock response in schizophrenia and in healthy CNS function.

Gene name	F	q
DNAJC5	228.23	7.68E-183
DNAJC6	213.06	1.66E-178
GAK	138.77	1.50E-151
DNAJC11	137.07	8.50E-151
DNAJC27	107.85	4.83E-136
DNAJB2	97.35	7.60E-130
DNAJB14	88.29	5.31E-124
DNAJC14	87.36	2.28E-123
DNAJC19	87.22	2.82E-123
DNAJB9	83.20	1.76E-120
DNAJC13	76.48	1.51E-115
SEC63	76.44	1.64E-115
HSCB	75.60	7.19E-115
DNAJA3	70.00	2.03E-110
DNAJC3	68.52	3.45E-109
DNAJC4	68.14	7.17E-109
DNAJC12	67.18	4.65E-108
DNAJA2	62.08	1.37E-103
DNAJC18	61.15	9.70E-103
DNAJC2	61.10	1.08E-102
SACS	60.13	8.45E-102
DNAJC10	59.92	1.33E-101
DNAJC1	59.15	7.03E-101
DNAJC16	58.33	4.15E-100
DNAJB4	52.14	6.11E-94
DNAJC7	49.69	2.39E-91
DNAJA1	49.38	5.17E-91
DNAJC22	38.53	4.11E-78
DNAJC30	38.07	1.67E-77
DNAJB12	35.79	1.92E-74
DNAJC21	31.97	5.47E-69
DNAJB6	31.01	1.52E-67
DNAJC8	30.11	3.57E-66
DNAJC28	26.90	5.03E-61
DNAJC25	24.71	2.60E-57
DNAJC24	24.57	4.63E-57
DNAJC15	24.48	6.71E-57
DNAJB1	22.75	8.58E-54
DNAJC17	21.23	6.03E-51
DNAJB11	19.63	7.75E-48
DNAJC5G	19.33	3.08E-47
DNAJC9	18.77	4.21E-46
DNAJB5	18.50	1.49E-45
DNAJA4	14.67	3.07E-37
DNAJB7	12.95	3.39E-33
DNAJB13	6.86	5.65E-17
DNAJC5B	2.31	0.0014
DNAJB8	1.30	0.2419
DNAJB3	0.98	0.6614

Table 5.3 Differential expression of DNAJ genes in the schizophrenia DL-PFC compared to controls.

All 49 DNAJ family genes were found to be expressed in the DL-PFC, with 46 of the 49 found to be differentially expressed in the schizophrenia group. DNAJC5 was the most significantly deregulated DNAJ gene in individuals with schizophrenia, with an F value of 228.23 and a q-value of 7.68 x10^{-183}.

5.4 Discussion

The region at chromosome 20q13.3 is known to play a role in healthy CNS functioning, with the DNAJC5 gene at this locus encoding a neuroprotective pre-synaptic vesicle protein, with roles in synaptic maintenance, excitability, and transmission (Lopez-Ortega et al. 2017, Ahrendt et al. 2014, Donnelier and Braun 2014, Johnson, Ahrendt and Braun 2010). The human-specific, brain-expressed miRNA, MIR941, also lies within the first intron of the major DNAJC5 transcript, and, while its function is currently unknown, altered expression of this miRNA has been observed in the blood of individuals experiencing psychosis and depression (Jeffries et al. 2016, Belzeaux et al. 2012). The region encompassing MIR941 is a CpG island, and is also known to be a VNTR which is polymorphic in the general population (Hu et al. 2012). In this study, we describe an alternative DNAJC5 transcript, AK128776, which utilises this VNTR as its first exon (Figure 5.1). Warburton et al. has previously shown that a polymorphic VNTR at the schizophrenia-associated MIR137 locus acts as an internal promoter and supports differential expression in an allele specific manner (Warburton et al. 2015b, Warburton et al. 2015a), while others have demonstrated that the MIR137 VNTR can alter the RNA structure of transcripts utilising this region as an exon (Mamdani et al. 2013). It is therefore likely that the DNAJC5/MIR941 VNTR will not only change the copy number of MIR941, but also has the potential to alter the regulation of both DNAJC5 and AK128776 in an allele dependent manner, as well as the potential to alter the structure of the AK128776 RNA based on VNTR copy number. As this region of AK128776 is annotated on the UCSC Genome Browser as an untranslated exon (likely the 5' UTR), it is unlikely that changes in MIR941 VNTR repeat number would alter protein structure, though they may have the potential to alter the function of the 5' UTR, for example with regard to translation of the RNA. Data on chromatin states

and histone modifications suggests that the region around the MIR941 VNTR is an active promoter in neuronal progenitor cells and in the hippocampus (Figure 5.2b), with transcriptional regulatory properties identified across a number of other brain regions, which would support the MIR941 VNTR acting as both a promoter and regulator of transcripts across this locus in the brain. We also identify a primate-specific ERVK and Alu, approximately 1.9 kb upstream of the AK128776/MIR941 transcriptional start site, with histone modification data suggesting that this region may act as a promoter and enhancer in the brain (Figure 5.3).

As evidence from histone modification data suggested regulatory function of the MIR941 VNTR in the brain, it is possible that MIR941 VNTR genotype could result in altered expression or regulation at this locus. Enrichment analysis provided insight into which molecular pathways may be affected by modulation of DNAJC5 levels. We demonstrated a role for DNAJC5 in numerous brain-related processes, including regulation of synaptic function and the levels, secretion, and transport of neurotransmitters, as well as being associated with visual hallucinations and depression in humans (Figure 5.4). This would be consistent with altered pathways identified in DNAJC5 knockout mice which exhibit severe synaptic dysfunction (Ninkina et al. 2012, Ruiz, Biea and Tabares 2014), and would further support findings demonstrating differential expression of DNAJC5 in the brain in response to anti-depressant medication in mouse and rat models of depression (Malki et al. 2012, Yamada et al. 2001).

MiRNAs are known to control the expression of large gene networks, particularly in the brain, with relevance to schizophrenia and other CNS conditions (Olde Loohuis et al. 2017, Wright et al. 2015, Cao et al. 2016, Xu et al. 2012). Similarly, VNTRs have been shown to possess regulatory properties, and have frequently been demonstrated

to differentially regulate CNS-expressed genes based on genotype, with likely implications in both behavioural and psychiatric conditions (Warburton et al. 2015b, Paredes et al. 2013, Klenova et al. 2004). While variation in MIR941 VNTR genotype is common across population (Figure 5.5), we hypothesised that particular genotypes may be linked to schizophrenia risk. Genotyping of the MIR941 VNTR in schizophrenia and control cohorts of German and Central European descent failed to identify any significant difference in genotype frequencies between the cohorts. Two genotypes, 15R/10R and 10R/9R, were observed only in the schizophrenia grouping, though further studies on independent cohorts with a higher sample number would be required before any significance could be determined (Figure 5.6).

Analysis of RNA-seq data from the DL-PFC of 155 individuals with schizophrenia and 196 controls demonstrated that DNAJC5 was within the top 0.05% of genes that were most significantly deregulated in this brain region in schizophrenia. DNAJC5 has been shown to interact with Palmitoyl-Protein Thioesterase 1 (PPT1), with individuals with NCL (caused by mutations in DNAJC5) showing globally decreased protein palmitoylation, particularly of synaptic and lysosomal proteins (Henderson et al. 2016). The recent finding that palmitoylation is decreased in the DL-PFC of individuals with schizophrenia may therefore hint at a role for the deregulation of DNAJC5 in this mechanism in the schizophrenia DL-PFC (Pinner et al. 2016).

Finally, extending this analysis to the wider DNAJ gene family demonstrated that 46 of the 49 DNAJ genes were deregulated in the DL-PFC of individuals with schizophrenia when compared to controls ($q < 0.001$) (Table 5.3). While DNAJC9 has previously been linked to a subgroup of schizophrenia with attention deficit phenotypes (Liu et al. 2011), DNAJ (HSP40) proteins are predominantly known for their interaction with, and regulation of, HSP70 proteins (Zhao, Braun and Braun 2008, Fan, Lee and

Cyr 2003, Cyr and Ramos 2015), SNPs in which have repeatedly been identified as risk factors for schizophrenia (Bozidis et al. 2014, Kowalczyk et al. 2014, Kim et al. 2008). Further, activation of the heat shock response has been shown to modulate expression of schizophrenia and autism candidate genes in induced pluripotent stem cell models, including altering the regulation of the schizophrenia GWAS gene, ZNF804A (Lin et al. 2014). It is therefore possible that deregulation of heat shock pathways in the DL-PFC of individuals with schizophrenia could modulate known schizophrenia-associated gene networks. Alternatively, given the role of the DNAJ gene family in neuroprotective processes and in a wide range of CNS conditions, deregulation of this gene set in general may be indicative of poor CNS health.

5.5 Summary

Polymorphic variation at the DNAJC5/MIR941 locus alters the copy number of the brain-expressed MIR941 miRNA, as well as presenting the potential for allele-specific regulatory effects altering the expression or structure of DNAJC5 transcripts across this locus. We demonstrated variation at the MIR941 locus across the population, with two genotypes that were found to be exclusive to the schizophrenia cohort in this study. Further, RNA-seq analysis revealed that DNAJC5 is within the top 0.05% of the most highly deregulated genes in the schizophrenia DL-PFC. Strong deregulation of DNAJC5 in the brains of individuals with schizophrenia may be indicative of deregulation across this locus, also encompassing MIR941, which others have shown to be downregulated in schizophrenia. Extending this analysis also highlighted deregulation of the wider DNAJ gene family in schizophrenia, and suggested a significant role for the heat shock response as an underlying mechanism in schizophrenia biology, or as a potential marker for assessing brain health.

Chapter 6

Primate-specific SVA and LINE-1 insertions in zinc finger and glutamate gene evolution.

The work in the following section is contained within a manuscript entitled:

1. Distribution of primate-specific SVA and LINE-1 retrotransposon insertions across the human genome demonstrates a role for retrotransposon-mediated zinc finger and glutamate gene evolution.

6.1 Introduction

Chapter 5 focused on a region at the DNAJC5 locus which included a region of repetitive DNA referred to as a VNTR. Other classes of repeated elements, some containing VNTR regions, and collectively termed 'retrotransposable elements', are the focus of this chapter.

Like VNTRs, retrotransposable elements are known possess regulatory properties which can influence the expression of nearby genes through methods such as binding transcription factors, altering local chromatin structure, or altering gene splicing (Cordaux and Batzer 2009, Erwin et al. 2014, Gianfrancesco, Bubb and Quinn 2016a). However, retrotransposons are also mobile DNA elements that can insert new copies of themselves across the genome, and as such are drivers of genomic diversity, with the ability to provide genes with large new regulatory domains through evolution. For example, Savage et al. has previously demonstrated the ability of SVA elements to alter gene expression both *in vitro* and *in vivo* in cell lines and chick embryo models (Savage et al. 2013b, Savage et al. 2014). Similarly, others have demonstrated and reviewed the effect of LINE-1 retrotransposons on gene expression (Klawitter et al. 2015, Denli et al. 2015, Ngamphiw, Tongsima and Mutirangura 2014). The regulatory potential of ERV and Alu transposable elements was touched on briefly in Chapter 5, identifying them as potential modulators of expression at the MIR941 locus. In this chapter, we focus on analysing the movement of SVA and LINE-1 elements across the human genome in recent evolutionary history.

Work by Bundo et al. suggested a small but statistically significant increase in expression of the LINE-1 ORF2 transcript in the pre-frontal cortex of individuals with a diagnosis of schizophrenia. This work was originally carried out on an initial sample of 13 individuals with schizophrenia and 13 controls, and was replicated in a second

sample set including 35 individuals with schizophrenia and 34 controls (Bundo et al. 2014). In a follow up experiment on samples from three individuals with schizophrenia and three controls, distribution analysis of brain-specific LINE-1 insertions demonstrated that new LINE-1 insertions in individuals with schizophrenia were more likely to be found within or around genes involved in CNS function compared to tissue-specific insertions in control brains (Bundo et al. 2014).

The original intention for this chapter was to test the findings outlined in Bundo et al. (2014) on a significantly larger sample set, working with Eli Lilly and Company to assess LINE-1 expression in the DL-PFC of individuals with schizophrenia and control cohorts using RNA-seq data from the Lieber Institute for Brain Development. This would then have been followed up by distribution analysis to extend and build upon the small study in Bundo et al. which was based on samples from only three individuals. However, due to a number of issues regarding data access, we were not able to carry out the planned experiments. Instead, the work presented in this chapter attempts to understand the distribution of retrotransposon insertions in the human genome across evolutionary time, using publicly available data from the UCSC Genome Browser.

It is known that retrotransposon insertions have allowed the evolution of unique patterns of tissue-specific expression in higher primates and humans (Robbez-Masson and Rowe 2015), and as such, human-specific retrotransposon insertions are considered to be one of the key driving forces in the evolution of human-specific regulatory networks (Glinsky 2017). For this reason, the distribution of primate- and human-specific retrotransposons across the genome can provide information as to which genes or pathways have undergone recent evolutionary change uniquely in higher primate species.

It is known that retrotransposons are distributed non-randomly across the human genome. For example, it has been demonstrated that the SVA class of retrotransposons are preferentially found at regions of high GC content (Wang et al. 2005), with up to 60% of SVAs residing within genes or up to 10 kb upstream (Savage et al. 2013b). Previous work by the group has also demonstrated that SVAs are found in and around genes that are involved in CNS functioning, and that this gene set is over-represented for genes associated with Parkinson's disease (Vasieva et al. 2016b, Savage, Bubb and Quinn 2013a). On the other hand, LINE-1 elements are known to cluster at the X chromosome, and their distribution across this chromosome is thought to play a role in the non-random spreading and patterning of X chromosome inactivation (Ross et al. 2005, Abrusán, Giordano and Warburton 2008).

SVAs are the youngest and smallest class of retrotransposon in the human genome and are specific to hominids, thought to have first evolved around 13.6 million years ago and having expanded to include approximately 2700 elements in the hg19 human reference genome (Wang et al. 2005). For this reason, SVAs are of particular interest when studying the most recent retrotransposon-mediated evolutionary changes across the human genome, and their small number (0.13% of the genome) (Savage et al. 2013b) makes them an ideal starting point for studying retrotransposons on a genome-wide scale, in comparison to, for example, the vast number of LINE elements which comprise up to 20% of the human genome (Lander et al. 2001).

Here, we study the distribution of SVA elements across the genome, as well as the distribution of the three most recently evolved LINE-1 subfamilies (L1HS, L1PA2, and L1PA3) for comparison. The SVA class of retrotransposons is sub-divided into seven groups, SVA A to SVA F1, based on their SINE regions and evolutionary age of the SVAs, with the oldest, SVA A, being 13.6 million years old, and the younger SVA F

being 3.2 million years old. We simplistically split the seven SVA subfamilies into an evolutionary ‘older’ grouping, comprising SVA A, B, and Cs, and an evolutionary ‘younger’ grouping of SVA D, E, F, and F1s. The older group of SVAs are present in multiple primate species including chimpanzees and humans, whereas more recently evolved SVA E, F, and F1 elements are present only in humans. SVA Ds can be either human-specific, or also present in chimpanzees.

For comparison to SVAs, we selected the three most recently evolved LINE-1 subfamilies for analysis due to their comparable evolutionary age to SVAs. The oldest of the three, L1PA3 is 12.5 million years old, while L1PA2 and L1HS are 7.6 and 3.1 million years, respectively. L1PA3 and L1PA2 are found in both chimpanzees and humans, whereas L1HS elements are specific to humans (Khan et al. 2006). The L1PA4 subfamily and others are 18 million years of age and older, and thus were not considered an appropriate comparison for study in this analysis.

We analysed the distribution of reference SVAs and recent LINE-1 subfamilies across the human genome with the aim of first identifying any patterns in their distribution in recent evolutionary history, and then used this knowledge to examine which specific gene loci or biological pathways had been the target of recent retrotransposon-mediated change. This analysis was next extended to known or predicted germline SVA and LINE-1 retrotransposon insertion polymorphisms (RIPs), which are variable for their presence or absence across the human population. This allowed the comparison of distribution patterns between reference and non-reference insertions, identifying pathways that may be modulated in a variable manner between individuals due to the presence or absence of retrotransposons, and thereby shedding light on potential mechanisms involved in retrotransposon-mediated disease risk.

The UCSC Genome Browser and Galaxy were used to download the co-ordinates of all reference genome SVA and recent LINE-1 subfamily insertions annotated in the 'RepeatMasker' data set for the hg19 human genome build, and their distribution and clustering analysed per megabase (Mb) of the genome. This was repeated for SVA and LINE-1 RIPs using data on non-reference germline insertions from the TEBreak programme. Overlaying the co-ordinates of both reference and RIP SVA and LINE-1 insertions with annotated genes allowed us to generate lists of genes that had likely undergone recent retrotransposon-mediated evolution. Enrichment analysis was performed on these gene lists to gain insight into molecular pathways and processes that may have been altered through retrotransposon insertion.

6.2 Aims

- To determine the distribution patterns of human reference genome SVA and recent LINE-1 subfamily insertions.
- To determine genes or pathways which may have undergone SVA or LINE-1 retrotransposon-mediated change throughout primate and human evolution.
- To determine the distribution patterns of non-reference genome, polymorphic SVA and LINE-1 insertions in humans and compare to reference genome distribution.
- To determine genes or pathways containing SVA or LINE-1 retrotransposon insertion polymorphisms which may be altered in a variable manner in individuals due to presence/absence polymorphisms.

6.3 Results

6.3.1 Developing an unbiased method to study retrotransposon distribution across the genome

Our previous method in the lab, of selecting a candidate gene and then using a genome browser to look for nearby retrotransposons, is a useful but biased method in identifying genes which may have undergone retrotransposon-mediated changes throughout primate evolution. In order to more widely and unbiasedly investigate the distribution of retrotransposons across the genome, and identify potential genes or regions which may have been targets of retrotransposon-mediated evolution, I (with computational support from Dr. Bethany Geary at the University of Manchester) developed a basic bioinformatic method to map retrotransposon distribution.

This was first carried out to look at the distribution of SVAs, as their relatively small number in the genome made them a useful set for initial testing. For simplicity in this initial attempt, we combined all SVA subfamilies into one larger group of SVAs. We used Galaxy to access the UCSC table browser, through which the co-ordinates of all SVA elements in the human genome build 19 'Repeat Masker' track were downloaded. The co-ordinates were then separated by chromosome and saved individually as text files. A short R script, written by Dr. Bethany Geary from the University of Manchester, was used to read in two text files (1) the SVA co-ordinates, and (2) a list of specified ranges, which in this case were 1 Mb windows, presented in the text file as 1-1,000,000, 1,000,001-2,000,000, and so on for the length of the chromosome. The R script then counted the number of SVA co-ordinates which fell into the specified ranges, and returned a text file for each chromosome detailing the number of SVAs in each Mb. This process can easily be adapted to provide higher resolution mapping by decreasing the size of the specified ranges, however, the lack of overlapping windows

is likely to mean that some retrotransposon-dense regions will unfortunately not be identified. However, in this case, counting the number of SVAs in static 1 Mb windows was adequate to provide a simple basis for further inspection of the most SVA-dense genomic regions. All retrotransposon count data generated using this method is contained within Supplementary Data 6.1.

6.3.2 Both older and more recent reference SVAs cluster at specific zinc finger loci, particularly on chromosome 19

We first aligned the SVA count data generated using the above method with counts for transcript number per Mb (Supplementary Data 6.1). Data on the number of transcripts at 1 Mb intervals across the genome was generated using a modified version of the publicly available R script, provided by Dr Giovanni M Dall'Olio through the BioStars website (https://www.biostars.org/p/169171/#169211). Briefly, this made use of the 'Homo.sapiens' Bioconductor package, through which one can access information on all the transcripts in the human genome build 19 'known gene' data set from UCSC, through the 'human.genes' object of the 'Homo.sapiens' library. Dall'Olio's script then specifies particular genomic ranges and counts the number of transcripts within each window for the whole genome, outputting the information as four data columns, representing the chromosome name, the co-ordinates used for counting, the DNA strand, and the total number of transcripts counted within the given window. Modifying this script for our work, we set the range to 1 Mb.

Savage et al. had previously demonstrated the positive correlation between SVA density and gene density at the genomic scale (Savage et al. 2013b), though this had not been studied on individual chromosomes. Therefore, after aligning genome-wide data on SVA counts with counts for the number of transcripts per Mb, we calculated

the correlation coefficient for the number of transcripts and the number of SVAs per Mb of each chromosome.

This analysis showed that 22 of the 24 human chromosomes did indeed follow the pattern of increased SVA number at increasingly transcript dense regions, to varying degrees (Table 6.1). However, chromosome 19 and chromosome Y stood out clearly as the only chromosomes which did not follow this pattern, instead showing no correlation, with a correlation coefficient for chromosome 19 of -0.055 and for chromosome Y of -0.140. As there are large regions of sequencing data missing for the Y chromosome on the UCSC Genome Browser, we were unable to confirm whether this trend was representative of the whole chromosome, and therefore did not continue our investigation into SVAs on the Y chromosome. In order to determine the reason behind the negative correlation between SVA and transcript number on chromosome 19, we plotted the number of transcripts versus the number of SVAs per Mb across chromosome 19 (Figure 6.1a). This demonstrated that the negative correlation on this chromosome was due to a small number of regions which were vastly over-represented for SVAs compared to their transcript number.

To get a clearer picture of whether these regions with high numbers of SVAs had undergone rapid or sustained change over evolutionary time, we reanalysed the distribution of SVAs by evolutionary age, first separating evolutionary older SVAs A-C, and comparing their distribution to the more recently evolved SVA D-F1 subfamilies in an ‘old vs new’ analysis. Overlaying the clustering patterns of the older and more recent subfamilies of SVAs and plotting this data as a heat map allowed rapid visualisation of these SVA dense loci across chromosome 19 and the wider genome (Figure 6.1b).

	Correlation coefficient (r)			
Region	Reference SVAs	RIP SVAs	Reference LINE-1	RIP LINE-1
Whole genome	0.351633878	0.305886765	-0.307915231	-0.230497235
Chr1	0.335103511	0.249894976	-0.211254833	-0.142039136
Chr2	0.410900907	0.362230128	-0.264629378	-0.22195854
Chr3	0.452071632	0.244541099	-0.366828346	-0.251843313
Chr4	0.336122289	0.260744281	-0.288887259	-0.215586953
Chr5	0.272138074	0.208388196	-0.361159143	-0.279509935
Chr6	0.544926431	0.439995731	-0.260330375	-0.078669674
Chr7	0.320827891	0.167051786	-0.309998692	-0.229521986
Chr8	0.156825407	0.089453876	-0.486696742	-0.255263382
Chr9	0.298495366	0.356033771	-0.256648181	-0.072318524
Chr10	0.284397605	0.421492217	-0.301079394	-0.325413473
Chr11	0.376946155	0.297180153	-0.332275693	-0.331016545
Chr12	0.535019589	0.299430888	-0.381738659	-0.316756022
Chr13	0.558096476	0.347398301	-0.284199123	-0.173069849
Chr14	0.322441334	0.303475371	-0.308662671	-0.351819178
Chr15	0.267194732	0.131471722	-0.206419007	-0.175963787
Chr16	0.504027865	0.267861315	-0.302575007	-0.12790296
Chr17	0.236569698	0.262782808	-0.375845698	-0.293426604
Chr18	0.453501425	0.25459194	-0.188478669	-0.280139556
Chr19	-0.055021488	0.144299263	-0.378120638	-0.314417114
Chr20	0.326204794	0.419643121	-0.247494996	-0.197290327
Chr21	0.419117555	0.138102718	-0.402388991	-0.333049121
Chr22	0.344448376	0.208382629	-0.069580754	-0.108982206
ChrX	0.219723753	0.159311425	-0.046493363	-0.146019181
ChrY	-0.140254542	0.056840401	0.102967532	-0.08759664

Table 6.1 Correlation coefficient comparing the correlation of retrotransposon and transcript number per chromosome.

The correlation coefficient (r) demonstrates the strength of correlation between retrotransposon number and transcript number. A positive r value indicates increased retrotransposon insertion at genic regions, whereas a negative r value indicates decreased retrotransposon insertion at genic regions. Overall, both reference genome and RIP SVAs are preferentially found to be increased at genic regions across all chromosomes except chromosome 19, whereas reference and RIP LINE-1 elements are preferentially found at gene poor regions.

Figure 6.1 SVA elements preferentially cluster on chromosome 19.

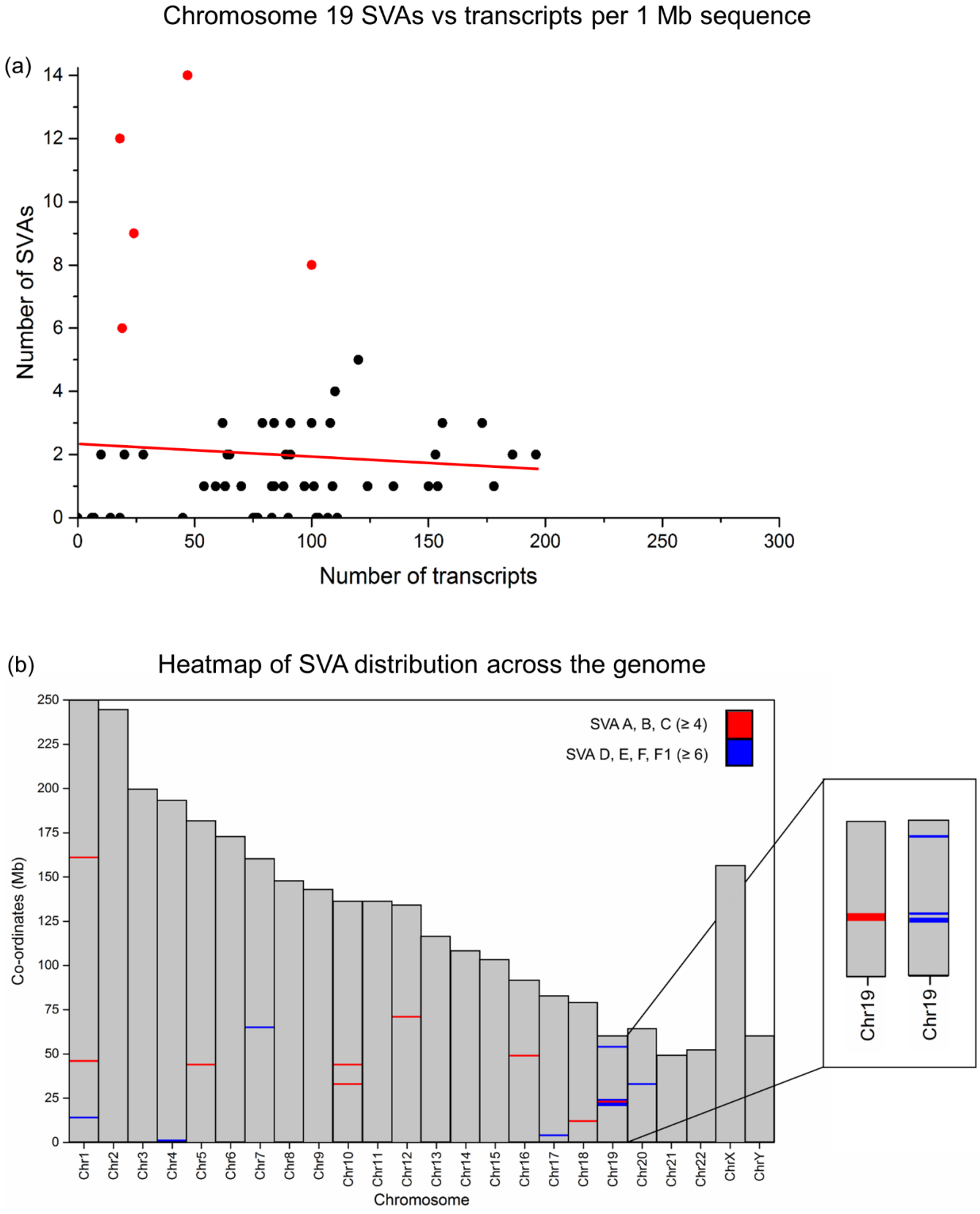


Figure 6.1 SVA elements preferentially cluster on chromosome 19.

(a) Chromosome 19 is the only autosome that displays no correlation between SVA and transcript number per Mb (correlation coefficient = -0.055). The apparent trend of decreasing SVAs with increasing transcript number is skewed by a small number of regions which are vastly over-represented for SVAs based on their gene density (red).

(b) Red bars indicate regions with four or more SVA A-Cs (0.43% of all SVA A-C elements), while blue bars identify regions with six or more SVAs of the more recent D-F1 subfamilies (0.34% of all SVA D-F1 elements). Both older and younger SVAs preferentially cluster at the chr19:20,000,000-24,000,000 locus, as demonstrated in the boxed region. We also note eight separate regions that have been significant targets of older subfamily SVA-mediated change, and six additional regions that have been preferential targets of SVA D-F1 subfamily insertion, including a second chromosome 19 locus. This suggests sustained SVA-mediated evolution at specific regions on chromosome 19, from the evolution of the earliest SVA class 13.6 million years ago to more recent human-specific changes.

This analysis revealed a four Mb locus at Chr19:20,000,000-24,000,000 that was the key region driving the negative correlation between SVA and transcript number on this chromosome. The four Mb chromosome 19 locus also represented the only region of the genome at which both older and more recent SVA groupings clustered together (Figure 6.1b).

That this region in particular is enriched for both older and younger SVA subfamilies suggests that the Chr19:20,000,000-24,000,000 locus has consistently been the target of SVA insertion and retention throughout the evolution and divergence of higher primates and humans, from the appearance of the SVA A subfamily approximately 13.6 million years ago (Mya) up until, and likely throughout, human evolution.

We identified 108 transcripts across the Chr19:20,000,000-24,000,000 locus encoded by 49 genes, of which 32 (65.3%) were zinc finger genes of the C_2H_2-type ZNF family (Figure 6.2). 41 SVAs (17 older SVAs, and 24 younger SVAs) were also identified across the region, which gave an average of one SVA per 2.63 transcripts, or one SVA per 1.2 genes. This was 11-fold higher than the total transcript average of one SVA per 29.45 transcripts, and a 9-fold increase over the total gene average of one SVA per 10.63 genes.

The remaining region highlighted on chromosome 19 (Chr19:53,000,000-54,000,000) was enriched only for the younger classes of SVAs (Figure 6.3a) and contained 33 genes (which encoded 100 transcripts), of which 25 (75.76%) were zinc finger genes of the ZNF family. This one Mb locus contained eight SVAs, with an average of one SVA per 12.5 transcripts or per 4.13 genes, a 2.36- or 2.58-fold increase over the average.

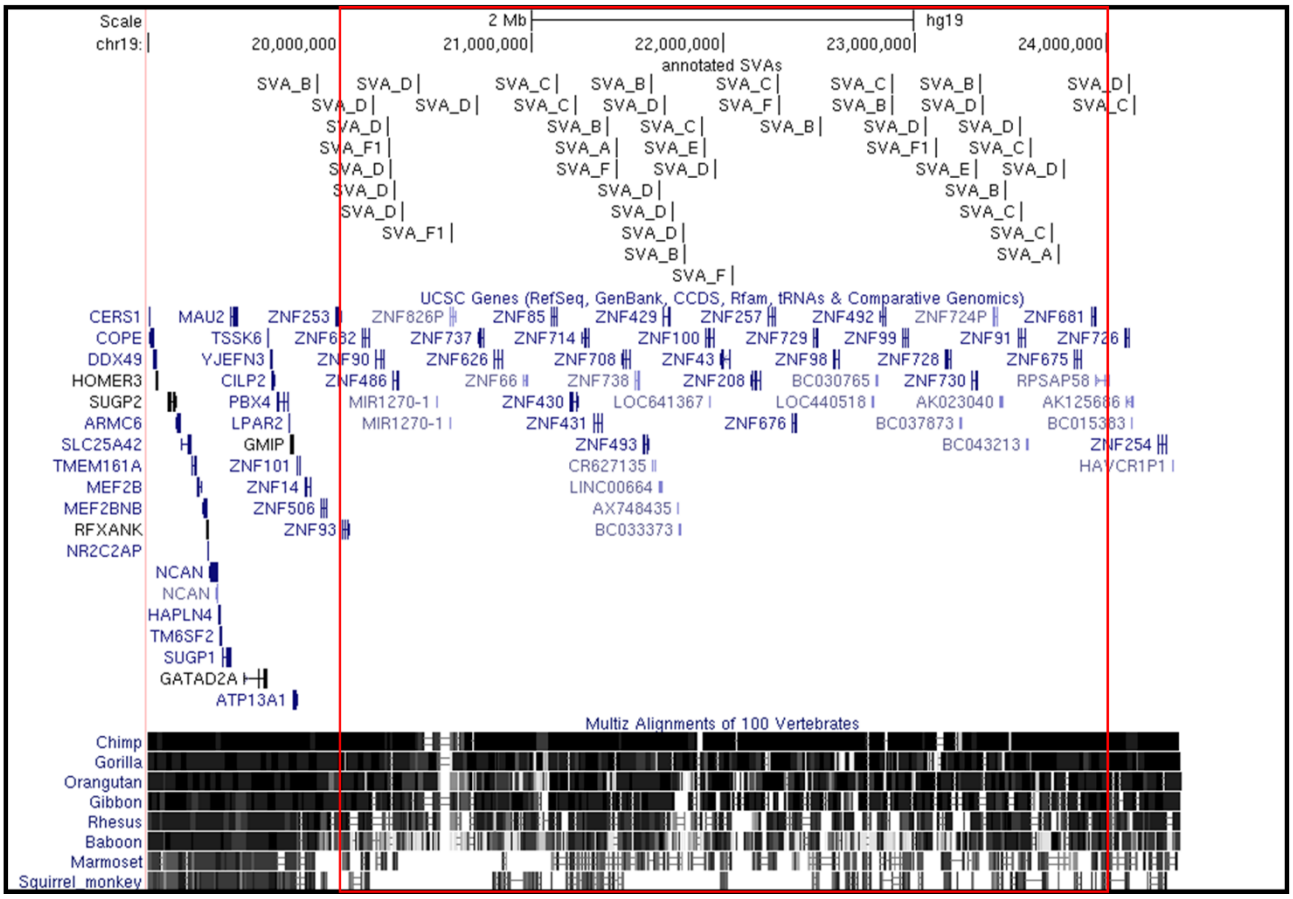


Figure 6.2 SVAs on chromosome 19 predominantly cluster at a four Mb ZNF zinc finger locus.

The red boxed region represents the SVA-dense chr19:20,000,000-24,000,000 locus, which contains 49 genes, 32 of which are zinc finger genes of the ZNF family, and 41 SVAs. Observation of the flanking regions demonstrates that the SVAs clearly cluster over the zinc finger genes, with minimal SVAs found outside of this locus. The Multiz alignment in this instance contains only primate species, as SVAs are found only in the primate lineage. This demonstrates the evolution of the zinc finger region in different primate species, which is likely due in part to SVA insertion. The flanking region to the left shows clear conservation as denoted by solid black colouring, while the large white gap on the right of the Multiz track represents the centromere of chromosome 19, over which no sequence data is available.

Figure 6.3 Three of the remaining five genome-wide regions enriched for reference SVA D-F1 subfamily insertions are ZNF zinc finger clusters.

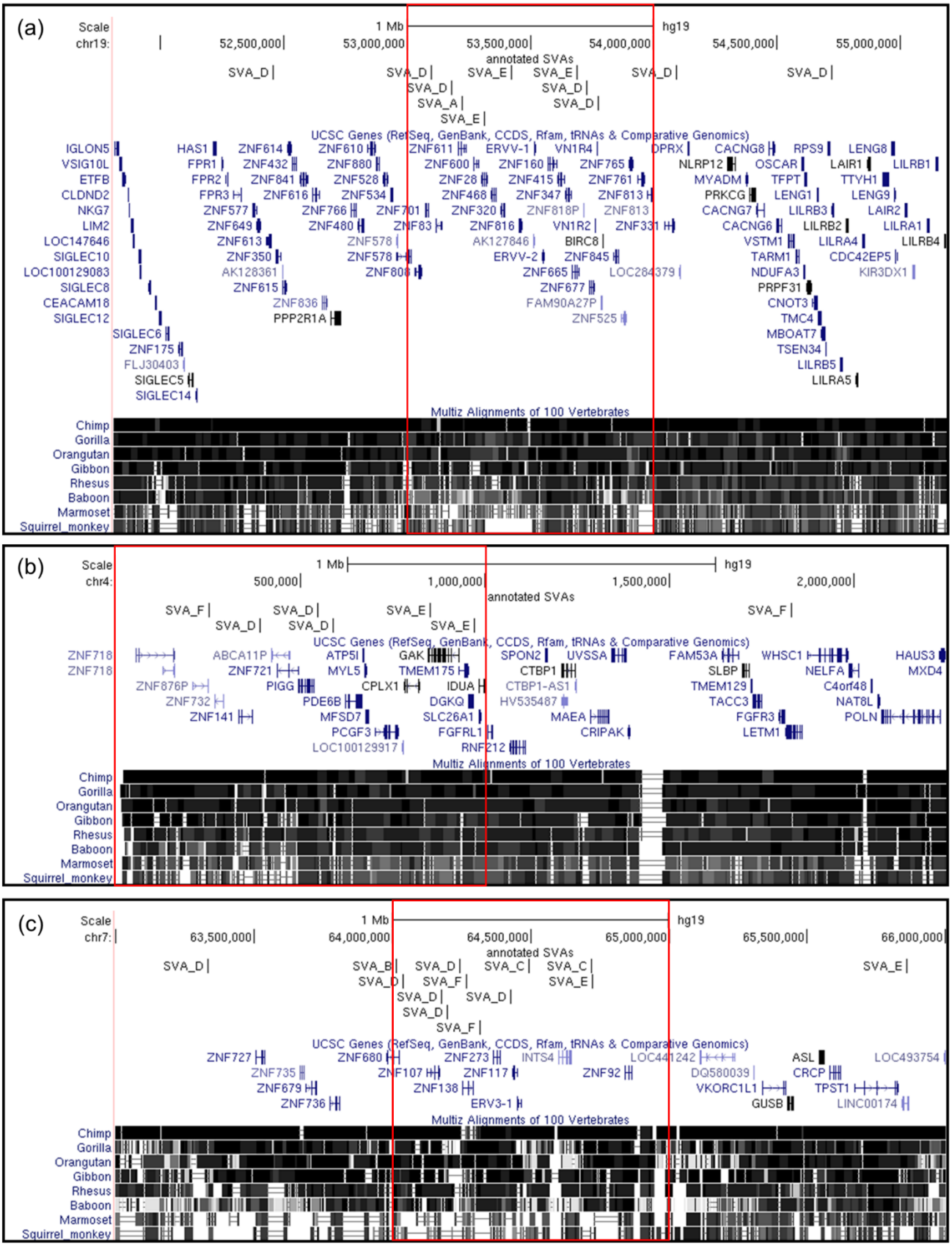


Figure 6.3 Three of the remaining five genome-wide regions enriched for reference SVA D-F1 subfamily insertions are ZNF zinc finger clusters.

(a) The chr19:53,000,000-54,000,000 locus (boxed) contains one SVA A, four SVA Ds, and three SVA Es, as well as 25 ZNF genes. Observation of the regions flanking the boxed locus demonstrate that the SVA clustering is limited to the zinc finger region, with few SVAs over non-ZNF genes at this locus.

(b) The chr4:1-1,000,000 locus (boxed), encompasses three SVA Ds, two SVA Es, and an SVA F, arranged over a stretch of six ZNF zinc finger genes on the tip of chromosome 4. Of note, two human-specific SVA Es are directly over a genome-wide associated region for Parkinson's Disease (PD) encompassing the GAK and DGKQ genes, with one SVA E within a GAK intron, and the second lying within 2 kb of the DGKQ transcriptional start site. SVAs at this region could modulate expression of both zinc finger and PD-related genes at this locus uniquely in primates and humans. Again, the SVAs are predominantly around the ZNF region, with few SVAs outside of this Mb.

(c) The chr7:64,000,000-65,000,000 locus (boxed) contains one SVA B, two SVA Cs, five SVA Ds, one SVA E, and two SVA Fs, spread across six zinc finger genes. While the flanking regions contain a number of additional ZNF genes, SVAs were not enriched outside of this 1 Mb region.

The visualisation of zinc finger loci which are over-represented for SVAs demonstrates that the SVAs at these regions are clustered specifically over the ZNF genes, with few SVAs in the flanking regions around non-ZNF genes.

Of the remaining five loci at which clustering of a high number of SVA D-F1s was seen, two of these, at Chr4:1-1,000,000 and at Chr7:64,000,000-65,000,000, were also ZNF cluster regions (Figure 6.3b, c). The former, at chromosome 4, encompasses six zinc finger genes and multiple SNPs for Parkinson's disease risk across the GAK and DGKQ genes, as well as containing six younger SVAs (Figure 6.3b), while the latter at chromosome 7 contains eight younger SVAs over a region containing six zinc finger genes (Figure 6.3c). This would suggest a sustained drive for SVA-mediated evolution at the Chr19:20,000,000-24,000,000 locus throughout primate evolution over the last 13.6 million years, and a more recent evolution uniquely in higher primates, including chimpanzees and humans, involving multiple other zinc finger loci.

The remaining three non-ZNF loci with an equally high number of SVA D-F1s are outlined in Table 6.2 and include the PRAME (preferentially expressed antigen in melanoma) gene family locus at Chr1:13,000,000-14,000,000 which encompasses 12 PRAMEF genes and seven SVAs. Expression of the PRAME gene family is primarily restricted to the testis, with aberrant expression noted in numerous cancers, including melanoma, leukaemia, breast cancer, head and neck cancer, Hodgkin's lymphoma, and many others, with increased PRAME expression typically associated with metastasis and overall poor prognosis (Ding et al. 2012, Ercolak et al. 2015, Field et al. 2016, Sun et al. 2016, Szczepanski et al. 2013). The PRAME gene family is highly variable between species, having undergone amplification and expansion across mammalian species (Chang et al. 2011). Particularly, the PRAME locus on chromosome 1, which we show to be over-represented for SVAs, is known to have duplicated approximately 3 Mya, thereby giving rise to hominin-specific changes at this locus (Birtle, Goodstadt and Ponting 2005).

Chromosome	Start	End	Total SVAs	SVA A-C	SVA D-F1	Subtype clustering	Genes
chr1	13000001	14000000	7	0	7	Young	LOC440563, LRRC38, PDPN, PRAMEF10, PRAMEF13, PRAMEF16, PRAMEF17, PRAMEF19, PRAMEF20, PRAMEF22, PRAMEF3, PRAMEF5, PRAMEF6, PRAMEF8, PRAMEF9
chr1	45000001	46000000	8	4	4	Old	BEST4, BTBD19, C1orf228, CCDC163P, EIF2B3, HECTD3, HPDL, KIF2C, LOC400752, MIR5584, MMACHC, MUTYH, PLK3, PRDX1, PTCH2, RNF220, RPS8, SNORD38A, SNORD38B, SNORD46, SNORD55, TCTEX1D4, TESK2, TMEM53, TOE1, UROD, ZSWIM5
chr1	160000001	161000000	4	4	0	Old	AK093299, ATP1A2, ATP1A4, CASQ1, CD244, CD48, CD84, COPA, DCAF8, F11R, IGSF8, ITLN1, ITLN2, KCNJ10, KCNJ9, LY9, NCSTN, NHLH1, PEA15, PEX19, PIGM, SLAMF1, SLAMF6, SLAMF7, SUMO1P3, VANGL2
chr4	1	1000000	6	0	6	Young	ABCA11P, ATP5I, BC020343, CPLX1, DGKQ, DKFZp547K2416, GAK, IDUA, LOC100129917, MFSD7, MYL5, PCGF3, PDE6B, PIGG, SLC26A1, TMEM175, ZNF141, ZNF595, ZNF718, ZNF721, ZNF732, ZNF876P
chr5	43000001	44000000	7	4	3	Old	ANXA2R, C5orf28, C5orf34, CCL28, DQ601842, HMGCS1, LOC100132356, LOC100506639, LOC100652772, LOC153684, LOC648987, NIM1, NNT, PAIP1, ZNF131
chr7	64000001	65000000	11	3	8	Young	AK057766, AK097702, BC044608, BC053669, CCT6P3, DQ596928, ERV3-1, INTS4, LOC100128885, LOC641746, ZNF107, ZNF117, ZNF138, ZNF273, ZNF680, ZNF92
chr10	32000001	33000000	5	4	1	Old	ARHGAP12, C10orf68, CCDC7, EPC1, KIF5B
chr10	43000001	44000000	7	4	3	Old	AK123067, BMS1, CSGALNACT2, FXYD4, HNRNPF, HNRPF, MIR5100, RASGEF1A, RET, ZNF33B, ZNF37BP, ZNF487P
chr12	70000001	71000000	5	4	1	Old	BC031864, BC042465, BEST3, C12orf28, CNOT2, KCNMB4, LRRC10, Mir_548, PTPRB, RAB3IP
chr16	48000001	49000000	5	4	1	Old	ABCC11, ABCC12, LOC100507577, LONP2, MIR548AE2, N4BP1, SIAH1, U6
chr17	3000001	4000000	6	0	6	Young	AB062083, ASPA, ATP2A3, C17orf85, CAMKK1, CTNS, DKFZp761G0818, EMC6, GSG2, ITGAE, OR1A1, OR1A2, OR1D4, OR1E1, OR1E2, OR1G1, OR3A1, OR3A2, OR3A3, OR3A4P, P2RX1, P2RX5, P2RX5-TAX1BP3, SHPK, SPATA22, TAX1BP3, TRPV1, TRPV3, ZZEF1
chr18	11000001	12000000	5	4	1	Old	CHMP1B, DQ570262, DQ570920, DQ571750, DQ572814, DQ573573, DQ575348, DQ575585, DQ576414, DQ576785, DQ577323, DQ577471, DQ577739, DQ578037, DQ578665, DQ579207, DQ579650, DQ581117, DQ582047, DQ582489, DQ583379, DQ584035, DQ584752, DQ584777, DQ585728, DQ586209, DQ586641, DQ587169, DQ587882, DQ588090, DQ588121, DQ588286, DQ589220, DQ589620, DQ590893, DQ590984, DQ591120, DQ591184, DQ594388, DQ594439, DQ594539, DQ595824, DQ596015, DQ596967, DQ597648, DQ599038, DQ599577, DQ600032, DQ600527, DQ600844, GNAL, IMPA2, MPPE1, PIEZO2
chr19	20000001	21000000	9	0	9	Young	MIR1270-1, ZNF253, ZNF486, ZNF626, ZNF66, ZNF682, ZNF737, ZNF826P, ZNF90, ZNF93
chr19	21000001	22000000	14	7	7	Both	AX746719, AX748435, BC033373, CR627135, LINC00664, LOC641367, ZNF100, ZNF429, ZNF43, ZNF430, ZNF431, ZNF493, ZNF708, ZNF714, ZNF738, ZNF85
chr19	22000001	23000000	6	4	2	Old	BC030765, LOC440518, ZNF208, ZNF257, ZNF43, ZNF492, ZNF676, ZNF729, ZNF98, ZNF99
chr19	23000001	24000000	12	6	6	Both	AK022793, AK023040, BC037873, BC038574, BC043213, RPSAP58, ZNF675, ZNF681, ZNF724P, ZNF728, ZNF730, ZNF91
chr19	53000001	54000000	8	1	7	Young	AK127846, BIRC8, ERVV-1, ERVV-2, FAM90A27P, TPM3P9, VN1R2, VN1R4, ZNF137P, ZNF160, ZNF28, ZNF320, ZNF321P, ZNF347, ZNF415, ZNF468, ZNF525, ZNF578, ZNF600, ZNF611, ZNF665, ZNF677, ZNF701, ZNF702P, ZNF761, ZNF765, ZNF808, ZNF813, ZNF816, ZNF816-ZNF321P, ZNF818P, ZNF83, ZNF845
chr20	32000001	33000000	6	0	6	Young	ACTL10, AHCY, ASIP, C20orf144, CBFA2T2, CHMP4B, E2F1, EIF2S2, ITCH, MIR4755, NECAB3, PXMP4, RALY, SNTA1, ZNF341

Table 6.2 Regions of the human genome with the highest SVA clustering.

Top 18 regions of the genome containing the highest number of SVAs per Mb, displaying the loci co-ordinates (hg19), number of older (SVA A-C) and younger (SVA D-F1) SVAs within each Mb, and the genes contained within these regions according to the UCSC Genome Browser's hg19 'known gene' data set.

Further, the chromosome 1 PRAME locus is thought to have undergone further duplication events and is not yet fixed within the population, remaining variable for copy number in modern humans (Birtle et al. 2005). The region around the TRPV1 and TRPV3 genes at Chr17:3,000,000-4,000,000 is also over-represented for younger SVAs, with six SVA Ds across this region. The TRPV1 and TRPV3 genes (transient receptor potential cation channel subfamily V) are activated by capsaicin and noxious heat stimuli, and signal thermal pain by activating sensory neurons (Chung, Jung and Oh 2011).

Further, the Chr20:3,2000,000-33,000,000 locus was found to contain four SVA Ds, one SVA F, and one SVA F1, with two SVA Ds found end to end 3' of the PXMP4 gene, and one SVA D, and the SVA F/F1s found between the EIF2S2 and ASIP genes (Table 6.2). PXMP4 (peroxisomal membrane protein 4), also known as PMP24, is a peroxisomal membrane protein (Reguenga et al. 1999), hypermethylation at which has been associated with prostate cancer (Wu and Ho 2004, Zhang et al. 2010b). EIF2S2, or EIF2beta (eukaryotic translation initiation factor 2 subunit beta), is involved in the early steps of protein synthesis (Singh and Wahba 1996, Singh, Aroor and Wahba 1994), while ASIP (agouti signalling protein) is involved in pigmentation (Kanetsky et al. 2002), with polymorphisms in this gene being linked to malignant melanoma (Duffy et al. 2010, Nan et al. 2009).

Aside from the four megabase region on chromosome 19, we did not observe the clustering of SVA A-C subfamilies at any further zinc finger loci. The remaining seven regions which were found to be overrepresented for older SVA subfamilies span chromosomes 1, 10, 12, 16, and 18, with the genes in these regions listed in Table 6.2.

ZNF zinc fingers are the largest family of transcription factors in the human genome (Vaquerizas et al. 2009). This gene family has a complicated evolutionary history due to multiple rounds of duplication between species, which has resulted in significant diversity (Emerson and Thomas 2009, Nowick et al. 2011). Further analysis of the four zinc finger clusters shown to be targeted by SVA insertions in this analysis demonstrated that two loci (Chr4:1-1,000,000; and Chr7:64,000,000-65,000,000) were part of the primate-specific ZNF91 subfamily originating from the Chr19:20,000,000-24,000,000 locus. This region is known to have partially duplicated numerous times across multiple chromosomes throughout primate evolution (Hamilton et al. 2006). The fact that ZNF91 resides within the most SVA-dense region of the genome, and a region which has duplicated to other chromosomes and acquired further SVA insertions, is of particular interest due to ZNF91's known ability to repress SVAs (Jacobs et al. 2014). We know that the SVA D-F1 SVAs inserted into these regions post-duplication, as the duplicated ZNF loci are present within primate species which pre-date the evolution of SVAs, such as the Rhesus macaque and baboon. The region encompassing ZNF91 has itself undergone SVA-mediated change, as was apparent from observation of Multiz vertebrate sequence alignments through the UCSC genome browser which showed gaps in primate species conservation where SVAs had inserted at this locus (Figure 6.4). ZNF91 had two chimp- and human-specific SVAs of the C and D class within the third intron of the gene, an additional SVA C approximately 24 kb upstream of the ZNF91 transcriptional start site, and an SVA B around 67 kb downstream of the 3' UTR. The location of these SVAs may impact ZNF91 expression and splicing uniquely in higher primate species such as chimpanzees and humans. Jacobs et al. demonstrate that ZNF91 underwent rapid evolution, including structural changes 8 to 12 million years ago which allowed it to

repress SVAs. This would coincide with the expansion of SVA subfamilies C and D, around 10.88 and 9.55 million years ago, respectively. While these SVA families have continued to move across the genome, it may be possible that the SVAs identified at this locus played a role in the rapid evolutionary changes alluded to by Jacobs (Jacobs, 2014; Wang, 2005).

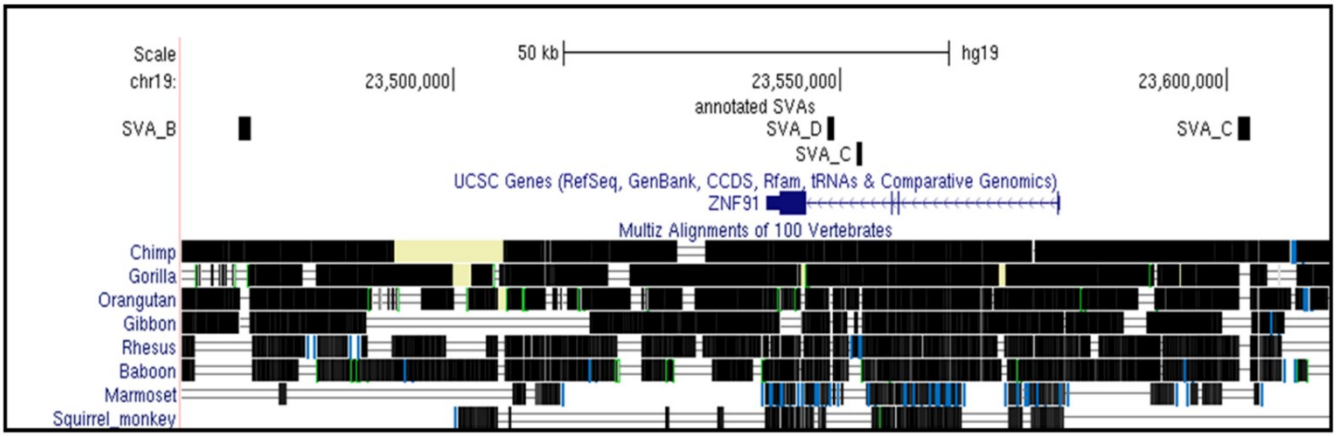


Figure 6.4 Visualisation of SVAs at the ZNF91 gene locus.

Visualisation of the chr19:23,464,801-23,613,281 region encompassing ZNF91. The third intron of ZNF91 contains one SVA C and one SVA D, with a further SVA C approximately 24 kb upstream of the ZNF91 transcriptional start site, and an SVA B around 67 kb downstream of the 3' UTR. From these locations, all four SVAs have the potential to modulate ZNF91 expression.

6.3.3 Reference LINE-1 subfamilies L1HS, L1PA2, and L1PA3 collectively cluster on the X chromosome

We next applied the above methods to the three most recent LINE-1 subfamilies, L1PA3, L1PA2, and L1HS. As there are approximately 900,000 LINE-1 elements in the human genome according to annotations on the UCSC Genome Browser (hg19), we narrowed this list down based on the evolutionary age of the LINE-1 subfamilies. The L1PA3 and L1PA2 subfamilies are present in both chimpanzees and humans, and are 12.5 and 7.6 million years old, respectively. L1HS is specific to humans and is believed to have evolved 3.1 million years ago (Khan et al. 2006). While L1PA3 is around 1 million years younger than the oldest SVA A (13.6 million years old), L1PA4 and other subfamilies were 18 million years old and older, and were therefore not considered as appropriate comparisons for this work. The evolutionary ages of the different retrotransposon subfamilies in this study are outlined in Table 6.3.

Co-ordinates for all L1HS, L1PA2, and L1PA3 subfamilies were downloaded through the UCSC table browser. Some elements in this data set were 'split' and incorrectly denoted as two elements, therefore co-ordinates of overlapping elements were combined to represent single elements. After correction, we found 1307 L1HS insertions, 4146 L1PA2 insertions, and 8904 L1PA3 insertions, totalling 14,357 LINE-1 elements in this analysis. The corrected co-ordinates of all LINE-1 insertions used in this analysis are available as .BED files in Supplementary Files 6.1, 6.2, and 6.3.

Mapping the 14,357 LINE-1 elements by Mb across the genome firstly suggested the opposite distribution pattern to SVAs. Where the number of SVAs per Mb was predominantly found to be increased at genic regions, comparing combined L1HS, L1PA2, and L1PA3 data with the number of transcripts per Mb demonstrated that the number of L1 insertions decreased with increasing transcript number.

Element subfamily	Age (Myrs)	Humans	Chimpanzees
SVA A	13.6	✓	✓
SVA B	11.6	✓	✓
SVA C	10.9	✓	✓
SVA D	9.6	✓	✓/✗
SVA E	3.5	✓	✗
SVA F	3.2	✓	✗
SVA F1	< 3.2	✓	✗
L1PA3	12.5	✓	✓
L1PA2	7.6	✓	✓
L1HS	3.1	✓	✗

Table 6.3 Names and evolutionary ages of retrotransposon subfamilies in this analysis.

All SVA subfamilies (A-F1) were used in this analysis, along with the most recent three LINE-1 subfamilies, with their names and evolutionary age detailed in this table. SVA subfamilies A-C and LINE-1 subfamilies L1PA2 and L1PA3 are present in species from chimpanzees to humans, whereas SVA subfamilies E-F1 and the LINE-1 subfamily L1HS are unique to humans. The 9.6 Myr old SVA D subfamily can be present in both chimpanzees and humans, and their conservation between these species varies for each SVA D element.

The genome-wide correlation coefficient for LINE-1 and transcript number was -0.31, demonstrating a significant negative correlation. This negative correlation held true for all individual chromosomes except chromosome Y, which showed no correlation (correlation coefficient = 0.10) (Table 6.1). However, as previously identified, sequencing data across much of the Y chromosome is missing and it therefore cannot be determined whether this would be an accurate reflection of the complete chromosome. By transforming the LINE-1 distribution data and plotting it as a heatmap, we see that reference LINE-1 elements from the most recent three subfamilies preferentially cluster on the X chromosome (Figure 6.5), with 9.2% of all elements in this analysis residing on the X chromosome. In particular, we find that more recent LINE-1 elements cluster at the larger regions chrX:74,000,001-76,000,000 and adjacent to the centromere on both the q and p arms at chrX:63,000,001-67,000,000 and chrX:56,000,001-57,000,000, with 57, 84, and 24 LINE-1 elements across two, four, and one Mb, respectively (Figure 6.6a, b, c).

Of note, the chrX:74,000,001-76,000,000 locus contains two genes from the MAGE family (melanoma antigen genes), MAGEE1 and MAGEE2. Similar to the PRAMEF genes, Type I MAGE genes (MAGE-A, MAGE-C) are ‘cancer/testis antigens’ (CTAs), however, the Type II MAGEE1 and MAGEE2 genes are ubiquitously expressed (Weon and Potts 2015, Li, Hughes and Wevrick 2015), with the highest expression in the brain according to RNA-seq from the Geneotype-Tissue Expression (GTEx) project (data not shown).

A further region of the X chromosome which stands out in relation to LINE-1 clustering is the chrX:49,000,001-50,000,000 locus, in which the LINE-1 elements cluster around the GAGE family gene locus (Figure 6.6d).

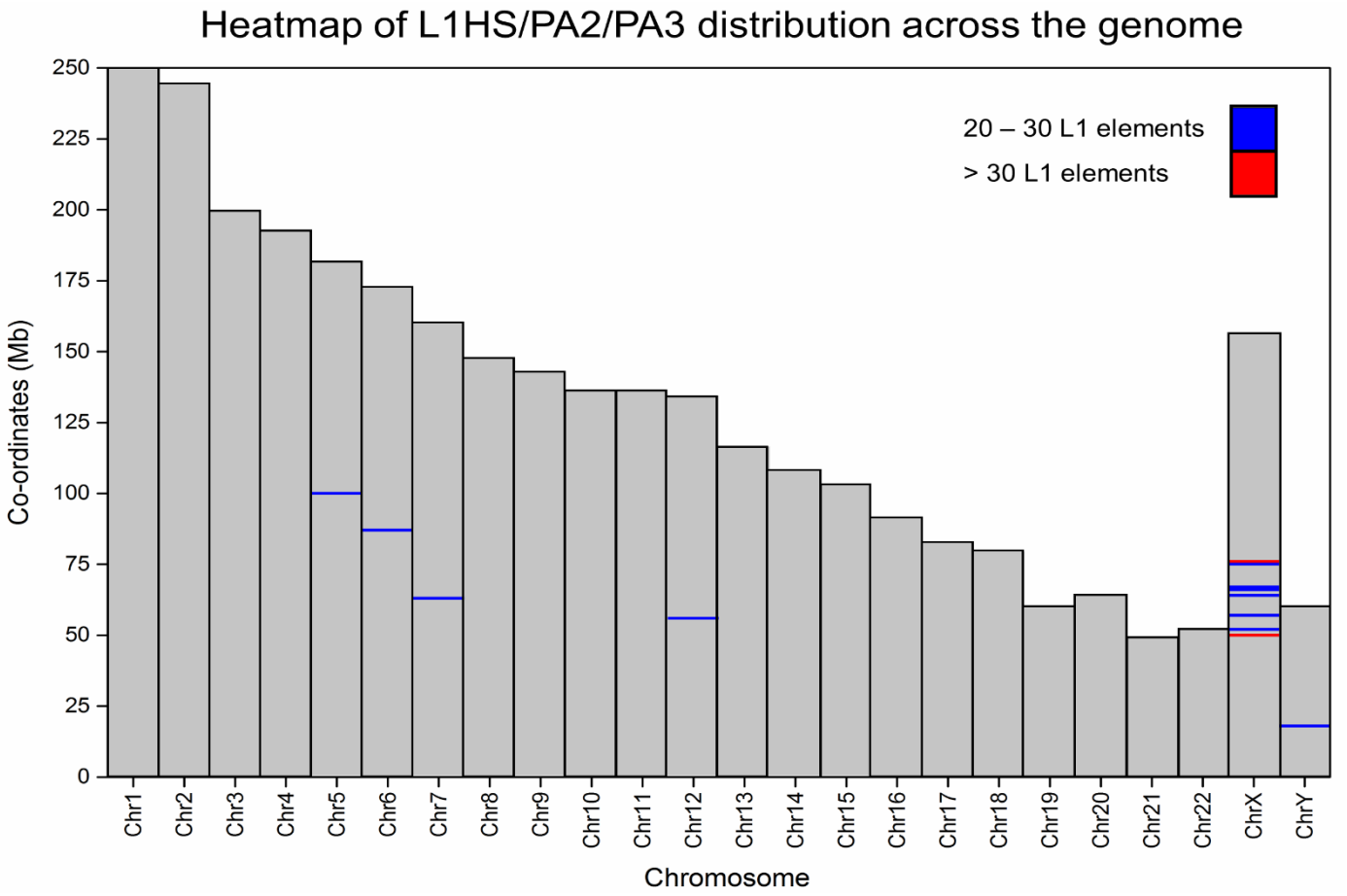


Figure 6.5 Heatmap of recent LINE-1 subfamily distribution across the human genome.

LINE-1 subfamilies L1HS, L1PA2, and L1PA3 cluster on the X chromosome, with 9.2% of elements in this study found on the X chromosome. Blue bars indicate regions which contain 20 to 30 recent LINE-1 subfamily insertions (0.14 to 0.21% of the total), while red bars indicate regions with over 30 insertions (over 0.21% of the total). The two most LINE-1 dense regions are found at chrX:75,000,000-76,000,000 and chrX:49,000,000-50,000,000 with 35 and 31 LINE-1 insertions, respectively. We also note 20 or more LINE-1 insertions at a small number of regions on chromosome 5, 6, 7, 12, and Y.

Figure 6.6 Regions of LINE-1 clustering on the X chromosome.

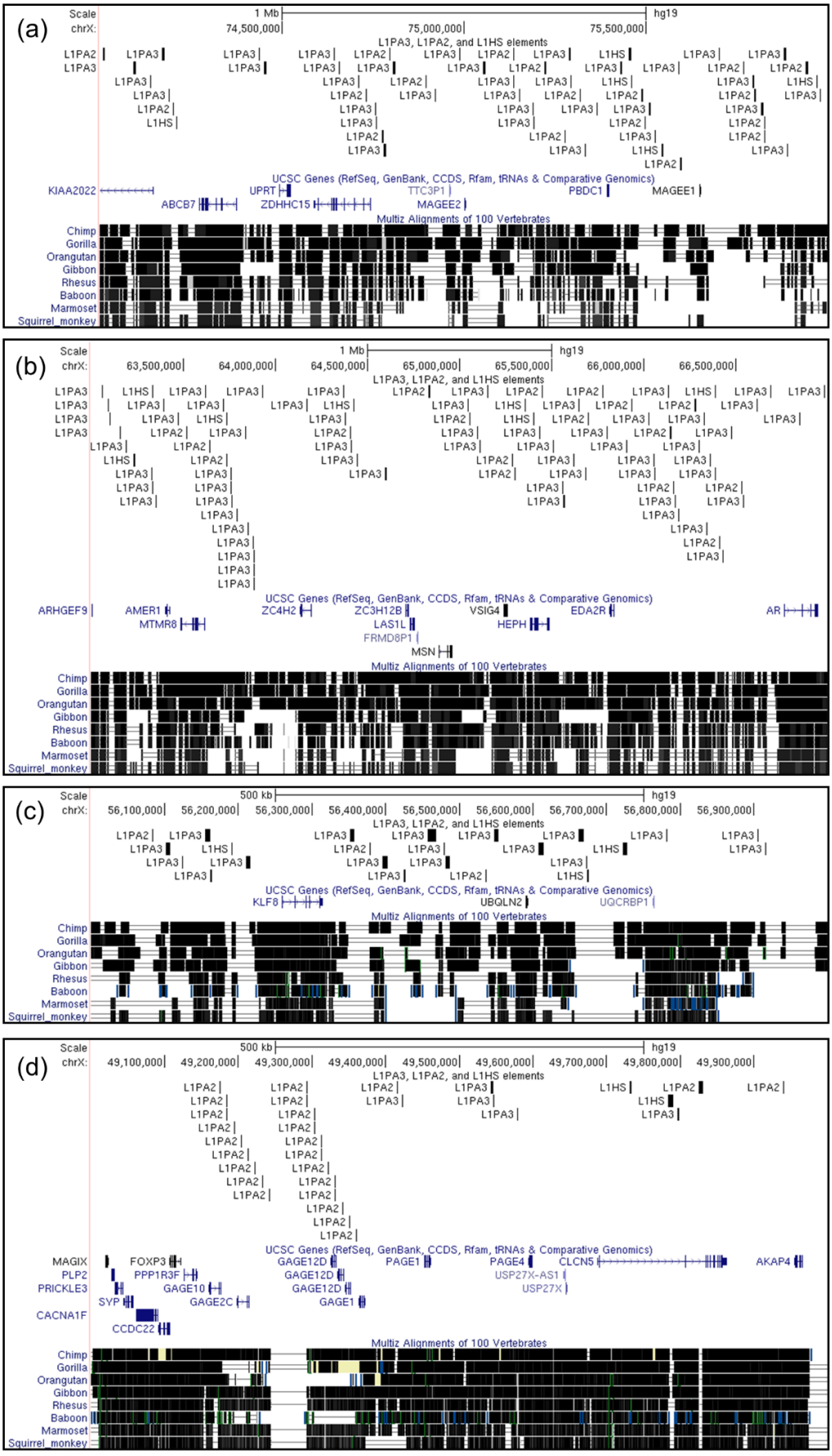


Figure 6.6 Regions of LINE-1 clustering on the X chromosome.

Regions with the highest amount of recent LINE-1 subfamily clustering include the loci at chrX:74,000,001-76,000,000 (a), with 57 elements across a region of two Mb, and the regions adjacent to the centromere on both the q and p arms, chrX:63,000,001-67,000,000 and chrX:56,000,001-57,000,000 (b, c), with 84 and 24 LINE-1 elements across four and one Mb, respectively. The locus at chrX:49,000,001-50,000,000 is also over-represented for LINE-1 elements, which predominantly cluster around the GAGE gene family at this region. This may be of interest, as the GAGE gene family are cancer/testis antigens, as are the PRAMEF genes within an SVA cluster locus, and the MAGE genes, two of which reside within the most LINE-1 dense region in (a).

The GAGE gene family is thought to be primate-specific, and has recently duplicated and evolved after the split between humans and chimpanzees (Liu, Zhu and Zhu 2008). Again, similar to the PRAMEF and MAGE gene families, the GAGE genes are CTA genes, expressed in the testis and overexpressed in a range of cancers (Chen et al. 2011b, De Backer et al. 1999, Kong et al. 2004, Zhang et al. 2010a).

Given the clear clustering of LINE-1 elements at specific regions on the X chromosome, we next separated the elements by subfamily to assess whether all three recent LINE-1 subfamilies were so clearly over-represented at the X chromosome when compared to the rest of the genome. We first compared the number of LINE-1 insertions per chromosome, divided by chromosome size in base pairs (hg19), in order to account for chromosome size. Secondly, as the size of LINE-1 elements in this analysis was highly variable, from small fragments in the tens of base pairs to full length elements upwards of 6 kb, we also compared the percentage of each chromosome that was made up of LINE-1 sequence. This would allow a more precise method of determining how much new genetic material had been added to each chromosome through the expansion of these subfamilies, based on genomic size rather than number of insertions.

Beginning with the older L1PA3 subfamily, we see a 1.5-fold increase of L1PA3 insertions on the X chromosome when compared to the next most L1PA3-dense chromosome, chromosome 4, with the X chromosome being the most L1PA3-dense based on the number of insertions. Similarly, we see that the X chromosome gained 1.5-fold more of its total size than chromosome 4, with each chromosome gaining 1.24 and 0.83% of its DNA from L1PA3 insertions, respectively (Figure 6.7a).

Figure 6.7 Clustering of recent LINE-1 subfamilies on the X chromosome appears to decrease with younger evolutionary age.

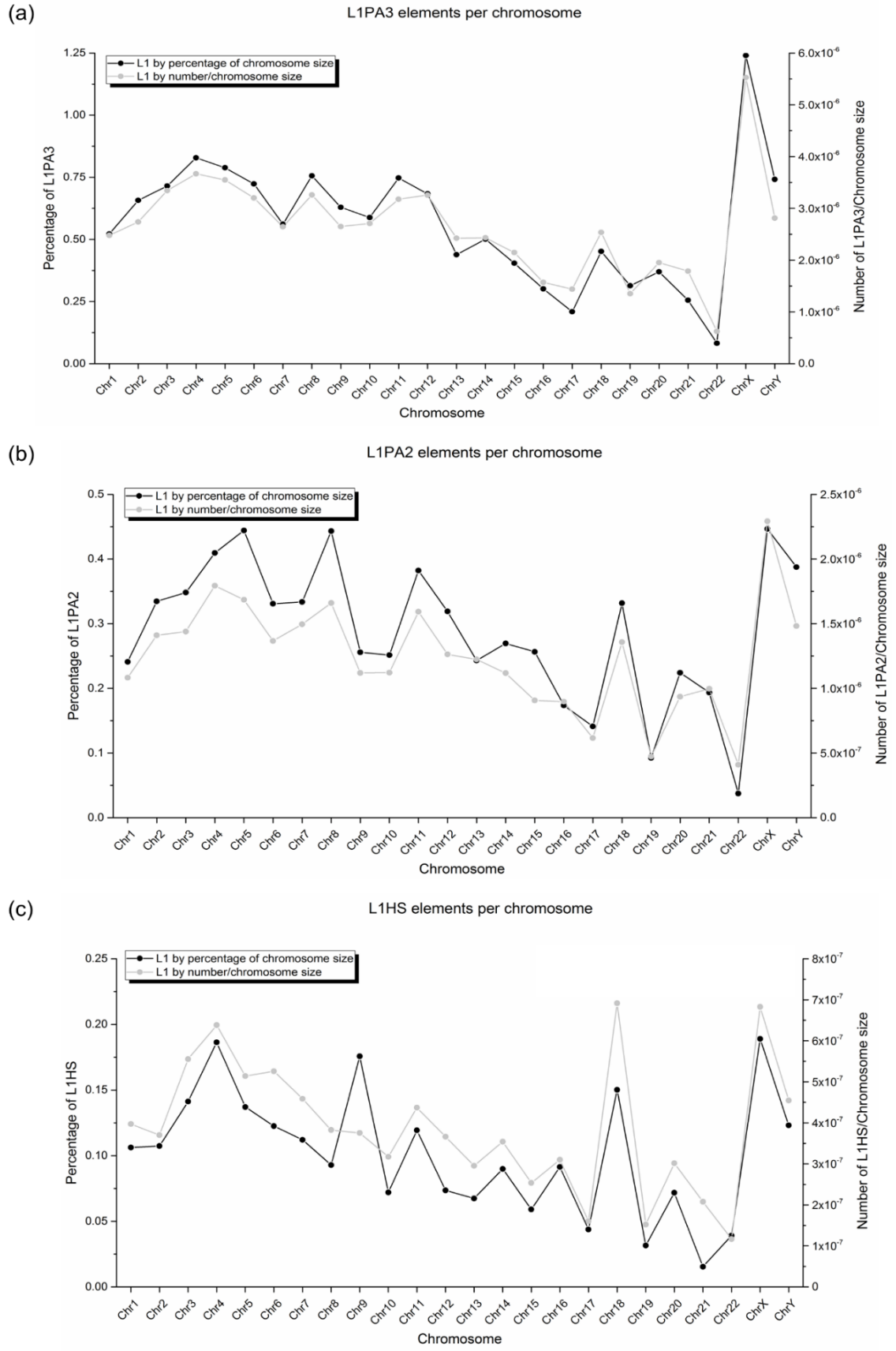


Figure 6.7 Clustering of recent LINE-1 subfamilies on the X chromosome appears to decrease with younger evolutionary age.

(a) The L1PA3 subfamily (12.5 Myrs old) consists of 8904 elements, which are preferentially found on the X chromosome. L1PA3 insertion on the X chromosome shows a clear bias over other chromosomes, which is evident when measuring both the percentage of each chromosome sequence made up of L1PA3 (black line), and based on the number of L1PA3 insertions divided by chromosome size (grey line). For example, chromosome X has 1.5-fold more insertions than the next most L1PA3-dense chromosome 4.

(b) The L1PA2 subfamily (7.6 Myrs old) contains 4146 elements. Assessing clustering by number of insertions divided by chromosome size demonstrates that the X chromosome contains the highest number of L1PA2 elements, however, when calculating how much of each chromosome is made up of L1PA2 sequence, chromosomes 4, 8, and X have all gained approximately 0.44% of their sequence from L1PA2 insertions.

(c) The human specific L1HS subfamily (3.1 Myrs old) contains 1307 elements, with the highest number of copies per chromosome size found on chromosomes 18 and X. However, chromosomes 4, 9, and X have gained the highest percentage of their sequence directly from L1HS insertion, with approximately 0.19% of their overall sequence being accounted for by L1HS elements.

For the L1PA2 subfamily, the X chromosome remains the most LINE-1 dense of all 24 chromosomes in terms of number of insertions, however this trend decreases from L1PA3, with only a 1.3-fold increase over the chromosome 4, which again remains the second most L1PA2-dense chromosome based on insertions counts (Figure 6.7b). However, when measuring the percentage of each chromosome that is accounted for by L1PA2 sequence, we find that chromosomes 5, 8, and X all gained approximately 0.44% of their total size through L1PA2 insertions. Finally, for the L1HS subfamily, we find chromosome 18 to be the site of the most L1HS insertions, overtaking both chromosome X and 4, which are second and third most L1HS-dense (Figure 6.7c). However, we find that, based on the percentage of chromosome size accounted for by these insertions, the X chromosome narrowly tops the list, gaining 0.189% of its size from L1HS insertions, closely followed by chromosome 4 which gained 0.186%.

From this, we can see that while the number of insertions at the X chromosome appears to be one of the highest across all 24 chromosomes even when corrected for size, there are numerous other chromosomes that have gained a similar percentage of their total genomic size from the insertion of these elements. While the literature demonstrates that evolutionary younger LINE-1 subfamilies cluster more strongly on the X chromosome than the older LINE-1 subfamilies (Bailey et al. 2000), the differing results across the LINE-1 subfamilies studied in this thesis may suggest that the number of recent LINE-1 subfamily insertions on the X chromosome has been slowly decreasing over recent evolutionary history through higher primates and humans (Figure 6.7).

6.3.4 LINE-1 subfamilies are over-represented at genes involved in brain-related pathways

In order to characterise the genes that have potentially been re-modelled through LINE-1 insertion across the last 12.5 million years, we used the UCSC Genome Browser to download the co-ordinates of all transcripts in the 'known gene' track for the hg19 genome build. We then added 5 kb upstream of each transcriptional start site in order to represent a stringently defined promoter region. The co-ordinates of all 78,807 transcripts plus 5 kb upstream were uploaded to the UCSC Genome Browser as a custom track (Supplementary File 6.4), as were three separate files containing the co-ordinates of all L1HS, L1PA2, and L1PA3 elements (Supplementary Files 6.1, 6.2, 6.3). Using the UCSC table browser tool, we overlaid the transcript co-ordinates with each of the three LINE-1 subtype co-ordinates to generate gene lists in which one or more transcripts contained a LINE-1 element either within or up to 5 kb upstream of their sequence.

This resulted in the following three gene lists with LINE-1 insertions either within or up to 5 kb upstream of one or more transcripts of each gene in the list:

L1PA3 – 2203 genes
L1PA2 – 1275 genes
L1HS – 463 genes

In order to gain insight into which particular processes and pathways, if any, had been targets of LINE-1 mediated evolution between higher primates and humans, we next ran each gene set through the Ma'ayan Lab Enrichr tool.

From 5192 available terms in the Gene Ontology biological processes data set, the gene list for the 2203 genes with one or more L1PA3 insertions (Supplementary Data

6.2), was found to be enriched almost exclusively for roles in cell adhesion and cyclic AMP (cAMP) related processes, with statistically significant Benjamini-Hochberg adjusted Fisher's exact p-values for:

- cAMP catabolic processes — adjusted $p = 1.32 \times 10^{-3}$
- Cyclic nucleotide catabolic processes — adjusted $p = 4.11 \times 10^{-4}$
- cAMP metabolic processes — adjusted $p = 2.53 \times 10^{-2}$
- Cyclic nucleotide metabolic processes — adjusted $p = 1.88 \times 10^{-2}$
- Cell-cell adhesion — adjusted $p = 1.20 \times 10^{-3}$
- Cell adhesion via membrane adhesion molecules — adjusted $p = 1.32 \times 10^{-3}$

We also noted enrichment for glutamate receptor signalling pathways and neuron recognition, with Benjamini-Hochberg adjusted Fisher's exact p-values of 7.79×10^{-3} and 5.00×10^{-2}, respectively (Figure 6.8a). Data from the 641 terms within the Gene Ontology cellular components data set presented evidence for enrichment of L1PA3-targeted genes in the brain, showing significant enrichment for the presence of genes in this list across the synapse and in making up various transporter complexes (Figure 6.8b), with Benjamini-Hochberg adjusted Fisher's exact test significance for terms including:

- Synapse part — adjusted $p = 2.34 \times 10^{-5}$
- Synaptic membrane — adjusted $p = 2.40 \times 10^{-5}$
- Pre-synaptic membrane — adjusted $p = 1.77 \times 10^{-2}$
- Post-synaptic membrane — adjusted $p = 1.80 \times 10^{-4}$
- Ion channel complex — adjusted $p = 1.25 \times 10^{-2}$
- Transmembrane transporter complex — adjusted $p = 2.60 \times 10^{-2}$
- Calcium channel complexes — adjusted $p = 2.60 \times 10^{-2}$

Figure 6.8 Genes targeted by L1PA3 insertions show enrichment for roles in cell adhesion, cAMP metabolism, and brain-related pathways.

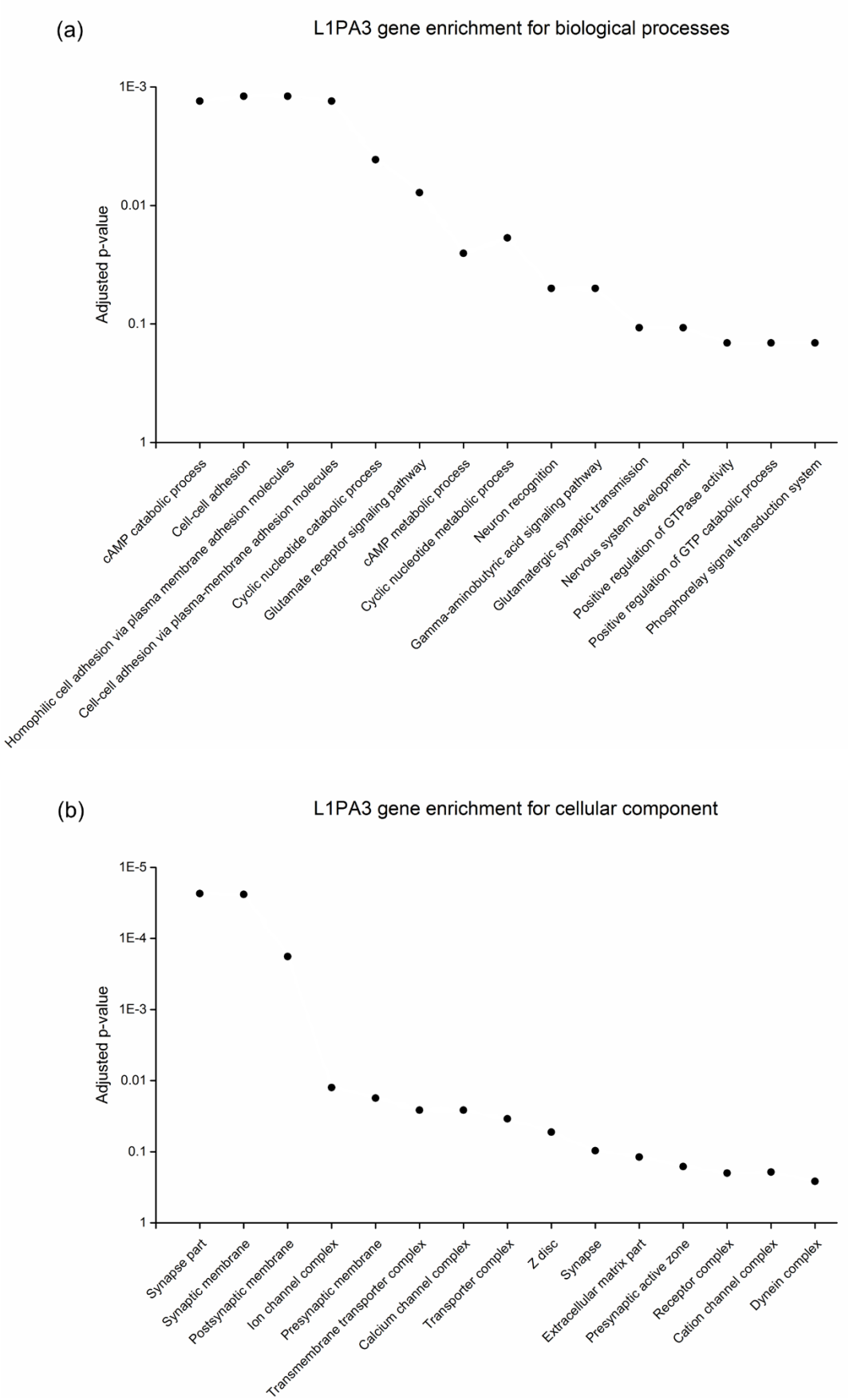


Figure 6.8 Genes targeted by L1PA3 insertions show enrichment for roles in cell adhesion, cAMP metabolism, and brain-related pathways.

(a) Biological processes terms for the 2203 genes targeted by L1PA3 insertions show enrichment for roles in cell adhesion (Benjamini-Hochberg adjusted Fisher's exact p-value = 1.20 x10^{-3} and 1.32 x10^{-3}) and cyclic AMP metabolism (Benjamini-Hochberg adjusted Fisher's exact p-value = 1.32 x10^{-3}, 4.11 x10^{-4}, 2.53 x10^{-2}, and 1.88 x10^{-2}), which may suggest roles in cell migration and signal transduction. We also see enrichment for specific brain-related processes including glutamate receptor signalling and neuron recognition (Benjamini-Hochberg adjusted Fisher's exact p-value = 7.79 x10^{-3} and 5.00 x10^{-2}).

(b) Analysis of the same gene set for cellular components showed enrichment for the localisation of these genes in the synapse and synaptic membranes (Benjamini-Hochberg adjusted Fisher's exact p-values = 2.34 x10^{-5}, 2.40 x10^{-5}, 1.80 x10^{-4}, and 1.77 x10^{-2}), as well as in channel complexes such as calcium ion transporters (Benjamini-Hochberg adjusted Fisher's exact p-values = 1.25 x10^{-2} and 2.60 x10^{-2}). This may suggest potential roles for this gene set in CNS membrane depolarisation and signalling.

Finally, enrichment analysis using data from 476 possible terms in the Mouse Genome Informatics (MGI) phenotype database suggested that mice with knockout or mutation of genes within this set were likely to display the phenotype of abnormal synaptic transmission (3.30×10^{-3}) (Figure 6.9), thus suggesting that these may be pathways modulated by retrotransposon insertion during evolution from mouse to primate.

When compared to the L1PA3 data, enrichment analysis for biological processes associated with the L1PA2 gene set began to show a clear move towards the targeting of brain-related pathways (Figure 6.10a, Supplementary Data 6.2), with strongest enrichment out of the available 5192 terms for:

- Neuron recognition — adjusted $p = 3.75 \times 10^{-3}$
- Behaviour — adjusted $p = 7.33 \times 10^{-3}$
- Synaptic transmission — adjusted $p = 9.25 \times 10^{-3}$
- Glutamate receptor signalling — adjusted $p = 1.27 \times 10^{-2}$
- Central nervous system development — adjusted $p = 1.75 \times 10^{-2}$

While significantly different from the enrichment results for the L1PA3 gene list, we saw some similarities in the enrichment for cell adhesion in genes targeted by L1PA2 insertions (cell-cell adhesion, cell-cell adhesion via plasma membrane adhesion molecules, and homophilic cell adhesion via plasma membrane adhesion molecules with Benjamini-Hochberg adjusted Fisher's exact p-values of 3.75×10^{-3}, 3.75×10^{-3}, and 1.30×10^{-2}).

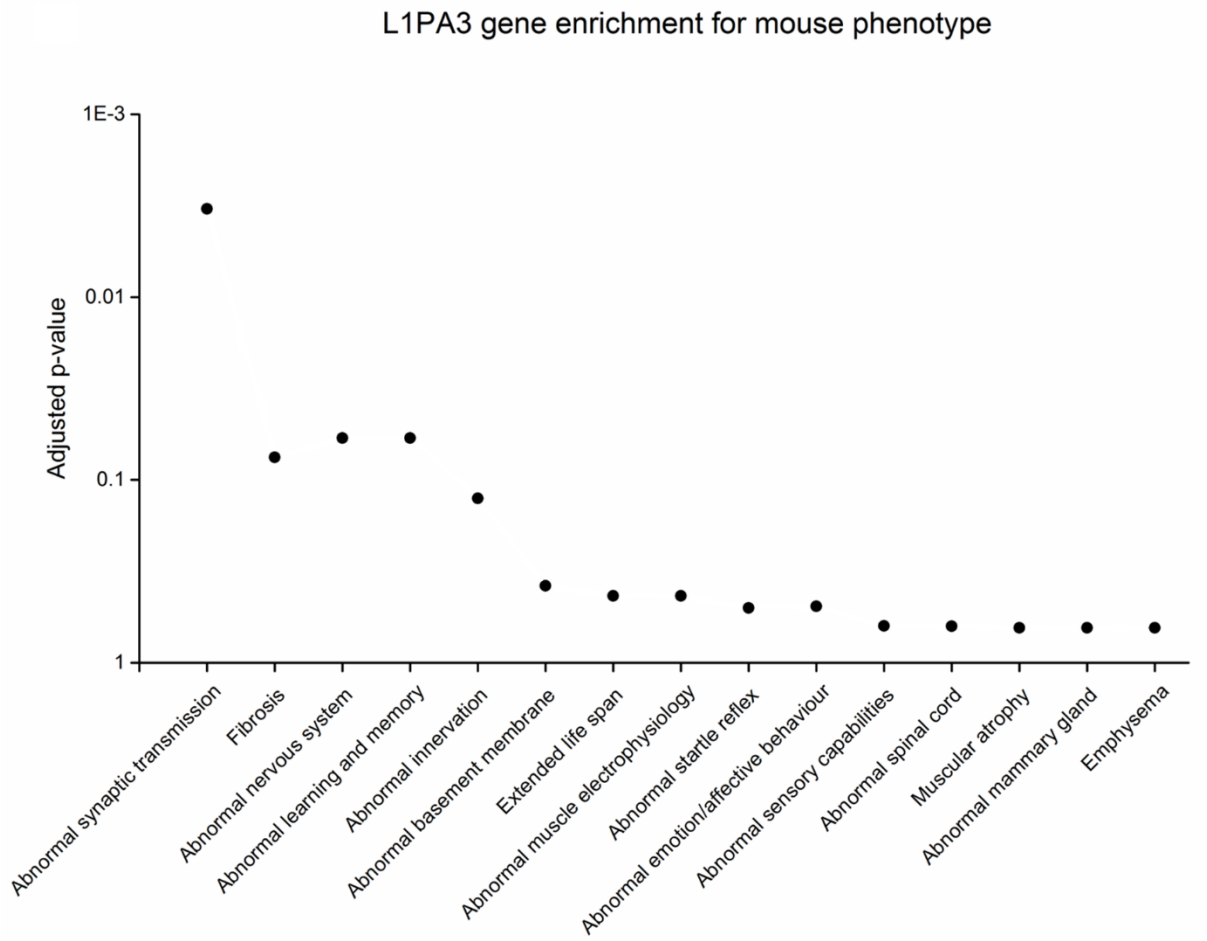


Figure 6.9 The L1PA3 target gene list is enriched for brain-related mouse phenotypes.

Enrichment analysis of L1PA3-containing genes using the Mouse Genome Informatics (MGI) knockout and mutation phenotype data showed significant enrichment only for abnormal synaptic transmission (Benjamini-Hochberg adjusted Fisher's exact p-value = 3.30 x10^{-3}). Multiple other CNS-related phenotypes such as abnormal nervous system and abnormal learning and memory were enriched at the Fisher's exact p-value level, but lost significance after Benjamini-Hochberg adjustment.

Figure 6.10 Genes targeted by L1PA2 insertions show enrichment for neuron recognition, behaviour, and synaptic transmission.

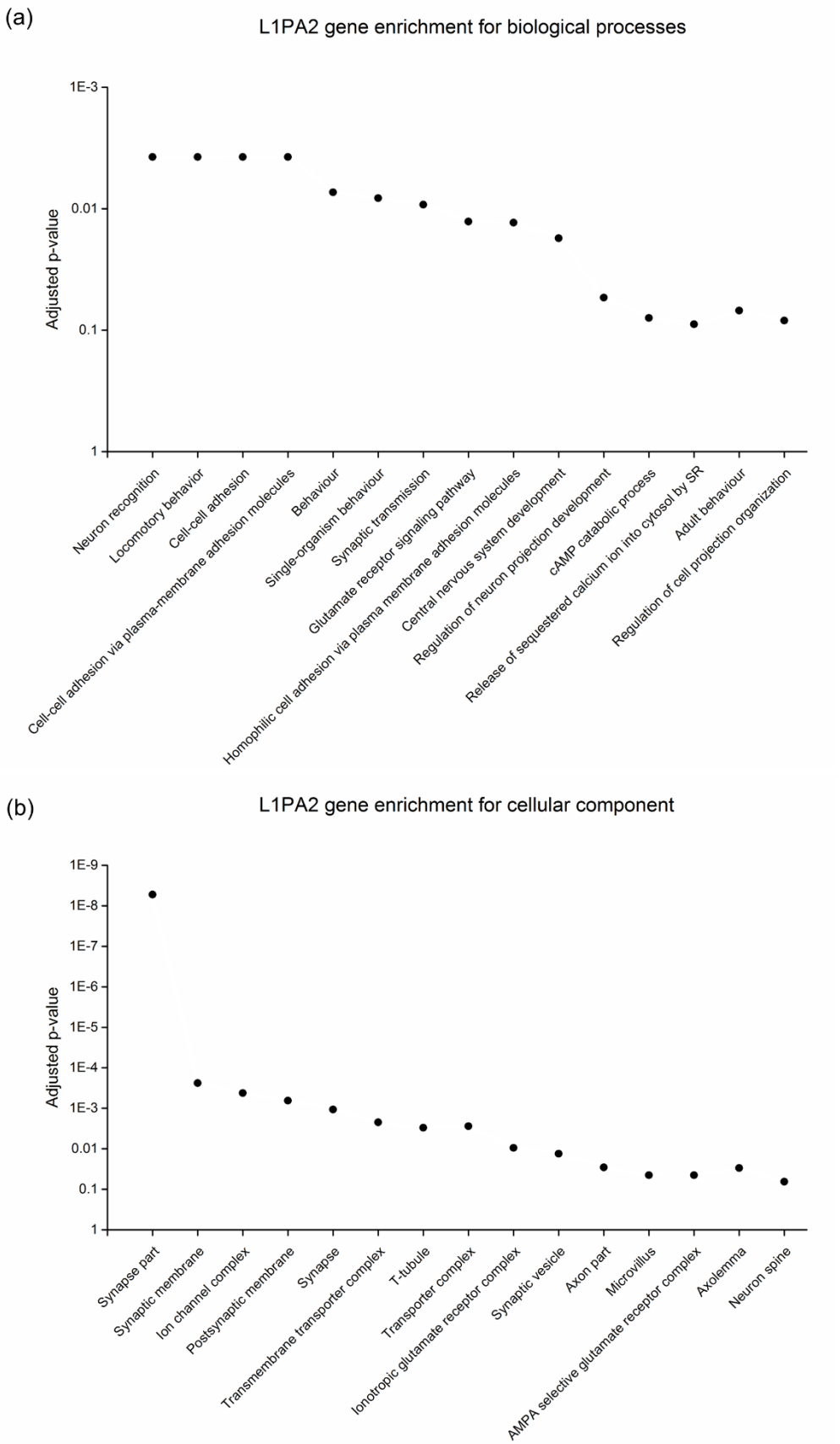


Figure 6.10 Genes targeted by L1PA2 insertions show enrichment for neuron recognition, behaviour, and synaptic transmission.

(a) Analysis of the 1275 genes with one or more L1PA2 insertion either within their sequence or up to 5 kb upstream showed enrichment for neuron recognition, behaviour, synaptic transmission, glutamate receptor signalling, and central nervous system development (Benjamini-Hochberg adjusted Fisher's exact p-values = 3.75 x10^{-3}, 7.33 x10^{-3}, 9.25 x10^{-3}, 1.27 x10^{-2}, and 1.75 x10^{-2}), though cell adhesion remained an enriched process (Benjamini-Hochberg adjusted Fisher's exact p-values = 3.75 x10^{-3} and 1.30 x10^{-2}).

(b) Enrichment for cellular localisation of this gene list supports a role for this gene set within the brain, with significance for localisation in the synapse part, synaptic membranes, and synaptic vesicles (Benjamini-Hochberg adjusted Fisher's exact p-values = 5.31 x10^{-9}, 2.38 x10^{-4}, 1.07 x10^{-3}, and 1.32 x10^{-2}), as well as ion channel and transporter complexes (4.20 x10^{-3} and 2.22 x10^{-3}), specifically the ionotropic and the AMPA-selective glutamate receptor complexes (9.47 x10^{-3} and 4.46 x10^{-2}). This highlights a role for L1PA2 targeting specific glutamate signalling pathways in the brain, as well as continuing to target more general processes associated with transmembrane transport within the synapse.

Similarly, enrichment for location in the cell provided strong evidence for this gene set's involvement in the brain (Figure 6.10b), with Benjamini-Hochberg adjusted Fisher's exact test significance for enrichment for synapse- and channel-related terms including:

- Synapse part — adjusted p = 5.31×10^{-9}
- Synaptic membrane — adjusted p = 2.38×10^{-4}
- Post-synaptic membrane — adjusted p = 6.47×10^{-4}
- Synapse — adjusted p = 1.07×10^{-3}
- Synaptic vesicle — adjusted p = 1.32×10^{-2}
- Ion channels — adjusted p = 4.20×10^{-4}
- Transmembrane transporters — adjusted p = 2.22×10^{-3}
- Ionotropic glutamate receptors — adjusted p = 9.47×10^{-3}

Further, analysis of the L1PA2 gene list using the 476 terms in the mouse phenotype data set demonstrated significant enrichment for abnormal synaptic function and abnormal learning and memory (Benjamini-Hochberg adjusted Fisher's exact p-values = 2.81×10^{-4} and 2.18×10^{-3}) (Figure 6.11).

For the 463 genes containing one or more L1HS insertions either within or up to 5 kb upstream of one or more transcripts (Supplementary Data 6.2), we found no significant enrichment for biological processes using the Benjamini-Hochberg corrected Fisher's exact test (Figure 6.12a).

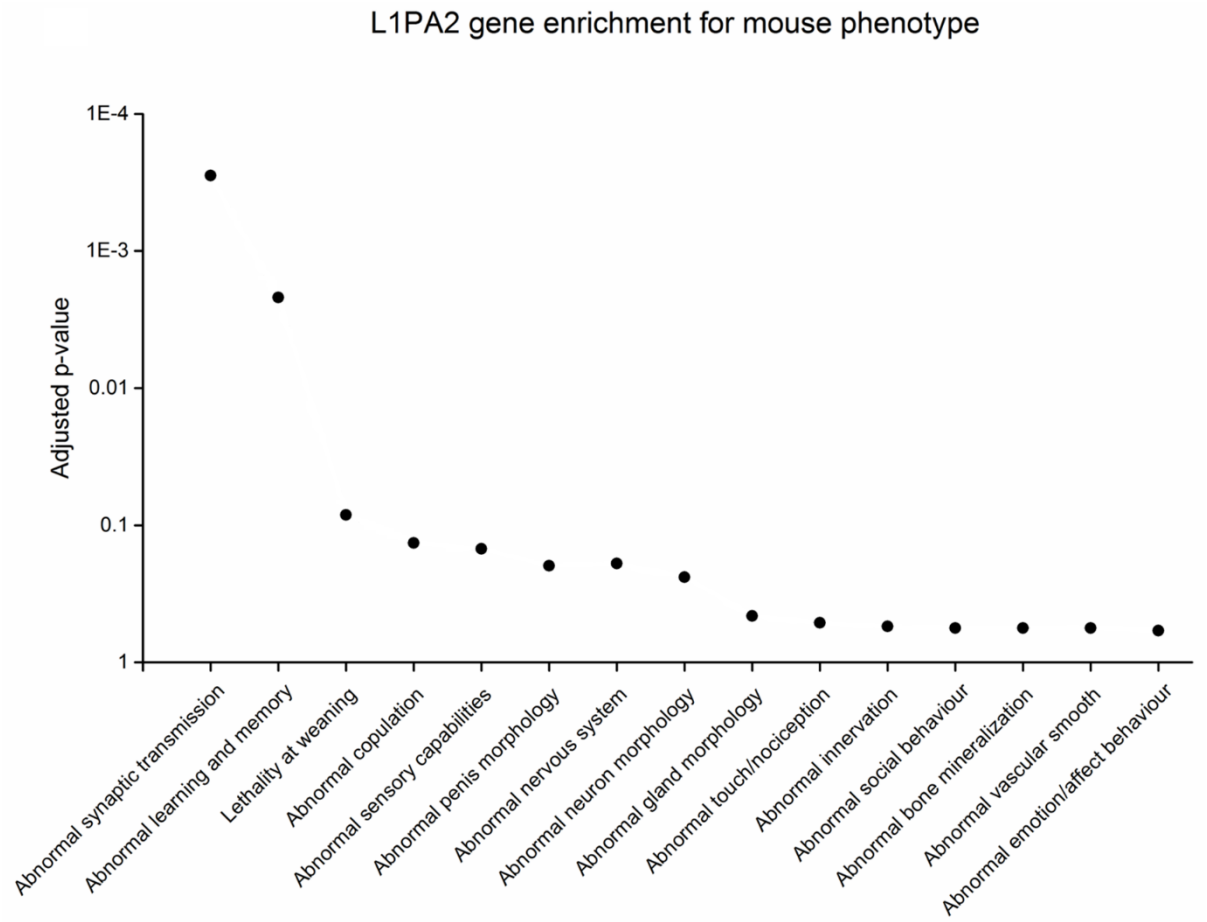


Figure 6.11 L1PA2 target gene list is enriched for brain-related mouse phenotypes.

Enrichment analysis using genes with one or more L1PA2 insertions demonstrated enrichment for mouse phenotypes involving abnormal synaptic transmission and abnormal learning and memory (Benjamini-Hochberg adjusted Fisher's exact p-values = 2.82 x10^{-4} and 2.18 x10^{-3}). These data support the enriched biological processes and cellular component data to suggest significant LINE-1 targeting of brain-related pathways and processes, with the potential to modulate CNS phenotype.

Figure 6.12 Genes with L1HS insertion show a trend towards enrichment for a wide range of CNS-related processes.

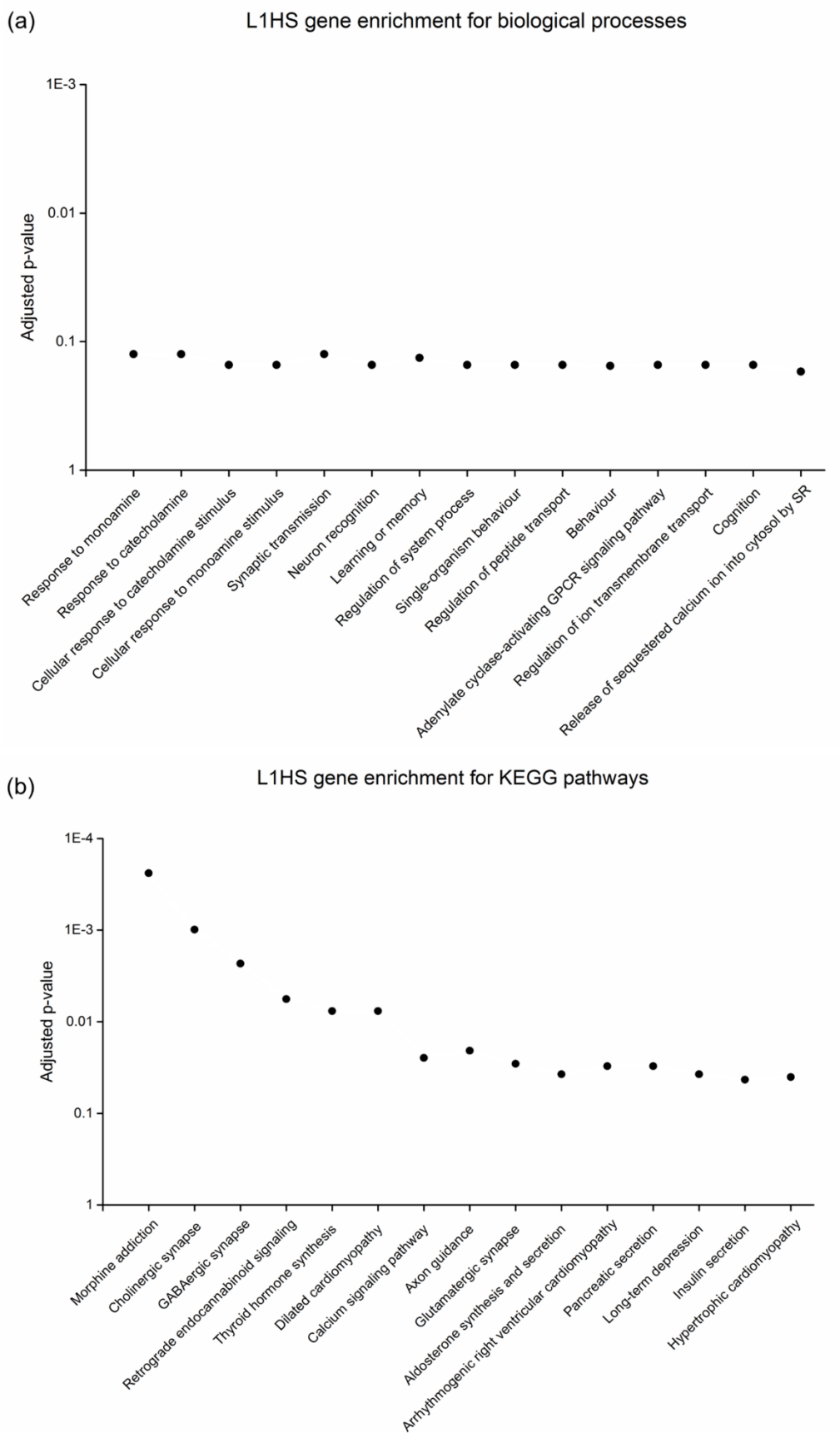


Figure 6.12 Genes with L1HS insertion show a trend towards enrichment for a wide range of CNS-related processes.

(a) Enrichment analysis for the L1HS gene set in the Gene Ontology biological processes database showed significance at the p-value level for numerous brain-related processes including response to monoamines and catecholamine (Benjamini-Hochberg adjusted Fisher's exact p-values = 1.55 x10^{-4} and 1.02 x10^{-3}), synaptic transmission, neuron recognition, learning or memory, behaviour, and cognition (Benjamini-Hochberg adjusted Fisher's exact p-values = 8.29 x10^{-5}, 9.04 x10^{-4}, 2.21 x10^{-4}, 1.28 x10^{-3}, and 8.71 x10^{-4}) (Supplementary Data 6.2). However, these terms did not retain significance after Benjamini-Hochberg adjustment, suggesting only a trend towards involvement in these processes.

(b) Analysis of this gene set using the KEGG database demonstrates significance for multiple brain-related pathways including morphine addiction, cholinergic synapse, GABAergic synapse, glutamatergic synapse, retrograde endocannabinoid signalling, axon guidance, and others (Benjamini-Hochberg adjusted Fisher's exact p-value = 2.39 x10^{-4}, 9.85 x10^{-4}, 2.32 x10^{-3}, 2.87 x10^{-2}, 5.67 x10^{-3}, and 2.07 x10^{-2}). We also note the significance of heart-related pathways such as dilated cardiomyopathy and arrhythmogenic cardiomyopathy (Benjamini-Hochberg adjusted Fisher's exact p-value = 7.64 x10^{-3} and 3.05 x10^{-2}), which may share some overlap with the brain through the similarly enriched calcium signalling pathway (Benjamini-Hochberg adjusted Fisher's exact p-value = 2.48 x10^{-2}).

However, enriched KEGG pathways again hinted at the targeting of brain-related pathways (Figure 6.12b) including those involved in:

- Morphine addiction — adjusted p = 2.39×10^{-4}
- Cholinergic synapse — adjusted p = 9.85×10^{-4}
- GABAergic synapse — adjusted p = 2.32×10^{-3}
- Retrograde endocannabinoid signalling — adjusted p = 5.67×10^{-3}
- Calcium signalling — adjusted p = 2.48×10^{-2}
- Axon guidance — adjusted p = 2.07×10^{-2}
- Glutamatergic synapse — adjusted p = 2.87×10^{-2}
- Cerebellar long-term depression — adjusted p = 3.74×10^{-2}
- Nicotine addiction — adjusted p = 3.53×10^{-2}

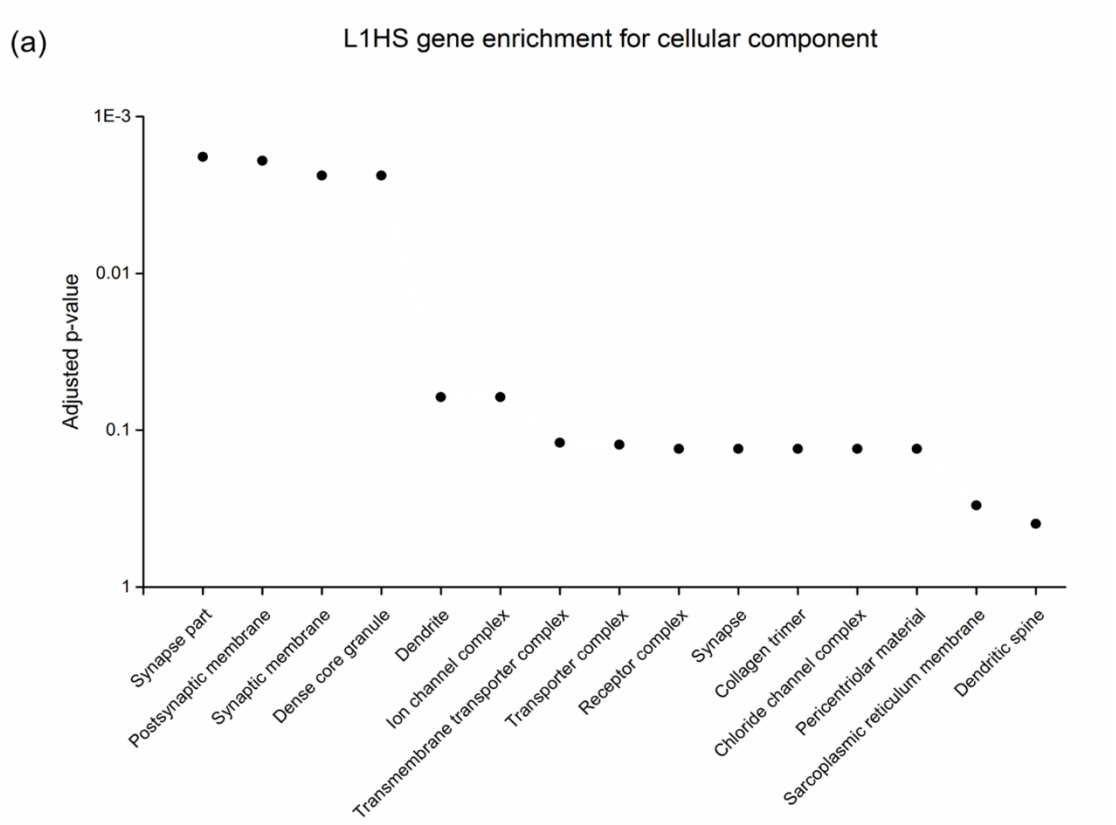
Again, enrichment for cellular components confirmed that genes with L1HS insertions were enriched for roles in the brain (Figure 6.13a), with significance for terms including:

- Synapse part — adjusted p = 1.81×10^{-3}
- Post-synaptic membrane — adjusted p = 1.92×10^{-3}
- Synaptic membrane — adjusted p = 2.38×10^{-3}
- Dense core granule — adjusted p = 2.38×10^{-3}

All p-values above are derived from the Fisher's exact test and adjusted using the Benjamini-Hochberg procedure. Similarly, we saw enrichment for abnormal synaptic transmission and abnormal learning and memory (Benjamini-Hochberg adjusted Fisher's exact p-value = 3.38×10^{-4} and 1.28×10^{-2}) in relation to mouse phenotypes involving knockout or mutation of genes in the L1HS input list (Figure 6.13b).

This data demonstrated that recently evolved LINE-1 subfamilies target genes involved in brain-related processes, and may suggest a significant shift towards the targeting of brain-related genes around the time approaching the split between chimpanzees and humans with the evolution of the L1PA2 subfamily.

Figure 6.13 Genes targeted by L1HS insertions show enrichment primarily for expression in the synapse and for CNS-related mouse phenotypes.



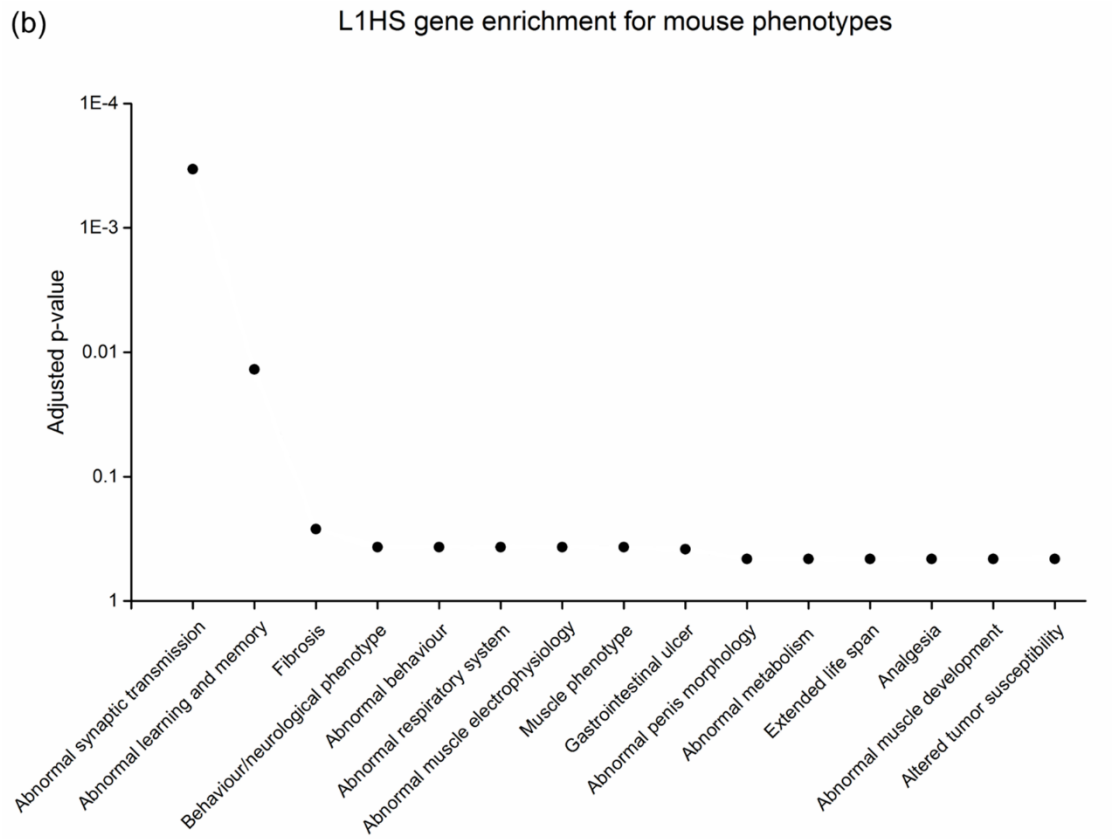


Figure 6.13 Genes targeted by L1HS insertions show enrichment primarily for expression in the synapse and for CNS-related mouse phenotypes.

(a) Enrichment analysis for the L1HS gene set in the cellular components data set supported a role for these genes acting within the brain, with enrichment for localisation in the synapse and synaptic membranes (Benjamini-Hochberg adjusted Fisher's exact p-value = 1.80 x10^{-3}, 1.92 x10^{-3} and 2.38 x10^{-3}), as well as in the dense core granule secretory organelle (Benjamini-Hochberg adjusted Fisher's exact p-value = 2.38 x10^{-3}).

(b) Enrichment for genes containing one or more L1HS elements using the MGI mouse phenotype data set further suggested roles for this gene set in the brain, with enrichment for abnormal synaptic transmission and abnormal learning and memory phenotypes in mouse models (Benjamini-Hochberg adjusted Fisher's exact p-values = 3.38 x10^{-4} and 1.38 x10^{-2}).

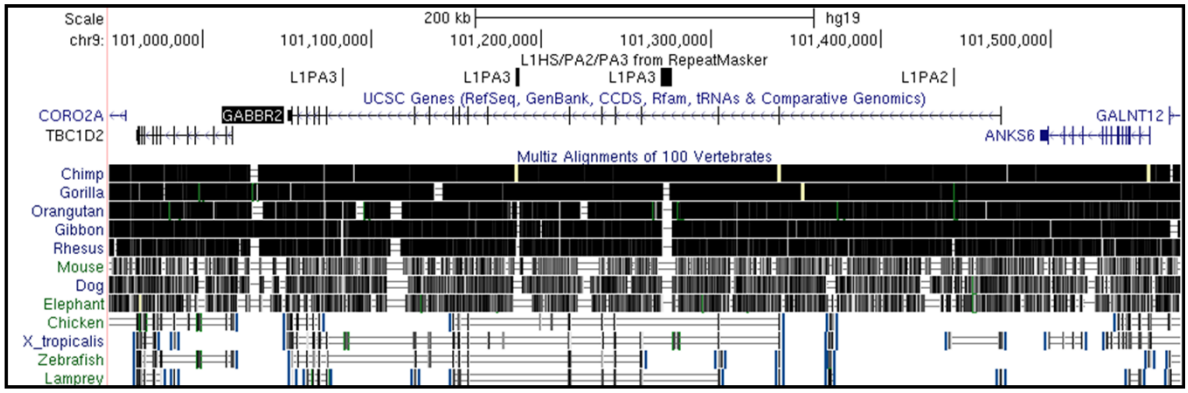
When considering the three LINE-1 subfamilies collectively, we found that the GABA and glutamate gene families have been significant targets of recent LINE-1 insertion. Of the 28,436 genes annotated on the UCSC 'known gene' track for the hg19 genome build, we found that 3039 had an L1HS, L1PA2, or L1PA3 insertion either within or up to 5 kb upstream of one or more of their transcripts. This gave an average of 10.7% of genes which contain one or more recent LINE-1 subfamily insertions. In contrast to this, of the 21 GABA receptor genes recognised by the HUGO Gene Nomenclature Committee, 10 (47.6%) had one or more recent LINE-1 insertion.

Similarly, 15 of the 26 glutamate receptor genes (57.7%) contained one or more L1HS, L1PA2, or L1PA3 insertion either in the gene or within 5 kb upstream (Figure 6.14). This constituted a 344.86% and 439.3% increase in recent LINE-1 subfamily insertion at the GABA and glutamate gene families, respectively, when compared to the total gene average. For comparison, we found three GABA genes with an SVA within or up to 5 kb upstream (GABRA2, GABRG3, and GABBR2) which constituted 14.29%, a 196.47% increase over the total gene average of 4.82% for SVA insertions. On the other hand, we found only one glutamate receptor gene, GRID1, with an SVA, constituting 3.85% of the glutamate receptor gene family.

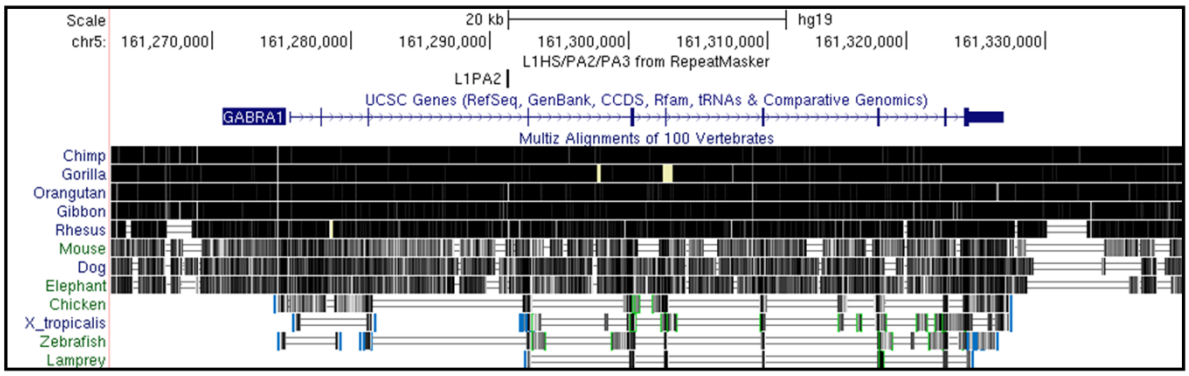
Here, we can see that, over evolutionary time, LINE-1 insertions appear to have either preferentially inserted, or been preferentially retained, increasingly around regions with involvement in brain-related processes. As we know that LINE-1 elements are able to confer regulatory properties at their site of insertion, this may have contributed in part to primate- and human-specific patterns of gene expression, regulation, and tissue distribution that may have influenced the evolution of higher cognitive abilities.

Figure 6.14 GABA and glutamate family genes with L1HS, L1PA2, and/or L1PA3 insertions.

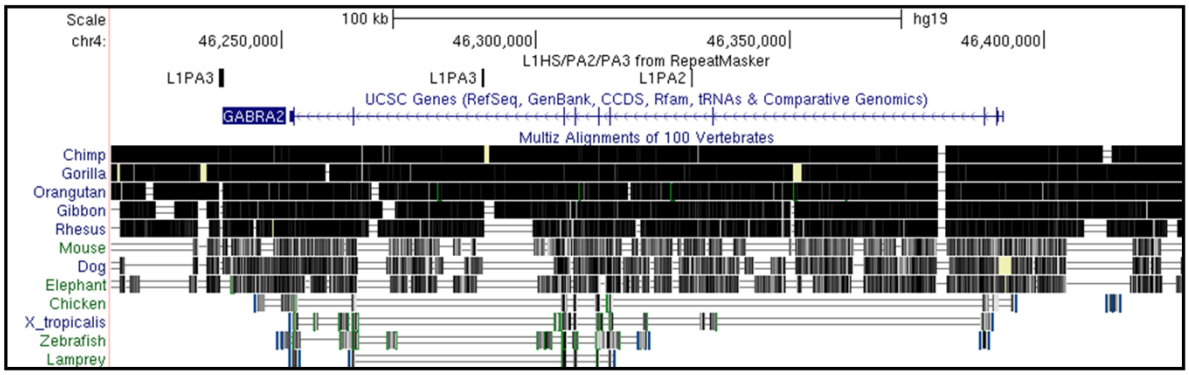
GABBR2



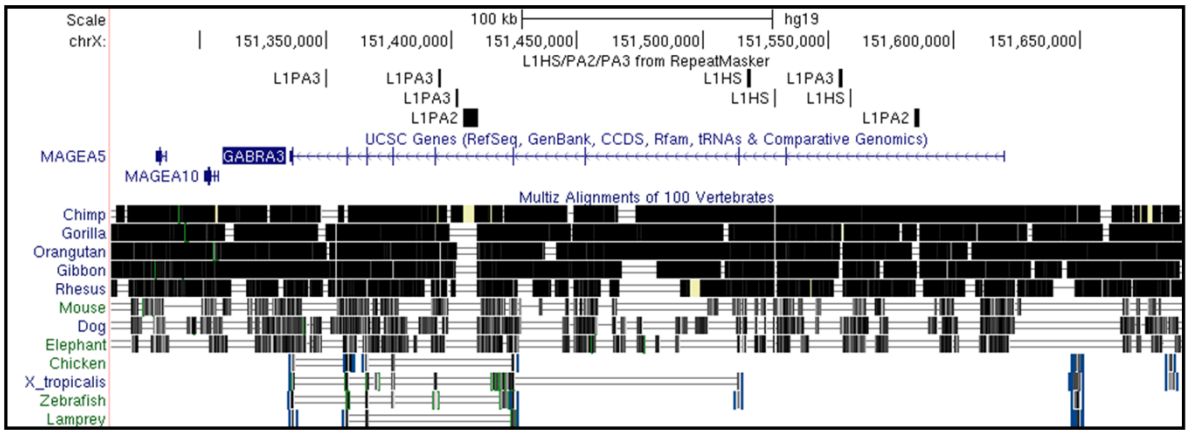
GABRA1



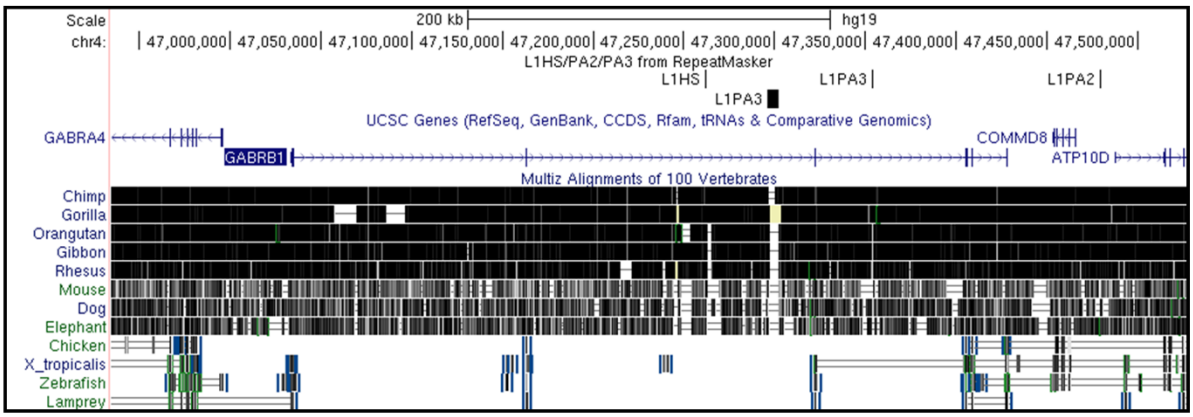
GABRA2



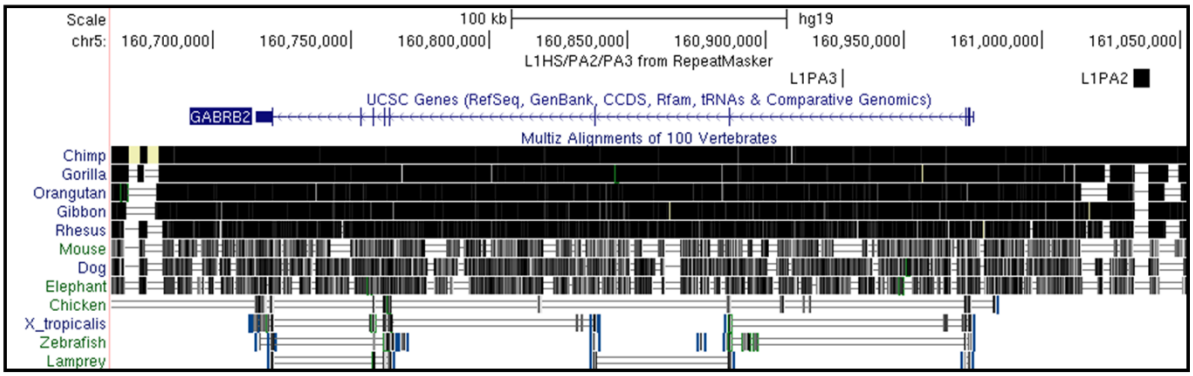
GABRA3



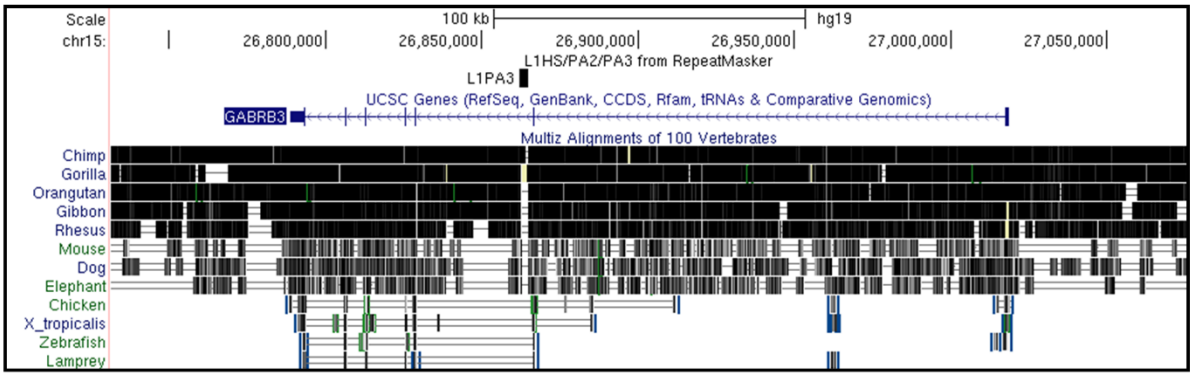
GABRB1



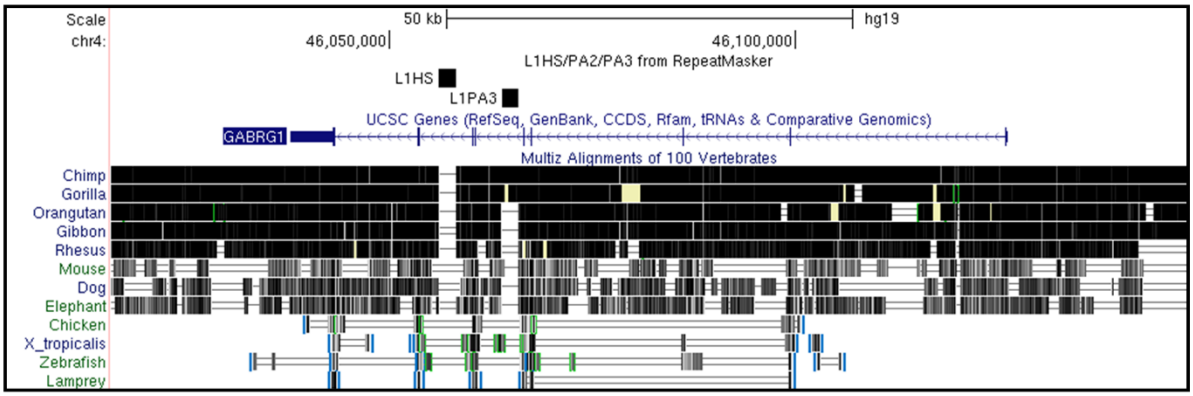
GABRB2



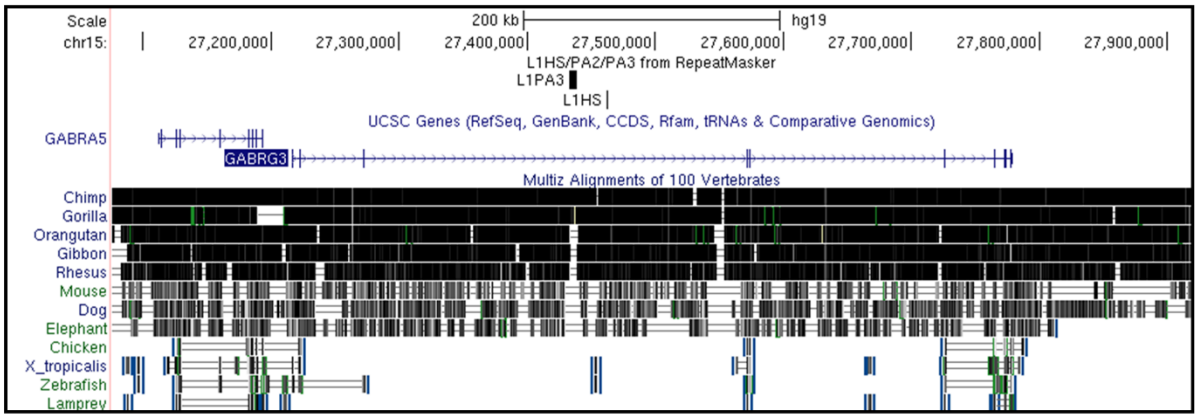
GABRB3



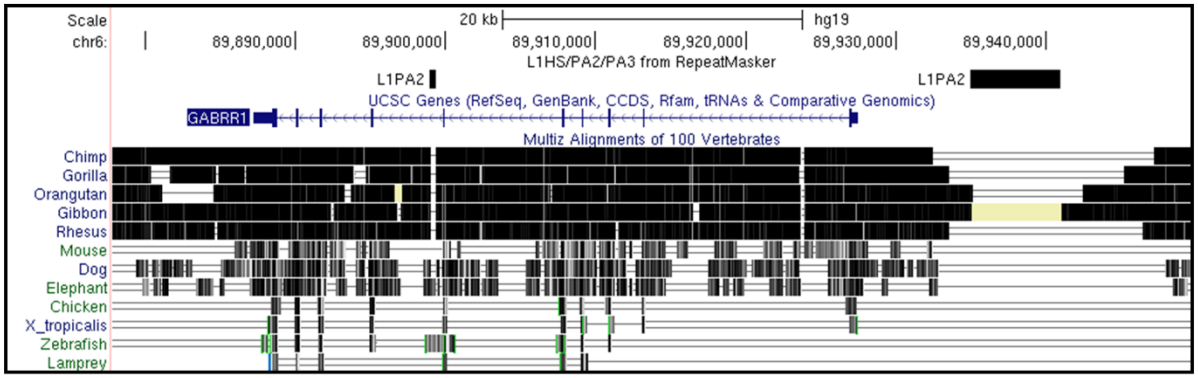
GABRG1



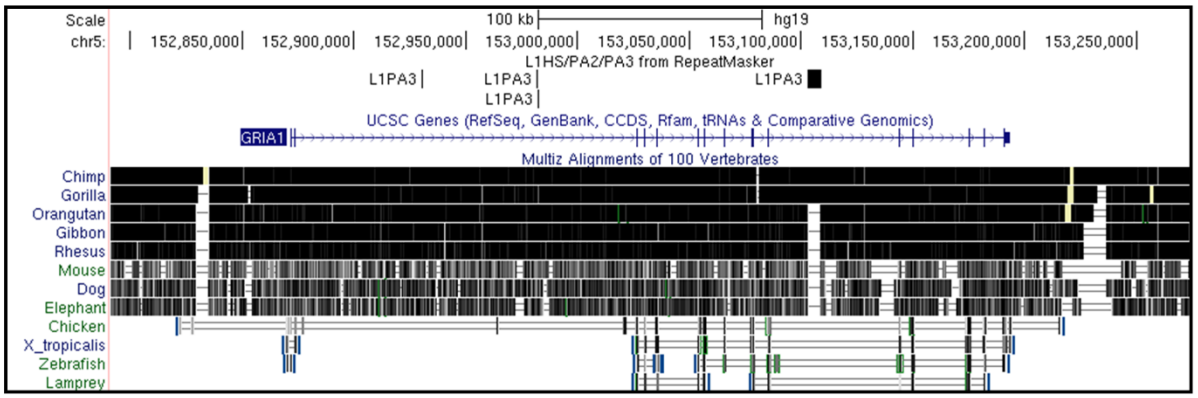
GABRG3



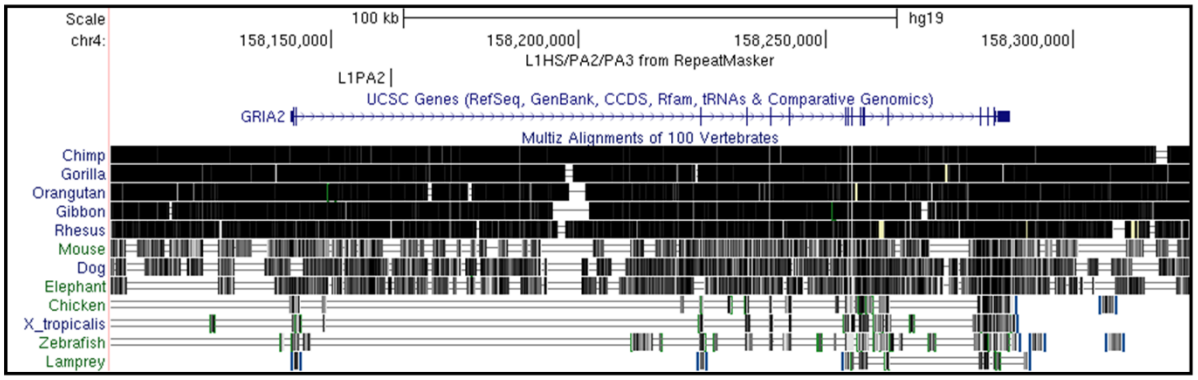
GABRR1



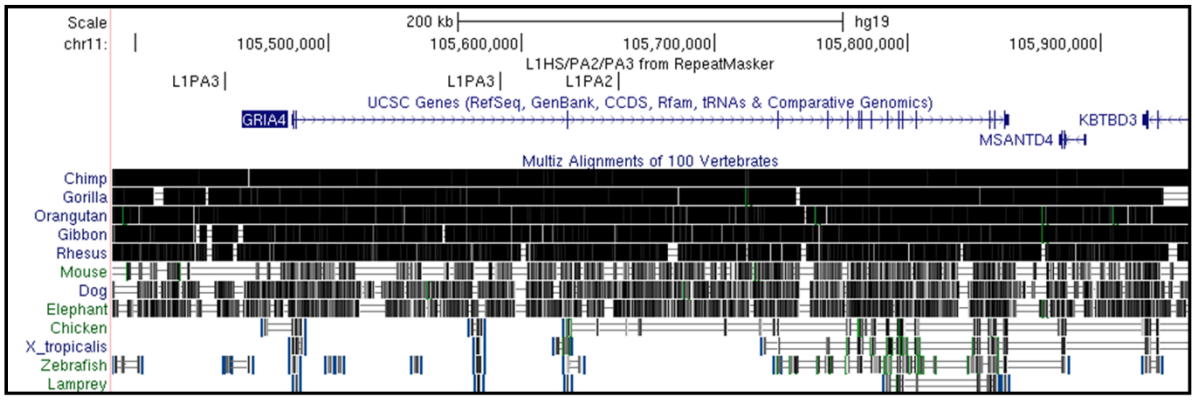
GRIA1



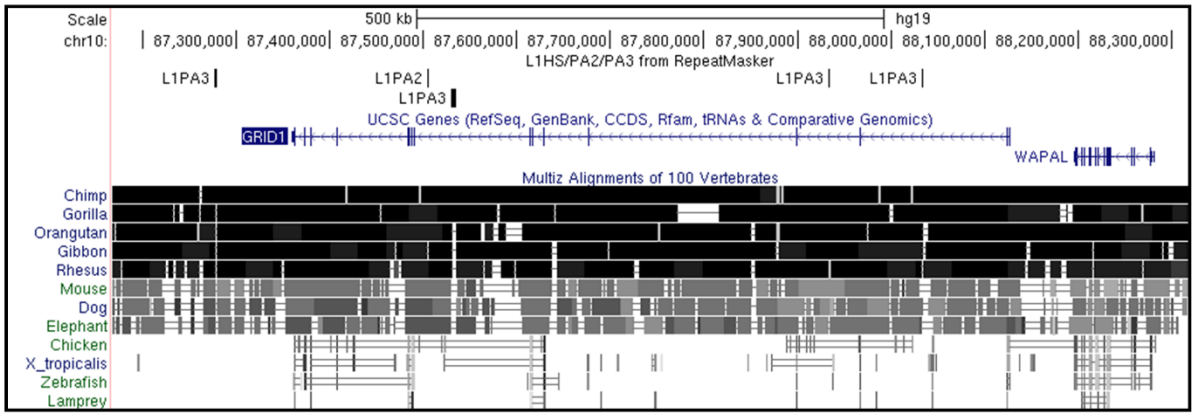
GRIA2



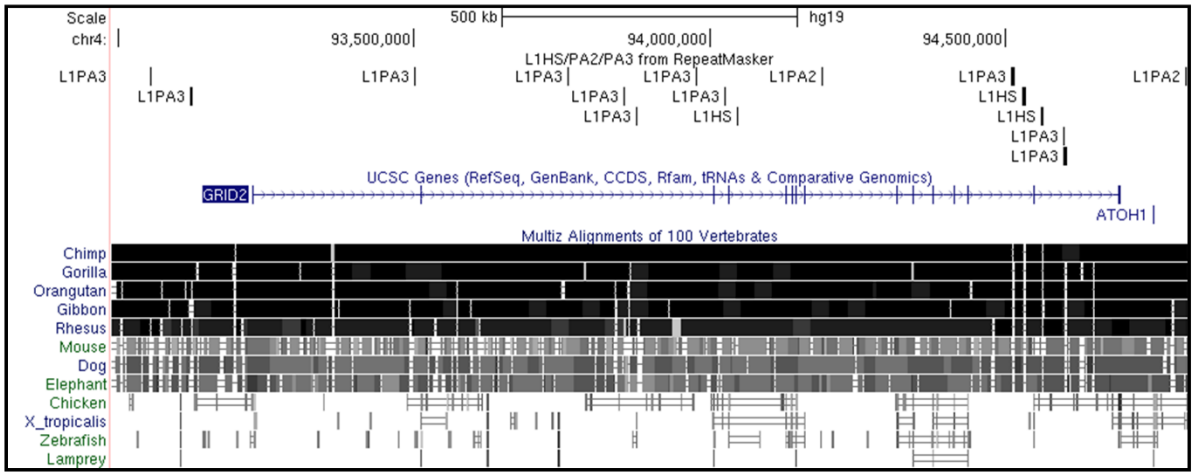
GRIA4



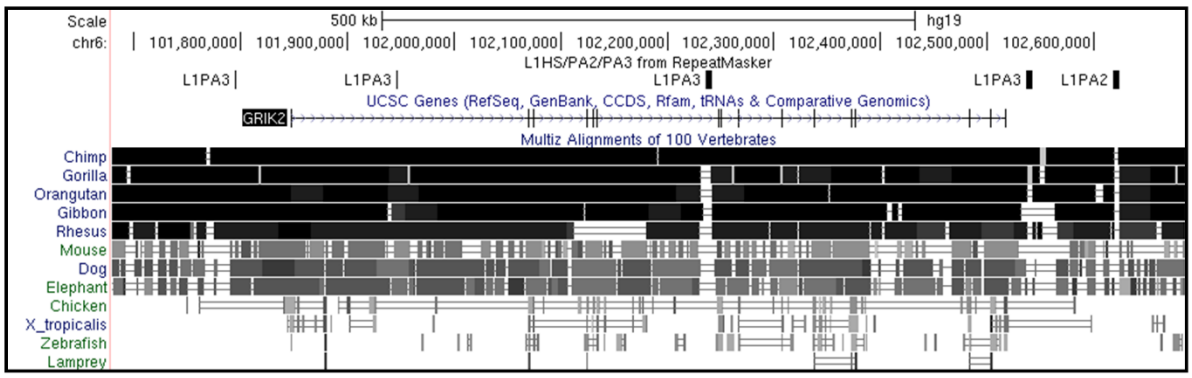
GRID1



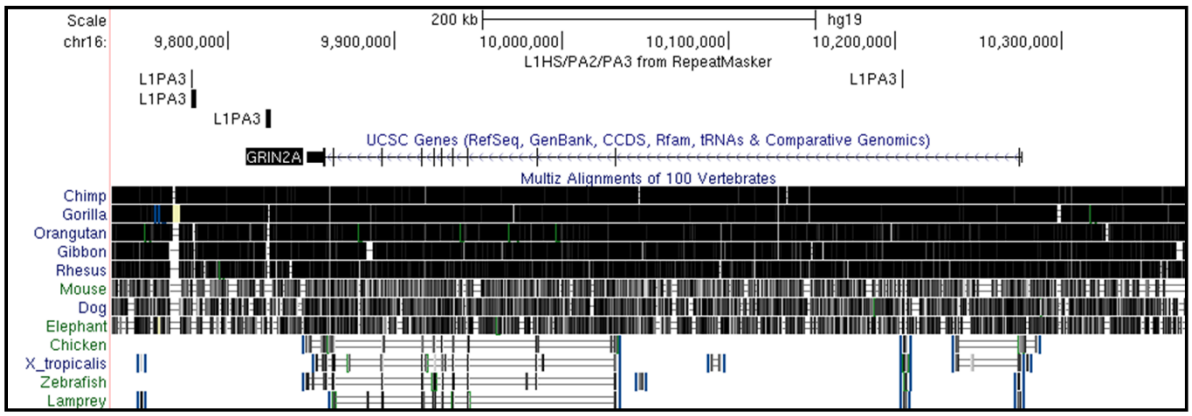
GRID2



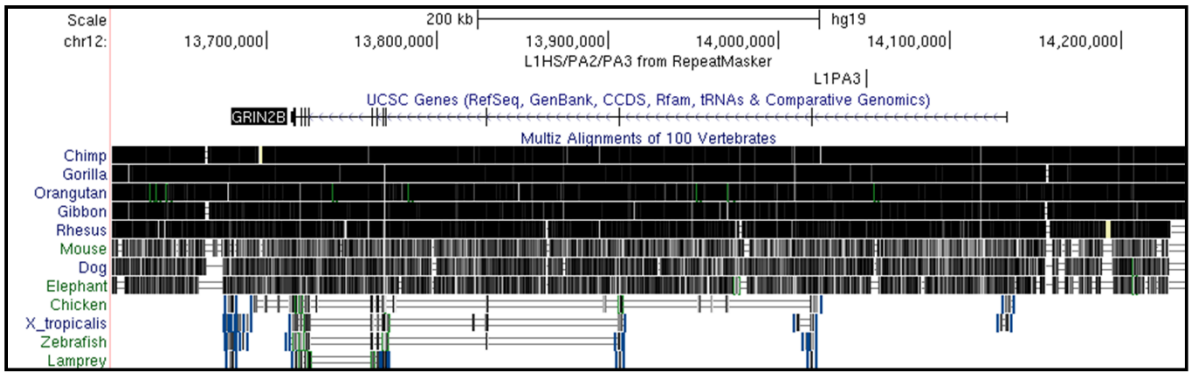
GRIK2



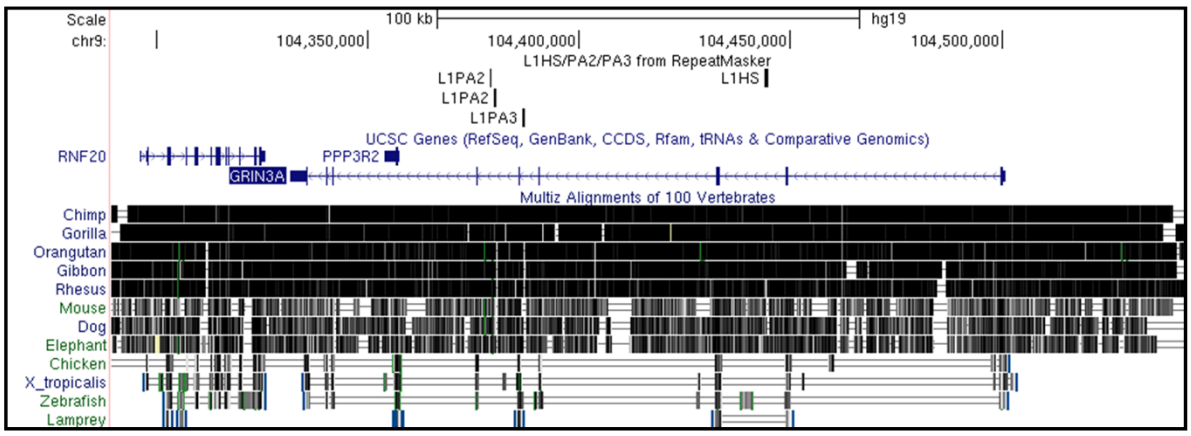
GRIN2A



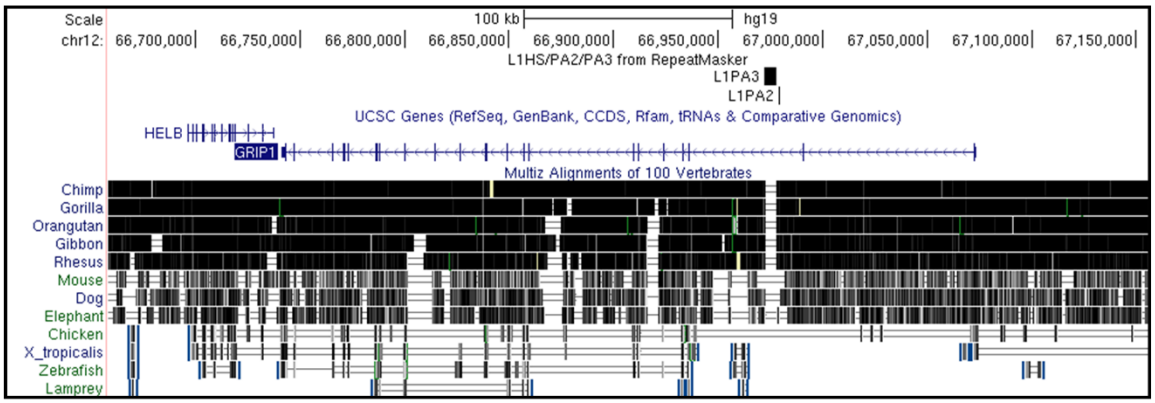
GRIN2B



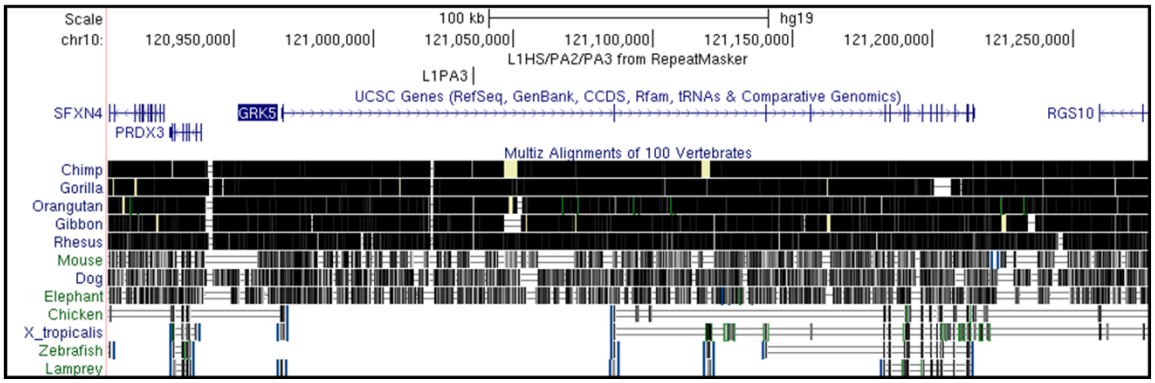
GRIN3A



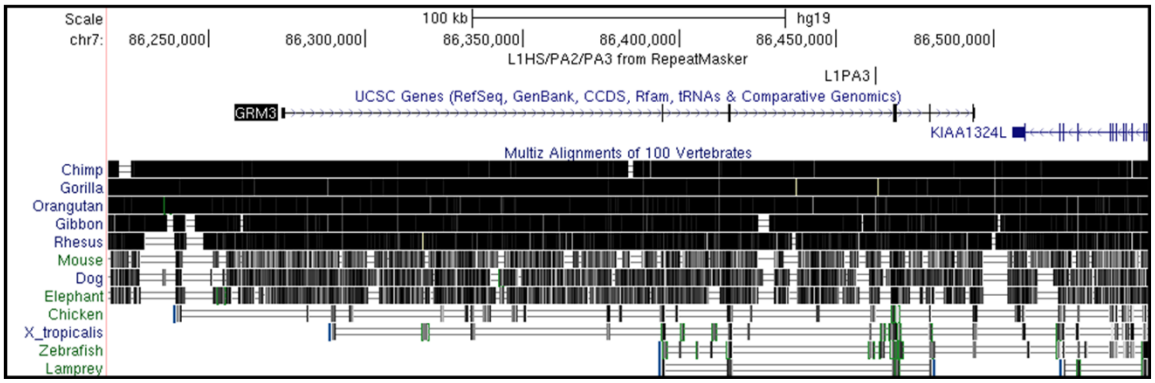
GRIP1



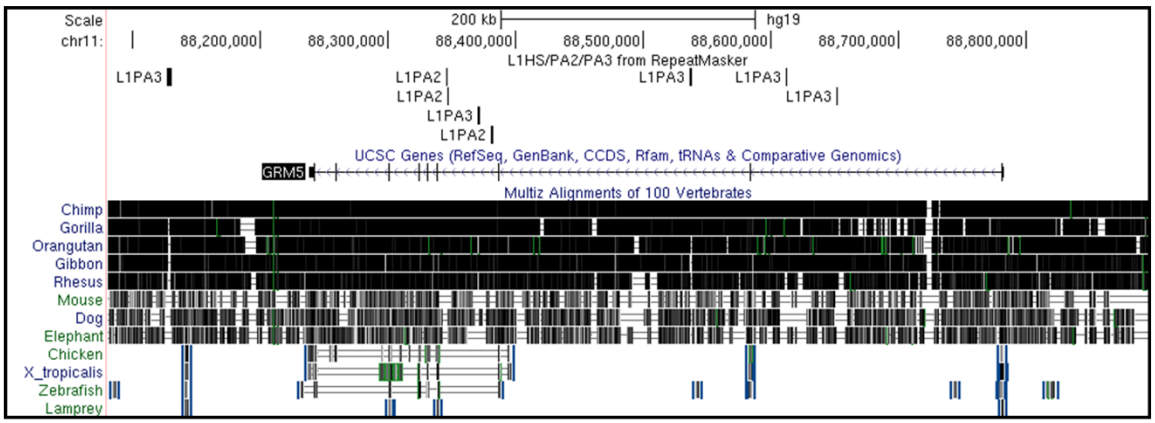
GRK5



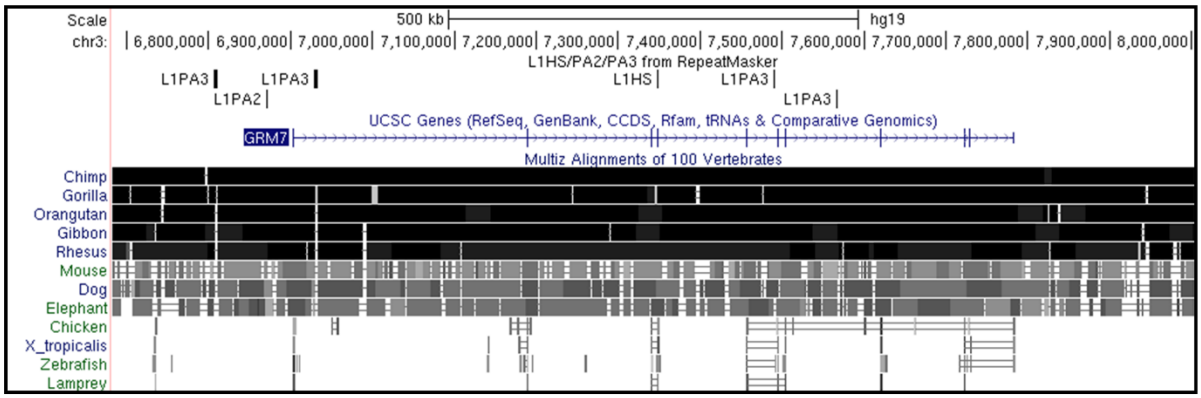
GRM3



GRM5



GRM7



GRM8

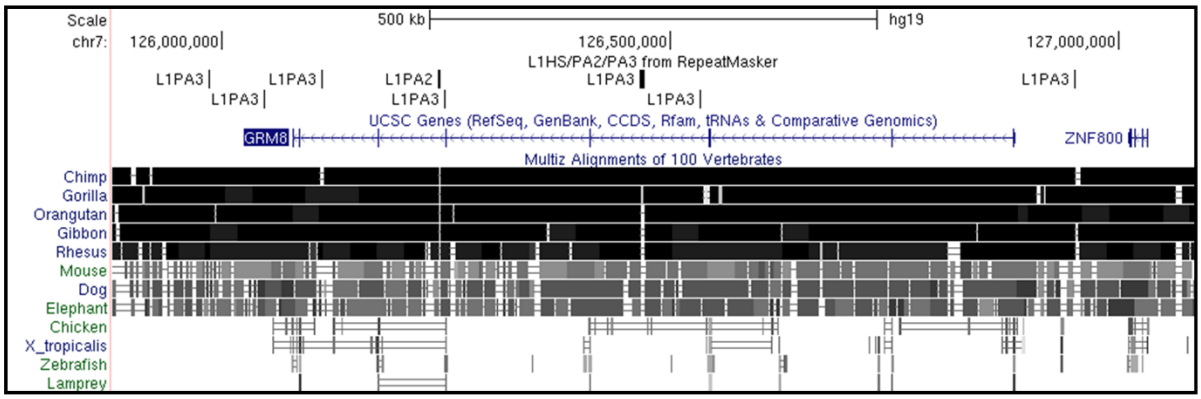


Figure 6.14 GABA and glutamate family genes with L1HS, L1PA2, and/or L1PA3 insertions.

Recent LINE-1 subfamily insertions are highly over-represented at GABA and glutamate family genes. The genome-wide average percentage of genes including one or more L1HS, L1PA2, or L1PA3 insertions is 10.7%. For GABA family genes, this is increased by nearly 400%, with 10 of the 19 genes in this family (52.6%) containing one or more recent LINE-1 subfamily insertion. Of those GABA genes with LINE-1 insertions, 50% have more than one insertion. GABRA3 particularly stands out with nine intronic LINE-1 insertions (four L1PA3, two L1PA2, and three L1HS). Of the glutamate family genes, 15 of 26 have one or more LINE-1 insertions (57.6%), which is a 440% increase compared to the genome-wide average. Of those with LINE-1 insertions, 67% have more than one insertion. Particularly, GRID2 stands out as containing 13 intronic LINE-1 insertions, with nine L1PA3, one L1PA2, and three L1HS. This provides evidence to suggest that GABA and glutamate gene families may have been significantly remodelled by recent LINE-1 subfamily insertions in higher primates and humans.

In particular, we find that the glutamate gene family appears to have been a consistent target of LINE-1 insertion, at least over the last 12-13 million years, with enrichment analysis repeatedly highlighting glutamate signalling pathways as enriched processes in the analysis of each of the three gene lists. In order to identify whether LINE-1 mediated remodelling of glutamate genes, or more generally of genes involved in brain-related processes, is still ongoing in humans, we repeated the above analysis using a collated list of non-reference genome retrotransposon insertion polymorphisms (RIPs) from multiple studies identifying new germline insertions in humans.

6.3.5 SVA and LINE-1 germline insertion polymorphisms suggest continued evolution of zinc finger and glutamate gene pathways

TEBreak is a tool created by Adam Ewing that can be used to find and characterise non-reference genome transposable element insertions from whole genome sequencing (WGS) data. Dr. Abigail Savage used this tool to access a list of germline, polymorphic retrotransposon insertions that had been identified in individuals from multiple studies (Helman et al. 2014, Shukla et al. 2013, Lee et al. 2012, Stewart et al. 2011, Wang et al. 2006, Sudmant et al. 2015). While these studies used differing methods and only a small number of insertions were validated, we can nonetheless make use of these known and predicted insertions as a starting point to investigate the distribution patterns of germline SVA and LINE-1 insertion polymorphisms across the genome.

A total of 1148 SVA RIPs were identified in the TEBreak list and used for distribution analysis in this section (Supplementary File 6.5). Plotting the number of transcripts per Mb versus the number of SVA RIPs per Mb confirmed that SVA RIPs follow the same general trend as reference SVAs, with a positive correlation between transcript

number and SVA RIP number (correlation coefficient = 0.31) (Figure 6.15 and Table 6.1). Clustering analysis demonstrated three loci across the genome with the highest number of SVA RIPs, with a total of five polymorphic insertions per Mb at Chr1:28,000,001-29,000,000, Chr9:134,000,001-135,000,000, and Chr19:44,000,001-45,000,000. Of the three regions, the Chr19:44,000,001-45,000,000 locus was determined to be an additional ZNF zinc finger gene cluster (Figure 6.16).

This region had three reference SVAs and was therefore not considered to be over-represented for reference SVA insertions. This may suggest SVA-mediated evolution at a further zinc finger locus which is ongoing in modern humans, independently of any clustering of reference SVA insertions at this locus.

In order to gain insight into particular genes or pathways that may be targets for SVA RIP insertions, the co-ordinates of all genes (plus 5 kb to capture the promoter region) from the UCSC Genome Browser's 'known gene' track (hg19) were overlaid with co-ordinates of the 1148 known or predicted SVA RIPs, which returned a list of 626 genes. Running this gene list through the Enrichr tool did not identify any significantly enriched biological pathways or mouse phenotypes associated with this gene set (Supplementary Data 6.3). However, analysis using the Gene Ontology cellular component data set demonstrated a small amount of significance for SVA RIPs around genes involved in transport vesicle membranes and the golgi membrane; an association that was primarily driven by multiple HLA genes in this gene set (Supplementary Data 6.3).

Figure 6.15 Both reference SVA insertions and SVA retrotransposon insertion polymorphisms are preferentially found at genic regions.

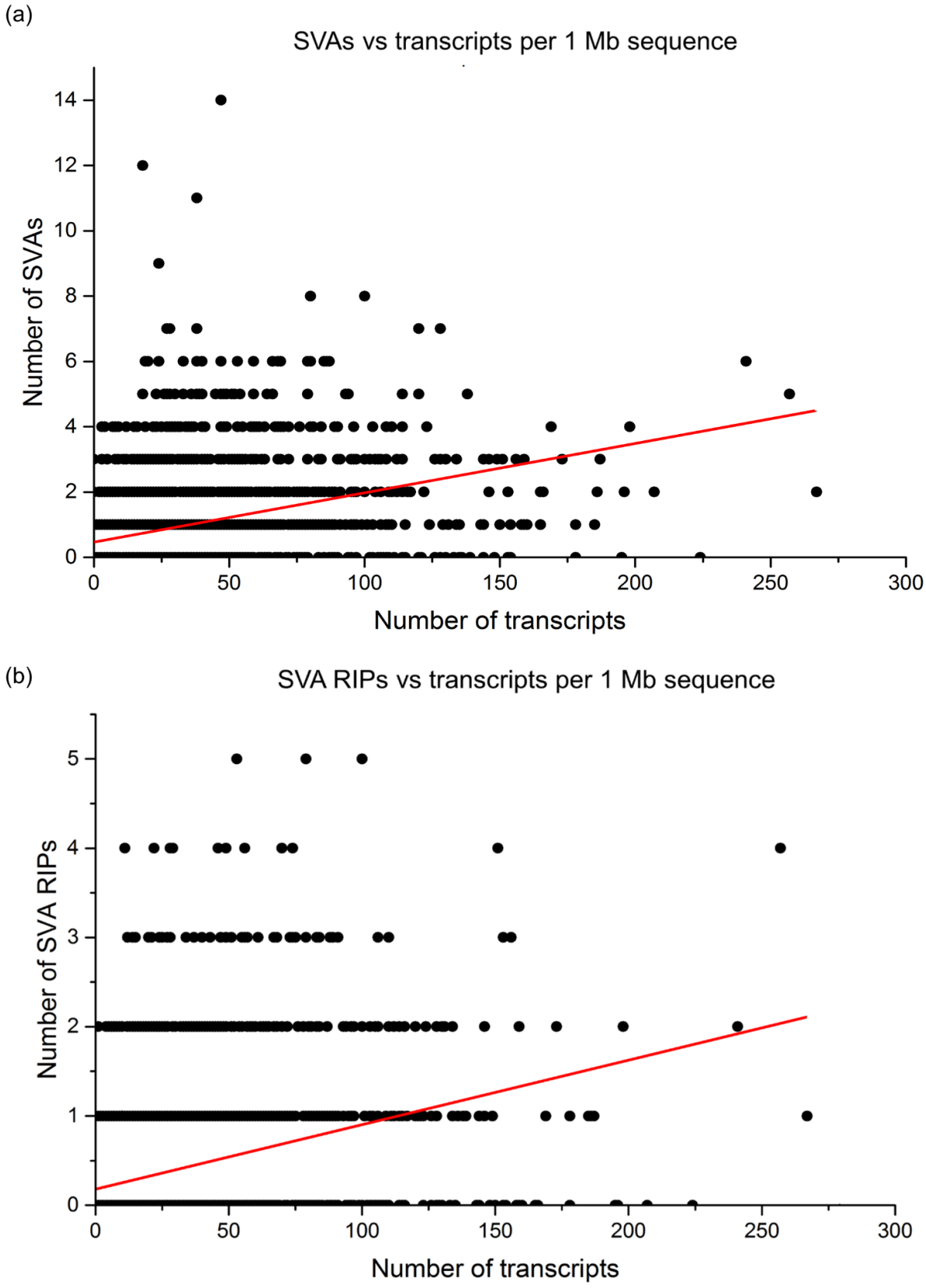


Figure 6.15 Both reference SVA insertions and SVA retrotransposon insertion polymorphisms are preferentially found at genic regions.

(a) Plotting reference SVAs vs transcripts per Mb across the whole genome shows that reference genome SVAs are preferentially found at genic regions, with higher transcript number per Mb correlating with higher SVA number (correlation coefficient = 0.352) (Table 6.1).

(b) Plotting the number of SVA RIPs vs transcript number per Mb shows that new and polymorphic SVA insertions follow the same trend as established reference SVAs, with a preference for genic regions (correlation coefficient= 0.306) (Table 6.1).

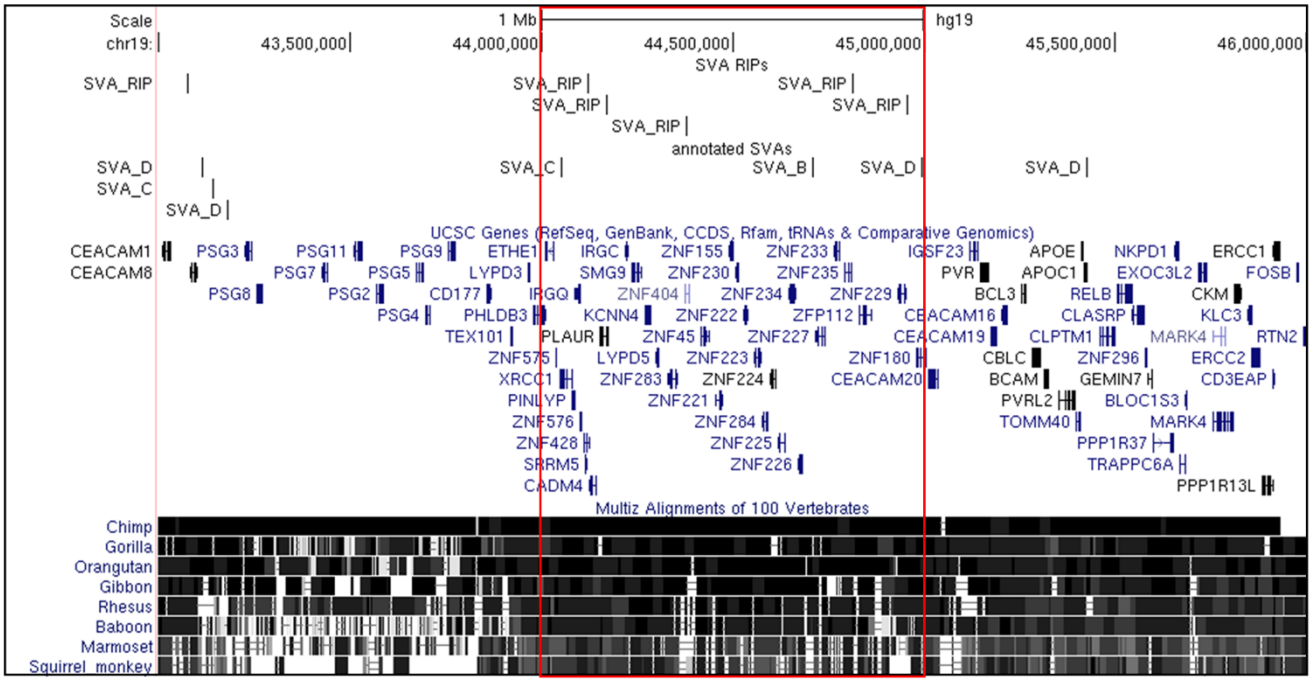


Figure 6.16 SVA RIPs cluster at a third zinc finger loci on chromosome 19, independently of reference SVAs.

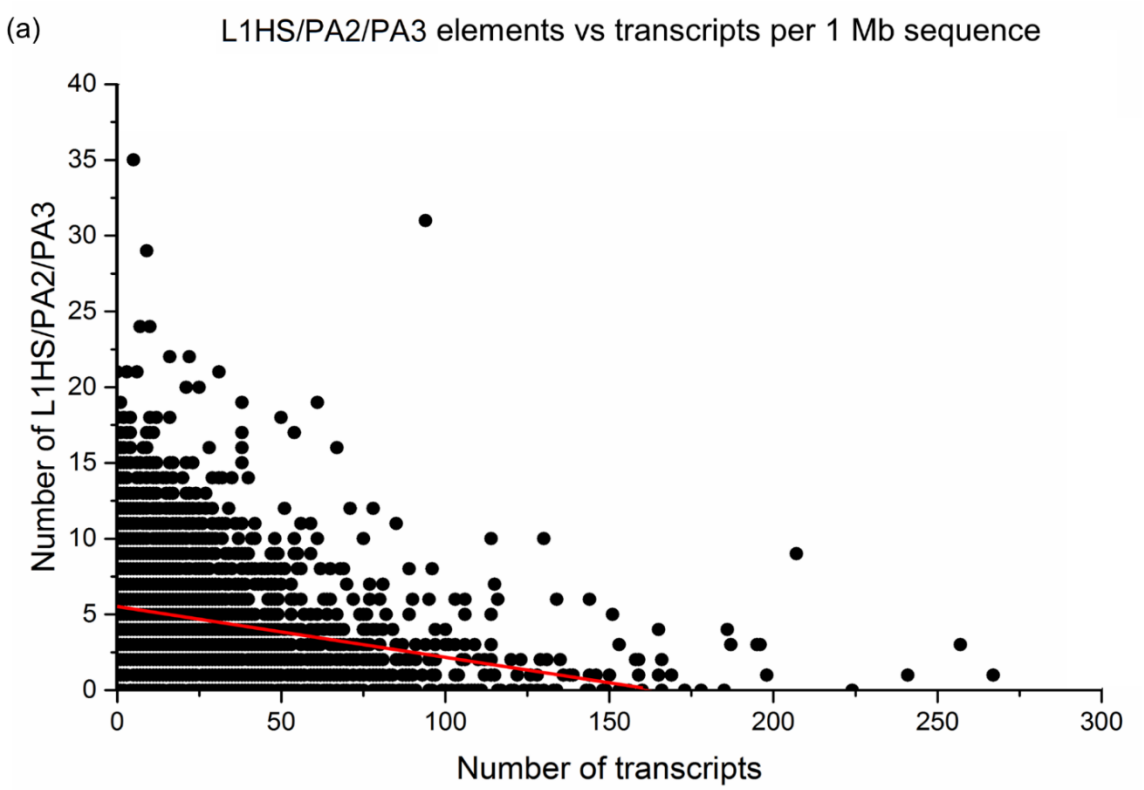
Visualisation of the chr19:44,000,000-45,000,000 locus (boxed) demonstrated that five SVA RIPs clustered over a third ZNF locus on this chromosome. With only three reference SVA insertions, this zinc finger cluster appears to be the target of SVA RIP insertion independently of previous reference genome SVA-mediated evolution. This could suggest that the chr19:44,000,000-45,000,000 ZNF locus may be continuing to undergo SVA-mediated change in modern humans.

Lack of significant enrichment in the analysis of genes with SVA RIPs may be a result of their young age, with their presence in the genome likely having not yet undergone evolutionary selection to fix or remove advantageous or deleterious insertions, with the group as a whole therefore retaining a degree of randomness in their distribution. Further to this, the current list of known or predicted germline SVA RIPs remains far from capturing the full extent of the likely variation across the genome and across the population. It may therefore be the case that a role for SVA RIPs in targeting specific genes or pathways could become more apparent when a wider range of data is available.

Repeating the same analysis on 6754 LINE-1 RIPs (Supplementary File 6.6) demonstrated that these insertions follow the same pattern as reference genome LINE-1 elements in that their rate of insertion is decreased around genic regions (correlation coefficient = -0.23; Figure 6.17; Table 6.2). However, when analysing this data by chromosome, we find that LINE-1 RIPs do not appear to preferentially target the X chromosome, as we had previously seen for reference LINE-1 insertions (data not shown). However, the validity of this result is not clear, as this may be due to differing methods or sample cohorts used in the numerous papers from which this data was collated, which may have included a sex bias in samples, or disregarded the sex chromosomes in their analysis.

We note that many of the top regions for LINE-1 RIP clustering are found at or adjacent to centromeres and telomeres, including the chromosome 4 telomere and the centromeres of numerous chromosomes, including chromosomes 2, 4, 5, 6, 7, 12, 16, 19, 20, and 21 (Supplementary Data 6.1).

Figure 6.17 Both recent reference LINE-1 subfamily insertions and LINE-1 retrotransposon insertion polymorphisms are preferentially found in gene poor regions.



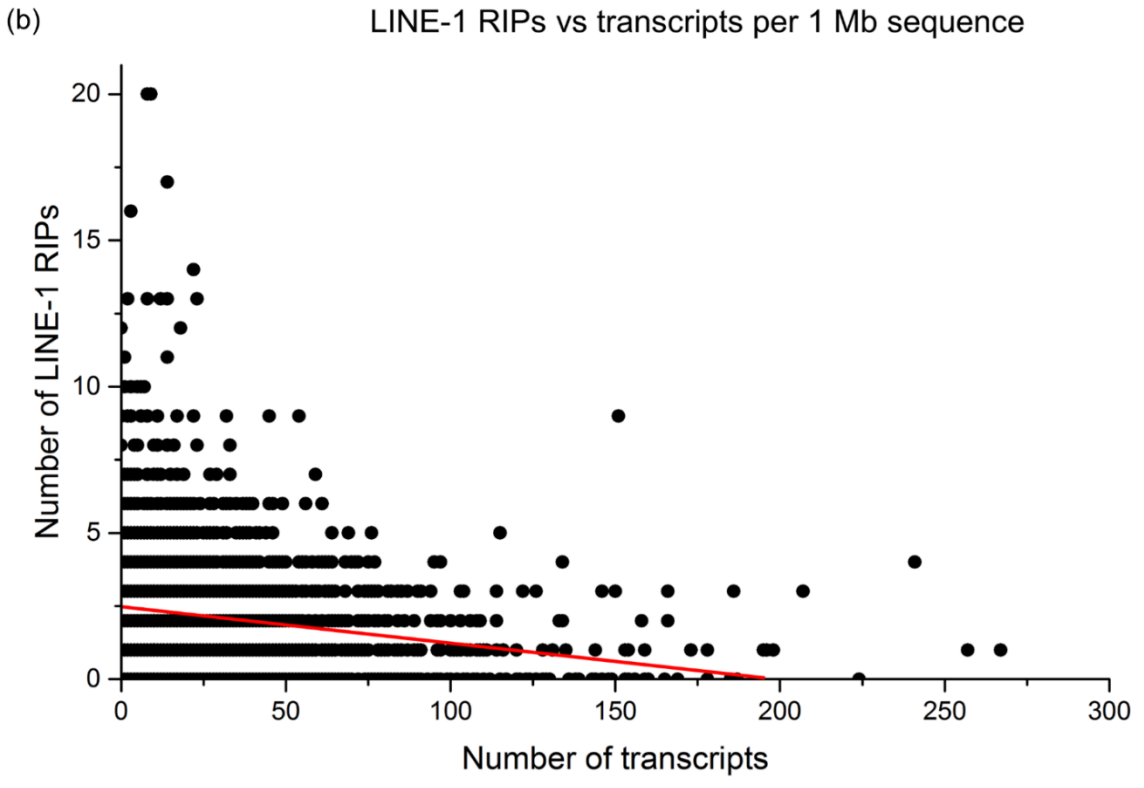


Figure 6.17 Both recent reference LINE-1 subfamily insertions and LINE-1 retrotransposon insertion polymorphisms are preferentially found in gene poor regions.

(a) Plotting reference LINE-1 (L1PA3, L1PA2, and L1HS) insertions vs transcripts per Mb across the whole genome shows that recent reference genome LINE-1s are preferentially found at gene poor regions, with higher transcript number per Mb correlating with lower LINE-1 number (correlation coefficient = -0.308) (Table 6.1).

(b) Plotting the number of LINE-1 RIPs vs transcript number per Mb shows that new and polymorphic LINE-1 insertions follow the same trend as established reference LINE-1s, with a preference for gene poor regions (correlation coefficient= -0.230) (Table 6.1).

Human centromeres are known to be made up of repetitive AT-rich α-satellite DNA (Manuelidis 1978), with over-representation of LINE-1 elements at ectopic centromeres having been noted in the literature (Chueh et al. 2009). In particular, LINE-1 DNA is increased by 2.5-fold at Centromere Protein-A (CENP-A) binding clusters. CENP-A is a marker of centromeres, being incorporated into centromeric chromatin where it replaces histone H3 in nucleosomes to generate a unique chromatin conformation and establish centromeric identity (Yoda et al. 2000, Sullivan and Karpen 2004). Chueh et al. have demonstrated that LINE-1 RNA binds CENP-A and becomes incorporated into CENP-A-associated chromatin, with knockdown of LINE-1 RNA revealing its critical role in maintaining normal mitotic function (Chueh et al. 2009).

Hypomethylation of LINE-1 elements at the centromere has further been shown to be correlated with chromosomal instability in head and neck squamous cell carcinoma (HNSCC), with cancerous cells frequently displaying chromosomal aberrations which are often associated with centromeric breaks (Martinez et al. 2012a, Martinez et al. 2012b). This would support the hypothesis that LINE-1 elements at the centromere play an important role in chromosome stability and normal mitotic functioning. Repetitive DNA at human centromeres is known to be polymorphic between individuals (Waye, Greig and Willard 1987, Dávila-Rodríguez et al. 2011), and polymorphic LINE-1 insertions have been identified previously at the Y chromosome centromere (Santos et al. 2000). Given that many regions of LINE-1 RIP clustering map to centromeric regions, this may add to inter-individual differences, and potentially to chromosome stability.

Using the UCSC Genome Browser to overlay the LINE-1 RIP co-ordinates with the co-ordinates of all transcripts in the genome plus 5 kb upstream, we generated a list of

2125 genes with one or more LINE-1 RIP insertion within or upstream of one or more of their transcripts. Running this gene list through Enrichr for Gene Ontology biological processes demonstrated significant enrichment for 34 processes, with the majority being directly brain-related (Figure 6.18a; Supplementary Data 6.3). Significantly enriched biological processes included:

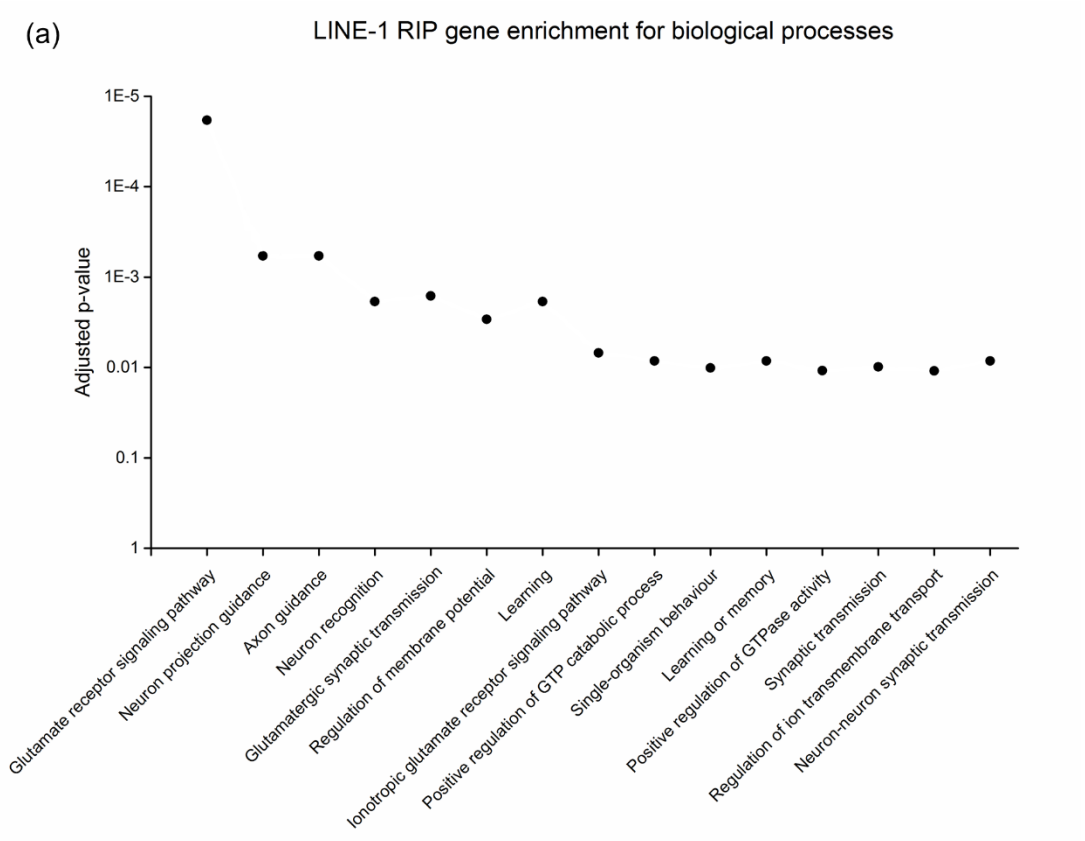
- Glutamate receptor signalling pathway — adjusted $p = 1.84 \times 10^{-5}$
- Glutamatergic synaptic transmission — adjusted $p = 1.62 \times 10^{-3}$
- Ionotropic glutamate receptor signalling pathway — adjusted $p = 6.89 \times 10^{-3}$
- Neuron projection guidance — adjusted $p = 5.85 \times 10^{-4}$
- Axon guidance — adjusted $p = 5.85 \times 10^{-4}$
- Neuron recognition — adjusted $p = 1.87 \times 10^{-3}$
- Synaptic transmission — adjusted $p = 9.85 \times 10^{-3}$
- Learning — adjusted $p = 1.87 \times 10^{-3}$
- Learning or memory — adjusted $p = 8.47 \times 10^{-3}$
- Behaviour — adjusted $p = 3.81 \times 10^{-2}$
- Cognition — adjusted $p = 3.81 \times 10^{-2}$

We also observed enrichment for numerous roles in transmembrane ion transport (particularly for calcium), membrane depolarization, and regulation of synaptic membrane potential, as well as synapse organisation and CNS development.

Enrichment for cellular components strongly confirms LINE-1 RIP insertions at genes enriched for roles in the brain, with 10 of the 18 enriched terms directly stating their presence in brain related cells or locations (Figure 6.18b), including the:

- Synapse part — adjusted $p = 7.77 \times 10^{-10}$
- Synaptic membrane — adjusted $p = 4.24 \times 10^{-8}$
- Pre-synaptic membrane — adjusted $p = 1.98 \times 10^{-3}$
- Post-synaptic membrane — adjusted $p = 3.79 \times 10^{-7}$
- Post-synaptic density — adjusted $p = 1.00 \times 10^{-2}$
- Pre-synaptic active zone — adjusted $p = 2.27 \times 10^{-2}$

Figure 6.18 Genes targeted by LINE-1 retrotransposon insertion polymorphisms show strongest association for brain-related pathways, particularly in glutamatergic signalling.



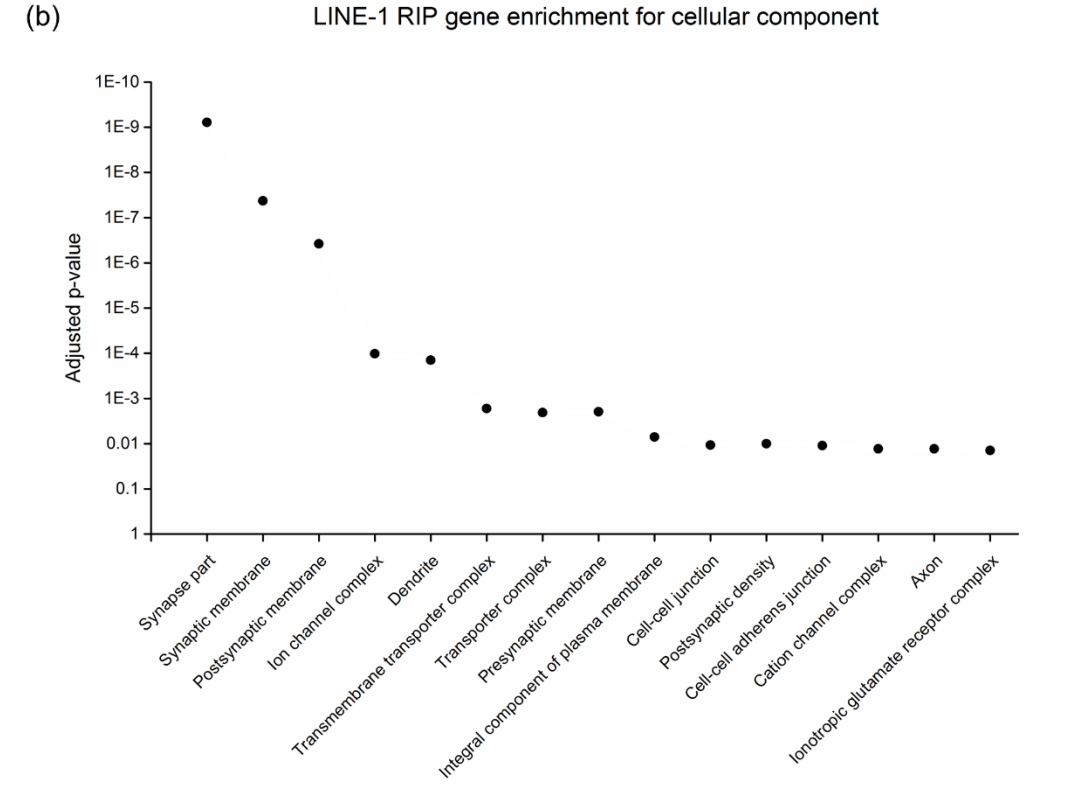


Figure 6.18 Genes targeted by LINE-1 retrotransposon insertion polymorphisms show strongest association for brain-related pathways, particularly in glutamatergic signalling.

(a) Enrichment analysis for biological processes using the set of 2125 genes with LINE-1 RIPs finds 34 significantly enriched terms, the majority of which were brain-related (Supplementary Data 6.3). Genes with LINE-1 RIPs were enriched for processes including glutamate signalling and glutamatergic synaptic transmission (Benjamini-Hochberg adjusted Fisher's exact p-value = 9.85 $x10^{-3}$, 8.47 $x10^{-3}$, and 1.62 $x10^{-3}$), neuron projection, guidance, and recognition (Benjamini-Hochberg adjusted Fisher's exact p-value = 5.85 $x10^{-4}$ and 1.87 $x10^{-3}$), and functions including learning, memory, behaviour, and cognition (Benjamini-Hochberg adjusted Fisher's exact p-value = 1.87 $x10^{-3}$, 8.47 $x10^{-3}$, and 3.81 $x10^{-2}$). This suggests a role for new and polymorphic LINE-1 insertions in the potential ongoing evolution or individual variation in CNS signalling and cognitive function.

(b) Enrichment for cellular localisation using the same gene set demonstrated enrichment in the synapse and synaptic membranes (Benjamini-Hochberg adjusted Fisher's exact p-values = 7.77 $x10^{-10}$, 4.24 $x10^{-8}$, 3.79 $x10^{-7}$, and 1.98 $x10^{-3}$), as well as transporter complexes (1.67 $x10^{-3}$, and 2.06 $x10^{-3}$) including ion and cation channel complexes (1.02 $x10^{-4}$ and 1.29 $x10^{-2}$) and glutamate receptor complexes (1.42 $x10^{-2}$).

- Synapse adjusted p = 3.15 $x10^{-2}$
- Dendrite adjusted p = 1.43 $x10^{-4}$
- Axon adjusted p = 1.29 $x10^{-2}$
- Ionotropic glutamate receptor adjusted p = 1.41 $x10^{-2}$

We also find a clear role for this gene set in ion transport (ion and cation channel complexes; 1.02 $x10^{-4}$ and 1.29 $x10^{-2}$), as well as in transmembrane transporter complexes (transmembrane transporter complex, transporter complex; 1.67 $x10^{-3}$ and 2.06 $x10^{-3}$).

Taken together, these data show a clear bias for both established, reference LINE-1 insertions and new, polymorphic LINE-1 insertions targeting genes involved in brain-related pathways, and particularly in glutamatergic signalling. This would suggest that glutamate genes and pathways have been consistently undergoing LINE-1-mediated evolution for at least 12-13 million years, with this trend appearing to continue in modern humans.

6.4 Discussion

The mobilisation and insertion of retrotransposons is a known driver of genetic diversity, with integration of retrotransposable elements within genic regions allowing new regulatory capacities through mechanisms such as transcription factor binding, alteration of local chromatin structure, or altered gene splicing (Cordaux and Batzer 2009, Erwin et al. 2014). New retrotransposon insertions are likely to enable the tissue-specific and stimulus-inducible regulation of genes uniquely in different species (Robbez-Masson and Rowe 2015), or in the case of polymorphic insertions, in different individuals. In this chapter, we investigated the distribution patterns of the most recent retrotransposon classes across the human genome, comparing the youngest class, SVAs, with the three most recent LINE-1 subfamilies, L1HS, L1PA2, and L1PA3, as a

match for evolutionary age. We then extended this analysis to non-reference, polymorphic SVA and LINE-1 RIPs to determine whether more recent insertions followed similar patterns to their reference genome counterparts. Alu elements were not considered in this analysis due to their large number within the human genome. This enabled us to trace retrotransposon insertions across the human genome over the last 13.6 million years of evolution and into the present day, allowing an understanding of the genes and pathways that have undergone retrotransposon-mediated alteration in higher primates and humans, and which may be variably affected between individuals within the human population.

We find that SVAs have repeatedly targeted C_2H_2-type ZNF zinc finger loci throughout recent primate and human evolution, consistently across the chr19:20,000,000-24,000,000 locus, with both older and more recent subfamilies enriched at this region (Figure 6.2). More recently, younger SVA subfamilies have targeted a further three zinc finger clusters on chromosomes 19, 7, and 4 (Figure 6.3), with SVA RIPs targeting an additional chromosome 19 zinc finger locus independently of reference SVAs at this locus (Figure 6.16). C_2H_2-type ZNF zinc fingers are the largest family of transcription factors in the human genome (Vaquerizas et al. 2009) and are known to have a complex evolutionary history due to repeated rounds of duplication, which has generated substantial diversity throughout many species (Emerson and Thomas 2009, Nowick et al. 2011). In particular, the ZNF91 subfamily is known to have duplicated across multiple chromosomes in primate genomes, giving rise to primate-specific zinc finger genes (Hamilton et al. 2006).

The association of SVAs with ZNF loci is of interest due to the known role of zinc finger genes in binding and silencing a wide range of retrotransposable elements, through which they are known to direct gene expression in both embryonic and adult tissues

(Ecco et al. 2016, Wolf et al. 2015, Castro-Diaz et al. 2014). In particular, the primate-specific zinc finger gene, ZNF91, is known to repress SVAs (Jacobs et al. 2014), which is of interest due to its location within the most SVA dense region of the human genome at chr20,000,000-24,000,000 (Figure 6.2), with the area encompassing ZNF91 having undergone recent primate-specific change through the insertion of four SVA elements (Figure 6.4).

It is therefore clear that SVAs have repeatedly and specifically been involved in the retrotransposon-mediated alteration of ZNF gene loci throughout recent primate history. KRAB-containing zinc finger proteins have recently been shown to bind transposable elements, and to use fragments of these elements as regulatory platforms from which they control species- and tissue-specific gene regulation (Imbeault, Helleboid and Trono 2017). Many C_2H_2-type zinc finger genes contain KRAB domains, and thus the addition of large numbers of retrotransposable elements around these gene clusters is likely to have resulted in significant changes uniquely in higher primates and humans in their ability to modulate zinc finger expression, which may have far reaching transcriptional consequences with regard to the gene networks that are controlled by zinc finger proteins (Emerson and Thomas 2009, Nowick et al. 2011). Indeed, the expression of zinc finger genes has been suggested to be one of the defining differences between chimpanzee and human brains, but not within other tissues tested, such as the liver or heart (Nowick et al. 2009), which our data may suggest is driven in part by SVA-mediated evolution and tissue-specific regulation of zinc finger genes uniquely in different primate species.

There is increasing support in the literature to suggest a role for retrotransposons in CNS conditions, with changes in the expression and regulation of transposable elements being observed in animal models of neurodegeneration with relevance to

Amyotrophic Lateral Sclerosis (ALS), Fronto-Temporal Lobe Degeneration (FTLD) and Fragile X (Krug et al. 2017, Li et al. 2012, Tan et al. 2012). In humans, increased LINE-1 expression has been observed in the brains of some individuals with a diagnosis of schizophrenia and autism (Bundo et al. 2014, Shpyleva et al. 2017). However, it should be noted that in the case of Bundo et al. this was a post-hoc finding after analysing LINE-1 expression in the brains of individuals with multiple psychiatric conditions, including schizophrenia, bipolar disorder, and major depression. The initial sample set included 13 healthy controls, 13 individuals with schizophrenia, 13 individuals with bipolar disorder, and 12 individuals with major depression. Alteration in LINE-1 expression was found to be small but significant ($p = 0.0295$ and $p = 0.0061$ after normalisation to HERVH and alpha satellite expression, respectively) only in individuals with schizophrenia. The schizophrenia subset was then selected for further study and the results replicated in samples from 35 individuals with schizophrenia compared to 34 healthy controls. Such work would ideally need to be replicated with a larger sample size to improve statistical rigour. Similarly, work by Shpyleva et al. tested LINE-1 expression in four brain regions (cerebellum, frontal cortex BA9, auditory cortex BA22, and anterior cingulate BA24) in 13 individuals with autism and 13 controls. This work demonstrated an increase in LINE-1 expression in only one of the four regions tested (cerebellum), and was limited by the small sample size. This work should be replicated in a larger sample set in order to confirm the findings described by Shpyleva et al.

More specifically in the case of schizophrenia, it has been demonstrated that LINE-1 methylation in peripheral blood leukocytes of individuals experiencing first episode psychosis is significantly decreased only in those who have experienced childhood trauma, while first episode psychosis individuals without trauma display similar LINE-

1 methylation levels to control individuals (Misiak et al. 2015). Supporting the idea that stress and trauma can affect LINE-1 methylation with implications in mental health, US military service members with a diagnosis of post-traumatic stress disorder (PTSD) display hypomethylation of LINE-1 in peripheral blood post-deployment compared to their pre-deployment methylation levels (Rusiecki et al. 2012). This would suggest that, in addition to CNS conditions such as autism and schizophrenia, exposure to environmental stress is also able to directly impact LINE-1 methylation in ways that may modulate risk for these conditions.

In this chapter, we have demonstrated that both the three most recent subfamilies of reference LINE-1 insertions, and non-reference LINE-1 RIPs, are preferentially found at genes enriched for roles in brain-related pathways, and particularly in genes involved in GABA and glutamate signalling (Figure 6.14; Figure 6.18). It would therefore follow that disease states or environmental exposures that alter LINE-1 regulation or mobilisation would be likely to disproportionately affect brain-related gene pathways. This may result either through changes to LINE-1 ability to regulate nearby gene expression, such as through changes to methylation or local chromatin structure, or through active expression of LINE-1 elements, which may interfere with the usual expression of the host gene, for example, through steric hindrance of active transcription by expression of an intronic anti-sense LINE-1 element. With LINE-1 elements being over-represented in brain-related pathways, the changes in LINE-1 expression and regulation seen in the brains of individuals with schizophrenia and autism may highlight one contributing mechanism to the retrotransposon-mediated risk in neurological conditions, which may be increased significantly in individuals with a higher number of LINE-1 RIPs within genes involved in CNS pathways.

6.5 Summary

We have demonstrated that recent primate-specific retrotransposon families, SVAs and the three most recent classes of LINE-1, are not distributed randomly across the human genome. Instead, SVAs are preferentially found at genic regions, and are over-represented at multiple zinc finger loci on chromosomes 4, 7, and 19. Both older and younger classes of SVAs cluster together at the Chr19:20,000,000-24,000,000 locus, whereas the remaining ZNF clusters highlighted in this chapter show over-representation only for the younger SVA subclasses. ZNFs are the largest group of transcription factors in the genome, and the likely modulation of expression of these genes by SVA insertions across multiple ZNF loci is likely to have influenced a large range of transcriptional networks uniquely in higher primates and humans. In contrast, LINE-1 subfamilies L1HS, L1PA2, and L1PA3 are preferentially found at regions with few genes. Enrichment analysis of genes containing one or more recent LINE-1 subfamily insertions within or up to 5 kb upstream of their sequence demonstrated consistent LINE-1 mediated remodelling of GABA and glutamate gene loci over recent evolutionary history. Extending this analysis to incorporate retrotransposon insertion polymorphisms confirmed that both SVA and LINE-1 RIPs follow similar patterns to their reference genome counterparts, with the association of LINE-1 RIPs being even stronger than reference LINE-1 for insertion at gene loci involved in brain-related pathways. Such findings may add to our understanding of LINE-1 related mechanisms in CNS conditions.

Chapter 7

Thesis summary

Maintaining the appropriate transcriptional balance in the cell, while allowing transient responses to environmental stimuli, is a complex process involving numerous mechanisms, including the action of regulatory gene pathways which can exert their effects via non-coding regulatory elements. This thesis primarily aimed to extend our understanding of transcriptional regulation at the MIR137 schizophrenia-associated locus and around other CNS-related genes including DNAJC5 and MIR941. Prior work identifying and characterising a VNTR at the MIR137 schizophrenia-associated locus (Warburton et al. 2015a, Warburton et al. 2015b) led us to apply similar approaches to the MIR941/DNAJC5 locus in order to study VNTRs at additional loci containing brain-expressed miRNAs and genes involved in CNS function. Studying a VNTR at the DNAJC5/MIR941 locus then led us to expand our work to address the role of other classes of repetitive elements, SVAs and LINE-1s, with the aim of adding to our understanding of the recent retrotransposon-mediated primate and human evolution of genes involved in CNS pathways. The work presented in this thesis provided the following observations:

- Seven ECRs at the MIR137 locus are active regulators of gene expression *in vitro*. Three of the ECRs are predicted to be active in embryonic cells, while three are predicted regulators in multiple brain regions, with five having schizophrenia GWAS SNPs either within or adjacent to their sequence. This may suggest that variation around these highly conserved regulatory elements may modulate their function in a way that could contribute to the schizophrenia association of this locus, with the potential to contribute to both the neurodevelopmental and adult risk for schizophrenia depending on the timepoint at which the ECRs are active.

- The lncRNA, EU358092, lies within the region of schizophrenia GWAS association at the MIR137 locus and has many similarities to MIR137, including degree of conservation, and near identical expression patterns across primates and the human brain, and in response to drug treatment *in vitro*. This would support our hypothesis of co-expression and co-regulation of the two non-coding RNAs at this locus. EU358092 is deregulated in the DL-PFC of individuals with schizophrenia compared to controls, which would be consistent with a role for this lncRNA in adding to the GWAS association at this locus.
- MIR137, REST, and EZH2 form a regulatory network which functions to regulate both the expression of one another, as well as downstream CNS- and schizophrenia-associated gene sets. MIR137, REST, and EZH2 are deregulated in the DL-PFC of individuals with schizophrenia compared to controls, which enrichment analysis suggests could alter the regulation of gene sets involved in behaviour and synaptic transmission.
- The miRNA, MIR941, lies within a VNTR, of which there are four common alleles in a control cohort. Different genotypes result in different copy number of the miRNA, with two genotypes being identified as unique to a schizophrenia cohort in this study. The MIR941 VNTR is also embedded in DNAJC5, and may be able to regulate gene expression in an allele-specific manner from this location. DNAJC5 is one of the most significantly deregulated genes in the schizophrenia DL-PFC, with the VNTR having the potential to act as a regulator and the first exon of a shorter DNAJC5 transcript.

- SVAs are clustered at multiple ZNF zinc finger gene loci on chromosomes 4, 7, and 19. ZNFs are the largest class of transcription factors in the genome, with likely SVA-mediated alterations in their regulation expected to have contributed to species-specific regulation of numerous gene networks.
- The three most recent LINE-1 subfamilies are predominantly clustered on the X chromosome, with enrichment analysis demonstrating recent LINE-1 mediated remodelling of genes involved in GABA and glutamate signalling. This may suggest a mechanism in LINE-1 mediated risk for CNS conditions.

This work highlights potential mechanisms involved in the GWAS association and regulation at the MIR137 locus, as well as the role of repetitive DNA and retrotransposons in altering the structure or regulation around CNS-expressed genes, throughout evolution and variably across the human population. Experiments to determine the validity of the mechanisms suggested in this thesis could include functional assessment of schizophrenia risk SNPs on ECR function at the MIR137 locus, through luciferase assay of ECRs with different SNP alleles, or through the identification of eQTL SNPs within these regions. Similar studies could be used to assess the regulatory capacity of the MIR941/DNAJC5 VNTR, and the potential for allele-specific regulatory effects based on VNTR genotype. This work on the MIR941/DNAJC5 locus is currently being carried out by Ana Illera-Lopez. Additional studies should determine the potential role of the MIR941/DNAJC5 VNTR to alter the sequence and structure of alternate DNAJC5 transcripts, in a similar manner to that demonstrated for the MIR137 VNTR (Mamdani et al. 2013). Dr. Abigail Savage, Kimberley Billingsley, Emma Price, Ben Middlehurst, Jack Marshall, and Ana Illera-Lopez are currently extending the hypotheses around retrotransposons as regulators of CNS gene expression using techniques including retrotransposon capture

sequencing (RC-seq) in Parkinson's disease brain samples, as well as eQTL, haplotype, and enrichment analyses, in addition to the continued use of the techniques utilised in this thesis.

A key theme throughout each chapter has been genetic variation, and how such variants could alter regulation and expression of genes involved in healthy brain function; from schizophrenia-associated SNPs in the MIR137 and EU358092 ECRs, to variable copy number at the MIR941 locus VNTR, to retrotransposon insertion polymorphisms within glutamate genes (Figure 7.1).

ECRs, VNTRs, and retrotransposons are all known to bind transcription factors, and have been shown to alter regulatory activity in an allele dependent manner. For example, previous studies have demonstrated the ability of multiple ECRs to modulate the expression of reporter genes, both *in vitro* and *in vivo*. Paredes et al. identified and characterised an ECR at the dopamine receptor D4 (DRD4) gene which supported expression of a reporter gene in rat frontal cortex primary cultures (Paredes et al. 2011). This work identified a binding site for the transcription factor, SP1, overexpression of which was shown to significantly decrease the expression supported by the ECR. Further, Davidson et al. and Hing et al. demonstrated two ECRs upstream of the galanin and BDNF promoters which were shown to have transcriptional regulatory properties in reporter gene experiments. Both ECRs were found to contain SNPs, with two SNPs in the galanin ECR in LD with a SNP for major depressive disorder (MDD), and a SNP in the BDNF ECR which had previously been linked to mood disorders and cognitive decline. The regulatory activity of each ECR was shown to be significantly altered based on the genotype of the SNPs, which would likely have implications in CNS gene expression in mental health conditions (Davidson et al. 2011, Hing et al. 2012).

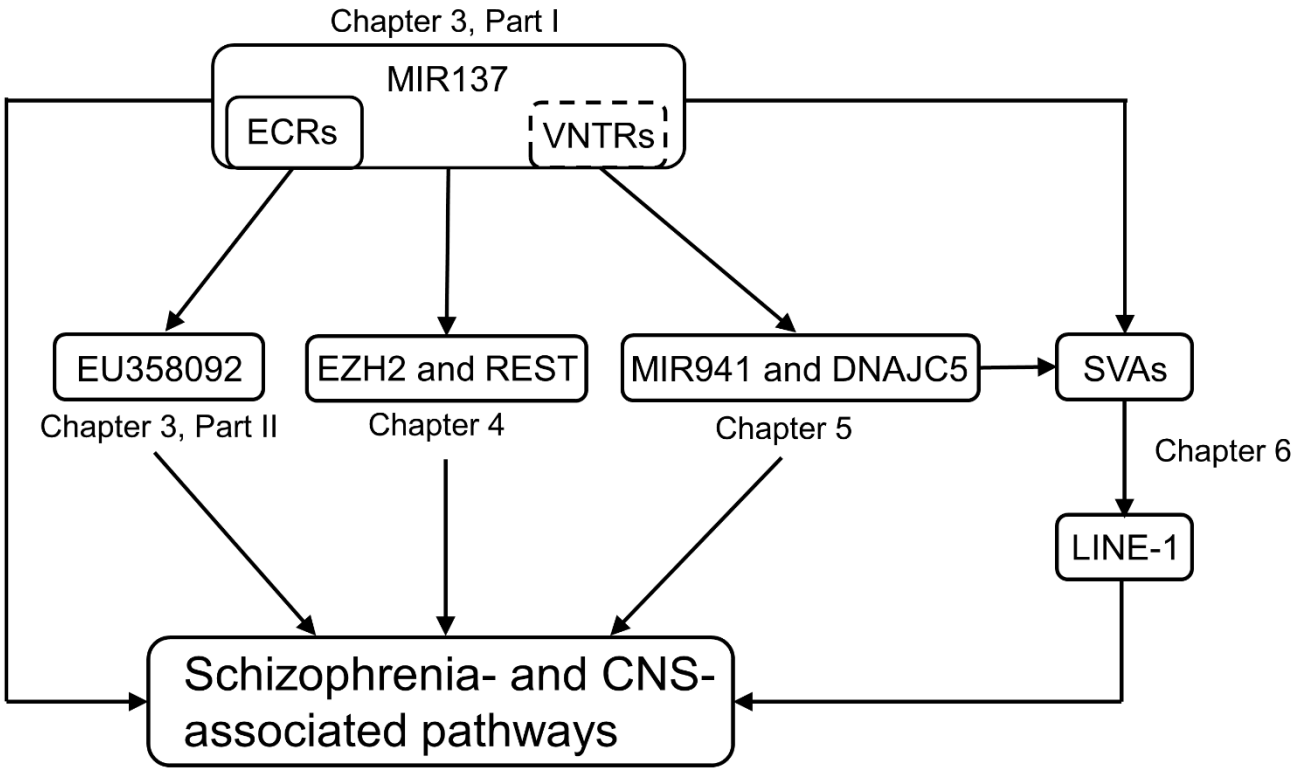


Figure 7.1 An overview of thesis structure and related chapters.

The work in each chapter of this thesis provided direction and inspiration for consecutive chapters as displayed above. Work on ECRs at the MIR137 locus in Chapter 3 Part I led to the identification of the lncRNA EU358092, the characterisation of which then became the focus of Chapter 3 Part II. Previous work by the group on the REST-mediated regulation of MIR137 led to an extended analysis of regulatory pathways in Chapter 4, which involved MIR137, REST, and EZH2. Again, previous work by the group on the MIR137 VNTR directed our interest to the MIR941 VNTR, and we broadened our focus to DNAJC5 and other members of the DNAJ family in the data presented in Chapter 5. As SVA elements also contain VNTRs, we extended our analysis to additional repetitive elements, including the retrotransposon classes, SVA and LINE-1, in Chapter 6. Each results chapter in this thesis converges on an overall theme of schizophrenia- and CNS-associated gene regulation, and genetic variation which may influence this.

The dashed box represents work that was carried out by others in the group and is not presented in this thesis.

Such work demonstrates that ECRs can respond to changes in transcription factor levels in the cell, and regulate reporter gene expression in an allele-specific manner based on the genotype of SNPs within their sequence. This has been demonstrated for SNPs within ECRs at the galanin and BDNF loci, whereby the alternate alleles in the galanin ECR resulted in a 40% decrease in reporter gene activity over the reference alleles in primary rat hypothalamic neurons, and the C/T SNP in the BDNF BE5.2 ECR demonstrated allele-specific reporter gene expression in response to treatment with potassium chloride in primary rat neuronal cultures from the hippocampus, cortex, and amygdala (Davidson et al. 2011, Hing et al. 2012).

In Chapter 3, we identified seven ECRs at the MIR137 locus which were shown to have transcriptional regulatory activity in the SH-SY5Y neuroblastoma cell line model (Figure 3.7) (Gianfrancesco et al. 2016b). Publicly available chromatin state and histone modification data demonstrated that three of these ECRs were predicted to be active in embryonic cells and a further three were predicted regulators in *ex vivo* samples from multiple human brain regions (Figure 3.4 and 3.5). Five of the seven ECRs had schizophrenia GWAS SNPs within or adjacent to their sequence (Figure 3.2 and 3.3). Specific genotypes of these schizophrenia GWAS SNPs may have the potential to modulate activity of the ECRs which could alter transcriptional regulation at the MIR137 locus, perhaps contributing to the schizophrenia GWAS association at this region. This was not tested in the work presented in this thesis, however, allele-specific regulation of gene expression based on SNP genotype of ECRs has been demonstrated in work by Davidson et al. and Hing et al. as described above (Davidson et al. 2011, Hing et al. 2012). Similarly, work by Duan et al. identified a rare enhancer SNP, 1:g.98515539A>T, at the MIR137 locus which was found to be associated with

schizophrenia and bipolar disorder, with the minor allele reducing reporter gene expression by over 50% in a neuroblastoma cell line model (Duan et al. 2014).

Given that half of the ECRs were predicted to be functional during embryonic development, with the other half active in the adult brain, schizophrenia risk SNPs at different ECRs could potentially contribute to both the neurodevelopmental or adult risk factors for schizophrenia depending on the time point at which they were active. Further ECRs in the extended schizophrenia GWAS locus downstream of MIR137 identified a brain-expressed lncRNA, EU358092 (Figure 3.9), with an ECR 2.6 kb upstream which was shown to be a positive regulator of reporter gene expression and also contained a schizophrenia GWAS SNP (Figures 3.16 and 3.10) (Gianfrancesco et al. 2017). EU358092 further contained multiple schizophrenia GWAS SNPs across its sequence (Figure 3.11a), with a SNP directly at the annotated splice site, which could have the potential to negatively affect splicing of this lncRNA. While EU358092 was found to be deregulated in the DL-PFC of individuals with schizophrenia in RNA-seq data from the Lieber Institute, SNP genotype data was unfortunately unavailable for the individuals in this study, and we were thus unable to correlate the genotype of the EU 2 ECR GWAS SNP with expression of the lncRNA, though the literature may suggest a differential regulatory effect based on SNP genotype with the potential for implications in schizophrenia.

Warburton et al. demonstrated that a schizophrenia GWAS SNP at the MIR137 internal promoter, rs2660304, modulated reporter gene expression from this promoter in an allele dependent manner based on the genotype of the SNP, with the major A allele significantly reducing the expression supported by this promoter (Warburton et al. 2015a). In Chapter 4, we described a regulatory network with relevance to schizophrenia, in which MIR137, REST, and EZH2 regulated each other, as well as

multiple CNS- and schizophrenia-associated gene sets (Figure 4.4). In Figure 4.1, we demonstrated that both REST and EZH2 bind across the MIR137 imir promoter, with the ENCODE ChIP-seq signal for the two transcription factors overlapping the schizophrenia GWAS SNP characterised by Warburton et al. In Chapter 4, we extended this work to demonstrate that the MIR137-REST-EZH2 network was deregulated in the DL-PFC of individuals with a diagnosis of schizophrenia. Enrichment analysis of REST and EZH2 targets identified through interrogation of ENCODE ChIP-seq data revealed that deregulation of the MIR137-REST-EZH2 network was likely to impact gene networks involved in behaviour and synaptic transmission. With ENCODE ChIP-seq signals for REST and EZH2 overlapping rs2660304 at the MIR137 imir promoter, which has a known impact expression from this promoter, the genotype of schizophrenia-associated SNPs at genes involved in this pathway is likely to add another layer of complexity to the association of the MIR137-REST-EZH2 network with schizophrenia. We would theorise that an individual's SNP genotype would likely modulate the effect that deregulation of this network had on the regulation of its target genes, such as in this example for MIR137. This could be tested in a number of ways. One method would involve generating reporter gene constructs containing promoters of schizophrenia-associated genes which are known to interact with EZH2 or REST and which contain SNPs. Multiple constructs containing promoter sequences with different SNP alleles could be assayed for reporter gene expression after over-expression of EZH2 or REST. This could test for potential allele-specific responses to altered expression of the genes outlined in this network, though this would be limited to selected regions. Alternatively, an unbiased approach such as ChIP-exo (a modification of ChIP-seq with an additional exonuclease step) could be employed in order to identify SNPs which result in altered

binding of EZH2 or REST (Perreault and Venters 2016). Variable binding SNPs could then be compared against schizophrenia GWAS SNPs in order to identify potential functional risk alleles, using methods similar to those employed by Gallone et al. to study VDR binding in immune conditions (Gallone et al. 2017).

Chapter 5 investigated a larger type of genetic variation comprising a VNTR at the DNAJC5/MIR941 locus on the tip of the chromosome 20 long arm at 20q13.3. Previous work has consistently demonstrated the regulatory activity of VNTRs, which is often allele-specific depending on the number of tandem repeats in the element. Indeed, the aforementioned MIR137 imir promoter characterised by Warburton et al. contained a VNTR which was variable within the population, with a promoter region containing the longer 12-copy VNTR supporting significantly higher reporter gene expression compared to the 4-copy allele (Warburton et al. 2015b), which may have implications in schizophrenia risk. Similarly, the regulatory properties of VNTRs around other CNS-expressed genes have been characterised with relevance to psychiatric and behavioural conditions, including VNTRs at dopamine receptor and transporter genes, certain genotypes of which have been linked to attention deficit hyperactivity disorder (ADHD), and drug and alcohol misuse (Tabatabaei et al. 2017, Vasconcelos et al. 2015, Stolf et al. 2014, Mallard et al. 2016, Paredes et al. 2013, Guindalini et al. 2006). Further, certain genotypes of the monoamine oxidase A (MAOA) VNTR (Sabol et al. 1998) have been linked to anxiety, aggression, and drug use (Voltas et al. 2015, Zhang et al. 2016, Pickles et al. 2013), and VNTRs at the SLC6A4 serotonin transporter also play a role in regulation with relevance to affective disorders (MacKenzie and Quinn 1999, Klenova et al. 2004, D'Souza et al. 2013). A number of these VNTRs have been shown to interact with environmental factors such as abuse or trauma in a way that modulates an individual’s risk of psychiatric or behavioural

conditions based on their genotype and environmental exposure. Further, some VNTRs, including the serotonin transporter linked polymorphic region (5-HTTLPR) and MAOA VNTR, are known to both interact with each other and with the environment in a so called ‘gene x gene x environment’ mechanism to modulate traits such as aggression after childhood maltreatment in a genotype dependent manner (Zhang et al. 2017).

The VNTR at the DNAJC5/MIR941 locus is of particular interest for a number of reasons, including its inclusion of the human-specific, brain expressed miRNA, MIR941 (Hu et al. 2012), its recent evolution, only being present in higher primate and human genomes, and finally its location within the gene encoding the neuroprotective and highly brain expressed pre-synaptic vesicle protein, DNAJC5 (Figure 5.1a). Interrogation of gene annotation data on the UCSC Genome Browser further demonstrated that the VNTR may act as a primate-specific first exon in an alternative DNAJC5 transcript, AK128776 (Figure 5.1b). From this position, the DNAJC5/MIR941 VNTR could have numerous roles. Firstly, the VNTR could exert regulatory influence on both the full length DNAJC5 and the shorter AK128776 transcripts, potentially in an allele specific manner. Histone modification data from *ex vivo* samples suggested that this region was an active regulator in multiple human brain regions (Figure 5.2b). Secondly, the VNTR may alter the RNA structure of AK128776 based on tandem repeat number. The same effect has been demonstrated for the MIR137 VNTR, with copy number of the repeat altering the structure of transcripts that utilise this region (Mamdani et al. 2013). Finally, we demonstrate that changes in VNTR copy number alters the copy number of the MIR941 miRNA (Figure 5.5).

Little is known about the function of MIR941, however work by Jeffries et al. noted that MIR941 is lost from the microRNA networks (in peripheral blood) of high risk

individuals after they progress to psychosis. This is in contrast to controls and high risk individuals who did not progress to psychosis in the two-year follow up period, whose peripheral blood miRNA networks were not altered. However, loss of MIR941 from the network did not translate to any significant change in expression in the peripheral blood of individuals that progressed to psychosis (Jeffries et al. 2016). Such work does not provide clear evidence to link MIR941 to psychosis, but the findings presented by Jeffries et al. may suggest that MIR941 is a microRNA of interest for further work in order to better elucidate any potential role for this transcript in CNS health and disease.

Genotyping a cohort of 340 controls and 342 individuals with schizophrenia demonstrated two rare genotypes that were specific to the schizophrenia group. These genotypes were found in nine individuals, accounting for 2.63% of the schizophrenia cohort (Figure 5.6 and Table 5.1), though the precise mechanism through which such genotypes may influence risk is unknown. Extending this analysis, we demonstrated that DNAJC5 was one of the most significantly deregulated genes in the schizophrenia DL-PFC (Table 5.3), which could potentially be influenced by VNTR genotype, and could mark the DNAJC5 locus as a region of importance in further understanding schizophrenia biology.

There is evidence in the literature to suggest that VNTRs around CNS-expressed genes can modulate an individual's risk for psychiatric or behavioural conditions in response to trauma or maltreatment. For example, longer alleles of the DRD4 exon three VNTR (seven or more copies) have been associated with increased severity of suicidal ideation in adolescents who have experienced sexual trauma (Doorley et al. 2017), and with more severe post-traumatic stress disorder (PTSD) symptoms in individuals who had survived natural disaster (Dragan and Oniszczenko 2009). Further, in conditions of high prenatal maternal stress, children with the seven-repeat

allele of the DRD4 VNTR were at higher risk of a conduct disorder or oppositional defiant disorder diagnosis (Zohsel et al. 2014). Similarly, meta-analysis demonstrated that the “low activity” MAOA VNTR allele resulted in significantly increased antisocial behaviour in males who had experienced childhood maltreatment and adversity (Byrd and Manuck 2014), with similar childhood experiences and maternal sensitivity being shown to interact with MAOA VNTR genotype to modulate anger and aggressive behaviour (Holz et al. 2016, Zhang et al. 2016, Pickles et al. 2013). For this reason, it is possible that, in addition to the molecular effects of MIR941 VNTR genotype such as altering miRNA copy number and potentially altering RNA structure of AK128776, the MIR941 VNTR genotype could also interact with an individual’s life experience to modulate DNAJC5 and/or AK128776 expression in a way which may alter the risk of psychiatric or behavioural conditions based on an individual’s experience and genotype.

Chapter 6 extended our work from VNTRs to retrotransposons, specifically studying the distribution patterns of SVAs and recent LINE-1 subfamilies across the human genome. Retrotransposons have repeatedly been shown to modulate gene expression, with work by Savage et al. demonstrating the regulatory activity of SVAs both *in vitro* and *in vivo* (Savage et al. 2013b, Savage et al. 2014). As SVAs are composite structures made up of multiple repetitive units, they often display variation in size across the human population. This can be due to different copy numbers of the 5’ $(CCCTCT)_n$ hexamer repeat, the VNTR, or the 3’ poly(A) tail. For example, an SVA upstream of the FUS gene is known to contain polymorphisms in the VNTR, and drives allele-specific expression in reporter gene models (Savage et al. 2014). Others have demonstrated the ability of LINE-1 elements to affect gene expression, with somatic insertion of an intronic LINE-1 element being shown to disrupt gene expression in

induced pluripotent stem cells (Klawitter et al. 2015), and studies in cancer demonstrating that changes in methylation of intronic LINE-1 elements correlated with altered expression of their host oncogene (Zhu et al. 2014, Zhu et al. 2015, Hur et al. 2014). Retrotransposons can also be polymorphic for their presence or absence at specific loci across the human population, with variation in the presence of such large regulatory elements having the potential to significantly alter gene expression between individuals. Studies have demonstrated that between 6-30% of capped human and mouse RNAs originate from transcriptional start sites within retrotransposable elements, and studies in humans suggest that 4% of transcripts originate from a LINE-1 antisense promoter (Faulkner et al. 2009, Criscione et al. 2016). Similarly, Kim et al. have demonstrated the capacity of SVAs to act as novel promoters, identifying 12 cases in which human-specific SVA insertions drove expression of human-specific transcripts originating from within the SVA (Kim and Hahn 2010, Kim and Hahn 2011). Taken together, reference genome SVA and LINE-1 elements are likely to play a role in regulating gene expression, while the addition of an SVA or LINE-1 RIP could significantly alter the expression of nearby genes, potentially even resulting in novel transcripts.

In Chapter 6, we demonstrated that both reference and RIP SVAs are preferentially found at genic regions (Figure 6.15) and particularly cluster at ZNF zinc finger gene loci, (Figure 6.2, 6.3, 6.16) most notably on chromosome 19 (Figure 6.1). On the other hand, recent reference and RIP LINE-1 elements are preferentially found at gene poor regions (Figure 6.17), yet are over-represented at genes involved in GABA and glutamate signalling (Figure 6.8 – 6.14, 6.18). We therefore proposed that environmental exposure or disease states that alter the regulation or expression of recent LINE-1 subfamilies would preferentially affect brain-related gene pathways.

Traumatic experiences are a robust risk factor for mental illness, including schizophrenia, and have been shown to correlate with decreased LINE-1 methylation. For example, Misiak et al. demonstrated that individuals with first episode psychosis who had experienced childhood trauma had decreased LINE-1 methylation in peripheral blood compared to controls or individuals with psychosis but no trauma (Misiak et al. 2015). Rusiecki et al. also showed hypomethylation of LINE-1 in the blood of US military individuals with PTSD post-deployment compared to controls (Rusiecki et al. 2012). The former would be in line with findings suggesting that LINE-1 expression is increased in the brains of individuals with schizophrenia (Bundo et al. 2014). Taken together with our findings, we propose that the trauma- and LINE-mediated risk for schizophrenia and other CNS conditions are likely to work in part through an overlapping mechanism, with stress-induced alterations in LINE-1 methylation and expression preferentially affecting key signalling pathways in the brain that would alter risk for schizophrenia and other brain-related conditions.

The regulatory elements and pathways characterised in this thesis, and the variation within or around them, would correlate clearly with the stress-vulnerability model of schizophrenia (Figure 1.7), which suggests a ‘gene x environment’ mechanism in predicting risk. Indeed, studies outlined in the introductory chapter (Section 1.3) have provided evidence to link early life stress to mental health conditions through a range of mechanisms, including modifying the trajectory of brain development or of molecular pathways in ways which would sensitise the individual to further stress, with stress sensitivity being a predicting factor for psychosis.

Some stress-related risk factors for mental health conditions (including schizophrenia) may be unclear in terms of the causal factors, or with regard to the direction of causal effect. For example, relative poverty has been associated with mental ill health

(Bjorkenstam et al. 2017, Loch et al. 2017, Rotenberg, Tuck and McKenzie 2017), though there are multiple potential factors within this that may influence risk, such as malnutrition (or maternal malnutrition), living in an urban environment, living in unstable housing, and others (Heinz, Deserno and Reininghaus 2013, Davis et al. 2016). Further, it has been suggested that experiencing mental illness such as schizophrenia may affect an individual's ability to work, and could in turn lead to reduced social mobility and potentially increased poverty. This is the basis of the 'social drift' model (Fox 1990, Sariaslan et al. 2016).

While it is difficult to control for confounding effects in human studies, evidence for stress as a causative risk for phenotypes associated with mental health conditions has been demonstrated in animal studies (Li et al. 2017, Sun et al. 2017, Weinstock 2017, Scott and Tamminga 2018). Indeed, that chronic mild stress (CMS), maternal stress or deprivation, and social isolation are used to model depression and schizophrenia in rodents makes clear the effects of stress on brain development and on behaviours that are associated with mental health conditions (Willner 2017, Jones, Watson and Fone 2011).

The evidence in this section details the roles of genetic variation - in the form of SNPs, VNTR copy number, and retrotransposon polymorphisms - and their association with altered gene regulation and altered risk for a multitude of diagnoses and behavioural traits in response to stress or trauma.

During the course of my PhD, I have also worked alongside Dr. Peter Taylor from the University of Manchester to compile and co-edit a book entitled *'Personal Experiences of Psychological Therapy for Psychosis'*, to be published through Routledge as part of the ISPS (International Society for Psychological and Social Approaches to Psychosis)

book series. The personal stories contained within this book all begin with the authors' experience of stress, trauma, and hardship, which they each feel was a large contributor to their psychosis. While the research described in this thesis removes the personal aspect of the experience of psychosis, the narratives presented within the book and the thesis complement each other in supporting a 'G x E' model of mental illness, whereby childhood trauma and/or a period of high stress in adolescence or adulthood could push certain schizophrenia-associated molecular pathways out of balance, likely in ways that would be modulated by an individual's genotype. We would hypothesise that the transient nature of the molecular response to stress may provide one explanation for the episodic nature of psychosis, and may explain why many cases of psychosis occur after highly stressful experiences, or why many people recover fully after a single episode of illness. Thinking in this 'G x E' way about psychosis would also advocate for the role of psychological therapy or other social support as ways to help a person cope with or resolve the stress in their life that may be preventing them from achieving recovery. Many contributors to the book described how therapy had allowed them to better understand and cope with their life experiences, which led to positive change in their mental health. For this reason, it could be suggested that such support may be of equal importance and benefit to the individual in the long term compared to treatments such as medication that focus on changing the 'G', while leaving the individual with a stressful and unresolved 'E' that could continue to negatively affect their health. Recent evidence is beginning to show that psychological therapy for individuals with experiences of psychosis is both safe and effective, with long term benefits even for individuals with medication resistant symptoms (Peters et al. 2015, Burns, Erickson and Brenner 2014, Mehl, Werner and Lincoln 2015, Hazell et al. 2016). Further, studies into the biological effects of therapy

have demonstrated that cognitive behavioural therapy can normalise the connectivity between brain areas that are associated with social threat in individuals with psychosis in ways which correlate with symptom improvement, and which also predict long-term recovery eight years after finishing therapy (Mason et al. 2016, Mason et al. 2017). Additionally, studies of psychological therapy for panic disorder and anxiety disorder have demonstrated that individuals who respond to therapy show changes in methylation over key genes such as MAOA and the serotonin transporter, SLC6A4 (Ziegler et al. 2016, Roberts et al. 2014).

The results in this thesis characterise regulatory elements and gene networks around CNS- and schizophrenia-associated genes which can influence gene expression, likely in ways that would allow the cell to respond to its environment. We further identify variants in these regulatory elements that may cause individuals to respond differently on the molecular level to the same environmental challenge, which could modulate risk for schizophrenia in a genotype-dependent manner. Experiences of stress and trauma are known to induce epigenetic changes, often modulated by genotype, that can alter such molecular pathways and put an individual at greater risk of mental illness. However, regulation within the cell is dynamic and ever-changing, in a constant state of interaction and modulation in response to the environment. It is for this reason that there is hope for recovery, and hope in the knowledge that this process also works in reverse, with positive change in a person's life having the ability to cause positive change at the molecular level.

Chapter 8

Appendix

Name	Primers (5' to 3')	Target	Application	PCR conditions
MIR137 ECR 1	Fwd: GCAGTGGCTGTAAGATGAGGA Rev: AGAGGCCTGGAGTCTGTGAC	chr1:98499831-98500934	Cloning into pGL3P	Phusion Polymerase 98 °C for 30 s 98 °C for 10 s Variable °C for 30 s 72 °C for 30 s 72 °C for 5 min 4 °C hold 25 cycles
MIR137 ECR 2	Fwd: CCCCATGATGTTCTCATACCA Rev: TACAGCCACTGCAAATACGG	chr1:98500923-98502814	Cloning into pGL3P	
MIR137 ECR 3	Fwd: AGCTCTTACGCGTGCTAGCTGCACTTTGCATTCCTC Rev: AGATCGCAGATCTCGAGCTCACACTTCCTAACTGGT	chr1:98525381-98525895	Cloning into pGL3P	
MIR137 ECR 4	Fwd: AGCTCTTACGCGTGCTAGTGCCCTTGTCTAATGAA Rev: AGATCGCAGATCTCGAGCATTCAGGACTCTAGTCT	chr1:98538705-98540267	Cloning into pGL3P	
MIR137 ECR 5	Fwd: AGCTCTTACGCGTGCTAGTGCCCTTGTCTAATGAA Rev: AGATCGCAGATCTCGAGCATTCAGGACTCTAGTCT	chr1:98552809-98554083	Cloning into pGL3P	
MIR137 ECR 6	Fwd: AGCTCTTACGCGTGCTAGAGAAGAGGATTTGTGGGCTAC Rev: AGATCGCAGATCTCGAGGGCTTGGGATACCTGACAATTAGCAAC	chr1:98567339-98567854	Cloning into pGL3P	
MIR137 ECR 7	Fwd: CGAGCTCTTACGCGTGCTAGACATTCTTGATTTGCATAA Rev: AGATCGCAGATCTCGAGTGCTTCAGTGTAACTACTG	chr1:98592252-98592661	Cloning into pGL3P	
EU 1 ECR	Fwd: AGCTCTTACGCGTGCTAGTGTAGCGAACCAACTGT Rev: GCAGATCGCAGATCTCGAGTCAAGGCTTATTGTCTTTGG	chr1:98398666-98399594	Cloning into pGL3P	
EU 2 ECR	Fwd: AGCTCTTACGCGTGCTAGAGGCTTCAATGAAAAGAG Rev: AGATCGCAGATCTCGAGTCATGTGTAATGTCCTGG	chr1:98395661-98396399	Cloning into pGL3P	
pGL3P sequencing	Fwd: CTTTATGTTTTTGGCGTGTTCC Rev: CTAGCAAAATAGGCTGTCCC	pGL3P vector	Sequencing	N/A
EU358092 mRNA	Fwd: GGTGGGAATTGGGTCTCACA Rev: CTTAACACAGCGTTGTCAAGGTTCATC	chr1:98399147-98407177	RT-PCR	Taq Polymerase 95 °C for 5 min 95 °C for 30 s 60 °C for 30 s 72 °C for 30 s 72 °C for 10 min 4 °C hold 35 cycles
Beta actin	Fwd: CACCCCTACAATGAGCTGCGTGTG Rev: ATAGCACAGCCTGGATAGCAACGTAC	chr7:5566779-5570232	RT-PCR	As above
MIR941 VNTR	Fwd: ACGTGTCCGGGGAGAGGACG Rev: CCCGGTCCGACGCAGGAC	chr20:62550716-62551708	PCR	ReddyMix 95 °C for 6 min 95 °C for 20 s 61 °C for 20 s 72 °C for 30 s 72 °C for 5 min 4 °C hold 35 cycles

Chapter 9

References

Abrajano, J. J., I. A. Qureshi, S. Gokhan, D. Zheng, A. Bergman & M. F. Mehler (2009) REST and CoREST modulate neuronal subtype specification, maturation and maintenance. *PLoS One,* 4**,** e7936.

Abrusán, G., J. Giordano & P. E. Warburton. 2008. Analysis of Transposon Interruptions Suggests Selection for L1 Elements on the X Chromosome. In *PLoS Genet*.

Ahrendt, E., B. Kyle, A. P. Braun & J. E. Braun (2014) Cysteine string protein limits expression of the large conductance, calcium-activated K(+) (BK) channel. *PLoS One,* 9**,** e86586.

Alarcon, M., B. S. Abrahams, J. L. Stone, J. A. Duvall, J. V. Perederiy, J. M. Bomar, J. Sebat, M. Wigler, C. L. Martin, D. H. Ledbetter, S. F. Nelson, R. M. Cantor & D. H. Geschwind (2008) Linkage, association, and gene-expression analyses identify CNTNAP2 as an autism-susceptibility gene. *Am J Hum Genet,* 82**,** 150-9.

Alarcón, M., R. M. Cantor, J. Liu, T. C. Gilliam & D. H. Geschwind (2002) Evidence for a Language Quantitative Trait Locus on Chromosome 7q in Multiplex Autism Families. *Am J Hum Genet,* 70**,** 60-71.

Aloia, L., B. Di Stefano & L. Di Croce (2013) Polycomb complexes in stem cells and embryonic development. *Development,* 140**,** 2525-34.

American Psychiatric Association. 2013. *Diagnostic and statistical manual of mental disorders (5th ed.)*. Arlington, VA: American Psychiatric Publishing.

Bailey, J. A., L. Carrel, A. Chakravarti & E. E. Eichler (2000) Molecular evidence for a relationship between LINE-1 elements and X chromosome inactivation: the Lyon repeat hypothesis. *Proc Natl Acad Sci U S A,* 97**,** 6634-9.

Bala Tannan, N., M. Brahmachary, P. Garg, C. Borel, R. Alnefaie, C. T. Watson, N. S. Thomas & A. J. Sharp (2014) DNA methylation profiling in X;autosome translocations supports a role for L1 repeats in the spread of X chromosome inactivation. *Hum Mol Genet,* 23**,** 1224-36.

Ballas, N., C. Grunseich, D. D. Lu, J. C. Speh & G. Mandel (2005) REST and its corepressors mediate plasticity of neuronal gene chromatin throughout neurogenesis. *Cell,* 121**,** 645-57.

Bantysh, O. B. & A. A. Buzdin (2009) Novel family of human transposable elements formed due to fusion of the first exon of gene MAST2 with retrotransposon SVA. *Biochemistry (Mosc),* 74**,** 1393-9.

Barrett, J. C., B. Fry, J. Maller & M. J. Daly (2005) Haploview: analysis and visualization of LD and haplotype maps. *Bioinformatics,* 21**,** 263-5.

Barry, G., J. A. Briggs, D. P. Vanichkina, E. M. Poth, N. J. Beveridge, V. S. Ratnu, S. P. Nayler, K. Nones, J. Hu, T. W. Bredy, S. Nakagawa, F. Rigo, R. J. Taft, M. J. Cairns, S. Blackshaw, E. J. Wolvetang & J. S. Mattick (2014) The long non-coding RNA Gomafu is acutely regulated in response to neuronal activation and involved in schizophrenia-associated alternative splicing. *Mol Psychiatry,* 19**,** 486-94.

Baskerville, S. & D. P. Bartel (2005) Microarray profiling of microRNAs reveals frequent coexpression with neighboring miRNAs and host genes. *RNA,* 11**,** 241-7.

Bejerano, G., M. Pheasant, I. Makunin, S. Stephen, W. J. Kent, J. S. Mattick & D. Haussler (2004) Ultraconserved elements in the human genome. *Science,* 304**,** 1321-5.

Belancio, V. P., D. J. Hedges & P. Deininger (2006) LINE-1 RNA splicing and influences on mammalian gene expression. *Nucleic Acids Res,* 34**,** 1512-21.

Belzeaux, R., A. Bergon, V. Jeanjean, B. Loriod, C. Formisano-Treziny, L. Verrier, A. Loundou, K. Baumstarck-Barrau, L. Boyer, V. Gall, J. Gabert, C. Nguyen, J. M. Azorin, J. Naudin & E. C. Ibrahim (2012) Responder and nonresponder patients exhibit different peripheral transcriptional signatures during major depressive episode. *Transl Psychiatry,* 2**,** e185.

Benitez, B. A., D. Alvarado, Y. Cai, K. Mayo, S. Chakraverty, J. Norton, J. C. Morris, M. S. Sands, A. Goate & C. Cruchaga (2011) Exome-sequencing confirms DNAJC5 mutations as cause of adult neuronal ceroid-lipofuscinosis. *PLoS One,* 6**,** e26741.

Bennett, E. A., H. Keller, R. E. Mills, S. Schmidt, J. V. Moran, O. Weichenrieder & S. E. Devine (2008) Active Alu retrotransposons in the human genome. *Genome Res,* 18**,** 1875-83.

Bentall, R. P., S. Wickham, M. Shevlin & F. Varese (2012) Do Specific Early-Life Adversities Lead to Specific Symptoms of Psychosis? A Study from the 2007 The Adult Psychiatric Morbidity Survey. *Schizophr Bull,* 38**,** 734-40.

Beveridge, N. J. & M. J. Cairns (2012) MicroRNA dysregulation in schizophrenia. *Neurobiol Dis,* 46**,** 263-71.

Bilgin Sonay, T., T. Carvalho, M. D. Robinson, M. P. Greminger, M. Krutzen, D. Comas, G. Highnam, D. Mittelman, A. Sharp, T. Marques-Bonet & A. Wagner (2015) Tandem repeat variation in human and great ape populations and its impact on gene expression divergence. *Genome Res,* 25**,** 1591-9.

Bilic, P., V. Jukic, M. Vilibic, A. Savic & N. Bozina (2014) Treatment-resistant schizophrenia and DAT and SERT polymorphisms. *Gene,* 543**,** 125-32.

Billingsley, K. J., M. Manca, O. Gianfrancesco, D. A. Collier, H. Sharp, V. J. Bubb & J. P. Quinn (2018) Regulatory characterisation of the schizophrenia-associated CACNA1C proximal promoter and the potential role for the transcription factor EZH2 in schizophrenia aetiology. *Schizophrenia Research*.

Birtle, Z., L. Goodstadt & C. Ponting (2005) Duplication and positive selection among hominin-specific PRAME genes. *BMC Genomics,* 6**,** 120.

Bjorkenstam, E., S. Cheng, B. Burstrom, A. R. Pebley, C. Bjorkenstam & K. Kosidou (2017) Association between income trajectories in childhood and psychiatric disorder: a Swedish population-based study. *J Epidemiol Community Health,* 71**,** 648-654.

Blackledge, N. P., N. R. Rose & R. J. Klose (2015) Targeting Polycomb systems to regulate gene expression: modifications to a complex story. *Nat Rev Mol Cell Biol,* 16**,** 643-9.

Blake, J. A. & M. R. Ziman (2014) Pax genes: regulators of lineage specification and progenitor cell maintenance. *Development,* 141**,** 737-51.

Boettger, M. K., D. Grossmann & K. J. Bar (2013) Increased cold and heat pain thresholds influence the thermal grill illusion in schizophrenia. *Eur J Pain,* 17**,** 200-9.

Borchert, G. M., W. Lanier & B. L. Davidson (2006) RNA polymerase III transcribes human microRNAs. *Nat Struct Mol Biol,* 13**,** 1097-101.

Bouttier, M., D. Laperriere, B. Memari, J. Mangiapane, A. Fiore, E. Mitchell, M. Verway, M. A. Behr, R. Sladek, L. B. Barreiro, S. Mader & J. H. White (2016) Alu repeats as transcriptional regulatory platforms in macrophage responses to M. tuberculosis infection. *Nucleic Acids Res,* 44**,** 10571-10587.

Bozidis, P., T. Hyphantis, C. Mantas, M. Sotiropoulou, N. Antypa, E. Andreoulakis, A. Serretti, V. Mavreas & K. Antoniou (2014) HSP70 polymorphisms in first psychotic episode drug-naive schizophrenic patients. *Life Sci,* 100**,** 133-7.

Bramness, J. G. & E. B. Rognli (2016) Psychosis induced by amphetamines. *Curr Opin Psychiatry,* 29**,** 236-41.

Brattas, P. L., M. E. Jonsson, L. Fasching, J. Nelander Wahlestedt, M. Shahsavani, R. Falk, A. Falk, P. Jern, M. Parmar & J. Jakobsson (2017) TRIM28 Controls a Gene Regulatory Network Based on Endogenous Retroviruses in Human Neural Progenitor Cells. *Cell Rep,* 18**,** 1-11.

Breen, G., D. Collier, I. Craig & J. Quinn (2008) Variable number tandem repeats as agents of functional regulation in the genome. *IEEE Eng Med Biol Mag,* 27**,** 103-4, 108.

Briggs, J. A., E. J. Wolvetang, J. S. Mattick, J. L. Rinn & G. Barry (2015) Mechanisms of Long Non-coding RNAs in Mammalian Nervous System Development, Plasticity, Disease, and Evolution. *Neuron,* 88**,** 861-77.

Brotons, O., O. G. O'Daly, C. Guindalini, M. Howard, J. Bubb, G. Barker, J. Dalton, J. Quinn, R. M. Murray, G. Breen & S. S. Shergill (2011) Modulation of orbitofrontal response to amphetamine by a functional variant of DAT1 and in vitro confirmation. *Mol Psychiatry,* 16**,** 124-6.

Brouha, B., J. Schustak, R. M. Badge, S. Lutz-Prigge, A. H. Farley, J. V. Moran & H. H. Kazazian. 2003. Hot L1s account for the bulk of retrotransposition in the human population. In *Proc Natl Acad Sci U S A*, 5280-5.

Bundo, M., M. Toyoshima, Y. Okada, W. Akamatsu, J. Ueda, T. Nemoto-Miyauchi, F. Sunaga, M. Toritsuka, D. Ikawa, A. Kakita, M. Kato, K. Kasai, T. Kishimoto, H. Nawa, H. Okano, T. Yoshikawa, T. Kato & K. Iwamoto (2014) Increased l1 retrotransposition in the neuronal genome in schizophrenia. *Neuron,* 81**,** 306-13.

Burgoyne, R. D. & A. Morgan (2015) Cysteine string protein (CSP) and its role in preventing neurodegeneration. *Semin Cell Dev Biol,* 40**,** 153-9.

Burns, A. M., D. H. Erickson & C. A. Brenner (2014) Cognitive-behavioral therapy for medication-resistant psychosis: a meta-analytic review. *Psychiatr Serv,* 65**,** 874-80.

Byrd, A. L. & S. B. Manuck (2014) MAOA, childhood maltreatment, and antisocial behavior: meta-analysis of a gene-environment interaction. *Biol Psychiatry,* 75**,** 9-17.

Cabili, M. N., C. Trapnell, L. Goff, M. Koziol, B. Tazon-Vega, A. Regev & J. L. Rinn (2011) Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. *Genes Dev,* 25**,** 1915-27.

Cadieux-Dion, M., E. Andermann, P. Lachance-Touchette, O. Ansorge, C. Meloche, A. Barnabe, R. I. Kuzniecky, F. Andermann, E. Faught, S. Leonberg, J. A. Damiano, S. F. Berkovic, G. A. Rouleau & P. Cossette (2013) Recurrent mutations in DNAJC5 cause autosomal dominant Kufs disease. *Clin Genet,* 83**,** 571-5.

Callinan, P. A. & M. A. Batzer (2006) Retrotransposable elements and human disease. *Genome Dyn,* 1**,** 104-15.

Camkurt, M. A., F. Karababa, M. E. Erdal, H. Bayazit, S. B. Kandemir, M. E. Ay, H. Kandemir, O. I. Ay, E. Cicek, S. Selek & B. Tasdelen (2016) Investigation of Dysregulation of Several MicroRNAs in Peripheral Blood of Schizophrenia Patients. *Clin Psychopharmacol Neurosci,* 14**,** 256-60.

Cao, D. D., L. Li & W. Y. Chan. 2016. MicroRNAs: Key Regulators in the Central Nervous System and Their Implication in Neurological Diseases. In *Int J Mol Sci*.

Castro-Diaz, N., G. Ecco, A. Coluccio, A. Kapopoulou, B. Yazdanpanah, M. Friedli, J. Duc, S. M. Jang, P. Turelli & D. Trono (2014) Evolutionally dynamic L1 regulation in embryonic stem cells. *Genes Dev,* 28**,** 1397-409.

Chang, T. C., Y. Yang, H. Yasue, A. K. Bharti, E. F. Retzel & W. S. Liu (2011) The Expansion of the PRAME Gene Family in Eutheria. *PLoS One,* 6.

Chen, E. Y., C. M. Tan, Y. Kou, Q. Duan, Z. Wang, G. V. Meirelles, N. R. Clark & A. Ma'ayan (2013) Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool. *BMC Bioinformatics,* 14**,** 128.

Chen, H. J., J. C. Mitchell, S. Novoselov, J. Miller, A. L. Nishimura, E. L. Scotter, C. A. Vance, M. E. Cheetham & C. E. Shaw (2016a) The heat shock response plays an important role in TDP-43 clearance: evidence for dysfunction in amyotrophic lateral sclerosis. *Brain,* 139**,** 1417-32.

Chen, L. L. (2016) Linking Long Noncoding RNA Localization and Function. *Trends Biochem Sci,* 41**,** 761-72.

Chen, S., X. Sun, W. Niu, L. Kong, M. He, W. Li, A. Zhong, J. Lu & L. Zhang (2016b) Aberrant Expression of Long Non-Coding RNAs in Schizophrenia Patients. *Med Sci Monit,* 22**,** 3340-51.

Chen, X., J. Wang, H. Shen, J. Lu, C. Li, D. N. Hu, X. D. Dong, D. Yan & L. Tu (2011a) Epigenetics, microRNAs, and carcinogenesis: functional role of microRNA-137 in uveal melanoma. *Invest Ophthalmol Vis Sci,* 52**,** 1193-9.

Chen, Y. T., R. Chiu, P. Lee, D. Beneck, B. Jin & L. J. Old (2011b) Chromosome X-encoded cancer/testis antigens show distinctive expression patterns in developing gonads and in testicular seminoma. *Hum Reprod,* 26**,** 3232-43.

Chendrimada, T. P., R. I. Gregory, E. Kumaraswamy, J. Norman, N. Cooch, K. Nishikura & R. Shiekhattar (2005) TRBP recruits the Dicer complex to Ago2 for microRNA processing and gene silencing. *Nature,* 436**,** 740-4.

Chow, J. C., C. Ciaudo, M. J. Fazzari, N. Mise, N. Servant, J. L. Glass, M. Attreed, P. Avner, A. Wutz, E. Barillot, J. M. Greally, O. Voinnet & E. Heard (2010) LINE-1 activity in facultative heterochromatin formation during X chromosome inactivation. *Cell,* 141**,** 956-69.

Chueh, A. C., E. L. Northrop, K. H. Brettingham-Moore, K. H. Choo & L. H. Wong (2009) LINE retrotransposon RNA is an essential structural and functional epigenetic component of a core neocentromeric chromatin. *PLoS Genet,* 5**,** e1000354.

Chung, M. K., S. J. Jung & S. B. Oh (2011) Role of TRP channels in pain sensation. *Adv Exp Med Biol,* 704**,** 615-36.

Collins, A. L., Y. Kim, R. J. Bloom, S. N. Kelada, P. Sethupathy & P. F. Sullivan (2014) Transcriptional targets of the schizophrenia risk gene MIR137. *Transl Psychiatry,* 4**,** e404.

Cordaux, R. & M. A. Batzer (2009) The impact of retrotransposons on human genome evolution. *Nat Rev Genet,* 10**,** 691-703.

Cosgrove, D., D. Harold, O. Mothersill, R. Anney, M. J. Hill, N. J. Bray, G. Blokland, T. Petryshen, C. Wellcome Trust Case Control, A. Richards, K. Mantripragada, M. Owen, M. C. O'Donovan, M. Gill, A. Corvin, D. W. Morris & G. Donohoe (2017) MiR-137-derived polygenic risk: effects on cognitive performance in patients with schizophrenia and controls. *Transl Psychiatry,* 7**,** e1012.

Coskun, V., R. Tsoa & Y. E. Sun (2012) Epigenetic regulation of stem cells differentiating along the neural lineage. *Curr Opin Neurobiol,* 22**,** 762-7.

Coulson, J. M., S. I. Ahmed, J. P. Quinn & P. J. Woll (2003) Detection of small cell lung cancer by RT-PCR for neuropeptides, neuropeptide receptors, or a splice variant of the neuron restrictive silencer factor. *Methods Mol Med,* 75**,** 335-52.

Coulson, J. M., J. L. Edgson, P. J. Woll & J. P. Quinn (2000) A splice variant of the neuron-restrictive silencer factor repressor is expressed in small cell lung cancer: a potential role in derepression of neuroendocrine genes and a useful clinical marker. *Cancer Res,* 60**,** 1840-4.

Criscione, S. W., N. Theodosakis, G. Micevic, T. C. Cornish, K. H. Burns, N. Neretti & N. Rodic (2016) Genome-wide characterization of human L1 antisense promoter-driven transcripts. *BMC Genomics,* 17**,** 463.

Cristobal-Narvaez, P., T. Sheinbaum, S. Ballespi, M. Mitjavila, I. Myin-Germeys, T. R. Kwapil & N. Barrantes-Vidal (2016) Impact of Adverse Childhood Experiences on Psychotic-Like Symptoms and Stress Reactivity in Daily Life in Nonclinical Young Adults. *PLoS One,* 11**,** e0153557.

Cui, X., W. Niu, L. Kong, M. He, K. Jiang, S. Chen, A. Zhong, W. Li, J. Lu & L. Zhang (2017a) Can lncRNAs be indicators for the diagnosis of early onset or acute schizophrenia and distinguish major depressive disorder and generalized anxiety disorder?-A cross validation analysis. *Am J Med Genet B Neuropsychiatr Genet,* 174**,** 335-341.

Cui, X., W. Niu, L. Kong, M. He, K. Jiang, S. Chen, A. Zhong, Q. Zhang, W. Li, J. Lu & L. Zhang (2017b) Long noncoding RNA as an indicator differentiating schizophrenia from major depressive disorder and generalized anxiety disorder in nonpsychiatric hospital. *Biomark Med,* 11**,** 221-228.

Cullinan, W. E., J. P. Herman, D. F. Battaglia, H. Akil & S. J. Watson (1995) Pattern and time course of immediate early gene expression in rat brain following acute stress. *Neuroscience,* 64**,** 477-505.

Cummings, E., G. Donohoe, A. Hargreaves, S. Moore, C. Fahey, T. G. Dinan, C. McDonald, E. O'Callaghan, F. A. O'Neill, J. L. Waddington, K. C. Murphy, D. W. Morris, M. Gill & A. Corvin (2013) Mood congruent psychotic symptoms and specific cognitive deficits in carriers of the novel schizophrenia risk variant at MIR-137. *Neurosci Lett,* 532**,** 33-8.

Cyr, D. M. & C. H. Ramos (2015) Specification of Hsp70 function by Type I and Type II Hsp40. *Subcell Biochem,* 78**,** 91-102.

D'Souza, U. M., G. Powell-Smith, K. Haddley, T. R. Powell, V. J. Bubb, T. Price, P. McGuffin, J. P. Quinn & A. E. Farmer (2013) Allele-specific expression of the serotonin transporter and its transcription factors following lamotrigine treatment in vitro. *Am J Med Genet B Neuropsychiatr Genet,* 162b**,** 474-83.

Daalman, K., K. M. Diederen, E. M. Derks, R. van Lutterveld, R. S. Kahn & I. E. Sommer (2012) Childhood trauma and auditory verbal hallucinations. *Psychol Med,* 42**,** 2475-84.

Dannlowski, U., H. Kugel, D. Grotegerd, R. Redlich, N. Opel, K. Dohm, D. Zaremba, A. Grogler, J. Schwieren, T. Suslow, P. Ohrmann, J. Bauer, A. Krug, T. Kircher, A. Jansen, K. Domschke, C. Hohoff, P. Zwitserlood, M. Heinrichs, V. Arolt, W. Heindel & B. T. Baune (2016) Disadvantage of Social Sensitivity: Interaction of Oxytocin Receptor Genotype and Child Maltreatment on Brain Structure. *Biol Psychiatry,* 80**,** 398-405.

David, A. P., E. Margarit, P. Domizi, C. Banchio, P. Armas & N. B. Calcaterra (2016) G-quadruplexes as novel cis-elements controlling transcription during embryonic development. *Nucleic Acids Res,* 44**,** 4163-73.

Davidson, S., M. Lear, L. Shanley, B. Hing, A. Baizan-Edge, A. Herwig, J. P. Quinn, G. Breen, P. McGuffin, A. Starkey, P. Barrett & A. MacKenzie (2011) Differential activity by polymorphic variants of a remote enhancer that supports galanin expression in the hypothalamus and amygdala: implications for obesity, depression and alcoholism. *Neuropsychopharmacology,* 36**,** 2211-21.

Davidson, S., K. A. Miller, A. Dowell, A. Gildea & A. Mackenzie (2006) A remote and highly conserved enhancer supports amygdala specific expression of the gene encoding the anxiogenic neuropeptide substance-P. *Mol Psychiatry,* 11**,** 323, 410-21.

Davidson, S., L. Shanley, P. Cowie, M. Lear, P. McGuffin, J. P. Quinn, P. Barrett & A. MacKenzie (2016) Analysis of the effects of depression associated polymorphisms on the activity of the BICC1 promoter in amygdala neurones. *Pharmacogenomics J,* 16**,** 366-74.

Davis, J., H. Eyre, F. N. Jacka, S. Dodd, O. Dean, S. McEwen, M. Debnath, J. McGrath, M. Maes, P. Amminger, P. D. McGorry, C. Pantelis & M. Berk (2016) A review of vulnerability and risks for schizophrenia: Beyond the two hit hypothesis. *Neurosci Biobehav Rev,* 65**,** 185-94.

De Backer, O., K. C. Arden, M. Boretti, V. Vantomme, C. De Smet, S. Czekay, C. S. Viars, E. De Plaen, F. Brasseur, P. Chomez, B. Van den Eynde, T. Boon & P. van der Bruggen (1999) Characterization of the GAGE genes that are expressed in various human cancers and in normal testis. *Cancer Res,* 59**,** 3157-65.

de Leede-Smith, S. & E. Barkus (2013) A comprehensive review of auditory verbal hallucinations: lifetime prevalence, correlates and mechanisms in healthy and clinical individuals. *Front Hum Neurosci,* 7**,** 367.

Deb, G., V. S. Thakur & S. Gupta (2013) Multifaceted role of EZH2 in breast and prostate tumorigenesis: epigenetics and beyond. *Epigenetics,* 8**,** 464-76.

Denli, A. M., I. Narvaiza, B. E. Kerman, M. Pena, C. Benner, M. C. Marchetto, J. K. Diedrich, A. Aslanian, J. Ma, J. J. Moresco, L. Moore, T. Hunter, A. Saghatelian & F. H. Gage (2015) Primate-specific ORF0 contributes to retrotransposon-mediated diversity. *Cell,* 163**,** 583-93.

Derrien, T., R. Johnson, G. Bussotti, A. Tanzer, S. Djebali, H. Tilgner, G. Guernec, D. Martin, A. Merkel, D. G. Knowles, J. Lagarde, L. Veeravalli, X. Ruan, Y. Ruan, T. Lassmann, P. Carninci, J. B. Brown, L. Lipovich, J. M. Gonzalez, M. Thomas, C. A. Davis, R. Shiekhattar, T. R. Gingeras, T. J. Hubbard, C. Notredame, J. Harrow & R. Guigo (2012) The GENCODE v7 catalog of human long noncoding RNAs: analysis of their gene structure, evolution, and expression. *Genome Res,* 22**,** 1775-89.

DeVylder, J. E., A. Koyanagi, J. Unick, H. Oh, B. Nam & A. Stickley (2016) Stress Sensitivity and Psychotic Experiences in 39 Low- and Middle-Income Countries. *Schizophr Bull,* 42**,** 1353-1362.

Dietrich, N., M. Lerdrup, E. Landt, S. Agrawal-Singh, M. Bak, N. Tommerup, J. Rappsilber, E. Sodersten & K. Hansen (2012) REST-mediated recruitment of polycomb repressor complexes in mammalian cells. *PLoS Genet,* 8**,** e1002494.

Ding, K., X. M. Wang, R. Fu, E. B. Ruan, H. Liu & Z. H. Shao (2012) PRAME Gene Expression in Acute Leukemia and Its Clinical Significance. *Cancer Biol Med,* 9**,** 73-6.

Djebali, S., C. A. Davis, A. Merkel, A. Dobin, T. Lassmann, A. Mortazavi, A. Tanzer, J. Lagarde, W. Lin, F. Schlesinger, C. Xue, G. K. Marinov, J. Khatun, B. A. Williams, C. Zaleski, J. Rozowsky, M. Roder, F. Kokocinski, R. F. Abdelhamid,

T. Alioto, I. Antoshechkin, M. T. Baer, N. S. Bar, P. Batut, K. Bell, I. Bell, S. Chakrabortty, X. Chen, J. Chrast, J. Curado, T. Derrien, J. Drenkow, E. Dumais, J. Dumais, R. Duttagupta, E. Falconnet, M. Fastuca, K. Fejes-Toth, P. Ferreira, S. Foissac, M. J. Fullwood, H. Gao, D. Gonzalez, A. Gordon, H. Gunawardena, C. Howald, S. Jha, R. Johnson, P. Kapranov, B. King, C. Kingswood, O. J. Luo, E. Park, K. Persaud, J. B. Preall, P. Ribeca, B. Risk, D. Robyr, M. Sammeth, L. Schaffer, L. H. See, A. Shahab, J. Skancke, A. M. Suzuki, H. Takahashi, H. Tilgner, D. Trout, N. Walters, H. Wang, J. Wrobel, Y. Yu, X. Ruan, Y. Hayashizaki, J. Harrow, M. Gerstein, T. Hubbard, A. Reymond, S. E. Antonarakis, G. Hannon, M. C. Giddings, Y. Ruan, B. Wold, P. Carninci, R. Guigo & T. R. Gingeras (2012) Landscape of transcription in human cells. *Nature,* 489**,** 101-8.

Donnelier, J. & J. E. A. Braun (2014) CSPα—chaperoning presynaptic proteins. *Front Cell Neurosci,* 8.

Doorley, J., C. Williams, T. Mallard, C. Esposito-Smythers & J. McGeary (2017) Sexual Trauma, the Dopamine D4 Receptor, and Suicidal Ideation Among Hospitalized Adolescents: A Preliminary Investigation. *Arch Suicide Res,* 21**,** 279-292.

Dragan, W. L. & W. Oniszczenko (2009) The association between dopamine D4 receptor exon III polymorphism and intensity of PTSD symptoms among flood survivors. *Anxiety Stress Coping,* 22**,** 483-95.

Duan, J., J. Shi, A. Fiorentino, C. Leites, X. Chen, W. Moy, J. Chen, B. S. Alexandrov, A. Usheva, D. He, J. Freda, N. L. O'Brien, A. McQuillin, A. R. Sanders, E. S. Gershon, L. E. DeLisi, A. R. Bishop, H. M. Gurling, M. T. Pato, D. F. Levinson, K. S. Kendler, C. N. Pato & P. V. Gejman (2014) A rare functional noncoding variant at the GWAS-implicated MIR137/MIR2682 locus might confer risk to schizophrenia and bipolar disorder. *Am J Hum Genet,* 95**,** 744-53.

Duffy, D. L., Z. Z. Zhao, R. A. Sturm, N. K. Hayward, N. G. Martin & G. W. Montgomery (2010) Multiple pigmentation gene polymorphisms account for a substantial proportion of risk of cutaneous malignant melanoma. *J Invest Dermatol,* 130**,** 520-8.

Dávila-Rodríguez, M., E. Cortés Gutiérrez, R. Cerda Flores, M. Pita, J. Fernández, C. López-Fernández & J. Gosálvez (2011) Constitutive heterochromatin polymorphisms in human chromosomes identified by whole comparative genomic hybridization. *Eur J Histochem,* 55.

Ecco, G., M. Cassano, A. Kauzlaric, J. Duc, A. Coluccio, S. Offner, M. Imbeault, H. M. Rowe, P. Turelli & D. Trono (2016) Transposable Elements and Their KRAB-ZFP Controllers Regulate Gene Expression in Adult Tissues. *Dev Cell,* 36**,** 611-23.

Edvardson, S., Y. Cinnamon, A. Ta-Shma, A. Shaag, Y. I. Yim, S. Zenvirt, C. Jalas, S. Lesage, A. Brice, A. Taraboulos, K. H. Kaestner, L. E. Greene & O. Elpeleg (2012) A deleterious mutation in DNAJC6 encoding the neuronal-specific clathrin-uncoating co-chaperone auxilin, is associated with juvenile parkinsonism. *PLoS One,* 7**,** e36458.

Elbarbary, R. A., B. A. Lucas & L. E. Maquat (2016) Retrotransposons as regulators of gene expression. *Science,* 351**,** aac7247.

Ellegood, J., S. Markx, J. P. Lerch, P. E. Steadman, C. Genc, F. Provenzano, S. A. Kushner, R. M. Henkelman, M. Karayiorgou & J. A. Gogos (2014) Neuroanatomical phenotypes in a mouse model of the 22q11.2 microdeletion. *Mol Psychiatry,* 19**,** 99-107.

Emerson, R. O. & J. H. Thomas (2009) Adaptive evolution in zinc finger transcription factors. *PLoS Genet,* 5**,** e1000325.
Ercolak, V., S. Paydas, E. Bagir, M. Ergin, G. Seydaoglu, H. Celik, B. Yavu, K. Tanriverdi, M. Gunaldi, C. U. Afsar & B. B. Duman (2015) PRAME Expression and Its Clinical Relevance in Hodgkin's Lymphoma. *Acta Haematol,* 134**,** 199-207.
Ernsberger, U. (2012) Regulation of gene expression during early neuronal differentiation: evidence for patterns conserved across neuron populations and vertebrate classes. *Cell Tissue Res,* 348**,** 1-27.
Erwin, J. A., M. C. Marchetto & F. H. Gage (2014) Mobile DNA elements in the generation of diversity and complexity in the brain. *Nat Rev Neurosci,* 15**,** 497-506.
Esteller, M. (2011) Non-coding RNAs in human disease. *Nat Rev Genet,* 12**,** 861-74.
Fan, C. Y., S. Lee & D. M. Cyr. 2003. Mechanisms for regulation of Hsp70 function by Hsp40. In *Cell Stress Chaperones*, 309-16.
Fan, H. M., X. Y. Sun, W. Niu, L. Zhao, Q. L. Zhang, W. S. Li, A. F. Zhong, L. Y. Zhang & J. Lu (2015) Altered microRNA Expression in Peripheral Blood Mononuclear Cells from Young Patients with Schizophrenia. *J Mol Neurosci,* 56**,** 562-71.
Fasching, L., A. Kapopoulou, R. Sachdeva, R. Petri, M. E. Jonsson, C. Manne, P. Turelli, P. Jern, F. Cammas, D. Trono & J. Jakobsson (2015) TRIM28 represses transcription of endogenous retroviruses in neural progenitor cells. *Cell Rep,* 10**,** 20-8.
Fatemi, S. H. & T. D. Folsom (2011) The role of fragile X mental retardation protein in major mental disorders. *Neuropharmacology,* 60**,** 1221-6.
Faulkner, G. J., Y. Kimura, C. O. Daub, S. Wani, C. Plessy, K. M. Irvine, K. Schroder, N. Cloonan, A. L. Steptoe, T. Lassmann, K. Waki, N. Hornig, T. Arakawa, H. Takahashi, J. Kawai, A. R. Forrest, H. Suzuki, Y. Hayashizaki, D. A. Hume, V. Orlando, S. M. Grimmond & P. Carninci (2009) The regulated retrotransposon transcriptome of mammalian cells. *Nat Genet,* 41**,** 563-71.
Feng, Q., J. V. Moran, H. H. Kazazian, Jr. & J. D. Boeke (1996) Human L1 retrotransposon encodes a conserved endonuclease required for retrotransposition. *Cell,* 87**,** 905-16.
Field, M. G., C. L. Decatur, S. Kurtenbach, G. Gezgin, P. A. van der Velden, M. J. Jager, K. N. Kozak & J. W. Harbour (2016) PRAME as an Independent Biomarker for Metastasis in Uveal Melanoma. *Clin Cancer Res,* 22**,** 1234-42.
Fiskerstrand, C. E., E. A. Lovejoy & J. P. Quinn (1999) An intronic polymorphic domain often associated with susceptibility to affective disorders has allele dependent differential enhancer activity in embryonic stem cells. *FEBS Lett,* 458**,** 171-4.
Forstner, A. J., F. Degenhardt, G. Schratt & M. M. Nothen (2013) MicroRNAs as the cause of schizophrenia in 22q11.2 deletion carriers, and possible implications for idiopathic disease: a mini-review. *Front Mol Neurosci,* 6**,** 47.
Fox, J. W. (1990) Social class, mental illness, and social mobility: the social selection-drift hypothesis for serious mental illness. *J Health Soc Behav,* 31**,** 344-53.
Friedman, R. C., K. K. Farh, C. B. Burge & D. P. Bartel (2009) Most mammalian mRNAs are conserved targets of microRNAs. *Genome Res,* 19**,** 92-105.
Gabriel, S. B., S. F. Schaffner, H. Nguyen, J. M. Moore, J. Roy, B. Blumenstiel, J. Higgins, M. DeFelice, A. Lochner, M. Faggart, S. N. Liu-Cordero, C. Rotimi, A. Adeyemo, R. Cooper, R. Ward, E. S. Lander, M. J. Daly & D. Altshuler (2002) The structure of haplotype blocks in the human genome. *Science,* 296**,** 2225-9.

Gallone, G., W. Haerty, G. Disanto, S. V. Ramagopalan, C. P. Ponting & A. J. Berlanga-Taylor (2017) Identification of genetic variants affecting vitamin D receptor binding and associations with autoimmune disease. *Hum Mol Genet,* 26**,** 2164-2176.

Gao, Z., K. Ure, P. Ding, M. Nashaat, L. Yuan, J. Ma, R. E. Hammer & J. Hsieh (2011) The master negative regulator REST/NRSF controls adult neurogenesis by restraining the neurogenic program in quiescent stem cells. *J Neurosci,* 31**,** 9772-86.

Gemayel, R., J. Cho, S. Boeynaems & K. J. Verstrepen. 2012. Beyond Junk-Variable Tandem Repeats as Facilitators of Rapid Evolution of Regulatory and Coding Sequences. In *Genes (Basel)*, 461-80.

Gemayel, R., M. D. Vinces, M. Legendre & K. J. Verstrepen (2010) Variable tandem repeats accelerate evolution of coding and regulatory sequences. *Annu Rev Genet,* 44**,** 445-77.

Gianfrancesco, O., V. J. Bubb & J. P. Quinn (2016a) SVA retrotransposons as potential modulators of neuropeptide gene expression. *Neuropeptides*.

Gianfrancesco, O., D. Griffiths, P. Myers, D. A. Collier, V. J. Bubb & J. P. Quinn (2016b) Identification and Potential Regulatory Properties of Evolutionary Conserved Regions (ECRs) at the Schizophrenia-Associated MIR137 Locus. *J Mol Neurosci,* 60**,** 239-47.

Gianfrancesco, O., A. Warburton, D. A. Collier, V. J. Bubb & J. P. Quinn (2017) Novel brain expressed RNA identified at the MIR137 schizophrenia-associated locus. *Schizophr Res,* 184**,** 109-115.

Glinsky, G. V. (2017) Mechanistically Distinct Pathways of Divergent Regulatory DNA Creation Contribute to Evolution of Human-Specific Genomic Regulatory Networks Driving Phenotypic Divergence of Homo sapiens. *Genome Biology and Evolution,* 8**,** 2774-2788.

Gonzalez-Giraldo, Y., R. E. Gonzalez-Reyes & D. A. Forero (2016) A functional variant in MIR137, a candidate gene for schizophrenia, affects Stroop test performance in young adults. *Psychiatry Res,* 236**,** 202-5.

Goodier, J. L. & H. H. Kazazian, Jr. (2008) Retrotransposons revisited: the restraint and rehabilitation of parasites. *Cell,* 135**,** 23-35.

Gorka, A. X., J. L. Hanson, S. R. Radtke & A. R. Hariri (2014) Reduced hippocampal and medial prefrontal gray matter mediate the association between reported childhood maltreatment and trait anxiety in adulthood and predict sensitivity to future life stress. *Biol Mood Anxiety Disord,* 4**,** 12.

Goulart, L. F., F. Bettella, I. E. Sønderby, A. J. Schork, W. K. Thompson, M. Mattingsdal, V. M. Steen, V. Zuber, Y. Wang, A. M. Dale, O. A. Andreassen & S. Djurovic (2015) MicroRNAs enrichment in GWAS of complex human phenotypes. *BMC Genomics,* 16.

Green, M. J., M. J. Cairns, J. Wu, M. Dragovic, A. Jablensky, P. A. Tooney, R. J. Scott & V. J. Carr (2013) Genome-wide supported variant MIR137 and severe negative symptoms predict membership of an impaired cognitive subtype of schizophrenia. *Mol Psychiatry,* 18**,** 774-80.

Gu, H. P., S. Lin, M. Xu, H. Y. Yu, X. J. Du, Y. Y. Zhang, G. Yuan & W. Gao (2012) Up-regulating relaxin expression by G-quadruplex interactive ligand to achieve antifibrotic action. *Endocrinology,* 153**,** 3692-700.

Guella, I., A. Sequeira, B. Rollins, L. Morgan, F. Torri, T. G. van Erp, R. M. Myers, J. D. Barchas, A. F. Schatzberg, S. J. Watson, H. Akil, W. E. Bunney, S. G. Potkin,

F. Macciardi & M. P. Vawter (2013) Analysis of miR-137 expression and rs1625579 in dorsolateral prefrontal cortex. *J Psychiatr Res,* 47**,** 1215-21.
Guindalini, C., M. Howard, K. Haddley, R. Laranjeira, D. Collier, N. Ammar, I. Craig, C. O'Gara, V. J. Bubb, T. Greenwood, J. Kelsoe, P. Asherson, R. M. Murray, A. Castelo, J. P. Quinn, H. Vallada & G. Breen (2006) A dopamine transporter gene functional variant associated with cocaine abuse in a Brazilian sample. *Proc Natl Acad Sci U S A,* 103**,** 4552-7.
Guo, Z., W. Niu, Y. Bi, R. Zhang, D. Ren, J. Hu, X. Huang, X. Wu, Y. Cao, F. Yang, L. Wang, W. Li, X. Li, Y. Xu, L. He, T. Yu & G. He (2016) A study of single nucleotide polymorphisms of GRIN2B in schizophrenia from Chinese Han population. *Neurosci Lett,* 630**,** 132-5.
Gur, R. E., A. S. Bassett, D. M. McDonald-McGinn, C. E. Bearden, E. Chow, B. S. Emanuel, M. Owen, A. Swillen, M. Van den Bree, J. Vermeesch, J. A. S. Vorstman, S. Warren, T. Lehner & B. Morrow (2017) A neurogenetic model for the study of schizophrenia spectrum disorders: the International 22q11.2 Deletion Syndrome Brain Behavior Consortium. *Mol Psychiatry*.
Hacıhamdioğlu, B., D. Hacıhamdioğlu & K. Delil (2015) 22q11 deletion syndrome: current perspective. *Appl Clin Genet,* 8**,** 123-32.
Hamilton, A. T., S. Huntley, M. Tran-Gyamfi, D. M. Baggott, L. Gordon & L. Stubbs (2006) Evolutionary expansion and divergence in the ZNF91 subfamily of primate-specific zinc finger genes. *Genome Res,* 16**,** 584-94.
Han, J., Y. Lee, K. H. Yeom, J. W. Nam, I. Heo, J. K. Rhee, S. Y. Sohn, Y. Cho, B. T. Zhang & V. N. Kim (2006) Molecular basis for the recognition of primary microRNAs by the Drosha-DGCR8 complex. *Cell,* 125**,** 887-901.
Hancks, D. C. & H. H. Kazazian, Jr. (2016) Roles for retrotransposon insertions in human disease. *Mob DNA,* 7**,** 9.
Hangauer, M. J., I. W. Vaughn & M. T. McManus (2013) Pervasive transcription of the human genome produces thousands of previously unidentified long intergenic noncoding RNAs. *PLoS Genet,* 9**,** e1003569.
Harrow, J., A. Frankish, J. M. Gonzalez, E. Tapanari, M. Diekhans, F. Kokocinski, B. L. Aken, D. Barrell, A. Zadissa, S. Searle, I. Barnes, A. Bignell, V. Boychenko, T. Hunt, M. Kay, G. Mukherjee, J. Rajan, G. Despacio-Reyes, G. Saunders, C. Steward, R. Harte, M. Lin, C. Howald, A. Tanzer, T. Derrien, J. Chrast, N. Walters, S. Balasubramanian, B. Pei, M. Tress, J. M. Rodriguez, I. Ezkurdia, J. van Baren, M. Brent, D. Haussler, M. Kellis, A. Valencia, A. Reymond, M. Gerstein, R. Guigo & T. J. Hubbard (2012) GENCODE: the reference human genome annotation for The ENCODE Project. *Genome Res,* 22**,** 1760-74.
Hazell, C. M., M. Hayward, K. Cavanagh & C. Strauss (2016) A systematic review and meta-analysis of low intensity CBT for psychosis. *Clin Psychol Rev,* 45**,** 183-92.
Heard, T. T., S. Ramgopal, J. Picker, S. A. Lincoln, A. Rotenberg & S. V. Kothare (2014) EEG abnormalities and seizures in genetically diagnosed Fragile X syndrome. *Int J Dev Neurosci,* 38**,** 155-60.
Heim, C., D. J. Newport, T. Mletzko, A. H. Miller & C. B. Nemeroff (2008) The link between childhood trauma and depression: insights from HPA axis studies in humans. *Psychoneuroendocrinology,* 33**,** 693-710.
Heinz, A., L. Deserno & U. Reininghaus (2013) Urbanicity, social adversity and psychosis. *World Psychiatry,* 12**,** 187-97.

Helman, E., M. L. Lawrence, C. Stewart, C. Sougnez, G. Getz & M. Meyerson (2014) Somatic retrotransposition in human cancer revealed by whole-genome and exome sequencing.

Henderson, M. X., G. S. Wirak, Y. Q. Zhang, F. Dai, S. D. Ginsberg, N. Dolzhanskaya, J. F. Staropoli, P. C. Nijssen, T. T. Lam, A. F. Roth, N. G. Davis, G. Dawson, M. Velinov & S. S. Chandra (2016) Neuronal ceroid lipofuscinosis with DNAJC5/CSPalpha mutation has PPT1 pathology and exhibit aberrant protein palmitoylation. *Acta Neuropathol,* 131**,** 621-37.

Herringa, R. J., R. M. Birn, P. L. Ruttle, C. A. Burghy, D. E. Stodola, R. J. Davidson & M. J. Essex (2013) Childhood maltreatment is associated with altered fear circuitry and increased internalizing symptoms by late adolescence. *Proc Natl Acad Sci U S A,* 110**,** 19119-24.

Hindorff, L. A., P. Sethupathy, H. A. Junkins, E. M. Ramos, J. P. Mehta, F. S. Collins & T. A. Manolio (2009) Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc Natl Acad Sci U S A,* 106**,** 9362-7.

Hing, B., S. Davidson, M. Lear, G. Breen, J. Quinn, P. McGuffin & A. MacKenzie (2012) A polymorphism associated with depressive disorders differentially regulates brain derived neurotrophic factor promoter IV activity. *Biol Psychiatry,* 71**,** 618-26.

Hirabayashi, Y., N. Suzki, M. Tsuboi, T. A. Endo, T. Toyoda, J. Shinga, H. Koseki, M. Vidal & Y. Gotoh (2009) Polycomb limits the neurogenic competence of neural precursor cells to promote astrogenic fate transition. *Neuron,* 63**,** 600-13.

Hoeffding, L. K., B. B. Trabjerg, L. Olsen, W. Mazin, T. Sparso, A. Vangkilde, P. B. Mortensen, C. B. Pedersen & T. Werge (2017) Risk of Psychiatric Disorders Among Individuals With the 22q11.2 Deletion or Duplication: A Danish Nationwide, Register-Based Study. *JAMA Psychiatry,* 74**,** 282-290.

Hoffman, G. E., M. S. Smith & J. G. Verbalis (1993) c-Fos and related immediate early gene products as markers of activity in neuroendocrine systems. *Front Neuroendocrinol,* 14**,** 173-213.

Hoffmann, A., M. Ziller & D. Spengler (2016) The Future is The Past: Methylation QTLs in Schizophrenia. *Genes (Basel),* 7.

Holz, N., R. Boecker, A. F. Buchmann, D. Blomeyer, S. Baumeister, S. Hohmann, C. Jennen-Steinmetz, I. Wolf, M. Rietschel, S. H. Witt, M. M. Plichta, A. Meyer-Lindenberg, M. H. Schmidt, G. Esser, T. Banaschewski, D. Brandeis & M. Laucht (2016) Evidence for a Sex-Dependent MAOAx Childhood Stress Interaction in the Neural Circuitry of Aggression. *Cereb Cortex,* 26**,** 904-14.

Honkaniemi, J., T. Kainu, S. Ceccatelli, L. Rechardt, T. Hokfelt & M. Pelto-Huikko (1992) Fos and jun in rat central amygdaloid nucleus and paraventricular nucleus after stress. *Neuroreport,* 3**,** 849-52.

Hu, H. Y., L. He, K. Fominykh, Z. Yan, S. Guo, X. Zhang, M. S. Taylor, L. Tang, J. Li, J. Liu, W. Wang, H. Yu & P. Khaitovich (2012) Evolution of the human-specific microRNA miR-941. *Nat Commun,* 3**,** 1145.

Hu, Z. & Z. Li (2017) miRNAs in synapse development and synaptic plasticity. *Curr Opin Neurobiol,* 45**,** 24-31.

Huang, Q. (2015) Genetic study of complex diseases in the post-GWAS era. *J Genet Genomics,* 42**,** 87-98.

Huppert, J. L. & S. Balasubramanian (2007) G-quadruplexes in promoters throughout the human genome. *Nucleic Acids Res,* 35**,** 406-13.

Hur, K., P. Cejas, J. Feliu, J. Moreno-Rubio, E. Burgos, C. R. Boland & A. Goel (2014) Hypomethylation of long interspersed nuclear element-1 (LINE-1) leads to activation of proto-oncogenes in human colorectal cancer metastasis. *Gut,* 63**,** 635-46.

Imbeault, M., P.-Y. Helleboid & D. Trono (2017) KRAB zinc-finger proteins contribute to the evolution of gene regulatory networks. *Nature,* 543**,** 550-554.

Jacobs, F. M. J., D. Greenberg, N. Nguyen, M. Haeussler, A. D. Ewing, S. Katzman, B. Paten, S. R. Salama & D. Haussler (2014) An evolutionary arms race between KRAB zinc-finger genes ZNF91/93 and SVA/L1 retrotransposons. *Nature,* 516**,** 242-245.

Jaffe, A. E., R. E. Straub, J. H. Shin, R. Tao, Y. Gao, L. C. Torres, T. Kam-Thong, H. S. Xi, J. Quan, Q. Chen, C. Colantuoni, B. Ulrich, B. J. Maher, A. Deep-Soboslay, T. B. Consortium, A. Cross, N. J. Brandon, J. T. Leek, T. M. Hyde, J. E. Kleinman & D. R. Weinberger (2017a) Developmental And Genetic Regulation Of The Human Cortex Transcriptome In Schizophrenia.

Jaffe, A. E., R. Tao, A. L. Norris, M. Kealhofer, A. Nellore, J. H. Shin, D. Kim, Y. Jia, T. M. Hyde, J. E. Kleinman, R. E. Straub, J. T. Leek & D. R. Weinberger (2017b) qSVA framework for RNA quality correction in differential expression analysis. *Proc Natl Acad Sci U S A*.

Jansen, A., R. Gemayel & K. J. Verstrepen (2012) Unstable microsatellite repeats facilitate rapid evolution of coding and regulatory sequences. *Genome Dyn,* 7**,** 108-25.

Jeffries, C. D., D. O. Perkins, S. D. Chandler, T. Stark, E. Yeo, J. Addington, C. E. Bearden, K. S. Cadenhead, T. D. Cannon, B. A. Cornblatt, D. H. Mathalon, T. H. McGlashan, L. J. Seidman, E. F. Walker, S. W. Woods, S. J. Glatt & M. Tsuang. 2016. Insights into psychosis risk from leukocyte microRNA expression. In *Transl Psychiatry*, e981-.

Johns, L. C., K. Kompus, M. Connell, C. Humpston, T. M. Lincoln, E. Longden, A. Preti, B. Alderson-Day, J. C. Badcock, M. Cella, C. Fernyhough, S. McCarthy-Jones, E. Peters, A. Raballo, J. Scott, S. Siddi, I. E. Sommer & F. Larøi (2014) Auditory Verbal Hallucinations in Persons With and Without a Need for Care. *Schizophr Bull,* 40**,** S255-64.

Johnson, J. N., E. Ahrendt & J. E. Braun (2010) CSPalpha: the neuroprotective J protein. *Biochem Cell Biol,* 88**,** 157-65.

Johnson, R., C. H. Teh, G. Kunarso, K. Y. Wong, G. Srinivasan, M. L. Cooper, M. Volta, S. S. Chan, L. Lipovich, S. M. Pollard, R. K. Karuturi, C. L. Wei, N. J. Buckley & L. W. Stanton (2008) REST regulates distinct transcriptional networks in embryonic and neural stem cells. *PLoS Biol,* 6**,** e256.

Jonas, R. K., C. A. Montojo & C. E. Bearden (2014) The 22q11.2 deletion syndrome as a window into complex neuropsychiatric disorders over the lifespan. *Biol Psychiatry,* 75**,** 351-60.

Jones, C., D. Watson & K. Fone (2011) Animal models of schizophrenia. *Br J Pharmacol,* 164**,** 1162-94.

Jones, S. & J. A. Kauer (1999) Amphetamine depresses excitatory synaptic transmission via serotonin receptors in the ventral tegmental area. *J Neurosci,* 19**,** 9780-7.

Josephson, S. A., R. E. Schmidt, P. Millsap, D. Q. McManus & J. C. Morris (2001) Autosomal dominant Kufs' disease: a cause of early onset dementia. *J Neurol Sci,* 188**,** 51-60.

Kanetsky, P. A., J. Swoyer, S. Panossian, R. Holmes, D. Guerry & T. R. Rebbeck (2002) A polymorphism in the agouti signaling protein gene is associated with human pigmentation. *Am J Hum Genet,* 70**,** 770-5.

Kang, H. J., Y. I. Kawasawa, F. Cheng, Y. Zhu, X. Xu, M. Li, A. M. Sousa, M. Pletikos, K. A. Meyer, G. Sedmak, T. Guennel, Y. Shin, M. B. Johnson, Z. Krsnik, S. Mayer, S. Fertuzinhos, S. Umlauf, S. N. Lisgo, A. Vortmeyer, D. R. Weinberger, S. Mane, T. M. Hyde, A. Huttner, M. Reimers, J. E. Kleinman & N. Sestan (2011) Spatio-temporal transcriptome of the human brain. *Nature,* 478**,** 483-9.

Kashi, K., L. Henderson, A. Bonetti & P. Carninci (2016) Discovery and functional analysis of lncRNAs: Methodologies to investigate an uncharacterized transcriptome. *Biochim Biophys Acta,* 1859**,** 3-15.

Kataoka, N., M. Fujita & M. Ohno (2009) Functional association of the Microprocessor complex with the spliceosome. *Mol Cell Biol,* 29**,** 3243-54.

Kejnovsky, E., V. Tokan & M. Lexa (2015) Transposable elements and G-quadruplexes. *Chromosome Res,* 23**,** 615-23.

Kendler, K. S., H. Ohlsson, J. Sundquist & K. Sundquist (2015) IQ and schizophrenia in a Swedish national sample: their causal relationship and the interaction of IQ with genetic risk. *Am J Psychiatry,* 172**,** 259-65.

Kenny, P. & S. Ceman. 2016. RNA Secondary Structure Modulates FMRP's Bi-Functional Role in the MicroRNA Pathway. In *Int J Mol Sci*.

Kessler, R. C., K. A. McLaughlin, J. G. Green, M. J. Gruber, N. A. Sampson, A. M. Zaslavsky, S. Aguilar-Gaxiola, A. O. Alhamzawi, J. Alonso, M. Angermeyer, C. Benjet, E. Bromet, S. Chatterji, G. de Girolamo, K. Demyttenaere, J. Fayyad, S. Florescu, G. Gal, O. Gureje, J. M. Haro, C. Y. Hu, E. G. Karam, N. Kawakami, S. Lee, J. P. Lepine, J. Ormel, J. Posada-Villa, R. Sagar, A. Tsang, T. B. Ustun, S. Vassilev, M. C. Viana & D. R. Williams (2010) Childhood adversities and adult psychopathology in the WHO World Mental Health Surveys. *Br J Psychiatry,* 197**,** 378-85.

Khan, H., A. Smit & S. Boissinot. 2006. Molecular evolution and tempo of amplification of human LINE-1 retrotransposons since the origin of primates. In *Genome Res*, 78-87.

Khursheed, K., T. P. Wilm, C. Cashman, J. P. Quinn, V. J. Bubb & D. J. Moss (2015) Characterisation of multiple regulatory domains spanning the major transcriptional start site of the FUS gene, a candidate gene for motor neurone disease. *Brain Res,* 1595**,** 1-9.

Kidd, S. A., A. Lachiewicz, D. Barbouth, R. K. Blitz, C. Delahunty, D. McBrien, J. Visootsak & E. Berry-Kravis (2014) Fragile X syndrome: a review of associated medical problems. *Pediatrics,* 134**,** 995-1005.

Kim, A. H., E. K. Parker, V. Williamson, G. O. McMichael, A. H. Fanous & V. I. Vladimirov (2012) Experimental validation of candidate schizophrenia gene ZNF804A as target for hsa-miR-137. *Schizophr Res,* 141**,** 60-4.

Kim, D. S. & Y. Hahn (2010) Human-specific antisense transcripts induced by the insertion of transposable element. *Int J Mol Med,* 26**,** 151-7.

Kim, D. S. & Y. Hahn (2011) Identification of human-specific transcript variants induced by DNA insertions in the human genome. *Bioinformatics,* 27**,** 14-21.

Kim, J. J., L. Mandelli, S. Lim, H. K. Lim, O. J. Kwon, C. U. Pae, A. Serretti, V. L. Nimgaonkar, I. H. Paik & T. Y. Jun (2008) Association analysis of heat shock protein 70 gene polymorphisms in schizophrenia. *Eur Arch Psychiatry Clin Neurosci,* 258**,** 239-44.

Kim, S., N. K. Yu & B. K. Kaang. 2015. CTCF as a multifunctional protein in genome regulation and gene expression. In *Exp Mol Med*, e166-.

Kim, Y. K. & V. N. Kim (2007) Processing of intronic microRNAs. *EMBO J,* 26**,** 775-83.

Klawitter, S., N. V. Fuchs, K. R. Upton, M. Muñoz-Lopez, R. Shukla, J. Wang, M. Garcia-Cañadas, C. Lopez-Ruiz, D. J. Gerhardt, A. Sebe, I. Grabundzija, S. Merkert, P. Gerdes, J. A. Pulgarin, A. Bock, U. Held, A. Witthuhn, A. Haase, B. Sarkadi, J. Löwer, E. J. Wolvetang, U. Martin, Z. Ivics, Z. Izsvák, J. L. Garcia-Perez, G. J. Faulkner & G. G. Schumann (2015) Reprogramming triggers endogenous L1 and Alu retrotransposition in human induced pluripotent stem cells. *Nature Communications, Published online: 8 January 2016; | doi:10.1038/ncomms10286*.

Klengel, T. & E. B. Binder (2015) Epigenetics of Stress-Related Psychiatric Disorders and Gene x Environment Interactions. *Neuron,* 86**,** 1343-57.

Klenova, E., A. C. Scott, J. Roberts, S. Shamsuddin, E. A. Lovejoy, S. Bergmann, V. J. Bubb, H. D. Royer & J. P. Quinn (2004) YB-1 and CTCF differentially regulate the 5-HTT polymorphic intron 2 enhancer which predisposes to a variety of neurological disorders. *J Neurosci,* 24**,** 5966-73.

Kong, U., J. Koo, K. Choi, J. Park & H. Chang (2004) The expression of GAGE gene can predict aggressive biologic behavior of intestinal type of stomach cancer. *Hepatogastroenterology,* 51**,** 1519-23.

Koppel, I. & T. Timmusk (2013) Differential regulation of Bdnf expression in cortical neurons by class-selective histone deacetylase inhibitors. *Neuropharmacology,* 75**,** 106-15.

Kornienko, A. E., P. M. Guenzl, D. P. Barlow & F. M. Pauler (2013) Gene regulation by the act of long non-coding RNA transcription. *BMC Biol,* 11**,** 59.

Kowalczyk, M., A. Owczarek, R. Suchanek, M. Paul-Samojedny, A. Fila-Danilow, P. Borkowska, K. Kucia & J. Kowalski (2014) Heat shock protein 70 gene polymorphisms are associated with paranoid schizophrenia in the Polish population. *Cell Stress Chaperones,* 19**,** 205-15.

Krakvik, B., F. Laroi, A. M. Kalhovde, K. Hugdahl, K. Kompus, O. Salvesen, T. C. Stiles & E. Vedul-Kjelsas (2015) Prevalence of auditory verbal hallucinations in a general population: A group comparison study. *Scand J Psychol,* 56**,** 508-15.

Krol, J., K. Sobczak, U. Wilczynska, M. Drath, A. Jasinska, D. Kaczynska & W. J. Krzyzosiak (2004) Structural features of microRNA (miRNA) precursors and their relevance to miRNA biogenesis and small interfering RNA/short hairpin RNA design. *J Biol Chem,* 279**,** 42230-9.

Krug, L., N. Chatterjee, R. Borges-Monroy, S. Hearn, W. W. Liao, K. Morrill, L. Prazak, N. Rozhkov, D. Theodorou, M. Hammell & J. Dubnau (2017) Retrotransposon activation contributes to neurodegeneration in a Drosophila TDP-43 model of ALS. *PLoS Genet,* 13**,** e1006635.

Kuleshov, M. V., M. R. Jones, A. D. Rouillard, N. F. Fernandez, Q. Duan, Z. Wang, S. Koplev, S. L. Jenkins, K. M. Jagodnik, A. Lachmann, M. G. McDermott, C. D. Monteiro, G. W. Gundersen & A. Ma'ayan (2016) Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. *Nucleic Acids Res,* 44**,** W90-7.

Kumar, G., S. L. Clark, J. L. McClay, A. A. Shabalin, D. E. Adkins, L. Xie, R. Chan, S. Nerella, Y. Kim, P. F. Sullivan, C. M. Hultman, P. K. Magnusson, K. A. Aberg & E. J. van den Oord (2015) Refinement of schizophrenia GWAS loci using methylome-wide association data. *Hum Genet,* 134**,** 77-87.

Kuswanto, C. N., M. Y. Sum, A. Qiu, Y. Y. Sitoh, J. Liu & K. Sim (2015) The impact of genome wide supported microRNA-137 (MIR137) risk variants on frontal and striatal white matter integrity, neurocognitive functioning, and negative symptoms in schizophrenia. *Am J Med Genet B Neuropsychiatr Genet,* 168B**,** 317-26.

Kwiatkowski, T. J., Jr., D. A. Bosco, A. L. Leclerc, E. Tamrazian, C. R. Vanderburg, C. Russ, A. Davis, J. Gilchrist, E. J. Kasarskis, T. Munsat, P. Valdmanis, G. A. Rouleau, B. A. Hosler, P. Cortelli, P. J. de Jong, Y. Yoshinaga, J. L. Haines, M. A. Pericak-Vance, J. Yan, N. Ticozzi, T. Siddique, D. McKenna-Yasek, P. C. Sapp, H. R. Horvitz, J. E. Landers & R. H. Brown, Jr. (2009) Mutations in the FUS/TLS gene on chromosome 16 cause familial amyotrophic lateral sclerosis. *Science,* 323**,** 1205-8.

Kwon, E., W. Wang & L. H. Tsai. 2013. Validation of schizophrenia-associated genes CSMD1, C10orf26, CACNA1C and TCF4 as miR-137 targets. In *Mol Psychiatry*, 11-2. England.

Lander, E. S., L. M. Linton, B. Birren, C. Nusbaum, M. C. Zody, J. Baldwin, K. Devon, K. Dewar, M. Doyle, W. FitzHugh, R. Funke, D. Gage, K. Harris, A. Heaford, J. Howland, L. Kann, J. Lehoczky, R. LeVine, P. McEwan, K. McKernan, J. Meldrim, J. P. Mesirov, C. Miranda, W. Morris, J. Naylor, C. Raymond, M. Rosetti, R. Santos, A. Sheridan, C. Sougnez, Y. Stange-Thomann, N. Stojanovic, A. Subramanian, D. Wyman, J. Rogers, J. Sulston, R. Ainscough, S. Beck, D. Bentley, J. Burton, C. Clee, N. Carter, A. Coulson, R. Deadman, P. Deloukas, A. Dunham, I. Dunham, R. Durbin, L. French, D. Grafham, S. Gregory, T. Hubbard, S. Humphray, A. Hunt, M. Jones, C. Lloyd, A. McMurray, L. Matthews, S. Mercer, S. Milne, J. C. Mullikin, A. Mungall, R. Plumb, M. Ross, R. Shownkeen, S. Sims, R. H. Waterston, R. K. Wilson, L. W. Hillier, J. D. McPherson, M. A. Marra, E. R. Mardis, L. A. Fulton, A. T. Chinwalla, K. H. Pepin, W. R. Gish, S. L. Chissoe, M. C. Wendl, K. D. Delehaunty, T. L. Miner, A. Delehaunty, J. B. Kramer, L. L. Cook, R. S. Fulton, D. L. Johnson, P. J. Minx, S. W. Clifton, T. Hawkins, E. Branscomb, P. Predki, P. Richardson, S. Wenning, T. Slezak, N. Doggett, J. F. Cheng, A. Olsen, S. Lucas, C. Elkin, E. Uberbacher, M. Frazier, et al. (2001) Initial sequencing and analysis of the human genome. *Nature,* 409**,** 860-921.

Lawrence, C. L. & A. S. Baldwin (2016) Non-Canonical EZH2 Transcriptionally Activates RelB in Triple Negative Breast Cancer. *PLoS One,* 11**,** e0165005.

Le Hellard, S., Y. Wang, A. Witoelar, V. Zuber, F. Bettella, K. Hugdahl, T. Espeseth, V. M. Steen, I. Melle, R. Desikan, A. J. Schork, W. K. Thompson, A. M. Dale, S. Djurovic & O. A. Andreassen (2017) Identification of Gene Loci That Overlap Between Schizophrenia and Educational Attainment. *Schizophr Bull,* 43**,** 654-664.

Lee, E., R. Iskow, L. Yang, O. Gokcumen, P. Haseley, L. J. Luquette, 3rd, J. G. Lohr, C. C. Harris, L. Ding, R. K. Wilson, D. A. Wheeler, R. A. Gibbs, R. Kucherlapati, C. Lee, P. V. Kharchenko & P. J. Park (2012) Landscape of somatic retrotransposition in human cancers. *Science,* 337**,** 967-71.

Lee, J. M., K. W. Cho, E. J. Kim, Q. Tang, K. S. Kim, C. Tickle & H. S. Jung (2015) A contrasting function for miR-137 in embryonic mammogenesis and adult breast carcinogenesis. *Oncotarget,* 6**,** 22048-59.

Lee, S. T., Z. Li, Z. Wu, M. Aau, P. Guan, R. K. Karuturi, Y. C. Liou & Q. Yu (2011) Context-specific regulation of NF-kappaB target gene expression by EZH2 in breast cancers. *Mol Cell,* 43**,** 798-810.

Lee, Y., M. Kim, J. Han, K. H. Yeom, S. Lee, S. H. Baek & V. N. Kim (2004) MicroRNA genes are transcribed by RNA polymerase II. *Embo j,* 23**,** 4051-60.

Legendre, M., N. Pochet, T. Pak & K. J. Verstrepen (2007) Sequence-based estimation of minisatellite and microsatellite repeat variability. *Genome Res,* 17**,** 1787-96.

Lencz, T., E. Knowles, G. Davies, S. Guha, D. C. Liewald, J. M. Starr, S. Djurovic, I. Melle, K. Sundet, A. Christoforou, I. Reinvang, S. Mukherjee, P. DeRosse, A. Lundervold, V. M. Steen, M. John, T. Espeseth, K. Raikkonen, E. Widen, A. Palotie, J. G. Eriksson, I. Giegling, B. Konte, M. Ikeda, P. Roussos, S. Giakoumaki, K. E. Burdick, A. Payton, W. Ollier, M. Horan, G. Donohoe, D. Morris, A. Corvin, M. Gill, N. Pendleton, N. Iwata, A. Darvasi, P. Bitsios, D. Rujescu, J. Lahti, S. L. Hellard, M. C. Keller, O. A. Andreassen, I. J. Deary, D. C. Glahn & A. K. Malhotra (2014) Molecular genetic evidence for overlap between general cognitive ability and risk for schizophrenia: a report from the Cognitive Genomics consorTium (COGENT). *Mol Psychiatry,* 19**,** 168-74.

Lewandowska, E., W. Lipczynska-Lojkowska, J. Modzelewska, T. Wierzba-Bobrowicz, H. Mierzewska, G. M. Szpak, E. Passenik & K. Jachinska (2009) Kufs' disease: diagnostic difficulties in the examination of extracerebral biopsies. *Folia Neuropathol,* 47**,** 259-67.

Lewis, B. P., C. B. Burge & D. P. Bartel (2005) Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. *Cell,* 120**,** 15-20.

Li, B. J., P. Liu, Z. Chu, Y. Shang, M. X. Huan, Y. H. Dang & C. G. Gao (2017) Social isolation induces schizophrenia-like behavior potentially associated with HINT1, NMDA receptor 1, and dopamine receptor 2. *Neuroreport,* 28**,** 462-469.

Li, J., Y. You, W. Yue, H. Yu, T. Lu, Z. Wu, M. Jia, Y. Ruan, J. Liu, D. Zhang & L. Wang (2016) Chromatin remodeling gene EZH2 involved in the genetic etiology of autism in Chinese Han population. *Neurosci Lett,* 610**,** 182-6.

Li, W., Y. Jin, L. Prazak, M. Hammell & J. Dubnau. 2012. Transposable Elements in TDP-43-Mediated Neurodegenerative Disorders. In *PLoS One*.

Li, X., S. C. Hughes & R. Wevrick (2015) Evaluation of melanoma antigen (MAGE) gene expression in human cancers using The Cancer Genome Atlas. *Cancer Genet,* 208**,** 25-34.

Li, Y., W. Tang, L. R. Zhang & C. Y. Zhang (2014) FMRP regulates miR196a-mediated repression of HOXB8 via interaction with the AGO2 MID domain. *Mol Biosyst,* 10**,** 1757-64.

Liang, Y., D. Ridzon, L. Wong & C. Chen (2007) Characterization of microRNA expression profiles in normal human tissues. *BMC Genomics,* 8**,** 166.

Liao, Q., Y. Wang, J. Cheng, D. Dai, X. Zhou, Y. Zhang, S. Gao & S. Duan (2015a) DNA methylation patterns of protein coding genes and long noncoding RNAs in female schizophrenic patients. *Eur J Med Genet,* 58**,** 95-104.

Liao, Q., Y. Wang, J. Cheng, D. Dai, X. Zhou, Y. Zhang, J. Li, H. Yin, S. Gao & S. Duan (2015b) DNA methylation patterns of protein-coding genes and long non-coding RNAs in males with schizophrenia. *Mol Med Rep,* 12**,** 6568-76.

Liao, Y., G. K. Smyth & W. Shi (2014) featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics,* 30**,** 923-30.

Lin, M., D. Zhao, A. Hrabovsky, E. Pedrosa, D. Zheng & H. M. Lachman (2014) Heat shock alters the expression of schizophrenia and autism candidate genes in an

induced pluripotent stem cell model of the human telencephalon. *PLoS One,* 9**,** e94968.
Lipska, B. K., A. Deep-Soboslay, C. S. Weickert, T. M. Hyde, C. E. Martin, M. M. Herman & J. E. Kleinman (2006) Critical factors in gene expression in postmortem human brain: Focus on studies in schizophrenia. *Biol Psychiatry,* 60**,** 650-8.
Liu, B., X. Zhang, B. Hou, J. Li, C. Qiu, W. Qin, C. Yu & T. Jiang (2014a) The impact of MIR137 on dorsolateral prefrontal-hippocampal functional connectivity in healthy subjects. *Neuropsychopharmacology,* 39**,** 2153-60.
Liu, C. M., C. S. Fann, C. Y. Chen, Y. L. Liu, Y. J. Oyang, W. C. Yang, C. C. Chang, C. C. Wen, W. J. Chen, T. J. Hwang, M. H. Hsieh, C. C. Liu, S. V. Faraone, M. T. Tsuang & H. G. Hwu (2011) ANXA7, PPP3CB, DNAJC9, and ZMYND17 genes at chromosome 10q22 associated with the subgroup of schizophrenia with deficits in attention and executive function. *Biol Psychiatry,* 70**,** 51-8.
Liu, Y., Q. Zhu & N. Zhu (2008) Recent duplication and positive selection of the GAGE gene family. *Genetica,* 133**,** 31-5.
Liu, Z., X. Li, N. Sun, Y. Xu, Y. Meng, C. Yang, Y. Wang & K. Zhang (2014b) Microarray profiling and co-expression network analysis of circulating lncRNAs and mRNAs associated with major depressive disorder. *PLoS One,* 9**,** e93388.
Loch, A. A., C. Chianca, T. M. Alves, E. L. Freitas, L. Hortencio, J. C. Andrade, M. T. van de Bilt, M. R. Fontoni, M. H. Serpa, W. F. Gattaz & W. Rossler (2017) Poverty, low education, and the expression of psychotic-like experiences in the general population of Sao Paulo, Brazil. *Psychiatry Res,* 253**,** 182-188.
Lopez-Ortega, E., R. Ruiz & L. Tabares (2017) CSPalpha, a Molecular Co-chaperone Essential for Short and Long-Term Synaptic Maintenance. *Front Neurosci,* 11**,** 39.
Loscher, W. (2002) Basic pharmacology of valproate: a review after 35 years of clinical use for the treatment of epilepsy. *CNS Drugs,* 16**,** 669-94.
Lu, J. & A. G. Clark (2012) Impact of microRNA regulation on variation in human gene expression. *Genome Res,* 22**,** 1243-54.
Ludwig, N., P. Leidinger, K. Becker, C. Backes, T. Fehlmann, C. Pallasch, S. Rheinheimer, B. Meder, C. Stahler, E. Meese & A. Keller (2016) Distribution of miRNA expression across human tissues. *Nucleic Acids Res,* 44**,** 3865-77.
Lugli, G., J. Larson, M. P. Demars & N. R. Smalheiser (2012) Primary microRNA precursor transcripts are localized at post-synaptic densities in adult mouse forebrain. *J Neurochem,* 123**,** 459-66.
Lugli, G., J. Larson, M. E. Martone, Y. Jones & N. R. Smalheiser (2005) Dicer and eIF2c are enriched at postsynaptic densities in adult mouse brain and are modified by neuronal activity in a calpain-dependent manner. *J Neurochem,* 94**,** 896-905.
Lugli, G., V. I. Torvik, J. Larson & N. R. Smalheiser (2008) Expression of microRNAs and their precursors in synaptic fractions of adult mouse forebrain. *J Neurochem,* 106**,** 650-61.
Lund, E., S. Guttinger, A. Calado, J. E. Dahlberg & U. Kutay (2004) Nuclear export of microRNA precursors. *Science,* 303**,** 95-8.
Ma, Y., R. Fan & M. D. Li (2016) Meta-Analysis Reveals Significant Association of the 3'-UTR VNTR in SLC6A3 with Alcohol Dependence. *Alcohol Clin Exp Res,* 40**,** 1443-53.
MacKenzie, A. & J. Quinn (1999) A serotonin transporter gene intron 2 polymorphic region, correlated with affective disorders, has allele-dependent differential

enhancer-like properties in the mouse embryo. *Proc Natl Acad Sci U S A,* 96**,** 15251-5.
MacKenzie, A. & J. P. Quinn (2004) Post-genomic approaches to exploring neuropeptide gene mis-expression in disease. *Neuropeptides,* 38**,** 1-15.
Mahmoudi Saber, M. & N. Saitou (2017) Silencing effect of Hominoid highly conserved non-coding sequences on embryonic brain development. *Genome Biol Evol*.
Malki, K., A. Lourdusamy, E. Binder, J. Paya-Cano, F. Sluyter, I. Craig, R. Keers, P. McGuffin, R. Uher & L. C. Schalkwyk (2012) Antidepressant-dependent mRNA changes in mouse associated with hippocampal neurogenesis in a mouse model of depression. *Pharmacogenet Genomics,* 22**,** 765-76.
Mallard, T. T., J. Doorley, C. L. Esposito-Smythers & J. E. McGeary (2016) Dopamine D4 receptor VNTR polymorphism associated with greater risk for substance abuse among adolescents with disruptive behavior disorders: Preliminary results. *Am J Addict,* 25**,** 56-61.
Mallo, M., D. M. Wellik & J. Deschamps (2010) Hox genes and regional patterning of the vertebrate body plan. *Dev Biol,* 344**,** 7-15.
Mamdani, M., G. O. McMichael, V. Gadepalli, V. Williamson, E. K. Parker, V. Haroutunian & V. I. Vladimirov (2013) Differential Regulation of Schizophrenia-associated microRNA Gene Function by Variable Number Tandem Repeats (VNTR) polymorphism. *Schizophr Res,* 151**,** 284-6.
Manuelidis, L. (1978) Chromosomal localization of complex and simple repeated human DNAs. *Chromosoma,* 66**,** 23-32.
Martin, S. L. & F. D. Bushman (2001) Nucleic acid chaperone activity of the ORF1 protein from the mouse LINE-1 retrotransposon. *Mol Cell Biol,* 21**,** 467-75.
Martinez, J. G., J. Perez-Escuredo, P. Castro-Santos, C. A. Marcos, J. L. Pendas, M. F. Fraga & M. A. Hermsen (2012a) Hypomethylation of LINE-1, and not centromeric SAT-alpha, is associated with centromeric instability in head and neck squamous cell carcinoma. *Cell Oncol (Dordr),* 35**,** 259-67.
Martinez, J. G., J. Perez-Escuredo, J. L. Llorente, C. Suarez & M. A. Hermsen (2012b) Localization of centromeric breaks in head and neck squamous cell carcinoma. *Cancer Genet,* 205**,** 622-9.
Marzi, M. J., F. Ghini, B. Cerruti, S. de Pretis, P. Bonetti, C. Giacomelli, M. M. Gorski, T. Kress, M. Pelizzola, H. Muller, B. Amati & F. Nicassio (2016) Degradation dynamics of microRNAs revealed by a novel pulse-chase approach. *Genome Res,* 26**,** 554-65.
Mason, L., E. Peters, S. C. Williams & V. Kumari (2017) Brain connectivity changes occurring following cognitive behavioural therapy for psychosis predict long-term recovery. *Transl Psychiatry,* 7**,** e1001.
Mason, L., E. R. Peters, D. Dima, S. C. Williams & V. Kumari (2016) Cognitive Behavioral Therapy Normalizes Functional Connectivity for Social Threat in Psychosis. *Schizophr Bull,* 42**,** 684-92.
Mathias, S. L., A. F. Scott, H. H. Kazazian, Jr., J. D. Boeke & A. Gabriel (1991) Reverse transcriptase encoded by a human transposable element. *Science,* 254**,** 1808-10.
Matsunami, M. & N. Saitou (2013) Vertebrate paralogous conserved noncoding sequences may be related to gene expressions in brain. *Genome Biol Evol,* 5**,** 140-50.
McCarthy-Jones, S., L. K. Oestreich, A. E. Lyall, Z. Kikinis, D. T. Newell, P. Savadjiev, M. E. Shenton, M. Kubicki, O. Pasternak & T. J. Whitford (2017) Childhood adversity associated with white matter alteration in the corpus callosum, corona

radiata, and uncinate fasciculus of psychiatrically healthy adults. *Brain Imaging Behav*.
McKetin, R., S. Dawe, R. A. Burns, L. Hides, D. J. Kavanagh, M. Teesson, D. Y. R. Mc, A. Voce & J. B. Saunders (2016) The profile of psychiatric symptoms exacerbated by methamphetamine use. *Drug Alcohol Depend,* 161**,** 104-9.
Meechan, D. W., T. M. Maynard, E. S. Tucker, A. Fernandez, B. A. Karpinski, L. A. Rothblat & A. S. LaMantia (2015) Modeling a model: Mouse genetics, 22q11.2 Deletion Syndrome, and disorders of cortical circuit development. *Prog Neurobiol,* 130**,** 1-28.
Mehl, S., D. Werner & T. M. Lincoln (2015) Does Cognitive Behavior Therapy for psychosis (CBTp) show a sustainable effect on delusions? A meta-analysis. *Front Psychol,* 6**,** 1450.
Metzgar, D., J. Bytof & C. Wills. 2000. Selection Against Frameshift Mutations Limits Microsatellite Expansion in Coding DNA. In *Genome Res*, 72-80.
Mighell, A. J., A. F. Markham & P. A. Robinson (1997) Alu sequences. *FEBS Lett,* 417**,** 1-5.
Mills, R. E., E. A. Bennett, R. C. Iskow & S. E. Devine (2007) Which transposable elements are active in the human genome? *Trends Genet,* 23**,** 183-91.
Minichino, A., R. Delle Chiaie, G. Cruccu, S. Piroso, G. Di Stefano, M. Francesconi, F. S. Bersani, M. Biondi & A. Truini (2016) Pain-processing abnormalities in bipolar I disorder, bipolar II disorder, and schizophrenia: A novel trait marker for psychosis proneness and functional outcome? *Bipolar Disord,* 18**,** 591-601.
Misiak, B., E. Szmida, P. Karpinski, O. Loska, M. M. Sasiadek & D. Frydecka (2015) Lower LINE-1 methylation in first-episode schizophrenia patients with the history of childhood trauma. *Epigenomics,* 7**,** 1275-85.
Mistry, S., J. R. Harrison, D. J. Smith, V. Escott-Price & S. Zammit (2017) The use of polygenic risk scores to identify phenotypes associated with genetic risk of schizophrenia: Systematic review. *Schizophr Res*.
Moncini, S., A. Salvi, P. Zuccotti, G. Viero, A. Quattrone, S. Barlati, G. De Petro, M. Venturin & P. Riva (2011) The role of miR-103 and miR-107 in regulation of CDK5R1 expression and in cellular migration. *PLoS One,* 6**,** e20038.
Monks, S., M. Niarchou, A. R. Davies, J. T. Walters, N. Williams, M. J. Owen, M. B. van den Bree & K. C. Murphy (2014) Further evidence for high rates of schizophrenia in 22q11.2 deletion syndrome. *Schizophr Res,* 153**,** 231-6.
Mothersill, O., D. W. Morris, S. Kelly, E. J. Rose, C. Fahey, C. O'Brien, R. Lyne, R. Reilly, M. Gill, A. P. Corvin & G. Donohoe (2014) Effects of MIR137 on fronto-amygdala functional connectivity. *Neuroimage,* 90**,** 189-95.
Mukherjee, S., R. Brulet, L. Zhang & J. Hsieh (2016) REST regulation of gene networks in adult neural stem cells. *Nat Commun,* 7**,** 13360.
Nan, H., P. Kraft, D. J. Hunter & J. Han (2009) Genetic variants in pigmentation genes, pigmentary phenotypes, and risk of skin cancer in Caucasians. *Int J Cancer,* 125**,** 909-17.
Nelson, P. T. & W. X. Wang (2010) MiR-107 is reduced in Alzheimer's disease brain neocortex: validation study. *J Alzheimers Dis,* 21**,** 75-9.
Ngamphiw, C., S. Tongsima & A. Mutirangura. 2014. Roles of Intragenic and Intergenic L1s in Mouse and Human. In *PLoS One*.
Nilsson, E. M., K. B. Laursen, J. Whitchurch, A. McWilliam, N. Odum, J. L. Persson, D. M. Heery, L. J. Gudas & N. P. Mongan (2015) MiR137 is an androgen regulated repressor of an extended network of transcriptional coregulators. *Oncotarget,* 6**,** 35710-25.

Ninkina, N., O. M. Peters, N. Connor-Robson, O. Lytkina, E. Sharfeddin & V. L. Buchman (2012) Contrasting effects of alpha-synuclein and gamma-synuclein on the phenotype of cysteine string protein alpha (CSPalpha) null mutant mice suggest distinct function of these proteins in neuronal synapses. *J Biol Chem,* 287**,** 44471-7.

Nowick, K., C. Fields, T. Gernat, D. Caetano-Anolles, N. Kholina & L. Stubbs (2011) Gain, loss and divergence in primate zinc-finger genes: a rich resource for evolution of gene regulatory differences between species. *PLoS One,* 6**,** e21553.

Nowick, K., T. Gernat, E. Almaas & L. Stubbs (2009) Differences in human and chimpanzee gene expression patterns define an evolving network of transcription factors in brain. *Proc Natl Acad Sci U S A,* 106**,** 22358-63.

Nuechterlein, K. H. & M. E. Dawson (1984) A heuristic vulnerability/stress model of schizophrenic episodes. *Schizophr Bull,* 10**,** 300-12.

Nuevo, R., S. Chatterji, E. Verdes, N. Naidoo, C. Arango & J. L. Ayuso-Mateos (2012) The Continuum of Psychotic Symptoms in the General Population: A Cross-national Study. *Schizophr Bull,* 38**,** 475-85.

Okbay, A., J. P. Beauchamp, M. A. Fontana, J. J. Lee, T. H. Pers, C. A. Rietveld, P. Turley, G. B. Chen, V. Emilsson, S. F. Meddens, S. Oskarsson, J. K. Pickrell, K. Thom, P. Timshel, R. de Vlaming, A. Abdellaoui, T. S. Ahluwalia, J. Bacelis, C. Baumbach, G. Bjornsdottir, J. H. Brandsma, M. Pina Concas, J. Derringer, N. A. Furlotte, T. E. Galesloot, G. Girotto, R. Gupta, L. M. Hall, S. E. Harris, E. Hofer, M. Horikoshi, J. E. Huffman, K. Kaasik, I. P. Kalafati, R. Karlsson, A. Kong, J. Lahti, S. J. van der Lee, C. deLeeuw, P. A. Lind, K. O. Lindgren, T. Liu, M. Mangino, J. Marten, E. Mihailov, M. B. Miller, P. J. van der Most, C. Oldmeadow, A. Payton, N. Pervjakova, W. J. Peyrot, Y. Qian, O. Raitakari, R. Rueedi, E. Salvi, B. Schmidt, K. E. Schraut, J. Shi, A. V. Smith, R. A. Poot, B. St Pourcain, A. Teumer, G. Thorleifsson, N. Verweij, D. Vuckovic, J. Wellmann, H. J. Westra, J. Yang, W. Zhao, Z. Zhu, B. Z. Alizadeh, N. Amin, A. Bakshi, S. E. Baumeister, G. Biino, K. Bonnelykke, P. A. Boyle, H. Campbell, F. P. Cappuccio, G. Davies, J. E. De Neve, P. Deloukas, I. Demuth, J. Ding, P. Eibich, L. Eisele, N. Eklund, D. M. Evans, J. D. Faul, M. F. Feitosa, A. J. Forstner, I. Gandin, B. Gunnarsson, B. V. Halldorsson, T. B. Harris, A. C. Heath, L. J. Hocking, E. G. Holliday, G. Homuth, M. A. Horan, et al. (2016) Genome-wide association study identifies 74 loci associated with educational attainment. *Nature,* 533**,** 539-42.

Olde Loohuis, N. F., N. Nadif Kasri, J. C. Glennon, H. van Bokhoven, S. S. Hebert, B. B. Kaplan, G. J. Martens & A. Aschrafi (2017) The schizophrenia risk gene MIR137 acts as a hippocampal gene network node orchestrating the expression of genes relevant to nervous system development and function. *Prog Neuropsychopharmacol Biol Psychiatry,* 73**,** 109-118.

Pan, T., X. Li, W. Xie, J. Jankovic & W. Le (2005) Valproic acid-mediated Hsp70 induction and anti-apoptotic neuroprotection in SH-SY5Y cells. *FEBS Lett,* 579**,** 6716-20.

Paredes, U. M., V. J. Bubb, K. Haddley, G. A. Macho & J. P. Quinn (2011) An evolutionary conserved region (ECR) in the human dopamine receptor D4 gene supports reporter gene expression in primary cultures derived from the rat cortex. *BMC Neurosci,* 12**,** 46.

Paredes, U. M., J. P. Quinn & U. M. D'Souza (2013) Allele-specific transcriptional activity of the variable number of tandem repeats in 5' region of the DRD4 gene is stimulus specific in human neuronal cells. *Genes Brain Behav,* 12**,** 282-7.

Paschou, M., M. D. Paraskevopoulou, I. S. Vlachos, P. Koukouraki, A. G. Hatzigeorgiou & E. Doxakis (2012) miRNA Regulons Associated with Synaptic Function. *PLoS One,* 7.

Payton, A., P. Sindrewicz, V. Pessoa, H. Platt, M. Horan, W. Ollier, V. J. Bubb, N. Pendleton & J. P. Quinn (2016) A TOMM40 poly-T variant modulates gene expression and is associated with vocabulary ability and decline in nonpathologic aging. *Neurobiol Aging,* 39**,** 217.e1-7.

Peitl, V., M. Stefanovic & D. Karlovic (2017) Depressive symptoms in schizophrenia and dopamine and serotonin gene polymorphisms. *Prog Neuropsychopharmacol Biol Psychiatry,* 77**,** 209-215.

Perreault, A. A. & B. J. Venters (2016) The ChIP-exo Method: Identifying Protein-DNA Interactions with Near Base Pair Precision. *J Vis Exp*.

Peters, E., T. Crombie, D. Agbedjro, L. C. Johns, D. Stahl, K. Greenwood, N. Keen, J. Onwumere, E. Hunter, L. Smith & E. Kuipers (2015) The long-term effectiveness of cognitive behavior therapy for psychosis within a routine psychological therapies service. *Front Psychol,* 6.

Phiel, C. J., F. Zhang, E. Y. Huang, M. G. Guenther, M. A. Lazar & P. S. Klein (2001) Histone deacetylase is a direct target of valproic acid, a potent anticonvulsant, mood stabilizer, and teratogen. *J Biol Chem,* 276**,** 36734-41.

Pickles, A., J. Hill, G. Breen, J. Quinn, K. Abbott, H. Jones & H. Sharp (2013) Evidence for interplay between genes and parenting on infant temperament in the first year of life: monoamine oxidase A polymorphism moderates effects of maternal sensitivity on infant anger proneness. *J Child Psychol Psychiatry,* 54**,** 1308-17.

Pinner, A. L., J. Tucholski, V. Haroutunian, R. E. McCullumsmith & J. H. Meador-Woodruff (2016) Decreased protein S-palmitoylation in dorsolateral prefrontal cortex in schizophrenia. *Schizophr Res,* 177**,** 78-87.

Piunti, A. & A. Shilatifard (2016) Epigenetic balance of gene expression by Polycomb and COMPASS families. *Science,* 352**,** aad9780.

Plessy, C., T. Dickmeis, F. Chalmel & U. Strahle (2005) Enhancer sequence conservation between vertebrates is favoured in developmental regulator genes. *Trends Genet,* 21**,** 207-10.

Prabhakar, S., F. Poulin, M. Shoukry, V. Afzal, E. M. Rubin, O. Couronne & L. A. Pennacchio (2006) Close sequence comparisons are sufficient to identify human cis-regulatory elements. *Genome Res,* 16**,** 855-63.

Qi, L., J. L. Cao, Y. Hu, J. G. Yang, Y. Ji, J. Huang, Y. Zhang, D. G. Sun, H. F. Xia & X. Ma (2013) The dynamics of polycomb group proteins in early embryonic nervous system in mouse and human. *Int J Dev Neurosci,* 31**,** 487-95.

Quilez, J., A. Guilmatre, P. Garg, G. Highnam, M. Gymrek, Y. Erlich, R. S. Joshi, D. Mittelman & A. J. Sharp (2016) Polymorphic tandem repeats within gene promoters act as modifiers of gene expression and DNA methylation in humans. *Nucleic Acids Res,* 44**,** 3750-62.

Quinn, J. P., A. Warburton, P. Myers, A. L. Savage & V. J. Bubb (2013) Polymorphic variation as a driver of differential neuropeptide gene expression. *Neuropeptides,* 47**,** 395-400.

Qureshi, I. A. & M. F. Mehler (2009) Regulation of non-coding RNA networks in the nervous system--what's the REST of the story? *Neurosci Lett,* 466**,** 73-80.

Raiz, J., A. Damert, S. Chira, U. Held, S. Klawitter, M. Hamdorf, J. Lower, W. H. Stratling, R. Lower & G. G. Schumann (2012) The non-autonomous retrotransposon SVA is trans-mobilized by the human LINE-1 protein machinery. *Nucleic Acids Res,* 40**,** 1666-83.

Rajendiran, S., L. D. Gibbs, T. Van Treuren, D. L. Klinkebiel & J. K. Vishwanatha (2016) MIEN1 is tightly regulated by SINE Alu methylation in its promoter. *Oncotarget,* 7**,** 65307-65319.

Rao, S. Q., H. L. Hu, N. Ye, Y. Shen & Q. Xu (2015) Genetic variants in long non-coding RNA MIAT contribute to risk of paranoid schizophrenia in a Chinese Han population. *Schizophr Res,* 166**,** 125-30.

Read, J., R. Fosse, A. Moskowitz & B. Perry (2014) The traumagenic neurodevelopmental model of psychosis revisited. *Neuropsychiatry,* 4**,** 65-79.

Read, J., B. D. Perry, A. Moskowitz & J. Connolly (2001) The contribution of early traumatic events to schizophrenia in some patients: a traumagenic neurodevelopmental model. *Psychiatry,* 64**,** 319-45.

Reguenga, C., M. E. Oliveira, A. M. Gouveia, C. Eckerskorn, C. Sa-Miranda & J. E. Azevedo (1999) Identification of a 24 kDa intrinsic membrane protein from mammalian peroxisomes. *Biochim Biophys Acta,* 1445**,** 337-41.

Reiner, A., I. Dragatsis & P. Dietrich (2011) Genetics and neuropathology of Huntington's Disease. *Int Rev Neurobiol,* 98**,** 325-72.

Ren, H., Z. Gao, N. Wu, L. Zeng, X. Tang, X. Chen, Z. Liu, W. Zhang, L. Wang & Z. Li (2015a) Expression of REST4 in human gliomas in vivo and influence of pioglitazone on REST in vitro. *Biochem Biophys Res Commun,* 463**,** 504-9.

Ren, X., X. Bai, X. Zhang, Z. Li, L. Tang, X. Zhao, Z. Li, Y. Ren, S. Wei, Q. Wang, C. Liu & J. Ji (2015b) Quantitative nuclear proteomics identifies that miR-137-mediated EZH2 reduction regulates resveratrol-induced apoptosis of neuroblastoma cells. *Mol Cell Proteomics,* 14**,** 316-28.

Ren, Y., Y. Cui, X. Li, B. Wang, L. Na, J. Shi, L. Wang, L. Qiu, K. Zhang, G. Liu & Y. Xu (2015c) A co-expression network analysis reveals lncRNA abnormalities in peripheral blood in early-onset schizophrenia. *Prog Neuropsychopharmacol Biol Psychiatry,* 63**,** 1-5.

Rhodes, D. & H. J. Lipps (2015) G-quadruplexes and their regulatory roles in biology. *Nucleic Acids Res,* 43**,** 8627-37.

Richardson, S. R., S. Morell & G. J. Faulkner (2014) L1 retrotransposons and somatic mosaicism in the brain. *Annu Rev Genet,* 48**,** 1-27.

Ripke, S., C. O'Dushlaine, K. Chambert, J. L. Moran, A. K. Kahler, S. Akterin, S. E. Bergen, A. L. Collins, J. J. Crowley, M. Fromer, Y. Kim, S. H. Lee, P. K. Magnusson, N. Sanchez, E. A. Stahl, S. Williams, N. R. Wray, K. Xia, F. Bettella, A. D. Borglum, B. K. Bulik-Sullivan, P. Cormican, N. Craddock, C. de Leeuw, N. Durmishi, M. Gill, V. Golimbet, M. L. Hamshere, P. Holmans, D. M. Hougaard, K. S. Kendler, K. Lin, D. W. Morris, O. Mors, P. B. Mortensen, B. M. Neale, F. A. O'Neill, M. J. Owen, M. P. Milovancevic, D. Posthuma, J. Powell, A. L. Richards, B. P. Riley, D. Ruderfer, D. Rujescu, E. Sigurdsson, T. Silagadze, A. B. Smit, H. Stefansson, S. Steinberg, J. Suvisaari, S. Tosato, M. Verhage, J. T. Walters, C. Multicenter Genetic Studies of Schizophrenia, D. F. Levinson, P. V. Gejman, K. S. Kendler, C. Laurent, B. J. Mowry, M. C. O'Donovan, M. J. Owen, A. E. Pulver, B. P. Riley, S. G. Schwab, D. B. Wildenauer, F. Dudbridge, P. Holmans, J. Shi, M. Albus, M. Alexander, D. Campion, D. Cohen, D. Dikeos, J. Duan, P. Eichhammer, S. Godard, M. Hansen, F. B. Lerer, K. Y. Liang, W. Maier, J. Mallet, D. A. Nertney, G. Nestadt,

N. Norton, F. A. O'Neill, G. N. Papadimitriou, R. Ribble, A. R. Sanders, J. M. Silverman, D. Walsh, N. M. Williams, B. Wormley, C. Psychosis Endophenotypes International, M. J. Arranz, S. Bakker, S. Bender, E. Bramon, D. Collier, B. Crespo-Facorro, et al. (2013) Genome-wide association analysis identifies 13 new risk loci for schizophrenia. *Nat Genet,* 45**,** 1150-9.

Ritchie, M. E., B. Phipson, D. Wu, Y. Hu, C. W. Law, W. Shi & G. K. Smyth (2015) limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res,* 43**,** e47.

Riva, P., A. Ratti & M. Venturin (2016) The Long Non-Coding RNAs in Neurodegenerative Diseases: Novel Mechanisms of Pathogenesis. *Curr Alzheimer Res,* 13**,** 1219-1231.

Robbez-Masson, L. & H. M. Rowe (2015) Retrotransposons shape species-specific embryonic stem cell gene expression. *Retrovirology,* 12**,** 45.

Roberts, J., A. C. Scott, M. R. Howard, G. Breen, V. J. Bubb, E. Klenova & J. P. Quinn (2007) Differential regulation of the serotonin transporter gene by lithium is mediated by transcription factors, CCCTC binding protein and Y-box binding protein 1, through the polymorphic intron 2 variable number tandem repeat. *J Neurosci,* 27**,** 2793-801.

Roberts, S., K. J. Lester, J. L. Hudson, R. M. Rapee, C. Creswell, P. J. Cooper, K. J. Thirlwall, J. R. Coleman, G. Breen, C. C. Wong & T. C. Eley (2014) Serotonin transporter methylation and response to cognitive behaviour therapy in children with anxiety disorders. *Transl Psychiatry,* 4**,** e444.

Rodriguez, A., S. Griffiths-Jones, J. L. Ashurst & A. Bradley. 2004. Identification of Mammalian microRNA Host Genes and Transcription Units. In *Genome Res*, 1902-10.

Rohrer, J. D., A. M. Isaacs, S. Mizielinska, S. Mead, T. Lashley, S. Wray, K. Sidle, P. Fratta, R. W. Orrell, J. Hardy, J. Holton, T. Revesz, M. N. Rossor & J. D. Warren (2015) C9orf72 expansions in frontotemporal dementia and amyotrophic lateral sclerosis. *Lancet Neurol,* 14**,** 291-301.

Ronan, J. L., W. Wu & G. R. Crabtree (2013) From neural development to cognition: unexpected roles for chromatin. *Nat Rev Genet,* 14**,** 347-59.

Roncero, C., C. Daigre, L. Grau-Lopez, C. Barral, J. Perez-Pazos, N. Martinez-Luna & M. Casas (2014) An international perspective and review of cocaine-induced psychosis: a call to action. *Subst Abus,* 35**,** 321-7.

Ross, M. T., D. V. Grafham, A. J. Coffey, S. Scherer, K. McLay, D. Muzny, M. Platzer, G. R. Howell, C. Burrows, C. P. Bird, A. Frankish, F. L. Lovell, K. L. Howe, J. L. Ashurst, R. S. Fulton, R. Sudbrak, G. Wen, M. C. Jones, M. E. Hurles, T. D. Andrews, C. E. Scott, S. Searle, J. Ramser, A. Whittaker, R. Deadman, N. P. Carter, S. E. Hunt, R. Chen, A. Cree, P. Gunaratne, P. Havlak, A. Hodgson, M. L. Metzker, S. Richards, G. Scott, D. Steffen, E. Sodergren, D. A. Wheeler, K. C. Worley, R. Ainscough, K. D. Ambrose, M. A. Ansari-Lari, S. Aradhya, R. I. Ashwell, A. K. Babbage, C. L. Bagguley, A. Ballabio, R. Banerjee, G. E. Barker, K. F. Barlow, I. P. Barrett, K. N. Bates, D. M. Beare, H. Beasley, O. Beasley, A. Beck, G. Bethel, K. Blechschmidt, N. Brady, S. Bray-Allen, A. M. Bridgeman, A. J. Brown, M. J. Brown, D. Bonnin, E. A. Bruford, C. Buhay, P. Burch, D. Burford, J. Burgess, W. Burrill, J. Burton, J. M. Bye, C. Carder, L. Carrel, J. Chako, J. C. Chapman, D. Chavez, E. Chen, G. Chen, Y. Chen, Z. Chen, C. Chinault, A. Ciccodicola, S. Y. Clark, G. Clarke, C. M. Clee, S. Clegg, K. Clerc-Blankenburg, K. Clifford, V. Cobley, C. G. Cole, J. S. Conquer, N. Corby, R. E. Connor, R.

David, J. Davies, C. Davis, J. Davis, O. Delgado, D. Deshazo, et al. (2005) The DNA sequence of the human X chromosome. *Nature,* 434**,** 325-37.
Rossbach, M. (2011) Non-Coding RNAs in Neural Networks, REST-Assured. *Front Genet,* 2**,** 8.
Rotenberg, M., A. Tuck & K. McKenzie (2017) Psychosocial stressors contributing to emergency psychiatric service utilization in a sample of ethno-culturally diverse clients with psychosis in Toronto. *BMC Psychiatry,* 17**,** 324.
Rowe, H. M., A. Kapopoulou, A. Corsinotti, L. Fasching, T. S. Macfarlan, Y. Tarabay, S. Viville, J. Jakobsson, S. L. Pfaff & D. Trono (2013) TRIM28 repression of retrotransposon-based enhancers is necessary to preserve transcriptional dynamics in embryonic stem cells. *Genome Res,* 23**,** 452-61.
Ruiz, R., I. A. Biea & L. Tabares (2014) alpha-Synuclein A30P decreases neurodegeneration and increases synaptic vesicle release probability in CSPalpha-null mice. *Neuropharmacology,* 76 Pt A**,** 106-17.
Rusiecki, J. A., L. Chen, V. Srikantan, L. Zhang, L. Yan, M. L. Polin & A. Baccarelli (2012) DNA methylation in repetitive elements and post-traumatic stress disorder: a case–control study of US military service members. *Epigenomics,* 4.
Sabol, S. Z., S. Hu & D. Hamer (1998) A functional polymorphism in the monoamine oxidase A gene promoter. *Hum Genet,* 103**,** 273-9.
Sadee, W., K. Hartmann, M. Seweryn, M. Pietrzak, S. K. Handelman & G. A. Rempala (2014) Missing heritability of common diseases and treatments outside the protein-coding exome. *Hum Genet,* 133**,** 1199-215.
Sagar, S. M., F. R. Sharp & T. Curran (1988) Expression of c-fos protein in brain: metabolic mapping at the cellular level. *Science,* 240**,** 1328-31.
Sahakyan, A. B., P. Murat, C. Mayer & S. Balasubramanian (2017) G-quadruplex structures within the 3' UTR of LINE-1 elements stimulate retrotransposition. *Nat Struct Mol Biol,* 24**,** 243-247.
San, L., M. Bernardo, A. Gomez & M. Pena (2013) Factors associated with relapse in patients with schizophrenia. *Int J Psychiatry Clin Pract,* 17**,** 2-9.
Sandelin, A., P. Bailey, S. Bruce, P. G. Engstrom, J. M. Klos, W. W. Wasserman, J. Ericson & B. Lenhard (2004) Arrays of ultraconserved non-coding regions span the loci of key developmental genes in vertebrate genomes. *BMC Genomics,* 5**,** 99.
Sandyk, R. (1981) Adult neuronal ceroid lipofuscinosis (Kufs' disease). A sporadic case. *S Afr Med J,* 60**,** 754-5.
Santarelli, D. M., N. J. Beveridge, P. A. Tooney & M. J. Cairns (2011) Upregulation of dicer and microRNA expression in the dorsolateral prefrontal cortex Brodmann area 46 in schizophrenia. *Biol Psychiatry,* 69**,** 180-7.
Santos, F. R., A. Pandya, M. Kayser, R. J. Mitchell, A. Liu, L. Singh, G. Destro-Bisol, A. Novelletto, R. Qamar, S. Q. Mehdi, R. Adhikari, P. de Knijff & C. Tyler-Smith (2000) A polymorphic L1 retroposon insertion in the centromere of the human Y chromosome. *Hum Mol Genet,* 9**,** 421-30.
Sariaslan, A., S. Fazel, B. M. D'Onofrio, N. Langstrom, H. Larsson, S. E. Bergen, R. Kuja-Halkola & P. Lichtenstein (2016) Schizophrenia and subsequent neighborhood deprivation: revisiting the social drift hypothesis using population, twin and molecular genetic data. *Transl Psychiatry,* 6**,** e796.
Savage, A., V. Bubb & J. Quinn (2013a) What role do human specific retrotransposons play in mental health and behaviour? *Current Trends in Neurology,* 7**,** 57 - 68.

Savage, A. L., V. J. Bubb, G. Breen & J. P. Quinn (2013b) Characterisation of the potential function of SVA retrotransposons to modulate gene expression patterns. *BMC Evol Biol,* 13**,** 101.

Savage, A. L., Bubb, V.J., Quinn, J.P. (2013) What role do human specific retrotransposons play in mental health and behaviour? *Current Trends in Neurology,* 7**,** 57-68.

Savage, A. L., T. P. Wilm, K. Khursheed, A. Shatunov, K. E. Morrison, P. J. Shaw, C. E. Shaw, B. Smith, G. Breen, A. Al-Chalabi, D. Moss, V. J. Bubb & J. P. Quinn (2014) An evaluation of a SVA retrotransposon in the FUS promoter as a transcriptional regulator and its association to ALS. *PLoS One,* 9**,** e90833.

Sawaya, S., A. Bagshaw, E. Buschiazzo, P. Kumar, S. Chowdhury, M. A. Black & N. Gemmell (2013) Microsatellite tandem repeats are abundant in human promoters and are associated with regulatory elements. *PLoS One,* 8**,** e54710.

Schizophrenia Psychiatric Genome-Wide Association Study (GWAS) Consortium (2011) Genome-wide association study identifies five new schizophrenia loci. *Nat Genet,* 43**,** 969-76.

Schizophrenia Psychiatric Genome-Wide Association Study, C. (2011) Genome-wide association study identifies five new schizophrenia loci. *Nat Genet,* 43**,** 969-76.

Schizophrenia Working Group of the Psychiatric Genomics Consortium (2014) Biological insights from 108 schizophrenia-associated genetic loci. *Nature,* 511**,** 421-7.

Schneider, A., C. Johnston, F. Tassone, S. Sansone, R. J. Hagerman, E. Ferrer, S. M. Rivera & D. Hessl (2016) Broad autism spectrum and obsessive-compulsive symptoms in adults with the fragile X premutation. *Clin Neuropsychol,* 30**,** 929-43.

Schneider, M., M. Debbane, A. S. Bassett, E. W. Chow, W. L. Fung, M. van den Bree, M. Owen, K. C. Murphy, M. Niarchou, W. R. Kates, K. M. Antshel, W. Fremont, D. M. McDonald-McGinn, R. E. Gur, E. H. Zackai, J. Vorstman, S. N. Duijff, P. W. Klaassen, A. Swillen, D. Gothelf, T. Green, A. Weizman, T. Van Amelsvoort, L. Evers, E. Boot, V. Shashi, S. R. Hooper, C. E. Bearden, M. Jalbrzikowski, M. Armando, S. Vicari, D. G. Murphy, O. Ousley, L. E. Campbell, T. J. Simon & S. Eliez (2014) Psychiatric disorders from childhood to adulthood in 22q11.2 deletion syndrome: results from the International Consortium on Brain and Behavior in 22q11.2 Deletion Syndrome. *Am J Psychiatry,* 171**,** 627-39.

Scott, D. & C. A. Tamminga (2018) Effects of genetic and environmental risk for schizophrenia on hippocampal activity and psychosis-like behavior in mice. *Behav Brain Res,* 339**,** 114-123.

Sham, P. C. & D. Curtis (1995) Monte Carlo tests for associations between disease and alleles at highly polymorphic loci. *Ann Hum Genet,* 59**,** 97-105.

Shevlin, M., J. E. Houston, M. J. Dorahy & G. Adamson (2008) Cumulative traumas and psychosis: an analysis of the national comorbidity survey and the British Psychiatric Morbidity Survey. *Schizophr Bull,* 34**,** 193-9.

Shevlin, M., J. Murphy, J. Read, J. Mallett, G. Adamson & J. E. Houston (2011) Childhood adversity and hallucinations: a community-based study using the National Comorbidity Survey Replication. *Soc Psychiatry Psychiatr Epidemiol,* 46**,** 1203-10.

Shi, C., L. Zhang & C. Qin (2017) Long Non-coding RNAs in Brain Development, Synaptic Biology, and Alzheimer's Disease. *Brain Res Bull*.

Shin, Y. J., V. Kumarasamy, D. Camacho & D. Sun (2015) Involvement of G-quadruplex structures in regulation of human RET gene expression by small

molecules in human medullary thyroid carcinoma TT cells. *Oncogene,* 34**,** 1292-9.
Shpyleva, S., S. Melnyk, O. Pavliv, I. Pogribny & S. Jill James (2017) Overexpression of LINE-1 Retrotransposons in Autism Brain. *Mol Neurobiol*.
Shukla, R., K. R. Upton, M. Munoz-Lopez, D. J. Gerhardt, M. E. Fisher, T. Nguyen, P. M. Brennan, J. K. Baillie, A. Collino, S. Ghisletti, S. Sinha, F. Iannelli, E. Radaelli, A. Dos Santos, D. Rapoud, C. Guettier, D. Samuel, G. Natoli, P. Carninci, F. D. Ciccarelli, J. L. Garcia-Perez, J. Faivre & G. J. Faulkner (2013) Endogenous retrotransposition activates oncogenic pathways in hepatocellular carcinoma. *Cell,* 153**,** 101-11.
Shyu, K. G., B. W. Wang, Y. H. Yang, S. C. Tsai, S. Lin & C. C. Lee (2004) Amphetamine activates connexin43 gene expression in cultured neonatal rat cardiomyocytes through JNK and AP-1 pathway. *Cardiovasc Res,* 63**,** 98-108.
Siegert, S., J. Seo, E. J. Kwon, A. Rudenko, S. Cho, W. Wang, Z. Flood, A. J. Martorell, M. Ericsson, A. E. Mungenast & L. H. Tsai (2015) The schizophrenia risk gene product miR-137 alters presynaptic plasticity. *Nat Neurosci,* 18**,** 1008-16.
Singh, L. P., A. R. Aroor & A. J. Wahba (1994) Translational control of eukaryotic gene expression. Role of the guanine nucleotide exchange factor and chain initiation factor-2. *Enzyme Protein,* 48**,** 61-80.
Singh, L. P. & A. J. Wahba (1996) Regulation of protein synthesis in eukaryotic cells by the guanine nucleotide exchange factor and chain initiation factor 2. *SAAS Bull Biochem Biotechnol,* 9**,** 1-8.
Smalheiser, N. R., G. Lugli, H. Zhang, H. Rizavi, E. H. Cook & Y. Dwivedi (2014) Expression of microRNAs and Other Small RNAs in Prefrontal Cortex in Schizophrenia, Bipolar Disorder and Depressed Subjects. *PLoS One,* 9.
Smrt, R. D., K. E. Szulwach, R. L. Pfeiffer, X. Li, W. Guo, M. Pathania, Z. Q. Teng, Y. Luo, J. Peng, A. Bordey, P. Jin & X. Zhao (2010) MicroRNA miR-137 regulates neuronal maturation by targeting ubiquitin ligase mind bomb-1. *Stem Cells,* 28**,** 1060-70.
Somel, M., X. Liu, L. Tang, Z. Yan, H. Hu, S. Guo, X. Jiang, X. Zhang, G. Xu, G. Xie, N. Li, Y. Hu, W. Chen, S. Paabo & P. Khaitovich (2011) MicroRNA-driven developmental remodeling in the brain distinguishes humans from other primates. *PLoS Biol,* 9**,** e1001214.
Spadaro, P. A., C. R. Flavell, J. Widagdo, V. S. Ratnu, M. Troup, C. Ragan, J. S. Mattick & T. W. Bredy (2015) Long Noncoding RNA-Directed Epigenetic Regulation of Gene Expression Is Associated with Anxiety-like Behavior in Mice. *Biol Psychiatry*.
Spencer, E. M., K. E. Chandler, K. Haddley, M. R. Howard, D. Hughes, N. D. Belyaev, J. M. Coulson, J. P. Stewart, N. J. Buckley, A. Kipar, M. C. Walker & J. P. Quinn (2006) Regulation and role of REST and REST4 variants in modulation of gene expression in in vivo and in vitro in epilepsy models. *Neurobiol Dis,* 24**,** 41-52.
Stewart, C., D. Kural, M. P. Stromberg, J. A. Walker, M. K. Konkel, A. M. Stutz, A. E. Urban, F. Grubert, H. Y. Lam, W. P. Lee, M. Busby, A. R. Indap, E. Garrison, C. Huff, J. Xing, M. P. Snyder, L. B. Jorde, M. A. Batzer, J. O. Korbel & G. T. Marth (2011) A comprehensive map of mobile element insertion polymorphisms in humans. *PLoS Genet,* 7**,** e1002236.
Stolf, A. R., C. M. Szobot, R. Halpern, G. C. Akutagava-Martins, D. Muller, L. S. Guimaraes, F. H. Kessler, F. Pechansky & T. Roman (2014) Crack cocaine users show differences in genotype frequencies of the 3' UTR variable number

of tandem repeats of the dopamine transporter gene (DAT1/SLC6A3). *Neuropsychobiology,* 70**,** 44-51.

Strazisar, M., S. Cammaerts, K. van der Ven, D. A. Forero, A. S. Lenaerts, A. Nordin, L. Almeida-Souza, G. Genovese, V. Timmerman, A. Liekens, P. De Rijk, R. Adolfsson, P. Callaerts & J. Del-Favero (2015) MIR137 variants identified in psychiatric patients affect synaptogenesis and neuronal transmission gene sets. *Mol Psychiatry,* 20**,** 472-81.

Strichman-Almashanu, L. Z., R. S. Lee, P. O. Onyango, E. Perlman, F. Flam, M. B. Frieman & A. P. Feinberg. 2002. A Genome-Wide Screen for Normally Methylated Human CpG Islands That Can Identify Novel Imprinted Genes. In *Genome Res*, 543-54.

Sudmant, P. H., T. Rausch, E. J. Gardner, R. E. Handsaker, A. Abyzov, J. Huddleston, Y. Zhang, K. Ye, G. Jun, M. H.-Y. Fritz, M. K. Konkel, A. Malhotra, A. M. Stütz, X. Shi, F. P. Casale, J. Chen, F. Hormozdiari, G. Dayama, K. Chen, M. Malig, M. J. P. Chaisson, K. Walter, S. Meiers, S. Kashin, E. Garrison, A. Auton, H. Y. K. Lam, X. J. Mu, C. Alkan, D. Antaki, T. Bae, E. Cerveira, P. Chines, Z. Chong, L. Clarke, E. Dal, L. Ding, S. Emery, X. Fan, M. Gujral, F. Kahveci, J. M. Kidd, Y. Kong, E.-W. Lameijer, S. McCarthy, P. Flicek, R. A. Gibbs, G. Marth, C. E. Mason, A. Menelaou, D. M. Muzny, B. J. Nelson, A. Noor, N. F. Parrish, M. Pendleton, A. Quitadamo, B. Raeder, E. E. Schadt, M. Romanovitch, A. Schlattl, R. Sebra, A. A. Shabalin, A. Untergasser, J. A. Walker, M. Wang, F. Yu, C. Zhang, J. Zhang, X. Zheng-Bradley, W. Zhou, T. Zichner, J. Sebat, M. A. Batzer, S. A. McCarroll, T. G. P. Consortium, R. E. Mills, M. B. Gerstein, A. Bashir, O. Stegle, S. E. Devine, C. Lee, E. E. Eichler & J. O. Korbel (2015) An integrated map of structural variation in 2,504 human genomes. *Nature,* 526**,** 75-81.

Sullivan, B. A. & G. H. Karpen (2004) Centromeric chromatin exhibits a histone modification pattern that is distinct from both euchromatin and heterochromatin. *Nat Struct Mol Biol,* 11**,** 1076-83.

Sun, J., G. Zheng, Z. Gu & Z. Guo (2015a) MiR-137 inhibits proliferation and angiogenesis of human glioblastoma cells by targeting EZH2. *J Neurooncol,* 122**,** 481-9.

Sun, L., L. Min, H. Zhou, M. Li, F. Shao & W. Wang (2017) Adolescent social isolation affects schizophrenia-like behavior and astrocyte biomarkers in the PFC of adult rats. *Behav Brain Res,* 333**,** 258-266.

Sun, X. Y., J. Lu, L. Zhang, H. T. Song, L. Zhao, H. M. Fan, A. F. Zhong, W. Niu, Z. M. Guo, Y. H. Dai, C. Chen, Y. F. Ding & L. Y. Zhang (2015b) Aberrant microRNA expression in peripheral plasma and mononuclear cells as specific blood-based biomarkers in schizophrenia patients. *J Clin Neurosci,* 22**,** 570-4.

Sun, Z., Z. Wu, F. Zhang, Q. Guo, L. Li, K. Li, H. Chen, J. Zhao, D. Song, Q. Huang & J. Xiao (2016) PRAME is critical for breast cancer growth and metastasis. *Gene,* 594**,** 160-164.

Swergold, G. D. (1990) Identification, characterization, and cell specificity of a human LINE-1 promoter. *Mol Cell Biol,* 10**,** 6718-29.

Szczepanski, M. J., A. B. DeLeo, M. Luczak, M. Molinska-Glura, J. Misiak, B. Szarzynska, G. Dworacki, M. Zagor, N. Rozwadowska, M. Kurpisz, A. Krzeski, A. Kruk-Zagajewska, T. Kopec, J. Banaszewski & T. L. Whiteside (2013) PRAME expression in head and neck cancer correlates with markers of poor prognosis and might help in selecting candidates for retinoid chemoprevention in pre-malignant lesions. *Oral Oncol,* 49**,** 144-51.

Szulwach, K. E., X. Li, R. D. Smrt, Y. Li, Y. Luo, L. Lin, N. J. Santistevan, W. Li, X. Zhao & P. Jin (2010) Cross talk between microRNA and epigenetic regulation in adult neurogenesis. *J Cell Biol,* 189**,** 127-41.

Tabatabaei, S. M., S. Amiri, S. Faghfouri, S. G. Noorazar, S. AbdollahiFakhim & A. Fakhari (2017) DRD4 Gene Polymorphisms as a Risk Factor for Children with Attention Deficit Hyperactivity Disorder in Iranian Population. *Int Sch Res Notices,* 2017**,** 2494537.

Tabor, H. K., N. J. Risch & R. M. Myers (2002) Candidate-gene approaches for studying complex genetic traits: practical considerations. *Nat Rev Genet,* 3**,** 391-7.

Tak, Y. G. & P. J. Farnham. 2015. Making sense of GWAS: using epigenomics and genome engineering to understand the functional relevance of SNPs in non-coding regions of the human genome. In *Epigenetics Chromatin*.

Tamminga, C. A. & R. S. Zukin (2015) Schizophrenia: Evidence implicating hippocampal GluN2B protein and REST epigenetics in psychosis pathophysiology. *Neuroscience,* 309**,** 233-42.

Tan, H., A. Qurashi, M. Poidevin, D. L. Nelson, H. Li & P. Jin (2012) Retrotransposon activation contributes to fragile X premutation rCGG-mediated neurodegeneration. *Hum Mol Genet,* 21**,** 57-65.

Tang, K. L., K. M. Antshel, W. P. Fremont & W. R. Kates (2015) Behavioral and Psychiatric Phenotypes in 22q11.2 Deletion Syndrome. *J Dev Behav Pediatr,* 36**,** 639-50.

Tempelaar, W. M., F. Termorshuizen, J. H. MacCabe, M. P. Boks & R. S. Kahn (2017) Educational achievement in psychiatric patients and their siblings: a register-based study in 30 000 individuals in The Netherlands. *Psychol Med,* 47**,** 776-784.

The ENCODE Project Consortium (2012) An Integrated Encyclopedia of DNA Elements in the Human Genome. *Nature,* 489**,** 57-74.

Tiwari, S. S., M. d'Orange, C. Troakes, B. N. Shurovi, O. Engmann, W. Noble, T. Hortobagyi & K. P. Giese (2015) Evidence that the presynaptic vesicle protein CSPalpha is a key player in synaptic degeneration and protection in Alzheimer's disease. *Mol Brain,* 8**,** 6.

Tobo, M., Y. Mitsuyama, K. Ikari & K. Itoi (1984) Familial occurrence of adult-type neuronal ceroid lipofuscinosis. *Arch Neurol,* 41**,** 1091-4.

Tsai, M. C., O. Manor, Y. Wan, N. Mosammaparast, J. K. Wang, F. Lan, Y. Shi, E. Segal & H. Y. Chang (2010) Long noncoding RNA as modular scaffold of histone modification complexes. *Science,* 329**,** 689-93.

Uchida, S., K. Hara, A. Kobayashi, H. Funato, T. Hobara, K. Otsuki, H. Yamagata, B. S. McEwen & Y. Watanabe (2010) Early life stress enhances behavioral vulnerability to stress through the activation of REST4-mediated gene transcription in the medial prefrontal cortex of rodents. *J Neurosci,* 30**,** 15007-18.

Ulitsky, I. & D. P. Bartel (2013) lincRNAs: genomics, evolution, and mechanisms. *Cell,* 154**,** 26-46.

Ullu, E. & C. Tschudi (1984) Alu sequences are processed 7SL RNA genes. *Nature,* 312**,** 171-2.

Vaiserman, A. M. (2015) Epigenetic programming by early-life stress: Evidence from human populations. *Dev Dyn,* 244**,** 254-65.

Vance, C., B. Rogelj, T. Hortobagyi, K. J. De Vos, A. L. Nishimura, J. Sreedharan, X. Hu, B. Smith, D. Ruddy, P. Wright, J. Ganesalingam, K. L. Williams, V. Tripathi,

S. Al-Saraj, A. Al-Chalabi, P. N. Leigh, I. P. Blair, G. Nicholson, J. de Belleroche, J. M. Gallo, C. C. Miller & C. E. Shaw (2009) Mutations in FUS, an RNA processing protein, cause familial amyotrophic lateral sclerosis type 6. *Science,* 323**,** 1208-11.

Vaquerizas, J. M., S. K. Kummerfeld, S. A. Teichmann & N. M. Luscombe. 2009. A census of human transcription factors: function, expression and evolution. In *Nat Rev Genet*, 252-63. England.

Varese, F., F. Smeets, M. Drukker, R. Lieverse, T. Lataster, W. Viechtbauer, J. Read, J. van Os & R. P. Bentall (2012) Childhood adversities increase the risk of psychosis: a meta-analysis of patient-control, prospective- and cross-sectional cohort studies. *Schizophr Bull,* 38**,** 661-71.

Vasconcelos, A. C., S. Neto Ede, G. R. Pinto, F. K. Yoshioka, F. J. Motta, D. F. Vasconcelos & R. Canalle (2015) Association study of the SLC6A3 VNTR (DAT) and DRD2/ANKK1 Taq1A polymorphisms with alcohol dependence in a population from northeastern Brazil. *Alcohol Clin Exp Res,* 39**,** 205-11.

Vasieva, O., S. Cetiner, A. Savage, G. G. Schumann, V. J. Bubb & J. P. Quinn (2016a) Primate specific retrotransposons, SVAs, in the evolution of networks that alter brain function. *arXiv:1602.07642 [q-bio.NC]*.

Vasiliou, S. A., F. R. Ali, K. Haddley, M. C. Cardoso, V. J. Bubb & J. P. Quinn (2012) The SLC6A4 VNTR genotype determines transcription factor binding and epigenetic variation of this gene in response to cocaine in vitro. *Addict Biol,* 17**,** 156-70.

Vavouri, T., K. Walter, W. R. Gilks, B. Lehner & G. Elgar. 2007. Parallel evolution of conserved non-coding elements that target a common set of developmental regulatory genes from worms to humans. In *Genome Biol*, R15.

Vilarino-Guell, C., A. Rajput, A. J. Milnerwood, B. Shah, C. Szu-Tu, J. Trinh, I. Yu, M. Encarnacion, L. N. Munsie, L. Tapia, E. K. Gustavsson, P. Chou, I. Tatarnikov, D. M. Evans, F. T. Pishotta, M. Volta, D. Beccano-Kelly, C. Thompson, M. K. Lin, H. E. Sherman, H. J. Han, B. L. Guenther, W. W. Wasserman, V. Bernard, C. J. Ross, S. Appel-Cresswell, A. J. Stoessl, C. A. Robinson, D. W. Dickson, O. A. Ross, Z. K. Wszolek, J. O. Aasly, R. M. Wu, F. Hentati, R. A. Gibson, P. S. McPherson, M. Girard, M. Rajput, A. H. Rajput & M. J. Farrer (2014) DNAJC13 mutations in Parkinson disease. *Hum Mol Genet,* 23**,** 1794-801.

Visel, A., S. Prabhakar, J. A. Akiyama, M. Shoukry, K. D. Lewis, A. Holt, I. Plajzer-Frick, V. Afzal, E. M. Rubin & L. A. Pennacchio (2008) Ultraconservation identifies a small subset of extremely constrained developmental enhancers. *Nat Genet,* 40**,** 158-60.

Viturawong, T., F. Meissner, F. Butter & M. Mann (2013) A DNA-centric protein interaction map of ultraconserved elements reveals contribution of transcription factor binding hubs to conservation. *Cell Rep,* 5**,** 531-45.

Voltas, N., E. Aparicio, V. Arija & J. Canals (2015) Association study of monoamine oxidase-A gene promoter polymorphism (MAOA-uVNTR) with self-reported anxiety and other psychopathological symptoms in a community sample of early adolescents. *J Anxiety Disord,* 31**,** 65-72.

von Schimmelmann, M., P. A. Feinberg, J. M. Sullivan, S. M. Ku, A. Badimon, M. K. Duff, Z. Wang, A. Lachmann, S. Dewell, A. Ma'ayan, M. H. Han, A. Tarakhovsky & A. Schaefer (2016) Polycomb repressive complex 2 (PRC2) silences genes responsible for neurodegeneration. *Nat Neurosci,* 19**,** 1321-30.

Walsh, J., O. Tighe, D. Lai, R. Harvey, M. Karayiorgou, J. A. Gogos, J. L. Waddington & C. M. O'Tuathaigh (2010) Disruption of thermal nociceptive behaviour in mice

mutant for the schizophrenia-associated genes NRG1, COMT and DISC1. *Brain Res,* 1348**,** 114-9.
Wan, R. P., L. T. Zhou, H. X. Yang, Y. T. Zhou, S. H. Ye, Q. H. Zhao, M. M. Gao, W. P. Liao, Y. H. Yi & Y. S. Long (2017) Involvement of FMRP in Primary MicroRNA Processing via Enhancing Drosha Translation. *Mol Neurobiol,* 54**,** 2585-2594.
Wang, H., J. Xing, D. Grover, D. J. Hedges, K. Han, J. A. Walker & M. A. Batzer (2005) SVA elements: a hominid-specific retroposon family. *J Mol Biol,* 354**,** 994-1007.
Wang, J., L. Song, D. Grover, S. Azrak, M. A. Batzer & P. Liang (2006) dbRIP: a highly integrated database of retrotransposon insertion polymorphisms in humans. *Hum Mutat,* 27**,** 323-9.
Wang, Q., W. H. Fang, J. Krupinski, S. Kumar, M. Slevin & P. Kumar (2008) Pax genes in embryogenesis and oncogenesis. *J Cell Mol Med,* 12**,** 2281-94.
Wang, X. D., T. M. Ou, Y. J. Lu, Z. Li, Z. Xu, C. Xi, J. H. Tan, S. L. Huang, L. K. An, D. Li, L. Q. Gu & Z. S. Huang (2010) Turning off transcription of the bcl-2 gene by stabilizing the bcl-2 promoter quadruplex with quindoline derivatives. *J Med Chem,* 53**,** 4390-8.
Wang, Y., X. Zhao, W. Ju, M. Flory, J. Zhong, S. Jiang, P. Wang, X. Dong, X. Tao, Q. Chen, C. Shen, M. Zhong, Y. Yu, W. T. Brown & N. Zhong (2015) Genome-wide differential expression of synaptic long noncoding RNAs in autism spectrum disorder. *Transl Psychiatry,* 5**,** e660.
Warburton, A., G. Breen, V. J. Bubb & J. P. Quinn (2015a) A GWAS SNP for Schizophrenia Is Linked to the Internal MIR137 Promoter and Supports Differential Allele-Specific Expression. *Schizophr Bull*.
Warburton, A., G. Breen, D. Rujescu, V. J. Bubb & J. P. Quinn (2015b) Characterization of a REST-Regulated Internal Promoter in the Schizophrenia Genome-Wide Associated Gene MIR137. *Schizophr Bull,* 41**,** 698-707.
Warburton, A., A. L. Savage, P. Myers, D. Peeney, V. J. Bubb & J. P. Quinn (2015c) Molecular signatures of mood stabilisers highlight the role of the transcription factor REST/NRSF. *J Affect Disord,* 172**,** 63-73.
Ward, L. D. & M. Kellis. 2012. HaploReg: a resource for exploring chromatin states, conservation, and regulatory motif alterations within sets of genetically linked variants. In *Nucleic Acids Res*, D930-4.
Ward, L. D. & M. Kellis. (2016) HaploReg v4: systematic mining of putative causal variants, cell types, regulators and target genes for human complex traits and disease. *Nucleic Acids Res,* 44**,** D877-81.
Waye, J. S., G. M. Greig & H. F. Willard (1987) Detection of novel centromeric polymorphisms associated with alpha satellite DNA from human chromosome 11. *Hum Genet,* 77**,** 151-6.
Wei, H., Y. Yuan, S. Liu, C. Wang, F. Yang, Z. Lu, H. Deng, J. Zhao, Y. Shen, C. Zhang, X. Yu & Q. Xu (2015) Detection of circulating miRNA levels in schizophrenia. *Am J Psychiatry,* 172**,** 1141-7.
Wei, W., N. Gilbert, S. L. Ooi, J. F. Lawler, E. M. Ostertag, H. H. Kazazian, J. D. Boeke & J. V. Moran (2001) Human L1 retrotransposition: cis preference versus trans complementation. *Mol Cell Biol,* 21**,** 1429-39.
Weinstock, M. (2017) Prenatal stressors in rodents: Effects on behavior. *Neurobiol Stress,* 6**,** 3-13.
Wellik, D. M. (2007) Hox patterning of the vertebrate axial skeleton. *Dev Dyn,* 236**,** 2454-63.
Weon, J. L. & P. R. Potts (2015) The MAGE protein family and cancer. *Curr Opin Cell Biol,* 37**,** 1-8.

Wicker, T., F. Sabot, A. Hua-Van, J. L. Bennetzen, P. Capy, B. Chalhoub, A. Flavell, P. Leroy, M. Morgante, O. Panaud, E. Paux, P. SanMiguel & A. H. Schulman (2007) A unified classification system for eukaryotic transposable elements. *Nat Rev Genet,* 8**,** 973-82.

Wickham, S., K. Sitko & R. P. Bentall (2015) Insecure attachment is associated with paranoia but not hallucinations in psychotic patients: the mediating role of negative self-esteem. *Psychol Med,* 45**,** 1495-507.

Williamson, V. S., M. Mamdani, G. O. McMichael, A. H. Kim, D. Lee, S. Bacanu & V. I. Vladimirov (2015) Expression quantitative trait loci (eQTLs) in microRNA genes are enriched for schizophrenia and bipolar disorder association signals. *Psychol Med,* 45**,** 2557-69.

Willner, P. (2017) The chronic mild stress (CMS) model of depression: History, evaluation and usage. *Neurobiol Stress,* 6**,** 78-93.

Wisneiwski, K. E., E. Kida, O. F. Patxot & F. Connell (1992) Variability in the clinical and pathological findings in the neuronal ceroid lipofuscinoses: review of data and observations. *Am J Med Genet,* 42**,** 525-32.

Wolf, G., P. Yang, A. C. Fuchtbauer, E. M. Fuchtbauer, A. M. Silva, C. Park, W. Wu, A. L. Nielsen, F. S. Pedersen & T. S. Macfarlan (2015) The KRAB zinc finger protein ZFP809 is required to initiate epigenetic silencing of endogenous retroviruses. *Genes Dev,* 29**,** 538-54.

Wolff, E. M., H. M. Byun, H. F. Han, S. Sharma, P. W. Nichols, K. D. Siegmund, A. S. Yang, P. A. Jones & G. Liang (2010) Hypomethylation of a LINE-1 Promoter Activates an Alternate Transcript of the MET Oncogene in Bladders with Cancer. *PLoS Genet,* 6.

Won, E. & Y. K. Kim. 2017. An Oldie but Goodie: Lithium in the Treatment of Bipolar Disorder through Neuroprotective and Neurotrophic Mechanisms. In *Int J Mol Sci*.

Woolfe, A. & G. Elgar (2008) Organization of conserved elements near key developmental regulators in vertebrate genomes. *Adv Genet,* 61**,** 307-38.

Woolfe, A., M. Goodson, D. K. Goode, P. Snell, G. K. McEwen, T. Vavouri, S. F. Smith, P. North, H. Callaway, K. Kelly, K. Walter, I. Abnizova, W. Gilks, Y. J. Edwards, J. E. Cooke & G. Elgar (2005) Highly conserved non-coding sequences are associated with vertebrate development. *PLoS Biol,* 3**,** e7.

Wright, C., V. D. Calhoun, S. Ehrlich, L. Wang, J. A. Turner & N. I. Bizzozero (2015) Meta gene set enrichment analyses link miR-137-regulated pathways with schizophrenia risk. *Front Genet,* 6**,** 147.

Wu, M. & S. M. Ho (2004) PMP24, a gene identified by MSRF, undergoes DNA hypermethylation-associated gene silencing during cancer progression in an LNCaP model. *Oncogene,* 23**,** 250-9.

Wu, X., P. S. Chen, S. Dallas, B. Wilson, M. L. Block, C. C. Wang, H. Kinyamu, N. Lu, X. Gao, Y. Leng, D. M. Chuang, W. Zhang, R. B. Lu & J. S. Hong (2008) Histone deacetylase inhibitors up-regulate astrocyte GDNF and BDNF gene transcription and protect dopaminergic neurons. *Int J Neuropsychopharmacol,* 11**,** 1123-34.

Xie, X., T. S. Mikkelsen, A. Gnirke, K. Lindblad-Toh, M. Kellis & E. S. Lander (2007) Systematic discovery of regulatory motifs in conserved regions of the human genome, including thousands of CTCF insulator sites. *Proc Natl Acad Sci U S A,* 104**,** 7145-50.

Xu, B., P. K. Hsu, M. Karayiorgou & J. A. Gogos (2012) MicroRNA dysregulation in neuropsychiatric disorders and cognitive dysfunction. *Neurobiol Dis,* 46**,** 291-301.

Yamada, M., S. Yamazaki, K. Takahashi, K. Nara, H. Ozawa, S. Yamada, Y. Kiuchi, K. Oguchi, K. Kamijima, T. Higuchi & K. Momose (2001) Induction of cysteine string protein after chronic antidepressant treatment in rat frontal cortex. *Neurosci Lett,* 301**,** 183-6.

Yang, Y., W. Li, H. Zhang, G. Yang, X. Wang, M. Ding, T. Jiang & L. Lv (2015) Association Study of N-Methyl-D-Aspartate Receptor Subunit 2B (GRIN2B) Polymorphisms and Schizophrenia Symptoms in the Han Chinese Population. *PLoS One,* 10**,** e0125925.

Ying, X., Y. Sun & P. He (2017) MicroRNA-137 inhibits BMP7 to enhance the epithelial-mesenchymal transition of breast cancer cells. *Oncotarget,* 8**,** 18348-18358.

Yoda, K., S. Ando, S. Morishita, K. Houmura, K. Hashimoto, K. Takeyasu & T. Okazaki (2000) Human centromere protein A (CENP-A) can replace histone H3 in nucleosome reconstitution in vitro. *Proc Natl Acad Sci U S A,* 97**,** 7266-71.

Yuan, L., Z. Song, X. Deng, W. Zheng, Y. Guo, Z. Yang & H. Deng (2016) Systematic analysis of genetic variants in Han Chinese patients with sporadic Parkinson's disease. *Sci Rep,* 6**,** 33850.

Zhang, F. & J. R. Lupski. 2015. Non-coding genetic variants in human disease. In *Hum Mol Genet*, R102-10.

Zhang, M. M., C. Xiao, K. Yu & D. Y. Ruan (2003) Effects of sodium valproate on synaptic plasticity in the CA1 region of rat hippocampus. *Food Chem Toxicol,* 41**,** 1617-23.

Zhang, S., X. Zhou, H. Yu & Y. Yu (2010a) Expression of tumor-specific antigen MAGE, GAGE and BAGE in ovarian cancer tissues and cell lines. *BMC Cancer,* 10**,** 163.

Zhang, X., M. Wu, H. Xiao, M. T. Lee, L. Levin, Y. K. Leung & S. M. Ho (2010b) Methylation of a single intronic CpG mediates expression silencing of the PMP24 gene in prostate cancer. *Prostate,* 70**,** 765-76.

Zhang, Y., M. Fan, Q. Wang, G. He, Y. Fu, H. Li & S. Yu (2015) Polymorphisms in MicroRNA Genes And Genes Involving in NMDAR Signaling and Schizophrenia: A Case-Control Study in Chinese Han Population. *Sci Rep,* 5**,** 12984.

Zhang, Y., Q. Ming, X. Wang & S. Yao (2016) The interactive effect of the MAOA-VNTR genotype and childhood abuse on aggressive behaviors in Chinese male adolescents. *Psychiatr Genet,* 26**,** 117-23.

Zhang, Y., Q. S. Ming, J. Y. Yi, X. Wang, Q. L. Chai & S. Q. Yao (2017) Gene-Gene-Environment Interactions of Serotonin Transporter, Monoamine Oxidase A and Childhood Maltreatment Predict Aggressive Behavior in Chinese Adolescents. *Front Behav Neurosci,* 11**,** 17.

Zhao, X., A. P. Braun & J. E. Braun (2008) Biological roles of neural J proteins. *Cell Mol Life Sci,* 65**,** 2385-96.

Zhao, Y., Y. Li, G. Lou, L. Zhao, Z. Xu, Y. Zhang & F. He (2012) MiR-137 targets estrogen-related receptor alpha and impairs the proliferative and migratory capacity of breast cancer cells. *PLoS One,* 7**,** e39102.

Zhao, Z., Y. Li, H. Chen, J. Lu, P. M. Thompson, J. Chen, Z. Wang, J. Xu, C. Xu & X. Li (2014) PD_NGSAtlas: a reference database combining next-generation

sequencing epigenomic and transcriptomic data for psychiatric disorders. *BMC Med Genomics,* 7**,** 71.

Zhu, C., T. Utsunomiya, T. Ikemoto, S. Yamada, Y. Morine, S. Imura, Y. Arakawa, C. Takasu, D. Ishikawa, I. Imoto & M. Shimada (2014) Hypomethylation of long interspersed nuclear element-1 (LINE-1) is associated with poor prognosis via activation of c-MET in hepatocellular carcinoma. *Ann Surg Oncol,* 21 Suppl 4**,** S729-35.

Zhu, J., Y. Ling, Y. Xu, M. Z. Lu, Y. P. Liu & C. S. Zhang (2015) Elevated expression of MDR1 associated with Line-1 hypomethylation in esophageal squamous cell carcinoma. *Int J Clin Exp Pathol,* 8**,** 14392-400.

Zhu, M. & S. Zhao (2007) Candidate Gene Identification Approach: Progress and Challenges. *Int J Biol Sci,* 3**,** 420-7.

Ziats, M. N. & O. M. Rennert (2013) Aberrant expression of long noncoding RNAs in autistic brain. *J Mol Neurosci,* 49**,** 589-93.

Ziegler, C., J. Richter, M. Mahr, A. Gajewska, M. A. Schiele, A. Gehrmann, B. Schmidt, K. P. Lesch, T. Lang, S. Helbig-Lang, P. Pauli, T. Kircher, A. Reif, W. Rief, A. N. Vossbeck-Elsebusch, V. Arolt, H. U. Wittchen, A. O. Hamm, J. Deckert & K. Domschke (2016) MAOA gene hypomethylation in panic disorder-reversibility of an epigenetic risk pattern by psychotherapy. *Transl Psychiatry,* 6**,** e773.

Zohsel, K., A. F. Buchmann, D. Blomeyer, E. Hohm, M. H. Schmidt, G. Esser, D. Brandeis, T. Banaschewski & M. Laucht (2014) Mothers' prenatal stress and their children's antisocial outcomes--a moderating role for the dopamine D4 receptor (DRD4) gene. *J Child Psychol Psychiatry,* 55**,** 69-76.

Zovoilis, A., C. Cifuentes-Rojas, H. P. Chu, A. J. Hernandez & J. T. Lee (2016) Destabilization of B2 RNA by EZH2 Activates the Stress Response. *Cell,* 167**,** 1788-1802 e13.

Zubin, J. & B. Spring (1977) Vulnerability--a new view of schizophrenia. *J Abnorm Psychol,* 86**,** 103-26.