THE UNIVERSITY
of EDINBURGH

DISC1, Differential Expression, and Deconvolution; human and mouse RNA-Seq investigation of a t(1;11) translocation that predisposes to major mental illness

THE UNIVERSITY *of* EDINBURGH

Shane O'Sullivan

This dissertation is submitted for the degree of Doctor of Philosophy in Genomics and Experimental Medicine

University of Edinburgh

February 2020

“Fight on my men”, says Sir Andrew Barton

“I am hurt, but I am not slain;

I’ll lay me down and bleed a while

And then I’ll rise and fight again”

DECLARATION

This dissertation is the result of my own work and includes nothing which is the outcome of work done in collaboration except where specifically indicated in the text. It has not been previously submitted, in part or whole, to any university or institution for any degree, diploma, or other qualification.

Signed:

Date: 10/1/2020

Shane O'Sullivan

ACKNOWLEDGEMENTS

Well, here we are. Thanks, like data, are plural.

It's a very strange thing to be writing this thesis, and it has been an interesting experience. I owe a great deal to a great many people. First, I thank my supervisor, Dr Kirsty Millar, for her advice and guidance during the research and writing of this thesis, and the opportunity to carry out research in the Millar group. I also thank Dr Kathy Evans, my secondary supervisor. Professor Colin Semple and Dr Tilo Kunath complete my thesis committee, and I am grateful to them all for the helpful corrections and constructive criticism I received during my reviews. I also thank the members of the Millar lab; Helen (most particularly!), Jonathan, Marion and Laura, for the comradery, advice, and assistance throughout the PhD. Everyone at the wider IGMM was also very good to me, and I particularly thank Susan Anderson, Philippe Gautier, Graeme Grimes, and Ailith Ewing for key advice or assistance. A great many people also provided humour, cheer, and technical tips throughout my PhD. I thank Amaka, Grant, Sophie, Francis, and Jilly. To Rodanthi and Georgios who housed me (during the festival, no less!) I am extraordinarily grateful.

I also thank my funders, and especially the family who this thesis is based upon.

In Edinburgh I was most dependent upon and cheered by the other musketeers Katerina and Fiona. I thank Katerina and Panos for attempting many times and unsuccessfully to teach me to smoke, and Fiona for unsuccessfully teaching me to be a beer drinking bloke. I judge their efforts not by their results but by their intentions. Thank you all so much for the friendship and the good times. You are wonderful friends and people. Now would be the perfect time to quote Nietzsche or Freud, if I had read them.

I missed my Irish friends greatly and was always cheered by their visits. Thank you to the rest of the Fab Four, Louise, Rosemary, and Padraig, who were a delight either when visiting me or when I returned to Cork. My friends Eilis, Siobhan, the Dalester Cathal, and Niki, lifted my spirits every time we met again. To Cian and Stephen, the 26 Cambridge Avenue boys, I can only say thanks for shepherding me through another degree. Looking forward to our next Europe trip. The Svn group must also be thanked, as must my temporary roomie Ryan.

My most inner circle is my family. My father John, who taught industry and perseverance by example. My mother Evelyn, who gave me hope and humour. Aoife, Ciara and Emmet, my siblings, I can’t thank you enough for the laughs and the memories, especially at Christmas.

Last and most, thank you Ailbhe. I would never have done this without you and there are no words that properly express my thanks. You’re the best, but you already knew that.

CONTENTS

LAY SUMMARY AND ABSTRACT

Lay Summary

Several decades ago, a family with an unusually high rate of multiple psychiatric disorders was discovered in Scotland. The disorders included schizophrenia, bipolar disorder, and recurrent major depressive disorder, which are highly debilitating and involve emotional and behavioural problems. It was subsequently found that the family also carries a unique genetic mutation, called the t(1:11) translocation, that involves the exchange of genetic material between the chromosomes, which carry genes. Chromosomes 1 and 11 are affected. The translocation is inherited with very high risk of developing the disorders and is a major factor in the family's risk.

The exchange of genetic material, DNA, between chromosomes 1 and 11 means that large pieces of DNA have broken off at both locations, then exchanged. This now means that part of chromosome 1 is on chromosome 11, and vice versa. This has disrupted three genes, two which are present and overlap on chromosome 1, and one which is on chromosome 11. Of these three, one on chromosome 1 can instruct cells to produce a protein, named DISC1. There is a high level of DISC1 protein in cells of the brain (neurons) and it increases in presence during the development of neuron cell models. The protein is also known to have a role in many cellular processes involving the strengthening of connections between neurons in the brain, which is known to be important in learning and memory. It also has a role in how the neurons organise during brain development, and in how they produce energy. Because of the exchange of genetic material between chromosomes 1 and 11, the second half of the *DISC1* gene is missing and in its place is DNA from chromosome 11. This appears to result in lower levels of the DISC1 protein, in addition to changes in movement of molecules around the neurons..

We now have access to a unique type of neuron which is generated from skin samples, donated by members of the family. These neurons are therefore genetically matched to the family members that donated them. Members both with and without the translocation have made donations, so we can compare the two groups. We also have access to a unique mouse model. Mice also have a version of the *DISC1* gene. Here, this unique mouse model has been artificially altered so that its *Disc1* gene is

also missing the second half, which has been replaced by human DNA from chromosome 11 at the correct breakpoint. We have used mice where either one or both of its original *Disc1* genes have been altered in this manner.

This thesis describes the study of these human and mouse models, which have been investigated for altered levels of other cellular molecules which could be changed due to the mutation. It is shown that these changes are more likely than chance to be overlapping with those highlighted by other researchers, and are more likely than chance to be involved in various neuron activities relating to strengthening connections and moving molecules around the neuron. We also report that there appear to be higher levels of DRD2, a protein which antipsychotic drugs block the action of. This thesis also describes an investigation to look for different proportions of various cell types between the mutant and control samples; little evidence for this was found in the human cells. A part of the brain called the cortex shows unusual cell type proportion changes in the mouse if both the *Disc1* genes are altered. It does also appear that some cell types will be worse affected by the mutation than others, in activity if not in total number. Overall, this thesis highlights the overlaps between the effects of this unique mutation and other more common ones which are known to increase risk of schizophrenia. It also highlights some activities of the cells which appear to be abnormal and have been previously suspected of being important in psychiatric illness, and confirms some differences in key interesting molecules.

Abstract

The t(1;11) translocation is a mutation unique to a Scottish pedigree, members of which have been diagnosed with schizophrenia, bipolar disorder, recurrent major depressive disorder and other related disorders. The translocation is significantly linked to increased risk of these diagnoses. It disrupts three genes, only one of which, *DISC1*, encodes a protein. A number of experiments have explored the function of DISC1 as a molecular scaffold and developmental regulator. DISC1 and its interactors have roles in processes of relevance to psychiatric disease. These include neuronal precursor proliferation, migration and integration in the developing and adult brain, neurite outgrowth, mitochondrial activity, which is particularly important in neurons due to their high energy demands, and intracellular trafficking, especially critical in neurons due to their highly elongated morphology. Although various *DISC1* mutations have been investigated in the past, it is only with advances in technology that neural cells derived directly from translocation carriers, and therefore carrying the translocation plus their genetic background, have been generated and analysed. In addition a recently described mouse model mimics the effects of the translocation upon *DISC1* expression. It does so by removing endogenous *Disc1* exons corresponding to those distal to the breakpoint in translocation carriers, and fusing the remaining endogenous 5' *Disc1* genomic sequence to human chromosome 11 genomic sequence distal to the translocation breakpoint. The result is a chimeric gene with 5' mouse *Disc1* joined to a segment of human *DISC1FP1*, the non-coding fusion partner of *DISC1* located on chromosome 11. This leads to loss of wild-type Disc1 and prediction of chimeric transcripts encoding aberrant C-terminally truncated forms of Disc1.

This thesis builds on the work of previous researchers to characterise the RNA-sequenced transcriptome of 'cortical' neurons derived from induced pluripotent stem cells from various members of the pedigree. Both heterozygous and homozygous mutant mice have also been utilised to generate RNA-sequencing data from the hippocampus and cortex. The thesis not only describes the differential expression of genes and exons, but also carries out a series of analyses to examine whether

proportions of certain cell types are altered, as well as whether differentially expressed genes are highly associated with specific cell types.

RNA-Seq data have been analysed for differential expression at the gene and individual exon level using DESeq2 and DEXSeq, respectively. This has revealed over 1,200 differentially expressed genes in human neurons carrying the translocation, which predict changes to functions relating to intracellular transport and synaptic activity. In addition, a number of genes have been verified by RT-qPCR as being differentially expressed in these neurons. These include genes of known relevance to schizophrenia such as *DRD2*, which encodes the D2 dopamine receptor, *NTRK2*, which encodes the BDNF receptor NTRK2, and *BBS1* which encodes the DISC1 interactor and centrosomal protein BBS1. The human neurons also show significant overlap with previously published dysregulated genes in human neurons carrying other *DISC1* mutations, as well as with genes associated with schizophrenia by large-scale genome wide association and copy number variation studies. Human neuron RNA-Seq data have also been examined for evidence of local effects of the translocation upon gene expression, and no obvious strong effect was found. The pattern of gene dysregulation in heterozygous mutant mouse cortex overlaps with that of the mutant human neurons. Gene expression changes in the mutant mouse cortex have also been verified by RT-qPCR in the genes *Arc* and *Avp*, and the list of implicated genes also shows overlap with genes associated with schizophrenia by large-scale genome wide association and copy number variation studies.

An RNA-Seq deconvolution analysis was carried out to look for evidence of altered proportions of cell types at both the broad and more specific cell type level. This compared the observed expression of hundreds of genes in in the RNA-Seq samples against their expression in publically available RNA-Seq data of specific cell types. There does not appear to be any strong and consistent effect of the t(1;11) or mouse mutation on cell proportions. However, the data indicate greater than expected dysregulation of genes that are highly enriched in specific cell types. This includes certain subtypes of astrocyte. Mutant mouse cortex also shows dysregulation of genes associated with several subtypes of interneuron and pyramidal neuron, including parvalbumin positive interneurons. This indicates that, while the

proportions of cell types appears to be unaffected by the translocation or mouse mutation, specialised cellular functions may be perturbed.

To conclude, this thesis highlights a number of processes which appear to be disturbed by the translocation and mouse mutation. In all models, RNA-Seq evidence suggests signalling pathways of known relevance to psychiatric disease have been affected without significant alteration of cell proportions. This concurs with histological analyses of the mouse model by previous researchers. This thesis also describes the overlap between genes implicated in the study of this unique mutation as well as those implicated by studies seeking common or rare mutations predisposing to schizophrenia, supporting the hypothesis that different genomic risk variants and mutations converge upon certain molecular pathways that are especially important in this illness. The implication that the t(1;11) may alter the activities of certain cell types is also notable and future work can elucidate the cell-specific effects of the translocation.

LIST OF ABBREVIATIONS

Genes and proteins

ACTB	Beta-Actin
APOE	Apolipoprotein E
APP	Amyloid Precursor Protein
ARC	Activity Regulated Cytoskeleton Associated Protein
AVP	Arginine Vasopressin
BBS1/2/5	Bardet-Biedl Syndrome 1/2/5
CACNA1C	Calcium channel, voltage-dependent, L type, alpha 1C
CHRNA	Neuronal acetylcholine receptor subunit alpha
CHRNB	Neuronal acetylcholine receptor subunit beta
CPT2	Carnitine palmitoyltransferase II precursor
DISC1	Disrupted in schizophrenia 1
DISC1FP1	Disrupted in schizophrenia 1 fusion partner 1
DISC2	Disrupted in schizophrenia 2
DLD	Dihydrolipoamide dehydrogenase
DLG2/4	Disks Large homolog 2/4
DPYSL2/3	Dihydropyrimidinase-Like 2/3
DRD1/2	Dopamine Receptor D1/D2
DVL1	Dishevelled 1
ECI2	Enoyl-CoA Delta Isomerase 2
ERBB4	Receptor tyrosine-protein kinase erbB-4
FMRP	Fragile X mental retardation protein
FOXP2	Forkhead box protein P2
GAPDH	Glyceraldehyde 3-phosphate dehydrogenase
GCSH	Glycine cleavage system H protein, mitochondrial
GLDC	Glycine decarboxylase
GLRA	Glycine receptor, alpha 1
GPC1/5	Glypican 1/5
GRIA	Glutamate Receptor Ionotropic AMPA
GRIK	Glutamate Receptor Ionotropic Kainate
HAP1	Huntington's Associated Protein 1
HIF1A	Hypoxia Inducible Factor 1 subunit alpha
KANSL1	KAT8 Regulatory NSL Complex Subunit 1
KIF	Kinesin family member
LIS1	Lissencephaly 1
LYNX1	Ly6/neurotoxin protein 1
LYPD	LY6/Plaur domain-containing protein
MAG	Myelin Associated Glycoprotein
MBP	Myelin Basic Protein

METRN	Meteorin
MT2/3	Metallothionein 2/3
MYO	Myosin
NDE1	Nuclear distribution protein nudE homolog 1
NDEL1	Nuclear distribution protein nudE-like homolog 1
NDST3	N-deacetylase/N-sulfotransferase 3
NMDAR	N-methyl-D-aspartate receptor
NRCAM	Neuronal cell adhesion molecule
NRG	Neuregulin
NRP	Neuropilin
NRX	Neurexin
NTRK	Neurotrophic Receptor Tyrosine Kinase
OXT	Oxytocin
PDE	Phosphodiesterase
PDYN	Prodynorphin
PSD95	Post synaptic density 95
QKI	Quaking
SHTN1	Shootin1
SLC	Solute Carrier family
SYT	Synaptotagmin
SOX	Sry-boxes
TRAK1	Trafficking Kinesin Protein 1
ZNF	Zinc finger nuclear

Biochemical molecules

AMPA	α-amino-3-hydroxy-5-methyl-4-isoxazolepropionic acid
ATP	Adenosine triphosphate
cAMP	Cyclic adenosine monophosphate
cDNA	Complementary DNA
CO_2	Carbon Dioxide
DNA	Deoxyribonucleic acid
EDTA	Ethylenediaminetetraacetic acid
FGF	Fibroblast Growth Factor
GABA	Gamma-Aminobutyric acid
mRNA	Messenger ribonucleic acid
NAA	N-acetylaspartate
NMDA	N-methyl-D-aspartate
O_2	Oxygen
RNA	Ribonucleic acid
RNAi	RNA interference
RNA-Seq	RNA Sequencing
ROS	Reactive Oxygen Species
rRNA	ribosomal Ribonucleic Acid
shRNA	short hairpin RNA
siRNA	small interfering RNA
SNP	Single Nucleotide Polymorphism

Physical measurements and currencies

$	United States Dollars
%	Percentage
£	Pounds sterling
€	Euro
°C	Degrees Celsius
Bp	Base pair
DALYs	Disability adjusted life years
g	Gram
Kb	Kilobase, 1000 base pair
L	Litre
M	Molar (mols/litre)
Mb	Megabase, 1,000,000 Bp
mm	Millimetre
mol	Mols
µg	Microgram
µm	Micrometre

Biological terms and other abbreviations

ANOVA	Analysis of variance
BP	Bipolar Disorder
CDCV	Common disease common variant
CDRV	Common disease rare variant
CNV	Copy number variant
CP(1/60/69)	Chimeric protein 1/60/69
CPM	Counts per million
Der1	Derived chromosome 1, referring to the construct which mimics the t(1;11) in mice
DMEM	Dulbecco's Modified Eagle Medium
DSM-4	Diagnostic and statistical manual of mental disorders 4
DSM-5	Diagnostic and statistical manual of mental disorders 5
E18	Embryonic day 18
EWCE	Expression Weighted Cell type enrichment
FCS	Foetal calf serum
FPKM	Fragments Per Kilobase of transcript per million mapped reads
GO	Gene ontology
GWAS	Genome wide associate study
HET	Heterozygote
HOM	Homozygote
IGMM	Institute of Genetics and Molecular Medicine
iPSC	Induced pluripotent stem cells
ITK	Insight
LOD	Logarithm of the odds
NCBI	National Center for Biotechnology Information
NPC	Neural precursor cell
Padj	Adjusted P value
PBS	Phosphate buffered saline
PCA	Principal component analysis
PCR	Polymerase chain reaction
PGC	Psychiatric Genetics Consortium
Poly-A	Polyadenylation
PPI	Pre-pulse inhibition
qRT-PCR	Quantitative reverse transcription polymerase chain reaction
rMDD	Recurrent major depressive disorder
RPKM	Reads Per Kilobase of transcript per million mapped reads
SZ	Schizophrenia
t(1:11)	Translocation 1(1;11)

UTR	Untranslated region
WHO	World Health Organisation
WT	Wild Type

1 Introduction and Literary Review

1.1 Psychiatric illness

Psychiatric illnesses are among the major causes of human misery. The World Health Organisation estimates that over 800,000 people die by suicide every year[1]. However, a more nuanced measure of the burden of psychiatric illness is given by Disability Adjusted Life Years (DALYs). DALYs are a useful, if limited, means of quantifying harm in terms of years of life lost (YLLs), and years lived with disability (YLDs), which take into account time lived with the condition and the negative effects on quality of life during that period. Psychiatric illnesses are highly significant DALY contributors, with the three relatively common conditions of recurrent major depressive disorder (rMDD) (2.5% world total DALYs), schizophrenia (SZ) (0.6%), and bipolar disorder (BP) (0.5%) each contributing substantial proportions to the world total of DALYs [2]. As a group, they are also the largest single contributor to the world total of YLDs[3]. The true contribution is higher when self-harm and increased mortality due to psychiatric illnesses is included[4]. As expected given their prevalence and severity, these illnesses have immense social and economic implications[5], with the three conditions above estimated to have cost England a combined 16.7 billion pounds in 2007, with 5.5 billion of this in service costs and the rest in lost earnings[6]. Despite their severity and chronic nature, psychiatric illnesses receive a disproportionately low level of funding, with the global median estimated at just under 3% of expenditure compared to the 10% of DALYs they contribute (including self-harm)[4]. Phenotypically, psychiatric illnesses are characterised by harmful behaviours or abnormal psychological functions, as well as altered cognition.

Schizophrenia is a chronic condition characterised by a multitude of behavioural and psychological symptoms. These can be broadly characterised into positive, negative and cognitive[7]. Positive symptoms relate to acquired phenotypes such as hallucinations and delusions. Negative symptoms are those which relate to a loss of normal function, such as apathy, anhedonia and a “flat/blunt affect” (unemotional responsiveness and flattened speech). Deficits in cognitive functioning have been more recently identified as being present in schizophrenia. The DSM-V has been criticised in some quarters, but the kappa values (measurement of the likelihood of

agreement on a diagnosis) for DSM-V schizophrenia are relatively high, indicating a reliable diagnosis. The editors of the American Journal of Psychiatry note that the 0.46 kappa value for schizophrenia equates to two clinicians agreeing on a diagnosis 85% of the time, if 10% of their patients in the clinic have the illness[8]. DSM-IV schizophrenia, which is highly similar to DSM-V schizophrenia, is a diagnosis which is reliable with 80-90% of individuals diagnosed retaining it for 1-10 years after diagnosis[9]. Schizophrenia is therefore a diagnosis which is long lasting, enjoys broad but not absolute agreement among clinicians, and is partially treatable with recognised symptoms. Active psychosis can be controlled via use of antipsychotics[7], but functional recovery allowing resumption of employment, independent living, etc., is less achievable, with estimates ranging from 30-40% of individuals achieving this a few years after their first episode of schizophrenia[10]. Bipolar disorder is another severe psychiatric illness characterised by behavioural abnormalities. It has previously been distinguished from schizophrenia in that patients do not present with psychotic symptoms; this is part of the so called-Kraepelian dichotomy separating the affective disorders involving episodes of altered mood and affect from the psychotic disorders which tend to present earlier and last longer[7,11]. This dichotomy has been challenged by some authors[12]. Bipolar disorder is generally diagnosed by the presence of manic episodes characterised by hyper-excitability and altered mood, along with the presence of depressive episodes involving anhedonia, apathy, and other symptoms. There are different subtypes of the disorder; including cyclothymia which appears to involve symptoms of reduced intensity; including hypomania rather than full mania and depressive symptoms which do not fit the criteria for a depressive episode[11]. Bipolar disorder is relatively common; adding the two major subtypes (type I and type II) together gives a lifetime prevalence of 1%[13]. The kappa values of type I and type II, 0.56 and 0.4, are similar to that of schizophrenia[8]. The exact diagnosis of bipolar disorder is difficult, as major depressive disorder has phenotypic overlap with bipolar disorder. This is particularly seen in those subtypes which are not characterised by periods of mania. Many individuals diagnosed with unipolar depression may actually have a form of bipolar disorder and 20% of these individuals experience a manic or hypomanic episode within five years[11]. It appears an elevated number of these may be found amongst treatment-resistant depression

cases[13]. Major depressive disorder itself is characterised by a number of biological changes (loss of appetite, loss of desire, changes in sleep patterns), sadness, suicidal thoughts, slowing of speech and action, which persist for weeks and cause significant disability. It is surprisingly common, with a lifetime prevalence of 17% according to the US National Comorbidity study, although this should be tempered with the understanding that the diagnosis is difficult to make[14]. The DSM-V kappa value is 0.28, described as of "questionable agreement" among clinicians[8]. The disorder tends to be lifelong in duration and adolescent or childhood presentation is not rare[14].

These three conditions cannot be regarded as entirely separate entities. It is clear that there is great phenotypic overlap between the manic/psychotic elements of bipolar disorder and schizophrenia, while depression is very hard to distinguish from bipolar disorder which has not yet been characterised by a manic phase. It is also difficult to describe any symptom which is highly specific and sensitive in the diagnosis of these psychiatric diseases. Van Os and Kapur 2009 state (italics mine) "Within the *cluster of diagnostic categories*, the term schizophrenia is applied to a syndrome characterised by long duration, bizarre delusions, negative symptoms, and few affective symptoms" which they contrast with bipolar disorder, which has less negative and more affective symptoms. They also note that many symptoms, even of schizophrenia, are present in the healthy population at reasonable prevalence. Experience of auditory hallucinations and paranoia are common, at 5-8%[7]. Clearly fine lines cannot be drawn around psychiatric conditions. Some authors have proposed that psychiatric illnesses should be described more as a spectrum[12]. In addition to their phenotypic similarities, the conditions have pharmacological overlap. Antipsychotics such as clozapine can be used to treat the positive symptoms of schizophrenia as well as the manic phase of bipolar disorder, particularly if the patient exhibits psychosis. Selective serotonin reuptake inhibitors are key in treating the depressive symptoms of bipolar disorder as they are unlikely to cause mania, but are also used in the treatment of major depressive disorder[13]. As I describe below, the genetics of these conditions are characterised by a similar overlap.

1.2 Genetics of psychiatric disease

The genetic architecture of schizophrenia and other psychiatric illnesses continues to be discussed and is mired in controversy. Psychiatric diseases are surprisingly common and heritable, yet were hypothesised to have an extreme effect upon reproductive fitness via increased mortality and lower fecundity ratios. Both of these observations have been proven to be true. Psychiatric diseases are highly injurious, reflecting their contribution towards global DALYs. According to one epidemiological review, the median mortality rate of individuals with rMDD is 1.7 times the norm, while it is 2.6 times the norm for individuals with schizophrenia or bipolar disorder[15]. The effect is particularly strong in schizophrenia with 13-15 years of life typically being lost according to a recent meta-analysis. The authors hypothesised that the effect was due to increased suicide rates, but also due to cardiovascular and diabetic complications which are known to be more prevalent in schizophrenia[16]. Regarding fecundity, Power *et al.* 2013 have used a cohort of $2.3x10^6$ individuals to investigate fecundity rates in individuals with autism or schizophrenia compared to the undiagnosed (and presumably largely healthy) population. Male individuals with autism or schizophrenia have a fecundity ratio ¼ of the norm, while female individuals have fecundity ratios about ½ of the norm. Bipolar disorder has a lesser effect, but also results in reduced fecundity, especially in males. Similarly, male but not female individuals with depression had reduced fecundity[17]. It is notable that schizophrenia has an earlier onset and poorer prognosis in males[18], and the incidence rate is higher[19]. Autism also shows gender differences, particularly in prevalence[20]. The body of literature discussing the evolution of psychiatric disorders is enormous[21] and discusses possible fitness effects (especially in relatives), "parental war" involving allelic imprinting, disorders as emergent consequences of the human mind or extremes of normal variation[22] , and mutation-selection balances [23,24]. The study described by Power *et al.* was interested in sibling effects, so as to explore the possibility that variants predisposing to psychiatric disease were maintained by balancing selection. They found slight decreases in male sibling fertility and increases in female sibling fertility in autism and a decrease in male sibling fertility in schizophrenia, which they proposed might reflect sex specific allele affects which help to maintain the risk alleles via balancing selection. However

this clear male-deficit female-benefit was not seen in bipolar disorder, only partially in schizophrenia, and in any case involved fecundity changes of >3%, far too small to mean positive selection is an explanation for why these disorders persist despite strong selection[17]. For now, it appears that no positive selection effect of these diseases has been discovered. One might ask why it is that these diseases therefore exist despite obvious negative effects.

Of relevance to this persistence of schizophrenia is the debate over the genetic architecture of the disease, and its evolutionary history. The consensus among researchers is that various psychiatric illnesses have moderate to strong heritability (typically >0.6, up to 0.8), with that of schizophrenia and autism greater than bipolar disorder, and bipolar disorder greater than rMDD[25]. This consensus rests upon classic twin studies[26], and is buttressed by large population studies confirming that the risks of schizophrenia and bipolar disorder run in families[27]. As of yet however, there is no known psychiatric illness that has a strictly Mendelian inheritance; therefore no allele can give diagnostic certainty. Psychiatric diseases therefore exhibit complex inheritance. The primary debate is whether this inheritance is made up of very many low risk but common (>5% in the population) variants, the basis for genome wide association studies (GWAS) due to their commonality[28,29], or fewer but more highly penetrant variants (<1% in the population, typically undetectable by GWAS)[30,31]. These hypotheses are referred to as the "common disease common variants" (CDCV) model and the "common disease rare variants" model (CDRV) respectively. The CDRV model predicts that rare variants increasing risk of disease are unlikely to be inherited due to selection against them, and are therefore likely to be caused by de novo mutation. Such mutations could run in families for some generations before selection inevitably takes its toll. Evidence now exists suggesting de novo mutation is higher in cases of psychiatric disease[32,33], and affects specific pathways known to be of interest in these conditions, namely synaptic processes and neurodevelopment[34–36]. In addition there are paternal-age effects on the genesis of psychiatric illness[37–39] and on *de novo* mutation[40]. The degree to which advanced paternal age causes psychiatric disease has been questioned, as the reverse can also be true for a behavioural condition and has been shown to only need a small effect to

explain the observation that older fathers are more likely to have offspring with these conditons[41]. However a *de novo* origin for variants predisposing to schizophrenia would also explain the equal prevalence across populations. We therefore have with the CDRV model an explanation for the prevalence of the disease; mutation-selection balance. The CDCV model is the basis of GWAS, which associate particular alleles with conditions. One of the largest GWAS performed so far (with approximately 37,000 cases and 113,000 controls) has associated 108 loci with SNPs increasing risk of schizophrenia[29]. The SNPs associated with schizophrenia are for the most part weakly so; selection would have a very mild effect on these variants as they are only infrequently associated with the disease state. The CDCV model therefore too explains the prevalence of psychiatric disease. Controversy continues over the utility of the candidate gene CDRV-driven approach versus that of the GWAS CDCV-driven approach, although it has been suggested that both hypotheses have merit to them and need not be mutually exclusive[42,43]. GWAS, for example, explain very little of the heritability of psychiatric disease, and many rare mutations appear to be unique or near unique. Together they might explain the genetics of psychiatric disease. There are analogous scenarios; breast cancer has at least 86 variants associated with moderately increased risk of the disease (typically <2 times the risk), yet several cancers are also familial conditions with rare mutations associated with vastly increased rates of the disease, such as those in *BRCA1*[44]. It is evident that both rare familial mutations and common variation are at play here, as is the impact of environmental risk factors. In addition, the two models can be reconciled with a typical hypothesis of how risk of schizophrenia and other psychiatric illnesses is understood[43]. This is called the threshold model and a representation is given in Figure 1. Under this model, risk of a psychiatric disease is normally distributed, and a few individuals are close to the threshold of disease. In reality the threshold does not need to be so binary; it may represent individuals who are at appreciable risk of developing illness, or those who require a secondary factor to induce it. A high risk red population in Figure 1 has a greater proportion of individuals who have crossed the threshold than the general population.

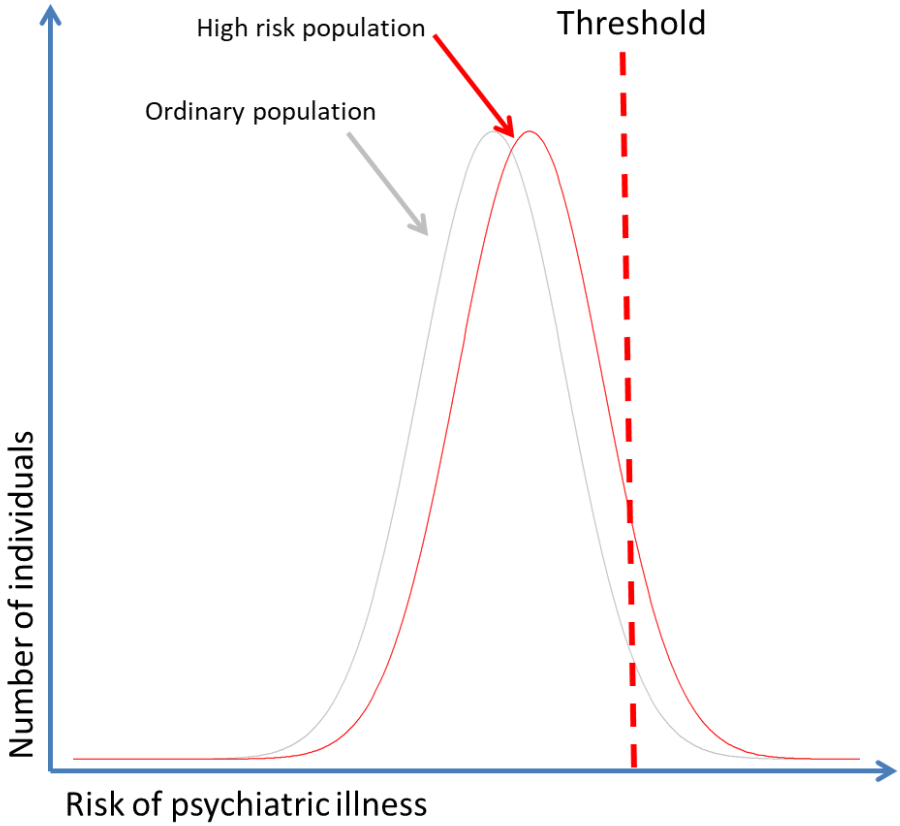


Figure 1. Representative image of threshold model for risk of psychiatric disease. A high risk population (red) and an ordinary population (grey) both have variation in underling risk, and both have individuals who have crossed the threshold for disease emergence. This proportion is far greater for the red population.

We can see how this can reconcile the two models of psychiatric genetics; the underlying population risk (possibly represented by an endophenotype, a genetically encoded phenotype which varies in the population and is disease associated) is underpinned by common variation, the load of which is likely to be normally distributed[43]. Meanwhile, the factor that converts a grey population or family to a red one can be a rare mutation. Alternatively, the rare mutation may function as a trigger factor for the individuals past the threshold. Environmental risk factors also undoubtedly play a role.

Many environmental risk factors for psychiatric disease act during development. Maternal stress during pregnancy increases risk of schizophrenia, bipolar disorder, and depression in the offspring[45,46], as does maternal infection in schizophrenia and bipolar disorder. The window for these risk factors appears to coincide with critical

periods of foetal brain development. Risks can also be postnatal; a wide variety of drugs appear to increase risk of bipolar disorder, for example, and hypoxia at birth increases the risk of schizophrenia[46,47]. Traumatic events and stress in childhood also increase the risk of psychiatric disease[45]. Exactly how environmental risk factors impact upon pathways of relevance to psychiatric disease is still being investigated. It has been noted that stress alters the HPA axis, a major neuroendocrinal system involved in the production of cortisol, a stress related hormone. Stress decreases hippocampal dendrites and hippocampal BDNF levels, which may be linked to depressive symptoms. Maternal stress has been linked to some cognitive dysfunction, hyperactivity, and diminished PPI in a mouse model, as well as altered GABAergic interneuron epigenetic markers[45]. Models for other risk factors show some schizophrenia related phenotypes as well; rats which underwent perinatal hypoxia showed diminished prepulse inhibition in adulthood, a schizophrenia-related phenotype which was treatable by the antipsychotic clozapine. Hypoxia may also affect myelination, which is a developmental process[45].

As mentioned previously, many psychiatric conditions display phenotypic and pharmacological overlap. The genetics of such closely related phenomena are characterised by similar overlap. Evidence of this is shown in large familial studies. Relatives of individuals with bipolar disorder have an enhanced risk of schizophrenia as well as bipolar disorder, even if raised in a different environment[27]. This strongly suggests that the same genetic lesions are responsible for risk predisposing towards multiple psychiatric disease. Copy Number Variants (CNVs) are one such source of genetic risk. An early paper by Guilmatre *et al.* searched for evidence of CNVs at several loci containing candidate genes for schizophrenia, autism, and mental retardation. They found CNVs likely to be causative for each disorder, and in many cases the CNVs appeared to be capable of predisposing to more than one disorder. They even detected comorbidity of schizophrenia and mental retardation in some individuals carrying certain rearrangements. Many of the CNVs were *de novo*, but a number had been inherited from an unaffected parent, typically from a mother to a son. We can see that this study shows risks run in families (some families found had multiple affected siblings), that risk is genetic, and that risk for multiple conditions

can be caused by the same mutation[48]. A large GWAS looking at variants predisposing to multiple disorders appears to have found SNPs which predispose to the five disorders of schizophrenia, bipolar disorder, major depressive disorder, autism, and ADHD. Of the four SNPs which predispose to all disorders, two are located within calcium channel genes, and one has been significantly associated in separate disorder specific GWAS for bipolar disorder, schizophrenia, and major depressive disorder[49]. The study described in this thesis is an opportunity to see if a rare mutation predisposing to schizophrenia displays any overlap with these common sources of risk. The question remains how exactly it is that the same mutation can associate with different disorders. Answers to the particular question of overlap for some pairs of disorders have been offered. For example, an explanation for the neuroanatomical, behavioural and epidemiological similarity between autism and schizophrenia has been proposed. This explanation suggests the timing and duration of neuroinflammation distinguishes schizophrenia from autism[50]. Blocking IL-6 mediated inflammation does appear to prevent neurological damage caused by prenatal LPS exposure. Similarly, an explanation for the differences between bipolar disorder and schizophrenia has been offered. Schizophrenia presents with more severe premorbid cognitive functioning, and this is apparent even in childhood. Brain abnormalities also appear to be greater in cases of schizophrenia, especially in the medial temporal lobe. Murray *et al.* have proposed that since prenatal and perinatal risk factors predispose to schizophrenia but not bipolar disorder, it is possible that this severe developmental insult places an individual already at high risk of bipolar disorder onto a trajectory towards schizophrenia. In this case, the environmental insult combines with the genetic liability to result in an exacerbated phenotype[51]. This “two-hit” hypothesis has surfaced elsewhere, in which a single CNV predisposed to developmental delay, while a second additional CNV within a minority of the cohort appeared to cause a more severe phenotype. This has led one reviewer to describe psychiatric disease inheritance as possibly being “omnigenic” in character[24]. Craddock and Owen have suggested that the overlap we see between many pairs of disorders should be viewed in another way; as an overarching spectrum of disorders, characterised by partial phenotypic and genetic overlap. In this model, psychiatric diseases are unified by some phenotypic overlap, but

distinguished by severity of developmental abnormality. Intellectual disability and autism are on one end of the spectrum, characterised by high developmental abnormality, high risk mutation (with larger CNVs), and severe pathology. On the other end are affective disorders such as bipolar and major depressive disorder. Schizophrenia is proposed as being somewhere in between, so mutation predisposing to either bipolar disorder or to autism can also predispose to it, but not typically to one another. Mutation and environmental risk can be either specific to a disorder or general. One of the strengths of this model is that it can incorporate our observations of phenotypic overlap, and give an explanation for why both rare and highly deleterious as well as common and less penetrant variation can both exist and cause psychiatric disease[12].

1.3 Pathways causing psychiatric disease

The biology of psychiatric disease is still mysterious, although great advances in understanding have been made. As mentioned previously, schizophrenia presents with anatomical differences including enlarged lateral ventricles and reduced cortical grey matter volume which are observable early in life. There are also premorbid problems in cognitive functioning. These observations, along with the lack of gliosis typically caused by neurodegeneration, have led to the conceptualisation of schizophrenia as a neurodevelopmental disorder, rather than a neurodegenerative one[52]. As stated earlier, this older conceptualisation fits in perfectly with the newer concept of all psychiatric diseases being distinguishable by differing degrees of neurodevelopmental pathology. The challenge now is to determine what processes are disturbed that might lead to altered development, or altered function in adulthood resulting from this altered development. One reasonable approach is genetic; to look for what genes carry the rare or common variants discussed earlier. Rare variant genes are typically found via investigation of pedigrees with high rates of disease (or parent-child trios searching for *de novo* mutation) while GWAS search for common variants. Rare variants often have accompanying mouse models and functional studies. A non-exhaustive discussion of genes related to psychiatric disease and the pathways they implicate is given below.

Given the importance of the synapse in learning, memory, and neuronal activity, we might expect that proteins involved in synaptic structure or activity might be implicated in psychiatric disease. It has been known for some time that neurotransmitters underlie some of the pathology of psychiatric disease. For example, high dopamine levels in the prefrontal cortex are believed to underlie the positive symptoms of schizophrenia, and blockade of the cognate receptors alleviates these symptoms[53]. Similarly, serotonin is believed to have some kind of link to signalling pathways involved in depression[54]. NMDAR antagonism by agents such as ketamine and PCP results in psychosis highly similar to schizophrenia[55]. Interestingly, NMDAR antagonism and agonism are being explored as anti-depressant therapy[56]. Although the links between NMDAR signalling and psychiatric diseases are not yet fully understood, we do know that NMDAR signalling is crucial for some types of long term potentiation (LTP), which underlies synaptic plasticity via remodelling of the synapse, insertion of AMPARs, and increases in synaptic scaffold proteins[57]. This underlies learning and memory. Many synaptic scaffold proteins have been linked to schizophrenia by candidate gene studies. Neurexins are a group of presynaptic proteins which form trans-synaptic bonds with neuroligins, a group of postsynaptic proteins. Both neurexins and neuroligins bind to MAGUK proteins such as PSD-95 and are capable of inducing postsynaptic and presynaptic specialities in adjacent neuronal cells if they are ectopically expressed in non-neuronal cells[58]. Mutations in several members of both families have been shown to segregate with several disorders ranging from Tourette's syndrome to schizophrenia and autism[58]. *De novo* CNV studies looking at parent-affected offspring trios have found that mutations in NMDAR signalling related genes are enriched in cases of schizophrenia, as are mutations in the genes encoding targets of FMRP. This RNA binding protein, mutated in Fragile X syndrome, is neuronally expressed and targets many genes involved in LTP and synaptic activity such as *ARC*[59].

One of the largest GWAS for schizophrenia so far found 108 single nucleotide polymorphism (SNP) loci to be implicated in the inheritance of the disease, with genes near these loci having roles in synaptic activity, glutamatergic transmission, and calcium signalling, although these have not been functionally validated[29]. A

later, larger GWAS found still more loci. These disproportionately included the targets of the FMRP protein, showing a convergence between CNV and SNP risk for schizophrenia. Genes involved in synaptic activity and calcium ion import were also overrepresented, further solidifying the evidence for these processes being involved in schizophrenia[60]. Calcium channel proteins have also appeared in GWAS for bipolar disorder alone, as well as in GWAS looking for SNPs predisposing to many disorders simultaneously[49]. Other neurotransmitter receptors or ion channels such as the NMDAR subunit *GRIN2A* (glutamate) and *SCN2A* (sodium) have been identified as significantly associated with bipolar disorder by the largest GWAS yet for this condition. *CACNA1C* in particular has been consistently associated with the condition[61]. The *CACNA* genes encode subunits of voltage gatedCa^{2+} channels; multimeric proteins prevalently expressed in the brain. There is evidence to suggest that calcium is depressed in manic bipolar patients and increased during the depression pole. Calcium channels are also upregulated in mouse models of addiction; agonists can help prevent withdrawal symptoms[62]. The signalling pathways are complex; but it is known that drug-mediated activation of dopamine receptor D1R activates signalling molecules such as CAMKII, leading to AMPAR insertion at the cell surface. As a calcium-activated molecule, this activation of CAMKII is via the L type voltage gated calcium channels. Antagonism of the channels reduces addiction behaviour in rats[63]. *CACNA1C* encodes one of the L type subunits; conditional null mice lacking hippocampal expression of it have LTP deficits. A mutation in the cognate human gene causes a condition characterised by autism and cognitive abnormalities[64]. Given its role in NMDAR-independent LTP, crucial for learning and memory, it is highly interesting that many studies converge on this gene.

1.4 DISC1

DISC1 is an example of a gene with rare mutations predisposing to psychiatric illness. The focus of much research, including this thesis, is on a t(1;11)(q42.1;q14.3) translocation disrupting this gene. The Scottish pedigree in which this translocation has been uniquely observed first came to the attention of researchers over four decades ago for the unusually high prevalence of psychiatric disease, ranging from

alcoholism and conduct disorder to diagnoses of schizophrenia. One major paper describing the clinical findings of the entire pedigree, stretching over four generations, is that of Blackwood *et al.*[65]. 67 individuals underwent karyotyping for the translocation, as well as psychiatric interviews (or their records were examined, were they available). Of 29 with the translocation, 21 were diagnosed with a psychiatric illness. A third of these diagnoses were of schizophrenia and approximately half were of major depressive disorder. Of 38 without the translocation, 5 had diagnoses; one case of alcoholism, one of conduct disorder, and three of minor depression. It is evident that both the prevalence and severity of psychiatric illness are greatly exacerbated by the translocation, although there may well be other genetic and environmental factors at play. A LOD score of 7.1 was obtained if the phenotype was modelled as being "schizophrenia, bipolar disorder, and major depressive disorder"; the LOD score was 3.6 if it was modelled as just "schizophrenia". The translocation itself involves the exchange of a large amount of genetic material between chromosomes 1 and 11; approximately 18Mb and 45Mb respectively. The translocation disrupts a total of three genes; two on chromosome 1, and one on chromosome 11. The chromosome 1 breakpoint was the first described in detail; it consists of two genes on opposite strands. *DISC1* is transcribed in the breakpoint proximal to distal direction and encodes a protein, DISC1. The opposite strand encodes the antisense *DISC2*, which has no hallmarks of a protein encoding gene (no long ORF, exceptionally long candidate 3'UTR) but appears to be expressed in some tissues[66] . The 5' end of *DISC1* remains on chromosome 1 in the t(1;11) condition, as does the 3' end of *DISC2*. DISC1 is a large protein, and the breakpoint is towards the end of the protein, after exon 8[66]. Its function has since been elucidated in more detail, in addition to its expression and distribution[67]. Multiple protein isoforms of DISC1 exist and it appears to be localised at the mitochondria, as disturbing microtubule formation results in aberrant mitochondria and DISC1 localisation[67]. However, *DISC1* products are also found at the centrosome, where they interact with other proteins involved in cell division and migration[68]. DISC1 immunoreactivity is also seen at synapses, particularly the post-synaptic density, and throughout neuropili. DISC1 immunoreactive neurons are seen in all layers of the human cortex[69]. Disturbing microtubule formation particularly

affects the DISC1 which is found along microtubules non-adjacent to the nucleus[67]. The protein is expressed in neural precursor cells as well as neurons, and is upregulated during neuronal differentiation[70]. It is expressed in a wide variety of human foetal tissues as well as the adult hippocampus. The mouse and monkey orthologues show a similar pattern of expression in a variety of brain areas and neuron types, suggesting a possible conservation of function[67,71]. Expression of *DISC1* proximal to the breakpoint is depleted in lymphoblastoid lines carrying the t(1;11), suggesting that haploinsufficiency might be responsible in part for the phenotype of increased disease risk[72]. Transcripts from the truncated gene would produce proteins lacking the C-terminal domain rich in coiled coil regions which aid protein assembly. However, studies suggest that the t(1:11) could also have gain of function affects.

Eykelenboom *et al.* found that the derived 1 chromosome of the t(1:11) translocation (consisting of the N terminal end of *DISC1* fused to C terminal *DISC1FP1*) can be translated, although given that *DISC1* regions proximal to the breakpoint are downregulated, this may not be at a high level[72]. The theoretical transcripts also contain features which tend to cause nonsense-mediated mRNA decay[72]. Hypothetically, the translated protein would consist of the first ~600 amino acids of DISC1 fused to either 60 or 69 amino acids from an open reading frame in *DISC1FP1*. These were labelled CP60 and CP69. They appear to have different properties to hypothetical truncation proteins containing just the first ~600 amino acids. A fragment of DISC1 fused to MBP displayed different properties if the extra 69 amino acids were included. Dynamic light scattering indicated that the presence of these residues caused the DISC1-MBP fusion protein to be larger (non-significantly $p=0.056$) and more resistant to heat-induced deformation. Ectopic expression of CP60/CP69 results in localisation to the mitochondria and loss of mitochondrial membrane potential. This does not co-present with cytosolic cytochrome c which would indicate increased apoptosis, however[72]. It must be noted that these proteins have not been detected in any of several cell models carrying the translocation, including neural precursor cells and neurons, and their existence is still hypothetical[70].

1.4.1 DISC1 function

The function of DISC1 protein involves the maturation and migration of neurons, synaptic activity, and centrosomal orientation, amongst many others. DISC1 shows no enzymatic capacity itself, but is known to interact with a large number of other proteins to exert its effects, which have been dubbed the DISC1-interactome. A Y2H screen for potential DISC1 interactors found a large number, with overrepresentation among these of GO terms including those relating to cytoskeletal organisation, transport, and cell division[73]. Many of these potential interactors are already candidate genes for schizophrenia[74], or are expressed in the developing brain, as is *DISC1*[75]. For example, GSK-3β is a target of lithium chloride, the mood stabiliser drug used to aid management of bipolar disorder and related disorders[76], and is important in synapse function. GSK-3β activity is mediated by DISC1 interaction, altering neural progenitor proliferation via the stabilisation of pro-growth signalling molecules such as β-catenin[77]. In the developing cortex, neural progenitors replicate for a time before producing postmitotic neurons and migrating through the cortex. As its interaction with GSK-3β shows, DISC1 appears to have a key role in maintaining developing cortical cells in their proliferating phase. Experiments show that *DISC1* knockdown increases the rate of cell cycle exit and subsequently increases the proportion of cells expressing *Cux2*, a cortical marker and neuronal transcription factor. This indicates that DISC1 acts as a check on early and improper cell differentiation. However, it also has a role in helping cells swap to migration via its interaction with the centrosome. The centrosome is a structure shown to be key in neurodevelopmental processes. Mutations in genes encoding centrosomal proteins often have severe consequences[78], and some are DISC1 interactors. *BBS4* is an example of a disease gene which encodes a centrosomal protein which is also a DISC1 interactor. Mutation in any *BBS* gene can cause BBS, a multisystem developmental disorder characterised by behavioural abnormalities, obesity, and retinal degradation among other phenotypes[79,80]. Its interaction with DISC1 is phosphorylation dependant. DISC1 interacts with and inhibits GSK-3β to enable neural precursor proliferation. However, DISC1 phosphorylation at Ser70 greatly reduces the GSK-3β interaction, blocking precursor proliferation. The phosphorylated DISC1 then recruits the BBS proteins to the centrosome to stimulate

neuronal migration. Mutations in *DISC1* alter cell proliferation and migration; those in *BBS1* only affect migration, consistent with this model[81]. Other DISC1 interactors linked to neural processes include LIS1 (causative of lissencephaly), NDEL1, and NDE1. These three proteins work together to affect neuronal migration via nucleokinesis. This process involves interactions between microtubules and the centrosome, where the protein products of *NDEL1* and *NDE1* reside and interact with dynein, gamma-tubulin, and other centrosomal proteins. They both can bind LIS1, which interacts with microtubules[82]. Mice with mutations in the *LIS1* homologue have an unusually patterned cortex and a disorganised hippocampus with more scattered cells. Cells carrying a mutation also migrate poorly[83]. The genes also have roles in other critical neurodevelopmental processes such as cell proliferation and neurite outgrowth, and are regulated by PDE4[84]. *PDE4B* is another gene linked to psychiatric disease via a translocation co-occurring with psychosis as well as significance in the most recent large GWAS looking for variants associated with schizophrenia[60]. It also linked to *DISC1* via direct protein-protein interaction[74]. *PDE4B* is crucial for cAMP regulation, which itself is vital in several neural processes, such as synaptic plasticity, memory formation and cognition. Rising cAMP levels can trigger dissociation of PDE4B and DISC1 via PKA-mediated phosphorylation, which results in higher PDE4B activity and presumably helps mediate cAMP signalling via cleavage of cAMP[74,85]. PKA also phosphorylates NDE1 on two sites (and may phosphorylate the similar NDEL1 region). This phosphorylation causes NDE1-LIS1 co-immunoprecipitation to decrease. This was also seen in a NDE1 phosphomimic mutant, which additionally displayed less neurite outgrowth. The suggested model is that DISC1's interaction with PDE4 allows cAMP signalling to block neurite outgrowth and LIS1/NDE1 interaction[84]. Especially relevant is that a selective inhibitor of PDE4, rolipram, acts as an antidepressant[74].

A new area of research related to DISC1 is its interactions with NMDARs. As described in 1.3, NMDARs are important in psychiatric disease. Their agonism is being explored as an anti-depressant therapy, and can induce psychosis[55,56]. They have also been linked to the aetiology of schizophrenia by CNV studies and

stimulation of NMDARs is crucial for the initiation of the changes leading to synaptic plasticity. GluN1, an obligate subunit of the NMDAR, is trafficked to the synapse in order to regulate glutamate sensitivity. Such trafficking is necessary for NMDAR-dependent synaptic plasticity. A direct link between NMDARs and DISC1 has now been revealed in a paper by Malavasi *et al.* using the same models as are utilised in this thesis[70]. For disclosure, I am a co-author on this paper.

GluN1, encoded by GRIN1, is an obligate subunit of all NMDARs. NMDAR subunits are seen in the ER, where they reside before assembly and trafficking. It has now been shown that DISC1 co-immunoprecipitates with GluN1 via amino acids which are encoded proximal to the t(1;11) breakpoint. Exogenous DISC1 co-localises with GluN1 in hippocampal neurons accordingly, in dendritic locations. It was suspected that DISC1 interactors involved in its trafficking role might also interact with NMDAR subunits. TRAK1, a DISC1 interactor and trafficking molecule which is targeted to the mitochondria, is also shown to co-precipitate with exogenous GluN2b in synaptosomes and light membranes when expressed exogenously. A portion of exogenous GluN1 co-localises with the exogenous GluN2b and TRAK1 in triple transfected cells. Since NMDAR are assembled in the ER, the implication is that the co-localising of GluN1 and GluN2B represents assembled NMDARs. It was also shown that DISC1 overexpression in hippocampal cells resulted in alterations of GluN1 trafficking. Increased fluorescence of distal fast moving GluN1 was seen, the speed of which corresponds with actively trafficked NMDAR-containing vesicles. The effect was also seen in mouse neurons carrying a mutation which models the translocation. Overexpression of a non-TRAK1 interacting DISC1 mutant resulted in reduction of this fast moving fluorescence. The mouse mutation leaves the GluN1 and TRAK1 interacting regions intact. These mice also had increased puncta density of GluN1 and GluN2B in the homozygous and heterozygous mutation, in addition to increased GluN2A puncta density and GluN1/PSD-95 co-localisation in the homozygous state. Total protein levels and PSD-95 puncta volume were unchanged, implying differences in NMDAR subunit trafficking and subsequent synaptic formation rather than expression. PSD-95

distribution was also altered. The result is likely to be aberrations in synaptic plasticity.

1.4.1.1 Mitochondria and DISC1

The role of mitochondria in psychiatric disease has been recognized. A case study has been reported in which mitochondrial DNA mutations co-occur with psychiatric disease as well as histories of psychiatric disease on the maternal side of the family[86]. Another has been reported in which mitochondria DNA mutations in the tRNALeu gene coincide with a maternal family history of psychiatric disease as well as cardiomyopathy in a proband[87]. More generally, there are many links between mitochondrial dysfunction and psychiatric disease. Mood stabiliser drugs protect against mitochondrial damage, while there appears to be aberrant expression of genes involved in ATP generation and storage in psychiatric disease, as well as deficits in oxidative phosphorylation[88]. Neurons exhibit an unusually high energy demand and therefore might be especially sensitive to mitochondrial dysfunction, while the brain must supply this demand exclusively through the oxidative phosphorylation of glucose, further increasing the importance of mitochondria to brain metabolism[88,89]. Given the localisation of DISC1 to the mitochondria, it was not unreasonable to expect it might have some kind of function there. It has now been shown by several groups that DISC1 has an impact on mitochondrial trafficking. Expression or siRNA knockdown of DISC1 respectively increase and decrease the number of motile mitochondria in neurons[90]. A particular disease-associated DISC1 polymorphism of a conserved residue also rendered the protein incapable of restoring the motility deficiency caused by knock down of wild type DISC1[90,91]. Some detail on how DISC1 influences trafficking has now been elucidated. A group of proteins on the mitochondrial outer membrane, the Miro GTPases, function in mitochondrial trafficking[92]. When these are bound by the kinesin and dynein adaptor TRAK1, mitochondria can be trafficked along the kinesin/microtubule based transport system[92]. TRAK1 and DISC1 co-immunoprecipitate, suggesting some kind of interaction occurs between the two[93]. TRAK1/DISC1 co-transfection results in an altered localisation of DISC1 compared to DISC1 transfection alone[93]. Mutation of a relatively well conserved[94] arginine-rich N-terminal sequence in DISC1 alters its

mitochondrial location as well as its interaction with TRAK1[95]. DISC1 appears to interact with other proteins involved in the motor activity of mitochondria, including MIRO1[93] and TRAK2[96]. It is therefore a confirmed interactor of mitochondrial membrane proteins and motor protein adaptors. The DISC1 interactors GSK-3β and NDE1 also associate with TRAK1, enhancing anterograde or retrograde mitochondrial movement respectively, although GSK-3β has been shown to have other effects in other studies[95,96]. Given that LIS1/NDE1/NDEL1 interact together with dynein and centrosomal proteins to effect microtubule-based nucleokinesis, it is unsurprising that these interactions are also seen in microtubule-based mitochondrial transport. LIS1 can bind to dynein to promote trafficking, and it had been suspected that this would include mitochondrial trafficking[89]. LIS1 or NDEL1 knockdown inhibits axonal mitochondrial transport in both directions or just retrograde, respectively[96]. It has also been reported that overexpression of LIS1 stimulates retrograde organelle transport, and as expected this requires dynein binding[97]. Given that TRAK1 co-immunoprecipitates with DISC1, NDE1, and GSK-3β, it has been suggested that DISC1 recruits GSK-3β into this complex, which likely contains NDEL1 and LIS1 as well[96]. DISC1 overexpression or knockdown increases or decreases mitochondrial motility, respectively, but a non-synonymous variant which disturbs DISC1-GSK-3β interaction prevents the stimulation of motility[90]. There are clearly extensive interactions between mitochondria, motor proteins, and DISC1-interactors which are also seen at the centrosome. The exact nature of these interactions is yet to be fully elucidated. GSK-3β certainly plays a role, as does the kinase Cdk5 in rat, which phosphorylates Ndel1 in a manner necessary for organelle transport[97]. Given the importance of mitochondrial trafficking it is likely that the proteins at the centre of it, including DISC1, have the ability to incorporate signalling from multiple pathways. This might include PDE4 enzyme/PKA signalling, which has already been shown to be important for AMPAR subunit trafficking. This was not via PDE4B[98]. In any case it is clear that a large number of proteins important for the linking of mitochondria to motor proteins, or in stimulating their movement, appear to interact with DISC1, which appears to have a role in scaffolding these proteins and facilitating interactions. As discussed earlier, it also may be the case that gain-of-function chimeric DISC1 proteins, with 60 or 69 extra residues, assemble at

the mitochondria and cause loss of membrane potential[72]. Either by toxic gain of function caused by chimeric DISC1-DISC1FP1 proteins, or by haploinsufficiency of *DISC1* necessary for assembling transport complexes, the t(1;11) is evidently capable of exerting effects upon mitochondrial activity with potential consequences for neuronal health.

1.4.2 *DISC1 mutations*

DISC1 mutations have been found elsewhere. An American family has been discovered with high rates of major mental illness, as well as a 4bp frameshift mutation in *DISC1*[99]. This family was originally discovered through sequencing of *DISC1* in schizophrenia probands, leading to the discovery of the pedigree. However, inspection of the pedigree revealed that although two siblings with schizophrenia had the mutation, as did one with schizoaffective disorder, three of their siblings without the mutation had major depressive disorder and schizotypal personality disorder. The family's unaffected father carried the mutation, while the mother, known to not have the mutation, had a family member with schizophrenia (who presumably did not have the 4bp frameshift of the proband's pedigree). Although it is highly interesting that another mutation in *DISC1* has been discovered, the evidence is not yet compelling for an association with schizophrenia. There are several key reasons as to why this is so, several of which were acknowledged directly or indirectly by the original paper. Firstly, the rate of diagnosis of any disorder is equal in carriers and non-carriers within the family. Secondly, individuals without the mutation display severe psychiatric disorders. Thirdly, the pedigree is too small for generation of LOD scores. Fourthly, the unaffected mother, whose family presumably do not have the 4bp mutation, has a family history of schizophrenia which might partially explain the high rate of diagnosis we see. Fifthly, another study showed that this mutation was found in none of several hundred schizophrenia cases, but was present in two anonymous blood donors in the control group. These individuals did not undergo psychiatric evaluation but were unlikely to have a psychiatric diagnosis as they were not taking medication[100]. Although the case is highly interesting and is worthy of follow up, I argue it should be regarded as a familial case of idiopathic

schizophrenia, until more evidence, including full familial genotyping and clinical examination, emerges.

Other mutations have been found not by pedigree investigations but by larger haplotype association studies. It should be noted that *DISC1* is not a hit in either of the largest schizophrenia GWAS, a finding which might indicate common variation in the gene is less important to schizophrenia risk[29,60]. Some research groups have attempted to assess variation within the gene, and whether this variation has any phenotypic effect in psychiatric disease. Crowley *et al.* sequenced *DISC1*, along with 9 other candidate genes (such as *DRD2, NRG1),* for variants in >700 schizophrenia cases and >700 controls[101]. Due to constraints, only limited regions of the gene were sequenced (exons, UTRs, promoters, splice sites, conserved introns) In addition to technical replication to ensure SNP validity, they chose a subset of 92 SNPs in *DISC1* and other genes to verify in a secondary dataset of >2,000 cases and >2,000 controls. *DISC1* had the highest case:control SNP ratio, and two nonsense SNP variants found only in cases (only three such SNPs were found in the 10 genes). Despite this, no SNPs were found significantly associated with schizophrenia. Crowley *et al.* noted that they were the fifth group to search association of schizophrenia with *DISC1* SNPs, and despite using a larger sample size had found no significant SNPs after multiple testing. Two previous groups had either failed to replicate findings in replicate samples[102] or had not found significance[103]. Another of the four groups had shown variants in patients which were not in 10,000 controls including a particularly interesting variant described below[104], while the findings of the last group involve the *DISC1* frameshift described above. The evidence suggests if variants other than the t(1;11) involve *DISC1* and predispose to schizophrenia, they are likely very rare indeed. In contrast, there is evidence suggesting rare mutations in *DISC1* might be important. Thomson *et al* sequenced 500kb around the *DISC1* locus in ~900 cases (schizophrenia, bipolar disorder, and rMDD) and ~650 controls, including the entirety of the gene as opposed to just coding and conserved regions[105]. They found 2,000 rare variants with a frequency of <1%. A single SNP was significant for rMDD in the original dataset and a combined original and replication dataset, but not any of three rMDD replication datasets. Thomson *et al.* noted the

abundance of rare variation and low power due to sequencing costs meant it is highly difficult to ascertain the current impact of rare variation, particularly for a gene such as *DISC1* in which variation might be pleiotrophic and only partially penetrant.

Nevertheless some SNPs are of particular interest and have been shown to have direct functional consequences as they encode point mutations. R37W and L607F are two such mutations. L607 is a conserved residue from mouse to zebrafish and the non-synonymous mutation is within two haplotypes both associated with schizoaffective disorder. It is associated with high risk ratio of the disease (>2.4) and was located in a region of *DISC1* that modulates interaction with ATF4[106]. Another study scanned 288 patients with schizophrenia for variation within *DISC1* and found several mutations not found in 10,000 control alleles, including the mammalian conserved R37W[104]. L607F was also found to be not associated with schizophrenia, but given that the original study associated it with the related but not identical schizoaffective disorder this is less worrying than it might initially appear. Both 37W and 607F were subsequently functionally investigated by Malavasi *et al.*[107]. ATF4 is a cAMP-response element binding protein which acts as a transcription factor, mediating the effects of cAMP. It has been shown that it is bound by DISC1 and acts to effect transcriptional changes involving apoptosis, mitochondrial function, synaptic plasticity, and repression of LTP[107]. Both mutations decrease the abundance of nuclear DISC1 by approximately 50%, as shown by both immunocytochemistry and western blotting. Both approaches also confirmed *DISC1* expression is unaltered, meaning that differential targeting of DISC1 was responsible. Exogenous DISC1 was also shown to inhibit the transcriptional activity of ATF4; both mutations decreased this inhibition, although it appeared that in 607F this was due to decreased protein-protein interaction, while with 37W it was likely due to nuclear exclusion of DISC1[107]. In any case this series of papers defined a number of risk-associated alleles with confirmed consequences for DISC1 biology.

1.4.3 IPSC and mouse *DISC1* models

This thesis expands upon previous work by investigating the RNA-Seq profiles of neuronal cells derived from iPSCs of members of the t(1;11) pedigree. It also looks

at RNA-Seq profiles of neural tissue from a corresponding mouse model, referred to as the *Der1* model. In this model 100kb of DNA has been removed from the mouse chromosome carrying *Disc1*, with this being removed from downstream of exon 8 and replaced with 115kb of human DNA corresponding to chromosome 11. The effects of the translocation upon *DISC1* are therefore mimicked in the altered mouse chromosome, which both heterozygous and homozygous carriers of exist[70]. The backgrounds of each of these models are discussed in turn with reference to other similar models. A summary of the models is given in Figure 2.

1.4.3.1 Mouse models of Disc1 mutations

Tomoda *et al.* have summarised some of the difficulties in exploring mutant *Disc1* mouse models; primarily the fact that it is not yet known exactly how the t(1;11) exerts its effects and how this alters *DISC1* in human. Numerous point mutation studies have indicated particular residues of the Disc1 protein as being particularly important in certain interactions, such as the L607 residue and ATF4 interaction described earlier[107]. Others have looked at Disc1 knockdown or frameshifts at particular developmental stages and the resulting effect on neuronal development, or behaviour, or have attempted to replicate some predicted effects of the t(1;11) on *DISC1*. Particularly relevant is the emergence of phenotypes related to psychiatric disease in some mouse models[108], the impact on dopaminergic signalling caused by *Disc1* abnormalities, and gene-environment interactions with *Disc1* mutation and known schizophrenia risk factors.

The early paper described by Clapcote *et al.* looked at two induced point mutations in exon 2, which encodes amino acids that are present in all isoforms and is proximal to the point where the t(1:11) occurs in the orthologous *DISC1* gene. These mutations were Q31L and L100P, and mice carrying these exhibited a large number of schizophrenia related phenotypes. Both mutations caused deficiencies in prepulse inhibition (PPI), a known phenotype of schizophrenia. Latent inhibition (the phenomenon by which a previously encountered stimulus takes longer to acquire a new meaning) was also decreased in both mouse models, as was the ability to bind the candidate psychosis factor and cAMP regulator PDE4B. Some of these behavioural phenotypes could be diminished in severity by antipsychotic or

antidepressant drugs, depending on genotype. Both mouse mutants also displayed decreased brain volume, another commonly observed phenotype in schizophrenia[108]. It was later demonstrated that maternal immune activation during gestation interacted with one of these *Disc1* mutations[109]. L100P offspring had enhanced IL-6 presence in foetal brains compared to WT or Q31L mice upon maternal immune activation. Similarly, genotype-alone PPI was normal, as was maternal immune activation alone, but both combined resulted in PPI deficits in offspring[109]. Lipina *et al.* noted that maternal immune activation is a risk factor for schizophrenia, which the L100P phenotype is putatively similar to, while it is not for depression, which the Q31L phenotype is putatively similar to. We therefore see a schizophrenia specific risk factor, interacting with a putative schizophrenia related mutation in a known disease gene *Disc1*. Not only this, but the L100P phenotype appears to cause increased dopamine release in response to amphetamine, and increased D2R striatal expression. Haloperidol, a D2R antagonist, curbed some of the earlier described phenotypes of latent inhibition and prepulse inhibition abnormalities[110], consistent with dopamine's role as a key molecule in the pathology of schizophrenia[53]. Finally, the L607F mutation found associated with schizoaffective disorder in humans and known to alter ATF4 interaction has recently been modelled in the mouse. The orthologous residue mutation L604F was induced by CRISPR and homozygous mice had PPI deficits. It will be interesting to see what future experiments with this schizoaffective risk allele model show[111].

A *Disc1* truncating mutation in mice, giving rise to an aberrant short form of Disc1, also exhibited phenotypes including reduced prefrontal cortex size and diminished short-term potentiation in the CA3:CA1 synapses of the hippocampus[112]. These shorter isoforms are of interest as Disc1 is known to self-associate. Truncated proteins might therefore exert a dominant negative effect. It must be cautioned however that expression of such truncated proteins is not confirmed in the t(1;11)[70,113]. The recent L604F mouse model appears to show increased Disc1 aggregation[111]. In any case, Tomoda *et al.* have noted that mutations modelling a truncated protein effect tend to have consequences for dopaminergic signalling[114]. A model in which the dominant Disc1 mutant is expressed in the cortex and

hippocampus displayed enhanced D2R receptor binding in the striatum, as well as corresponding hypersensitivity to methamphetamine[115]. The group also reported a decrease in parvalbumin-positive interneuron staining in the prefrontal cortex[115]. These interneurons are notable as both genetic and environmental risk factors for schizophrenia have been shown to cause their ablation[116]. Erbb4, encoded by a schizophrenia candidate gene, has been shown to play a role in the synaptic pruning of excitatory synapses onto these very interneurons during monkey adolescence, with lesser but stronger synaptic inputs remaining post pruning[117]. Dominant-negative Disc1 displays gene by environment interactions in common with point mutations, and as expected for an apparent schizophrenia-risk modelling mutation[118]. Immune activation provoked IL-1β, IL-4, and IL-5 release in wild type mice, but IL-2 in mutant Disc1 mice. Phenotypes such as hyperactivity, increased swim test immobility, and decreased sociability emerged only upon mutant *Disc1* and maternal immune activation. Abnormalities in brain structure were also detected[118].

The final posited effect of the translocation, and one which could conceivably be altered by epigenetic factors and the local transcriptomic environment, is altered DISC1 expression. In mice, this has been typically modelled by shRNA mediated knockdown. Niwa *et al.* experimented with *in utero* transfection of shRNA against *Disc1* in E14 foetal mice and showed that the resulting knockdown of Disc1 was transient and confined mainly to pyramidal neurons of the prefrontal cortex[119]. Cell migration and proliferation was impaired, as previously shown in *Disc1* mutant mice, but postnatal mice also displayed dopaminergic phenotypes. Adult mice displayed deficiencies in tyrosine hydroxylase positive neurons and dopamine levels in the prefrontal cortex, but not until P56. Dopamine levels were corrected by clozapine administration, and the altered mice also displayed hypersensitivity to methamine administration, a phenotype which did not emerge until P56. One conclusion is that the developmental lack of Disc1 results in inappropriate dopaminergic neuronal maturation, and a consequential adaptation of the brain to function on this “lower dopamine” level. Subsequently dopaminergic stimulation by methamphetamine, or readjustment of the system during adolescence (synaptic pruning), might cause an abnormal state. Most interesting was the finding of Niwa *et al.* finding that

parvalbumin immunoreactivity at P56 was decreased in the prefrontal cortex; implying a reduction of parvalbumin positive interneurons[119]. Disc1 deficiency can cause developmental abnormalities even without an exacerbating factor such as maternal inflammation, or lead/cannabinoid poisoning, which have also been shown to interact with *Disc1* in a gene by environmental fashion[114]. A summary of the point mutation, truncated, and knockdown models is given in Figure 2.

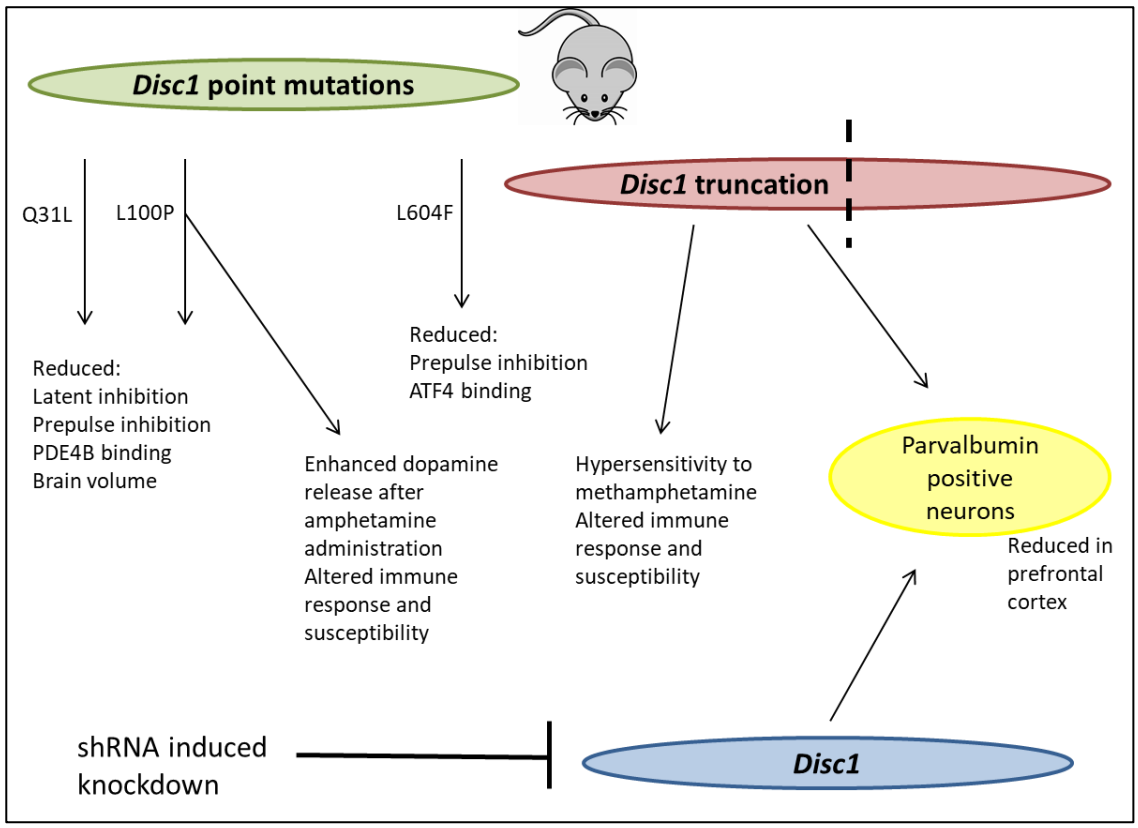


Figure 2. Illustrated findings utilised mouse models of *Disc1* point mutation, truncation, or knockdown. References given in text.

Although the exact effect of the t(1;11) upon DISC1 is unclear, there is evidence that point mutations, haploinsufficiency, and truncations all can cause phenotypes of interest including dopaminergic and parvalbumin positive interneuron abnormalities. It will be particularly interesting to see what role DISC1 dominant negative aggregation may play. These phenotypes are of known relevance to psychiatric disease, particularly in conjunction with a secondary risk factor.

1.4.3.2 IPSC models of DISC1 mutations

Induced pluripotent stem cells (IPSCs) are a cellular model derived from a non-pluripotent cell type which has been 'reprogrammed' into a pluripotent, or stem cell-like, state. iPSCs were first generated using mouse fibroblasts in 2006 and from human fibroblasts in 2007. They exhibited all the hallmarks of pluripotent cells including teratomagenicity, proliferative capacity, stem cell gene expression, telomerase activity, and appeared to retain these phenotypes throughout division[120]. Although initially iPSCs were created by retroviral transfection of Yamanaka factors (a group of key genes expressed by stem cells including *Oct3/4, Sox2, c-Myc, Klf4* in the original mouse experiment), subsequent experiments have utilised small molecules or plasmid constructs to minimise retroviral-induced insertional mutagenesis[121,122]. The research potential of induced pluripotent stem cells (IPSCs) has become increasingly realised in the post-Yamanaka world[120,122]. iPSCs have been used to model neurological and muscular diseases[123], but also schizophrenia[124]. iPSCs are specific to the individual and can generate otherwise hard to obtain neural cells, two factors making them of prime importance in the study of pedigrees with psychiatric illnesses. This will allow more "true to life" phenotypic analysis. This is a must, as laboratory recreations of biological events are not hypothesis free. As discussed earlier, several mouse models investigate aspects of *DISC1* biology using RNAi against all isoforms[125], point mutations, or truncations to produce dominant negative isoforms. The hypothesis that *DISC1* knockdown (or dominant negative effects) is of relevance to the effects of the t(1:11) translocation is a valid one. Yet the model, however efficacious, will not model any other effects such as differential methylation of the t(1:11). It has been shown that there is differential methylation associated with the t(1;11) in blood samples from carriers and controls, with most of the significant loci being at chromosomes 1 and 11 and close to the breakpoints[126]. It is unknown if *DISC1* disruption alone would replicate these and other effects. Indeed, given the importance of epistasis, and the relative ignorance we have of psychiatric disease aetiology, the creation of a model with the entire genetic background is a must, in order to accurately model any "two-hit" effects. To date the t(1;11) remains a well evidenced example of a psychiatric disease causing mutation involving *DISC1*. As discussed already, other mutations involving *DISC1* do not

have the same calibre of evidence, or are not as penetrant. *DISC1* is of valid biological interest by itself; but *DISC1* disruption is only one of the outcomes of the t(1;11). iPSCs allow the close approximation of a real biological scenario, including not only well-studied events such as the t(1;11) but all genomic information. In any case the complexity of psychiatric disease, involving multiple cell types and interacting brain regions[127], as well as a neurodevelopmental trajectory, will make complete reconstitution of phenotypes impossible. There are also significant drawbacks to iPSC-derived cell models. Yamanaka summarised some of these issues in a review[122]. The first major problem is inherent variability in the generation of iPSCs. The process generates mutation, and in any case the cells are derived in a clonal process typically from fibroblasts. Whether by production of mutation or selection of cells which carry mutation, iPSC-derived cells inevitably carry some variants not in the host genome. Yamanaka's review noted that there is some reported variation in iPSC differentiation efficiency and gene expression, but suggested much of this might be due to inherent variation from one laboratory to another and noted larger studies were less likely to find differences between iPSCs and embryonic stem cells[122]. However a later paper did find some differences *within* iPSC lines, which is of more relevance to my approach than iPSC and embryonic stem cell differences[128]. 40 iPSC lines were differentiated to neurons, with approximately 75% of these showing less than 1% undifferentiated cells after 14 days. However, a minority, approximately 20% of lines, showed more than 10% undifferentiated cells and were designated as "defective". These lines had differential expression of retroviral associated elements. Subsequently, lines were differentiated to dopaminergic neurons and implanted in mouse brains. Mice that received a "defective" transplant had greater graft sizes and had teratomas in 85% of cases, compared to 22% of cases in non-"defective" lines. Finally, many but not all "defective" grafts continued to express pluripotent markers despite a 30 day differentiation protocol and a 60 day implantation period. We can clearly see that there are issues with line variability in iPSCs. The intrinsic variability of iPSC-derived models is one problem; the other is that many of the phenotypes of interest (synaptic strengthening, etc.) change with time. Therefore, differences between cell lines may represent differences in culture time, which must be carefully controlled

for. Another problem is the relative immaturity of iPSC-derived neurons. Dolmetsch and Geschwind noted that these neurons are poorly characterised, are usually synaptically immature, and few fire action potentials[129]. Since this research project began, it has been reported that extended differentiation protocols result in a high proportion of electrophysiologically active neurons; such protocols require between 56 and 70 days of culturing and result in mixed cultures primarily of astrocytes and neurons[130]. Another group also reported that they could produce iPSC-derived neurons which appeared to have dendritic spines; this production required an equally long differentiation time, twice as long as that used in our protocol[131]. Although there have evidently been advances in iPSC-neuron techniques, the issues of variability remain, and the protocol utilised to generate the neurons described in this thesis does not produce neurons with spines.

Previous researchers have generated and utilised an iPSC-model carrying the t(1;11), as well as controls from the family without the translocation. They were subsequently differentiated to neural precursor cells, which were then differentiated to neurons and harvested for RNA utilised in RNA-Seq (see Materials and Methods).

1.5 Deconvolution

RNA-Seq data can be mined for a wealth of information other than the quantification of transcript levels for any one gene. Typically, as described in this thesis, RNA-Seq data is generated not solely from a unique cell type of interest but from a heterogeneous mixture of cells. If cells are collected from tissue, they will be heterogeneous given that tissues contain multiple cell types. If grown *in vitro* primary cell culture rarely results in a single cell type. Differentiation of neurons from iPSCs generates multiple neuronal cell types as well as non-neuronal cells[132]. This creates a challenge; we cannot directly distinguish between transcriptional changes caused by a relative change in cell proportions from those changes caused by a relative change in cell properties and/or activity. Both scenarios have biological relevance. The depletion of a particular subset of cells is of relevance to diseases as diverse as type 1 diabetes, Parkinson's, and possibly schizophrenia[116]. Inappropriate or diminished activity of cells, without their relative proportions changing in any

way, can also be a pathological mechanism. Common variants predisposing to schizophrenia have been suggested to converge on only a few cell types, for example. Both schizophrenia associated mutations and environmental risk factors affect parvalbumin positive interneurons[116], while a recent analysis by Skene *et al.* found that SNPs associated with schizophrenia are highly likely to affect genes which are specific to, or highly enriched in, certain cell types[3]. In deconvolution, we have a single signal comprised of multiple, individual components, a problem also presenting in audio signal and image processing, although in the case of RNA-Seq the components are cell types[134]. The goal is to "deconvolute" the data, changing it from one single convoluted signal into multiple deconvoluted signals. In a mixed cell RNA-Seq, each signal will correspond to a particular cell type. This will allow gene expression changes caused by variation in cell type to be distinguished from those caused by changed cell properties, in theory.

The mathematics of deconvolution is relatively straightforward. If we have RNA-Seq samples for each of the pure cell types that make up the mixed sample, it is assumed that the mixed sample will have a gene expression value equivalent to the sums of the gene expression values of each pure cell type, weighted by their proportion in the mixture. If we have the expression of the gene in every pure cell type, then each gene produces an equation which indicates what proportions each cell type are in. If more genes than cell types are measured, then these equations can be solved to find out what the proportions are. In practice proportions usually cannot be found which satisfy the equation for all genes. However this problem is a well-established one in mathematics and a method known as the non-negative least linear squares method can be utilised to give the proportions that sum to one and minimise error for each separate equation. I utilised the DeconRNASeq package developed by Gong *et al* to carry out deconvolution[135]. A further discussion of deconvolution can be found in Chapter 5.

1.6 Key papers

Various experimental approaches have been taken by several other research groups interested in the biological effects of *DISC1* mutation. In this thesis, I compare and

contrast some of their findings with my own. A discussion of some key papers is contained below, and the result of comparisons to those papers follows in the main body of the thesis.

1.6.1 "Disrupted in Schizophrenia 1 Interactome: evidence for the close connectivity of risk genes and a potential synaptic basis for schizophrenia"-Camargo *et al.* 2007[73]

This paper was the first to utilise a yeast two-hybrid screen to identify potential DISC1 protein interactors. 34 high-confidence interactors were identified in the initial study of full length DISC1 as a bait protein; a second round of yeast two-hybrid screening was carried out with 8 of these as well as the N-terminal 350aa of DISC1 encoded proximal to the breakpoint. In total, 127 proteins were identified as interacting with DISC1 or one of the 8 interactors, with some multiple interactors identified. As expected, many established interactors such as PDE4B, NDEL1, and LIS1 were re-identified by this approach. Overrepresented GO terms among the potential DISC1-interactome related especially to cytoskeletal processes such as tubulin and dynein interactions, as well as actin based transport. These processes are especially relevant to multiple stages of neuronal migration[82]. DISC1 had a particularly large number of such overrepresented GO terms among its potential interactors. Finally, the group also screened using a DISC1 construct that was truncated at the translocation breakpoint. This construct's interactors were quite different from that full length DISC1; only 16 were in common, while it interacted with 15 proteins that the full length DISC1 did not. It also lost interactions with over 20 binding partners, and presumably their binding partners which full length DISC1 indirectly interacts with (although it is of course possible that some of its novel partners interact with these secondary partners). This did not unambiguously suggest either a gain of function or loss of function mechanism was at play for the putative effects of truncated DISC1, but was a very interesting experiment highlighting that either could be responsible.

1.6.2 "Modelling schizophrenia using human induced pluripotent stem cells"-Brennand *et al.* 2011[124]

This paper was the first account of iPSCs being used to study the phenotypes of schizophrenia. Researchers at the Gage lab utilised the classic lentiviral transfection approach to generate iPSCs from the fibroblasts of individuals with idiopathic schizophrenia, as well as age matched controls. Subsequently, the cells were differentiated to neural precursor cells, as well as neurons. This has been the usual method of generating neurons and our cells have followed a similar route. Brennand *et al.* aimed to determine whether any of the phenotypes observed in post-mortem studies were replicated, such as reduced spine density, or whether receptors of relevance to schizophrenia were dysregulated. The majority of these phenotypes were not replicated, which may be due to the limitations of neural development in a 2D culture. Levels of synaptic PSD95, as well as VGAT, VGLUT1, GLUR1, and SYN were normal or not significantly different, although unlike some other papers they did not report a PSD95/SYN1 colocalisation assay. Electrophysiology and calcium imaging also did not reveal any difference between cells derived from individuals with SZ and those derived from control individuals. However, an RNA microarray indicated dysregulation of several hundred genes, 25% of which were previously implicated in SZ by post-mortem dysregulation or by association. The paper also included an experiment utilising the rabies virus, which spreads via synaptic connections. Connectivity (as measured by the ratio of initially infected cells to that of secondary infected cells, which could not spread the virus) appeared to be lower in the non-control cells, indicating possible deficits in neuronal organisation in SZ.

Although its importance as a new application of iPSC technology to psychiatric disease research cannot be doubted, the Brennand *et al.* paper does have its limitations. Many of the disease phenotypes were not replicated, which might be due to the high variability of the cell lines (indicated by varying prevalence of GAD67+ cells between all lines). Alternatively, phenotypes may be more subtle. The pathways implicated by the RNA-array analysis carried out by Brennand *et al.* included Wnt and cAMP signalling, which are of relevance to psychiatric disease. A number of

these changes in genes such as *WNT7A*, *TCF4*, *AXIN2*, *RAP2A* and several phosphodiesterases (PDE4 family) were verified by qPCR. The Wnt pathway is involved in processes such as β-catenin signalling, inhibited by GSK-3β (itself inhibited by Lithium, a mood stabiliser[136]), while cAMP signalling has well documented effects on neural transcription and memory[85]. Changes in these pathways could lead to phenotypes not entirely obvious in cell culture. DISC1 of course, exerts effects on both pathways.

A second paper from the Gage lab was published in 2014 and focused on the same cells, looking at neurotransmitter release[137]. The study was an analysis of basal and post KCl-stimulated neurotransmitter release, with technical replications. The cells stained positively for enzymes involved in catecholamine processing (such as dopamine decarboxylase, dopamine-β-hydroxylase, prohormone convertases and cathepsins), and showed that the iPSC-derived neurons originally from schizophrenic patients had an increase in the proportion of tyrosine hydroxylase positive neurons. This was mirrored by an increase in both basal and KCl stimulated catecholamine release. However, the paper also highlighted once again the variation between cell lines, even from the same patient source. Although the averages showed clear differences between SZ and WT lines, intragroup variation was high and the two groups were not cleanly distinguished from one another. This is a recurring issue with iPSC-derived models, and is a stumbling point for research. It is also difficult to draw conclusions from the tyrosine hydroxylase cell increases; Niwa *et al.* 's earlier discussed paper showed that a model of *DISC1* mutation had reduced cells of this type[119]. Of course, the pathology of schizophrenia is more complex than a simple increase or decrease, and what cell type and when these changes occur will be important to eventual pathology.

1.6.3 "Synaptic dysregulation in a human iPS cell model of mental disorders"-Wen *et al.* 2014[132]

Further papers were to follow on the heels of those of Gage and colleagues. "Synaptic dysregulation in a human iPS cell model of mental disorders" by Wen *et al.* was published in 2014 and employed advances in iPSC generation methods, using

non-integrating plasmids as gene vectors. The group harvested fibroblasts from a small pedigree with a *DISC1* frameshift and a high rate of psychiatric disease, inducing pluripotency and differentiating the resulting cells to NPCs and neurons. The use of a *DISC1* frameshift carrying genotype is of obvious relevance to the t(1;11), and the group were able to establish that DISC1 protein levels were depleted, with a corresponding increase in DISC1 ubiquitination. The group also took control and carrier iPSC lines and induced or corrected the *DISC1* frameshift, before differentiating these lines to neurons. TALEN mediated correction of the frameshift in carrier lines replenished levels of DISC1, as expected. In a similar manner, TALEN mediated causation of the frameshift caused loss of DISC1 protein in neurons. Given DISC1's role as a scaffold protein, and its presence in the synapse, it is not surprising that Wen *et al.* elected to investigate synaptic phenotypes. They observed phenotypes including deficits in PSD95/SYN1 co-localisation (indicating decreased levels of synaptic maturity) and reduced levels of synaptic vesicle protein 2, both capable of being induced or corrected by the presence of absence of the *DISC1* frameshift. This is strong evidence in favour of DISC1 having an important role in synaptic strengthening and formation. The group also carried out RNA-Seq on their cells, although the number of lines was limited. Over 2,000 genes were differentially regulated, with the top three overrepresented gene ontology terms being "synaptic transmission", "nervous development", and dendritic spine". Given the problem of variability, the near linearity between the RNA-Seq and qPCR (which used different differentiations of the same cell lines) is encouraging and the changes they verified by qPCR point towards synaptic dysregulation, matching their protein level phenotypic analysis. Changes in our RNA-Seq analysis which point towards the synapse and agree with Wen *et al.* should therefore be regarded with a greater measure of trust, especially given the similarities between the t(1;11) and a *DISC1* frameshift. It is also impressive that gene-corrected/altered controls displayed the same synaptic phenotypes; although these controls were not subjected to RNA-Seq.

1.6.4 "Genomic DISC1 Disruption in hiPSCs Alters Wnt Signaling and Neural Cell Fate"-Srikanth *et al.* 2015[138]

An interesting paper by Srikanth *et al.* also looked at the effects of *DISC1* frameshifts. The group induced frameshifts in either exon 2 or exon 8 of *DISC1* using targeted nucleases, resulting in iPSC lines with premature stop codons. The resulting iPSC lines were homozygous for exon 2 frameshifts (ex2mm) or had one (ex8wm) or two (ex8mm) exon 8 frameshifts. The lines appeared to retain all their characteristic iPSC features. A particular strength of this paper was its breadth: the group looked at the above three genotypes at two developmental time points, neural precursor cell and neuron. They noted that all mutant neural precursor cells displayed extensive NMD (nonsense mediated decay) on the more 3' exons (9, 11, 12/13), with only the ex8 mutants continuing to display late exon NMD into the neuronal stage. Of particular relevance to the t(1;11) was the discovery that ex8wm cells (d40, versus the d50 timepoint for neurons and d17 for neural precursor cells) had approximately ½ the wildtype level of DISC1 protein, while ex8mm appeared to display a complete absence. Although the ex8wm is not an exact replica of the t(1;11) it is interesting that they saw such a clear relation between genotype and phenotype. Most interesting of all however, was the observation that ex2mm cells had no protein at the 85kb band (WT DISC1 size) but had a novel band at ~64kb. Novel transcripts have been detected in the t(1;11), although a corresponding protein has not been detected.

DISC1 frameshifts also appear to impact development. Srikanth *et al.* found that ex8wm, ex8mm, and ex2mm NPCs all had downregulated levels of the cortical genes *FOXG1* and *TBR2*, with even more changes found solely in the ex2mm. At the neuronal stage the changes in *FOXG1* and *TBR2* were found to persist, and once again an additional set of changes were found in the ex2mm genotype involving decreases in neuronal receptors (*VGLUT1*, *GRIN1*) and cortical markers (*CTIP2*, *TBR1*, *FEZF2*). Srikanth *et al.* hypothesised that this represented a subtle shift in cell fate, and utilising RNA-Seq found that some ventral progenitor markers were decreased while dorsal ones were increased, although it should be noted that the changes were not universally significant. Believing that the changes may be due to

disinhibited Wnt activity, the group displayed by means of a TCF responsive luciferase assay that basal and stimulated Wnt signalling was higher in the mutant cells, especially the ex2mm. The scenario is evidently quite complex however: the Hh inhibitor cyclopamine did not alter the basal or stimulated signalling levels, although interestingly it did alter FOXG1 expression. Both Wnt agonism and antagonism (applied from days 7-17 of differentiation) exhibited the ability to alter cell fate markers in NPCs, although their effects on Wnt signalling were more complex, with agonism actually appearing to decrease signalling. Both Wnt agonism and antagonism exhibited the ability to alter cell fate markers in NPCs. Wnt antagonism increased FOXG1 and TBR2 in these cells, while agonism further decreased them and also increased MAP2 (neuronal marker) expression, more drastically in the ex8wm cells compared to WT ones. It appears that although the relationship between Wnt signalling and cell fate is not exactly a linear, simple one, Wnt signalling abnormalities prompted by DISC1 disruption appear to have effects on cell fate. The exact timing of the abnormalities appears to be important, with altering of signalling resulting in shifted cell fate marker expression even though the Wnt signalling levels later appeared to not be greatly affected.

To summarise, it appears as though several of the papers utilising iPSC-derived neurons as a cellular model for SZ agree on some core concepts. Synaptic dysregulation is a prevalent theme, while evidence supporting a corresponding electrophysiological dysregulation is less strong. In the work of Brennand *et al.*, these experiments had the greatest power and showed no positive results, while in Wen *et al.* electrophysiological abnormalities were inconsistent, and usually only evident in comparison to a certain control line. Wnt dysregulation is emerging as a common theme in DISC1 disruption, with cells often displaying abnormal expression of key Wnt signallers. The paper from Srikanth *et al.* is perhaps the best displayer of this trend, although it should be noted that the phenotype is complicated and that Wnt antagonism and agonism did not give expected, binary phenotypes. It appears that temporal factors are of importance. To an extent the prediction of Brennand *et al.* in 2011 is bearing out, as more data becomes available an ever-narrowing number of

genes are being consistently disrupted across all models of schizophrenia, perhaps hinting at pathways which are ubiquitously disrupted in schizophrenia.

1.6.5 "Biological insights from 108 schizophrenia-associated genetic loci"-Schizophrenia Working Group of the Psychiatric Genomics Consortium 2014[29]

The Common Disease Common Variant model of schizophrenia genetics predicts that many variants, each only mildly predisposing to schizophrenia, will exist in the population. These will be significantly enriched in schizophrenia cases compared to controls but by no means will they be absent from the control population. These variants will require large power to detect, and correspondingly large sample sizes. At the time, this paper was the largest genome-wide association study (GWAS) for schizophrenia. Utilising approximately 37,000 cases and 113,000 controls, this study looked at 9.5 million single nucleotide polymorphisms (SNPs). After disregarding SNPs in low linkage disequilibrium with more significant SNPs, they found 128 SNPs implicating 108 loci. SNPs were significantly more likely to be transmitted to an affected offspring in parent-offspring trios and more likely to be found significant in the replication cohort. Over 300 genes were implicated in these loci, although of course it is likely that only one of the genes associated with SNP locus is actually affected by the SNP in such a way to increase risk for schizophrenia. It is also, of course, possible that the affected gene is at a great distance from the SNP (loci boundaries were defined by SNPs which highly correlate with the putative risk SNP). 40% of loci contained only one gene, and genes included *DRD2*, many voltage gated calcium channels such as *CACNA1C, CACNB2,* and glutamatergic receptors such as *GRIA1, GRIN2A, GRM3,* some of which have roles in synaptic plasticity. Of particular interest was that the genes implicated had significant overlap with those found with *de novo* non-synonymous mutations in schizophrenia, intellectual disability, and autism spectrum disorder. As well as offering support to the "spectrum" model suggested by Craddock and Owen[12], this also suggests that rare and common mutation converge on certain genes, a finding which bolsters the relevance of this thesis.

1.6.6 "Common schizophrenia alleles are enriched in mutation-intolerant genes and maintained by background selection"-Pardiñas *et al.* 2016[60]

This new study, also by the Psychiatric Genomics Consortium, built upon their previous paper and utilised some of the same data. Here, the total sample was smaller (approximately 11,260 cases and 24,500 controls) but was phenotypically more consistent. This sample (derived from the CLOZUK cohort) had been filtered for high homogeneity based on genomic ancestry, limiting effects that could be due to population differences between cases and controls. Most patients were taking clozapine regularly, the exact figure was not given but 96% of the initial 15,000 were, meaning that of the post-filtered group of 11,260 at least 94.7% must have been taking clozapine. 18 loci were discovered as significant. Approximately half of the cases and three quarters of the controls were also in the previous study (the PGC sample) with 37,000 cases and 113,000 controls; removing these from that group gave the PGC independent sample. The PGC independent sample and the CLOZUK derived one had high genetic correlation and agreement on the direction of SNP effects. Meta-analysis of the CLOZUK sample and the PGC independent sample gave 177 significant SNPs at 143 loci, 50 of which were novel. 98 of the loci appeared to implicate only a single gene; these included *PDE4B, ERBB4, NRXN1,* as well as *CACNA1D* and *GABBR2,* implicating calcium and GABA signalling. Many genes from the previous GWAS were also implicated again. Finally, gene set enrichment analysis revealed that genes of the sets "targets of FMRP" (discussed earlier in this introduction), "$5HT_{2C}$-receptor complex", "voltage -gated calcium channel complexes", and those of "abnormal long term potentiation" were significantly enriched for hits. This paper not only reaffirmed the importance of calcium signalling, but also gave several hits novel for GWAS which implicated classic candidate genes and *PDE4B,* which encodes a DISC1 interactor.

1.6.7 "Contribution of copy number variants to schizophrenia from a genome-wide study of 41,321 subjects"-CNV analysis group of the Psychiatric Genomics Consortium 2017[139]

SNPs had been studied with large sample sizes, and the work of Guilmatre *et al.* and others had shown the importance of CNVs to the genetics of schizophrenia[48]. This

large study utilised ~21,000 cases and ~20,000 controls for the study of CNVs relating to schizophrenia. They found that in general schizophrenia cases had more CNVs (x1.03), which were larger (x1.1), and which contained more genes (x1.2). Most CNVs had a modest effect, and a number of rarer CNVs had a novel association with schizophrenia. Of 28 gene sets of relevance to schizophrenia, 15 were enriched for excess CNV loss in cases and four for excess CNV gain in cases. None of 8 control sets were enriched in any way. Genes associated with GO terms "synapse" and "ARC complex" were highly significantly enriched for CNV loss and appeared to drive most of the significance of the other sets. Genes from these sets overlapping with CNVs had extensive protein-protein interactions with synaptic molecules including pre and postsynaptic markers, as well as glutamatergic receptors. They attempted to further delineate the exact loci which drove CNV significance and described 8 of these as being of genome-wide significance.

1.6.8 "Genetic identification of brain cell types underlying schizophrenia"-Skene *et al.* 2018[140]

Skene *et al.*attempted to determine what cell types the mutations discovered in the above papers might exert their effects in. They used a superset of scRNA-Seq data from the Karolinska Institute. Their hypothesis was simple; if schizophrenia affects certain cell types, mutations which predispose to schizophrenia should be overrepresented in genes which are exclusive or near exclusively expressed in those cells. Schizophrenia "genes" could of course be more akin to housekeeping genes which are universally expressed. But this did not appear to be the case. Using two types of analyses and the entirety of the schizophrenia associated SNPs from Pardiñas *et al.*, they showed that enrichment for genes associated with these SNPs was found in several groups of genes. These included those highly specific to hippocampal CA1 pyramidal cells, striatal medium spiny neurons (MSNs), neocortical somatosensory pyramidal cells and cortical interneurons. More specifically, the group of genes highly specific to MSNs expressing *Drd2*, the group highly specific to MSNs expressing *Drd1*, as well as the group highly specific to parvalbumin positive interneurons were all heavily enriched for schizophrenia hits. This finding appeared to be specific for schizophrenia; depression, years of

education, and height related hits produced different significantly enriched groups (although MSNs-related genes were significant in the education investigation as well). In general, these findings were replicated using other datasets, and the same cell types were highlighted as having gene enrichment when instead of GWAS hits the input was "genes affected by antipsychotics" or gene sets associated with schizophrenia such as "NMDAR complex", "PSD-95 complex", and "FMRP associated genes". The effect was not found when gene sets related to Alzheimer's disease or migraine were used. The paper was a highly interesting look at what cell types schizophrenia genetics converges on; the emergence once again of dopaminergic and parvalbumin positive interneurons as being important to schizophrenia is striking.

1.7 Hypothesis and Aims of the PhD

It is clear that the genetics of psychiatric illness, although characterised by both common and rare variation, have still not been fully elucidated. It is also not entirely evident how the t(1;11) exerts its effects. Research has however highlighted particular pathways related to synaptic activity and neurodevelopmental as being important in pathogenesis. Large CNV and GWAS studies have found some candidate genes, while studies of iPSCs and *DISC1* have indicated the possible importance of cAMP and Wnt signalling in t(1;11) pathogenesis. In addition, new computational techniques can be applied to RNA-Seq profiles of iPSC-derived neurons, so as to go beyond the gene level changes and see if particular patterns emerge.

The aim of this PhD was to study iPSC-derived neurons of the Scottish pedigree, and compare and contrast RNA-Seq profiles of lines with and without the translocation. I also planned to utilise the *Der1* mouse model. By this I aimed to see

1) What pathways or genes are altered by the translocation, with particular interest in those which are amenable to further investigation. Further experiments would depend on the nature of the altered genes.
2) If there were convergences between genes altered by this unique mutation and those altered by more common SNPs and CNVs predisposing to

psychiatric illness. This would indicates that the t(1;11) is a good model for psychiatric illness more generally, and that insights from its investigation might be applicable to the field as a whole.

3) If there were convergences with *DISC1* models, which would be more evidence that the translocation exerts effects via effects on DISC1.
4) If the changes observed could be linked to changes in relative proportions of certain cell types.
5) If the changes observed could be related dysfunctions in particular cell types and what this might mean on the molecular level.

To answer questions 1-3, I utilised RNA-Seq investigation, comparing and contrasting the list of differentially expressed genes to those lists from major papers discussed earlier. I verified changes of particular importance utilising qRT-PCR with the idea of producing further experimental ideas. I aimed to answer question 4 using my deconvolution approach, and question 5 utilising the EWCE approach of Skene *et al.*

2 MATERIALS AND METHODS

2.1 Generation of induced pluripotent stem cells (iPSC) from dermal fibroblasts

The below text is replicated exactly (excepting formatting and clarification of some statements) from the Supplementary Materials of Malavasi et al[70]. I did not contribute to the production of the iPSC lines. As stated in my thesis declaration, only methods which have an explicit declaration of non-contribution as above were not carried out by the author (me). All these methods are clearly noted in the text and referenced. For convenience I have stated at the start of each section whether it was carried out by me, by collaborators, or in collaboration together.

Ethical consent relating to the translocation family is as follows: Prior to 2014, Lothian Research and Development (2011/P/PSY/09), Scotland A Research Ethics Committee (09/MRE00/81); 2014 onwards, Lothian Research and Development (2014/0303), Scotland A Research Ethics Committee (14/SS/0039). Fibroblasts were cultured in DMEM with 10% FBS (all media and supplements used in 2.1 and 2.2 from Life Technologies unless stated otherwise) at 37^{o}C with 5 % CO_2. Fibroblasts were reprogrammed by non-integrating methods, using episomal plasmids[141]. For episomal reprogramming of some lines we used a protocol adapted by Tilo Kunath, (University of Edinburgh) and Roslin Cells (roslincells.com). Other lines were reprogrammed by Roslin Cells. Plasmids incorporating Oct3/4, shRNA to p53, SOX2, KLF4, L-MYC and LIN28 were electroporated into fibroblasts using Nucleofection (Amaxa, Lonza). The episomal plasmids pCXLE-hOCT3/4-shp53-F (for OCT3/4 and p53 knockdown), pCXLE-hSK (for SOX2 and KLF4) and pCXLE-hUL (for L-MYC and LIN28) were a gift from Shinya Yamanaka. (These correspond to Addgene plasmids 27077, 27078, 27080). 5 x 10^5 fibroblasts were transfected with 1.7ug of pCXLE-hOCT3/4-shp53-F, 1.6ug of pCXLE-hSK and 1.7ug of pCXLE-hUL using the NHDF Nucleofector kit and Amaxa Nucleofector protocol U-023 (1,650 V, 10 ms, 3 time pulses), according to the manufacturer's instructions. Cells were then seeded into one well of a gelatin-coated 6 well tissue culture grade plate in Opti-MEM (ThermoFisher Scientific) supplemented with 10% FCS and 1% Antibiotic Antimycotic Solution (Life Technologies 15240062). All stages of cells were maintained in media supplemented with Antibiotic Antimycotic Solution

thereafter. The medium was replaced every 2-3 days before cells were replated into a 10cm Vitronectin/Geltrex (ThermoFisher Scientific) or Matrigel (Life Technologies)-coated tissue culture grade dish after 6-7 days.

The following day the medium was changed to Essential 6 medium (ThermoFisher Scientific) with added 100 ng/ml bFGF (Peprotech). The medium was changed every 2 days until colonies were ready to be picked, at approximately day 25-30. Individual colonies were picked and expanded into 12, then 6, well Vitronectin/Geltrex or Matrigel-coated tissue culture grade plates in Essential 8 medium (ThermoFisher Scientific) with daily medium changes. Cells were passaged using 0.5mM EDTA in PBS. iPSCs were generated and cultured at 37^{o}C, 5% CO_2 and 21% O_2. Quality control of iPSC lines was performed after clonal passage 10.

Pluripotency of iPSC lines was assessed using markers SSEA-1 PE, SSEA-4-AlexaFluor647 and Oct3/4 PerCP-Cy5.5 and isotype controls using the BD Stemflow Human and Mouse Pluripotent Stem Cell Analysis Kit (BD Biosciences, 560477) according to the manufacturer's instructions, or with SSEA-1-APC (301907), SSEA-4-FITC (330409), TRA-1- 60-PE (330609), TRA-1-81-PE (330707) and isotype controls (all BioLegend) as follows: iPSCs were dissociated using StemPro Accutase (Life Technologies) and washed with Essential 8 medium. 1x105 cells were incubated with antibodies in 2% Fetal Bovine Serum in PBS for 1 hour on ice. Cells were washed once with 2% BSA/PBS, centrifuged at 200 x g for 5 min and resuspended in 200 ul 2% BSA/PBS. FACS analysis was performed on single cell suspensions using a FACS Aria cell sorter (BD Biosciences). Data were analysed using FlowJo v10 software.

EBNA-1 primer sequences and amplification protocol were taken from the Epi5 Episomal iPSC Reprogramming Kit (Life Technologies, A15960). Genomic DNA was extracted from iPSCs using the DNeasy kit (Qiagen). Cells that had not been in contact with episomes, were used as negative controls. Positive controls were low passage iPSC lines where episomes were still present. A non-template control (NTC) was also used. iPSC lines were only taken forward once episomal clearance had been confirmed by this method (data not shown).

The human Cytoscan 750 K Array (Affymetrix) was used to identify genomic abnormalities in the iPSC lines. The array consists of 550,000 unique non-polymorphic probes and 200,000 SNPs for accurate genotyping. Genomic DNA was extracted from iPSC clones using the DNeasy kit (Qiagen) according to manufacturer's instructions. Samples were sent to the NHS Cytogenetics Laboratory (Western General Hospital, Edinburgh) for processing of the arrays. Chromosome analysis was performed using Chromosome Analysis Suite version 2.0 (Affymetrix). Copy number, breakpoints and Loss Of Heterozygosity (LOH) regions were determined using the models and algorithms incorporated within the software package. To exclude possible false positives due to inherent microarray noise the CNV threshold of gains and losses for inclusion in analyses was 10 kilobase pairs (kbp) and 10 consecutive markers. iPSC lines with deletions or duplications greater than 5 MB, the limit typically applied by G-banding, were excluded from further studies.

2.2 NPC culture and neuronal differentiation

The below text is replicated exactly (excepting formatting) from the Supplementary Materials of Malavasi et al[70]. I did not contribute in the production of the neurons which were utilised and described in this thesis.

iPSCs were converted into neuroectoderm by dual-SMAD signalling inhibition[142]. Long-term anterior neural precursor cells were generated and maintained under physiological normoxia (3% O_2) and in the absence of EGF3. NPCs were cultured at 37^{o}C, with 5% CO_2 and 3% O_2 on Matrigel (Life Technologies)-coated 6 well tissue culture grade plates in Advanced DMEM/F-12 (Life Technologies) with 1% Glutamax-1 (Life Technologies), 1% N2 supplement (Life Technologies), 0.1% B27 supplement (Life Technologies), 10 ng/ml bFGF (PeproTech) and 1% antibiotic/antimycotic solution (Life Technologies). NPCs were maintained up to passage 30 with feeding every 2-3 days and weekly passages using StemPro Accutase (Life Technologies). All NPC lines were tested every week for mycoplasma infection. For differentiation into cortical forebrain-like neurons[143], NPCs were plated into Matrigel (Life Technologies) and Laminin (Sigma-Aldrich)-

coated 12 well tissue culture grade plates in Advanced DMEM/F-12 with 0.5% Glutamax-1, 0.5% N2 supplement, 0.2% B27 supplement, 2 µg/ml Heparin and 1% antibiotic/antimycotic solution (Life Technologies). Neurons were maintained for 5 weeks with feeding as necessary. During weeks 2 and 3 the neuronal differentiation medium was supplemented with Forskolin (Tocris Bioscience). During weeks 4 and 5 the Forskolin was removed, and the medium was supplemented with BDNF (Life Technologies) plus GDNF (Life Technologies) to 5ng/ml each.

2.3 Human cDNA synthesis

I carried out synthesis of cDNA to produce standard curves and test primers. cDNA was synthesised from human cerebral cortex RNA using a 40µl reaction mix comprised as follows: 4µl GeneAmp 10x PCR buffer II, 8.8µl MgCl2 solution (both #N8080130, Life Technologies, Paisley, Glasgow, PA4 9RF), 2µl GeneAMP dNTP solution (#4303442, Life Technologies, Paisley, Glasgow, PA4 9RF), 0.8µl RNase Inhibitor (#N8080119, Life Technologies, Paisley, Glasgow, PA4 9RF), 1µl Multiscribe Reverse Transcriptase (#4311235, Life Technologies, Paisley, Glasgow, PA4 9RF), 4µl Random Hexamers (#N8080127, Life Technologies, Paisley, Glasgow, PA4 9RF), 18.4µl Ultrapure DNase/RNase-Free Distilled Water (#10977035, Life Technologies, Paisley, Glasgow, PA4 9RF), and 1µl of human cerebral cortex RNA solution (#636561, Takara Bio Europe, Saint-Germane-En-Layn, France) as supplied by Takara Bio Europe, containing 1µg of RNA. The following protocol was utilised.

The combined RNA and Ultrapure water solution (19.4µl total) was first denatured using the following protocol;

65^{o}C 10 minutes

4^{o}C 5 minutes

And was subsequently placed on ice. The rest of the reagents were added and the tube subjected to the following protocol;

25^{o}C 10 minutes

48^{o}C 30 minutes

95^{o}C 5 minutes

Tubes were then kept at 40^{o}C before being checked for genomic DNA contamination via PCR of a suitable region of the genome which gives a different product in cDNA and genomic DNA. cDNA preparations without contamination were subsequently stored at -20^{o}C until use, whereupon they were stored at 4^{o}C.

2.4 Polymerase Chain Reaction

I carried out PCRs to determine if primers were producing solely the desired product. PCRs were carried out using a 20µl reaction mix comprised as follows: 15.4µl U Ultrapure DNase/RNase-Free Distilled Water, 2µl 10X PCR Buffer with $MgCl_2$ (#P2192, Sigma-Aldrich), 0.4µl DNA Taq Polymerase (#18038018, Life Technologies), 0.4µl of 10µM forward primer, 0.4µl of 10µM reverse primer, 0.4µl of 10mM dNTP mix, and 1µl of cDNA solution (or Ultrapure water as a negative control). Unmodified salt purified primers were purchased from Sigma-Aldrich and stored as a 100µM solution at -20^{0}C. Working solutions of 10µM primer were kept at 4^{0}C. Ultrapure DNase/RNase-Free Distilled Water (#10977035, ThermoFisher Scientific (Life Technologies)) was used as the solvent. This mixture was subjected to the following PCR protocol, where X is between 55 and 65 depending on the primer pairs:

95^{o}C 1 minute

Then the following three steps repeated 10 times, with X decreasing by 1^{o}C every time;

95^{o}C 20 seconds

X+10^{o}C 30 seconds

72^{o}C 1 minute

Then the following three steps repeated 30 times;

95^{o}C 20 seconds

X^{o}C 30 seconds

72^{o}C 1 minute

Then;

72^{o}C 10 minutes

PCR products were then subjected to DNA electrophoresis for purposes of examination.

2.5 DNA Electrophoresis

I carried out DNA Electrophoresis. DNA electrophoresis was used to assess the size of DNA fragments generated by PCR. Gels were created using LMP agarose, diluted in TBE Buffer to a concentration of 2%. After melting and casting of the agarose gel in a mould, it was left to solidify at room temperature. Subsequently, 5-10µl (depending on well width) of DNA loading solution was added to each well. This solution consisted of 9 parts PCR product to 1 part DNA Loading Buffer. 5-10µl of DNA marker (1 Kb Plus DNA Ladder, Invitrogen) was also pipetted into a well. The gel was then placed in an electrophoresis tank (Bioscience Service) and completely submerged in TBE buffer. A 100 volt current was applied and the gel was visualised after 1 hour, with subsequent visualisations if necessary. The DNA fragments in the gel were visualised by UV light illumination using an Uvidoc Lightbox (Uvitec) and photographed with the built-in camera.

If the primers were intended to be used for qPCR, PCR products were subjected to DNA electrophoresis for purposes of examination. If the product size matched the expected size, and was the only product present, the reaction was deemed “clean” and the product was sequenced to ensure primer specificity.

2.6 Sequencing of Polymerase Chain Reaction products

I prepared products for sequencing. Clean PCR products were subjected to sequencing. The following reagents were added to the wells of a 96-well plate; 1µl of PCR product, 1µl of ExoSapIT, and 3µl of Ultrapure DNase/RNase-Free Distilled Water (#10977035, ThermoFisher Scientific (Life Technologies). These were then subjected to the following protocol:

37°C 60 minutes

80°C 20 minutes

Subsequently, the following was added to each well; 1µl BigDye, 1µl BigDye x5 sequencing buffer, 1µl of 3.2µM primer solution, and 2µl Ultrapure DNase/RNase-Free Distilled Water (#10977035, ThermoFisher Scientific (Life Technologies)).

Unmodified salt purified primers were purchased from Sigma-Aldrich and stored as a 100µM solution at -20oC. Working solutions of 10µM primer were kept at 4oC. Ultrapure DNase/RNase-Free Distilled Water (#10977035, ThermoFisher Scientific (Life Technologies)) was used as the solvent. 3.2µM primer solutions were generated from 10µM solutions and the solvent was the sane. The wells were then subjected to the following protocol:

96°C 1 minute

Then the following three steps repeated 30 times:

96°C 10 seconds

50°C 5 seconds

60°C 4 minutes

Then each well had the following added, in the order in which they are listed; 2.5µl of 125mM EDTAs solution, 30µl of 95% ethanol solution. Then the plate was sealed and the following protocol was carried out:

Inversion of plate 4 times

Incubation at room temperature for 15 minutes

Centrifuge 30 minutes at 3000rpm, at 8oC

Removal of excess ethanol by inversion of opened plate onto absorbent tissue

Addition of 30µl of 70% ethanol solution

Centrifuge 15 minutes at 3000rpm, at 8oC

Removal of excess ethanol by inversion of opened plate onto absorbent tissue

Open air drying of plate for 3-5 minutes, then resealing

Sequencing on a 3730 Genetic Analyzer (Applied Biosystems). This was carried out by appropriately trained staff at the IGMM Sequencing Facility, chiefly Stephen Brown. https://www.ed.ac.uk/igmm/facilities/dna-sequencing-facility.

Ultrapure DNase/RNase-Free Distilled Water (#10977035, ThermoFisher Scientific (Life Technologies)) was used as the solvent in all cases.

Primers which did not produce a single, intended sequence were discarded for the purposes of qPCR.

2.7 Quantitative Polymerase Chain Reaction

This method is as in Malavasi *et al.* 2018[70]. qPCRs were carried out both by me and by Helen S. Torrance, and the section describing primers indicates gene by gene contributions by Helen S. Torrance. Note that housekeeping gene stability was assessed as described in Malavasi *et al.* and I did not contribute to the selection or quantification of housekeeping genes. I contributed solely to the selection and quantification of genes of interest.

Non-template and minus reverse transcriptase controls were included in all experiments with three technical replicates for all samples. To control for inter-plate

variation a calibrator sample was included on every plate for normalisation purposes. Melting curve analysis was carried out for each primer pair to optimise amplification conditions and confirm amplification specificity. Specificity was also confirmed by PCR and sequencing (2.4 and 2.6). Melting curves of significant genes and housekeeping genes are provided in the Appendix, 9.2 and 9.3. PCR efficiency was assessed by running standard curves using serial dilutions of NPC or mouse brain samples for human and mouse primers, respectively. All samples were run in triplicate, and replications with more than 1 CT difference were discarded. Gene expression levels were calculated using the relative standard curve method; with normalisation to the geometric mean of the reference housekeeping genes (see 2.8 and 2.9). They were then averaged for each set of replicates to give the expression score for that sample.

From Malavasi *et al.* and relating to housekeeping genes which I did not contribute towards identifying[70]; For quantification of human gene expression, housekeeping gene stability was assessed across samples taken from NPCs through to five week neurons for several genes using geNorm (genorm.cmgg.be/). *ACTB* and *GAPDH* were subsequently selected as the most stable housekeeping genes for use in quantitative RT-PCR in these samples. *ACTB* was used as a reference gene for human iPSC-described neuron RNA-Seq follow up. For quantification of mouse gene expression, Cyclophilin (*Ppib*) and *Hmbs* were found to be stable in the mouse brain samples analysed.

A typical qRT-PCR was as follows and closely follows that recommended by the manufacturer. Quantitative Polymerase Chain Reactions were carried out in triplicate in a 384-well plate. Each well was loaded with 4.8µl Power Sybre Green PCR Master Mix (#4367659, ThermoFisher Scientific (Life Technologies)), 0.6µl 10mM forward primer, and 0.6µl 10mM reverse primer, as well as 4µl cDNA sample. Non-template control replaced cDNA with Ultrapure DNase/RNase-Free Distilled Water (#10977035, ThermoFisher Scientific (Life Technologies). Unmodified salt purified primers were purchased from Sigma-Aldrich and stored as a 100µM solution at -20^{o}C. Working solutions of 10µM primer were kept at 4^{o}C. Ultrapure DNase/RNase-Free Distilled Water (#10977035, ThermoFisher Scientific (Life Technologies)) was

used as the solvent. The plate was spun on a centrifuge for at least 3 minutes to force reagents to the ends of the wells.

The plate was then run on a 7900HT Fast Real-Time PCR System with 384-Well Block Module (ThermoFisher Scientific (Applied Biosystems)) using the following protocol;

50°C 5mins

95°C 10 mins

Then the following two steps repeated 40 times, with X varying between 57 and 63 depending on primer pair;

95°C 15 sec

X°C 45 sec

Dissociation curve step;

95°C 15 sec

X°C 15 sec

95°C 15 sec

QPCR products were occasionally subjected to DNA electrophoresis for purposes of examination.

2.8 Human primers used in this thesis

All primers were designed using UCSC hg38, at https://genome.ucsc.edu/, and were then checked for specificity using BLAST, at https://blast.ncbi.nlm.nih.gov/Blast.cgi. Primers which had a product greater than 250bp, or an offsite potential product less than 1kb were discarded. All products were examined by DNA electrophoresis to ensure correct product size and sequenced to ensure specificity. All primers span an exon-exon boundary containing an intron of at least 1kb in size. Note that *DRD2*,

ACTB, and *GAPDH* primer design was carried out by Helen S.Torrance. *ACTB* was selected as a housekeeping gene due to its stable expression; selection of this gene and other details are described in Malavasi *et al.* 2018 and I did not carry out the corresponding PCRs[70].

Gene	**Forward**	**Reverse**
ACTB	GTTACAGGAAGTCCCTTGCCATCC	CACCTCCCCTGTGTGGACTTGGG
BBS1	TGAAACTCAATGTGCCCCGA	GTAGGCGCAGCAGGTATAGG
CALB1	AATTTCCTGCTGCTCTTCCGA	TCTATGAAGCCACTGTGGTCAG
CHRNA4	GGCCGAAGACACAGACTTCTC	AGCAGGCAGACGATGATGAAC
DRD2	AAGGGCACGTAGAAGGAGAC	GGTCACCGTCATGATCTCCA
GAPDH	GAGTCCACTGGCGTCTTCAC	ATGACGAACATGGGGGCATC
GPC1	CCATGCTTGCCACCCAG	GCTCTGAGTACAGGTCCCG
HAP1	GCCCCTAAGCTGATTTCGCA	AGAGTGTCGAGTTGAGAGGC
HIF1A	TTTTGGCAGCAACGACACAG	GTGCAGGGTCAGCACTACTT
KANSL1	GTTACAGCCAGCACATCGC	AGACTGAAGATCAACCTCCCG
METRN	GCGACTTCGTAATTCACGGG	TGGGGTACGAATGGAGGTCA
NRP2	GGTGGACCCCTCAACAAAGC	GCCGGTACACCATCCAGTC
NTRK2	CTGGCCTGGAATTGACGATGG	CGAGAGATGTTCCCGACCG
PDYN	TGTAAAGACCCAGGATGGTCC	AGTCCTCCTTGTCATTGAGCC
QKI	CTAATCACTGTGGAAGATGCTCA	TTCTTCAGGCTGTCTTCTCCTT

Exon	Forward	Reverse	Associated isoform accession number	Location in UCSC hg19
DLG2	TGCATGTTACT GTGCACTCCG	GGGCCTCTTACT TCGTGGGT	Isoforms 2 and 6, NM_001364.3 and NM_001300983.1	-chr11:84843812-84844167
DPYSL2	CTTCCCGGCAGT TTTTGCCT	ATATGTCTGCAT AGAACGACTGGT	NM_001197293	+chr8:26371791-26372195
DPYSL3	TAGAGATCCGG AGCGCCACC	TCATTGACGATT CTGCCTCCCT	NM_001197294.1	-chr5:146889041-146889098
DVL1	CACCAGCTCCTC CTCACTAACC	GGCGCTCATGTC ACTCTTCAC	NM_001330311.2	-chr1:1274962-1275029
GRIA4	CAACTCTTGGCA ATGACACAGC	TTAGGAATGGTC GAACAGCG	NM_001112812.1 and NM_001077244.1	+chr11:105782602-105783836
NTRK2	TTCTGCTTAAGT TGGCAAGACAC	GCACTTCCCGGG ATAAGCCA	Isoforms b and f, NM_001007097, and NM_001291937.2	+chr9:87425457-87430617
NTRK3	ATGAGGAACCT GAGGTCCAG	AAAAGCCATGAC GTCCTTTG	NM_001007156 and NM_001320134.1	-chr15:88520598-88520822
SHTN1	TTCGAAAGGCT GCGAAAGTG	CCAACACTGGCA TGGATTTGG	See explanation below	-chr10:118661276-118661468
SLC12A2	CACAAGAGAAA TCTCCTGGCACC	TTGAGTTGCAGT CTTGCCATCC	Transcript variant 1, NM_001046.2 and NR_046207.1	+chr5:127512797-127512844

Note that *DVL1*'s exon matches a second isoform which is poorly annotated and has a retained intron. It has neither an associated mRNA nor an EST. The primers of

SHTN1 were designed to detect an exon found in multiple isoforms. In humans the exon is found in several isoforms distinguished by altered C and N terminals, and is only absent from one. See section 3.9.9 for more details on the properties of the analogous gene in rodents. NM_001127211, NM_001258298, NM_001258299, NM_001258300 are the isoforms listed in NCBI which contain the exon. Locations are as given by DEXSeq, with +/- indicating the strand as in UCSC. Note that the exons are visually identified in a series of images displayed in theAppendix 9.1.

2.9 Mouse primers used in this thesis

All primers were designed using UCSC mm10, at https://genome.ucsc.edu/ and were then checked for specificity using BLAST, at https://blast.ncbi.nlm.nih.gov/Blast.cgi. Primers which had a product greater than 250bp, or an offsite potential product less than 1kb were discarded. All primers span an exon-exon boundary containing an intron of at least 1kb in size. All products were examined by DNA electrophoresis to ensure correct product size and sequenced to ensure specificity. Note that *Hmbs* and *Ppib* (cyclophilin) primer design was carried out by Helen S. Torrance. These were selected as housekeeping genes due to stable expression; details are described in Malavasi *et al.* 2018 and I did not carry out the corresponding PCRs[70].

Gene	Forward	Reverse
Apoe	GTTGGTCACATTGCTGACAGG	CCAGCGCAGGTAATCCCAG
Arc	GGTGAGCTGAAGCCACAAATG	ACTTCTCAGCAGCCTTGAGAC
Avp	CACAGTGCCCACCTATGCTCG	TTGGTCCGAAGCAGCGTCC
Hap1	CTAAGGCTGAGACAGCGCAC	ATAGCCTTCCAGCCTCAACAC
Hmbs	CCCTGAAGGATGTGCCTACCATA	AAGGTTTCCAGGGTCTTTCCAA
Metrn	CCTTCCGTTTTGAACTGCACG	AGCTCGGCATCACTGC
Mt2	CCTGCAAATGCAAACAATGCAA	TCGGAAGCCTCTTTGCAGAT
Nrp2	AGTGAGAAGCCAGCAAGATCC	GTTCGGGGGCGTAGACAATC
Ppib	GGAGATGGCACAGGAGGAAAG	GCCCGTAGTGCTTCAGCTTGAA
Slc1a1	ATGATCTCGTCCAGTTCGGC	GGTCCAAGCCATTCAGTTGC

2.10 Harvesting of RNA from iPSC-derived neurons

I did not carry out the culture or harvesting of the cells used to generate RNA but have exactly replicated the relevant text written by collaborators. I claim no credit in the production, harvesting, RNA-Sequencing, or initial data processing of the samples, which was the work of the following individuals; Kirsty Millar, Helen S. Torrance, Marion Bonneau, Susan Anderson, Daniel McCartney, Philippe Gautier, and the Wellcome Trust Edinburgh Clinical Research Facility.

2.10.1 Culture and maintenance of iPSC-derived neurons

High density neuronal cultures (approximately 2 million cells per well in 12 well plates) were differentiated for 5 weeks. Immunofluorescence staining of parallel cultures was used to confirm correct NPC morphology and Nestin expression at the time of plating, and successful neuronal differentiation by assessing morphology and acquisition of βIII-tubulin expression at the time of harvesting, for every culture used. Three independent neuronal differentiations were performed per NPC line. Neurons were harvested in RNAlater (ThermoFisher Scientific), stored at -80^{o}C, then processed in batches to extract the RNA. Each batch consisted of one triplicate per line to minimise batch effects.

2.10.2 Processing of RNA samples for RNA sequencing

All subsequent steps were performed by the Wellcome Trust Edinburgh Clinical Research Facility (www.wtcrf.ed.ac.uk) . Total RNA samples were assessed on the Agilent Bioanalyser (Agilent Technologies, G2939AA) with the RNA 6000 Nano Kit (5067-1511) for quality and integrity of total RNA, and then quantified using the Qubit 2.0 Fluorometer (Thermo Fisher Scientific Inc, Q32866) and the Qubit RNA BR assay kit (Q10210). Samples were also assessed for DNA contamination using the Qubit DNA HS assay Kit (catalogue Q32851).

Libraries were prepared from each total-RNA sample using the TruSeq Stranded Total RNA with Ribo-Zero Gold kit (RS-122-2301) according to the provided protocol.

500ng of total-RNA was processed to deplete rRNA before being purified, fragmented and primed with random hexamers. Primed RNA fragments were reverse transcribed into first strand cDNA using reverse transcriptase and random primers. RNA templates were removed and a replacement strand synthesised incorporating dUTP in place of dTTP to generate double-stranded cDNA. AMPure XP beads (Beckman Coulter, A63881) were then used to separate the double-stranded cDNA from the second strand reaction mix, providing blunt-ended cDNA. A single 'A' nucleotide was added to the 3' ends of the blunt fragments to prevent them from ligating to another during the subsequent adapter ligation reaction, and a corresponding single 'T' nucleotide on the 3' end of the adapter provided a complementary overhang for ligating the adapter to the fragment. Multiple indexing adapters were then ligated to the ends of the double-stranded cDNA to prepare them for hybridisation onto a flow cell, before 15 cycles of PCR were used to selectively enrich those DNA fragments that had adapter molecules on both ends and amplify the amount of DNA in the library suitable for sequencing.

Libraries were quantified by PCR using the Kapa Universal Illumina Library Quantification kit complete kit (KK4824) and assessed for quality using the Agilent Bioanalyser with the DNA HS Kit (5067-4626). Libraries were combined in three equimolar pools and sequencing was performed using the NextSeq 500/550 High-Output v2 (150 cycle) Kit (FC-404-2002) on the NextSeq 550 platform (Illumina Inc. SY-415-1002). Sequences were aligned to the human reference genome Hg19 using the RNA-Seq Alignment v1.0 application (Illumina Inc.).

2.11 Harvesting of RNA from mouse brain regions

I did not carry out the harvesting of the mouse RNA but have exactly replicated the relevant text of the thesis of Marion Bonneau, a collaborator and fellow student, here to illustrate the sample preparation. I claim no credit in the production or harvesting of the samples. I have also replicated text from Malavasi *et al* in 2.11.1[70]. 2.11.2 and

2.11.3 are replicated exactly from Bonneau's thesis with the exception of formatting. 2.11.4 was carried out by Yasmin Singh (CeGaT GmbH, Paul-Ehrlich-Straße 23, Tübingen, Germany).

2.11.1 Mouse colony production and maintenance

I did not carry out this protocol and the text is replicated from Malavasi *et al.* (2018).

VelociMouse® technology (Regeneron) was used to target embryonic stem cells and microinject them into mouse embryos[144]. In brief, F1H4 (129S6SvEv/C57BL6F1) embryonic stem cells were electroporated with the linearized vector construct and positive clones were microinjected into 8-cell stage mouse C57BL6 embryos. Microinjected embryos were transferred to uteri of pseudopregnant recipient females, weaned pups were scored, and high percentage chimera males were selected for mating with flp-positive C57BL6 females to remove the selection cassette, to prove germ-line transmission, and to generate F1 animals for further breeding.

Because there is already a mutation (25bp deletion) at the Disc1 allele in exon 6 in the 129/Sv strain which causes a truncation of Disc15, F1 progeny were generated and a PCR assay which distinguishes the C57BL/6 allele versus the 129/Sv allele was employed to determine which F0 mice were correctly targeted to the C57BL/6 locus (Supplementary Figure 10). Mice which carried the translocation on the C57BL/6 allele were then crossed to CMV-Cre mice to remove the Neo cassette via Cre-mediated recombination at the flanking loxP sites. Genotyping results were confirmed by Loss-of-Native-Allele assay.

The exclusion of the differentially spliced DISC1FP1 exon 3a6 that is present in a minority of transcripts (www.genome.ucsc.edu) precludes production of transcripts encoding CP1. The exclusion of the differentially spliced DISC1FP1 exon 7b does not affect the potential production of CP60/69 proteins since the stop codon in chimeric transcripts encoding these proteins occurs in exon 66. Since the Disc1 allele was modified on a mixed background of 129 and C57BL/6J, a congenic breeding strategy was adopted to purify the strain background. Following repeated crossing to C57BL/6J mice, genotyping of polymorphic markers carried out by the Jackson

Laboratory found the mice to be >99.5% C57BL/6J. These mice were then mated to C57BL/6J for one final round and the progeny used for subsequent experiments. Mice were housed in the Biomedical Research Facility at the University of Edinburgh. All mice were maintained in accordance with Home Office regulations, and all protocols were approved by the local ethics committee of the University of Edinburgh.

2.11.2 Collection of tissue

The below text is replicated from the thesis of Marion Bonneau (2018), and I did not participate in the collection of the tissue.

The tissue was collected from 9 weeks old mice. Each group (wild type, heterozygotes, homozygotes) were composed of 4 males and 4 females. Mice were culled under the schedule 1 procedure by trained staff at the animal facility. The brains were then directly removed and washed in ice-cold PBS. Hippocampi, and cortices minus hippocampus, cerebellum and olfactory bulbs, were dissected from the right brain hemisphere mice at nine weeks of age. The tissues from the right hemisphere were incubated overnight at 4ºC, in 5 volumes of RNA later (Ambion). After 24h, the RNA later was discarded to prevent the formation of salt crystals and the samples were snap frozen in liquid nitrogen and stored at -80^{o}C, then processed in batches of mixed genotypes to extract the RNA.

2.11.3 RNA preparation from tissue samples

The below text is replicated from the thesis of Marion Bonneau (2018), and I did not participate in the preperation of the RNA.

The samples were purified using QIAGEN RNA extraction kit, according to the manufacturer's instructions. To homogenise the tissues, the Tissueruptor was used. Then insoluble materials were removed and nucleoprotein complex dissociated. The RNA was re-dissolved in 100 µl of RNA free water. At that point, the RNA concentration was assessed using the nanodrop for a first time to assess its quality. To obtain the purest RNA possible, RNA clean-up was performed, according to the manufacturer's instructions. Additional On-column DNase digestion was then done

by treating the samples with DNase I, as indicated by the manufacturer's instructions. At the end, the RNA was eluted once for the hippocampal samples and twice for the cortical samples using 30 µl of RNAse free water, therefore 60 µl of pure cortical RNA and 30 µl of pure hippocampal RNA were obtained.

2.11.4 Sequencing and initial processing of mouse RNA samples

The below text is replicated from the thesis of Marion Bonneau (2018) and I did not contribute.

Total RNA samples were assessed with a Fragment Analyser (Agilent) for quality and integrity of total RNA. Libraries were prepared using 100ng of each total RNA sample using the TruSeq Stranded mRNA Library Prep Kit (Illumina). Single end RNA Sequencing was carried out to a depth of approximately 60 to more than 100 million reads. Sequencing was performed on a HighSeq4000 on HighOutput mode. Demultiplexing of the sequencing reads was performed with Illumina CASAVA (1.8.2). Adapters were trimmed with Skewer (version 0.1.116)[145]. Raw reads were mapped to the reference genome mm10 with STAR (Version 2.4.0h)[146]. Further analyses were performed with the Cufflinks Tool Suite (Version 2.1.1)[147,148]: Cufflinks was used to count mapped reads. FPKM values were computed with Cuffdiff using the "pooled-variance" model, "geometric" normalization and "multi-read-correct" option[149]. The quality of fastQ files was analyzed with FASTQC (Version 0.10.1)[150].

Cortical reads were as follows; Total average raw read number was 75.39 ±11.9 $x10^6$ for WT, 89.2 ±6.3 $x10^6$ for heterozygotes, and 98.7 ±17.3 $x10^6$ for homozygotes.

Hippocampal reads were as follows; Total average raw read number was 83.77 ±13.58 $x10^6$ for WT, 82.8 ±13.8 $x10^6$ for heterozygotes, and 79.9 ±14.08 $x10^6$ for homozygotes.

2.12 Bioinformatics

The version of R used varied as updates were released, but was between 3.4. & 3.5.1. R Studio versions used varied from to 1.0.153 to 1.1423. A number of different R packages were also utilised and details of particular packages are mentioned below.

2.12.1 Gene and exon analysis

Philippe Gautier at the IGMM carried out the analyses looking at gene differential expression between translocations and controls, or between mice of different *Der1* status. This was done utilising DESeq2 package version 1.2.. This was used with default settings enabled. As described in the main body of the thesis, DESEq2 compares favourably to alternative packages and is effective at detecting genes which have been verified as differentially expressed by qRT-PCR. It was developed by Love *et al.*[151] . When padj is referred to in the text of the thesis, this referes to p-values produced by DESeq2 which have been adjusted according to the Bejamini-Hochberg method. I also utilised DESeq2 to compare all human lines against one another sequentially; the idea here was to produce lists of genes which might be highly reliably differentially expressed.

Translocation status/genotype was utilised as the factor of interest and as recommended non-normalised counts were inputted for each replicate of each cell line, resulting in a 9 vs 9 comparison for the iPSC-derived neuronal cell lines and 6/8 sex-balanced mice per genotype in the mouse analyses. Mouse wild types per compared against heterozygotes for each sample set in turn, then against homozygotes.

Exon level differential expression was assessed by Philippe Gautier using DEXSeq[152].

2.12.2 Deconvolution analysis

Deconvolution of the mouse and human samples is described in detail in the appropriate chapter, particularly the aspects relating to optimisation and troubleshooting. All deconvolution was carried out by me in R using the DeconRNASeq R package version 1.24.0. The accuracy of DeconRNASeq was

assessed by generating pseudosamples from the cell types being utilised. This consisted of generating a series of numbers for each pseudosample, which sum to 1. The length of the series corresponds to the number of pure cell lines which DeconRNASeq is assessing the proportions of. The expression profile of each pure line was then multiplied by its corresponding number, and the resulting profiles were summed to give a single pseudosample. Series were retained so as to compare the predicted proportions according to DeconRNASeq to the actual weightings. If pure cell profiles were removed from the roster (for example, during the Zeisel deconvolutions looking at the removal of Interneuron 5), fresh pseudosamples using the new roster were produced.

Housekeeping gene normalisation was carried out as follows; a geometric mean of the expression of all utilised housekeeping gene was produced for each pseudosample/sample/pure cell line profile, then the profile was divided by that factor. Selection of housekeeping genes is described in the appropriate chapter. All genes except the utilised marker genes were then removed from the profiles prior to deconvolution. The datasets utilised in the deconvolution are described below.

The deconvolution was then carried out with the DeconRNASeq function described in the package "DeconRNASeq" by Gong *et al.* DeconRNASeq implements an nonnegative decomposition by quadratic programming, and the function was used in R as below;

DeconRNASeq(datasets, signatures, proportions = NULL, checksig = FALSE, known.prop = FALSE, use.scale = TRUE, fig = TRUE)

datasets were the housekeeping normalised RNA-Seq samples, signatures were the housekeeping normalised expressions of the pure cell types. These two datasets were filtered so as to only contain the marker genes. The other options mostly relate to illustrating the findings.

The "Zhang dataset" is described in Zhang *et al.*[153]. This is deposited in the Gene Expression Omnibus (GEO) database, www.ncbi.nlm.nih.gov/geo under accession no. GSE52564. For each pure cell profile the two provided profiles of that cell type

were averaged. Marker genes for the Zhang deconvolution were obtained from https://web.stanford.edu/group/barres_lab/brain_rnaseq.html which allows selection of genes enriched in one of the cell types versus a selection of other cell types. Markers were selected so as to be enriched in one pure cell profile versus all others utilised in the deconvolution.

The comparison datasets in the Zhang deconvolution are as follows;

The Zhang Two dataset is described in Zhang *et al*. and the data are deposited in the NCBI Gene Expression Omnibus (GEO) database, www.ncbi.nlm.nih.gov/geo under accession no. GSE73721[154].

The Darmanis dataset is described in Darmanis *et al.* and the data are deposited in the NCBI Gene Expression Omnibus (GEO) database, www.ncbi.nlm.nih.gov/geo under accession no. GSE67835[155].

The Dorsal Root Ganglion sensory neuron dataset is described in Li *et al.* and the data are deposited in the NCBI Gene Expression Omnibus (GEO) database, www.ncbi.nlm.nih.gov/geo under accession no. GSE63576[156].

The Allen Brain Atlas datasets are described in detail in their white paper as well as at the web address http://celltypes.brain-map.org/rnaseq (accessed on 16/10/2018), where they are available to download. The human dataset I utilised is comprised of single nucleus RNA-Seq of the middle temporal gyrus and is available in CPM form.

The "Zeisel dataset" is described in Zeisel *et al.* and is available at http://linnarssonlab.org/cortex/ [157]. The cortical samples were extracted and for each pure cell profile all provided profiles of that cell type were averaged. Markers were selected so as to be enriched in one pure cell profile versus all others utilised in the deconvolution, this was achieved by producing an SI value for each, a measure of the proportion of expression across all profiles attributable to that profile. Different SI values were experimented with for efficient pseudosample deconvolution.

The comparison datasets in the Zeisel deconvolution are as follows;

The Allen Brain Atlas datasets are described in detail in their white paper as well as at the web address http://celltypes.brain-map.org/rnaseq (accessed on 16/10/2018), where they are available to download. The human dataset I utilised is comprised of single nucleus RNA-Seq of the middle temporal gyrus and is available in CPM form. I also utilised the mouse primary visual cortex dataset for the mouse deconvolutions.

2.12.3 EWCE analysis

EWCE package version 1.2 was utilised, following the manual and script provided by the author at https://github.com/NathanSkene/EWCE/. These analyses were carried out by me. Two separate datasets were utilised, the first being the Zeisel dataset mentioned above, and the second being the Karolinska Institute (KI) superset. These are described in detail in Chapters 6 and 7, respectively. EWCE essentially quantifies the "enrichment", represented by gene specificity, in a target list for all queried cell types. It then compares these to the enrichments of a large number of lists, each of the same length as the query list but comprised of genes randomly selected from the list of background expressed genes.

The "Zeisel dataset" is described in Zeisel *et al.* and is available at http://linnarssonlab.org/cortex/[157].There was some initially some difficulty in utilising the package effectively, and the Zeisel dataset was loaded from local storage (originally obtained from http://linnarssonlab.org/cortex) rather than using that portion of the script. Reading in the Zeisel data manually must be done with care. The function read_celltype_data will only operate on a dataset which is identical in dimensions to the Zeisel cortical dataset described in the manual. To read in the hippocampal data, false genes and samples were manually added to the hippocampal data until it was the same dimensions as the cortical dataset. After this, the false genes and samples are removed along with their corresponding data, and SI values are subsequently calculated by the instruction in the provided script. The cortical and hippocampal datasets were retrieved as above.

The KI superset data is described in Skene *et al.* 2018[140] and data are available at http://www.hjerling-leffler-lab.org/data/scz_singlecell/. The KI superset was loaded

as an Rda file into R. Embryonic cell types were then removed, and SI values were then recalculated by dividing by the sum of SI values for that gene.

EWCE was carried out as follows. In the case of human gene list to mouse dataset comparisons genes were filtered for 1:1 homology using the list provided in the EWCE package. Query lists were the same as those utilised in the GO term analysis; utilising adjusted p value<0.05, and BaseMean at least half that of *DISC1/Disc1*. The background lists utilised were the remaining genes expressed in each utilised dataset.A total of 100,000 background lists of equal length to the query list were randomly selected and used to produce a distributuion of enrichment to which the query list could be compared for each cell type. The p values were subjected to Bonferroni corrections for multiple testing at both the class and subclass level.

2.12.4 Data visualisation

Heatmaps of gene expression were generated by me using R (version 3.4.2) and RStudio (version 1.0.143). Raw count data for all samples were together subjected to a regularised logarithm transformation using the DESeq2 package version 1.16.1. For each heat map, the transformed counts for each gene were normalised to Z-scores across all samples and subsequently visualised using the pheatmap package version 1.0.8 (cran.r-project.org/package=pheatmap).

GOrilla results were displayed using Microsoft PowerPoint. Deconvolution pseudosample results were displayed using R's plot function. qPCR results and deconvolution results were displayed using GraphPad Prism 6. EWCE plots were generated using R's "ggplot2". Details as to numbers of genes and scales are displayed in or under each image.

Volcano plots were produced in R using the "with" and "points" functions.

2.12.5 Prism

Figure legends detail the exact statistical test used in each analysis.

Multiple testing was corrected for using the "Sidak-Bonferroni" option in t-tests, which is described at the following link, accessed on 31/7/19.

https://www.graphpad.com/guides/prism/6/statistics/index.htm?stat_the_method_of_bonferroni.htm

Multiple testing was corrected for using the “Sidak-Bonferroni” option in ANOVAs, which is described at the following link accessed on 31/7/19.
https://www.graphpad.com/guides/prism/6/statistics/index.htm?stat_the_methods_of_tukey_and_dunne.htm

3 GENERATION AND INITIAL ANALYSIS OF HUMAN RNA-SEQ DATA

3.1 Introduction

RNA-Seq allows insights into differential gene expression at both the gene and exon level, which can implicate particular isoforms of a gene. Differentially expressed genes of particular interest include those which link to known pathological processes, whether this link is by pedigree studies, gene ontology, or GWAS/CNV predisposing to psychiatric illness. Candidate genes can then be examined and the differential expression verified by quantitative PCR. The next two chapters describe this process for the human iPSC-derived neuron, mouse heterozygous/homozygous *Der1* cortex and hippocampus RNA-Seq samples. I also hypothesised that there might be some regional effects on transcription around the breakpoints in the t(1;11) neurons. It has previously been shown that there is differential methylation around the breakpoints in the blood of t(1;11) carriers[126]. Chromosomal translocations are also known to cause some disturbances in local expression, although these do not affect the entirety of the chromosome. This is possibly as only cis-acting elements are directly disturbed. Harewood *et al.* showed that a translocation resulted in local expression disturbances, as well as an altered cellular location of the chromosomes[158], and I therefore hypothesised that there might be some similar effects caused by the t(1;11). Finally, the effects of the translocation on *DISC1* can also be examined by RNA-Seq.

3.2 Pedigree and neuron generation

The pedigree and effects of the translocation are described in greater detail in the introduction. A subset of the t(1;11) Scottish pedigree is shown in Figure 3 indicating prevalence of psychiatric illness. Fibroblasts from three individuals carrying the translocation and fibroblasts from three individuals without it were previously used to produce iPSCs. Figure 4 illustrates the familial relations between these individuals.

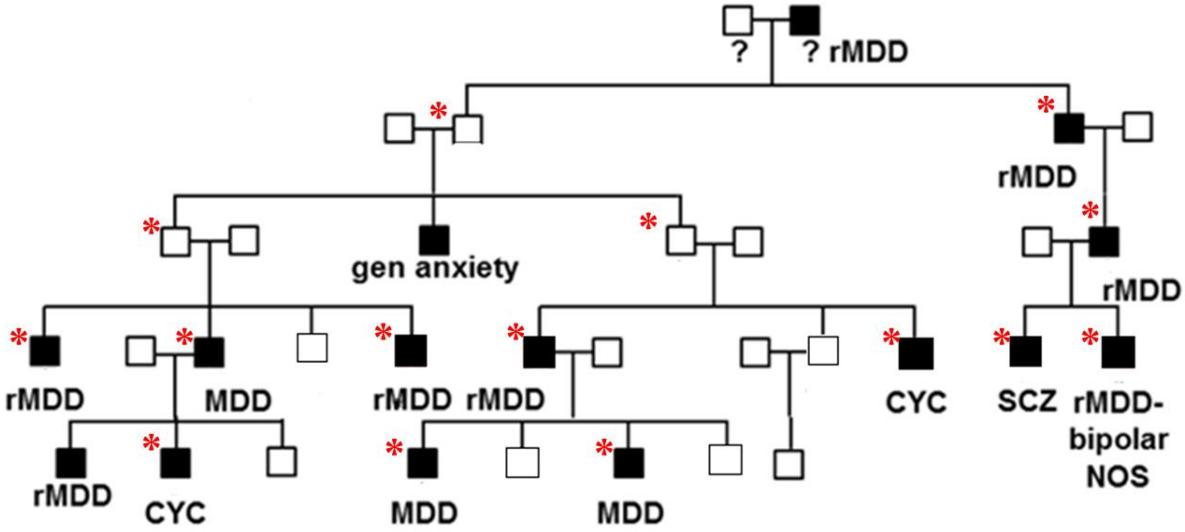


Figure 3. Subset of the Scottish pedigree carrying the t(1;11) translocation. Individuals with a psychiatric diagnosis are in solid black, those without one are in white. Individuals with a red asterisk carry the translocation. Diagnoses are given by acronyms; rMDD=recurrent major depressive disorder, CYC=cyclothymia, SCZ=schizophrenia, bipolar NOS=bipolar disorder not otherwise specified, gen anxiety=generalised anxiety disorder, ? =translocation status unknown

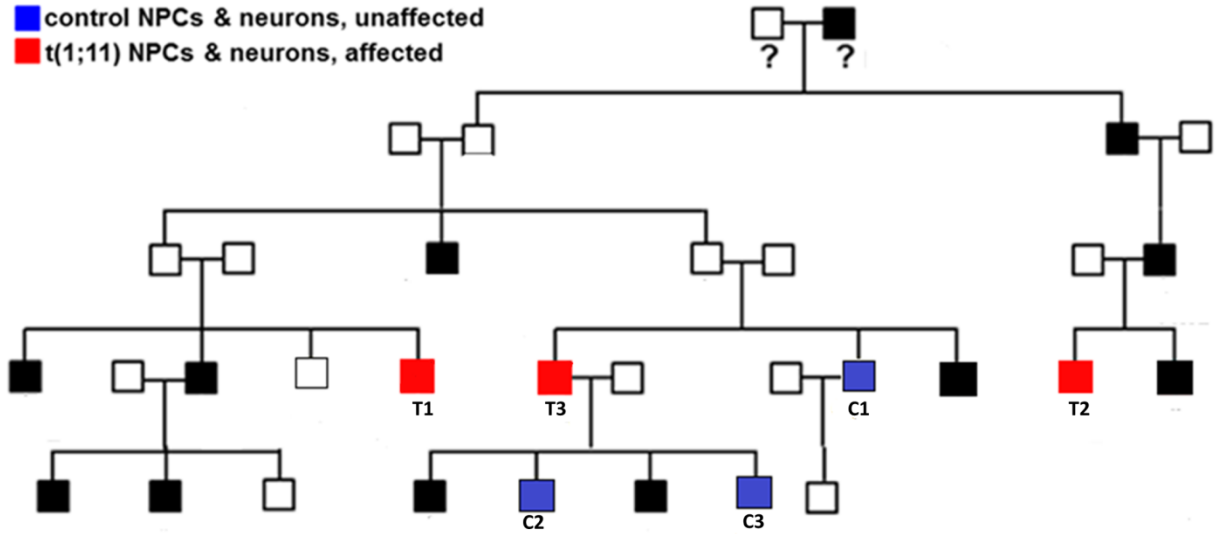


Figure 4. Subset of Scottish pedigree with illustration of lines selected for iPSC, NPC, and neuron generation. Names of each cell line are given below the individuals they are derived from. Red indicates translocation lines (T), blue indicates controls (C).

IPSCs were differentiated to neural precursor cells (NPCs). Each of these NPC lines was subsequently cultured in triplicate and differentiated to neurons in independent experiments. These neurons were harvested for RNA, which was sequenced, quality controlled, and aligned to the human genome hg19 commercially. Further details are

in the Materials and Methods where the appropriate contributors to each section are described. I did not contribute to the stages prior to RNA-Seq data analysis.

3.2.1 Effects of sex on iPSC-derived neuronal expression

There is a clear issue with sex imbalances between the three translocation and three control pedigree members selected for iPSC-derived neuron generation. Lines C1, C2, C3, and T3 were derived from female individuals, while lines T1 and T2 were derived from males. Translocation status is therefore highly associated with sex and the effects of the two may be difficult to distinguish. I attempted to resolve this complication by searching for studies of genes known to be differentially expressed between iPSC-derived neurons of individuals of different sexes. I did not find any such studies. However, after the submission of this thesis a study by Tiihonen *et al.* was published which examined this very phenomenon[159]. This study could not have informed my investigation, but it can critically inform my findings. It will be especially useful in identifying possible false positive genes which are altered by sex. Genes which may be affected by sex and which are discussed in the chapter have also been highlighted so as to be clear about any confounding effects; although as this is *post hoc* it can only exclude experiments I did do rather than inform new ones.

Tiihonen *et al.* analysed the iPSC-derived neuronal lines of pairs of monozygotic twins discordant for schizophrenia. Like our samples, these neurons were derived from iPSC-derived NPCs and are described as “cortical neurons”. Five twin pairs (three female pairs, two male pairs) and six unrelated controls (four females, two males) were utilised. This study allowed the comparison of twin pairs (schizophrenia vs schizophrenia risk background), of unaffected twins vs controls (schizophrenia risk background vs no risk background), of affected twins vs controls (schizophrenia vs no risk background), and of male controls vs female controls (sex effects). In addition, Tiihonen *et al.* made comparisons of the twin pairs utilising only males and only females in turn, to identify if the pathology of the disease differs between the sexes. Comparisons of all males vs all females were not carried out. It should be noted the neurons were likely more mature than ours, with an 8-12 week differentiation protocol vs our five week protocol. It should be noted that the

numbers of lines are equivalent to or lower than our 3 vs 3 comparisons, e.g. the male only comparisons are only 2 vs 2. Each of our lines was also differentiated three times; similar replications were not reported by Tiihonen *et al.*

Tiihonen *et al*. state that 12% of genes were differentially expressed between the male and female control samples (a 2 vs 4 comparison), however they only included genes that had a twofold or greater change. Including all genes which meet the standard I used (an adjusted pvalue of <0.05 and an average expression at least half that of *DISC1*) the number is 4,337, or 22%.

360 of these genes are also differentially expressed in our study; however in the majority of these genes the putative sex effect is in the *opposite* direction to the putative translocation effect and sex is therefore unlikely to be a factor in the significance of the gene. However it is possible that sex could affect these genes to change in one direction in the Tiihonen *et al.* study and in the other direction in my study. This cannot be ruled out but is difficult to assess the likelihood of. 94 of the genes change in the same direction in both studies and these are highly likely to be significant on the basis of sex as opposed to the translocation. These are 7% of all significant genes. Where any of these genes are discussed they are specifically highlighted as potentially being problematic.

3.3 DESeq2

The first step in analysing the data was to determine what differences existed between the control and translocation carrying neurons. A number of programs have been developed to analyse differential gene expression, including DESeq2. DESeq2 is a popular and user-friendly tool for which a large amount of literature exists, making it an ideal option as many further analyses and discussions are available. A study carried out by Costa-Silva *et al.* looked at human brain samples with over 400 genes verified by RT-qPCR as differentially expressed. They found that DESeq2 offered a satisfactory balance between low proportion of false positives and detection of true positives. Its “True Positive Rate” (true positives/all positives) was the second highest, its “Specificity” (true negatives/all negatives) the joint highest, and its “Accuracy” (correct predictions/all predictions) the highest of 9 different methods of

determining differential gene expression[160]. Although it would be incorrect to describe DESeq2 as the "best" package for all situations, it appears that it will be among the most satisfactory methods of determining differential gene expression in my samples.

DESeq2 was used to analyse differential expression between all control translocation samples, with all XY genes removed from the analysis prior to running DESeq2.This resulted in the comparison of 21,916 genes in 9 control samples vs 9 translocation samples. It was reported that 1,372 genes were differentially expressed at the whole gene level. On examination of the list, it was noted that the BaseMean (a measure of the mean of the sequence depth normalised counts of all samples) was very low for a number of these genes. *GRM6*, for example, with BaseMean 4.5, has 0 counts in 4 of the 18 samples, and 12 out of 18 have less than 5 counts. At low levels of expression, differences between the translocation and controls could easily be due to minor differences in sequencing rather than true biological differences. There is also an issue of practicality; poorly expressed genes will be very difficult to investigate further via qPCR, western blotting, and other experiments. I therefore restricted the analysis to genes which had a reasonable level of expression. This was defined as being at least approximately half the expression of *DISC1*, a gene which has detectable expression at the transcript and protein level in these cells. The BaseMean of *DISC1* is 21.8 and the threshold for expression was therefore set at 10. The threshold is somewhat arbitrary but has practical relevance considering future experiments, and analyses testing this list of differentially expressed genes compared it to the list of all expressed genes, not those with BaseMean over 10. A total of 1,252 genes had whole gene differential expression and passed the BaseMean threshold. A volcano plot of the genes analysed by DESeq2 is available in Figure 5.

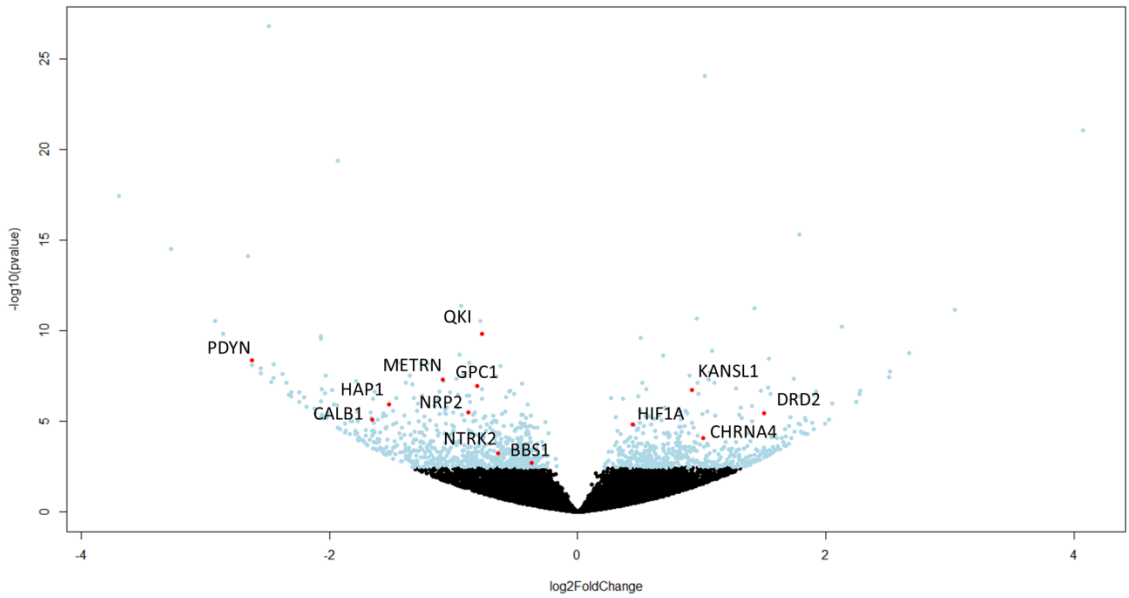


Figure 5. A volcano plot of the DESeq2 for the iPSC-derived neuronal lines for all genes with BaseMean>10. X-axis represents the log2 fold change between WT and translocation lines, while the Y axis represents significance (-log base 10 of p value). Black dots have an adjusted p value above 0.05, blue dots are significant with an adjusted value below 0.05. Red dots with labels represent genes for which a qPCR was carried out.

See Figure 6 for a heatmap of the normalised counts of the samples for all differentially expressed genes. We can clearly see that the samples cluster by line, and then by t(1;11) status. There is some variation in many of the genes but many show a reliable change depending on translocation status.

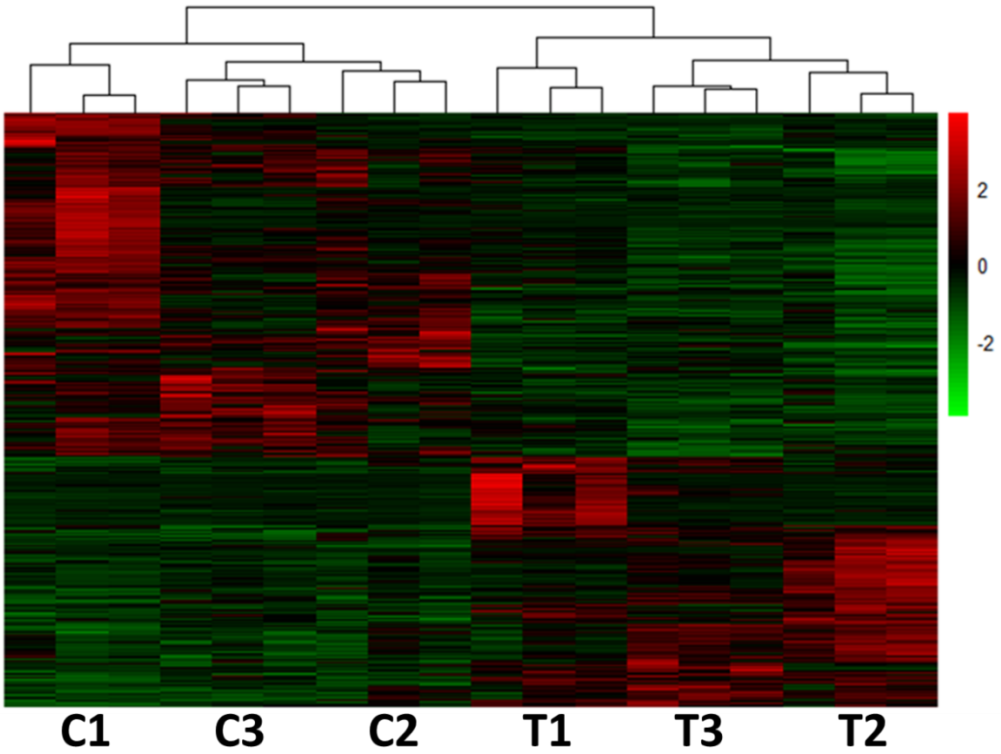


Figure 6. Heatmap of all differentially expressed genes with p<0.05 and BaseMean>10. Counts were normalised using the "rlog" function, which transforms counts to the log2 scale, normalises for library size, and minimises variation in poorly expressed genes. They were then antilogged, and changed to z scores by gene before generation of the heatmap. Red indicates z score above the mean, green indicates z scores below the mean. Each row is a gene.

DISC1 was not a differentially expressed gene. However, transcription of the *DISC1* encoding regions from both the *DISC1-DISC1FP1* and *DISC1FP1-DISC1* loci on chromosomes 1 and 11 could mask effects of the translocation on *DISC1* expression. The only accurate measure of intact *DISC1* RNA is the number of reads which contain exon 8-9, and therefore cross the t(1;11) breakpoint. These reads could only come from the intact *DISC1* allele. The transcript quantification was analysed in detail by Philippe Gautier using DEXSeq[152] and Integrative Genomics Viewer[161], and it was shown that the number of reads which spanned the breakpoints were significantly lower in the translocation carriers. See for a comparison of the reads that cross the breakpoint and therefore could only come from an intact *DISC1* gene. In addition, transcripts which could only have come from the derived chromosomes have been shown to exist in these neurons, and were found when searched for by Philippe Gautier (private correspondence)[70]. Reads from either side of the breakpoints are therefore likely contributing to the apparent non-significance of

DISC1. Differential expression of intact DISC1 protein was also confirmed in these neurons[70]. For the reasons given above, I do not have concerns about the apparent non-significance of *DISC1*.

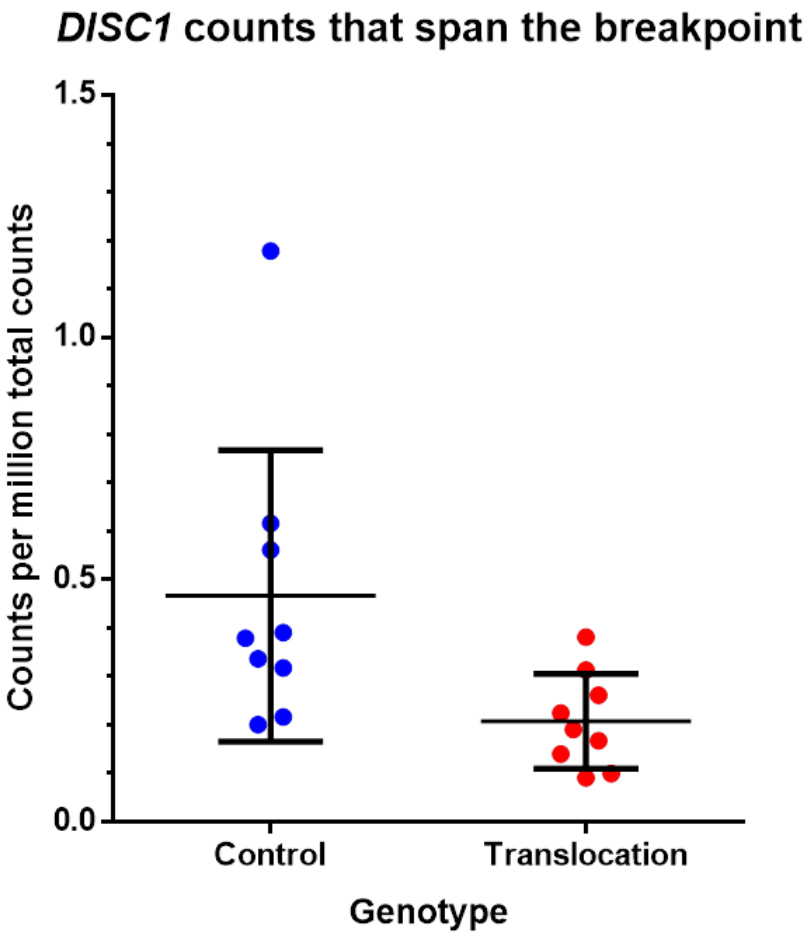


Figure 7. Cross comparison of normalised counts that span the t(1;11) breakpoint and therefore could only have come from an intact chromosome 1. Lines indicate the mean with smaller bars indicating the standard deviation. Counts normalised by dividing total counts spanning the *DISC1* breakpoint by millions of total counts. Total spanning counts: Control=l90, Translocation=81, p=0.033. Total normalised counts: Control=4.19, Translocation=1.86, p=0.0078. P values calculated by Mann-Whitney t test. Counts examined and quantified by Philippe Gautier using DEXSeq and Integrative Genomics Viewer. The control with ~1.2 normalised counts is not an outlier, taking outliers as being more than 3 standard deviations from the mean.

The heatmaps show that the samples cluster by cell identity and subsequently by translocation identity. Philippe Gautier also generated a Principle Component Analysis plot from the DESeq2 data, a modified version of which is displayed in Figure 8. We can see that once again the samples cluster by cell identity.

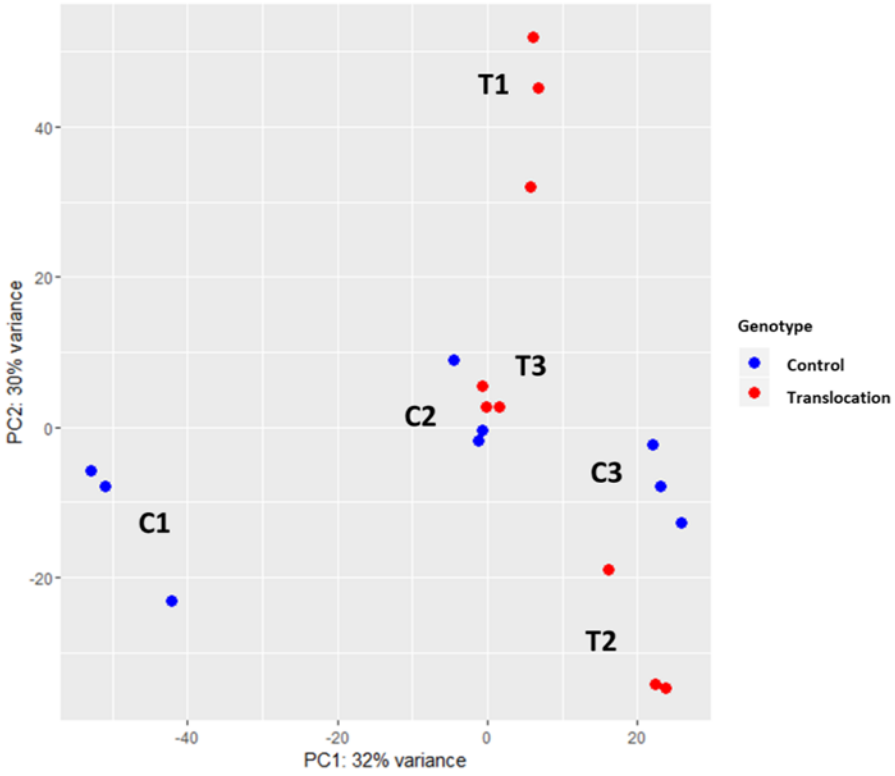


Figure 8. PCA of the translocation (red) and control (blue) lines. We can see that the triplicates cluster together for the most part, with some digressions. We can also see that PC1 is the separating factor for the controls and PC2 is the primary separating factor for the translocations. C2 and T3 are close together; these lines were derived from a mother and daughter. This PCA utilised all differentially expressed genes

In order to inform me as to what genes showed the most reliable differential expression, I also carried out DESeq2 comparing all possible combinations of cell lines in triplicate vs triplicate comparisons. This resulted in 15 comparisons, 9 translocation vs control, 3 control vs control, and 3 translocation vs translocation. See Table 1 for numbers of differentially expressed genes between translocation and control lines. 24 genes were differentially expressed in every translocation vs control comparison and 7 of these were not differentially expressed in any other comparison. These 7 genes were *PDYN*, *RELB*, *NUTM2F*, *MIR4458HG*, *LRRC37A2*, *GPC1*, and *BEST1*. See Figure 9 for a heatmap of the 24 genes. Of course the greatly reduced power of a 3 vs 3 comparison means that these results should only be seen as an aid to selecting genes for further analysis. It is also the case that I am searching for effects of the t(1;11) generally; the control vs translocation cell lines have different genetic backgrounds (some lines are XX and some are XY) and these will be reflected in any 3 vs 3 comparison, even when XY genes are removed.

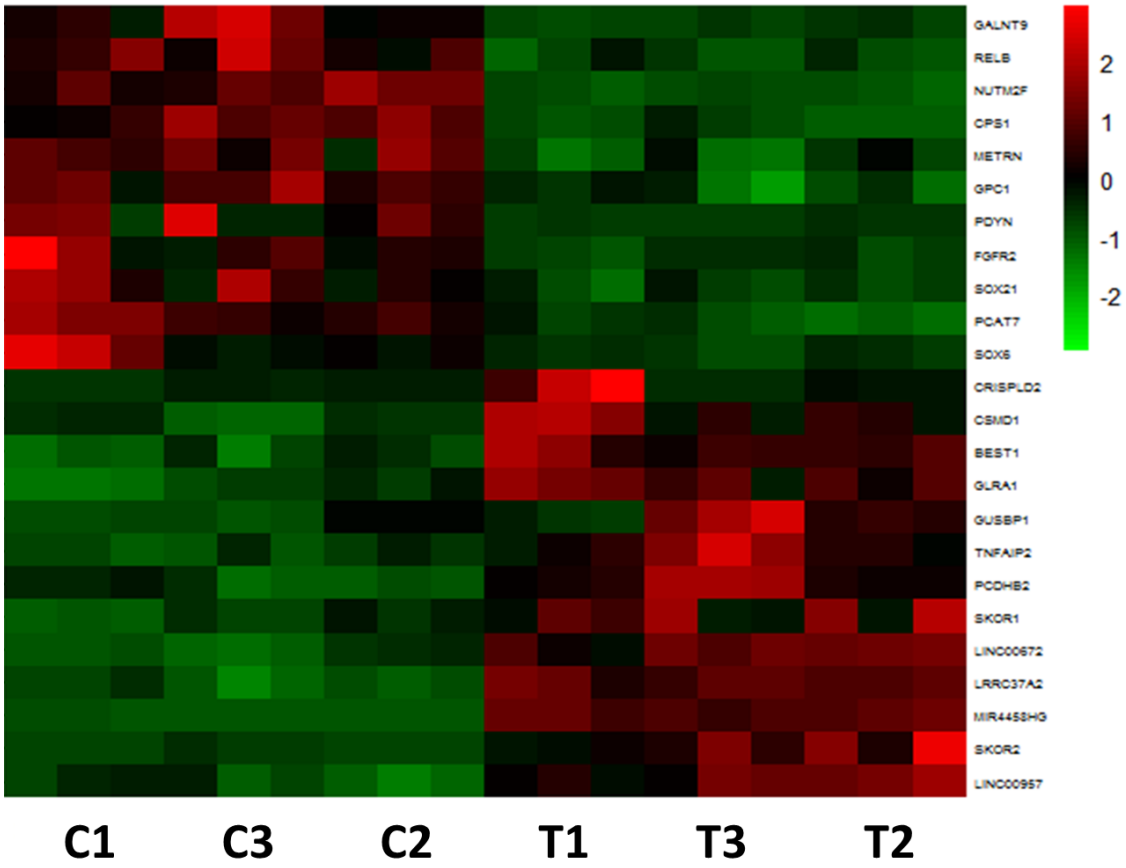


Figure 9. Heatmap of genes that are marked by all control vs translocation DESeq2 analyses as being differentially expressed. Counts were normalised using the "rlog" function, which transforms counts to the log2 scale, normalises for library size, and minimises variation in poorly expressed genes. They were then antilogged, and changed to z scores by gene before generation of the heatmap. Red indicates z score above the mean, green indicates z scores below the mean. Each row is a gene.

Line	C1 vs	C2 vs	C3 vs
T1	4,912	3,995	5,003
T2	8,633	6,942	7,235
T3	6,138	1,835	6,688

Table 1. Number of genes differentially expressed between pairs of triplicates according to DESeq2.

3.4 DEXSeq

DEXSeq analyses data for exon-level differential expression, which can be subsequently inspected to see if this provides evidence for differential expression of certain transcripts[152]. The ideal exon is found in only one transcript or a subset of transcripts, which have a defined biological role. The 9 vs 9 DEXSeq was carried out by Philippe Gautier, and I carried out subsequent analyses.

A total of 2,574 exons were described as differentially expressed at the p=0.1 level, with 1,932 of these unambiguously mapping to one gene. At the p=0.05 level 1,368 exons were differentially expressed with 1,020 exons unambiguously mapping to one gene.

As with the DESeq2 analysis, I carried out DEXSeq analyses comparing all possible combinations of lines to see which exons had the highest level of support for differential expression. At the p=0.05 level, 5 exons, all of which mapped unambiguously to 5 different genes, were significant in all translocation vs control comparisons. Three of these were also not significant in any control vs control or translocation vs translocation comparison. These three were from the genes *GUSBP3*, *NRG1*, and *LRRC37A2*.

Results of DESeq2 and DEXSeq are summarised in Table 2.

	Total Detected	**Significantly differentially expressed**
Genes	21,916	1,252
Exons		1,368 in 1,020 genes

Table 2. Summary of differential expression findings of non-XY genes when utilising DESeq2 (gene level analysis) and DEXSeq (exon level analysis), with "Significantly differentially expressed" meaning adjusted p value<0.05 and BaseMean>10 for genes.

3.5 Local expression

I analysed local expression around the breakpoints, as there is evidence that large chromosome rearrangements can result in altered chromosomal position within the nucleus. This appears to coincide with transcriptional effects[158]. I analysed this in two ways; by looking to see if more differentially expressed genes were found around the breakpoints than by chance, and whether expression showed an overall change regardless of gene significance.

3.5.1 Chromosome 1

3.5.1.1 Differentially expressed genes

One way to see if the breakpoints are affecting local expression would be to see if there is an unusually high number of differentially expressed genes around the breakpoint. I carried out a Monte Carlo simulation. To assess the background level of differentially expressed genes I selected 100,000 loci at random. These all came from chromosome 1 so as to ensure the background level of gene richness is an appropriate comparison. I then calculated the proportion of differentially expressed genes around those points within a "window". I initially set the window size to 40Mb but reduced this to 37.5Mb as *DISC1* is too close to the end of the chromosome. I then repeated this with windows of 20Mb and 10Mb. The results of this analysis are in Table 3. There are 273 genes within 20Mb either side of the chromosome 1 breakpoint.

Chromosome 1	**18.7Mb**	**10Mb**	**5Mb**
Proportion of genes significant around *DISC1*	0.069	0.038	0.051
Average proportion of genes significant around 100,000 points	0.069	0.070	0.067
P value for *DISC1* window	0.43	0.91	0.58

Table 3. Assessment for significantly increased proportion of differentially expressed genes in 18.7, 10, and 5Mb windows around *DISC1*, by comparison to 100,000 windows around chromosome 1.

We can see that for all analyses, shrinking the gene window by half reduces the average number of significant genes by about 50%, but does not affect the proportion of significant genes at all. It is evident that there is no significant enrichment of differentially expressed genes around the breakpoint on chromosome 1.

3.5.1.2 Local expression in FPKMs

Expression in terms of raw counts, FPKMs, and normalised counts using rlog (a normalisation method included in the DESeq2 package) was assessed around the breakpoints. However, I realised that there were several arguments against assessing local expression in any of these terms except FPKMs.

Raw counts would seem initially an attractive option for mapping the expression, but they do not adjust for sequencing depth and are therefore less than ideal. Since I am looking for relative change within a sample as I look around the breakpoints, with the idea of comparing this change to that of other samples, FPKMs are ideal. I am looking for regional changes within each sample, and then comparing these regional changes within each sample to those in other samples. I therefore made maps of genomic expression using FPKMs for each sample, which allow gene to gene comparisons to see areas of high and low expression normalised to sequence depth and gene length.

I selected all the genes within approximately 18.7Mb on both sides of the *DISC1* gene and graphed their expression against their location. A graph was made for each of the 18 samples, and they were compared by eye to see local expression changes. A condensed version of the result, where averages for each genotype are compared, can be seen in Figure 10. There appears to be no pattern of depressed or increased expression and I therefore did not investigate further. Looking at the graphs for each of the 18 samples, there appears to be no clear pattern of change between translocation and control samples. For the sake of brevity these are not reproduced here.

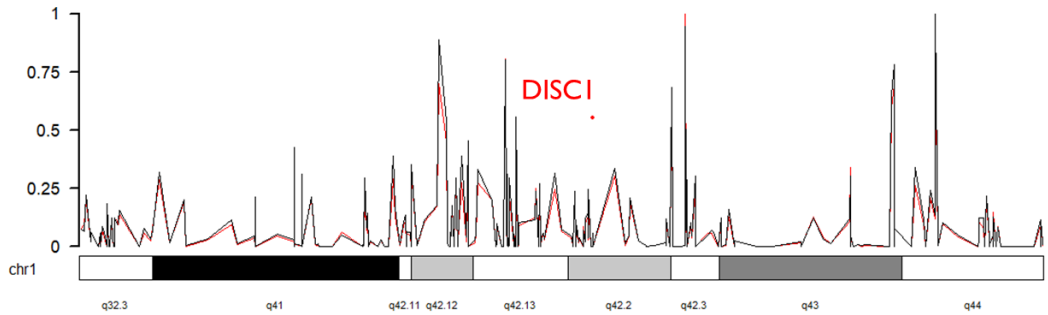


Figure 10. Genomic expression map of 37.5Mb around *DISC1*. Note that the chromosome ends approximately 17.5Mb after *DISC1*.Black indicates the average of the control FPKMs, red the average of the translocation FPKMs. Y-axis is in local minimum to local maximum expression, each point is a gene.

3.5.2 Chromosome 11

3.5.2.1 Differentially expressed genes

I applied the same methods as in 3.5.1.1. The results are in Table 4. Note that a 20Mb window was used as the breakpoint is not too close to the end of the chromosome to preclude this. There are 260 genes within 20Mb either side of the chromosome 11 breakpoint.

Chromosome 11	**20Mb**	**10Mb**	**5Mb**
Proportion of genes significant around *DISC1FP1*	0.088	0.16	0.14
Average proportion of genes significant around 100,000	0.065	0.068	0.069
P value for *DISC1FP1* window	0.18	0.016	0.11

Table 4. Assessment for significantly increased proportion of differentially expressed genes in 20, 10, and 5Mb windows around *DISC1FP1*, by comparison to 100,000 windows around chromosome 11.

The only significant result is the 10Mb window. There are two variants in chromosome 11 which may influence psychiatric disease in this family, which are located 3Mb upstream and 10Mb downstream from the chr11 breakpoint[162]. Neither of the two genes present in these loci which could be examined, *CNTN5* and *GRM5*, were significantly differentially expressed[70]. The effect of these loci is not entirely

known but they may influence the diversity of phenotypic presentation in the Scottish pedigree. We can conclude that the translocation may have an effect on medium distance gene expression, although the evidence is weak considering one would expect the 5Mb window to also be perturbed if this was the case. One way to see if the effect is only on genes a certain distance away (rather than on genes *up* to a certain distance away) would be to redo the analysis with "bands" rather than windows, but given the slim evidence for any regional effect I elected not to do this.

3.5.2.2 Local expression in FPKMs

I selected all the genes within 20Mb on both sides of the *DISC1FP1* gene and graphed their expression against their location. A graph was made for each sample, and they were compared by eye to see local expression changes. A condensed version of the result, where averages for each genotype are compared, can be seen in Figure 11. The two averages track one another closely and there appears to be no pattern of changed regional expression.

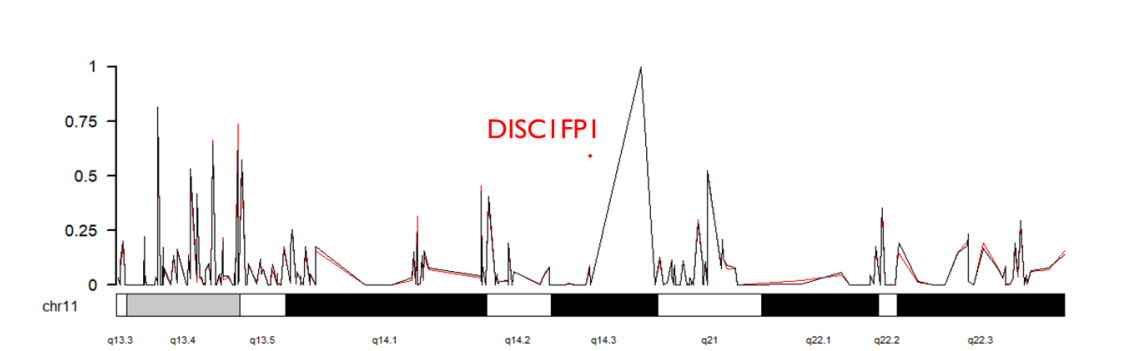


Figure 11. Genomic expression map of 40Mb around *DISC1FP1*. Black indicates the average of the control FPKMs, red the average of the translocation FPKMs. Y-axis is in local minimum to local maximum expression, each point is a gene.

3.6 GOrilla

To analyse the data for gene ontology terms that are overrepresented among the differentially expressed genes, I utilised GOrilla[163,164]. I compared the differentially expressed genes meeting the BaseMean>10 and $p<0.05$ criteria together with the list of genes with a differentially expressed exon at $p<0.05$. The background was the list of all genes detected at the whole gene level. GOrilla analyses overrepresentation of terms at the Process, Function, and Component levels. The top 10 significant results

by p-value are displayed in Figure 12, Figure 13, and Figure 14 respectively. A full discussion of candidate genes for further investigation is elsewhere, but summaries for terms of interest in each category are below. It should be noted that I usually do not discuss the most significant GO terms, but have selected terms of interest to discuss. These terms are usually outside the top 10 GO terms but are far more specific. The top GO terms are usually broad in nature, and contain dozens of genes, many of which only loosely relate to the term. However even granted this it is true that I have been selective in my discussion of GO terms; my primary reason for this is that I originally generated GO terms to help produce candidate genes. Topics such as "ionotropic glutamate receptor binding" or "intracellular transport" seem more amenable to biological experimentation than "cell adhesion" or "nervous system development", which are highly broad terms. In addition, I have provided p values, and all GO terms discussed are of course significant both nominally and at FDR-corrected p values (Benjamini and Hochberg method). All significantly differentially expressed genes have also been provided in the supplementary information of Malavasi *et al*[70].

3.6.1 GOrilla Process

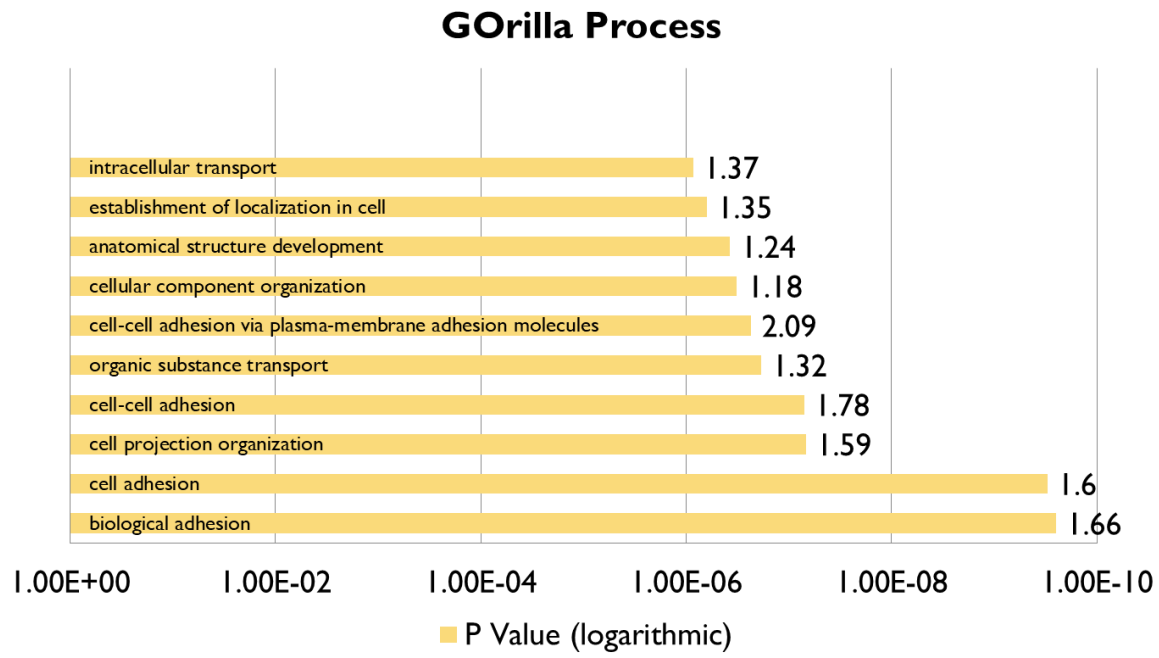


Figure 12. Significantly overrepresented Process GO terms for the genes which have a differentially expressed exon or are differentially expressed (p<0.05). The enrichment figure is given after each bar and the scale is logarithmic.

The top GO term is "biological adhesion", which contains the gene *DAG1*, discussed in the context of "actinin binding" in the GO Function section below. It also contains the gene *CLDN5*, which is potentially involved in schizophrenia and downregulated. It is notable for being contained within a region of the genome 22q11.2, which is found deleted in patients with a syndrome presenting with cardiac abnormalities, craniofacial abnormalities, and an increased risk of schizophrenia[165]. We can see that a number of the other entries are similar, but of interest are "cell projection organisation" and "intracellular transport". Intracellular transport of mitochondria is known to be affected by DISC1[95], and the transport of synaptic vesicles containing neurotransmitters is necessary for neuronal activity. DISC1 is a known trafficker and can also affect $GABA_AR$ and NMDAR surface presentation[70,166]. A number of the genes with this GO term were good candidates for further investigation, including *HAP1, SPIRE1, SLIT1,* and *SYT6.* However as stated at the start of this chapter, a recent paper has highlighted genes which may be affected by sex. These genes include *HAP1* and *SLIT1*. *HAP1* is an interactor of the product of the *HTT* gene, which causes Huntington's disease if mutated. *SLIT1* is a guidance molecule which

assists neuronal axon directions, while *SYT6* is a synaptotagmin, a family of genes involved in vesicular exocytosis[167,168].

Of particular interest to me was the presence of the *BBS2* and *BBS5* genes, which are mutated in Bardet-Biedl syndrome (BBS). *BBS1* is also differentially expressed and a fuller discussion is in the introduction. GO terms with significant p values included "synapse assembly" ($p=2.38x10^{-6}$, enrichment factor=2.99), which included genes such as *APP*, *NRG1*, *BSN*, *DRD2* and *ERBB4*, some of which have relevance as they are schizophrenia candidates (*ERBB4*, *NRG1)* or are the targets of antipsychotic medication (*DRD2*). *APP* is best known as the gene encoding the amyloid precursor protein, while *BSN* is a large presynaptic protein involved in vesicle exocytosis and trafficking[169,170]. Nervous system development was another GO term ($p=5.69x10^{-6}$, enrichment factor=1.86) of interest. However, it is difficult to assess the accuracy of this as two Hox genes (*DLX2* and *DLX6*) are likely to be differentially expressed according to sex. However two other Hox genes *VAX1* and *VAX2* in a different GO term group were differentially expressed.

3.6.2 GOrilla Function

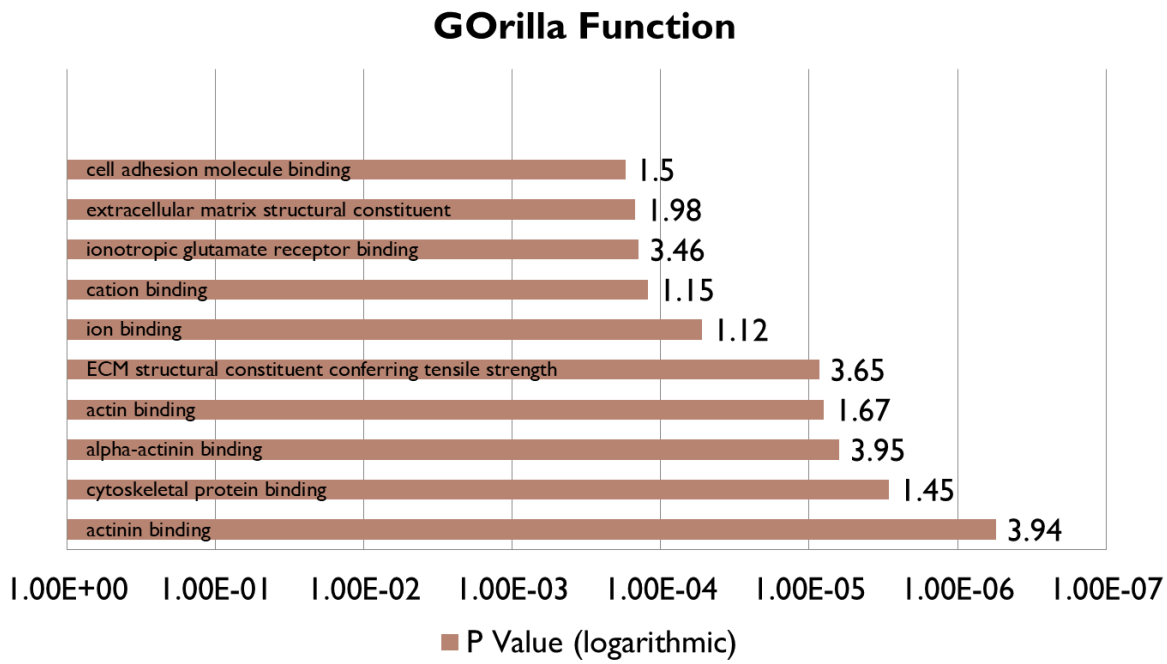


Figure 13. Top 10 significantly overrepresented Function GO terms for the genes which have a differentially expressed exon or are differentially expressed ($p<0.05$). The enrichment figure is given after each bar and the scale is logarithmic. ECM=extracellular matrix

The most significant term is "actinin binding". A number of genes were contained within this term, including a pair of potassium receptors *KCNA5* and *KCNN2*. Both potassium receptors and NMDARs have been shown to interact with alpha-actinin[171]. Another associated gene *DAG1* encodes an interactor of the protein encoded by *NRXN1*, which has mutations associated with autism and schizophrenia[172]. Under the "actin binding" term, the fourth most significant, a number of myosin and kinesin genes were found, indicating that trafficking molecules are abnormally expressed in these cells. An actin nucleator, *SPIRE1*, was also differentially expressed. Actin is important in growth cone organisation and initiating neuronal migration. Changes in these actin and cytoskeletal terms are interesting and may reflect possible migratory defects, as also evidenced by the dysregulation of centrosomal proteins[82]. The term "ionotropic glutamate receptor binding" with an enrichment factor of 3.46 is also relevant. Glutamate receptors such as the AMPARs and NMDARs are known to play a role in LTP, necessary for learning and memory. NMDAR antagonists such as

ketamine also induce psychosis[173]. Mutations in these families of genes have also been linked to psychiatric disease[174].

3.6.3 GOrilla Component

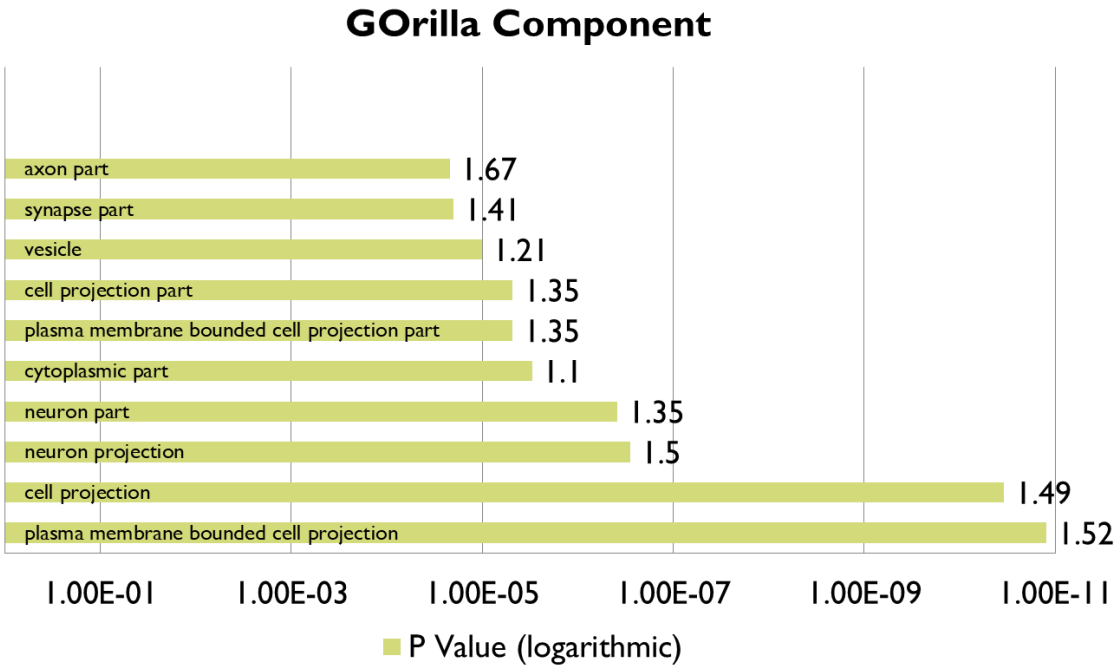


Figure 14. Top 10 significantly overrepresented Component GO terms for the genes which have a differentially expressed exon or are differentially expressed ($p<0.05$). The enrichment figure is given after each bar and the scale is logarithmic.

GO overrepresented component terms mostly relate to cell parts that are unique to the neuron, such as “axon part”, “synapse part”, and terms which relate to cell parts which have a special role in neuronal activity such as “vesicle”. This is unsurprising. It could be concerning if component terms uniquely related to some non-brain cell types had emerged as significant.

3.7 Comparison to other studies

It would be of interest to see if the genes differentially expressed in our experiment are also changed in analogous experiments. Comparisons can be made to other RNA-Seq experiments (particularly involving *DISC1* mutations and iPSC-derived neurons), GWAS, and CNV studies. Microarray studies are also a comparative option but have mostly been supplanted by RNA-Seq experiments. A search of the

literature was conducted and a number of studies which were apt comparisons were utilised.

3.7.1 Studies used

These studies are described in greater detail in the Introduction. All studies are in human cells or with human samples.

3.7.1.1 Psychiatric Genomics Consortium GWAS

GWAS search for genes predisposing to schizophrenia based on the concept of the "common disease, common variant" model, where many variants have a small but detectable impact on the risk of developing disease. The PGC have published multiple papers with ever-larger samples sizes. PGC1 refers to the first paper, implicating 108 loci and 348 genes[29]. PGC2-1 refers to the association analysis of the second, implicating 145 loci and 481 genes[60], while PGC2-2 refers to the list of genes implicated by MAGMA in that paper (535 genes).

3.7.1.2 Psychiatric Genomics Consortium CNV study

The PGC have also searched for copy number variations (CNVs) predisposing to schizophrenia. PGC3 refers to this study[139]. I utilised the gene list associated with schizophrenia by the Gene-Association analysis carried out in this paper.

3.7.1.3 Camargo *et al.* DISC1 interactors

Camargo *et al.* utilised a yeast two hybrid screen to identify potential interactors of the DISC1 protein[73]. All genes indicated by this study were searched for.

3.7.1.4 Brennand *et al.* idiopathic schizophrenia

Brennand *et al.* were the first to look at differential expression in iPSC-derived neurons[124]. Although the numbers were small, lines were established from patients with idiopathic schizophrenia. It utilised a microarray approach to look at expression differences and found 596 genes differentially expressed at $p<0.05$ and fold-change >1.3. B refers to this study.

3.7.1.5 Wen *et al. DISC1* frameshift

Wen *et al.* looked at iPSC-derived neurons from a family carrying a *DISC1* frameshift mutation[132]. The number of genotyped family members is small however. The frameshift cannot be unambiguously linked with psychiatric illness, while it is also the case that members of the pedigree have psychiatric illness but no *DISC1* frameshift. They also did not use as large a number of RNA samples (one control and two mutants, all in triplicate). A total of 3,697 genes were differentially expressed in their study, although as we have already seen from my 3 vs 3 DESeq2 comparisons that smaller numbers appear to give more differentially expressed genes, likely due to false positives. W refers to this study.

3.7.1.6 Srikanth *et al. DISC1* truncations

Srikanth *et al.* looked at two different timepoints in the production of neurons directly from iPSCs[138], although their protocol is not identical to ours. They induced mutations in the iPSC lines prior to neuron differentiation, in exon 2 and exon 8 of *DISC1*. The first mutation should remove all *DISC1* isoforms while the second is closer in effect to the t(1;11) translocation in its effect on *DISC1* by inducing a truncation close to the breakpoint. They also looked at heterozygous and homozygous carriers of these mutations. S refers to this study, x2 and x8 to the mutations, w/m to wild type/mutant status, and 18 or 50 to the timepoints. For example, Sx8wm50 refers to the heterozygous carriers of the exon 8 truncation at the neuron stage.

3.7.2 Results

Human t(1;11) Paper	Number of genes	P value	Selected genes of interest
PGC1	32	7E-06	*CHRNA4*, *DRD2*
PGC2-1	37	1E-03	*CHRNA4*, *DRD2*
PGC2-2	47	1.4E-03	*DRD2*, *CHRM4* (gene)
PGC3	13	0.36	*HOMER2*, *SHANK1*, *SYT6*
DI	10	0.31	
B	87	1.8E-03	*NRP2*, *NQO1*, *COBL*
W	300	4E-07	*VAX1*, *DRD2*, *COBL*, *DLX2*,
Sx2mmd50	200	5.6E-07	*BBS5*, *LRRTM1*, *SLIT1*,
Sx8mmd50	17	0.077	
Sx8wmd50	73	6.4E-04	*SEMA3F*

Table 5. Summary of overlap with other papers. Each paper is indicated by the acronym given in 3.7.1. The number of genes significant in both our study and the indicated one is given in the first row. The hypergeometric probability is given in the second, and a subset of interesting genes within this list of overlapping genes is within the third.

3.8 Gene level RT-qPCR

To confirm the results of the RNA-Seq, a number of RT-qPCRs were performed. All 68 genes with an absolute fold change >2 were considered as candidates. Most had no known relevance to major mental illness, while the remainder had poor evidence of line difference (e.g. a single line or sample drove significance). Seven were TRIM genes (tripartite motif containing), four were olfactory receptor family members, while three were prame family members. None of these appeared to be directly relevant to psychiatric illness. The largest fold change gene was MIR4458HG, a microRNA encoding gene about which little is known. A number of interesting candidates were discarded in cases where one line clearly drove the significance of the gene. These included genes such as *VIPR2*, *DDC*, *CCK* and *NRCAM*. All 24

genes which were significant in every WT vs t(1;11) line DESeq2 analysis (i.e., three samples from one WT line compared against three from a translocation line) were also assessed as candidates. Having exhausted these obvious possibilities and finding only a few genes, I attempted to select candidates from the remaining bulk of genes. Genes were chosen on the basis of high fold change, clear difference between genotypes, convergence with other papers, existing evidence of relevance to psychiatric disease, and contributing to a GO term of relevance. OMIM and Pubmed were also examined for disease associated variants or papers of interested. Of some use in selecting candidates was an approach searching for convergences between papers; a table displaying the numbers of genes significant between any pair of papers described in 3.7.1 as well as in the DESeq2 analysis is displayed as Table 6. Descriptions are given for each gene in turn and a table summarising the rationales is given in Table 7. Primer design is described in Materials and Methods.

DESeq2 crossovers in t(1;11) neurons	PGC1	PGC2-1	PGC2-2	PGC3	Exon level changes	DI	B	W	*Der1* Het cortex	*Der1* Het hippocampus	Sx2mm D50	Sx8mm D50	Sx8wm D50
PGC1	16	13	11	0	1	0	0	4	5	0	5	0	0
PGC2-1	13	20	12	0	3	0	0	5	4	0	5	0	2
PGC2-2	11	12	29	0	1	0	1	6	6	1	4	0	1
PGC3	0	0	0	5	0	0	0	1	1	0	0	0	1
Exon level changes	1	3	1	0	152	1	12	41	14	2	17	3	9
DI	0	0	0	0	1	2	0	1	1	0	0	0	0
B	0	0	1	0	12	0	68	22	9	2	8	2	6
W	4	5	6	1	41	1	22	213	23	4	36	7	17
Der1 heterozygous cortex	5	4	6	1	14	1	9	23	126	2	21	0	5
Der1 heterozygous hippocampus	0	0	1	0	2	0	2	4	2	12	1	0	0
Sx2mmD50	5	5	4	0	17	0	8	36	21	1	144	6	11
Sx8mmD50	0	0	0	0	3	0	2	7	0	0	6	13	6
Sx8wmd50	0	2	1	1	9	0	6	17	5	0	11	6	52

Table 6. Table of overlaps. Numbers represent the number of genes significant in our DESeq2 study of human neurons, in addition to the two papers in the corresponding row and column. Grey blocks indicate genes from only one paper (the same row and column index). Abbreviations as in 3.7.1.

	Expression as % of WT	Padj	Human exon changes	Mouse cortex changes	Mouse hippocampus changes	PGC1	PGC2.1	PGC2.2	PGC3	DI	B	W	Sx2mmd50	Sx8mmd50	Sx8wmd50
BBS1	77	3.50E-02		Exon level											
CALB1	23.4	7.37E-04	Yes								TRUE				
CHRNA4	215.8	4.26E-03			Gene level	TRUE	TRUE								
DRD2	339.5	4.31E-04				TRUE	TRUE	TRUE			TRUE				
GPC1	56.2	3.08E-05		Exon level	Gene level						TRUE				
HAP1	29.5	1.68E-04		Gene and exon level											
HIF1A	136.6	1.19E-03		Exon level											
KANSL1	194.8	4.40E-05	Yes												
METRN	45.2	1.61E-05		Gene level							TRUE				
NRP2	52.9	4.10E-04		Gene level						TRUE					
NTRK2	62.9	1.55E-02	Yes	Exon level									TRUE		
PDYN	3.6	2.60E-06									TRUE		TRUE		

Table 7. Highlighted information about candidate genes selected for qPCR. TRUE indicates that the gene is differentially expressed in the model of interest. Paper abbreviations are as in 3.7.1.

3.8.1 BBS1

Bardet Biedl Syndrome 1, *BBS1,* is a one of a number of genes which when mutated cause Bardet-Biedl syndrome, characterised by developmental delay, obesity, intellectual disability, and retinal problems. These genes participate in the construction of the BBSome, a protein complex seen at cilia and centrosomes[175]. In cilia it participates in trafficking, and interestingly BBS4 is seen to interact with dynactin, the component of the dynein motor complex HAP1 interacts with. At the centriole many of the BBS proteins interact with DISC1 along with other centrosomal proteins such as PCM1[176]. A number of these proteins are DISC1 interactors, including BBS1 and BBS4. *BBS1*, *BBS2*, and *BBS5* are all significantly downregulated according to the RNA-Seq data. The interaction with BBS1 is particularly interesting; phosphorylation on a key S70 residue of DISC1 results in a functional change. Upon phosphorylation of this residue, DISC1 loses its ability to enhance Wnt signalling and aids the centrosomal location of BBS1, which aids neuronal migration. Blocking this phosphorylation event, knocking down *DISC1*, or *BBS1*, hinders mouse neuronal migration in the developing cortex[81]. Given its role in migratory neurons and direct interaction with DISC1, as well as the downregulation of other BBS proteins, *BBS1* stands as a good candidate for qPCR analysis as well as future study.

3.8.2 CALB1

Calbindin is encoded by this gene and is a calcium binding protein abundant in the brain. It is found downregulated here[177]. Given calcium's importance as a secondary messenger this presented as an interesting candidate. The protein has a known role in buffering calcium increases which may be protective against excitotoxicity[178]. Expression of *CALB1* in neurons marks them out for survival against neurotoxic drugs in a mouse model[179]. Calbindin is known to exert anti-apoptotic roles via inhibition of caspase-3 and of calpain, which activates BAX[180]. Its protective effect against hyperdopaminergic signalling may be particularly relevant given the upregulation of *DRD2*. Finally, calbindin is known to interact with and enhance the activity of inositol monophosphatase, which is a target of lithium. This mechanism is hypothesised to be relevant in bipolar disorder treatment using lithium[177]. Myo-

inositol production catalysed by IMPase is important as inositol is a key substrate for the production of many signalling molecules. The hypothesis of inositol and bipolar disorder holds that lithium-induced inositol depletion helps reduce deleterious signalling. However, in the case of the t(1;11), this fits oddly with the finding that *CALB1* is downregulated. If the inositol hypothesis of bipolar disorder is correct, downregulated calbindin would be beneficial[181]. To further complicate matters, calbindin deficient mice show deficits in LTP, possibly via a failure to appropriately buffer calcium. In any case there are several lines of evidence tying *CALB1* to psychiatric pathology, although the evidence is perhaps not as strong as one would like and is occasionally contradictory. For example, one group found that calbindin had no anti-apoptotic effect and appears to mark surviving cells rather than aid cells to survive apoptotic challenge[182]. Nevertheless, this seemed an interesting gene to examine.

3.8.3 CHRNA4

This gene encodes neuronal acetylcholine receptor subunit alpha-4, a subunit of the nicotinic acetylcholine receptor. It was the first gene discovered to be causative of a type of frontal lobe epilepsy when mutated. All causative mutations appear to be gain-of-function, causing increased acetylcholine sensitivity, and some mutations (but not all) are linked to aberrant cognition[183]. The gene is upregulated in our t(1;11) samples compared to the WT. Especially interesting is the finding that a particular familial mutation in *CHRNA4* frequently co-presents with frontal-lobe epilepsy and schizophrenia, although the pedigree is small and no LOD score can be calculated. This mutation does not appear to be a loss-of-function (it is characterised by an extra Leucine residue) and other families with different *CHRNA4* mutations have epilepsy but no increased incidence of schizophrenia[184]. The product of *CHRNA4*, α4, is a protein that cooperates with a number of other subunits to form functional nicotinic acetylcholine receptors. α4β2 is one of the better studied nicotinic receptors consisting of alpha-4 and beta-2 protein subunits. These receptors allow the entry of cations such as Ca^{2+} and are triggered by acetylcholine, and by the exogenous ligand nicotine. This NAcR-triggered calcium influx in particular has been shown to help stimulate synaptic LTP[185]. The α4β2 receptor is widely expressed in the brain and is

highly expressed in the cerebral cortex, substantia nigra, thalamus, and hippocampus. There are also differences in receptor expression across some regions of the brain in post-mortem studies of individuals with schizophrenia, with greater expression in the striatum and some layers of the cingulate cortex. It is also notable that the prevalence of smoking is far greater among schizophrenic patients than the general population[186]. Although our cell culture is neither comprised primarily of GABAergic interneurons, nor dopaminergic ones, the α4β2 receptor formed by the product of *CHRNA4* seems to have effects on both these cell types that may be of relevance to the pathophysiology of schizophrenia. VTA dopaminergic neurons are stimulated greatly by nicotine, via α4-containing receptors, and dopaminergic signalling has long been implicated in schizophrenia. However, it cannot be said if the α4 increase in our neurons would also be apparent in VTA dopaminergic neurons. More directly however, there is evidence that activation of the receptor in cortical tissue alters circuit excitability. This could be of relevance to the pathophysiology of schizophrenia and other disorders[185].

CHRNA4 was indicated by Tiihonen *et al*. as being differentially expressed according to sex in iPSC-derived neurons. The putative change is upregulation in females, and here it is downregulated in translocation samples (which are disproportionately male). Therefore sex may well be a factor in *CHRNA4*'s significance and the results must be judged critically in this regard.

3.8.4 DRD2

DRD2 encoding Dopamine Receptor D2 is perhaps one of the best known genes in schizophrenia research. All known antipsychotics that treat the positive symptoms of schizophrenia antagonise this dopamine receptor, and it has been recently implicated by GWAS[60]. In addition, iPSC-derived neurons from patients with schizophrenia show increased secretion of dopamine compared to control neurons, although the effects are very variable[137]. It also has relevance to bipolar disorder; both this and schizophrenia respond to dopamine bloackade[51]. Finally, other studies of *DISC1* mutants have shown elevated levels of *DRD2* binding in mouse brain[115]. All of these lines of evidence link perfectly with the observation that DRD2 is significantly

upregulated in neurons carrying the t(1;11). Dopamine is a neurotransmitter important in diverse neurological activity leading to movement, pleasure, and memory[53]. Given the pharmacological relevance, its dysregulation in our cortical mouse model and others, as well as the support from GWAS, *DRD2* is a natural choice for further investigation.

DRD2 was indicated by Tiihonen *et al*. as being differentially expressed according to sex in iPSC-derived neurons. However, the putative change is upregulation in females, whereas here it is upregulated in translocation samples (which are predominantly male). Therefore sex does not appear to be a factor in *DRD2*'s significance.

3.8.5 ERBB4

The analysis of this gene was carried out by Helen Torrance and Kirsty Millar, but is also reported here due to its relevance. Results were previously described in Malavasi *et al.* 2018[70] .

ERBB4 is a receptor tyrosine kinase; its ligand is NRG1. The gene encoding NRG1 is also differentially expressed. It is expressed in the glutamatergic synapse and overexpression enhances AMPA currents, while RNAi reduces dendritic spine density and size[187]. ERBB4 appears to be the primary receptor for much of NRG1's effects, as null mouse neural progenitors are immune to NRG1-induced proliferation[188]. *Erbb4* mouse mutants also have behavioural phenotypes, while mice null for *Erbb2* or *Erbb3* do not[189].

3.8.6 GPC1

Glypican-1 is encoded by this gene. It is one of a group of heparan sulfate proteoglycans, a type of modified protein which has been studied for its roles in development. Of the six glypicans, the one encoded by *GPC1* appears to be the most widely and highly expressed in the CNS. It appears to be expressed broadly within the cell body in $TUJ1^+$ cells[190]. *GPC1* is perhaps most notable for the phenotype of *GPC1* homozygote and heterozygote mouse nulls as described by Jen *et al.*; they observed a reduction in brain size proportional to the number of null alleles.

Homozygote nulls have an approximately 20% reduction in brain size due to lower numbers of cells, not evident at E8.5, but evident at E9.5 through to adulthood. There was some evidence suggesting that Fgf signalling was decreased, meaning that proliferative neural precursors could be decreased in number, leading to the later lower number of cells. Correspondingly, increased numbers of $Tuj1^+$ cells were seen in the developing brain at E9.5 with no increase of apoptosis, implying that cells were differentiating aberrantly[191]. Glypican-1 is also a known interactor of Slit2, known for its roles in neuronal guidance and axonal repellence, and the two proteins are co-expressed in many areas of the rat brain including cerebral cortex and hippocampus[192].

What relevance might the downregulation of GPC1 have to the pathology of the t(1;11)? As with the case of *SLC12A2*/Nkcc1 (see 3.9.10), it appears that a developmental change is occurring earlier than expected. Oikari *et al.* looked at proteoglycan expression during differentiation of NSCs and found that *GPC1* was highly upregulated during neuronal differentiation to a $TUJ1^+$ phenotype (*DCX* and *NEFM* were also upregulated during this differentiation). Interestingly, a NSC *GPC1* knockdown line, at an early stage of differentiation, had downregulations of both the NPC markers Nestin and MSI1, and of the neuronal markers TUJ1 and NEFM[190]. One hypothesis is that *GPC1* downregulation causes two aberrant processes: the first being early exit from cell cycle, as elaborated in Jen *et al.* and causing downregulation of NSC markers as in Oikari *et al.*, resulting in less neurons due to a reduced precursor pool. The second is possible entrance into an unusual cell fate, resulting in less neuronal markers. This would explain the downregulation of both marker types. Such a process could well be occurring in our neuronal cultures, resulting in premature development leading to a lower number of neurons, or of an unusual fate for some of those cells. The gene was significant in every WT vs t(1;11) DESeq2 analysis, and never significant in a t(1;11) vs t(1;11) or WT vs WT comparison.

3.8.7 HAP1

Huntingtin Associated Protein 1 (HAP1) is a neuronally enriched protein found in many areas of the cell, including cell bodies and axons. It was first discovered as an interactor of the Huntingtin protein, a protein in which glutamine repeat expansions are causative of Huntington's Disease (HD), a progressive neurological disorder characterised by chorea, cognitive decline, and behavioural abnormalities. *Hap1* knockout mice display reduced body weight and thalamic degeneration, as do human HD patients and transgenic HD mouse models. HAP1 appears to have roles in cellular trafficking. It interacts with dynactin, an essential component of the dynein motor complex, as well as KLC2 and synaptic vesicles[193]. The homologue in Drosophila is also critical for kinesin dependant anterograde transport of mitochondria, although this has yet to be investigated in human cells[193]. Transfected HAP1 increases numbers of surface GABARs, apparently by drastically reducing the rate of internalized receptor degradation[194]. It is especially interesting that DISC1 appears to promote $GABA_AR$ surface presentation[166]. This is conceivably via HAP1. This concept is summarised in Figure 15.

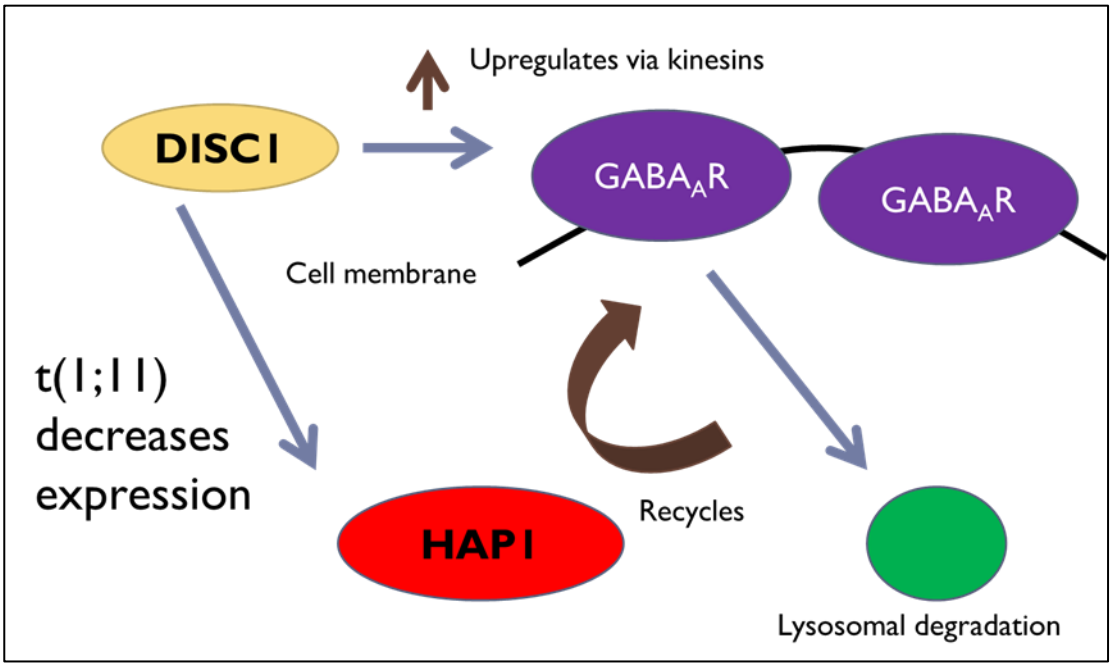


Figure 15. Illustration of theoretical DISC-HAP1-$GABA_AR$ interactions. Links between DISC1 and $GABA_ARs$ described by Wei *et al.*[166],links between HAP1 and $GABA_AR$ described by Kittler *et al.*[194]. DISC1-HAP1 link described in this thesis.

Particularly of note are the links between HAP1 and BDNF signalling described by Gauthier *et al.*[195]. siRNAs against Htt (the Huntingtin protein) increase the proportion of stationary BDNF-containing vesicles in neuronal cells, and cause slower movement of the vesicles overall. Transfection of Htt increases vesicular trafficking in cortical neurons, while empty vectors or the glutamine-expanded deleterious Htt isoform do not aid trafficking. The increase was especially evident in neurites and was not displayed by htt lacking the HAP1 interaction domain, strongly suggesting that the increase in transport is mediated by HAP1[195]. Htt also interacts with the aforementioned dynactin protein, an interaction which is potentiated by HAP1. It has been shown that direct *HAP1* knockdown or indirect destabilisation decreases levels of TrkB, the BDNF receptor encoded by *NTRK2*, in the cerebellum and brainstem of postnatal mice, as well as the phosphorylation of its key signalling molecules[196].

Given the necessity of transport for extended cells such as neurons, as well as its ties to a pathological process, *HAP1* presents as a strong candidate for qPCR investigation. It is also downregulated in the mouse model of the t(1;11), lending further credence to it being genuinely changed and perhaps functioning as the "missing link" between DISC1 and $GABA_AR$. Particularly interesting is the role HAP1 has in effecting the Htt protein's functions. Lack of HAP1 has been shown to abrogate the stimulatory effects of increased Htt on cellular trafficking; it is possible that lack of HAP1 in cells with normal levels of Htt similarly disrupts trafficking.

I had examined *HAP1* via qRT-PCR, but *HAP1* was later indicated by Tiihonen *et al*. as being differentially expressed according to sex in iPSC-derived neurons. The putative change is upregulation in females, and here it is downregulated in translocation samples (which are disproportionately male). Therefore sex may well be a factor in *HAP1*'s significance and the results must be judged critically in this regard.

3.8.8 HIF1A

Hypoxia-inducible factor 1-alpha (HIF1A) is a transcription factor which acts as the master regulator of response to hypoxia[197]. Prenatal and perinatal hypoxia have been

established by epidemiological studies as a schizophrenia risk factor, and a large number of candidate genes for schizophrenia are related to hypoxia signalling in some way. These include *BDNF*, the ligand for NTRK2, and *NRG1*, the ligand for ERBB4 (described in 3.8.13 and 3.8.11 respectively). They also include *COMT* encoding the dopamine degrading catechol O-methyl transferase, found in the 22q11.2 deletion region mentioned previously. Finally, these include *CHRNA7* (Cholinergic Receptor Nicotinic Alpha 7 Subunit, mice lacking this gene have deficits in parvalbumin positive interneurons[198]), and *RELN* encoding reelin[199]. Many of these appear to be disturbed in the t(1;11) neurons as well. However, it remains to be seen in if this remains true in the post-GWAS era. It also appears that neonatal hypoxia and genetic propensity for schizophrenia can interact, lending more evidence to the importance of hypoxia signalling[200]. It therefore seemed of interest to investigate its apparent upregulation further.

3.8.9 KANSL1

KANSL1 encodes KAT8 regulatory NSL complex subunit 1. Gene expression is increased in the neurons carrying the translocation. It encodes a chromatin modifying protein. Haploinsuffiency or point mutation in humans causes a developmental phenotype including intellectual disability, hypotonia, and distinctive facial morphology[201]. Differential expression of a large number of genes related to cell-cell signalling and synaptic transmission is found in cell lines carrying the human mutation. Mutation of the *Drosophila* homolog results in deficits in learning, and decreased binding of the protein to chromatin around genes related to those same functions, including synaptic transmission[201]. It is interesting that the gene is upregulated in our cells, but it could be some kind of compensatory mechanism. The gene appears to be involved in endosomal maturation, which could be relevant given the role of endosomes in the recycling and clearance of neurotransmitters. Genes related to synaptic trafficking, such as *SYT6*, are also differentially expressed in our neurons.

3.8.10 METRN

This gene encodes Meteorin and is highly downregulated in our samples, as well as in the model described by Wen *et al.*[132]. *METRN* was primarily of interest due to its ability to promote axonal extension in neurons. It is expressed in neuronal progenitors, and is highly expressed in myelinating oligodendrocytes. It appears to have a role in promoting glial differentiation[202,203]. It therefore is of importance to neuronal differentiation and the formation of networks, which are processes of interest. It is also differentially expressed in the mouse heterozygote cortical model.

3.8.11 NRG1

The analysis of this gene, encoding Neuregulin 1, was carried out by Helen Torrance and Kirsty Millar, but is reported here due to its relevance. Results were previously described in Malavasi *et al.* 2018[70] .

NRG1 and its cognate receptor *ERBB4* are both downregulated in the t(1;11) neurons, as described in the *ERBB4* section 3.8.5.*NRG1* has many roles of relevance to psychiatric disorder processes; it assists in the migration of cortical neurons, progenitor proliferation, and axonal guidance. It also aids synapse formation via induction of PSD95, and plasticity[188]. Mouse mutants of *Erbb4* or *Nrg1* display hyperactivity and impaired prepulse inhibition, phenotypes which bear relevance to psychiatric disease in humans[204]. *NRG1* also has a role in parvalbumin-positive interneuron migration, a cell type which is thought to be of relevance to schizophrenia in particular[116]. Given its historic relevance as a schizophrenia candidate, and its links to multiple processes of relevance to synaptic activity and neuronal development, *NRG1* is a good choice for qRT-PCR.

3.8.12 NRP2

Neuropilin-2 is encoded by this gene and is one of a family of receptors which help mediate the chemo-attractive and chemo-repulsive effects of the semaphorin ligands. These effects are necessary for neuronal extensions to reach their eventual destination and make synaptic connections. *Nrp2* deficient mice have been studied, and are seizure prone, although this phenotype is of more relevance to epilepsy. They

also have fewer interneurons, including parvalbumin positive and GABAergic interneurons, which many schizophrenia risk factors converge on[116]. CA1 pyramidal neuron dendrites also show decreased length and complexity[205]. It is possible that the downregulation of *NRP2* causes similar phenotypes here. However, it is unusual that the *DLX1/2* Hox genes are also downregulated in the t(1;11) neurons. Mouse nulls of the DLX homologues are embryonic lethal, with a malformed cortex with impaired GABAergic interneuron migration. The *Dlx* genes appear to promote GABAergic interneuron migration via the repression of *Nrp2*, and downregulation of them causes a concurrent upregulation of *Nrp2*[206]. It is difficult to reconcile these contradictory findings of *Nrp2*'s effects on GABAergic interneuron placement, although the answer may have something to do with timing of expression or the importance of signalling cues being balanced. However a consistent result from investigation of *Dlx* null mutants is impairment of GABAergic interneuron development, a finding which is likely to be of relevance to schizophrenia.

3.8.13 NTRK2

NTRK2 encodes a protein named TrkB, a receptor tyrosine kinase which binds the neurotrophic ligands BDNF and NT-3. BDNF in particular has important roles in neuronal survival, dendritic outgrowth, and synaptic strengthening (LTP). Like all receptor tyrosine kinases TrkB must dimerise to become enzymatically active, and propagates intracellular signalling cascades via a series of phosphorylation events. Isoforms lacking the intracellular phosphorylation domain can inhibit full length isoforms and prevent the BDNF-TrkB effects of neurite outgrowth, calcium efflux, and gene expression. The full length isoform is well expressed throughout the CNS[207,208].

Two separate qPCRs were performed for this gene. One differentially expressed exon is the unique C terminal exon for a pair of isoforms f and b which truncate early and do not have an intracellular signalling domain (referred to as TrkB-T1 in PubMed). It is highly downregulated, while the gene itself is also significantly downregulated at the whole gene level. Two primer pairs were designed, one pair to

match transcripts of isoforms f and b, and one pair for a ubiquitous set of extracellular domain exons found in all isoforms.

One paper has shown that TrkB-T and TrkB have distinct roles in dendritic extension in the ferret visual cortex. Transfected full length TrkB promoted proximal dendritic branching, while truncated TrkB promoted distal dendrite extension[209]. However, even without distinct roles for different isoforms, TrkB oversignalling could have pathological consequences. BDNF signalling can make neurons vulnerable to excitotoxicity, despite its necessity for neuronal survival and avoidance of apoptosis. One hypothesis has proposed that TrkB signalling is enhanced in fragile X syndrome not as a pathological process, but as a compensatory one due to the lack of LTP[210]. This is a highly interesting idea, and it must be borne in mind that cells are living organisms employing homeostatic mechanisms. It will be difficult to discern whether enhanced TrkB signalling is a cause or effect of the pathologies we observe in the t(1;11) neurons, particularly as BDNF trafficking may also be affected due to *HAP1* dysregulation. Finally, BDNF signalling has been shown to alter the phosphorylation of DPYSL2 and DPYSL3, two proteins encoded by genes which have altered exon expression in our neurons and which appear to mediate some of the effects of BDNF-induced neurogenesis[211].

NTRK2 was indicated by Tiihonen *et al*. as being differentially expressed according to sex in iPSC-derived neurons. However, the putative change is downregulation in females, whereas here it is downregulated in translocation samples (which are disproportionately male). Therefore sex does not appear to be a factor in *NTRK2*'s significance.

3.8.14 PDYN

PDYN's product prodynorphin is the precursor protein for several opioid peptides, all of which are ligands for the K opioid receptor, which is encoded by a significantly downregulated gene (*OPRK1*). Both genes are expressed in many regions of the brain, including in certain layers of the cortex and especially in the prefrontal cortex. Within the medial prefrontal cortex, their products are seen at presynaptic axon terminals and activation decreases release of dopamine, serotonin, and other

neurotransmitters[212]. Diminished inhibition of dopaminergic transmission could have profound effects on our cultured neurons, especially in synergy with increased *DRD2* expression. The result would be overactive dopamine signalling. However it is notable that OPRK1 agonism (rather than antagonism) results in visual disturbances, dissociative effects, and other symptoms[213]. This apparently counterintuitive observation could be due to the fact that PDYN products mediate the release of a number of various neurotransmitters including GABA, and would be present in many parts of the brain during administration to human subjects. There is also some evidence suggesting that genetic variance in *PDYN* and *OPRK1* predisposes to some neuropsychiatric disorders[212]. The gene was significant in every WT vs t(1;11) DESeq2 analysis, and never significant in a t(1;11) vs t(1;11) or WT vs WT comparison.

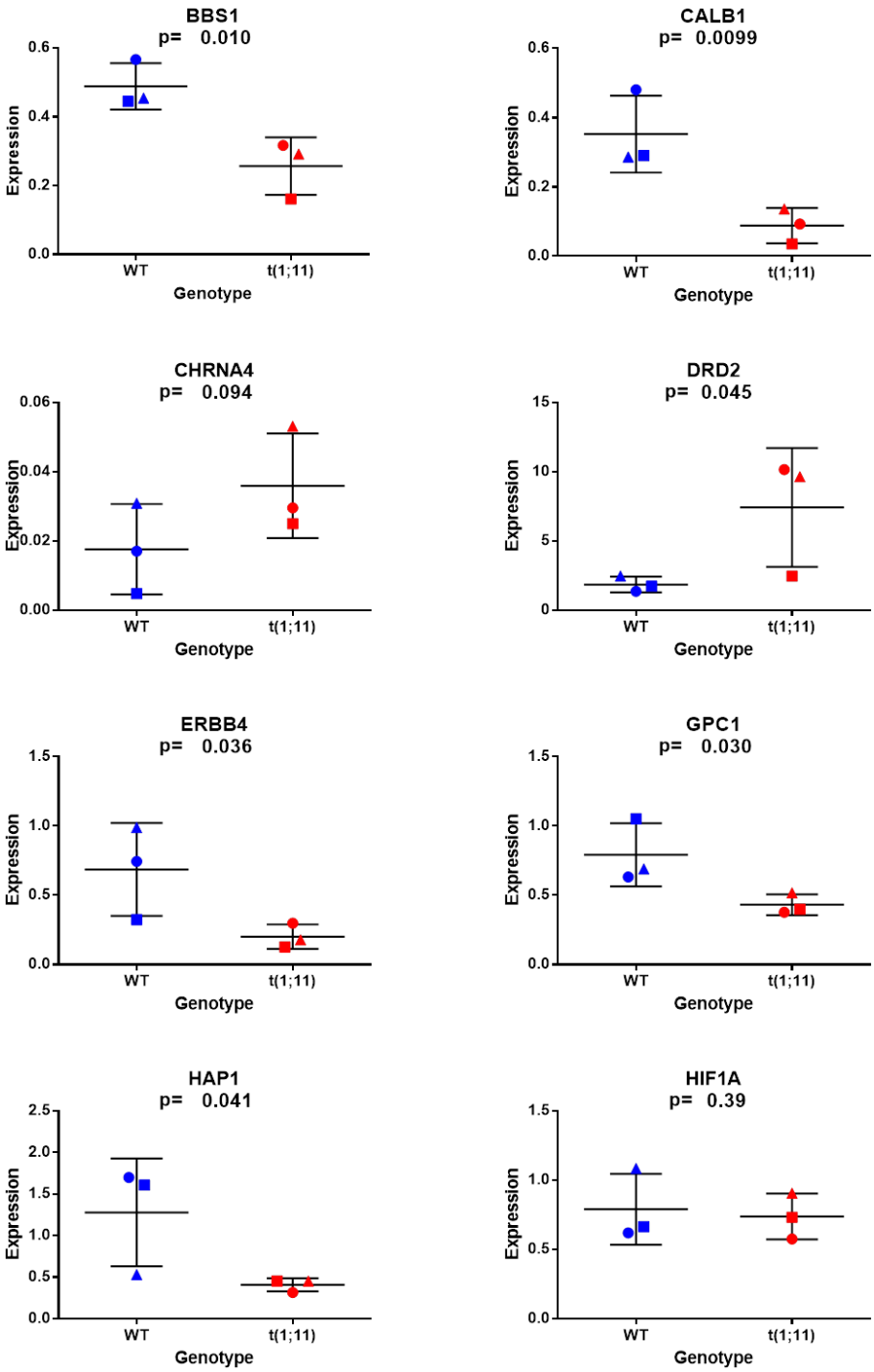
3.8.15 QKI

QKI encodes a protein named Quaking for its mouse mutant phenotype, which includes body tremor and CNS myelination deficits. Quaking protein binds RNA and integrates cell signalling by interacting with receptors. It contains a number of SH3 domains for this purpose and can be phosphorylated by Src family kinases, modulating its RNA binding ability[214]. The dysmyelination phenotype is evident even in mice which are just lacking two *QKI* isoforms solely in glial cells. Deletion of *QKI* is embryo lethal. Other mutants have cranial defects and altered development, or seizures and possible neurodegeneration[214,215]. Developmentally, the gene is initially expressed in many neural precursors but decreases in expression in neurons and increases in glial cells, with isoform specific expression changes. The number of RNAs Quaking binds is unknown but a possible consensus sequence has been generated and in post mortem human frontal cortex expression of *QKI* is correlated with the expression of a number of oligodendrocyte/myelin related genes, the expression of which is perturbed in schizophrenia[216]. *QKI* is downregulated in our neurons, leading to a few possible scenarios. Myelination deficits could lead to impaired neuronal activity. One scenario is that *QKI* is downregulated as the cell type in which it is highly expressed, oligodendrocytes, is lacking in our cell culture. However, our cell model should not contain a large number of oligodendrocytes,

although a reasonable expression of immature markers such as *PDGFRA* and *CSPG4* along with small amounts of mature oligodendrocyte markers such as *OLIG1*, *OLIG2*, *OLIG3*, and *SOX10* are all expressed. None of these are differentially expressed. It is also possible that the oligodendrocytes that are present will be non-functional due to poor *QKI* expression. Although *QKI* is expressed by neural precursors, our neurons are past this stage and the dysregulation should be unrelated to the typical downregulation seen in the precursor to neural cell fate change[217]. As I show in a later chapter, the cell proportions of our various t(1;11) models do not appear to be abnormal, implying the deficit is in existing oligodendrocytes rather than a lack of the cell type itself.

3.8.16 Results of RT-qPCR

qPCRs were carried out on all the aforementioned genes for the three samples of each of the six lines. Results were fitted to a standard curve and normalised to the scores for *BACT* expression as described in Materials and Methods. Owing to poor expression and the inability of primers to reliably detect it in all cell lines, the results of *PDYN* are not displayed. A summary of the results is shown in Table 8



and the expression plots are shown in Figure 16 and Figure 17. Samples were averaged by line before calculation of p values; a more conservative approach. Note that the results of *DRD2* have been previously published[70]. We can see that 10 of the 14 genes were confirmed at the RT-qPCR level, a good level of consistency. The gene

NRP2 was also near significant, an observation that reoccurred with the mouse orthologue in the mouse cortical heterozygous mutant samples. However it should be noted that for some genes the samples may be affected by sex imbalances, as the t(1;11) sample which was originally derived from a female patient has the t(1;11) value closest to the WT (all originally derived from females). This occurs with the genes *CALB1, GPC1*, *METRN* and *NRP2.* This is 4 of 14 genes, close to the expected number that would appear by chance (as one of the three t(1;11) samples must be closest to the WT samples). Nevertheless, should information emerge that any of these genes are also differentially expressed according to sex in iPSC-derived neurons, these two facts together are strong evidence in favour of the changes observed here being related to sex rather than to translocation status. In addition, the significance of *HAP1* is suspect and should be disregarded. This is because putative sex effects are in the same direction as the putative translocation effects.

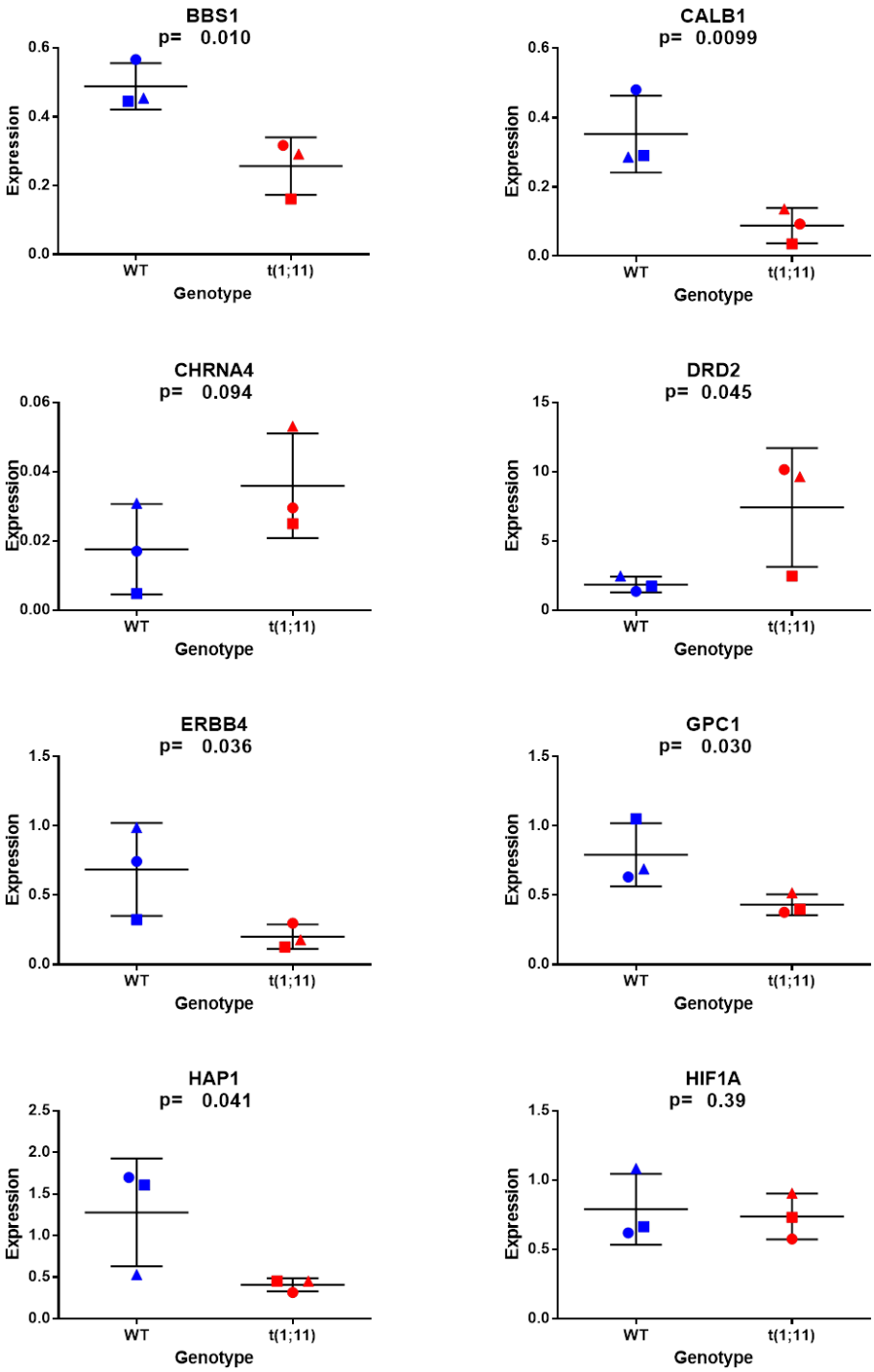


Figure 16. Expression plots for RT-qPCR results of *BBS1, CALB1, CHRNA4, DRD2, ERBB4, GPC1, HAP1,* and *HIF1A* differential whole gene expression. P values are given underneath each gene name. Each trio of neuronal samples from each line has been averaged. Lines indicate overall genotype average expression with smaller lines indicating one standard deviation above and below the mean. Colours indicate genotype and shape indicates line number. Blue=C line, Red=T line. Circles indicate 1, squares indicate 2, and triangles indicate 3. Lines C1, C2, C3, and T3 were derived from females.The RT-qPCR of *ERBB4* was carried out by Helen S. Torrance. N=3.

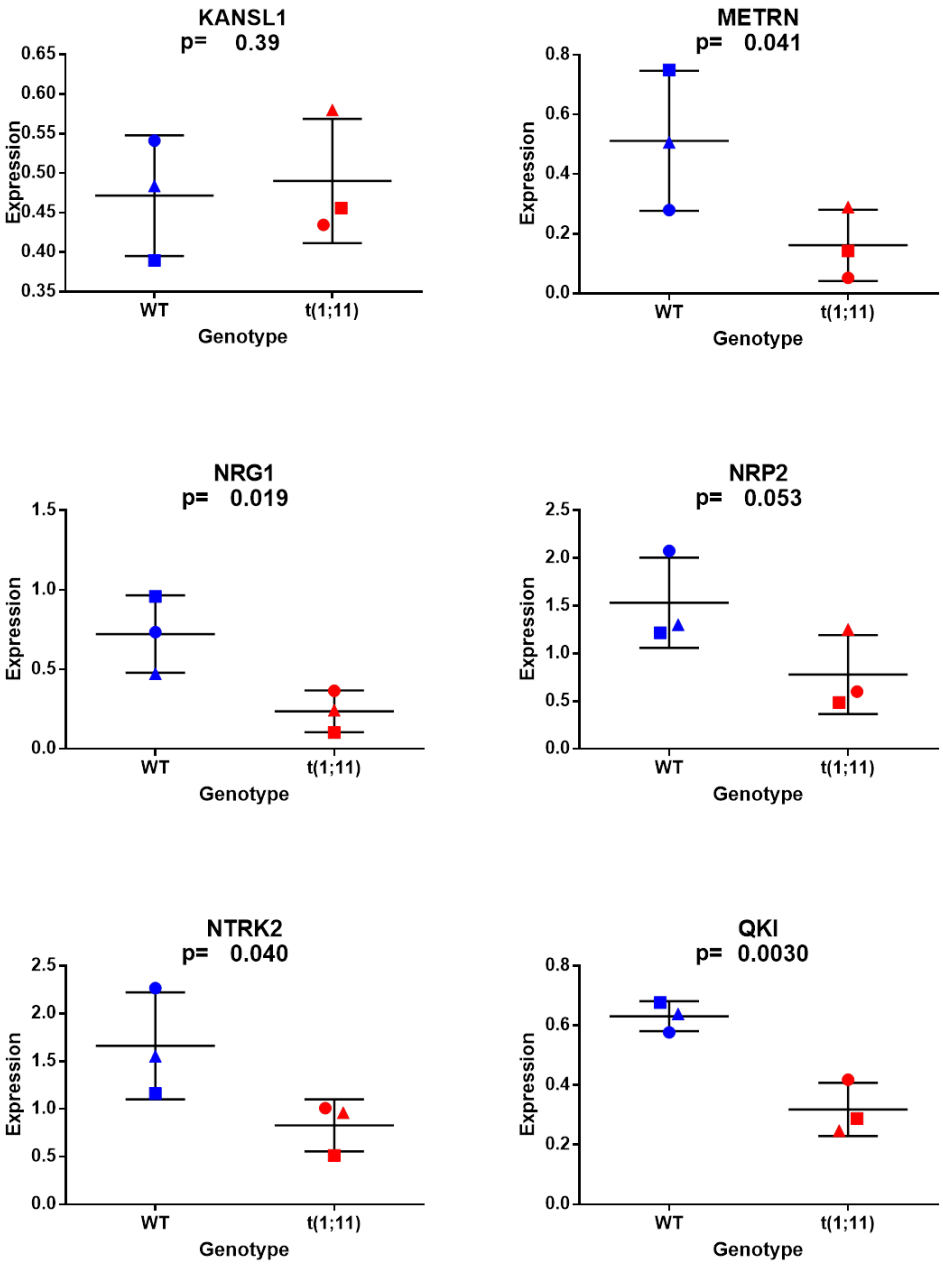


Figure 17. Expression plots for RT-qPCR results of *KANSL1, METRN, NRG1, NRP2, NTRK2 PDE4B* and *QKI* differential whole gene expression. P values are given underneath gene name. Each trio of neuronal samples from each line has been averaged. Lines indicate overall genotype average expression with smaller lines indicating one standard deviation above and below the mean. Colours indicate genotype and shape indicates line number. Blue=C line, Red=T line. .Circles indicate 1, squares indicate 2, and triangles indicate 3. Lines C1, C2, C3, and T3 were derived from females. The RT-qPCRs of was carried out by Helen S. Torrance. The primers used for the RT-qPCR of *NTRK2* shown here were designed to detect all isoforms of the gene. N=3.

RT-qPCR	P value	t(1;11) Fold change	Significance
BBS1	0.010	0.52	*
CALB1	0.0099	0.25	**
CHRNA4	0.09	2.03	
DRD2	0.045	3.98	*
ERBB4	0.036	0.29	*
GPC1	0.030	0.54	*
HAP1	0.041	0.32	*
HIF1A	0.39	0.93	
KANSL1	0.39	1.03	
METRN	0.041	0.32	*
NRG1	0.019	0.33	*
NRP2	0.053	0.50	
NTRK2	0.040	0.50	*
QKI	0.0030	0.50	**

Table 8. Summary of RT-qPCR gene level expression results for human neurons. * =p<0.05, **=p<0.01, ***=p<0.001. t(1;11) expression is given as a percentage of the WT expression rounded to the nearest %. P values calculated by t test, with each trio of individual differentiation of each iPSC-derived neuronal line being treated as a single averaged sample. The three genes *ERBB4, NRG1, PDE4B,* were analysed by Kirsty Millar and Helen S. Torrence. n=3.

A measure of how reliable the RNA-Seq findings are is given by how closely the RT-qPCR results track their respective samples' RNA-Seqs scores. In theory, these should be linearly related. I looked at the log2 foldchange between translocation and WT samples, seen in Figure 18.

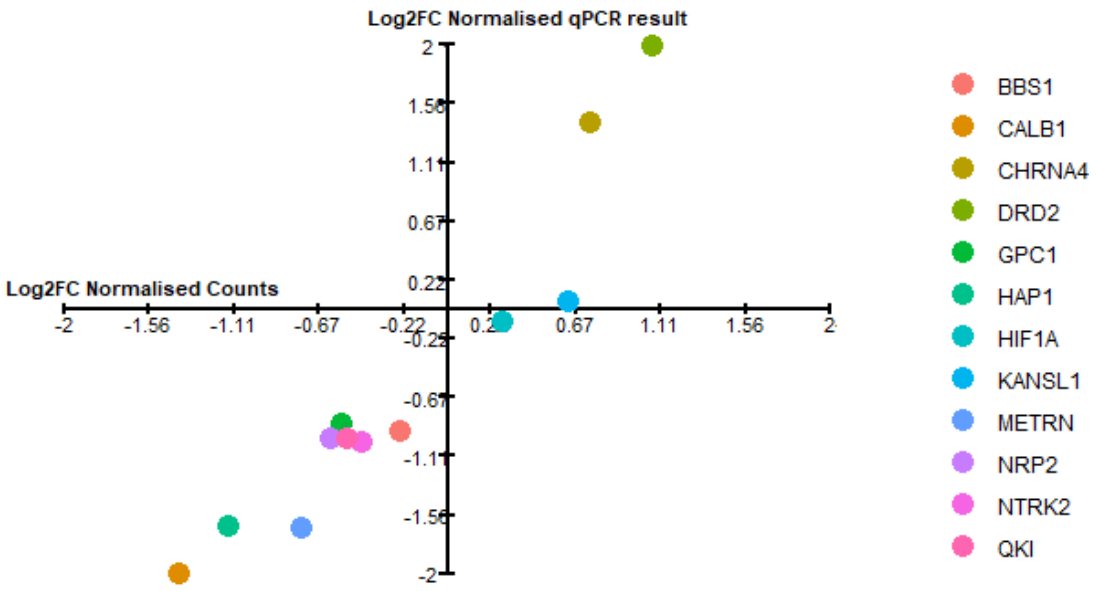


Figure 18. Graph displaying log2 fold change of normalised counts vs log2 fold change of normalised PCR score for each of the 12 genes I carried PCRs out on. Log2FC=Log2 fold change between WT and t(1;11) samples. Positive values indicate the t(1;11) samples have a higher score on average. Gene legends on right, differentiated by colour. Counts normalised using the "rlog" function in DESeq2 package. R^2=0.899

We can see that most genes show an apparently linear relationship between fold change of counts and fold change of qPCR score, showing that although there is indeed variance within genotypes, the overall trend is apparent (R^2=0.899). However, *HIF1A* and *KANSL1* show no apparent relationship. Indeed; these are the genes, along with *CHRNA4*, which were not found significant at the qPCR level. Individual qPCR results are displayed in Figure 16 and Figure 17. It should be noted that I designated one of the three C2 *CHRNA4* qPCR results as an outlier; it had a qPCR measurement which was 15 times greater than any other with a value 3.6 standard deviations from the mean, and if excluded changes the log2fold change from 1.41 to -0.68. The gene is non-significant in either case.

3.9 Exon level RT-qPCR

Genes were chosen for investigation on the same basis as the gene level candidates. All exons which belonged to genes implicated by the PGC GWAS or CNV studies were analysed as potential RT-qPCR candidates, as were exons belonging to genes encoding possible *DISC1* interactor. Subsequently exons were also inspected to see where in the gene they are. Exons that implicated a unique transcript, especially one that encoded a protein isoform with a known physiological role, were prioritised. N-terminal and C-terminal exons which implicated particular transcripts were also prioritised. A minority of potential targets were discarded due to the impossibility of designing specific primers, or because there is only one isoform in RefSeq according to the UCSC Genome Browser, genome hg19 (https://genome.ucsc.edu/). Of some use in selecting candidates was an approach searching for convergences between papers; a table displaying the numbers of genes significant between any pair of papers described in 3.7.1 as well as in the DESeq2 analysis is displayed as Table 9. Descriptions of each gene are given in turn and a table summarising the rationales is displayed as Table 10.

DEXSeq crossovers of (1;11) neurons	PGC1	PGC2-1	PGC2-2	PGC3	Gene changes	DI	B	W	*Der1* Het cortex	*Der1* Het hippocampus	Sx2mm D50	Sx8mm D50	Sx8wm D50
PGC1	8	4	5	0	0	0	1	1	1	0	0	0	1
PGC2-1	4	18	6	0	1	1	1	3	2	2	4	0	0
PGC2-2	5	6	15	0	1	1	1	2	2	0	3	0	1
PGC3	0	0	0	3	0	0	0	0	1	0	0	0	0
Gene changes	0	1	1	0	103	1	6	27	11	1	10	0	5
DI	0	1	1	0	1	8	0	1	2	1	2	0	0
B	1	1	1	0	6	0	26	6	2	0	1	0	1
W	1	3	2	0	27	1	6	99	13	0	13	2	0
Der1 Het cortex	1	2	2	1	11	2	2	13	80	1	91	2	1
Der1 Het hippocampus	0	2	0	0	1	1	0	0	1	8	0	0	0
Sx2mmD50	0	4	3	0	10	2	1	13	91	0	29	0	4
Sx8mmD50	0	0	0	0	0	0	0	2	2	0	0	3	1
Sx8wmd50	1	0	1	0	5	0	1	0	1	0	4	1	22

Table 9. Table of overlaps. Numbers represent the number of genes significant in our DEXSeq study of human neurons, in addition to the two papers in the corresponding row and column. Grey blocks indicate genes from only one paper (the same row and column index). Abbreviations as in 3.7.1.

	Expression as % of WT	Padj	Human gene changes	Mouse cortex changes	Mouse hippocampus changes	PGC1	PGC2.1	PGC2.2	PGC3	DI	B	W	Sx2mmd50	Sx8mmd50	Sx8wmd50
DLG2	351.8	1.83E-02													
DPYSL2	118.4	1.65E-04		Gene and exon level											
DPYSL3	181.5	4.28E-02		Exon level											
DVL1	115.0	1.25E-02		Exon level											
GRIA4	117.7	3.59E-02													
NTRK2	72.0	6.70E-04	Yes	Exon level											
NTRK3	137.0	3.78E-02		Exon level								TRUE			
SHTN1	238.1	2.88E-03										TRUE			
SLC12A2	158.7	1.15E-02		Exon level											

Table 10. Highlighted information about candidate exons selected for qPCR. TRUE indicates that the gene is differentially expressed in the model of interest. Paper abbreviations are as in 3.7.1.

3.9.1 DLG2

This gene encodes the protein PSD-93, and a total of three exons are differentially expressed. One leads to a region annotated by UCSC as being intronic, while the other two are a unique N terminal exon for a shorter isoform and an exon found within but not exclusive to said isoform. The isoform is described in PubMed as being the "alpha" isoform. PSD-93 is a synaptic protein that helps the clustering of various signalling proteins. It associates with other synaptic proteins such as the neuroligins, and associates with NMDARs, increasing their surface expression[218,219]. It also forms supercomplexes with NMDARs and PSD-95[220]. It is henceforth important to LTP and also associates with AMPARs via proteins like stargazin, which facilitate the increased synaptic strength following LTP[221]. Mutant mice display impaired LTP[222]. Furthermore, CNVs deleting *DLG2* are associated with schizophrenia[35]. Although not all the functional effects of the various isoforms are fully understood as of yet, *DLG2α* specifically being upregulated is of interest and should be investigated further.

3.9.2 DPYSL2

A number of exons are also dysregulated in the homologous mouse gene in the cortical heterozygous model. The pattern of exon dysregulation is the same as in the related gene *DPYSL3* described in the next section; one N terminal exon is upregulated and the alternative one is downregulated. The upregulated exon is only found in an isoform containing a number of metal binding residues, while the downregulated one is only found in isoforms lacking these residues. *DPYSL2* binds to tubulin molecules to promote microtubule assembly, and promotes axonal outgrowth in hippocampal cell culture[223]. Indeed, *DPYSL2* overexpression promotes the formation of excess axons, a phenotype shared by the overexpression of *SHTN1* (see 3.9.9) [224]. It is also a possible DISC1 interactor, and may have mutations linked to schizophrenia[73,223]. Finally, knockdown of *Dpysl2* and *Dpysl3* leads to motor neuron misplacement, a phenotype also produced by *Cdk5* and rescued by phosphomimetic Dpysl2[225]. It also incorporates signalling from GSK-3β and fails to inhibit axonal growth unless phosphorylated by it[211], while antipsychotic drugs inhibit the phosphorylation of this site, conceivably via GSK-3β[226]. It is clear that

DPYSL2 is a gene of importance to axonal development, kinase signalling, and is even a possible DISC1 interactor. The rationale for investigating further was clear, especially given that the related gene *DPYSL3* shows the same pattern of change. A PCR was carried out to detect changes in the upregulated exon.

3.9.3 DPYSL3

The protein encoded is DPYSL3, or CRMP4 (Collapsin response mediator protein 4). It is a possible DISC1 interactor according to a Y2H screen[73]. It is localised in neurites and axonal growth cones, where it appears to oligomerise and bundle F-actin[227]. It is also phosphorylated by GSK-3β, as is DPYSL2, and the phosphorylation of this site can be blocked by antipsychotic drugs[211,226]. Expression is at its highest during peak axonal growth, and mutations in the highly conserved nematode homologue cause severe axon defects in all neurons[228]. Cleavage of DPYSL3 post excitotoxic NMDA signalling prevents oligomerization and therefore actin bundling, which could have implications for the maintenance of axonal growth cones[227]. The exon which is upregulated corresponds to the N-terminal exon of an isoform which possesses a number of metal ion binding residues, while the alternative isoform without these residues has its N terminal exon downregulated. The corresponding mouse gene also has some differentially expressed exons. Given the alternating N terminal exon pattern and the fact that the isoforms have known differences, this is a strong candidate for RT-qPCR. Both DPYSL2 and DPYSL3 appear to incorporate signalling from multiple kinases known to promote neuronal development, which can be modulated by antipsychotic drugs, and are both involved in axonal growth. They present as a pair of highly interesting candidates. A PCR was carried out to detect transcripts containing the upregulated exon.

3.9.4 DVL1

Two isoforms of the encoded protein, Dishevelled-1, are described in the UCSC genome browser and are differentiated by differing versions of a central exon. The unique segment of one version is upregulated. *DVL1* is a highly conserved and broadly expressed developmental gene important for Wnt signalling and in establishing planar cell polarity. A null *Dvl1* mouse model was viable but exhibited

reduced sociality, altered behaviour, and reduced prepulse inhibition in response to acoustic or tactile startles[229]. Given the similarity to some known schizophrenia phenotypes, as well as the known links between *DISC1* and Wnt signalling, *DVL1* presented as a very interesting candidate for qPCR.

3.9.5 GRIA4

A number of exons are differentially expressed in this gene. One in is a C-terminal exon found only in isoform 3 of the gene. It is important to note that the "flip" and "flop" exons are not altered. *GRIA4* encodes GluR4, a glutamate receptor subunit found in AMPARs. AMPAR expression and trafficking is believed to underlie synaptic plasticity and from this many physiological aspects such as learning and memory[230]. The expression of an alternative isoform of *GRIA4* may have some biological impact on these functions, although the isoform is currently not well characterised in terms of a unique biological role. Nevertheless given the relevance of the gene it makes sense to investigate expression further.

3.9.6 NTRK2

The rationale for investigating *NTRK2* is described in detail in 3.8.13. The primers used in the following analysis were designed to only detect isoforms with a unique C terminal exon resulting in an early truncation of the protein and a corresponding lack of an intracellular signalling domain.

3.9.7 NTRK3

Like *NTRK2*, *NTRK3* encodes a receptor tyrosine kinase that dimerises to form active signalling complexes. As with *NTRK2*, the exons found only in truncated isoforms lacking the intracellular signalling domain are downregulated. *NTRK3* encodes a protein called TrkC which is the receptor for the neurotrophin NT-3. The truncated isoforms may inhibit the full-length ones, but importantly appear to have roles of their own. The extracellular domain of TrkC can bind both NT-3 and PTPσ, resulting in formation of glutamatergic excitatory synapses. Overexpression of the truncated isoforms results in increases in VGLUT1 expression but not VGAT, while knockdown of the gene results in decreased VGLUT1-PSD95 co-localisation. This is

rescuable by expression of the truncated isoforms. This implies a special role for the truncated isoforms of TrkC, so it is highly interesting to find these downregulated. The ratios of both TrkB and TrkC truncated to non-truncated isoforms increase at the peak of synaptogenesis[231]. Overexpression of the full length isoforms in mice results in behavioural abnormalities including abnormal responses to threats (increase flight response, frozen response, and decreased approach)[232]. The decrease in the truncated isoforms' unique exon is therefore very interesting, and might imply deficits in excitatory synapse formation, conceivably linked to altered behaviour.

3.9.8 PDE4B

The analysis of this gene was carried out by Helen Torrance and Kirsty Millar, but is reported here due to its relevance. Results were previously described in Malavasi *et al.* 2018[70] .

PDE4B is another gene linked to psychiatric disease, with a balanced translocation in the gene segregating with schizophrenia risk[74]. It is crucial for cAMP regulation, which itself is vital in several neural processes, such as memory[74,85]. PDE4B interacts with DISC1, and its inhibitor rolipram is a prototypical antidepressant. Mutations in the corresponding Drosophila gene cause learning deficits. PDE4B has also been shown to have a role in the functions of a number of other DISC1 interactors such as LIS1, NDE1, NDEL1, the genes of which have been found mutated in cases of lissencephaly (a developmental brain malformation) or are linked to crucial processes such as neurite outgrowth. Interaction with DISC1 appears to inhibit PDE4B's phosphodiesterase ability, while the subsequently high cAMP levels stimulate PKA-mediated phosphorylation of NDE1. NDE1 phosphodead NS-1 cells have inferior neurite outgrowth to wild type[84]. We can see that control of PDE4B via DISC1 is important for neural processes, yet we know from Drosophila and the PDE4B translocation that insufficiency is also problematic. The gene therefore presented as an interesting candidate for RT-qPCR.

3.9.9 SHTN1

This gene is also known as *KIAA1598* and produces a protein called Shootin1. It has a role in axonal sprouting. Studies involving rat hippocampal neurons showed that Shootin1 is initially expressed in neurites sprouting from cells. It fluctuates in expression but eventually is only expressed in one neurite, which is destined to become the neuron's axon. Knockdown delays axon formation, while overexpression results in accelerated accumulation of the protein in neurites and a corresponding chance to form multiple axons. It persists in the axonal growth cone where it activates PI3K. The authors theorised a model where Shootin1 is actively trafficked to neurite termini and passively diffuses back. The longer the neurite, the longer Shootin1 stimulated activity continues, and the greater the corresponding effect on neurite extension and axon formation. Shootin1 presence therefore determines axonal identity in a self-propelling loop[233]. Another paper showed that the formation of multiple axons caused by *Cdkl5* overexpression in mouse could be mitigated by *Shtn1* knockdown and that the two genes are found expressed in cortical neurons[234]. It also interacts with cortactin, an actin bundling protein. This is enhanced by an axonal chemoattractant, netrin-1[235]. Although data on the human isoforms is sparse, the isoforms appear to be analogous in the mouse and rat. The differentially expressed exon is downregulated in the rat PC12 line, and the isoform containing the analogous exon is constitutively expressed. The isoform without the analogous exon is expressed after NGF signalling and is necessary for subsequent neurite extension. Protein expression of both isoforms subsequently decreases[236]. Given its evident importance in neurite outgrowth, as well as a verified different expression pattern for the isoform containing the exon (in the rat PC12 line at least), I deemed *SHTN1* worthy of investigation via qRT-PCR.

3.9.10 SLC12A2

The product of this gene is known as NKCC1, a protein which is highly expressed in developing cortex but declines in expression over the first year of life to a baseline level equivalent to that of the adult cortex. This occurs in both rat and human[237]. The protein is a chloride transporter maintaining high levels of intracellular Cl^-, which when coupled with low levels of intracellular K^+ results in a high transmembrane

potential. Upon activation of GABA receptors the efflux of charged chloride results in neuronal depolarization and firing. The reverse effect is seen in adult tissue. High KCC2 expression (a potassium importer) and low NKCC1 expression keeps intracellular Cl^- low and K^+ high. GABAR activation here allows chloride ion influx resulting in hyperpolarization. Therefore in adult tissue GABA acts as an inhibitory neurotransmitter rather than an excitatory one. This GABAergic switch is a key developmental step and neonates who have yet to complete the switch are vulnerable to GABA-mediated excitatory neuronal activity causing seizures[238].

The differentially expressed exon is found in one isoform, NKCC1A, and not the other, NKCC1B, although both form functional transporters. NKCC1A is expressed in both foetal and adult prefrontal cortex and appears to follow a general trend of increasing in expression postnatally, being very poorly expressed foetally[239]. In the mouse, the excitatory activity of GABA (and high Nkcc1 expression) is required for immature hippocampal neurons to develop properly. Nkcc1 knockdown or Kcc2 upregulation both result in defective dendrite growth and synapse formation. *Disc1* knockdown, meanwhile, enhances dendrite outgrowth at this stage, which can be undone by concurrent Nkcc1 knockdown preventing excitatory GABAergic signalling. Complementary experiments showed that GABAR agonists or inhibitors of GABA degradation enzymes further enhanced the *Disc1* knockdown induced dendrite outgrowth. The effect of *Disc1* knockdown is eventually lost, but can be regained if Kcc2 expression is kept low[240]. Although NKCC1 isoforms have not yet been fully characterised, NKCC1A downregulation could indicate that our neuronal model is developing aberrantly, with a possible early switch in the role of GABA which would have severe consequences.

3.9.11 Results of RT-qPCR

qPCRs were carried out on all the aforementioned genes for the three samples of each of the six lines. Results were fitted to a standard curve and normalised to the scores for *BACT* expression. A summary of the results is given in and the expression plots are given in Figure 19 and Figure 20.

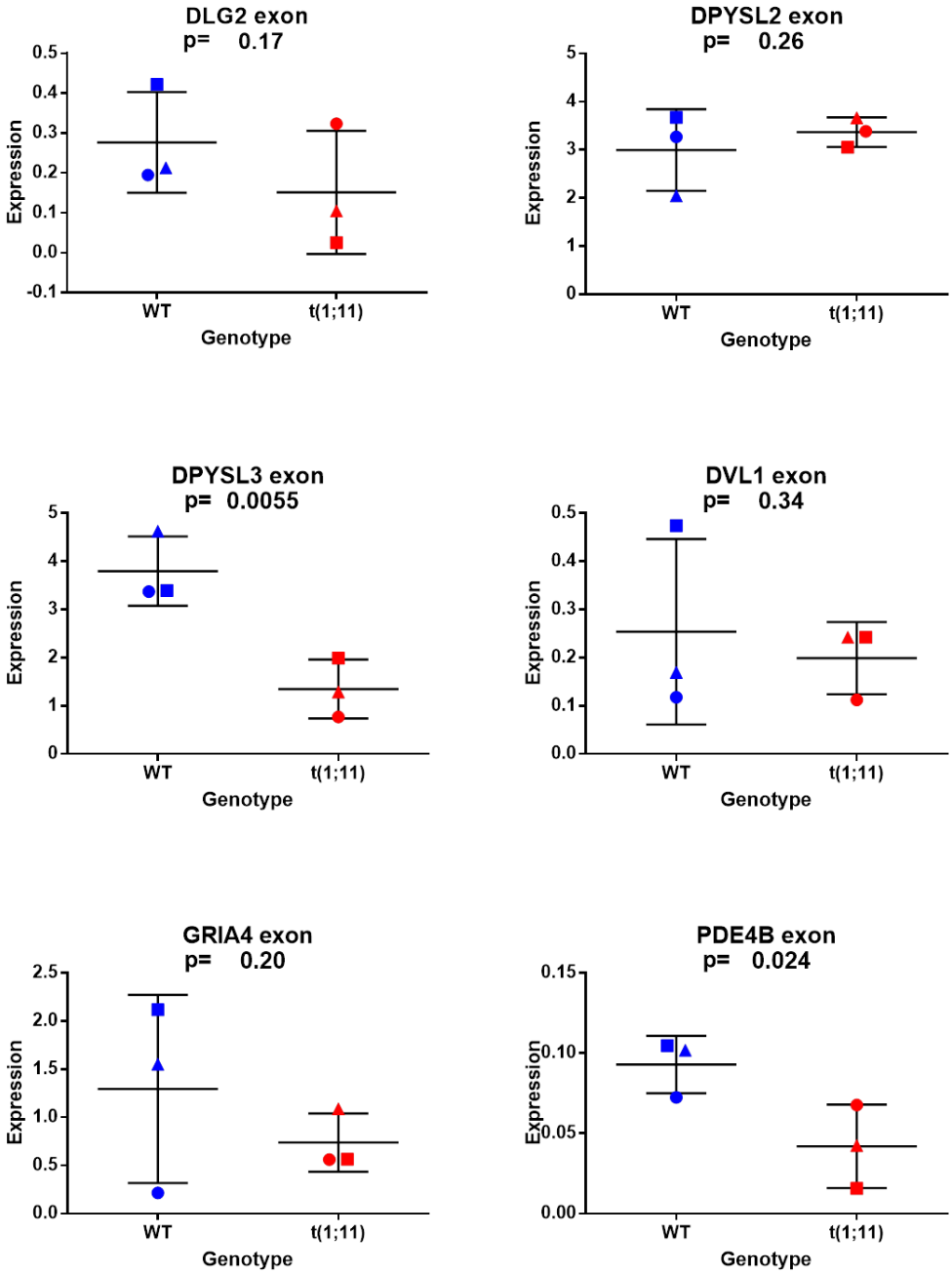


Figure 19. Expression plots for RT-qPCR results of *DLG2, DPYSL2, DPYSL3, DVL1*,*GRIA4* and *PDE4B* differential exon expression. P values are given underneath each gene name. Lines indicate overall genotype average expression with smaller lines indicating one standard deviation above and below the mean. Colours indicate genotype and shape indicates line number. Blue=C line, Red=T line. Circles indicate 1, squares indicate 2, and triangles indicate 3. Each trio of neuronal samples from each line has been averaged. N=3.

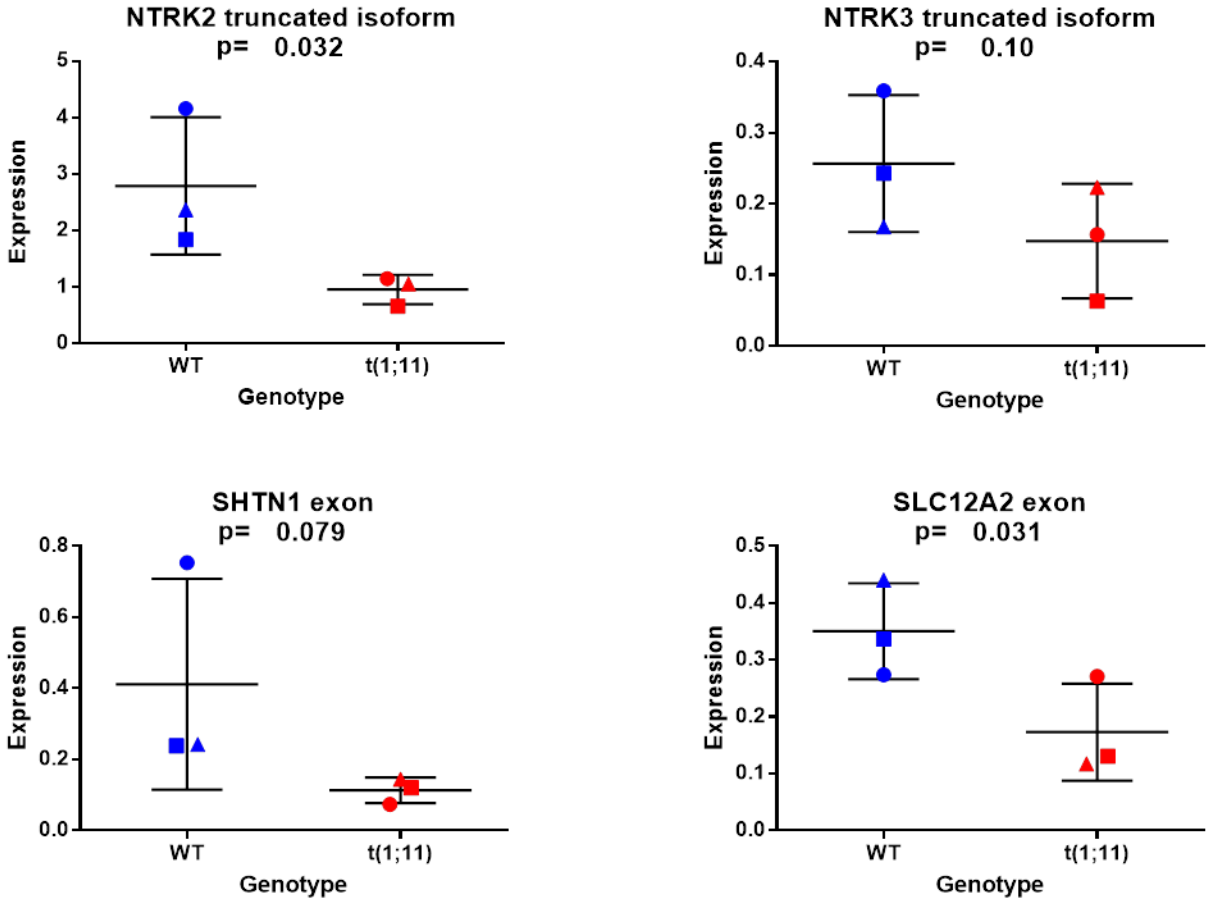


Figure 20. Expression plots for RT-qPCR results of *NTRK2, NTRK3, SHTN1* and *SLC12A2* differential exon expression. P values are given underneath each gene name. Each trio of neuronal samples from each line has been averaged. Lines indicate overall genotype average expression with smaller lines indicating one standard deviation above and below the mean. Colours indicate genotype and shape indicates line number. Blue=C line, Red=T line. Circles indicate 1, squares indicate 2, and triangles indicate 3.The primers used for the qRT-PCR of *NTRK2* shown here were designed to only detect the f and b isoforms, which truncate early.n=3.

RT-qPCR results	Exon identity	P value	t(1:11) Fold Change	Significance
DPYSL3	Alternative N-terminal exon	0.0054	0.35	**
NTRK2	Alternative C-terminal exon	0.031	0.34	*
PDE4B	Central exon found in isoforms with an alternative promoter start	0.024	0.45	*
SLC12A2	Alternative central exon	0.031	0.49	*

Table 11. Summary of RT-qPCR exon level expression results. * =p<0.05, **=p<0.01, ***=p<0.001. t(1;11) expression is given as a percentage of the WT expression. Only significant results are displayed.

We can see that a much lower proportion of the exon level changes have been confirmed; four out of ten as opposed to 10 out of 14 for the gene level changes. The reason is difficult to discern but may be due to the fact the counts for any exon are likely to be much lower than the counts mapping to a gene, given the vastly increased size of a gene compared to an exon. The lower number of counts may allow chance variation a greater role, resulting in falsely indicated differentially expressed genes. As before it should be noted that for some genes the samples, if judged by sex, cluster together (the red triangle with all blue shapes). This occurs with the qPCRs for exons in genes *NTRK3* and *SHTN1*. This is 2 of 10 genes, lower than the expected number that would appear by chance (as one of the three t(1;11) samples must be closest to the WT samples). As before the same conclusion applies. If any of these exons are differentially expressed according to sex in iPSC-derived neurons, this is evidence in favour of the changes observed here being related to sex rather than to translocation status.

As with the gene level qPCRs, I looked at how closely the RT-qPCR results track their respective samples' RNA-Seqs scores. Figure 21 shows the relation between qPCR score and normalised count for each of the 18 samples for 9 genes

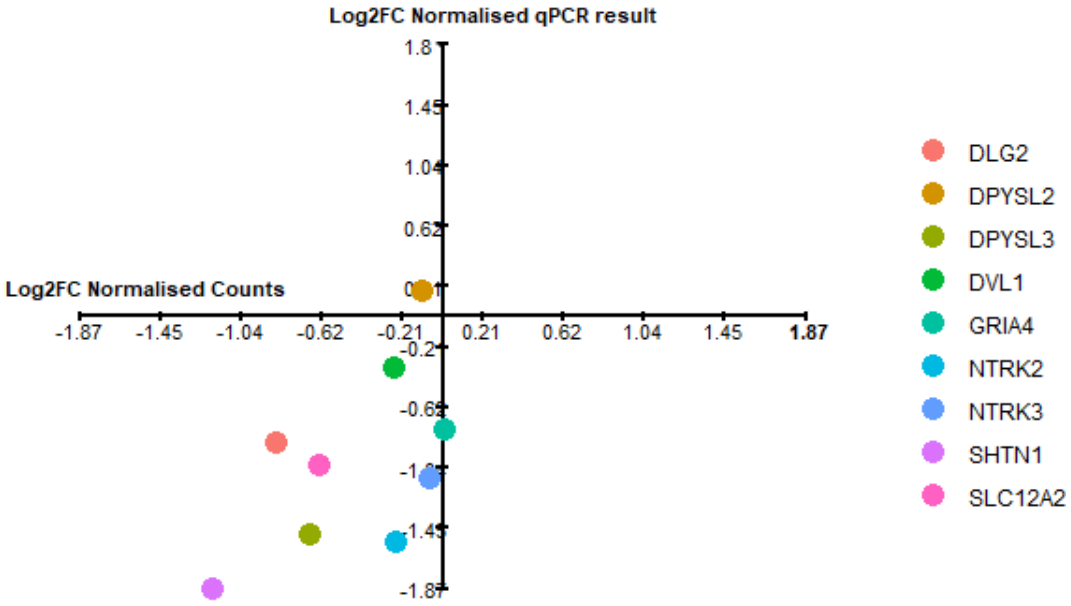


Figure 21. Graph displaying log2 fold change of normalised counts vs log2 fold change of normalised PCR score for each of the 8 exons I carried PCRs out on. Log2FC=Log2 fold change between WT and t(1;11) samples. Positive values indicate the t(1;11) samples have a higher score on average. Gene legends on right, differentiated by colour. Counts normalised using the "rlog" function in DESeq2 package.

There appears to be a more linear relationship between the qPCR and normalised count scores when samples are averaged, as in the gene level analysis. There is more scatter, and correspondingly less significant genes. However the overall linearity indicates that the changes may be genuine, and that the relatively small number of samples, as well as the low magnitude of the changes, may be the reason for the lack of significance at the qPCR level. As always, larger sample sizes would be ideal. Individual qPCR results are shown in Figure 19 and Figure 20.

3.9.11.1 NTRK2 results

The qPCR results suggest that both "all transcripts" and truncated transcripts of *NTRK2* are downregulated. It is entirely possible that the change in one is responsible for the apparent change in the other. An estimation of the ratio of truncated NTRK2 to full length NTRK2 can be obtained by looking at RNA-Seq

counts for unique exons. The C-terminal exon of the fully truncated isoforms b and f has an average of over 19,000 reads per WT sample, while the averages of the C-terminal exons for partial length isoforms d and e and full length isoforms a and c are approximately 2,000 and 2,100, respectively. The samples do not display drastically different counts for the partial and full length C-terminal exons, while those for the truncated C-terminal exon have been confirmed by qPCR as significantly downregulated to approximately one third of the levels of the WT samples.

It appears likely that the change at the "whole gene" level is driven mostly, if not entirely, by the downregulation of the truncated transcripts, given that it is the major isoform and is confirmed as downregulated. The reduction in the truncated isoforms is approximately 70%, while "all isoforms" are reduced by approximately 50%. The hypothesis that the change in "all isoforms" is driven entirely by a 70% reduction in the truncated isoform bears some relation to the observed ratios of the average counts of each C-terminal exon, although there may be some minor changes in partial length transcripts. See Table 12.

C-terminal exon counts average	Truncated NM_001007097.2 and NM_001291937.1	Partial Length NM_001018065.2 and NM_001018066.2	Full Length NM_006180.4 and NM_001018064.2	Total counts	qPCR of ubiquitous exon	qPCR Truncated
WT	19644	1977	2109	23730	100%	100%
t(1;11)	6100	1251	2011	9362	50%	30%

Table 12. Examination of counts of C-terminal exons unique to each isoform of *NTR2*. Note that the Total counts decreases by 14,000 counts, while Trk-T1 decreases by 13,000. The truncated isoforms are referred to as "b" and "f", the partial length isoforms as "d" and "e", and the full length isoforms as "a" and "c". Note that the accession numbers are not necessarily an exhaustive list of all accession numbers that may contain the exon. Only 6 transcripts have a verified mRNA in Ensembl, 2 have neither an mRNA nor an EST and are similar to isoforms "a" and "c".

We can conclude that the ratio of truncated: full NTRK2 transcripts is lower in the t(1;11) samples. This would imply that TrkB signalling is enhanced, particularly as the full length isoforms appear to be relatively unchanged.

3.10 Discussion

To summarise, over 1,200 genes appear to be differentially expressed between the WT and t(1;11) human neurons. There is little evidence to suggest regional effects of the translocation, but *DISC1* displays the expected phenotype of cross-breakpoint reads being halved. Differential expression at the protein level was also confirmed by other researchers[70]. The GO terms overrepresented among the differentially expressed genes relate to psychiatric disease and there is extensive overlap with other researchers using both iPSC-derived neuronal models and GWAS/CNVs. A number of candidate genes have been confirmed at the qRT-PCR level and the results are generally in good alignment with the RNA-Seq.

3.10.1 Evaluating the evidence

There are consistent areas of interest which are implicated by this investigation into t(1;11) pathology. These areas have been implicated by GO terms, have associated genes which are relevant to disease and/or are DISC1 interactors, and have associated genes confirmed at the RT-qPCR level.

- Intracellular trafficking. Although *HAP1* may be sex regulated,other differentially expressed genes include *KIF1A*, *MYO10*, *MYH11* encoding kinesin-related and myosin proteins. *DISC1* has already been associated with this function via recent papers (see Introduction), as well as interactome studies[73]. Potential DISC1 interactors include dynactin, NDEL1, FEZ1, and KIF1B[241]. DISC1's confirmed trafficking of $GABA_A$Rs is also relevant. A review by Devine *et al.* also suggested that HAP1 and DISC1 might cooperate in the trafficking of AMPARs[241]. This balance between excitatory and inhibitory synaptic activity is relevant to cell-specific disturbed GO terms, as I elaborate in the Discussion. Trafficking of mitochondria is important for neuronal energy demands, as receptor trafficking is for synaptic formation. Both are essential to neuronal function.

- Neuronal migration and placement, including *BBS1* (and other BBS genes, not subjected to qPCR), *NRG1, ERBB4.* A well established *DISC1* function (see Introduction).

- Neuronal developmental in terms of timing of developmental switches and dendritic outgrowth, including *SLC12A2, SHTN1, GPC1* and *METRN* , *PDE4B, DPYSL2/3* respectively. GOrilla Functions which are overrepresented are heavily cytoskeletal in nature; including actin nucleators such as *SPIRE1*. This has relevance as the cytoskeleton undergoes reorganisation to form axonal growth cones. *SHTN1* and *DPYSL2/3* have functions directly related to this.

- Synaptic activity including altered plasticity and strengthening including *NTRK2, DRD2.* As elaborated in the Discussion and Introduction, altered synaptic plasticity is highly relevant to psychiatric illness. Our research group (published as Malavasi *et al.*) showed that the *Der1* mice have altered PSD95 distribution[70]. Differentially expressed genes include those related to neurotransmtter release and receptors, such as *SYNJ2, SYT4, SYT6, DNM2* (synaptojanin, two synaptotagmins, and dynamin), *GLRA1, GABRD,GRIN2D* (receptors for glycine, GABA, and glutamate).

- Particularly interesting is the emergence of paired genes such as *OPRK1+PDYN, ERBB4 + NRG1, SHANK1 + HOMER2* (possibly), and several examples of paired developmental cue genes such as the semaphorins, netrins, and plexins.

Other interesting genes which I did not carry out a qPCR on include *APP, BBS2, BBS5, BSN, VAX1, VAX2,* amongst others. Many of these are synaptic, are Hox genes which play a role in development, are trafficking molecules, or are actin/microtubule organisers which will alter dendritic outgrowth. These are all functions of great relevance to psychiatric disease aetiology and are further explored in tandem with the results of other chapters in the Discussion.

4 GENERATION AND INITIAL ANALYSIS OF MOUSE RNA-SEQ DATA

4.1 Generation and initial analysis of mouse cortical RNA-Seq data

4.1.1 Introduction

The mouse data was for the most part analysed using the same methods as the human data, utilising differential expression analyses and using GOrilla to analyse gene ontology. Local expression changes were not analysed as there is no translocation. The cortical samples included much of the brain excepting the hippocampus. Philippe Gautier found that one of the WT mice was an outlier in RNA-Seq profiles; it was removed from the analyses. A second WT mouse was randomly chosen and also removed to balance the sex ratios.

4.1.2 WT vs heterozygous

4.1.2.1 DESeq2

The six WT and eight heterozygous samples were analysed using DESeq2. Since sex ratios were balanced, X and Y reads were not removed. A total of 2,112 genes were described as significantly differentially expressed between the two sample groups with a BaseMean at least half that of *Disc1*'s. As shown by the volcano plot in Figure 22, most of these genes showed only a mild difference in fold change with very few showing a twofold change. This is particularly evident if this volcano plot is contrasted with that of Figure 5. I also produced a heatmap of all differentially expressed genes between the wildtype and heterozygous mouse cortices, seen in Figure 23. In general, genotypes cluster together but there is an exception.

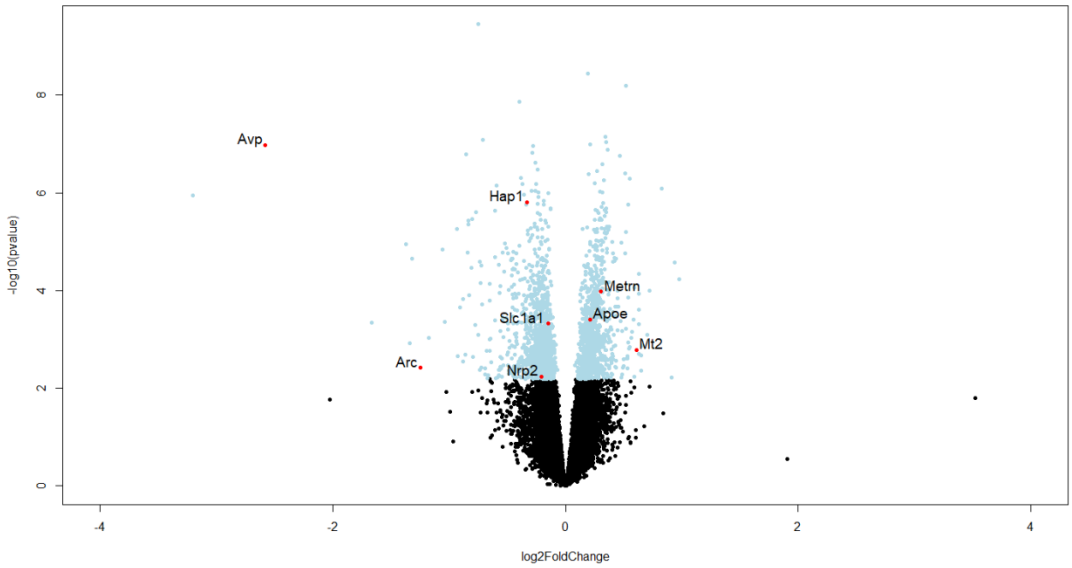


Figure 22. A volcano plot of the cortical DESeq2 for all genes with BaseMean>44 (half that of *Disc1*). X-axis represents the log2 fold change between WT and heterozygous *Der1* lines, while the Y axis represents significance (-log base 10 of p value). Black dots have an adjusted p value above 0.05, blue dots are significant with an adjusted value below 0.05. Red dots with labels represent genes for which a qPCR was carried out.

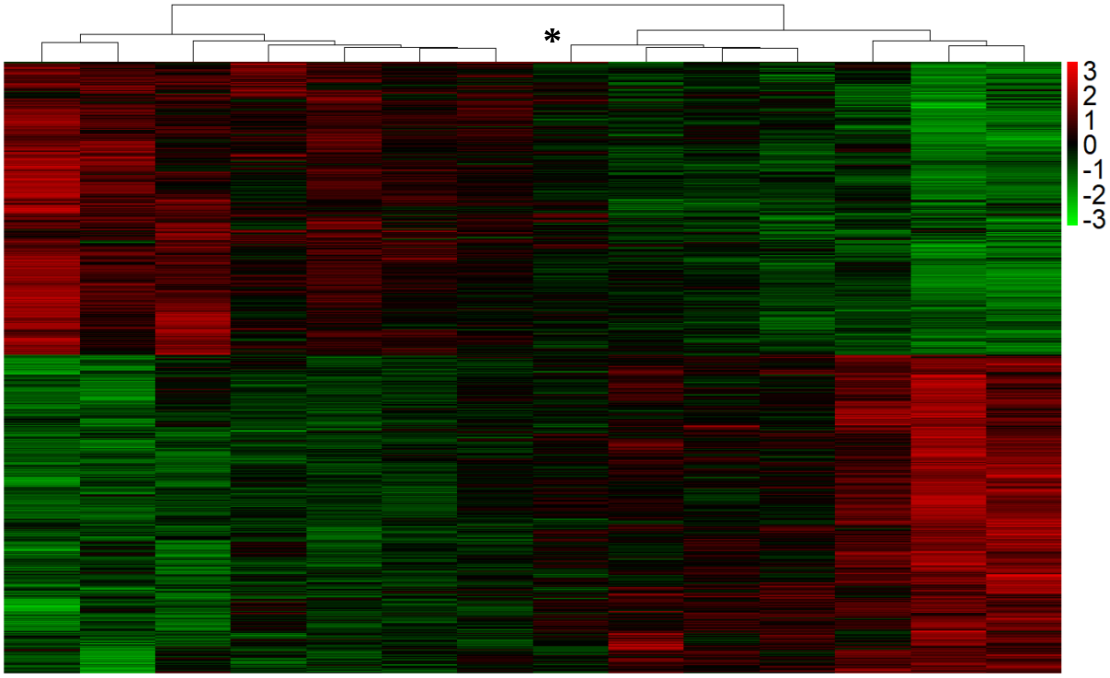


Figure 23. Heatmap of all differentially expressed genes with p<0.05 and at least half the expression of *Disc1*. Counts were normalised using the "rlog" function, which transforms counts to the log2 scale, normalises for library size, and minimises variation in poorly expressed genes. They were then antilogged, and changed to z scores by gene before generation of the heatmap. Red indicates z score above the mean, green indicates z scores below the mean. Each row is a gene. The groups do not cleanly separate by genotype; 7 heterozygotes are the left cluster, while the right cluster consists of 6 WTs and one heterozygote designated by an asterisk (*).

4.1.2.2 DEXSeq

At adjusted p value<0.05, a total of 8,993 exons were differentially expressed, found in 3,570 different genes.

4.1.2.3 GOrilla

As in the human analysis in 3.6, GOrilla was utilised to analyse gene ontologies overrepresented among the differentially expressed genes. This analysis was carried out by Marion Bonneau and Kirsty Millar and is described in the corresponding thesis.

4.1.2.4 Comparison to other papers

An analysis identical to that described in 3.7. was carried out by Kirsty Millar using the PGC and CNV papers which found significant overlaps between PGC-1 and the list of differentially expressed genes. I carried out an analysis utilising the other

papers mentioned in section 3.7. I also searched for RNA-Seq experiments utilising *Disc1* mouse models but did not find any suitable published papers. In all cases, the genes implicated by DESeq2 (BaseMean>10, $p<0.05$) and DEXSeq ($p<0.05$) were combined, and duplicate gene caused by multiple significant exons, as well as duplicates caused by overlap between DESeq2 and DEXSeq, were removed. Where DEXSeq assigned one exon to multiple genes, I used UCSC to manually assign the exon to the correct genes. Since some genes will have diverged in function, significance in overlap between the genes implicated by the *Disc1* mutation and those implicated in the various papers is less informative than in the human translocation study. Nevertheless many genes will have retained their functions and may be of relevance to the processes disturbed by psychiatric illness. A summary of the results is shown in Table 13.

Cortex Heterozygous			
Paper	Number of genes	P value	Genes of interest
DI	18	4.3E-02	*Dpysl2,*
B	79	1.3E-02	*Nrp2,*
W	272	1.3E-05	*App,*
Sx2mmd17	174	5.8E-03	*Slc1a1*
Sx2mmd50	202	6.3E-03	*Apoe*
Sx8mmd17	74	8.8E-03	~
Sx8mmd50	14	6.6E-02	~
Sx8wmd17	133	2.7E-03	*Apoe*
Sx8wmd50	53	3.9E-03	~

Table 13. Summary of overlap with other papers. Each paper is indicated by the acronym given in 3.7.1. The number of genes significant in both our study and the indicated one is given in the first row. The hypergeometric probability is given in the second, and a subset of interesting genes within this list of overlapping genes is within the third.

It is clear that there is significant overlap with other *Disc1* mutant models. We can see that many genes which have homologues differentially expressed in the human cells such as *Gpc1* and *Nrp2* appear, as do the potential Disc1 interactors encoded by *Dpysl2* and *Dpysl3.* Genes which are of interest to neuronal processes, appeared in a GO term, were also changed in the human cells, and were implicated by one of the above papers were particularly prioritised when looking to validate the RNA-Seq via RT-qPCR. A full summary of the overlaps is given in the Appendix.

4.1.2.5 Gene level RT-qPCR

To confirm the results of the RNA-Seq, a number of RT-qPCRs were performed. Genes were chosen on the same basis as the human neuron gene candidates but special attention was paid to genes differentially expressed in both models and many of these were examined. A summary of the overlaps between any pair of papers described in 3.7.1 and the list of genes significant according to DESeq2 is given in Table 14. A description of each gene is given in turn with a summary table given as Table 15.

DESeq2 crossovers of *Der1* het cortex	PGC1	PGC2-1	PGC2-2	PGC3	Exon changes	DI	B	W	T(1;11) human neurons	*Der1* Het hippocampus	Sx2mm D50	Sx8mm D50	Sx8wm D50
PGC1	42	29	35	0	14	2	1	8	5	0	4	1	1
PGC2-1	29	42	34	0	13	2	0	8	4	0	2	0	0
PGC2-2	35	34	72	0	22	2	1	13	7	3	3	1	0
PGC3	0	0	0	9	2	0	0	1	1	0	0	0	0
Exon changes	14	13	22	2	479	7	12	56	22	3	44	5	2
DI	2	2	2	0	7	14	1	2	1	2	1	0	11
B	1	0	1	0	12	1	70	15	9	1	14	0	5
W	8	8	13	1	56	2	15	241	23	2	47	8	11
T(1;11) human	5	4	7	1	22	1	9	23	127	2	21	0	5
Der1 Het hippocampus	0	0	3	0	3	2	1	2	2	27	5	1	0
Sx2mmD50	4	2	3	0	44	1	14	47	21	5	181	3	5
Sx8mmD50	1	0	1	0	5	0	0	8	0	1	3	13	2
Sx8wmd50	1	0	0	0	2	11	5	11	5	0	5	2	43

Table 14. Table of overlaps. Numbers represent the number of genes significant in our DESeq2 study of *Der1*heterozygous cortical samples, in addition to the two papers in the corresponding row and column. Grey blocks indicate genes from only one paper (the same row and column index). Abbreviations as in 3.7.1.

	Expression as % of WT	Padj	Exon gene changes	iPSC-derived neurons	Mouse hippocampus heterozygous changes	PGC1	PGC2.1	PGC2.2	PGC3	DI	B	W	Sx2mmd5	Sx8mmd5	Sx8wmd5
Apoe	115.5	1.06E-02											TRUE		
Arc	42.2	3.67E-02	TRUE		Exon level		TRUE	TRUE				TRUE			
Avp	16.7	1.75E-04													
Hap1	79.6	6.41E-04	TRUE	Gene level											
Metrn	123.6	5.35E-03		Gene level								TRUE			
Mt2	153.0	2.36E-02													
Nrp2	86.6	4.65E-02		Gene level							TRUE				
Slc1a1	90.4	1.18E-02	TRUE												

Table 15. Highlighted information about candidate genes selected for qPCR. TRUE indicates that the gene is differentially expressed in the model of interest. Paper abbreviations are as in 3.7.1.

4.1.2.5.1 Arc

Arc is an immediate early gene, capable of being rapidly translated to protein upon cellular signalling. It is expressed following learning, seizures, or LTP caused by

BDNF or high frequency afferent stimulation. Antisense *Arc* oligonucleotides inhibit the maintenance but not the induction of LTP. Its mRNA is targeted to the dendrite before translation; implying an important role there. It appears to have a role in organising actin at the dendrite, which is needed to enlarge dendritic spines. It also appears to have a role in AMPA receptor trafficking, particularly in AMPAR internalisation in LTD[57]. CNVs predisposing to schizophrenia have also been reported as converging on genes involved in Arc signalling[35]. All of this adds up to this gene having a crucial role in synaptic plasticity. A number of other immediate early genes are also downregulated, including *Egr1, Egr2,* and *Egr4*.

BDNF's receptor *NTRK2* is differentially expressed in the human neurons; it has an important role in LTP and appears to stimulate expression of *ARC.* Interestingly Aβ in cortical neurons appears to attenuate this expression increase, and in this context it is notable that the *Apoe* gene, known for a genotype which increases the risk of Alzheimer's disease, is also dysregulated in the mouse cortex[242]. Apoe appears to increase the formation of Aβ[243].

4.1.2.5.2 Apoe

Apoe is a lipoprotein known for transporting cholesterol, and is the most abundant lipoprotein in the brain. It is perhaps best known for a common variant which predisposes to Alzheimer's disease, with homozygote carriers of this ε4 variant suffering from the disease at a rate 8 times higher than the base rate. Apoe is important for carrying lipids around the brain and may even bind amyloid peptides[244]. Apoe also appears to have effects on neurons which do not appear to be directly related to its role in Alzheimer's pathology. Mouse cortical adult and embryonic neurons from *Apoe* KO mice have shorter dendrites, an effect which may be mediated *in vivo* by the expression of Apoe isoforms from astrocytes. Apoe protein is also capable of stimulating neurite outgrowth, and is upregulated here[245].

4.1.2.5.3 Avp

This gene encodes arginine vasopressin, a neuropeptide which is both necessary and sufficient for pair bonding in a species of vole[246]. Given its immense importance as a social neuropeptide I investigated further whether it had any other roles in brain

activity. Interestingly, the gene *Oxt*, encoding oxytocin is also dysregulated in our mouse heterozygote cortex. Avp receptor activation is necessary in the ventral palladium to allow pair bonding in males, while Oxt receptor activation is necessary in the nucleus accumbens to allow the same in females. The two are complementary in this regard and both are downregulated in our mouse cortex. Differential expression of the cognate receptors across related species appears to determine monogamous behaviour in the vole, and the two peptides control other behaviours as well, such as parental, aggressive, and in the case of Avp, social recognitive[247]. An older study looking at Avp co-administration during ethanol administration in mice found that Avp could maintain acquired ethanol tolerance (which otherwise lasted less than 6 days). Cessation of Avp administration began the process of losing acquired tolerance. The relevance is unknown but it is an interesting finding[248]. In any case, I thought it of special interest that both the Avp and Oxt peptides were dysregulated and thought this might be indicative of wider dysfunction in behaviours, especially social ones.

4.1.2.5.4 Hap1

The analogous human gene, *HAP1*, is differentially expressed in the human neurons and the function is more fully explained in the relevant section, 3.8.7.

However, the apparent convergence of significance in both mouse and human samples may be due to sex effects in the human cells. *Hap1* was not found significant in these mice samples, perhaps suggesting that the sex effects in humans and chance effects in mice converged, and that the gene is not truly involved in DISC1 pathology.

4.1.2.5.5 Metrn

The analogous human gene, *METRN*, is differentially expressed in the human neurons and the function is more fully explained in the relevant section, 3.8.10.

4.1.2.5.6 Mt2

This gene encodes metallothionein 2, a protein involved in metal-binding and control of oxidative stress. The related gene *Mt3* was also differentially expressed. *Mt2* mouse mutants have impaired spatial learning, and the protein appears to be involved

in the protective response against brain damage or chemical insult. Correspondingly, the mice have higher mortality post brain ischaemia[249]. It and the other metallothioneins can be induced by metal exposure. Lack of these proteins results in reduced oxidative stress gene expression post arsenic exposure, as well as increased cell lethality[250]. They are also involved in the response to non-metallic stress agents such as kainic acid, which induces seizures. Lack of metallothioneins here results in increased seizure phenotype as well as increased neuronal apoptosis[251]. It has a similar protective effect against dopamine toxicity[252]. Clearly the metallothioneins are important in protection against a variety of environmental and excitotoxic agents.

The gene was also analysed as a standard bearer for the other genes; it has one of the best separations between the genotypes, as well as a reasonable fold change. RT-qPCR for this gene would also gauge how reliably differential expression in the RNA-Seq can be confirmed at the qRT-PCR level in addition to analysing its expression in its own right.

4.1.2.5.7 Nrp2

The analogous human gene, *NRP2*, is differentially expressed in the human neurons and the function is more fully explained in the relevant section, 3.8.12.

4.1.2.5.8 Slc1a1

Slc1a1 encodes the protein Eaat3, a glutamate transporter expressed in neurons which is important in preventing excitotoxicity, over-signalling, and neuronal desensitization. It has been found that membrane presentation of the transporter, and subsequent increased glutamate uptake, occurs post-LTP[253]. Conversely, amphetamine causes the internalization of the receptor in dopaminergic neurons, which would presumably cause higher levels of glutamine at the synapse and subsequent excitatory signalling[254]. Excessive excitatory signalling in dopaminergic neurons has been proposed as being critical to psychosis[53]. Aged *Slc1a1* null mice have behavioural alterations consistent with neurodegeneration[253].

4.1.2.5.9 Results of RT-qPCR

The expression plots of the genes are displayed in while a summary is given in Table 16.

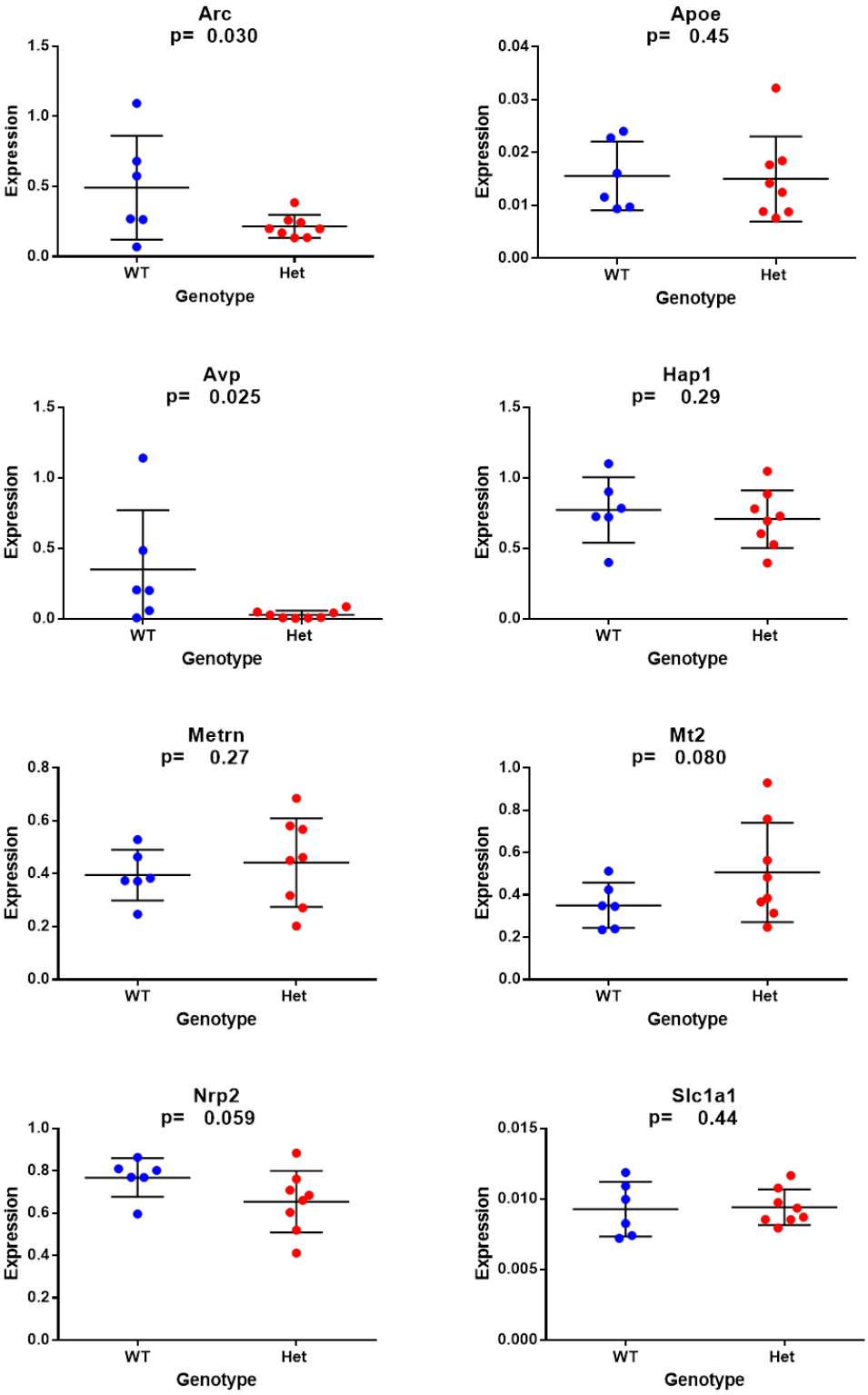


Figure 24. . Expression plots for RT-qPCR results of Arc, Apoe, Avp, Hap1, Metrn, Mt2, Nrp2, and Slc1a1 differential gene expression. P values given underneath and gene names above. Results have been normalised to the geomean of three housekeeping genes.

RT-qPCR results	P value	Heterozygous Fold Change	Significance
Arc	0.010	0.44	*
Avp	0.0099	0.09	*

Table 16. Summary of RT-qPCR gene level expression results for mouse cortical samples. * =p<0.05, **=p<0.01, ***=p<0.001.Mutant expression is in percentage of the WT expression. Only significant results are displayed.

A low proportion of the genes differentially expressed in the mouse heterozygous cortex were confirmed by RT-qPCR. Only two of 8 genes were confirmed. This is a surprisingly low number. However, the results of the RT-qPCR are not exceptionally different from the RNA-Seq. Only 10 genes suffer a reduction in expression of 50% or greater in the mouse samples, while no genes double in expression (see for detail). Only a tiny minority of gene expression plots show the heterozygous and WT samples separating into two groups which do not overlap. Only 46 genes (including *Avp*) have counts which fulfil this criterion. So it appears as though the effects on the mouse cortex are relatively mild or subtle, and do not reliably appear except in a few cases. This may be the reason for their failure to replicate in the RT-qPCR. This hypothesis is borne out by more detailed analysis of the relationship between qPCR and counts shown in Figure 25.

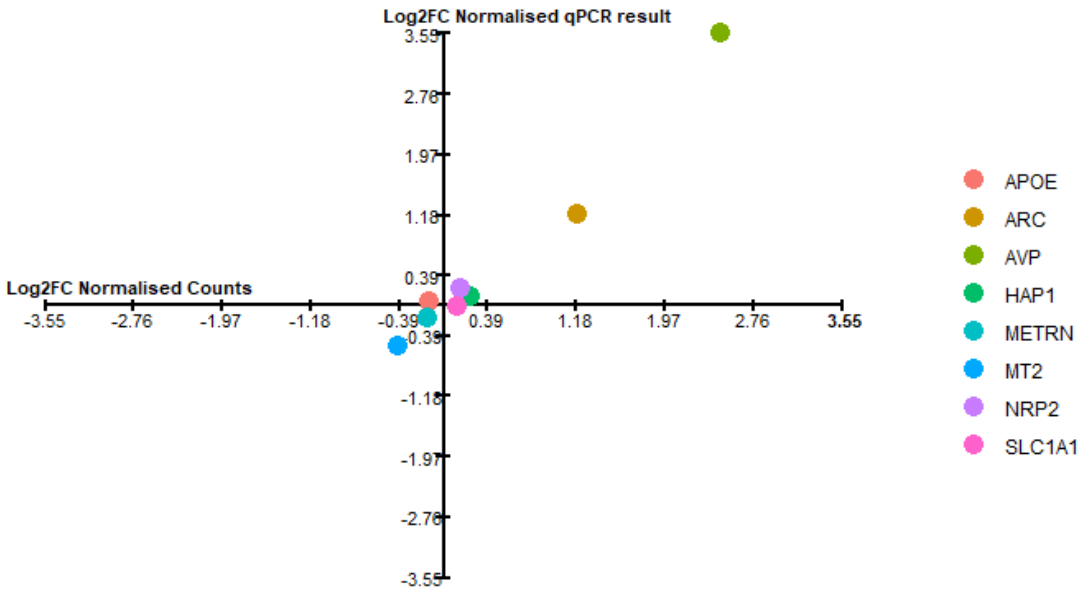


Figure 25. Graph displaying log2 fold change of normalised counts vs log2 fold change of normalised PCR score for each of the 8 genes I carried PCRs out on. Positive values indicate the mouse Der1 samples have a higher score on average.

Gene legends on right, differentiated by colour. Counts normalised using the "rlog" function in DESeq2 package. R^2=0.971

Figure 25 also shows a relatively linear relationship, although for most samples the fold changes are very small. Only for those two with larger fold changes (*Apoe* and *Arc*) was a significant difference found. This therefore lends some credence to the hypothesis that the mouse adult brains show relatively subtle changes.

4.1.3 WT vs homozygous

4.1.3.1 DESeq2

The six WT and eight homozygous samples were analysed using DESeq2. Since sex ratios were balanced, XY reads were not removed. Only five genes were found differentially expressed between WT and homozygous samples. However, it was noted by Kirsty Millar that the homozygous samples show a very unusual pattern and the results of the subsequent investigation are discussed here with permission. As a PCA reveals, clustering of the samples is very unusual. The PCA plot of the normalised counts can be seen in Figure 26 and utilises the top 500 most divergent genes, as five genes would be uninformative. We can see that the WT samples are separated by sex primarily, as in the previous section looking at WT and heterozygous mice. A new PCA using only the homozygous samples can be seen in Figure 27, where we again see a pattern of 4 pairs of samples. PC1, explaining 43% of the variance, separates samples 9, 10, 13 and 14 from samples 11, 12, 15 and 16.

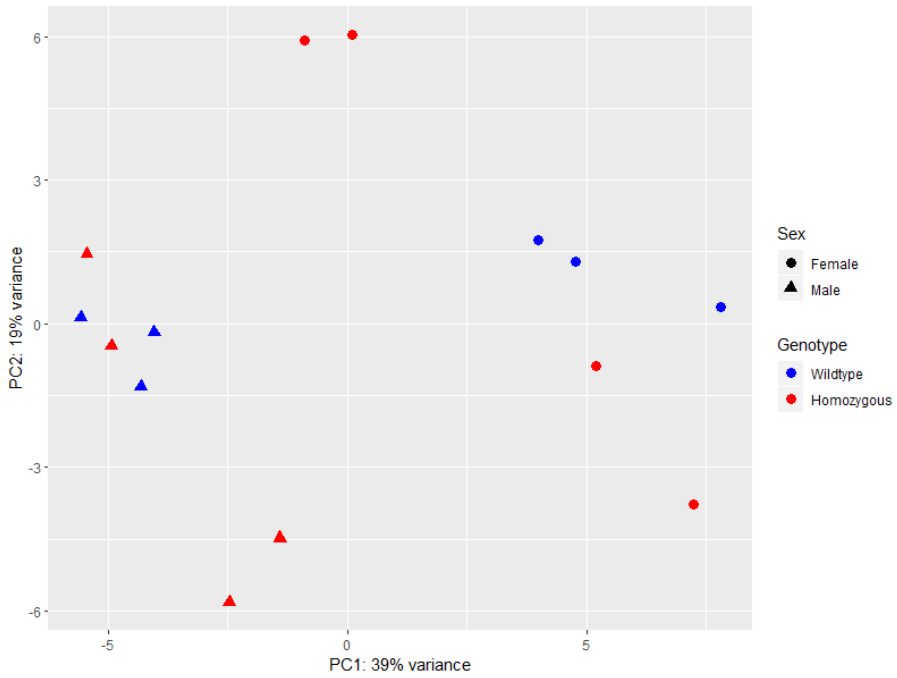


Figure 26. PCA of normalised counts for 6 WT (blue) and 8 homozygous (red) mouse hippocampal samples. Triangles are male. Circles are female. PCA generated using the top 500 most divergent genes.

The split is not related to sex as each group is comprised of two males and two females. The split in these homozygous cortical samples is interesting. Kirsty Millar found that the divergence is due to approximately 200 genes, which are differentially expressed in both groups from the WT, but in opposite directions. A DESeq2 analysis carried out by Philippe Gautier confirmed this and it is discussed in 4.1.3.3.

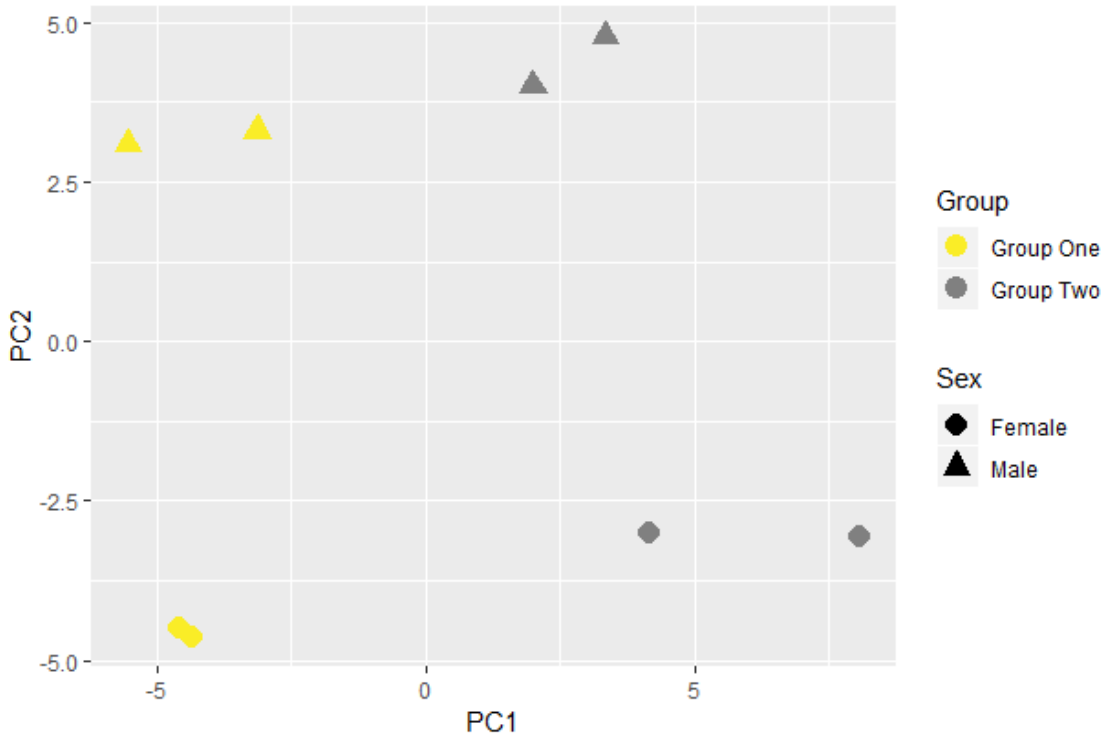


Figure 27. PCA of normalised counts only utilising homozygous samples. Samples 9-12 are male, 13-16 are female. Group One is samples 9,10, 13,14, while group Two is 11,12,15,16. PCA generated using the top 500 most divergent genes. PC1=43%, PC2=29% variance

In total, only five genes were differentially expressed between WT and homozygous samples and had a BaseMean higher than half that of *Disc1*'s. The high level of in-group variation in the homozygous sample is probably contributing to the low number of genes found. This is further discussed in section 4.1.3.3. One of the differentially expressed genes was *Disc1*. Given the exceptionally low number of differentially expressed genes I opted to concentrate my efforts on the heterozygous cortical samples and did not carry out RT-qPCRs.

4.1.3.2 DEXSeq

As with DESeq2, few changes are observed between the WT and homozygous samples. A total of six exons in three genes had an adjusted p value below 0.05, which rose to 10 exons in six genes at adjusted p value<0.1.

4.1.3.3 Divergence of two homozygote groups

Philippe Gautier carried out two further comparisons of the WT samples vs each of the homozygote groups using DESeq2. I then subsequently used GOrilla to analyse

the list of differentially expressed genes in each case, comparing and contrasting the two lists against each other and against the heterozygote list. 692 genes were differentially expressed by group one, and 2,619 by group two, applying the threshold of halved *Disc1* expression and $P_{adj}<0.05$. 249 of these genes are in common and the patterns within these were examined by Kirsty Millar. The results of this examination are printed in 4.1.3.3.3.

I searched for potential explanations for the divergence in two homozygote groups. Two putative explanations were litter effects (uterine environment and co-dissection effects) and circadian rhythms. A full picture of the pedigree is available in Figure 28. Group One consisted of mice 9, 10, 13, and 14, while Group Two Consisted of mice 11, 12, 15, and 16. These are sex-balanced. As seen in Figure 28, there are no litters which contain members from both groups. In addition, two litters make up each group. These contain mice 9/10/13 and mouse 14 for Group One, in addition to mice 11/16 and mice 12/15 for Group Two.

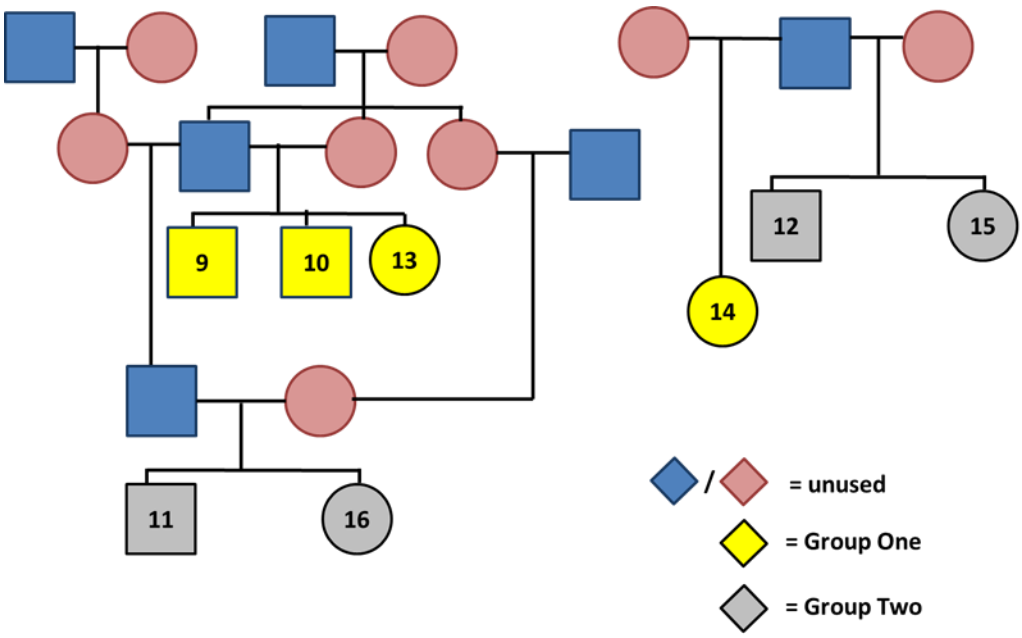


Figure 28. Pedigree of mice used to generate homozygous *Der1* cortical samples. Squares indicate male, circles female. Blue and pink represent unutilised mice, yellow indicates Group One homozygotes and grey Group Two homozygotes.

Regrettably, as the dissections for the mice were done by a variety of collaborators not all the dissection times are available. These times do point towards potential

effects either of circadian rhythms or dissection factors, however. The times are given in Table 17.

Mouse number	Group	Dissection time	Notes
9	One	2pm	Dissected with mouse 10
10	One	2pm	Dissected with mouse 9
11	Two	Not recorded	Dissected with mouse 16
12	Two	Not recorded	Dissected with mouse 15
13	One	11:45am	
14	One	12am (Noon)	
15	Two	Not recorded	Dissected with mouse 12
16	Two	Not recorded	Dissected with mouse 11

Table 17. Mouse numbers and dissection times, carried out by collaborators Marion Bonneau, Laura Murphy, and Elise Malavasi. Times provided by Marion Bonneau (private correspondence).

Group One litters were dissected either close to midday or at 2pm. The times for Group Two mice were not recorded and circadian rhythms can neither be implicated nor ruled out from the known times. Given that mouse 13 is part of the litter containing 9 and 10 there must have been a delay in processing these two mice. However, we can see that for most litters all constituent mice were dissected simultaneously, as expected. It is highly possible that litter effects (uterine environment, dissection effects), and possibly circadian rhythms, are responsible for the observed phenomenon of two groups.

4.1.3.3.1 Group One

4.1.3.3.1.1 GOrilla Process

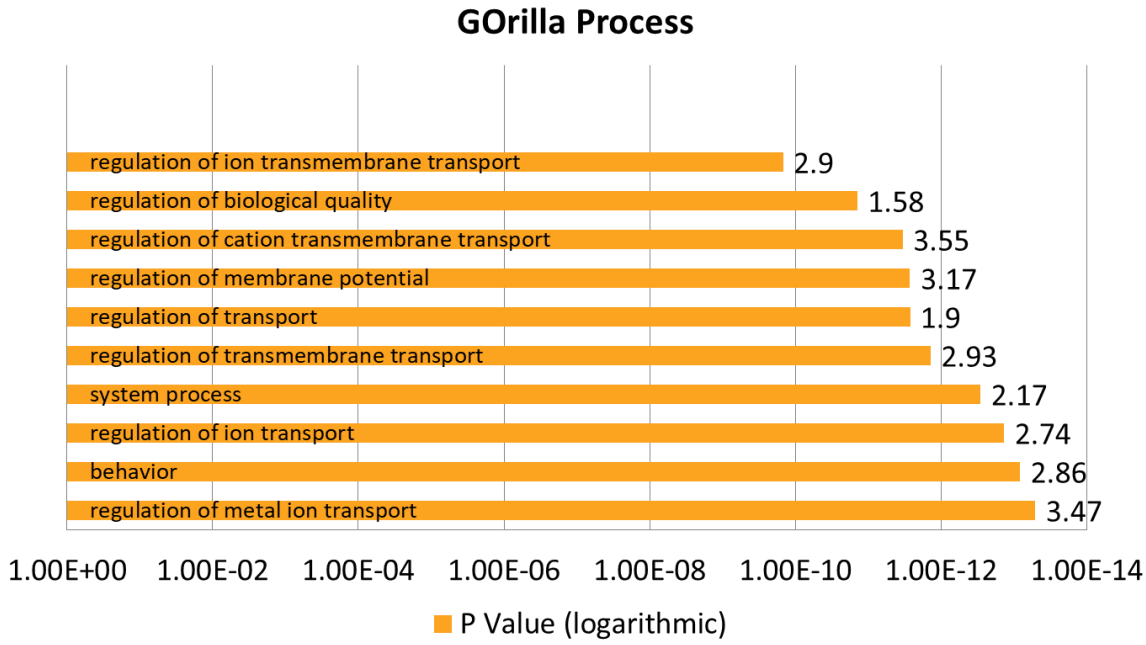


Figure 29. Top 10 significantly overrepresented Process GO terms for the genes which are differentially expressed ($p<0.05$). The enrichment figure is given after each bar and the scale is logarithmic.

There are a number of interesting GO terms, with "behaviour" having perhaps the most relevance to our model The genes associated with this GO term include interesting candidates, including many that were also seen in other comparisons. These included *Arc, Apoe, Avp, Oxt,* (all mouse heterozygous cortex), *Drd2* (human neurons), *Drd1a, Chrna7* (mouse heterozygous cortex, CNV studies), *Synj1, Cacnb4* (mouse heterozygous hippocampus), *Grid1*. "Regulation of cation transmembrane transport" included many of these genes, but also *Shisa6* and five potassium channel proteins *Kcns2, Kcnc2, Kcna1, Kcnab1, Kcnj2* and the interacting protein *Kcnip2.* Three sodium channels, *Scn2b, Scn4b, Scn8a* are also seen in the "regulation of ion transport" term.

4.1.3.3.1.2 GOrilla Function

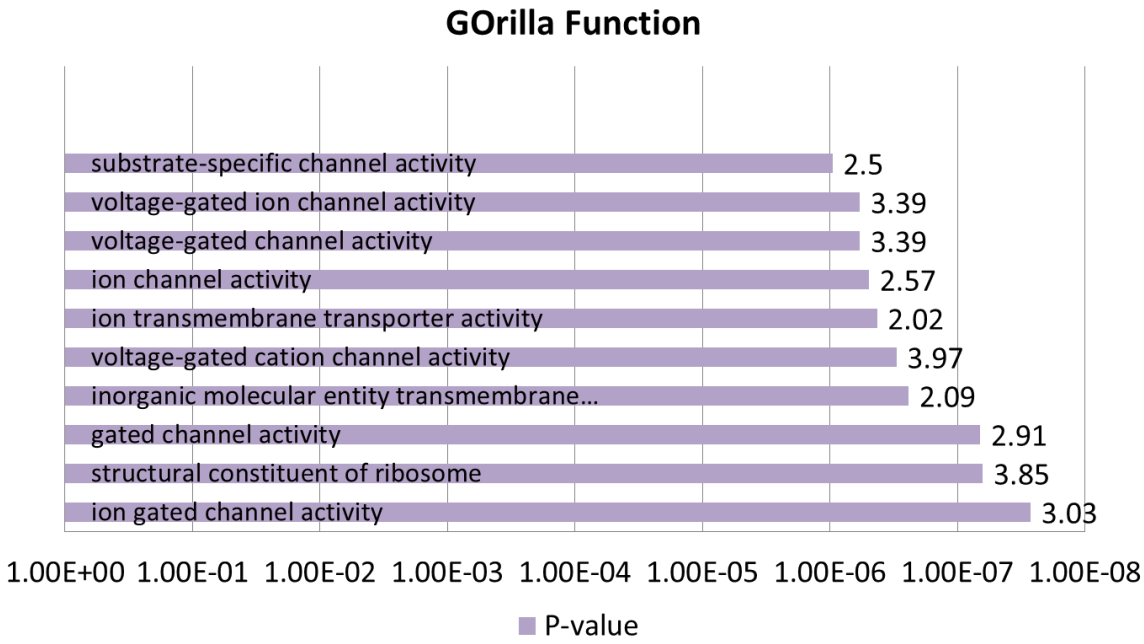


Figure 30. Top 10 significantly overrepresented Function GO terms for the genes which are differentially expressed (p<0.05). The enrichment figure is given after each bar and the scale is logarithmic.

Many of the terms contain the same genes found in the term "GABA-A receptor activity", p value= $3.8x10^{-4}$. This has the five genes *Gabre, Gabrg1, Gabrb2, Gabra1, Gabrq,* all subunits of the GABAA receptor which mediates inhibitory synaptic activity. *Gabre* and *Gabrq* encode the subunits ε and θ, which are less abundant but appear to be assembled into GABAARs with unusual potency[255].

4.1.3.3.1.3 GOrilla Component

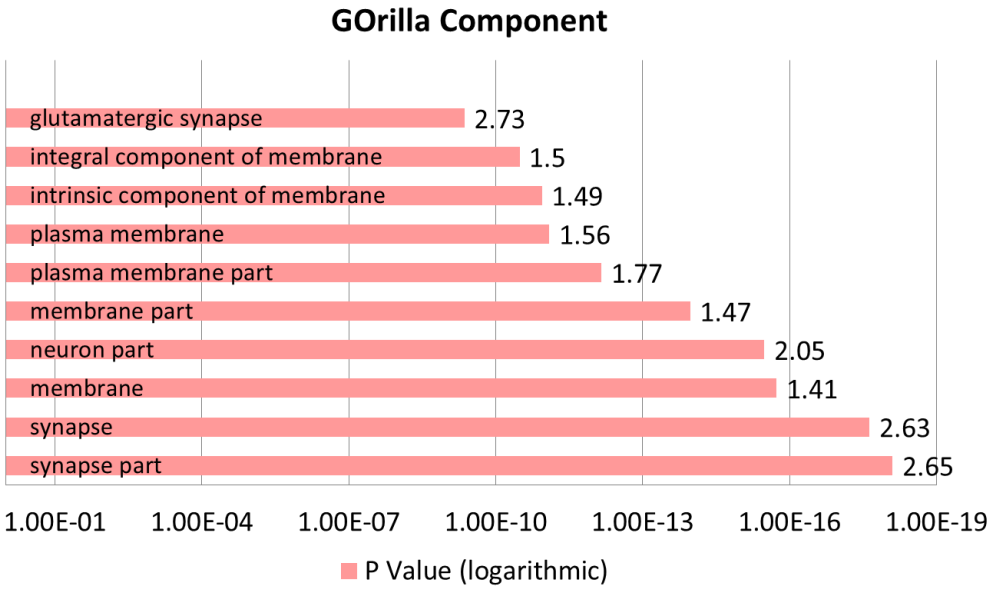


Figure 31. Top 10 significantly overrepresented Component GO terms for the genes which are differentially expressed ($p<0.05$). The enrichment figure is given after each bar and the scale is logarithmic.

As before, the terms relate to specialised structures of the neuron. *Grin2a* is one of the genes highlighted in "synapse", and its location on the neuron is known to be abnormal in these mice. It is also highlighted by GWAS and is crucial in long term potentiation[70].

4.1.3.3.2 Group Two

4.1.3.3.2.1 GOrilla Process

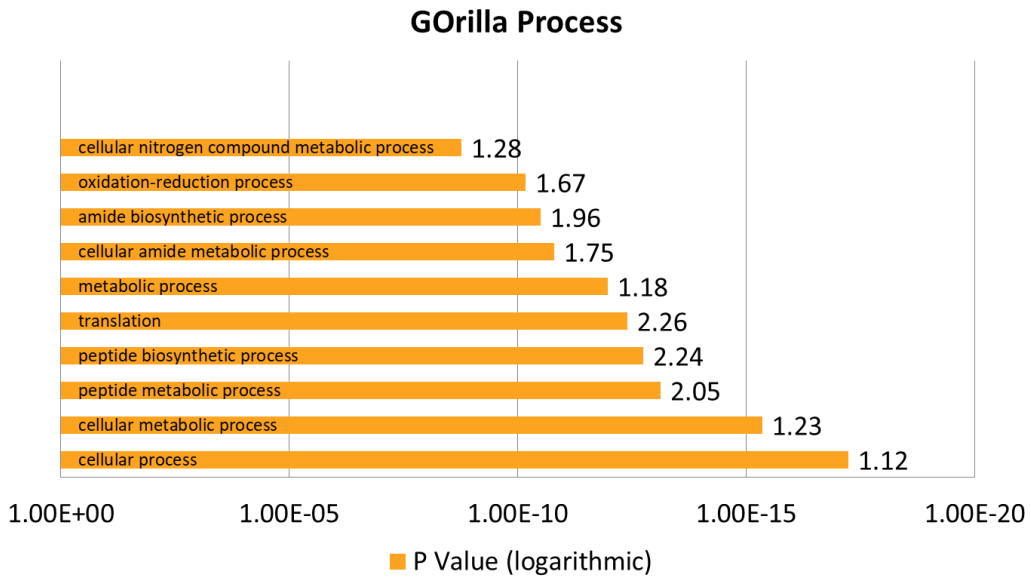


Figure 32. Top 10 significantly overrepresented Process GO terms for the genes which are differentially expressed ($p<0.05$). The enrichment figure is given after each bar and the scale is logarithmic.

One term of interest is “translation”, which has 53 genes encoding either mitochondrial ribosomal or ribosomal small and large protein subunits. This is a highly interesting change and may be of relevance to the needs of neurons, which require localised protein translation. “Oxidation-reduction process” may have some relevance to mitochondrial dysfunction.

4.1.3.3.2.2 GOrilla Function

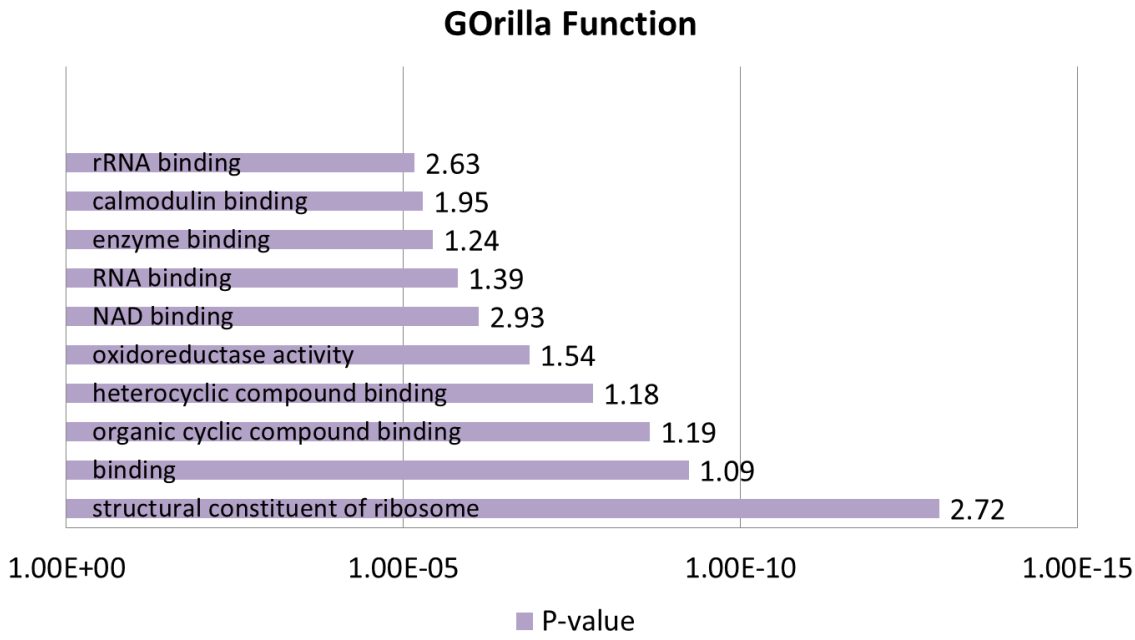


Figure 33. Top 10 significantly overrepresented Function GO terms for the genes which are differentially expressed ($p<0.05$). The enrichment figure is given after each bar and the scale is logarithmic.

These functions closely relate to the processes described in the previous section, and highlight ribosomal and mitochondrial dysfunction as themes, with 12 genes relating to oxidoreductase activity.

4.1.3.3.2.3 GOrilla Component

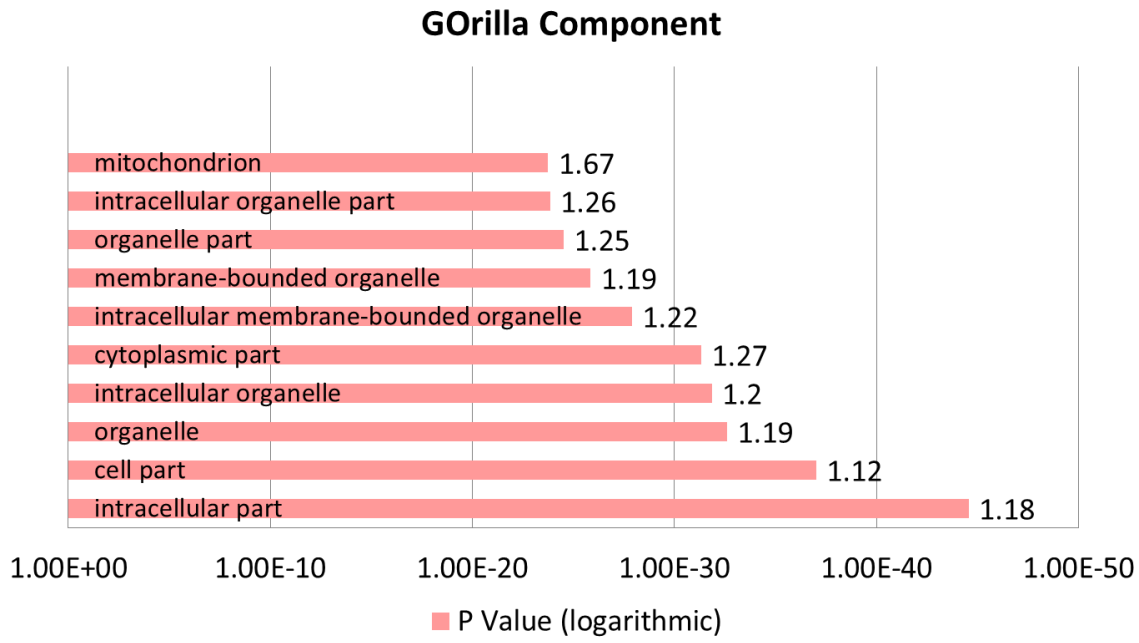


Figure 34. Top 10 significantly overrepresented Component GO terms for the genes which are differentially expressed ($p<0.05$). The enrichment figure is given after each bar and the scale is logarithmic.

The top terms are broad, but terms 10-13 are mitochondria related, and the significance of the top terms appears driven by genes within these terms.

4.1.3.3.3 238 genes in common that diverge

Of the 249 genes that were differentially expressed between the each group and the wild-type, and in common, only 11 change in the same direction in both homozygote groups. The following of these 11 are downregulated; *Arc*, *Disc1*, *Dusp5*, *Gadd45b*, *Junb*, *Per1*, *Plk3*, *Sertad1*, *Zfp948*. The following are upregulated; *Gm2115*, *Nkx3-1*.

Therefore the majority of the other genes in common are differentially regulated in opposite directions, with a difference from the WT being 30% on average. Taking the absolute % deviation from the wildtype expression, the difference between Group One and Group Two was on average 1.4%. This change was an average of 10% for the 11 genes that were in common, and 1% for the 238 which are different in sign. Therefore, the changes in opposite directions are strikingly similar in magnitude, being about 1% different from one another in terms of wildtype expression, and deviating around 30% from the wild type. 115 were increased in Group One, with

123 decreased, and these were decreased and increased in Group Two compared to wild-type respectively. The 249 genes are highly enriched for GO terms relating to "cognition" (5.82, $p=1.16x10^{-11}$), "learning or memory" (5.93, $p=6.87x10^{-11}$), and "behaviour" (3.79, $p=3.7x10^{-10}$), among other terms. In addition, 120 are shared with the list of genes differentially expressed in the heterozygous *Der1* cortex, with the vast majority of these having an unexpected division in sign change. Of the 120, three changed in the same direction in all lists; Group One and Group Two of the homozygotes, and the Heterozygous *Der1*. These were *Disc1, Nkx-3.1,* and *Zfp948,* which are downregulated, upregulated, and downregulated respectively. However, all 120 genes were changed in the same direction in Group Two and in the Heterozygous *Der1*. We therefore see a surprising overlap between these three groups of genes, and a highly unusual convergence of sign change between one of the homozygous and one of the heterozygous mutations. The changes were also similar in magnitude between Group Two and the Heterozygous *Der1,* with the mean difference between the changes being 2% of wildtype expression, and the maximum being 35%. This was in *Disc1,* which is at only 18% of wildtype expression in Group Two, but at 53% in the heterozygous *Der1*. It is interesting that the expression remains at 18% despite both loci being mutated; 9% per loci is higher than the 3% of expression the damaged locus appears to produce in the *Der1* condition. The next largest change is 10%, in *Zfp948*.

4.1.4 Discussion

The cortical heterozygous samples show differential expression of a large variety of genes compared to the wild type samples. These converge on some particularly interesting pathways. The differential expression of AMPAR and NMDAR subunits, proteins such as Arc (confirmed by RT-qPCR) and Egr1 involved in learning and memory, and trafficking molecules such as the kinesins and myosins indicates that there may well be synaptic abnormalities in the mouse cortex. Overlaps with other papers highlight Neurexin-1, protocadherins, and other synaptic structural proteins. There is therefore evidence for LTP related abnormalities in these cells; from the nucleus, to dendritic trafficking, to the synapses themselves. It should be noted that the majority of changes are low in absolute level. Many of the differentially

expressed genes have low fold changes and the noise inherent to RT-qPCR makes detecting them unlikely. As I showed, the qPCR is a good reflection of the RNA-Seq, but the small changes make verifying the differential expression difficult, and it is quite likely that given the small effect sizes chance is playing a role in the apparent differential expression of some genes.

In contrast, there are very few differentially expressed genes between the wild type and homozygote samples when taken as a whole. The minority of genes that are hold interest however; *Junb* is a transcription factor while *Per1* encodes a gene key in circadian rhythms. One aspect of particular interest is the splitting of the homozygous cortical samples into two groups which are not separated by sex, and the distinctive phenomenon of about 238 genes with the same fold changes but in opposite directions in each group. This set of genes includes many of those candidates which were looked at in the cortical heterozygous analysis, which are changed in the same direction in Group Two and the cortical heterozygote compared to Group One.

We therefore have a scenario with several groups of genes which must be clearly delineated. These are

1. Genes altered solely in the heterozygous *Der1* cortex
2. Genes altered solely in Group One of the homozygous *Der1* cortex
3. Genes altered solely in Group Two of the homozygous Der1 cortex
4. Genes which overlap between Group One and Group Two which are changed in the same direction. These are the minority, numbering 11, and are theoretically the invariant aspects of *Der1* homozygous mutation. Three also overlap with the heterozygous *Der1* cortex mutation and change in the same direction there. One is *Disc1*.
5. Genes which overlap between Group One and Group Two with changes in different directions. These are the majority of overlapping genes, numbering 238 and the magnitude is nearly the same. They are overrepresented for genes involved in behaviour, cognition, and learning.

6. Of these 238, there are 117 that overlap with those changed in the heterozygous *Der1* cortex and move cleanly as a group. 71 are upregulated in Group Two and the heterozygous cortex, and downregulated in Group One, with vice versa for the other 46. 3 other genes change in the same direction in all three.

There are some theoretical explanations.

The first is that the aberrant *Der1* locus is giving rise to mutant Disc1 proteins, which are interfering in the activities of oligomeric Disc1. This hypothesis requires that oligomeric and monomeric Disc1 have differing functions which at least in part are non-overlapping, and that the mutant Disc1 proteins cannot perturb the wild-type Disc1 in its monomer form, only when it oligomerises. The effect of the lost oligomeric functions is represented by the differentially expressed gene list in common between the homozygote and heterozygote functions, as these will be disturbed in both models. These genes included *Mt2, Ntrk3, Metrn, Drd2, Chrna7, Apoe, Slc1a1*. The monomeric functions that full length Disc1 carries out should be partially disrupted in the heterozygotes, and fully in the homozygotes, so under this model some of these genes may be related to Disc1's monomeric functions as well.

However, it is very difficult to explain the very clear phenomenon of a large group of genes changing in with the same magnitude but in different directions in the two homozygote groups, and that the majority of differentially expressed genes are not overlapping between the homozygote groups. As discussed in 4.1.3.3, both litter effects and circadian rhythms may be confounding or even causing the observed two group phenomenon. It is difficult to identify exactly how many genes might be altered by circadian rhythms; a recent study found that 43% of mouse genes showed circadian rhythms in translation in at least one of 12 organs, while the database CGDB names 9,580 gene as having daily oscillating expression[256,257]. However, I did find a list of genes which, when mutated, give behaviour phenotypes in mice relating to circadian rhythms[258]. These genes numbered 28. They were listed in Lowrey *et al.* as *Bmal1, Bmal2, Ccrn41, Clock, Cry1, Cry2, Cry3, Csnk1a1, Csnk1d, Csnk1e, Dbp, Dec1, Dec2, Fbx13, Mtnr1a, Mtnr1b, Npas2, Nrld2, Opn4, Per1, Per2, Per3, Prok2,*

Rora, Rorb, Rorc. Vip, Vipr2. I searched both Group One and Group Two for these genes, using both the above names and the synonyms given in Lowrey *et al*. *Per1* and *Per2* were differentially expressed in Group One, while *Cry1*, *Csnk1a1, Per1, Per2,* and *Rorb* were differentially expressed in Group Two. This is evidence in favour of circadian rhythms being partially or wholly responsible for the presence of two distinct groups.

4.2 Generation and initial analysis of mouse hippocampal RNA-Seq data

4.2.1 Introduction

The mouse hippocampal data was analysed in the same way as the mouse cortical data.

4.2.2 WT vs heterozygous

4.2.2.1 DESeq2

As previously mentioned in the cortical analyses, two WT mice had to be removed from the analyses due to one being an outlier. The second was randomly chosen and removed to balance the sex ratios. This outlier mouse was also used to generate hippocampal RNA-Seq data, so two WT mice have therefore been removed from the hippocampal analyses as well. A total of 184 genes were found differentially expressed, which are displayed in a volcano plot in Figure 35.

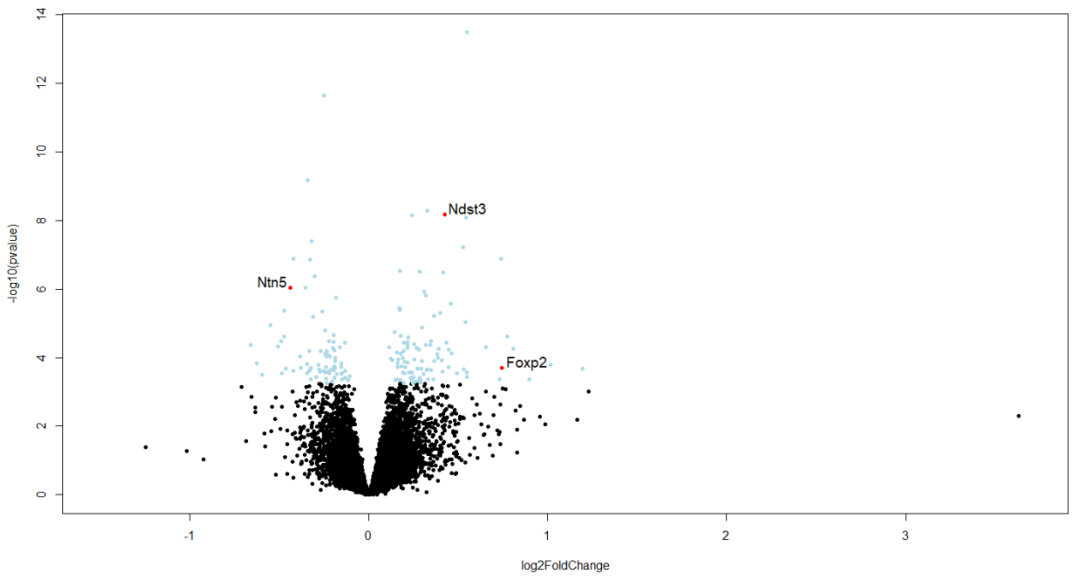


Figure 35. A volcano plot of the hippocampal RNA-Seq data for all genes with BaseMean>55. X-axis represents the log2 fold change between WT and heterozygoys *Der1* mice, while the Y axis represents significance (-log base 10 of p value). Black dots have an adjusted p value above 0.05, blue dots are significant with an adjusted value below 0.05. Red dots with labels represent genes for which a qPCR was carried out.

A PCA of the normalised counts for the 14 samples is in Figure 36. The result is very similar to that of the WT vs heterozygous comparison for the cortical samples. The samples are separated by a factor which corresponds to sex, and then by a factor which corresponds to *Disc1* mutant status for the females. However this second factor does not correspond to mutant status for the males. WT and heterozygotes are clearly interspersed. The translocation does not appear to exert as strong an effect here as in the cortical samples, where translocation status clearly was associated with PC2. This is in some way unsurprising; there were over 2,000 genes differentially expressed in the cortical heterozygotes, while here the number is closer to 200.

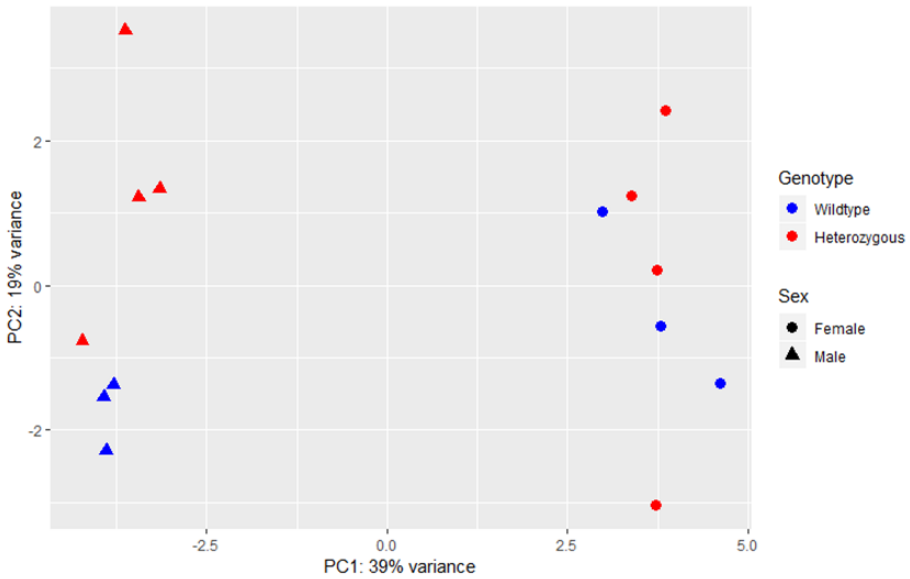


Figure 36. PCA of normalised counts for 6 WT (blue) and 8 heterozygous (red) mouse hippocampal samples. Triangles are male. Circles are female. PCA generated using the top 184 most divergent genes.

175 genes were significantly differentially expressed and had a BaseMean greater than half that of *Disc1*'s.

4.2.2.2 DEXSeq

A total of 136 exons in 131 genes were found differentially expressed between the two groups of samples, with adjusted p value below 0.10. This fell to 52 exons in 50 genes when the filter of $p<0.05$ was applied.

4.2.2.3 GOrilla

As in the human analysis in 3.6, GOrilla was utilised to analyse gene ontologies overrepresented among the differentially expressed genes. This analysis was carried out by Marion Bonneau and Kirsty Millar and is described in the corresponding thesis.

4.2.2.4 Comparison to other papers

I carried out an analysis identical to that described in 3.7. using the same papers. Removal of duplicate genes and determination of a match was carried out in the same manner as previously. A summary of the results is given in Table 18.

Hippocampus	**Heterozygous**		
Paper	Number of	P value	Selected genes
PGC1	10	1.2E-04	*Cacnb2*
PGC2-1	21	2.9E-10	*Erbb4, Ntn5*
PGC-2	15	0.99	*Erbb4, Ntn5*
PGC3	2	1.8E-1	*Nrxn1*
DI	4	2.2E-03	*Disc1, Snap91*
B	18	7.7E-08	*Grin2a*
W	56	1.6E-19	*Gpc1, Calb2,*
Sx2mmd17	22	2.3E-05	*Erbb4*
Sx2mmd50	30	1.9E-08	~
Sx8mmd17	9	8.9E-03	~
Sx8mmd50	4	5.8E-03	*Nrxn1*
Sx8wmd17	37	1.9E-17	*Cacnb2*
Sx8wmd50	4	1.7E-01	~

Table 18. Summary of overlap with other papers. Each paper is indicated by the acronym given in 1.8.1. The number of genes significant in both our study and the indicated one is given in the first row. The hypergeometric probability is given in the second, and a subset of interesting genes within this list of overlapping genes is within the third.

There were some highly interesting findings in this analysis. As with the mouse cortical analysis, almost all sets of genes from the papers had significant overlap with the list of differentially expressed features, with the exception of a Srikanth *et al.* neuronal model. It is also noteworthy that many genes appear in several studies, such as *Erbb4*, *Nrxn1,* and *Cacnb2.* In many cases, the genes are implicated by both a GWAS/CNV study and by a RNA-Seq analysis of an iPSC-derived model, indicating convergence between these different approaches. Some gene orthologues are also differentially expressed in the human t(1;11) neurons, such as *GPC1* and *ERBB4.* *Snap91* is also differentially expressed in the hippocampus of the homozygous mutant. We can see that there is overlap between both models of the t(1;11), as well as with other investigations of psychiatric illness.

4.2.2.5 Genes of interest

Although I did not carry out qPCRs, I had selected some candidates which are described in turn. A summary table of fold changes, etc. is also given in Table 19. One primary reason for not carrying out qPCRs was the low fold change (never >35%) of samples, as well as the lack of clear distinctions between the WT and *Der1* samples (in no cases did they cluster into two separate groups).

	Expression as % of WT	Padj	Exon gene changes	iPSC-derived neurons	Mouse cortex heterozygous changes	PGC1	PGC2.1	PGC2.2	PGC3	DI	B	W	Sx2mmd50	Sx8mmd50	Sx8wmd50
Foxp2	120.1	3.10E-02										TRUE			
Ndst3	126.7	1.99E-05		Gene level	Exon level						TRUE	TRUE	TRUE		
Ntn5	79.9	8.05E-05					TRUE	TRUE							

Table 19. Highlighted information about possible candidates. TRUE indicates that the gene is differentially expressed in the model of interest. Paper abbreviations are as in 3.7.1.

4.2.2.5.1 Foxp2

Foxp2 encodes a transcription factor known for its role in language; it is mutated in hereditary language disorders and the human homologue appears to have undergone human specific evolution. Foxp2 can induce neurite outgrowth, migration, and progenitor proliferation[259]. These are all processes which DISC1 is well known to have a role in, and the two genes are further linked by the fact that in human cells FOXP2 can repress *DISC1* expression. Lack of this repression is found in two alleles of the *FOXP2* gene which segregate with developmental verbal dyspraxia[260]. It seemed interesting that the gene was dysregulated here; it appears to be upregulated. Although no direct regulation of *FOXP2* by *DISC1* has been shown, it is interesting that both genes have overlapping functions. It could be possible that the two are capable of repressing one another in order to prevent over-stimulation of the relevant pathways; this would explain the upregulation of the gene in the mutant mice.

4.2.2.5.2 Ndst3

NDST3 (N-acetylase and N-sulfotransferase 3), the human homologue, is differentially expressed in the studies of Wen *et al.*, Brennand *et al.*, and the mouse heterozygous hippocampus model. It has also previously been implicated by a schizophrenia GWAS[124,132,261]. It therefore has support for its role in schizophrenia by a wide variety of investigative methods. The function of the encoded protein is to alter heparan sulfate, a molecule which is often found attached to various extracellular proteins. The modification it carries out is the first step for all other heparan sulfate modifications. Heparan sulfate proteoglycans usually have roles in the extracellular matrix; one example is GPC1, which is differentially expressed in the human neurons, see 3.8.6. It has been suggested that the heparanase/heparan sulfatase balance may alter the in/out trafficking of these altered proteins[262]. Heparan sulfate can also potentiate FGF-FGF receptor signalling and glycoproteins have roles in many other cell signalling processes, including endocytosis and cellular adhesion[263].

4.2.2.5.3 Ntn5

Netrin-5 is one of a family of proteins involved in axonal guidance and neuronal development. *Ntn5* is expressed particularly highly in regions of the brain which

undergo neurogenesis after development, including the hippocampus. It is co-expressed with *Dcx* and *Mash1* in neuroblasts. Expression decreases as the cells mature, and some of these cells are destined to become GABAergic interneurons[264]. Mutant mice had a phenotype similar to other axonal guidance genes nulls such as *Nrp2* or *Sema6a*; ectopic migration of motor neuron cell bodies[265]. Given the dysregulation of other axonal guidance molecules, as well as its possible role in the hippocampus, *Ntn5* seemed like an interesting gene.

4.2.3 WT vs homozygous

4.2.3.1 DESeq2

A PCA of the normalised counts is in Figure 37. The sex of the mice separates the samples. The number of differentially expressed genes is one, *Disc1*; even lower than in the WT vs heterozygote comparison, where mutant status did not correspond with principalcomponent 2. We can see in Figure 37 that mutant status does not correspond with principal component 2 here either.

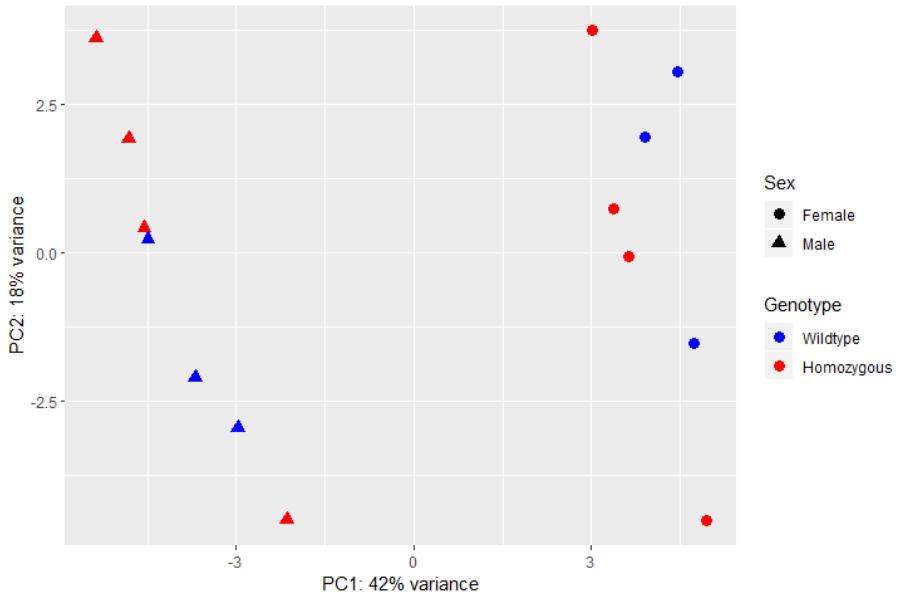


Figure 37. PCA of normalised counts for 6 WT (blue) and 8 homozygous (red) mouse hippocampal samples. Triangles are male. Circles are female. PCA generated using the top 500 most divergent genes.

4.2.3.2 DEXSeq

At adjusted p value<0.1, 305 exons in 251 genes were differentially expressed between the two groups of samples. 118 exons in 103 genes met the criterion of adjusted p value<0.05. *Disc1* has two differentially expressed exons. The two *Disc1* downregulated exons are on both sides of the breakpoint; however the changes are more severe after the breakpoint, indicating expression is more severely affected here.

4.2.3.3 GOrilla

The genes from the DEseq2 and DEXseq were combined to compare against the background list of expressed genes. The only gene differentially expressed at the whole gene level also has exons differentially expressed. It is *Disc1*. A total of 103 genes at the p<0.05 level were implicated by DEXSeq, and these were compared against the background list of 27,957 genes detected at the whole gene level. Significance was set at $p<1x10^{-3}$.

4.2.3.3.1 GOrilla Process

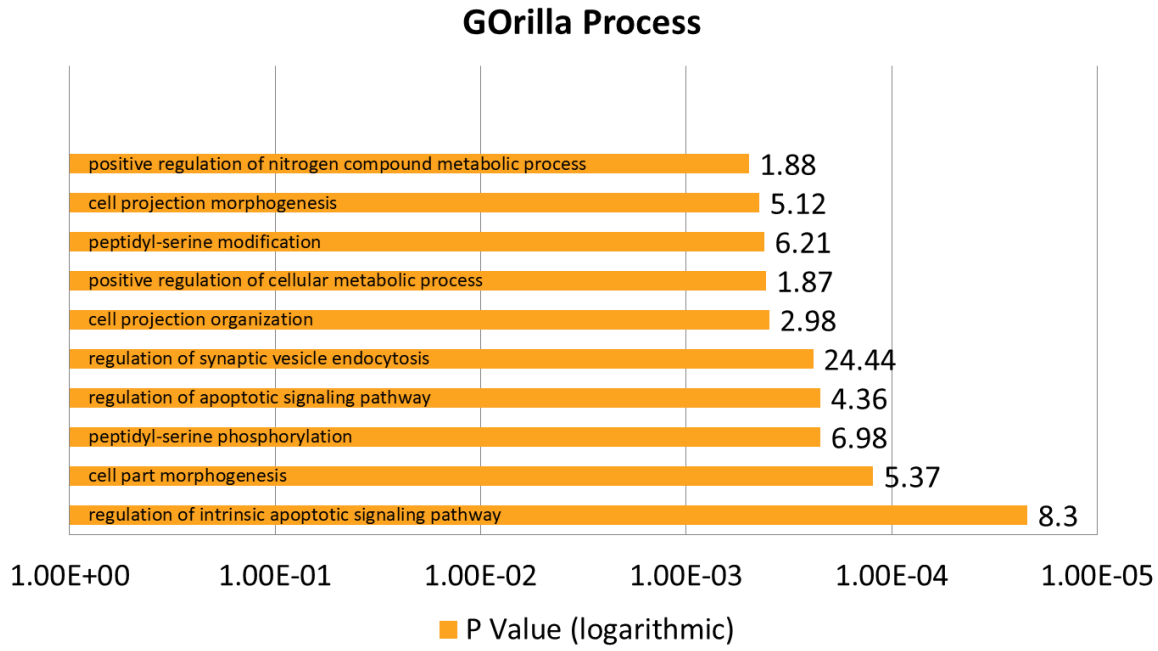


Figure 38. Top 10 significantly overrepresented Process GO terms for the genes which have a differentially expressed exon or are differentially expressed (p<0.05). The enrichment figure is given after each bar and the scale is logarithmic.

The top 10 GO Process terms are displayed in Figure 36. We can see that some of the terms in the top 10 are of particular interest to synaptic activity, such as "regulation of synaptic vesicle endocytosis". The genes within this group were *Snap91*, a synaptosome associated protein, *Pip5k1*, a phosphatidylinositol-4-phosphate 5-kinase, and *Mff*, a gene named "mitochondrial fission factor". The GO terms of relevance to apoptosis also seemed of possible relevance and included the well-known oncogene *Src*.

4.2.3.3.2 GOrilla Function

Two terms were significant. One, "binding", is too vague to be of any use. The other is far more specific, "acetylcholine receptor inhibitor activity" (p value=5.8×10^{-4}, enrichment factor=53.54), containing the genes *Ly6e* and *Ly6h*.

4.2.3.3.3 GOrilla Component

Three terms were significant. These were "extrinsic component of endosome membrane" (p value=3.1×10^{-5}, enrichment factor=46), "postsynaptic density" (p value=2.5×10^{-4}, enrichment factor=4.8) and "postsynaptic specialization" (p value=2.8×10^{-4}, enrichment factor=4.8). All of these relate to the synapse, and therefore may point towards some sort of synaptic alteration in the mouse hippocampus.

4.2.3.4 Comparison to other papers

I carried out an analysis identical to that described in 3.7 using the same papers. Removal of duplicate genes and determination of a match was carried out in the same manner as previously. A summary of the results is given in Table 20.

Hippocampus Homozygous Paper	Number	P value	Selected genes of interest
PGC1	6	6E-04	~
PGC2	13	1.5E-08	~
PGC2-2	7	0.99	*Ino80e, Snap91*
PGC3	2	6E-2	*Ino80e*
DI	2	2.4E-02	*Snap91*
B	6	6.4E-03	*Wnt7a*
W	16	3.2E-04	*Wnt7a*
Sx2mmd17	14	1.5E-05	*Kif1a*
Sx2mmd50	10	6.5E-03	*Kif1a*
Sx8mmd17	8	2.4E-04	~
Sx8mmd50	0		~
Sx8wmd17	12	3.4E-05	~
Sx8wmd50	4	3.3E-02	~

Table 20. Summary of overlap with other papers. Each paper is indicated by the acronym given in 3.7.1. The number of genes significant in both our study and the indicated one is given in the first row. The hypergeometric probability is given in the second, and a subset of interesting genes within this list of overlapping genes is within the third.

All the implicated genes have only exon level differential expression, perhaps indicating subtle splicing changes rather than whole gene differential expression. Nevertheless, we see some significant overlaps with other papers, although not as many as with the heterozygous mutation hippocampal or cortical samples. Of interest is *Snap91*, which as already mentioned is also differentially expressed in the heterozygous mouse hippocampus.

I also investigated the differentially expressed exon locations within their genes. *Snap91*'s exon is an N-terminal exon which appears to be in all isoforms. *Kif1a*'s is the 40th exon on UCSC, which appears to be in all isoforms. *Wnt7a*'s exon is the N-terminal exon which appears to be in all isoforms. *Ino80e*'s is an alternative N-terminal exon found in two isoforms; these appear to encode longer proteins than other transcripts but there is little functional information.

4.2.4 Discussion

The changes in the hippocampus appear to be of a far lesser number than the changes in the human neuron or mouse cortical models, with the total number being only about 10% of that in the other models. The heterozygous hippocampus does show some interesting changes, and the overlapping genes with other papers are of known interest. They are also seen in the mouse cortex as well (*Nrxn1*) or the human neurons (*Erbb4*, *Gpc1*). The overall enrichment of synaptic genes corresponds with the findings of the mouse heterozygous cortex.

In the case of the mouse homozygous hippocampus, only *Disc1* was differentially expressed at the whole gene level, while 103 genes showed splicing differences. It was also seen that the PCA did not identify mutation status as being one of the two principal components and male and female samples differed greatly from the WT samples without clustering together, although sex appeared to be a major component of sample differences. The splicing differences in *Snap91* (in both heterozygous and homozygous samples) and *Kif1a* might have suggested some kind of alteration with synaptic vesicles, but the exons do not appear to have any unique functional significance, as they appear in all isoforms. Overall it is difficult to identify a clear effect of the translocation in the homozygotes.

5 DECONVOLUTION OF THE RNA-SEQ DATA USING ZHANG *ET AL.* CELL TYPE ENRICHED DATASETS

5.1 Introduction

Deconvolution is an approach to extract data from a complex RNA-Seq sample. Each gene is a point of information which helps inform the estimated proportions of pure cell types making up the complex sample. Deconvolution algorithms are widely available. They typically assume that the contribution of each cell type to the mixed cell type expression is linearly related to the proportion of the mixed sample that is that pure cell type. This is described by the following equation;

$$X_{aj} = \sum_{s=1}^{q} p_s(X_{sj})$$

Where the transcriptional value X for gene *j* in pseudosample *a* is equivalent to the linear sum of the respective gene's values from pure samples $X_{s....q}$, multiplied by their relative proportion $p_{s....q}$, with all proportions adding to 1. This of course assumes that the transcriptional activity of pure cell types does not alter in the presence of other cell types. Each measured gene gives an equation similar to that above; if there are more measured genes than cell types the proportions can be determined. However, in practice the equations are not correct as it is unlikely that proportions can be found which solve the above equation for every gene. Therefore, close approximations are given as answers. These optimal proportions can be found using a non-negative least linear squares approach, although Cobos *et al.* discuss other approaches and deconvolution as a problem in a useful review[266,267].s

To briefly summarise, RNA-Seq deconvolution has three components.

- The mixed transcriptional profile, **M**
- The reference profiles **G**, each corresponding to one of the pure cell types present in the mixed cell population. Most genes are unlikely to be specifically expressed in a single cell type and are therefore not informative, so a subset of highly informative genes are used for the deconvolution.
- The relative proportions of the cell types, **C,** which have their expression profiles described by **G.**

- Given any two of these components we can derive the third, assuming that there is a relationship between the proportion of cell types and their representation in the sequenced sample. Such a relationship is likely to be linear.

DISC1 immuno-reactive neurons have been found throughout all layers of the human cortex, but also in rat cortical astrocytes, neurons, oligodendrocytes, and microglia[69,268]. Hence, there is potential for the t(1;11)/*Der1* to impact a wide variety of cell types, not just neurons. I first sought to deconvolute the wild-type and heterozygous *Der1* mouse RNA-Seq profiles. My interest in this was primarily to see if there were any differences in relative cell types caused by the *Der1* mutation. This could be due to altered cell development and differentiation, or degeneration, resulting in unusual levels of some cell types. Due to the importance of neurodevelopment to the aetiology of psychiatric illness, it seemed a plausible method by which the *Der1* exerted its effects. I initially aimed to deconvolute the profiles on a gross scale, looking for changes not within subtypes of neurons but for changes in neuronal vs various non-neuronal cell types such as astrocytes, oligodendrocytes, microglia, and endothelial cells.

I next sought to deconvolute the iPSC-derived cell profiles. Although the cultures are primarily neuronal in nature, it has been shown by Bilican *et al*. that the efficiency of the neuronal differentiation is not total. About 86% of cells are $TUJ1^+$ neurons, and glial cells are present. $GFAP^+$ cells (astrocytes) constitute 5-10% of cells[143]. I hypothesised changes in broad classes (astrocytes, oligodendrocytes, etc.) might be present, and describe my efforts to investigate this in this chapter. In the next chapter I looked at more detailed subclasses of neuron, interneuron, and other cell types.

5.2 Deconvolution datasets and program

5.2.1 Selection of an appropriate reference dataset

Regardless of the program used, successful deconvolution of mixed RNA-Seq data samples requires pure cell RNA-Seq data, **G**. For practical reasons I determined to use freely available data sets, which also fulfilled the following criteria:

Deconvolution of the RNA-Seq data using Zhang et al. Cell type enriched datasets

- They utilized a variety of brain cell types.
- The identification of each cell type was trustworthy. This is important so as to determine which genes are "markers" for the cell types of interest. It is likely that there will be some contaminating cells even within each "single-origin" RNA-Seq profile.
- The sequencing depth was comparable to our own.

Sequencing depth is especially important. Even post normalisation to total read number, samples sequenced to different depths are not directly comparable. There are two main reasons. It can be immediately appreciated that with increased sequencing depth comes an increase in the number of detected genes, as poorly expressed RNA features now have a greater likelihood of being sequenced. This alone can lead to false positive claims of differential expression between two biologically identical samples of different sequencing depth, as it will now appear that the gene is expressed in one sample and not in the other. For example, Tarazona *et al.* noted that regardless of total sequencing depth (from 20 million up to 200 million), increased depth always increased the number of genes with at least five counts[269]. Furthermore, this increase was especially pronounced the lower the total number of reads. The comparison of a 20 million sequencing depth sample to a 60 million depth sample will therefore be more problematic than the comparison of a 60 million depth sample to a 100 million depth sample. Secondly, genes with lower expression are more dramatically affected by increased sequencing depth, even post normalisation. Mortazavi *et al.* showed that after only 8 million reads, over 95% of the most highly expressed genes had RPKM normalised values within 5% of the values they would have at 40 million reads. Of the genes with the lowest expression, approximately only 50% had RPKM values within 5% of the value they would have at 40 million reads [270]. Since RPKM values are normalised to total sequencing depth, a comparison between sequencing at eight and 40 million reads might lead to the erroneous conclusion that these poorly expressed genes are differentially expressed, given that over half have differences in RPKM of over 5%. We can conclude that even normalising for the increase in sequencing depth, genes with low expression

show changes that are more dramatic in apparent expression as sequencing depth increases.

I conducted a search of the literature to find a comparable data set for deconvolution of mouse samples, and chose the data set described in Zhang *et al.*[153]. This is accessible through the NCBI GEO under accession number GSE52564. These data were obtained via sequencing of 100bp paired-end reads, with 65.6 ± 5.4 million reads sequenced (mean ± standard deviation). In comparison, the data from our mouse cortices were obtained via sequencing of 100bp single-end reads, so the length and therefore the number of mapped reads should be similar. For six WT samples, 75.3 ± 11.9 million reads were sequenced, for eight heterozygous *Der1* samples 89.2 ± 6.3 million reads were sequenced, and for eight homozygous *Der1* samples 98.7 ± 17.3 million reads were sequenced.

The sample generation method described by Zhang *et al.* is summarised in Figure 39. They pooled dissected cerebral cortices from three to twelve mice for each cell type, before purifying cell types via immunopanning and FACS, which they stated were equally viable and effective methods of cell purification with no discernible differences in expression profiles in purified cells. Astrocytes and endothelial cells were FACS-purified, while neurons, microglia, and oligodendrocyte lineage cells were purified by immunopanning. Oligodendrocyte lineage cells were isolated from P17 mouse brains, while astrocytes, endothelial cells, and neurons were isolated from P7 mouse brains. In total seven different RNA-Seq profiles were generated, from astrocytes, neurons, oligodendrocyte precursor cells, newly formed oligodendrocytes, myelinating oligodendrocytes, microglia, and endothelial cells. Two pools were generated for each cell type and I averaged these for each of the seven enriched RNA-Seq profiles. It should be noted that the ages of the mice utilised by Zhang *et al.* are not the same as our own; it is the case with astrocytes at least that age alters transcriptional profiles; Zhang *et al.* produced a follow up paper which segregated astrocytes by age and found differences[271]. This is likely to be the case with other cell types as well.

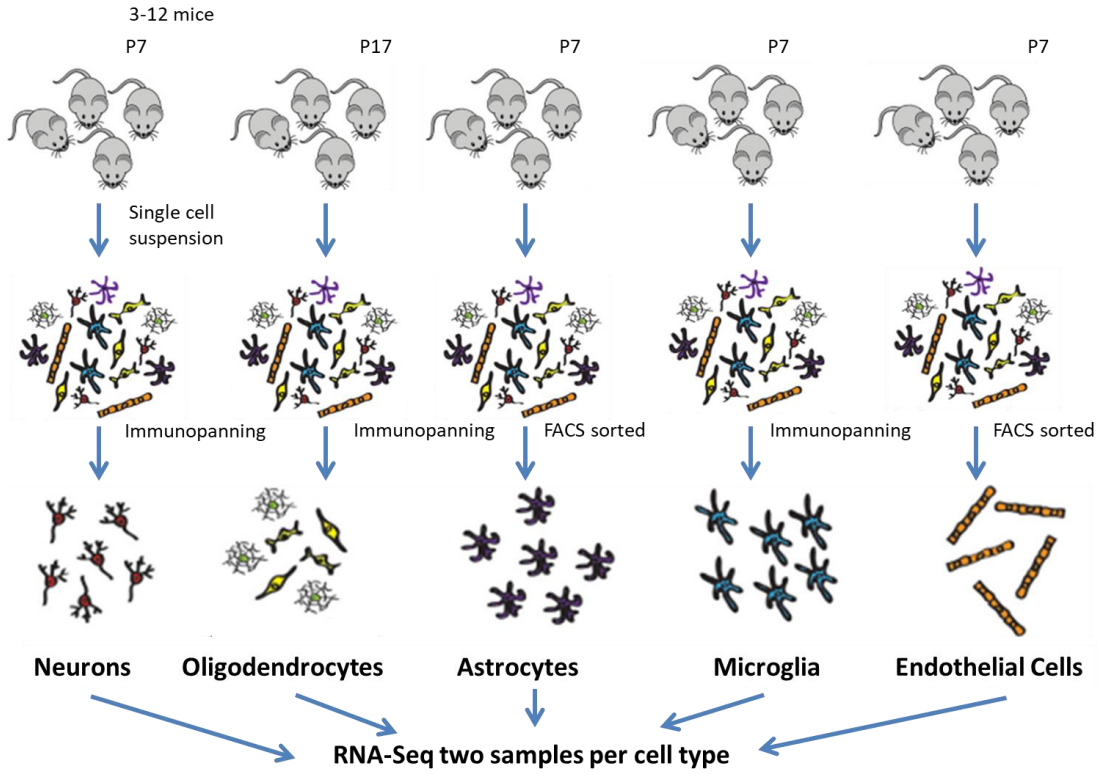


Figure 39. Adapted from Zhang *et al.* 2014[203]. Summary of Zhang *et al.* sample generation method. The Oligodendrocyte samples were further purified into Oligodendrocyte Precursor Cells, Newly Formed Oligodendrocytes, and Myelinating Oligodendrocytes before sequencing.

5.2.1.1 Comparing units across datasets

I sought to use values of RNA-Seq expression which are comparable across samples in an experiment. Zhang *et al.* have provided their data as Cufflinks-generated FPKMs (version 1.3.0) as well as in raw format. Cufflinks-generated FPKMs (version 2.2.1) were also provided for all our WT and *Der1* mouse RNA-Seq experiments as part of the commercial RNA-Seq data generation process. I therefore initially opted to use the conveniently generated FPKMs, which are comparable across samples as they have been normalised to total millions of reads for each sample. FPKM estimations appear to be near-identical regardless of Cufflink version number, see (http://cole-trapnell-lab.github.io/cufflinks/benchmarks/ visited on 14/6/18). However, deconvolution carried out on Cufflinks-generated FPKMs underperforms compared to deconvolution carried out on many other measurements of gene expression. Jin *et al.* undertook a comparative analysis of deconvolution utilising several expression quantification methods including Cufflinks. They utilised

RNA-Seq of pure RNA (universal reference human RNA and brain RNA), as well as RNA-Seq of mixed RNA of the two pure samples, mixed in a 3:1 ratio. They then compared projected values of expression (0.75)(Universal reference expression) + (0.25)(Brain expression) against the actual value of expression of the mixture. See Figure 40[272]. Jin *et al.* have here analysed the RNA-Seq profiles with a number of diffierent quantification methods including EdgeR, Salmon, Kallisto, and FPKMs generated by Cufflinks. They then display the rank of each gene or isoform's expression as projected by the 3:1 ratio of universal:brain against the rank of the gene in the RNA-Seq of the 3:1 mixed RNA. We can see that for some quantification methods, such as RSEM.count, the predicted by deconvolution and actual rank are highly similar for all genes. In general there is a strong linear relationship between predicted and actual rank, but this is weaker in isoforms than in genes and especially weak in FPKM-generated cufflinks. By this we can state that the use of Cufflinks as a quantification tool greatly deviates predicted gene expression from actual expression in deconvolution approaches. It can also be observed that more poorly expressed genes/isoforms are more likely to deviate from linearity.

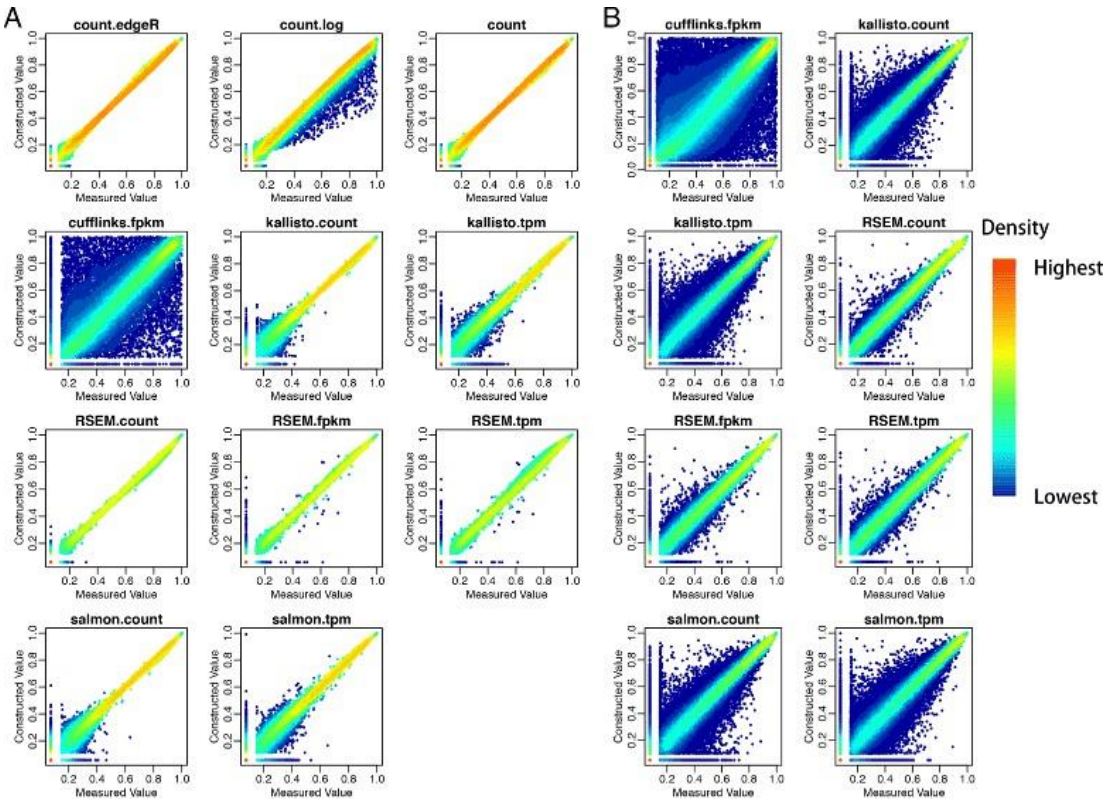


Figure 40. Comparison of expected versus observed gene expression generated by different programs at the gene level (A) and isoform level (B). Here, a gene/isoform's expression rank has been normalised by total number of genes/isoforms. The Y axis indicates the normalised rank as predicted by linear weighting of the pure reference sample expression according to their proportion in the mixture, while the X axis indicates the actual normalised rank of a mixture utilising those proportions. Figure replicated from Jin *et al.*[272].

However, upon using the FPKMs, I discovered some error had evidently occurred during the production of FPKMs from the raw counts. I describe this in more detail in the relevant section.

5.2.1.2 Selection of marker genes

Deconvolution is faster, more accurate, and less variable if only the most useful subset of the pure cell transcriptome is utilised. Many transcripts do not inform deconvolution and may in fact introduce noise, making their exclusion desirable. The most informative transcripts of all for deconvolution are those which are uniquely expressed in a cell type of interest, and which also display low variation across replicates. A minimum level of expression, so as to be free of low count biological noise, is also to be desired. Transcripts such as these are referred to as "markers", although given the paucity of such clear identifiers, in practice transcripts which

show high expression in a single cell type and reliably low expression in others are used as markers[266]. Expression values at both the gene and isoform level can be utilised. Isoform-level markers may add discriminating power, especially where the overall transcriptomic profile of the cells within the convoluted sample is quite similar and may be distinguished more by alternative splicing rather than by alternative gene transcription. I used genes as identifiers, given that the Zhang *et al.* cell types are broadly distinguished and the gene level data is already available.

Zhang *et al.* identified marker genes as having a FPKM >5 and carried out analyses using this benchmark. Given the ready availability of their data I used these marker gene lists as a starting point and further narrowed the lists to increase accuracy.

5.2.1.3 Range filtering of data

Filtering of the upper and lower bounds of marker gene expression prior to deconvolution has been investigated as a means to improve deconvolution accuracy. Genes with low expression can be unreliable, as the variation between cell types might be due to poor transcript detection rather than absence of the transcripts. A practical approach to this was taken by Mohammadi *et al.* in their review of deconvolution methods[134]. They noted that for microarray data, a theoretical upper and lower bound can be established. Upper limits are bounded by microarray sensitivity, while below a certain threshold, the assumption of linearity between transcript prevalence and measured expression has been shown to not exactly hold. However these limits do not offer any practical guidance as most values are not close to them and cannot be excluded on this basis. RNA-Seq data is even more untethered from a fixed standard, as the sequencing of a particular transcript does not become "saturated" as a fluorescent signal can. Mohammadi *et al.* attempted to establish upper and lower limits for marker expression filtering, but often found that these limits diminished the quality of the deconvolution rather than enhancing it. Even a method they described as "adaptive filtering", in which limits were delineated based on sudden changes of expression between a gene and the next highest/lowest expressed gene, was often detrimental in deconvolution quality, although it did occasionally outperform bluntly selecting cut off points without prior information. It appeared that approximately half of the datasets showed improved deconvolution,

while half showed inferior deconvolution (measured by the mean difference between all estimated cell frequencies and their respective true frequencies for all samples).

It is intuitive that range bounding could have a negative effect on deconvolution. The optimal situation would be to have very many moderately expressed, cell type unique, low cell to cell variance transcripts as markers. However, markers can also be highly expressed. A marker gene may well have a biological role in the cell type it marks, and therefore can be highly expressed in order to carry out that role. Range bounding therefore is highly likely to exclude the very genes which are informative marker genes. If the optimal moderately expressed markers were extremely prevalent, then bounding would likely have a positive effect on deconvolution. Improvements in deconvolution from range bounding likely result from the exclusion of genes which might well be cell-specific but are too variable in expression within the cell type they mark, introducing inaccuracy. It is therefore imperative to adopt a heuristic, flexible approach.

Given that even sophisticated range bounding is often detrimental to deconvolution, I opted for a simple approach. For all sets of marker genes for all cell types, I imposed different range bounding, or none, and observed what combinations resulted in the superior deconvolution of *in silico* convoluted pseudosamples (see 5.2.2.1). This approach is conceptually simple and is easily carried out, allowing a full comparison of possible combinations of range bounding.

5.2.2 Selection of deconvolution programme

I searched the literature for a deconvolution software package that I could utilise. DeconRNASeq described by Gong *et al.* was one example. Like most deconvolution methods it assumes that the formula described below accurately describes the relationship of a mixed sample's transcriptional values to those of the pure samples of which it is a mixture[135]. Their algorithm then determines the values for all cell proportions by solving the non –negative least squares problem for each marker transcript. Since all proportions are assumed to add up to 1, the algorithm cannot account for "missing" cell types. The data are also scaled to prevent highly expressed genes from becoming overwhelmingly weighty (since there is a squaring

component). Their paper showed good deconvolution for samples with over 5% prevalence of each cell type.

5.2.2.1 *In silico* generation of 100 convoluted samples

In order to assess the accuracy of the deconvolution, I generated pseudosamples. By producing these pseudosamples, I could assess the accuracy of deconvolution on samples of known mixed proportions. Each pseudosample was comprised of values for all genes measured in the samples described in Zhang *et al.*, so that

$$X_{aj} = \sum_{s=1}^{q} p_s(X_{sj})$$

Where the transcriptional value X (FPKMs, reads per million (RPMs), etc) for gene *j* in pseudosample *a* is equivalent to the linear sum of the respective gene's values from pure samples $X_{s....q}$ provided by Zhang *et al.*, multiplied by their relative proportion $p_{s....q}$, with all proportions adding to 1. The generation of pseudosamples operates under the same linearity assumption which underlies most deconvolution models. Relative proportions were generated randomly in R.

The concept behind utilising pseudosamples was to benchmark the performance of the deconvolution. Although it is impossible to know the true extent of cell proportion changes (if any), the use of pseudosamples is proof that the deconvolution of similar depth samples can be carried out, and that the error between the deconvolution predicted proportions and actual pseudosample proportion can be quantified. As I show in this chapter, use of optimal settings brought the mean absolute difference between predicted proportion and actual proportion to >7% of the actual proportion for deconvolution of mouse cortical, mouse hippocampal, and human iPSC-derived neuronal samples.

5.3 Initial DeconRNASeq and troubleshooting using FPKMs

My initial investigation utilised FPKMs, but did not examine comparison datasets as I did not initially know how to compare these datasets on different scales.

5.3.1.1 Optimising the approach by utilising pseudosamples

I performed DeconRNASeq using different sets of marker genes from Zhang *et al.*. I also applied a number of different "top and tail" filters to the marker gene data, excluding marker genes if their expression in the cell type they mark fell outside of the filter. Different marker gene sets produced by different filters were evaluated on the basis of how well the deconvolution performed on the *in silico* pseudosamples for each cell line.

5.3.1.1.1 Pseudosample results

I used DeconRNASeq to deconvolute my 100 pseudosamples, each generated from random proportions of the enriched cell RNA-Seq FPKM profiles from Zhang *et al.*. The cell types included were all of those investigated by Zhang *et al.*, namely astrocytes, neurons, oligodendrocyte precursor cells, newly myelinating oligodendrocytes, mature oligodendrocytes, microglia, and endothelial cells. I used different marker genes for each cell type, and used varying numbers of marker genes. This first deconvolution was performed three times, once with 40 marker genes per cell type, once with 125, and once with 500. This meant that my "pure" RNA-Seq profiles generated from the data of Zhang *et al.* utilised 280, 875, and 3500 genes in each comparison, as there are seven different cell profiles.

The results of the deconvolution were initially very poor, both overestimating and underestimating the actual proportions for each cell type. In astrocytes, the maximum overestimation was by a factor >300, in a pseudosample with relatively low astrocyte proportions. The standard deviation of the estimated/actual ratio was also very high. An optimum would be an average ratio of 1, and a standard deviation of 0, indicating perfect prediction. The average ratio was typically very high across the 100 samples and across cell types. I found that the standard deviation varied by cell type but no value was lower than 3, indicating the vast majority of samples had extremely poorly predicted levels of each cell type. Briefly experimenting with maximum and minimum expression filters did improve the results, but the range of ratios still varied. Two examples are given in Table 21 and Table 22, for the deconvolutions using 125 marker genes. We can see that the results are initially extremely incorrect, largely overestimating cell types. With filtering, they begin to fail to predict some

cell types' presence at all. We can see as an example that oligodendrocyte precursor cell estimated/actual ratios vary from 0.02 to 3.16 across the 100 pseudosamples. Biologically, variance in cell types to this degree would be grossly pathological. I did not find these results satisfactory and decided that a greater number of measures needed to be taken to ensure accuracy.

	Astrocytes	**Neuron**	**OPC**	**MO**	**Microglia**	**Endothelial Cells**
Max	314.7	66.2	8.57	219.73	23.3	66.09
Min	0.63	0.58	0.102	0.35	0.73	0.53
Standard deviation	31.2	6.61	1.23	21.85	2.62	7.22

Table 21. Average maximum, minimum, and standard deviations of the ratio of estimated:actual proportions of each cell type for 100 pseudosample deconvolutions using 125 marker genes. 1 is optimum for max and min, 0 is optimum for standard deviation. OPC=Oligodendrocyte precursor cell, MO=Myelinating oligodendrocyte

	Astrocytes	**Neuron**	**OPC**	**MO**	**Microglia**	**Endothelial Cells**
Max	3.11	2.1	3.16	3.26	1.16	1.88
Min	0	0	0.03	0.08	0.05	0.09
Standard deviation	0.65	0.48	0.72	0.71	0.23	0.44

Table 22. Average maximum, minimum, and standard deviations of the ratio of estimated:actual proportions of each cell type for 100 pseudosample deconvolutions using 125 marker genes, filtering marker gene expression for those with a value between 5 and 5000 FPKMs in the cell type they mark. 1 is optimum for max and min, 0 is optimum for standard deviation. OPC=Oligodendrocyte precursor cell, MO=Myelinating oligodendrocyte

5.3.1.1.2 Pseudosample results restricting cell type proportions to greater than 0.1

Given the highly inaccurate results of the initial deconvolution, some means of improving the predictions were necessary. Since the estimates were extremely inaccurate, I reasoned that selecting different numbers of marker genes for each line would not by itself improve deconvolution to an acceptable degree.

However I noted that if cell proportions were very low (<5%) the estimated proportions were highly incorrect. Gong *et al.* have also struggled to detect cell types which are less than 5% of the mixed cell population[267]. I therefore decided to optimise deconvolution with the assumption that all cell types would be greater than 10% of the overall population. I believe that this is the best approach as cells with a low proportion will be incorrectly predicted anyway and many research groups have experienced difficulties with predicting these low proportion cell types at all. The increased difficulty in optimising for the <10% scenario offsets the potential gains in accuracy, which would not be great in any case. Finally, I am comparing genotypes for relative changes to cell types, not for absolute prevalence of each cell type. I therefore determined to limit all proportions in my pseudosample generation, so that each cell type contributed a minimum of 10% towards the *in silico* RNA-Seq profile of each pseudosample. I also removed the newly myelinating oligodendrocytes from the pseudosamples, reasoning that these cells would not be prevalent at greater than 10%.

I then reassessed the profiles for marker genes, using data from Zhang *et al.* to find the genes with the largest fold changes between the marked cell type and all other cell types. See Figure 41, Figure 42, and Figure 43 for heatmaps of the marker gene profiles using 40, 125, and 500 markers per cell type (240, 750, and 3,000 markers total). Each marker gene block is distinguished by high expression in the cell type it marks, and is characterised by low expression in other cell types.

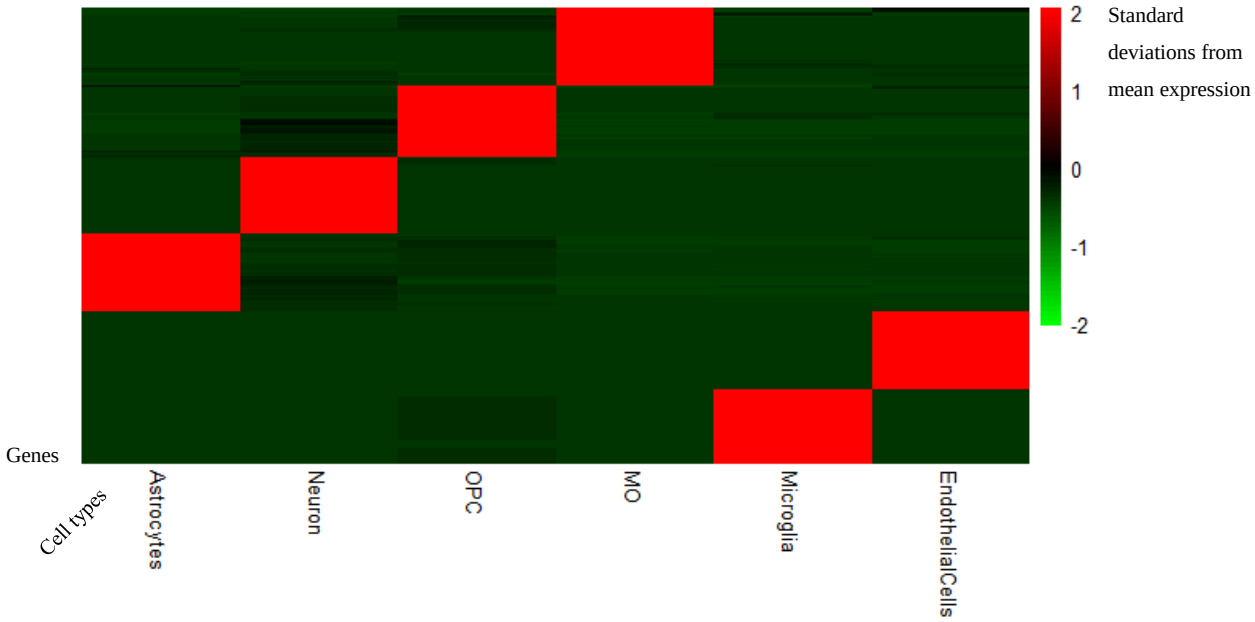


Figure 41. Heatmap of expression for the top 240 marker genes for cell types from Zhang *et al.* with each row representing a gene and columns indicating cell types. Scale is in Z scores (standard deviations from the mean), with red indicating expression higher than the mean of all cell types and green lower. Note that a minority of marker genes for astrocytes and OPCs in particular are not quite as specific. OPC=Oligodendrocyte precursor cell, MO=myelinating oligodendrocyte.

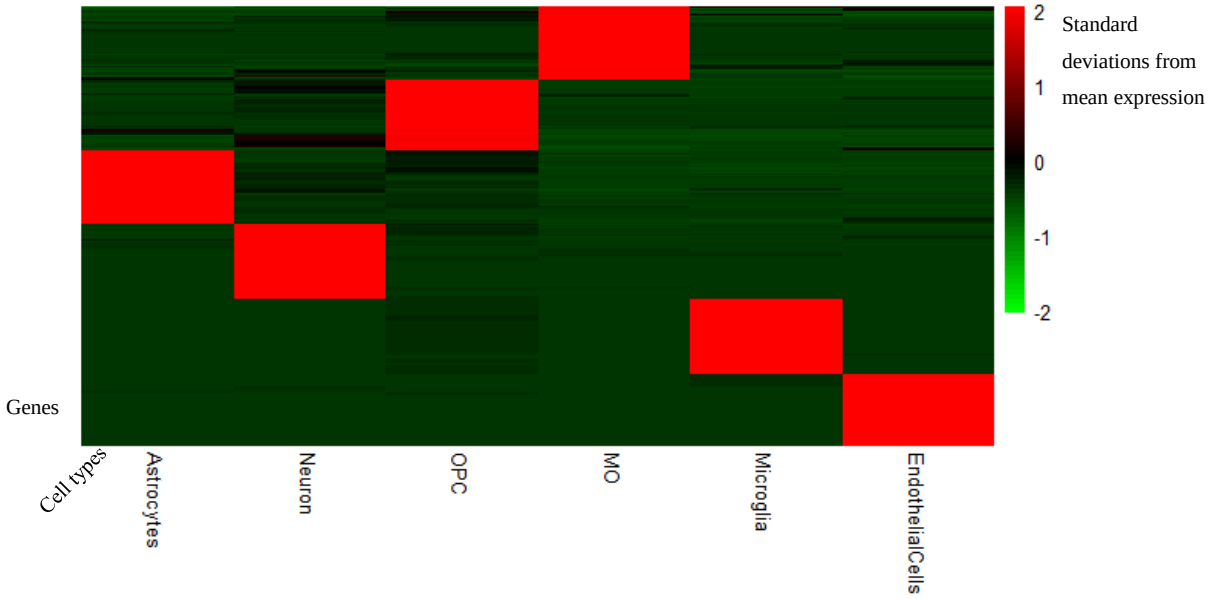


Figure 42. Heatmap of expression for the top 750 marker genes for cell types from Zhang *et al.* with each row representing a gene and columns indicating cell types. Scale is in Z scores (standard deviations from the mean), with red indicating expression higher than the mean of all cell types and green lower. Note that a greater number of genes have moderate expression in cells other than the cell type they mark. OPC=Oligodendrocyte precursor cell, MO=myelinating oligodendrocyte.

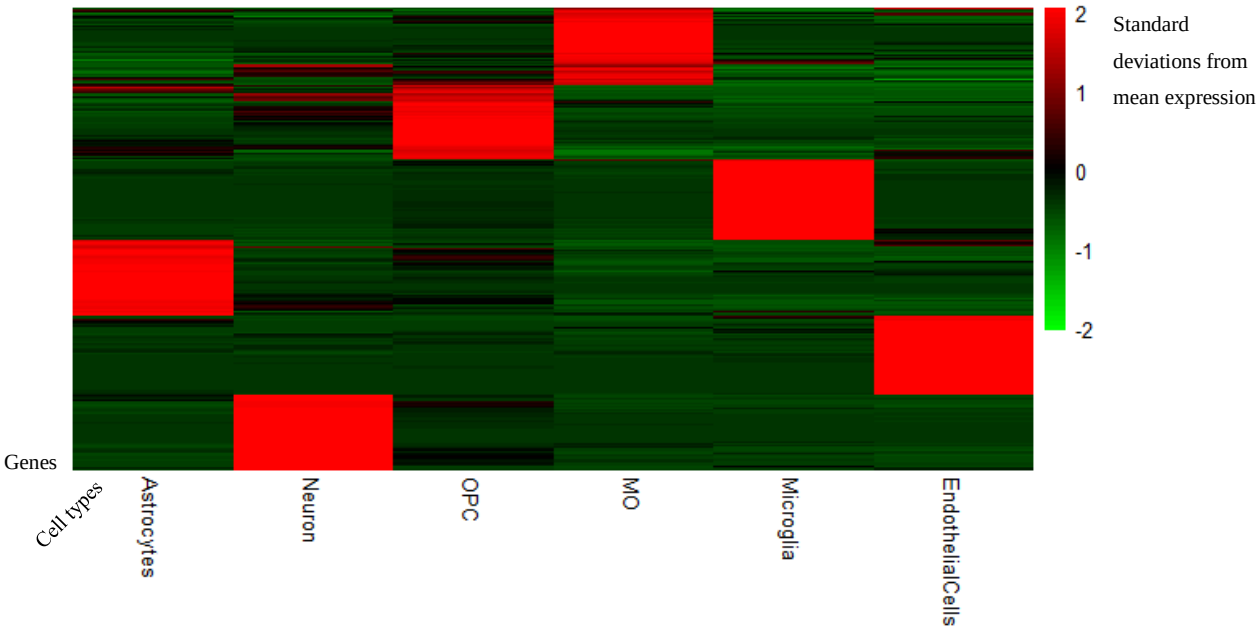


Figure 43. Heatmap of expression for the top 3,000 marker genes for cell types from Zhang *et al.* with each row representing a gene and columns indicating cell types. Scale is in Z scores (standard deviations from the mean), with red indicating expression higher than the mean of all cell types and green lower. Note that a large number of genes have reasonable expression in many cell types. It is unlikely that this will be the number of marker genes which will give optimum deconvolution.

Figure 44 displays the results of the deconvolution, using varying numbers of marker genes for each of the six cell lines. Several facts can be observed. Firstly, a larger number of marker genes appears to introduce greater variation in the ratio of estimated/actual proportions of each cell line. At 40 genes per line, there is very little variation in five out of six cell line estimations, whereas variation is greater at 125 genes (indicated by broader peaks) and is even larger at 500 genes. We can also see that microglia, indicated by orange, are often overpredicted, except when 500 genes are used. Finally, when using 500 genes for each line, DeconRNASEQ fails to detect endothelial cells (indicated by green) at all.

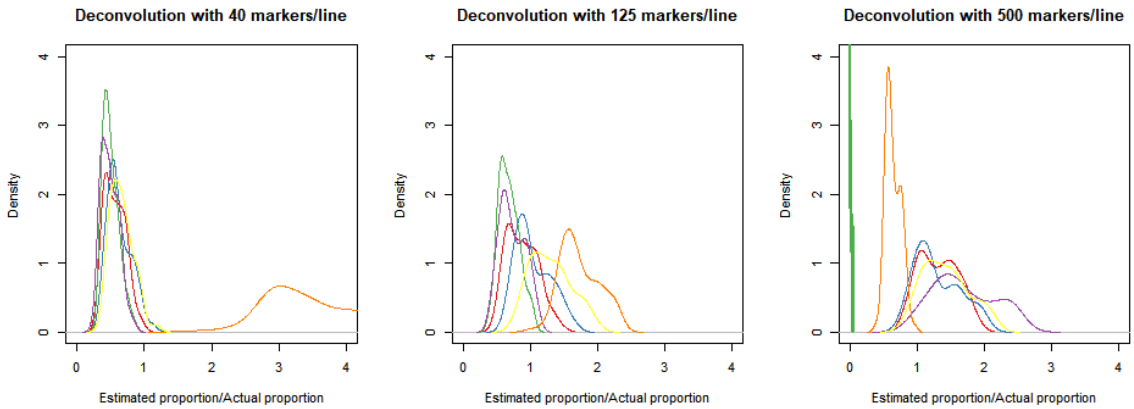


Figure 44. Deconvolution of 100 pseudosamples, shown as a density plot. Each cell type is represented by a different colour; Astrocytes=Red, Neurons=Blue, Oligodendrocyte Precursor Cells=Green, Myelinating Oligodendrocytes=Purple, Microglia=Orange, Endothelial Cells=Yellow. X axis = deconvolution prediction of cell proportion divided by actual cell proportion. Optimum is 1. Y axis = frequency of this ratio, displayed as a density plot. Each graph indicates the results using a different number of marker genes for each cell line, always using the markers with the largest fold changes.

We can conclude that restricting each cell type's proportion to 10% or greater exerted a large effect on the accuracy of the deconvolution, and that an increase in the number of marker genes appeared to initially bring the deconvolution ratios closer to one, but also introduced other variation.

5.3.1.1.3 Pseudosample results filtering marker genes, restricting cell type proportions to greater than 0.1

I aimed to further increase the accuracy of the deconvolution by combining the previous steps, but using differing numbers of marker genes, with a new step, filtering marker genes for high and low expression. I evaluated deconvolution with different variations of these criteria. Mohammadi *et al.* had shown that expression filtering improved deconvolution only about half of the time, so it was doubtful whether I would observe any improvements[134]. A subset of the results is displayed in Figure 45.

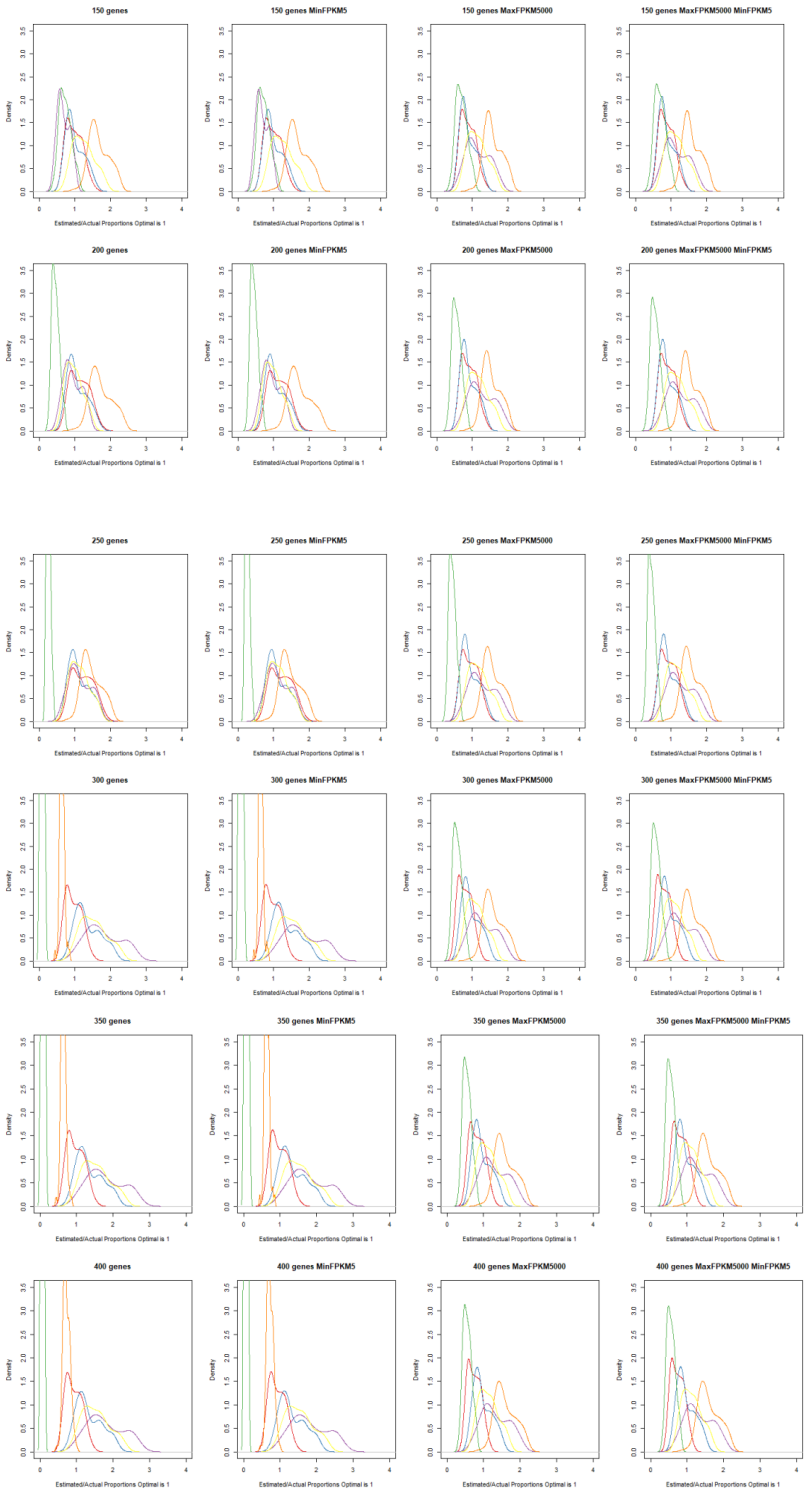


Figure 45. Results of the deconvolution using between 150 and 400 marker genes per line. In all cases the genes with the greatest fold changes between the marked line and other lines were utilised. Filters for expression are indicated in the titles of each graph by MaxFPKM, indicating the maximum expression allowed for a marker gene in the cell it marks, and MinFPKM, the relative minimum expression. Each cell type is indicated by a different colour; Astrocytes=Red, Neurons=Blue, Oligodendrocyte Precursor Cells=Green, Myelinating Oligodendrocytes=Purple, Microglia=Orange, Endothelial Cells=Yellow.

There are several findings to be observed. Firstly, if the number of marker genes per cell line is low, microglia, indicated by orange, are overestimated. Secondly, if the number of marker genes per cell line is high, oligodendrocyte cells, indicated by green, are not detected at all, indicated by the line being at 0. This is undone by excluding marker genes with an expression above 5000 FPKM in the line they mark. Excluding genes with less than 5 FPKMs appeared to result in no change. On further investigation it was evident this was due to a lack of marker genes with expression below this level. Several settings are about equal in accuracy, not overestimating or underestimating any lines to a degree of >4 or <0.25. These include all 100, 150, and 200 gene settings, and all higher marker gene number settings which have the FPKM 5000 maximum expression filter. Given the ambiguous results of expression filtering, I opted to not continue using it.

5.3.1.1.4 Initial deconvolution of mouse cortical samples using FPKMs

I deconvoluted my mouse cortical RNA-Seq profiles to investigate whether any of the six cell types were changed due to the *Der1* mutation. I used different numbers of marker genes as there were no clearly superior sets. I used both Cufflinks-generated FPKM values to deconvolute my data. Using FPKMs is a natural choice as the data described in Zhang *et al.* are in FPKMs. However, as we shall see, it became evident to me that a serious error had occurred in the generation of the *Der1* mouse cortical FPKMs.

In the process of the FPKM deconvolution I noted that a minority (<10%) of the marker genes had expression in all the mouse samples which was greater than the expression in my pure Zhang *et al.* samples. This is likely due to a number of factors; different sequencing depths, different origin of RNA (brain tissue vs enriched plated cells). Given that these markers are a minority I opted to remove them, as they clearly invalidate the assumption that the markers will have their maximum expression in the cell type that they mark. My concerns about this process were a major factor in later deciding to utilise housekeeping gene normalisation (see 5.4 for a full discussion). I emphasise that no particular cell type was singled out to have certain markers removed-all markers with greater expression in the samples versus the cell type they mark were removed.

Deconvolution of the RNA-Seq data using Zhang et al. Cell type enriched datasets

5.3.1.1.4.1 Results

Regardless of the number of marker genes utilised, we can see a distinctive split in the samples (see Figure 46). One group of samples has a moderate proportion of many cell types, and is reported as being approximately 50% neurons; close to reported biological reality[273,274]. The second is reported as being approximately 70% myelinating oligodendrocytes. The two groups are not separated by genotype, with the oligodendrocyte heavy group consisting of two wild-type, three heterozygous, and two homozygous mouse samples. The two groups are also not distinguished by the sex of the mice. We can also see that there is little variation in the estimation of all cell types between samples of different genotypes within the two groups. We can also see that the number of marker genes seems to have little if any effect on the estimations of the various cell type proportions.

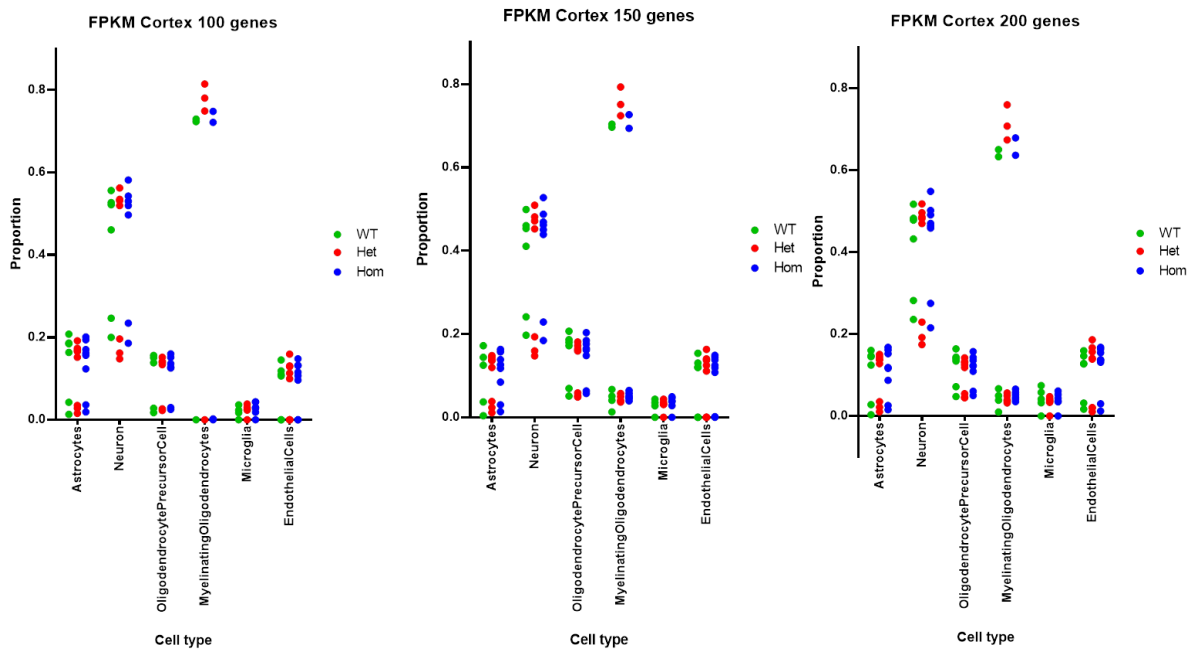


Figure 46. Deconvolution of the *Der1* mouse cortical sample FPKMs using 100, 150, or 200 marker genes. Genotypes are differentiated by colours.

The reported proportion of a number of the samples being mostly myelinating oligodendrocytes when using FPKMs is very surprising. This proportion is not close to most histological estimates by researchers looking at mouse cortex, and seems biologically impossible. It also seems unlikely that such enormous variation could

present in such a binary fashion within genotypes. I investigated the samples further and found that the FPKM expression of the myelinating oligodendrocyte marker, *Mbp*, encoding myelin basic protein, varied enormously and in a binary fashion. The samples reporting high myelinating oligodendrocyte proportions had high FPKM expression of this gene (it is the highest expressed gene of all), while those reporting lower proportions had near nil expression. The fold change between the two groups was over 800. I subsequently inspected the RNA-Seq counts directly and found that all samples had similar count values (variation <5%). It is not possible that the FPKMs of the samples are genuinely different to this degree. It was evident that some error had been made in the processing of the RNA-Seq data which caused *Mbp* to have large variation between samples despite the counts varying very little.

Removal of *Mbp* from the list of marker genes resulted in predicted proportions which were biologically plausible and similar in all samples, regardless of genotype (data not shown). The extremely potent power of *Mbp* is likely due to its high expression, as well as the stark contrast between the samples. I have not yet discerned what the processing error is and the miscounting of *Mbp* calls into question the Cufflinks-generated FPKMs from *Der1* cortex as the units of deconvolution, as many other genes may also have been affected.

5.3.1.1.4.2 Rejection of FPKM deconvolution as a valid approach

The *Der1* FPKM deconvolution is clearly untrustworthy. Although I have noted a problem with *Mbp*, other problems may still be present. It is also true that Cufflinks-FPKMs are the worst possible units to carry out deconvolution with, as shown by the research of Jin *et al.* (see Figure 40) showing these units give the highest variation in deconvolution estimates compared to other measurements[272].

However, RPM units may not be ideal either, as the data described in Zhang *et al.* are in the form of Cufflinks-generated FPKMs. Since FPKMs are normalised to both transcript length and sequencing depth, the scale of FPKMs will be entirely different to the scale of RPMs which are normalised to the number of base reads. It is also not a simple matter to relate the two, as Cufflinks does not utilise a predetermined set of transcript lengths but rather empirically determines the length of transcripts from the

data. It is possible given Mbp's complex splicing arrangement that the error in quantifying the transcript occurs due to different estimated transcript lengths.

Finally, there is the question of the <10% of genes that were removed because their expression in the samples was greater than their expression in pure cell types. This was a concerning finding, and was a major reason for my redoing the deconvolution in the next section, using housekeeping gene based normalisation, which I believed would result in all genes being on a similar scale of expression. This would also allow me to look at other datasets to assess the accuracy of the deconvolution.

It was also evident to me that there were other issues. Although I had generated pseudosample data and carried out the deconvolution using DeconRNASeq, I had not at any stage verified the deconvolution by testing it on a secondary dataset. Going forward, I would also need to do the deconvolution using RPMs, as there had clearly been an error with FPKM generation. I would need to develop a method allowing the comparison of multiple data types so as to continue using the Zhang *et al.* dataset. I describe in the next sections my solutions for these three challenges.

5.4 Mouse Cortical RNA-Seq deconvolution

5.4.1 Housekeeping gene normalisation to compare multiple datasets

5.4.1.1 Introduction

To compare the Zhang *et al.* dataset to others, I needed some method of normalisation. I decided to see if housekeeping gene based normalisation could work to ensure easy comparison between pseudosamples and an exterior dataset. I could empirically test this by comparing the efficiency of housekeeping gene normalised pseudosample deconvolution to non-normalised deconvolution. I utilised an objective method to do this.

5.4.1.2 Selection of housekeeping genes

Housekeeping genes (HKs) should be as uniformly expressed as possible, as well as being expressed in all cell types. In order to gauge the suitability of genes as housekeeping genes, I calculated the coefficient of variation (standard deviation

divided by mean, referred to as CV) for each gene across all Zhang cell types and within my *Der1* mouse samples. Each Zhang *et al.* cell type has only two high depth replicates, so it seemed wise not to average by cell type. I calculated two CVs: the ZhangCV, the CV for all of Zhang's samples for a gene, and the mouseCV, the CV of the gene across my cortical *Der1* mouse samples. The geometric mean of these two values was calculated and used to rank the genes as putative housekeeping genes. I also restricted housekeeping gene selection to those genes which had a no greater than fourfold difference between the maximum and minimum values in both the mouse and Zhang sets. This seemed like a good approach to filter out otherwise good housekeeping genes which had extraordinary expression in a single mouse sample or cell culture, which would result in this cell type or sample giving erroneous results. Genes which were not universally expressed were also discarded as potential housekeeping genes. The criteria for expression were more counts than *Disc1* in the case of the cortical mouse samples and at least 5 FPKM in the Zhang samples (I chose this as Zhang *et al.* had chosen it as the cut off level for marker gene expression). Of the >19,000 genes expressed in both sample sets, approximately 3800 m*et al.*l the criteria. The geometric means varied from 0.004 to 0.356. I selected housekeeping genes from this list, starting with those with the lowest geometric mean.

I subsequently used GOrilla to check the ranked housekeeping gene list for overrepresented ontologies. I expected that overrepresented gene functions would relate to typical housekeeping gene functions such as translation, transcription, and cell metabolism. If ontologies related to particular brain cell types appeared, such as synapse formation (related to neurons, but not to microglia) then this would be alarming and would indicate that my selection of housekeeping genes was faulty.

For the Cortical vs Zhang comparison, the top 10 ontologies overrepresented at the top of the list of ranked potential housekeeping genes can be seen in the Appendix. Although Process and Component were as expected for housekeeping genes, I was unsure of what some of the Functional ontologies related to. The BAT3 complex appears to be involved in the insertion of proteins into the ER membrane[275]. Agmatinase is an enzyme which produces precursors for polyamines, while its target

agmatine is brain-expressed and may have a role in depression. However Agmatinase is expressed in a wide variety of cell types according to Meylan *et al.*[276], so I decided to retain it. The two genes related to angiostatin binding are ATP-synthase subunits, which would likely be expressed in a non-cell specific manner. To summarise, with the possible exception of Agmatinase, there did not appear to be any alarming gene ontologies among my putative housekeeping genes. I also manually looked at the top 100 genes, which housekeeping genes were selected from. Three seemed initially alarming as their functions might be related to neuronal activity; these were *Wisp1*, *Creb3*, and *Fxr2*. However the first two appear to function in stress response, while *Fxr2'*s functions are not yet well understood. Nevertheless the low variation in all the genes across all samples means that there are unlikely to be any issues in normalisation.

5.4.1.3 Measures of error

I needed a more objective method of analysing pseudosample deconvolution. In all my deconvolution optimizations, I analysed error using two measures. The mean absolute difference, MAD, is given by the following formula:

$$MAD_i = \sum_{s=1}^{q}((x_{si}/p_{si}) - 1)/q$$

So that the MAD for pseudosample i, MAD_i , is equivalent to the sum of the deviations of the predicted to actual ratios of cell types s…q from one, divided by the total number of cell types. However, since I used 100 pseudosamples, MAD will always, unless specifically stated otherwise, refer to the average MAD of 100 pseudosamples.

The other measure I used was root mean squared error, RMSE, given by the following formulae:

$$\mathrm{RMSE_i} = \sum_{\mathrm{s=1}}^{\mathrm{q}}(\mathrm{x_{si}}/\mathrm{p_{si}})^2$$

$$\mathrm{RMSE} = \left(\frac{\sum_{i=1}^{j} \mathrm{RMSE_i}}{qj}\right)^{1/2}$$

Therefore the RMSE for a deconvolution is the square root of the sum of the squares of the relative difference between the predicted and actual proportions across all cell types across all pseudosamples. Both MAD and RMSE are discussed in Mohammadi *et al.*[134]. MAD in particular seemed like an intuitive way to describe the accuracy of a deconvolution in a single figure.

5.4.2 Deconvolution using housekeeping gene normalised data of Zhang et al.

5.4.2.1 Initial deconvolution

Initially, I used a number of marker genes, ranging from 25 to 500 in increments of 25, and numbers of housekeeping genes ranging from 10 to 100 in increments of ten. This meant that a total of 200 deconvolutions were carried out to determine the optimum settings. Since range filtering had not been especially helpful previously, I elected not to use this. This decision is in line with the evidence provided by Mohammadi *et al.*[134].

The results showed superior deconvolution to non-housekeeping gene normalisation. Figure 47 shows the effect of varying Marker Gene#, while Figure 48 shows the effect of varying Housekeeping Gene#. The best deconvolution is shown in detail in Figure 49. The minimum MAD was 0.059 (marker #=25, HK=40) and the maximum was 0.096 (marker #=500, HK=90).

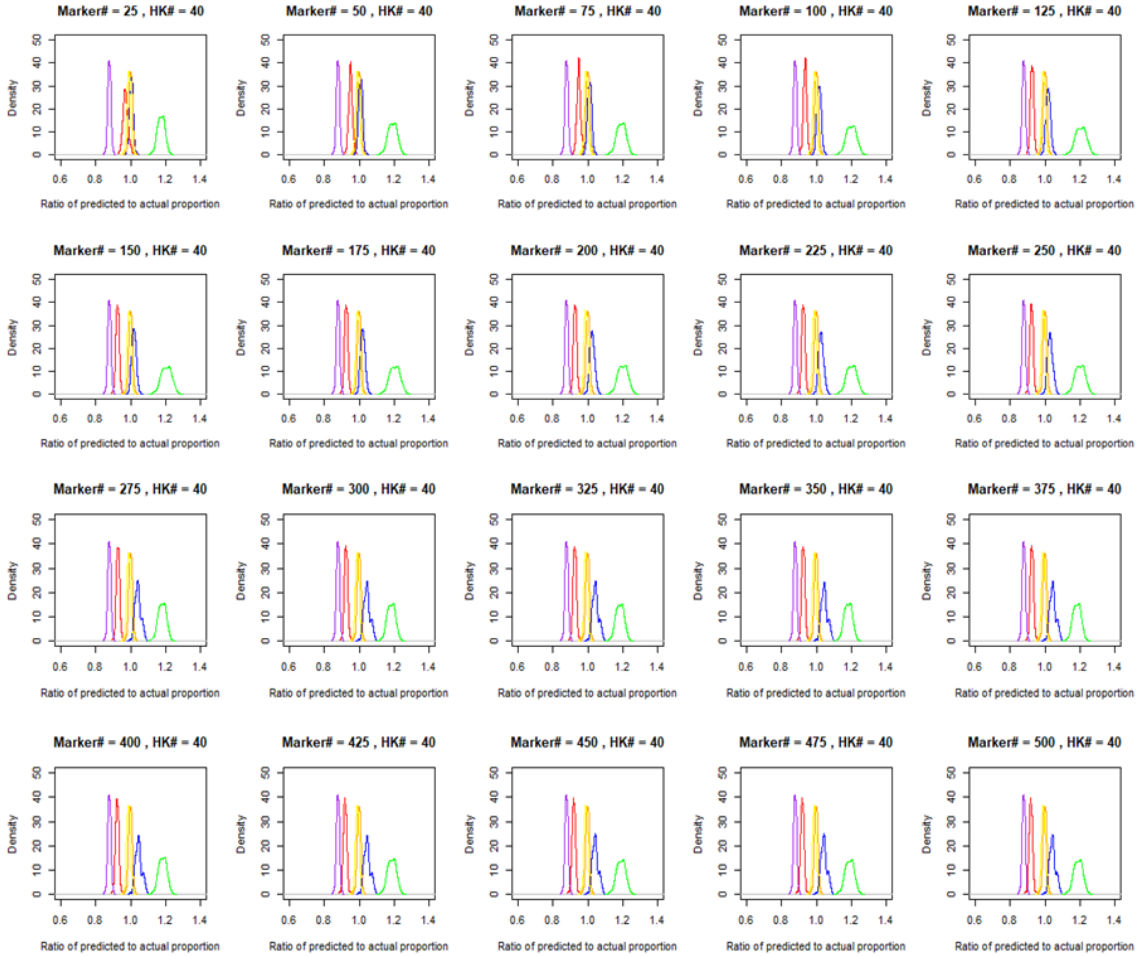


Figure 47. Set of ratio of predicted to actual proportions for 100 pseudosamples against density of estimates, for the deconvolutions where HK#=40 and Marker gene # varies from 25 to 500. Optimum is 25 as shown in more detail in Figure 49. Each cell type is indicated by a different colour. Astrocytes=Red, Neurons=Blue, Oligodendrocyte Precursor Cells=Green, Myelinating Oligodendrocytes=Purple, Microglia=Orange, Endothelial Cells=Yellow. The ideal scenario would be a straight line at 1, indicating perfect invariant deconvolution.

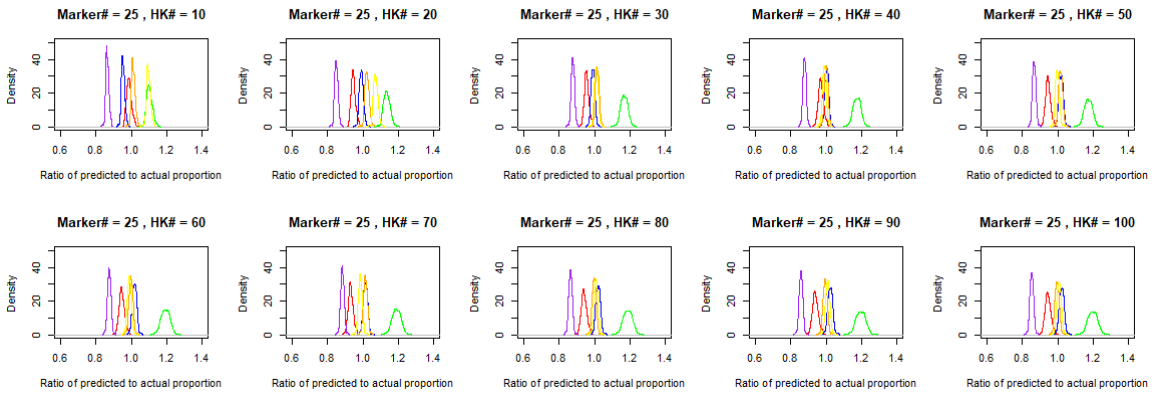


Figure 48. Set of ratio of predicted to actual proportions for 100 pseudosamples against density of estimates, for the deconvolutions where Marker gene #=25 and HK # varies from 10 to 100. Optimum is 40 as shown in more detail in Figure 49. Each cell type is indicated by a different colour. Astrocytes=Red, Neurons=Blue, Oligodendrocyte Precursor Cells=Green, Myelinating Oligodendrocytes=Purple, Microglia=Orange, Endothelial Cells=Yellow. The ideal scenario would be a straight line at 1, indicating perfect invariant deconvolution.

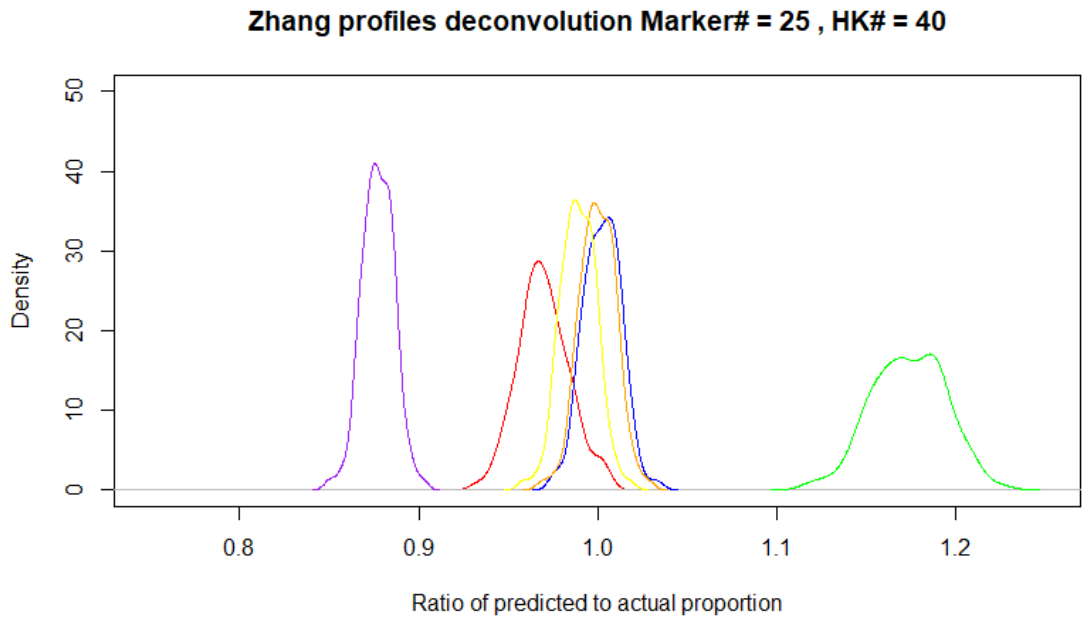


Figure 49. Detailed look at ratio of predicted to actual proportions for 100 pseudosamples against density of estimates, for the optimum deconvolution where HK#=40 and Marker gene#=25. Each cell type is indicated by a different colour. Astrocytes=red, Neurons=Blue, Oligodendrocyte Precursor Cells=Green, Myelinating Oligodendrocytes=Purple, Microglia=Orange, Endothelial Cells=Yellow.

As the number of marker genes increased the optimum HK decreased. The MAD values were low enough that I felt further optimisation was not necessary; an error range within 6% is quite good, and the limiting factors beyond this are unlikely to be resolved by better deconvolution parameters. Importantly, the estimations also

cluster relatively well, even when inaccurate. For example, in the marker#=25 KH=40 deconvolution, myelinating oligodendrocytes are consistently underestimated, but the estimates are always between 84 and 91% of the actual proportion, a quite small range. Since it is differences in relative cell type proportions between samples I am looking for, not absolute differences, it is more important that the deconvolution treats highly similar cells in the same way, although estimates varying between 97 and 104% would obviously be more ideal. We can also see in Figure 47 that the deconvolution of astrocytes, neurons, endothelial cells and microglia is very good, with the estimates for these four cell types varying between 93% (an astrocyte estimate) and 104% (a neuron estimate) of the actual proportions.

5.4.3 Deconvolution of comparison datasets to verify deconvolution

To ensure that my deconvolution was accurate it was imperative to test it on other datasets where the cellular composition has already been determined. Since the composition of my experimental samples was unknown, there was no other way to know whether the deconvolution was accurate. I searched for comparison datasets which fulfilled some of the criteria described in 5.2. They had to have similar RNA-Seq depth, and they had to have identified the cells in a trustworthy manner. I deemed it of less importance that the same dataset provide all cell types, although this would obviously be optimal. I could not find a truly independent dataset providing data from all brain cell types at an appropriate sequencing depth. I did find several which, together, provided pure RNA-Seq profiles of neurons, oligodendrocytes, astrocytes, microglia, and endothelial cells, from both human and mouse and at various ages and disease states. I decided to use these data sets to determine whether my deconvolution was accurate. If it was, it should predict each pure cell profile from each of these experiments as being entirely or almost entirely of that cell type.

5.4.3.1 Zhang Two

I had decided to use the dataset described in 5.2 by Zhang *et al.* for carrying out deconvolution[153]. The same research group released a follow up paper in 2016, looking at a number of human and mouse astrocytes from various disease states.

They also used immunopanning to isolate neurons, oligodendrocyte precursor cells, myelinating oligodendrocytes, and endothelial cells and subjected these to RNA-Sequencing (at a lower sequencing depth than previously). In some cases, different antibodies were used in the immunopanning process, but they were often the same as in the prior paper (CD45 for microglia, BSL-1 for endothelial cells, O4 for oligodendrocyte precursor cells). In particular, it should be noted that the antibody used to select astrocytes was specifically selected using the information from the previous paper by Zhang *et al.*. They also found, as expected, that each cell uniquely expressed a set of classic markers along with the marker used to isolate it. For example, neurons isolated using Thy1 expressed *Vglut1*, *Stmn2, Syt1*, and *Syn1* at high levels, while other cells did not express these.

The second paper by Zhang *et al.* is not a completely independent comparison. Being from the same research group, many of the culture techniques and methods of cell isolation are identical, and the data from the first paper even informed the selection of the astrocyte marker. Accurate deconvolution of these cells will show that the deconvolution is applicable to another closely related dataset, although testing against a more independent dataset is also necessary to show it is also applicable to our samples. Nevertheless this is an important step to show basic reliability as well as aid in troubleshooting. For the sake of clarity,the second dataset described by the 2016 paper by Zhang *et al.*[271] will be referred to as "Zhang Two".

I carried out a series of deconvolutions of the Zhang Two dataset using all possible settings and examined the predictions of each cell types. In total, there are 6 astrocyte samples of varying ages and sorting methods, along with two of each other cell type and three whole cortex samples. I made several observations. Firstly, regardless of marker gene or housekeeping gene number, neurons, oligodendrocyte precursor cells, myelinating oligodendrocytes, microglia, and endothelial cells are highly predicted as being their respective cell type. Usually the prediction was >95%, with the exception of one myelinating oligodendrocyte sample which consistently remained at 80%. The main changes observed were in astrocytes and in the whole cortical samples. The differences between deconvolutions with the same marker gene number but different housekeeping gene numbers was difficult to discern. Changing

housekeeping gene number appeared to have relatively little effect at any stage, unlike marker gene number.

At 300 marker genes and over, astrocytes began to be poorly predicted, with predictions between 50-90%. Above 425 markers, this dropped to 50-70%. This was also seen for the 125 marker gene deconvolutions, regardless of housekeeping gene number, where the astrocyte prediction for astrocytes was 60-90%. The rest of the prediction was typically a mix of endothelial cells and microglia. A few astrocytes (the immunopanned ones) always had less than optimal deconvolution, even at lower marker gene numbers. They were typically predicted as 60-80% astrocytes. However between 150-275 markers they tended to be better predicted with less variation between immunopanned samples and values above 80%. Examples of all cell types are given in Figure 50.

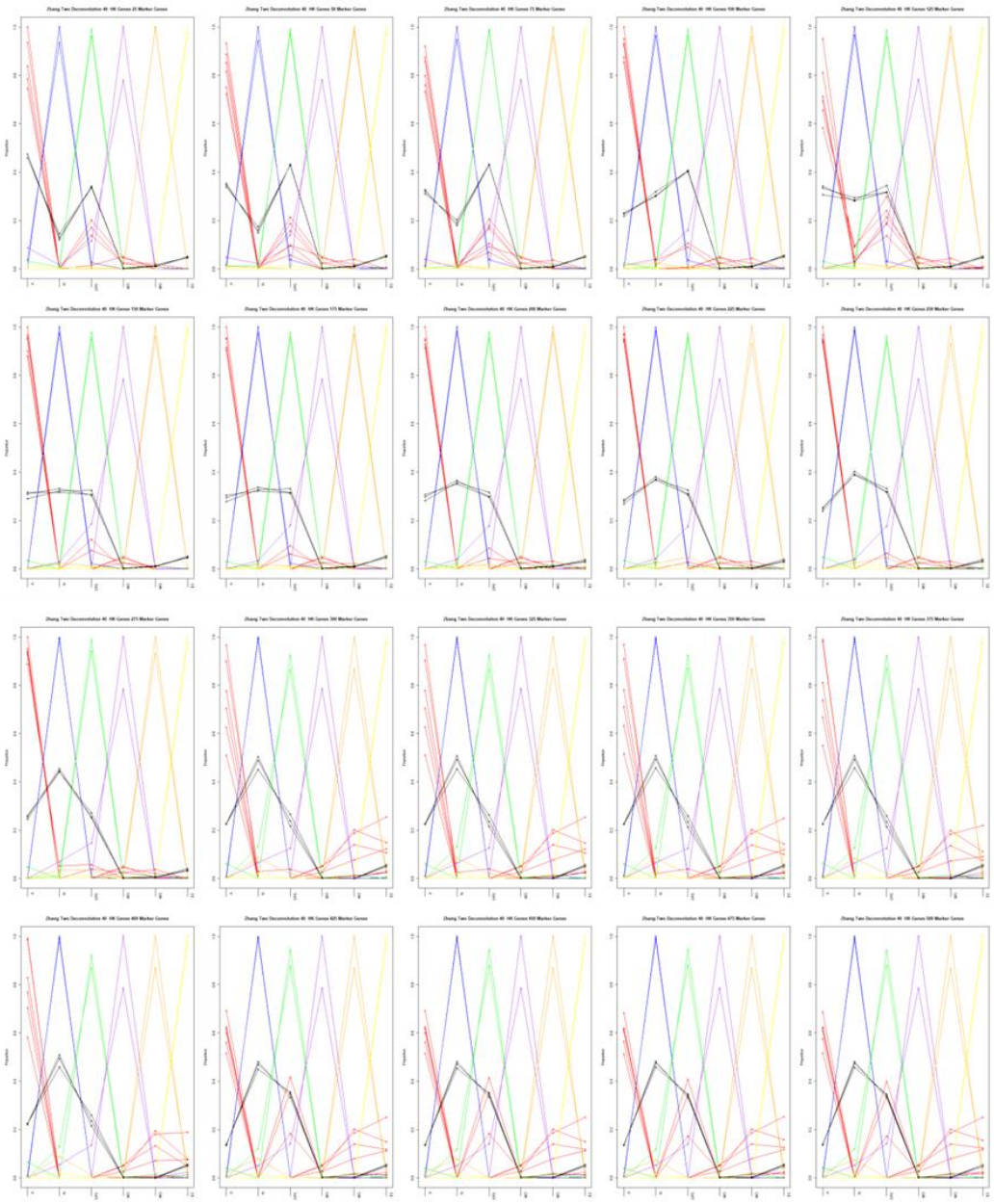


Figure 50. Set of predictions of Zhang Two dataset for the deconvolutions where HK#=40 and Marker gene # varies from 25 to 500. Optimum is 25. Each cell type is indicated by a different colour. Astrocytes=Red, Neurons=Blue, Oligodendrocyte Precursor Cells=Green, Myelinating Oligodendrocytes=Purple, Microglia=Orange, Endothelial Cells=Yellow, Whole Cortex=Black. Predicted cell types are on X axis and proportions are on the Y.

Whole cortex samples followed a pattern as well, which is described in Table 23.

Deconvolution of the RNA-Seq data using Zhang et al. Cell type enriched datasets

Marker Genes	Astrocyte	Neuron	OPC
25 – 75	45% dropping to 30%	12%	35% increasing to 42%
100	20%	30%	40%
125	30%	30%	30%
150 – 500	30% gradually dropping to	40% to 45%	20% to 30%

Table 23. Description of proportions predicted by DeconRNASeq when varying marker gene number. OPC= oligodendrocyte precursor cell. All percentages approximate averages of the three samples across all housekeeping gene numbers.

The results of the deconvolution are displayed in Table 24 using the optimum settings of marker gene number=25 and HK#=40.

Sample cell type	Predicted proportion of this cell
FACS sorted astrocytes	0.9995959
FACS sorted astrocytes	0.9997599
Immunopanned astrocytes 1 month	0.9261619
Immunopanned astrocytes 4 month	0.7781610
Immunopanned astrocytes 7 month	0.8459271
Immunopanned astrocytes 9 month	0.7388576
Neuron	0.9220760
Neuron	1.0000000
OPC	0.9897386
OPC	0.9650586
MO	1.0000000
MO	0.7890155
Microglia	0.9984338
Microglia	0.9997204
Endothelial Cell	0.9654983
Endothelial Cell	1.0000000

Table 24. Results of the deconvolution of the Zhang Two dataset using cortical optimum settings. OPC=Oligodendrocyte precursor cell, MO=Myelinating oligodendrocyte

We can see that in general the cells are predicted very well, with 10 of 16 cells predicted as a mixture comprised of >95% of the actual cell type. Three of the most

problematic predictions are of aged astrocytes, so it is possible that the maturation process results in more distinctive cell profiles due to differentiation. The best deconvolution setting for the 16 Zhang Two samples is HK=10, Marker Gene=225, which gives the lowest deviation from 1 for the 16 cell samples (average deviation 3.8%). Regardless of HK number, Marker Gene=225 is always best.

Three datasets from whole cortex from the Zhang Two paper were also deconvoluted and had highly similar profiles to one another. For the optimal deconvolution, the estimates were of a mixed cell profile consisting of an average of 49% astrocytes, 12% neurons, 30% oligodendrocyte precursor cells, 1.6% myelinating oligodendrocytes, 1.4% microglia and 5.7% endothelial cells (does not sum to 1 due to rounding up). The average difference between the highest and lowest estimates for the three samples for all cell types was 0.9% and the largest was 2.5%, for neurons.

However, these levels are at odds with the expected proportions for mouse cortical samples. It has been estimated by modern counting methods, using isotrophic fractionation and calculating the resulting neuronal:non-neuronal nuclei ratio, that the mouse cerebral cortex consists of 54% neurons by number, and 68% neuron by mass[274]. Other estimates looking at a variety of sources, including staining, estimate that mouse cortical neurons are slightly less, at 42%. It must be noted that the staining densities used to estimate this have high variability[273]. It is also true that more endothelial cells should be predicted; no setting predicts more than a small minority of endothelial cells. It is possible that they do not contribute many reads to the RNA-Seq, either due to loss during tissue harvesting or little transcriptomic activity. With this in mind, I reassessed the results of the Zhang Two deconvolution and have given a full discussion in the section before the *Der1* deconvolution.

The best deconvolution of the Zhang pseudosamples was obtained using HK=40, Marker Gene=25, MAD=0.059. In the Zhang Two dataset this gave a deviation of 6.7% from perfect prediction of the 16 individual cell samples, and predicted whole cortex as being ~12% neuron, ~45% astrocyte, and ~35% oligodendrocyte. The best deconvolution of the Zhang Two individual cell samples was obtained using HK=10, Marker Gene=225. This gave a MAD=0.073, and a deviation of 3.8% from perfect

prediction of the 16 Zhang Two samples. However, using these settings the prediction of mouse whole cortex samples was closer to values suggested by empirical data and informed the settings I eventually used.

5.4.3.2 Dorsal Root Ganglion Neurons

I also found a dataset generated by Li *et al.* in 2016, describing the RNA-Seq results of approximately 200 dorsal root ganglion neurons from WT mice[156]. Dorsal root ganglion neurons are a type of sensory neuron found in the spine; so these cells are clearly distinct from the cortical and hippocampal cells we have sequenced. The comparison is therefore less than ideal, but if the deconvolution can accurately identify these neurons it is evidence that the general neuron markers are acceptable. It also indicates that the markers for other cell types do not have appreciable expression either in the cortical/hippocampal cells they will be optimised for, or in these more distinct sensory neurons. There are other arguments for using this dataset as a comparison for the Zhang deconvolution. The sequencing depth is an average of 58.2 $x10^6$ mapped reads, highly comparable to the average of 65.6 $x10^6$ sequenced reads described by Zhang *et al.* (on average 87% of these map, giving around $57x10^6$ mapped reads). The comparison is therefore very apt. Another reason is the quantity of cells. Li *et al.* carried out this high depth sequencing on nearly 200 neurons. I can therefore be confident in the accuracy of the deconvolution, should it reliably identify these neurons as being a "mixture" comprised mainly of neurons.

I carried out deconvolution of the roughly 200 neurons described in the Li *et al.* 2016 paper[156] using the same spread of HK and marker gene numbers. Each RNA-Seq sample was normalised to total mapped reads, and then to the geomean of the housekeeping gene expressions. Some samples could not be included in the analysis. Since the housekeeping normalising factor is the geometric mean of multiple housekeeping expressions, if any one of these is zero then the normalising factor is nonsensical. A minority of samples, between 10-20% depending on housekeeping gene number, do not express every housekeeping gene and therefore become excluded from the analysis. With increasing housekeeping gene number, this proportion increases. The designation of a gene as "housekeeping" was defined by

the original Zhang sample and the mouse *Der1* cortical samples, so it is inevitable that some cells are excluded.

Regardless of the number of housekeeping genes used, a distinct trend was evident. All ~200 samples had very similar values for the predicted proportion of neuron across all settings. This proportion was typically low, especially if only a few markers were used. If 25 markers (150 total since there are 6 cell types) were used, the overall average predicted neuron proportion across all samples was 0.102 to 0.111, depending on the housekeeping gene number. This increased to an average of 0.164 at 50 markers (the average of approximately 1800 measurements, 180 in each housekeeping gene number group). At 75 this was 0.217, but at 100 it leaped to 0.5948 and increased unevenly, with another leap to 0.899 at 300 markers. A graph of all averages is seen in Figure 51.

The number of housekeeping genes has little effect on the average amount predicted to be neurons. This is a reasonable finding; 10 is a large number of housekeeping genes to normalise to initially. Since these genes are chosen specifically because of minimal variation in Zhang *et al.* and in our samples, it is reasonable that they show less variation in this new group of samples too (as all three sample groups are of the same tissue and species). I was surprised that there seemed to be no effect of housekeeping gene number at all; to investigate further I looked at the standard deviation of the neuron proportion across the >150 neuron deconvolutions. The result of this is displayed in Figure 52. We see that in all cases, having only 10 housekeeping genes for normalisation results in greater deviation in the proportion predicted to be neuron, although the average does not change much. It also appears that having 100 is inferior to having 20-80. It is also notable that the standard deviation drops considerably at the 300 marker gene level, where the prediction of average neuron proportion reaches its peak.

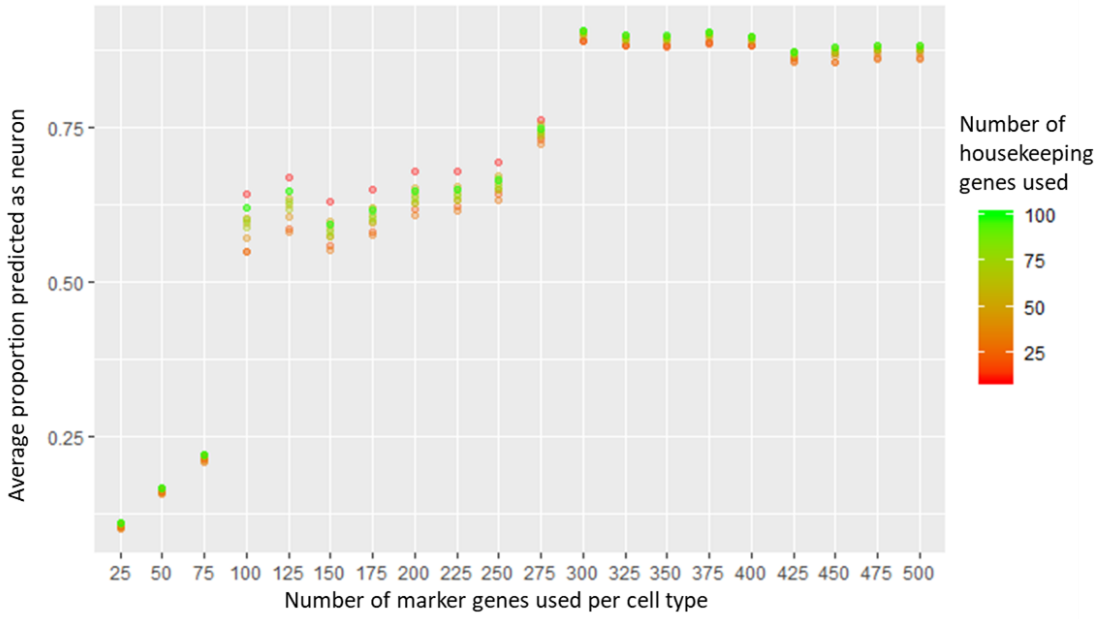


Figure 51. Graph displaying how neuronal prediction of >150 dorsal root ganglion RNA-Seq profiles using Zhang *et al.* cell types varies with housekeeping and marker gene number. The Y axis indicates the average proportion predicted to be neuron across >150 dorsal root ganglion RNA-Seq profiles; if identification is always perfect, it would be 1. The X-axis indicates the quantity of marker genes (per cell type) used in the deconvolution, while colour indicates the number of housekeeping genes utilised. Since six cell types were used, the total marker gene number varies from 150 to 3,000. Housekeeping genes chosen using Zhang and mouse cortical datasets.

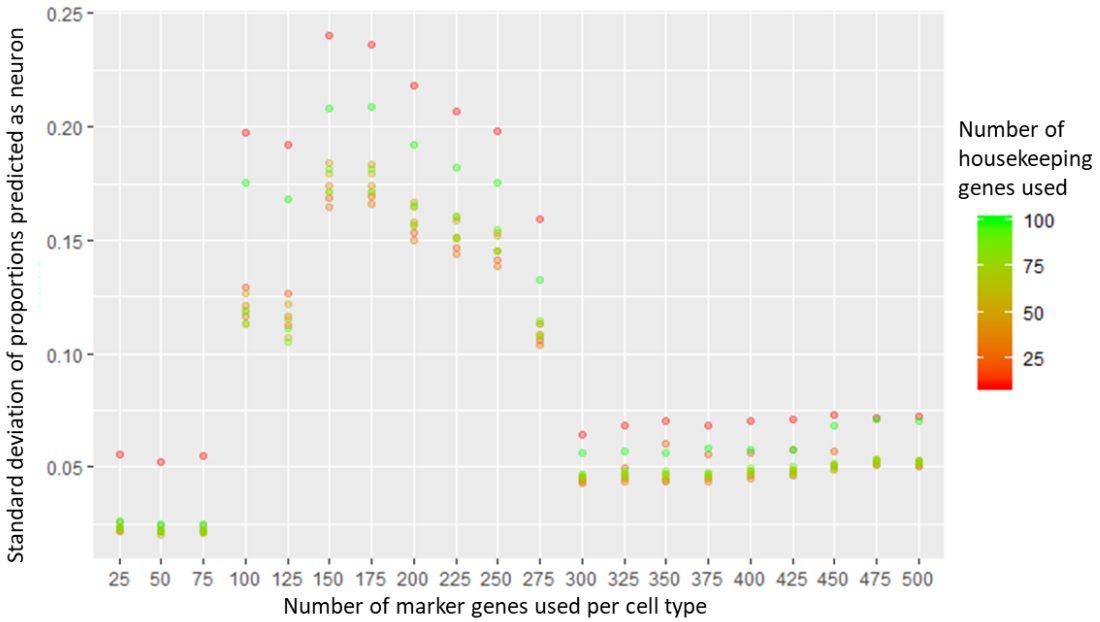


Figure 52. Graph displaying how neuronal prediction of >150 dorsal root ganglion RNA-Seq profiles using Zhang *et al.* cell types varies with housekeeping and marker gene number. The Y axis indicates the standard deviation in the proportion predicted to be neuron across >150 dorsal root ganglion RNA-Seq profiles; if identification is always perfect, it would be 1. The X-axis indicates the quantity of marker genes (per cell type) used in the deconvolution, while colour indicates the number of housekeeping genes utilised. Since six cell types were used, the total marker gene number varies from 150 to 3,000. Housekeeping genes chosen using Zhang and mouse cortical datasets.

Similar graphs to Figure 51 and Figure 52 looking at the maximum and minimum predicted proportion show similar findings. Although the maximum predicted proportion is of course 1 (with most HK numbers achieving this at 100 markers and all by 175), the minimum predicted proportion remains below 0.4 until 300 markers, where it increases to an average of about 0.7 across HK numbers, with the outlier of 10 HK which remains with a minimum of about 0.5. We can conclude that in general more markers are better, with a particular leap at the 300 marker level. As to why this is, it is possible that the markers selected prior to the 300 level are those which relate to function found in cortical neurons specifically. Markers were chosen on the basis of the Zhang *et al.* dataset; comparing cortical neurons to cortical astrocytes, glia, etc. The top neuron markers include genes such as *Reln*, *Dlx2*, *Nkx2.1* as well as other *Lhx* and *Dlx* genes. These genes are known to have roles in the formation of the cortex and therefore may not be particularly well expressed in terminally differentiated sensory neurons. This appears to be the case; when ranking genes by expression, the neuronal marker genes occupy much higher rankings in the Zhang

dataset than they do in this sensory neuron dataset compared in both cases to the list of all expressed genes (see Table 25).

	Rank of expression in Zhang Neurons	**Rank of expression in Sensory Neurons**
Reln	2	17
Sst	6	18
Npy	8	50
Uchl1	18	58
Stmn3	25	71
Stmn2	27	80
Atp1a3	31	92
Gap43	32	134
Tubb3	36	139
Nsg2	37	142
Snhg11	43	147

Table 25. Comparative ranking of marker genes in Zhang and sensory neuron datasets, for the top 10 marker genes expressed in both datasets. It can clearly be seen that the genes have higher expression in the Zhang dataset. The average position of the first 10 genes is 22.2 vs 80.1, the first 100 is 526 vs 1333, and all genes have average rankings of 4746 vs 8251 in the Zhang and sensory neuron datasets, respectively.

The pressing question is how to integrate this information into the selection of the optimal deconvolution. Looking at the DRG deconvolution using 300 marker genes, housekeeping gene number does not greatly alter the average neuron proportion. However, the optimal number of marker genes in the deconvolution of my pseudosamples was 25, with HK=40, at a MAD of 0.059. There is no conflict with HK number as 40 is not inferior to any other number in the DRG deconvolution. Using 300 marker genes takes the MAD up to 0.074 (this is across all cell types, not just neurons).

I conclude that my explanation regarding neuron marker genes is likely why more markers are better in the deconvolution of the DRG neurons, and less is better for the pseudosamples. We can see that the change in deconvolution of the pseudosamples is not severe (MAD goes from 5.9% to 7.4%) with these extra markers. A side by side

comparison is given in Figure 53. It is likely that if I found a more apt dataset, I could use fewer markers and securely know that the deconvolution is adequate. However a full discussion in the context of the Zhang Two deconvolution is given in the next section.

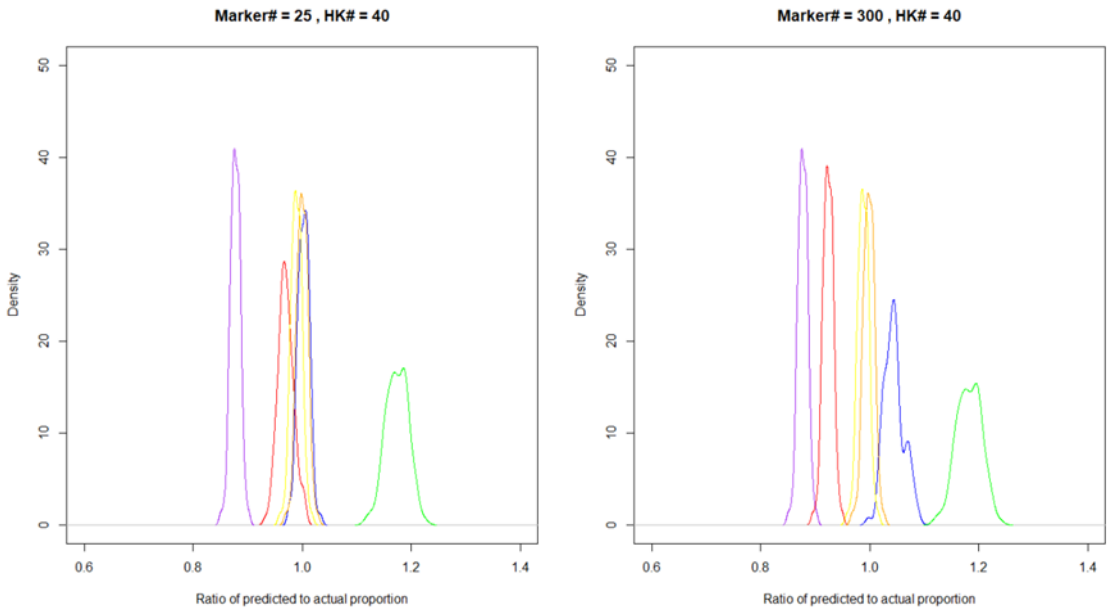


Figure 53. Detailed look at ratio of predicted to actual proportions for 100 pseudosamples against density of estimates for two deconvolutions. Labels above indicate housekeeping gene (HK) and marker gene numbers. HK#=40 and Marker gene#=25 is the overall optimum in terms of MAD. Each cell type is indicated by a different colour. Astrocytes=Red, Neurons=Blue, Oligodendrocyte Precursor Cells=Green, Myelinating Oligodendrocytes=Purple, Microglia=Orange, Endothelial Cells=Yellow.

5.4.4 Deconvolution of mouse Der1 cortical samples

It was important to evaluate all the evidence I had received in choosing the optimal settings for the deconvolution. There are three sets; the Zhang pseudosample deconvolution, the Zhang Two deconvolution of comparative samples and whole cortex, and the Dorsal Sensory neuron deconvolution. There are several factors to consider; how a deconvolution setting alters predictions of the 100 pseudosamples, how it alters identification of the Zhang Two samples and neurons, how it predicts the whole cortical samples and whether these predictions are similar to biological reality.

The Zhang pseudosample deconvolution revealed good deconvolution with all settings, from the best (HK=40, Marker Gene=25) with a MAD of 0.059, to the worst (HK=90, Marker Gene=500) with a MAD of 0.095. In general, MAD increased slowly and steadily with increasing Marker Gene number and with deviation of HK from 40. The conclusion is that the settings should be kept as close to HK=40, Marker Gene=25 as possible but the worst possible settings are only 50% less accurate than the best.

The Zhang Two dataset gives the conclusion that most cells are well identified, but over 300 marker genes results in poor astrocyte identification. Correspondingly, there is a major jump in inaccuracy with the increase from 275 to 300 marker genes (from average of 5.2% to 13.3% across housekeeping gene settings). More important is the relation to whole cortex samples, which our *Der1* samples will presumably be similar to. In general, higher markers give more neuronal proportion in the cortical samples, with 275 markers giving 40% neuron and the highest, 500 markers, giving 45% where HK=10.

The Dorsal Sensory neuron deconvolution confirms a finding already seen in the Zhang and Zhang Two deconvolutions; changing housekeeping gene number does not substantially alter the results in any way. However, it states that neuronal prediction leaps up at 300 markers, and above. As discussed in that section, this is possibly due to the lower expression of the markers in these more distinctive sensory neurons.

Given the uncertainties inherent in the deconvolution, I concluded that the best approach was to select a variety of high accuracy settings, starting with the marker gene number first, then the housekeeping gene number. A finding that appeared in all of the settings would be reliable. 275 marker genes gives the most accurate prediction of whole cortical samples with neuronal proportions approaching biologically plausible levels, but does not veer into underestimation of astrocytes. 275 is also a better predictor of the dorsal sensory neurons than lower marker gene numbers, although 250 is nearly as accurate. For both the Zhang and Zhang Two deconvolutions, the optimal HK is 10 for 275 marker genes, with a MAD of 6.9%

and a deviation from perfect prediction of 4.5%, respectively. HK=40 is a close contender in each case.

I therefore deconvoluted the *Der1* mouse cortical samples using six different settings, HK=10, HK=40 and Marker gene #=250, 275, or 300. I deconvoluted the depth normalised count data. The results can be seen in Figure 54.

Importantly, we can see that the samples are behaving the same way, proving that the existence of two groups seen in the FPKM deconvolution, one with many MOs, is an artefact of the process of generating Cufflinks-FPKMs. Removing *Mbp* has little effect on the deconvolution (data not shown); as expected given its low variance. This confirms that the issue seen before was due to that effects of FPKM generation. The higher accuracy of this deconvolution, as shown by superior performance in pseudosamples, makes this a trustworthy result.

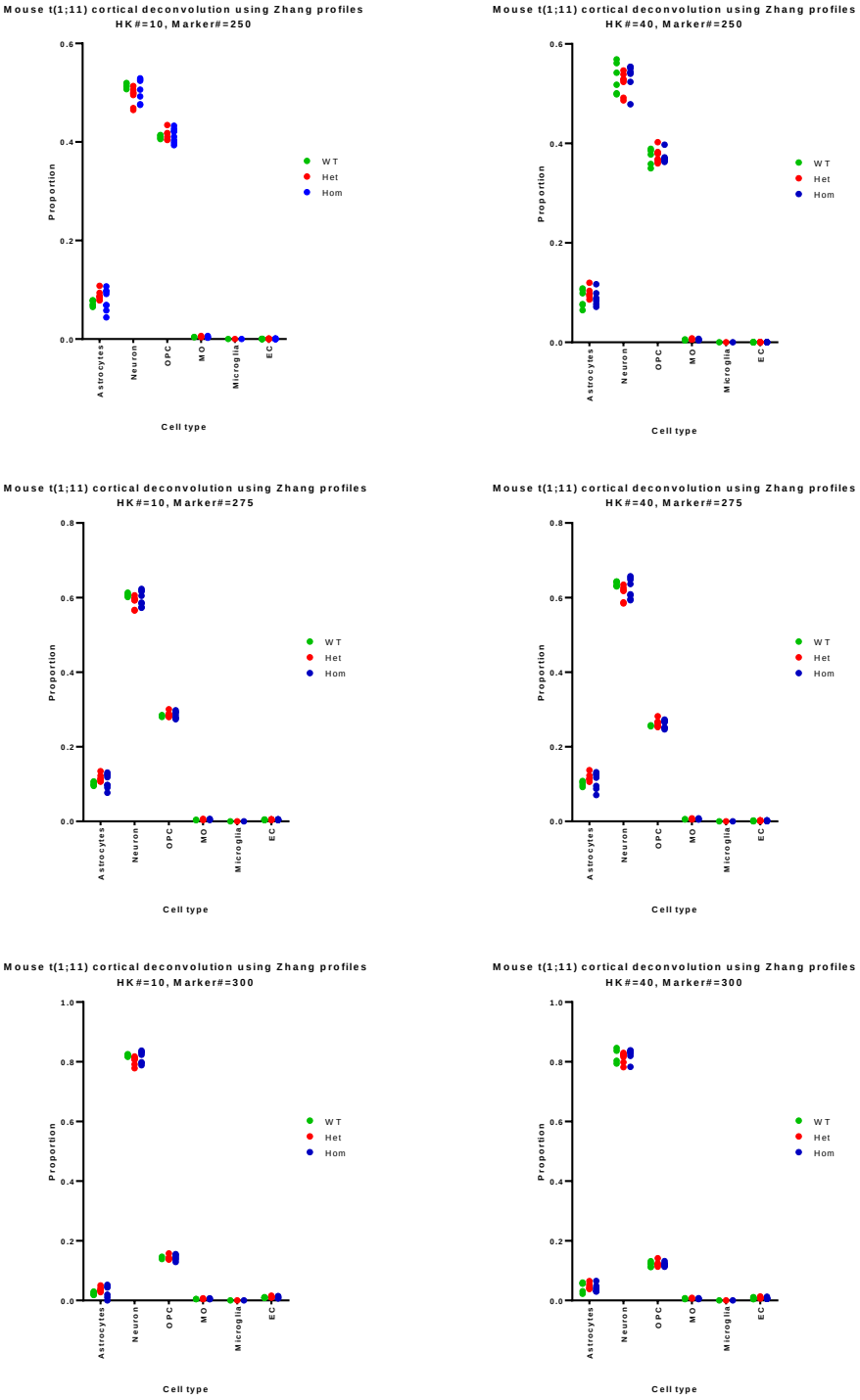


Figure 54. Deconvolution of mouse *Der1* cortical samples, showing cell types against proportions for four settings varying in marker gene and housekeeping gene numbers. WT=Wild-type, Het=Heterozygous, Hom=Homozygous, colours are green red and blue respectively. OPC=Oligodendrocyte precursor cell, MO=Myelinating oligodendrocyte, EC=Endothelial Cell

An ANOVA for each cell type was performed to determine the effects of genotype on each of the cell proportions. There was a significant effect of genotype on the proportion of astrocytes in all comparisons except for the HK=40, Marker Gene=250, and of neuron proportions in half of the deconvolutions. Post-hoc testing using Tukey's multiple comparisons test found no significance for pairwise comparisons between any genotype in any deconvolution.

As described in the corresponding chapter, there had appeared to be an internal structure within the homozygous samples. I therefore looked at these two groups, distinguished as shown by PCA in Chapter 2. Figure 54 is repeated in Figure 55 with the split in the homozygotes highlighted. I again carried out ANOVAs for each cell type and subsequent pairwise comparison testing, splitting the Homozygotes into two groups. There was a significant effect of genotype on the proportion of oligodendrocyte precursor cell proportions in the HK=10, Marker Gene=250 comparison, but post-hoc testing using Tukey's multiple comparisons test found no significance for pairwise comparisons between any genotype.

The two homozygote groups are shown alone in Figure 56. Pairwise t-tests for each cell type revealed significant differences in neurons, astrocytes, and myelinating oligodendrocytes in all HK=10 deconvolutions, and the HK=40, Marker Gene=275 deconvolution.

P values	**HK10M250**	**HK10M275**	**HK10M300**	**HK40M250**	**HK40M275**	**HK40M300**
Astrocyte	0.0012	0.0008	0.0004	ns	0.0012	ns
Neuron	0.0018	0.00043	< 0.0001	ns	0.00019	ns
MO	0.0028	0.0027	0.0025	ns	0.0030	ns
EC	ns	ns	0.0004	Ns	ns	ns

Table 26. Results for homozygote pairwise t-tests, with Sidak-Bonferroni correction for multiple testing within each deconvolution. MO=Myelinating Oligodendrocytes, EC=Endothelial Cells, ns=non significant

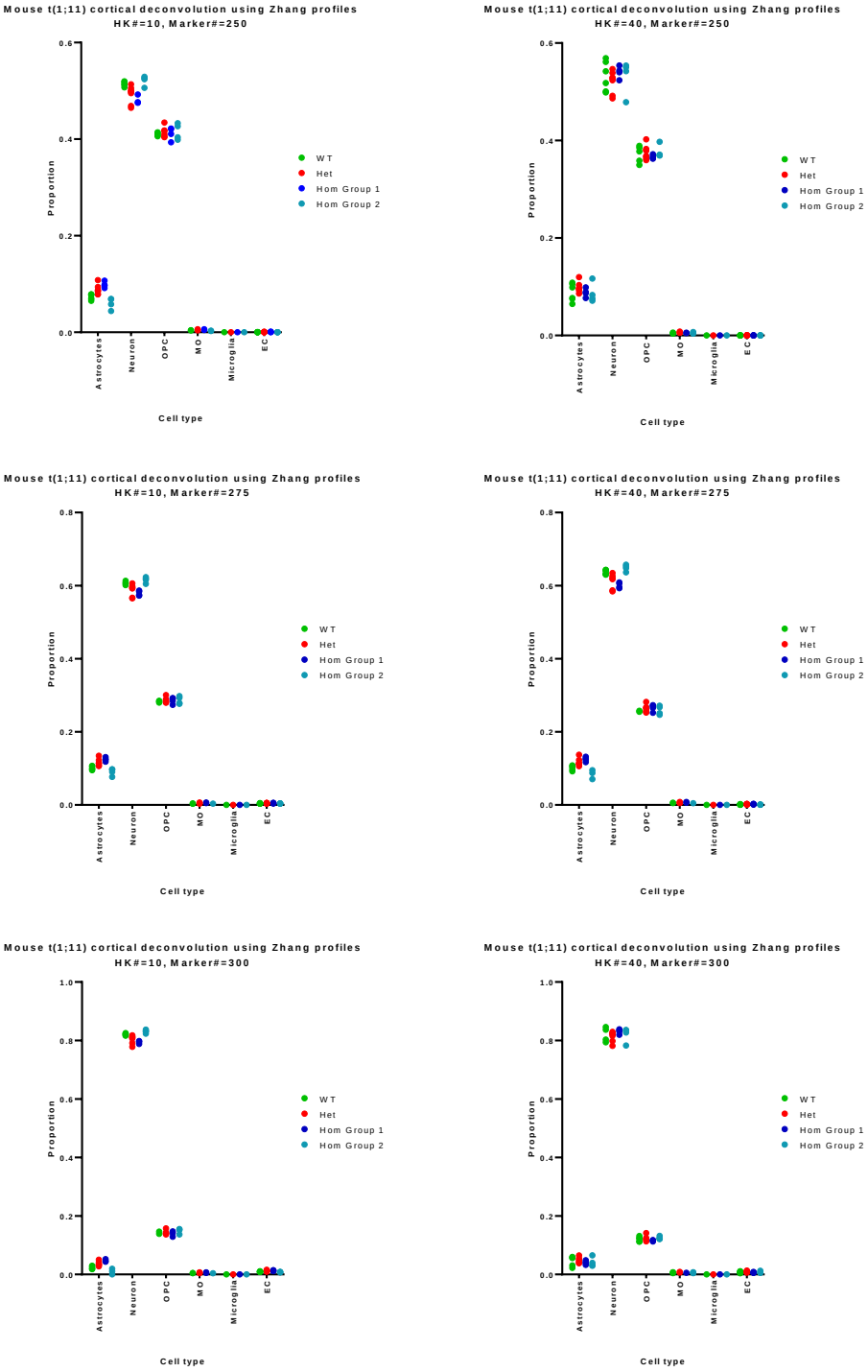


Figure 55. Deconvolution of mouse *Der1* cortical samples, showing cell types against proportions for four settings varying in marker gene and housekeeping gene numbers. WT=Wild-type, Het=Heterozygous, Hom Group 1=Homozygous Group 1, Hom Group 2=Homozygous Group 2, colours are green red, dark blue and light blue respectively. OPC=Oligodendrocyte precursor cell, MO=Myelinating oligodendrocyte, EC=Endothelial Cell.

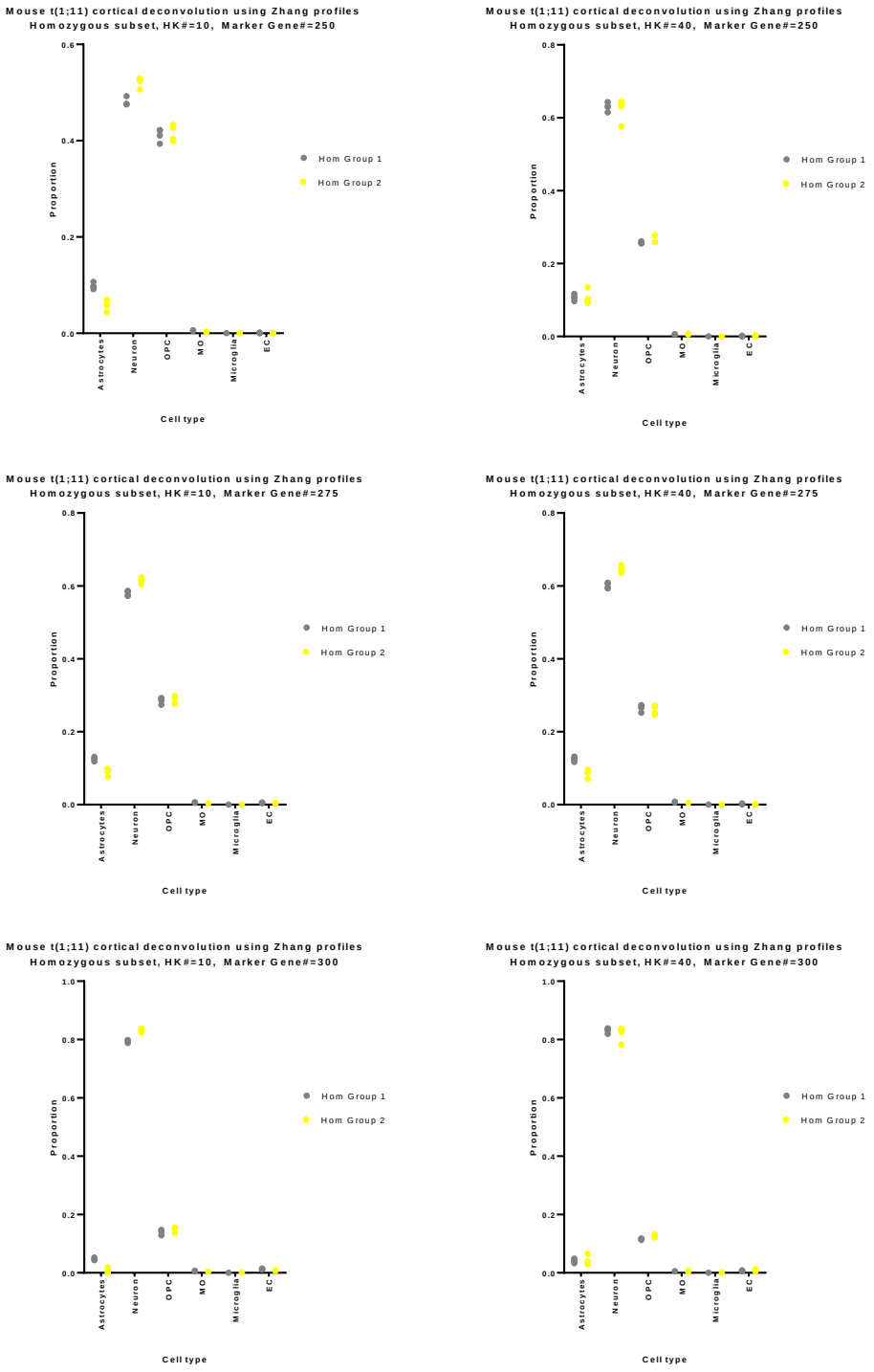


Figure 56. Deconvolution of mouse *Der1* cortical samples, showing cell types against proportions for four settings varying in marker gene and housekeeping gene numbers. Hom Group 1=Homozygous Group 1, Hom Group 2=Homozygous Group 2, colours are grey and yellow respectively. OPC=Oligodendrocyte precursor cell, MO=Myelinating oligodendrocyte, EC=Endothelial Cell.

5.5 Mouse Hippocampal RNA-Seq deconvolution

Since the cortical deconvolution had low MAD for the pseudosamples and Zhang Two dataset, I continued on to the *Der1* hippocampal samples, utilising the same methods. The difference here is that I am comparing the hippocampal samples to cortical datasets, and I am making the assumption that the cell types will closely relate across these two brain regions. Since FPKMs were confirmed as not being reliable, I moved straight to housekeeping-normalisation based deconvolution of the depth normalised counts. I also did not use marker gene range filtering as it had not been useful in the cortical deconvolution. Although the marker genes are the same, the housekeeping genes will be different as they are selected based on the enriched cell profiles and the mouse samples. Both marker genes and housekeeping genes were selected as in the cortical deconvolution.

For six WT samples 83.877 ± 13.58 million reads were sequenced, for eight heterozygous *Der1* samples 82.81 ± 13.84 million reads were sequenced, and for eight homozygous *Der1* samples 79.9 ± 14.08 million reads were sequenced.

Housekeeping genes were filtered for universal expression, less than fourfold variation, and an average minimum expression. A similar number of genes matched these criteria as in the cortical housekeeping gene selection. Of approximately 17,000 genes expressed in both sample sets, 3,832 m*et al.*l the criteria and were ranked in order of the geometric mean of both coefficients of variation. This list was subjected to ranked GO term analysis using GOrilla. The results of this are displayed in the Appendix. As with the cortical set, there were no alarming results. None of the terms related to specialised components of nervous cells, particular processes unique to a subset of them, or functions which are other than general housekeeping ones.

5.5.1 Deconvolution using housekeeping gene normalised data from Zhang et al.

The measure of inaccuracy used was MAD. Since Zhang *et al.* only used a few cell types, graphs were also utilised and examined by eye.

5.5.1.1 Initial deconvolution

As in the cortical deconvolution, I used housekeeping gene numbers from 10,20,30…100, and marker gene numbers 25,50,75….500 for a total of 200 deconvolutions. The results were of a similar quality to the cortical deconvolution and the min and max MAD values were similar. The minimum MAD was 0.0498 (marker=25, HK=10) and the maximum was 0.103 (marker=500, HK=100). Figure 57 and Figure 58 show the effects of varying marker gene number and HK, respectively.

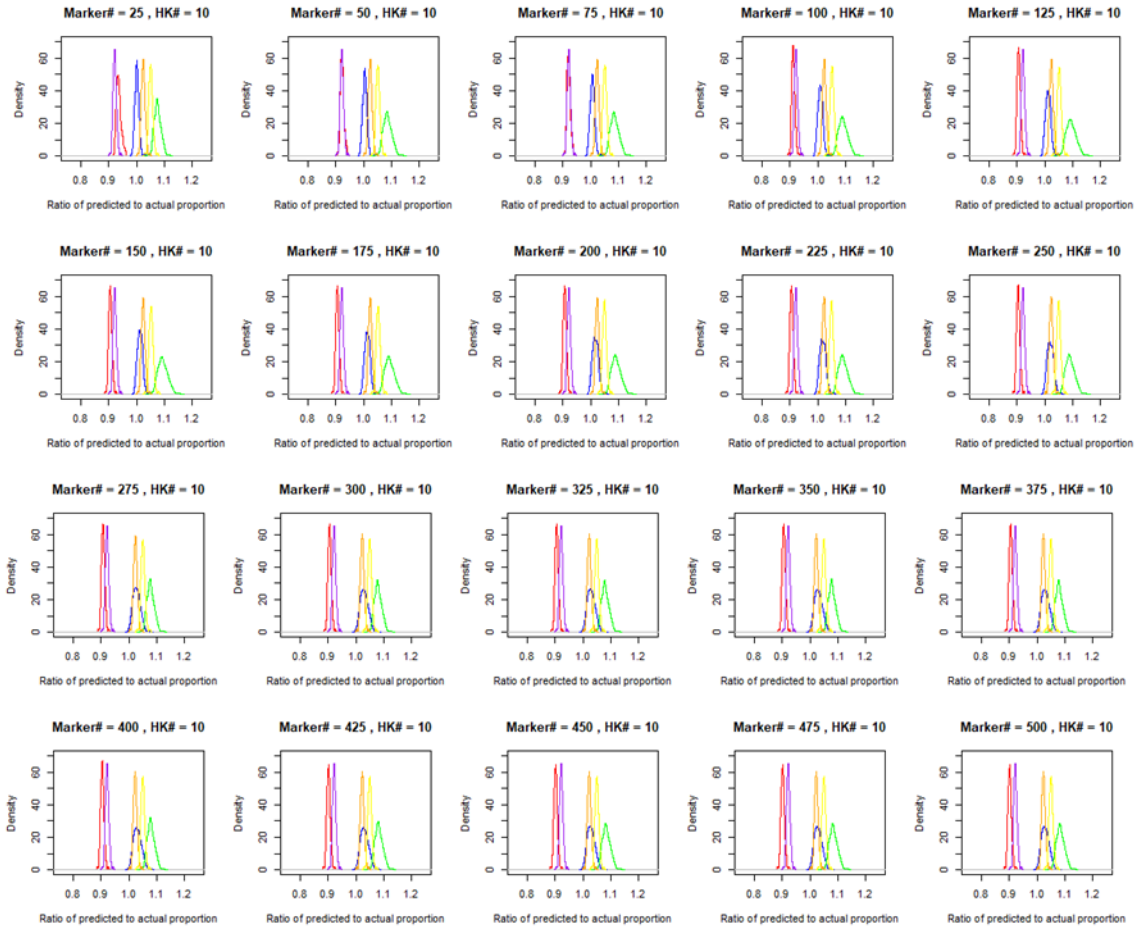


Figure 57. Set of ratio of predicted to actual proportions for 100 pseudosamples against density of estimates, for the deconvolutions where HK#=10 and Marker gene # varies from 25 to 500. Optimum is 25 as shown in more detail in Figure 59. Each cell type is indicated by a different colour. Astrocytes=Red, Neurons=Blue, Oligodendrocyte Precursor Cells=Green, Myelinating Oligodendrocytes=Purple, Microglia=Orange, Endothelial Cells=Yellow. The ideal scenario would be a straight line at 1, indicating perfect invariant deconvolution.

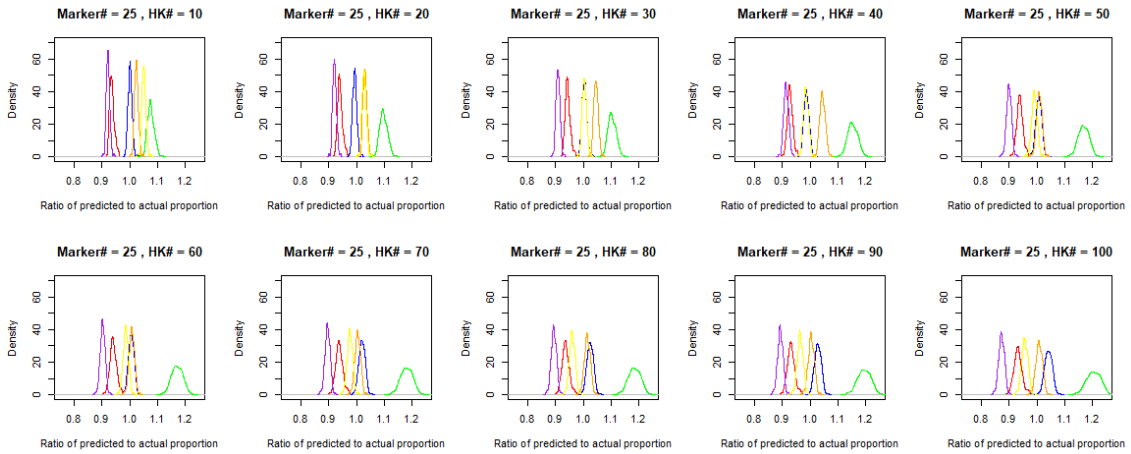


Figure 58. Set of ratio of predicted to actual proportions for 100 pseudosamples against density of estimates, for the deconvolutions where Marker gene #=25 and HK # varies from 10 to 100. Optimum is 40 as shown in more detail in Figure 59. Each cell type is indicated by a different colour. Astrocytes=Red, Neurons=Blue, Oligodendrocyte Precursor Cells=Green, Myelinating Oligodendrocytes=Purple, Microglia=Orange, Endothelial Cells=Yellow. The ideal scenario would be a straight line at 1, indicating perfect invariant deconvolution.

The results of the best deconvolution can be seen in Figure 59.

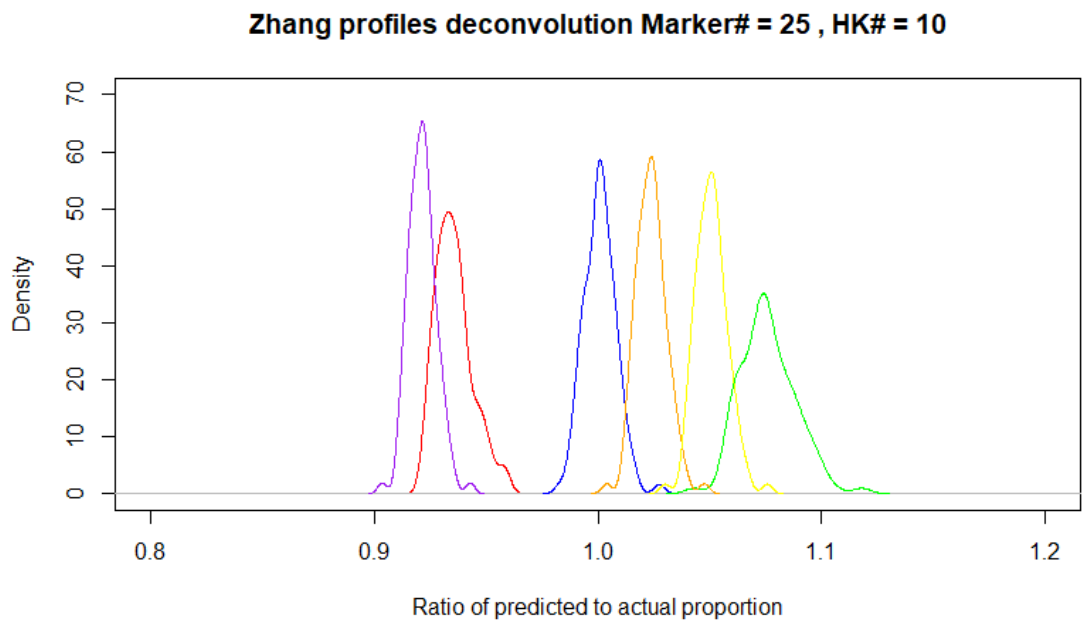


Figure 59. Set of ratio of predicted to actual proportions for 100 pseudosamples against density of estimates, for the deconvolution in which marker=25, HK=10. Each cell type is indicated by a different colour. Astrocytes=Red, Neurons=Blue, Oligodendrocyte Precursor Cells=Green, Myelinating Oligodendrocytes=Purple, Microglia=Orange, Endothelial Cells=Yellow. The ideal scenario would be a straight line at 1, indicating perfect invariant deconvolution.

It can be seen that the difference between the maximum and minimum estimate for each cell type is relatively close. With the exception of oligodendrocyte precursor cells, the difference is less than 0.05, while the actual estimates themselves vary with averages of 0.93 (astrocytes), 1 (neurons), 1.07 (oligodendrocyte precursor cells), 0.92 (myelinating oligodendrocytes), 1.02 (microglia) and 1.05 (endothelial cells).

5.5.2 Deconvolution of comparison datasets

5.5.2.1 Zhang Two

I carried out a series of deconvolutions of the Zhang Two dataset using all possible settings and examined the predictions of each cell types, as in the cortical analysis. Since many of the housekeeping genes are the same, I expected the results would be similar. With the exception of poorer astrocyte predictions if there were 300 or more marker genes, all settings were highly predictive. The results were near identical to the cortical deconvolution. All trends were the same and the conclusions reached in the corresponding cortical section apply here too.

Using the optimum settings of marker=25 and HK=10, the deconvolution of the "Zhang Two" dataset is shown in Table 27.

Sample cell type	Predicted proportion of matching cell type
FACS sorted astrocytes	0.9995959
FACS sorted astrocytes	0.9997599
Immunopanned astrocytes 1 month	0.9261619
Immunopanned astrocytes 4 month	0.7781610
Immunopanned astrocytes 7 month	0.8459271
Immunopanned astrocytes 9 month	0.7388576
Neuron	0.9220760
Neuron	1.0000000
OPC	0.9897386
OPC	0.9650586
MO	1.0000000
MO	0.7890155
Microglia	0.9984338
Microglia	0.9997204
Endothelial Cell	0.9654983
Endothelial Cell	1.0000000

Table 27. Results of the deconvolution of the Zhang Two dataset using hippocampal optimum settings. OPC=Oligodendrocyte precursor cell, MO=Myelinating oligodendrocyte.

The results are highly similar to those of the cortical deconvolution, with 11 of 16 cells predicted as >95% of the cell type that they are. As before, astrocytes were increasingly poorly predicted with age.

Zhang Two's three cortical datasets were also deconvoluted using the optimal hippocampal settings and had highly similar profiles to those predicted by the optimal cortical deconvolution. Profile estimates averaged at 51% astrocytes (2% more from cortical), 11% neurons (1% less), 30% oligodendrocyte precursor cells (same), 1.6% myelinating oligodendrocytes (same), 2.8% microglia (1.4% more) and 5.7% endothelial cells (1.4% less). The average difference between the highest and lowest estimates for all cell types was 0.8% and the largest was 1.7%, for neurons. In general these results are very similar to the cortical data, although they are less variable. The issues of predicting whole cortex samples will likely apply here as well.

5.5.2.2 Dorsal Root Ganglion Neurons

As in the cortical deconvolution testing, I carried out deconvolution of the roughly 200 neurons described in the Li *et al.* 2016 paper[156] using the same spread of HK and marker gene numbers. Each RNA-Seq sample was normalised to total mapped reads, and then to the geometric mean of the housekeeping gene expression. Since some samples (usually 10-20%) did not have all the housekeeping genes expressed, some samples could not be included in the analysis.

Given that the marker genes are the same, I expected a similar result to that of the cortical settings deconvolution; housekeeping genes not having much impact, and a sudden jump in neuron identification at 300 markers. This is exactly what I observed in Figure 60 and Figure 61, looking at the average and standard deviation neuron proportion in the deconvolution of >150 DRG neurons using the hippocampal settings. One difference is that having a low number of housekeeping genes (red) appears to give better neuron prediction at the 100-275 marker gene settings. However, the results are essentially the same and the same conclusions can be drawn as in the analysis which looked at the cortical settings.

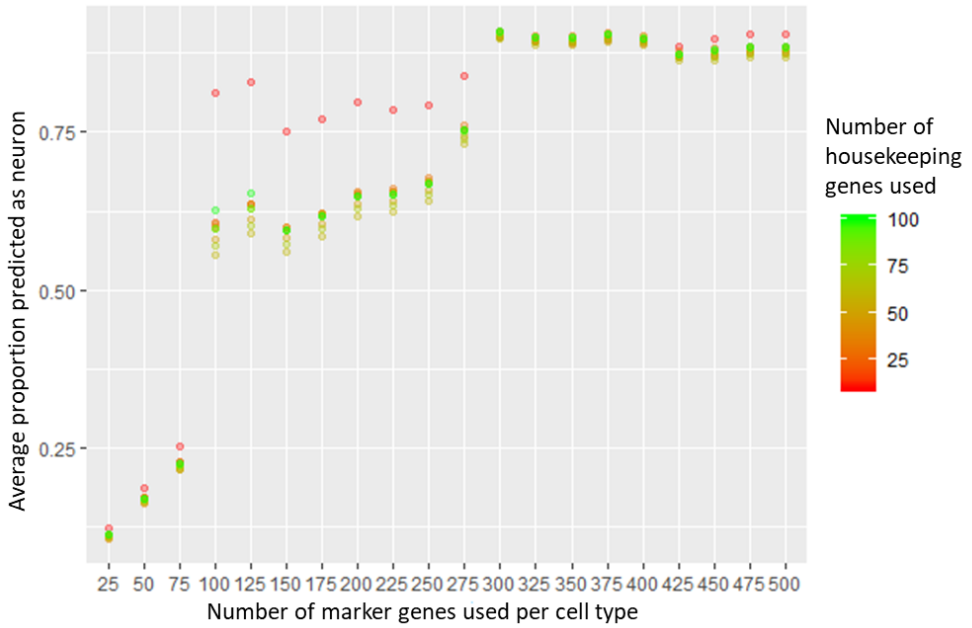


Figure 60. Graph displaying how neuronal prediction of >150 dorsal root ganglion RNA-Seq profiles using Zhang *et al.* cell types varies with housekeeping and marker gene number. The Y axis indicates the average proportion predicted to be neuron across >150 dorsal root ganglion RNA-Seq profiles; if identification is always perfect, it would be 1. The X-axis indicates the quantity of marker genes (per cell type) used in the deconvolution, while colour indicates the number of housekeeping genes utilised. Since six cell types were used, the total marker gene number varies from 150 to 3,000. Housekeeping genes chosen using Zhang and mouse hippocampal datasets.

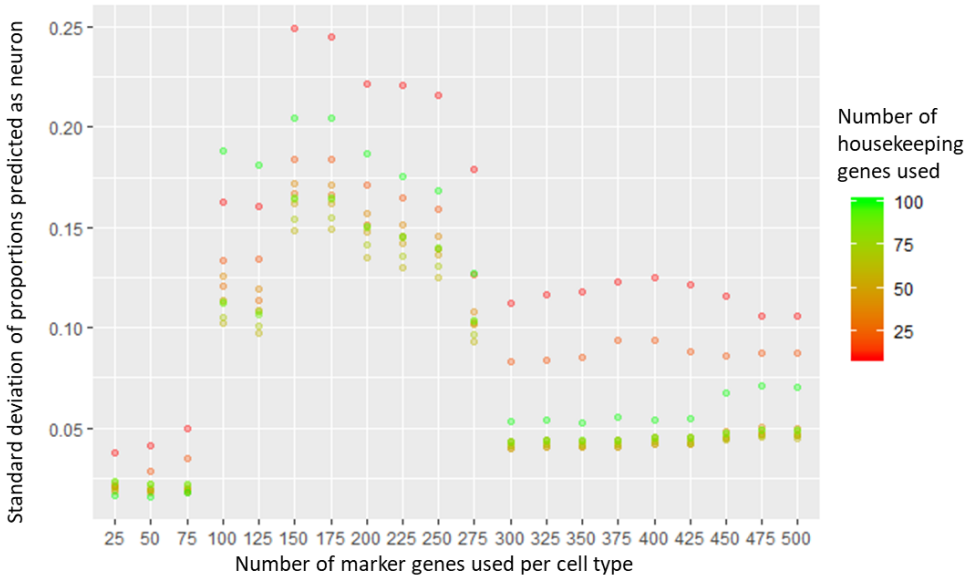


Figure 61. Graph displaying how neuronal prediction of >150 dorsal root ganglion RNA-Seq profiles using Zhang *et al.* cell types varies with housekeeping and marker gene number. The Y axis indicates the standard deviation in the proportion predicted to be neuron across >150 dorsal root ganglion RNA-Seq profiles; if identification is always perfect, it would be 1. The X-axis indicates the quantity of marker genes (per cell type) used in the deconvolution, while colour indicates the number of housekeeping genes utilised. Since six cell types were used, the total marker gene number varies from 150 to 3,000. Housekeeping genes chosen using Zhang and mouse hippocampal datasets.

The ramifications for the hippocampal deconvolution are similar to those of the cortical deconvolution. The 300 marker gene number appears to give optimal neuron deconvolution as in the cortical analysis. This is unsurprising given that these are the same 300 genes as in the cortical deconvolution. Looking at the deconvolution of the Zhang pseudosamples, the optimal setting always had 10 housekeeping genes regardless of marker gene number. There is no major issue here, although the standard deviation of the estimates with >300 markers is larger if 10 housekeeping genes were utilised. Taking HK=10, the minimum MAD was 0.049, at marker=25. At marker=300, the MAD is 0.059, a not tremendous increase in error. The comparison is displayed in Figure 62 where it is evident that there are only minor changes, mainly in neuronal prediction.

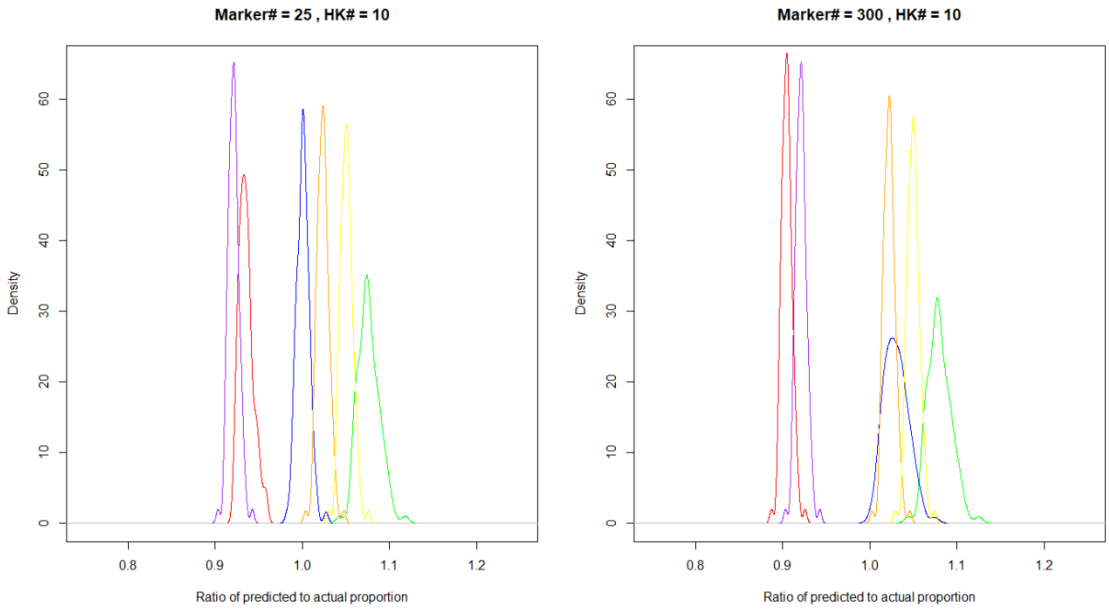


Figure 62. Detailed look at ratio of predicted to actual proportions for 100 pseudosamples against density of estimates for two deconvolutions. Labels above indicate housekeeping gene (HK) and marker gene numbers. HK#=10 and Marker gene#=25 is the overall optimum in terms of MAD. Each cell type is indicated by a different colour. Astrocytes=Red, Neurons=Blue, OligodendrocytePrecursorCells=Green, Myelinating Oligodendrocytes=Purple, Microglia=Orange, Endothelial Cells=Yellow

5.5.3 Deconvolution of mouse Der1 hippocampal samples

Given that all the conclusions from each set of deconvolutions are the same as in the cortical deconvolution (unsurprisingly as most housekeeping genes and all markers are the same), the same rationales apply. Correspondingly, there is a major jump in inaccuracy with the increase from 275 to 300 marker genes (from average of 4.3% to 13.6% across housekeeping gene settings). As before, higher markers give more neuronal proportion in the cortical samples, with 275 markers giving 45% neuron and the highest, 500 markers, giving 47% where HK=40, but at the cost of poor astrocyte prediction. The rationales are therefore the same as before in choosing 275 markers, with the best housekeeping gene set. Zhang pseudosamples have MADs of 5.8% if HK=10, with a mild increase of MAD with HK. Zhang Two's best deconvolution is if HK=40, with a deviation of 3.91% from perfect identification. The second best is if HK=10, with a deviation of 3.92%, a minor difference. Therefore the optimal settings are marker=275, HK=10, the same as in the cortical

deconvolution. This maximises pseudosample deconvolution accuracy and is almost identical in quality to the Zhang Two deconvolution.

I deconvoluted the mouse hippocampal samples using the same spread of marker and housekeeping genes as in the cortical deconvolutions, given the similarities. The results can be seen in Figure 63. An ANOVA for each cell type was performed to determine the effects of genotype on each of the cell proportions. There was no significant effect of genotype on the proportion of any cell type in any deconvolution.

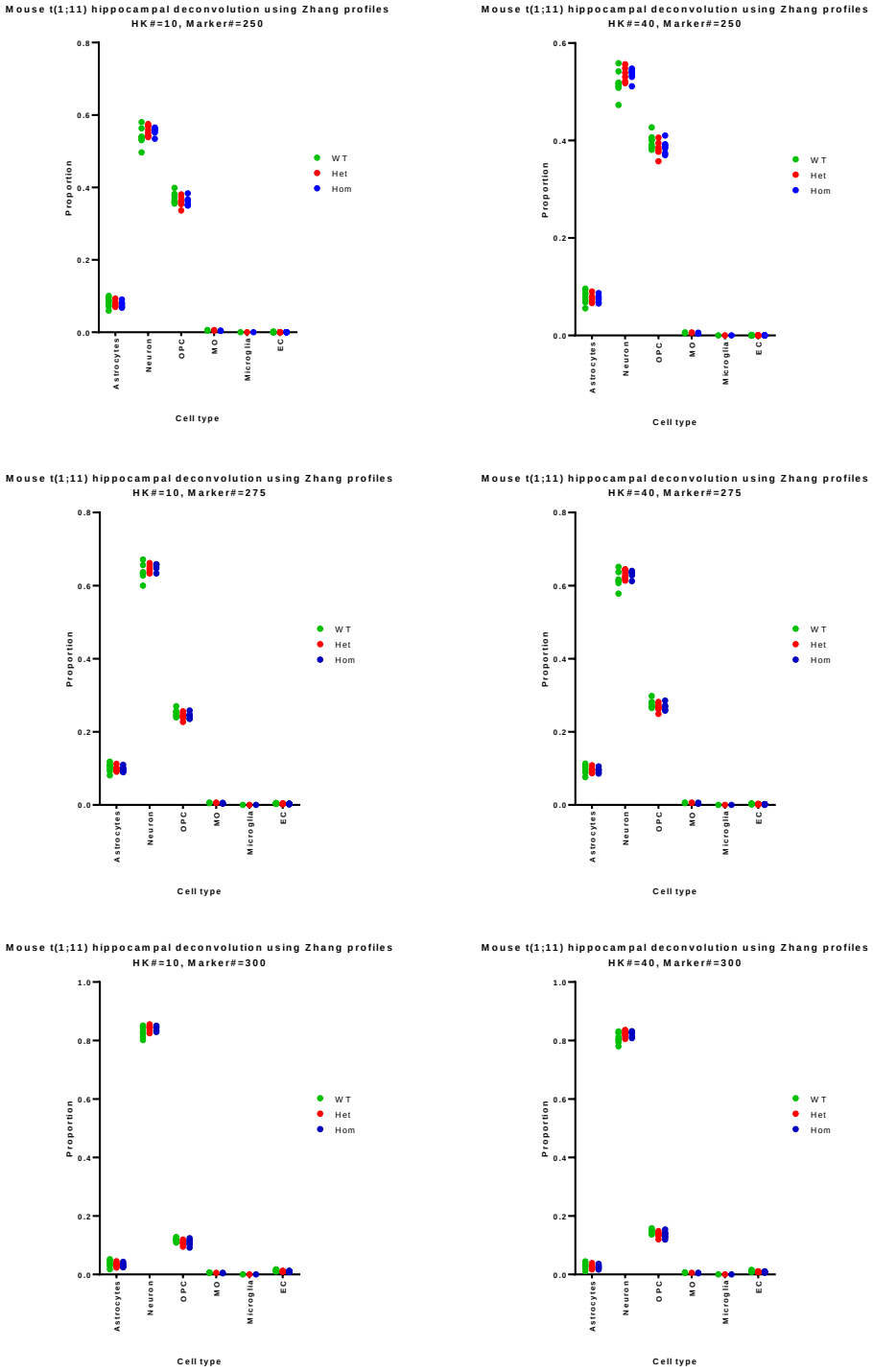


Figure 63. Deconvolution of mouse *Der1* hippocampal samples, showing cell types against proportions. WT=Wild-type, Het=Heterozygous, Hom=Homozygous, colours are green red and blue respectively. OPC=Oligodendrocyte precursor cell, MO=Myelinating oligodendrocyte, EC=Endothelial Cell

5.6 Human iPSC-derived neuron deconvolution

I moved on to deconvoluting the human t(1;11) samples. Housekeeping gene normalisation of the datasets prior to deconvolution was carried out exactly as before. I excluded genes with a fourfold difference between the maximum and minimum samples values, took the geometric mean of the coefficient of variations for both datasets, and used the geometric mean of several genes as a normalisation quotient. In addition, if the deconvolution reference dataset was mouse, then only orthologous genes were utilised.

5.6.1 Selection of appropriate datasets for human deconvolution

5.6.1.1 Zhang *et al.*

I decided to use the Zhang *et al.* dataset as it has several advantages over other datasets I examined. The read depth is high and comparable to that of our human neurons, unlike the Darmanis *et al.* dataset which I next describe. There are relatively few cell types, but their identification is trustworthy and they have many marker genes, ensuring the deconvolution should be relatively accurate. However, the main disadvantage is that this dataset was derived using mouse cells.

5.6.1.2 Darmanis *et al.*

I did not find a high read depth dataset of human cortical cell types, although many datasets exist which employ single cell RNA-Seq of hundreds or even thousands of human cortical cells. One such dataset is described by Darmanis *et al.*[155]. In this paper, they describe their analysis of several hundred cortical cells obtained via surgical resection of adult human brain, as well as a complementary set of cells obtained from foetal cortex. The adult humans were undergoing surgery for mesial temporal sclerosis and associated intractable seizures. RNA-Seq data was generated from 466 cells with a minimum sequencing depth of $4x10^5$ reads, with an average of $2.83x10^6$. Reads were 75bp and paired-end.

Darmanis *et al.* performed biased and unbiased clustering analyses, which were in broad agreement about cell cluster identities. They found that unbiased clustering gave 10 groups. The biased clustering approach used the top 50 cell markers from

Zhang *et al.* to separate the cells into a total of 7 groups, which matched to 8 of the unbiased groups, with the other two appearing to consist of a hybrid of cell types. The cell groups are distinguished by expression of many marker genes and both of their analyses show very broad agreement on which group a cell clusters with. Subsequent sub-clustering of the neuronal group (defined as cells clustered as neurons by both analyses) gave rise to two excitatory and five interneuron groups[155]. Regrettably neither the paper nor its supplementary information is sufficient to identify which subgroup the neuronal cells belong to, so this paper will not be useful for identifying neuronal subtypes.

To utilise the dataset, I obtained the RNA-Seq counts for all 466 cells described in the Darmanis *et al.* paper. I then removed those which were classified as "foetal quiescent" or "foetal replicating", leaving 331 cells in classes astrocyte, neuron, oligodendrocyte, oligodendrocyte precursor cell (OPC), hybrid, microglia, endothelial. I then normalised to read depth for each single cell RNA-Seq sample. I then generated pseudosamples using the Darmanis *et al.* dataset in the same manner as with the Zhang *et al.* dataset.

The dataset has one key advantage over the Zhang *et al.* dataset, in that it is generated from human cells. The cells are resected from living brain tissue rather than grown in cell culture like our own, and in contrast to the Zhang *et al.* cells which have spent some time in cell culture. However, the sequencing depth is far lower than our dataset. With this in mind, I decided to utilise both the datasets described by Zhang *et al.* and Darmanis *et al.*.

5.6.1.3 Allen *et al.* datasets

The Allen Brain Atlas datasets are described in detail in their white paper as well as at the web address http://celltypes.brain-map.org/rnaseq (accessed on 16/10/2018). The human dataset I utilised is comprised of single nucleus RNA-Seq of the middle temporal gyrus. Single nucleus RNA-Seq is somewhat less than ideal as many transcripts in human neurons are locally translated at dendrites and other non-nuclear locations. These transcripts number as high as 2,550, although of course many of these may also be partially translated in the nucleus and will appear in the datasets[277].

It is clear that single nucleus sequencing, as opposed to single cell sequencing, will not capture all the information available. There are 15,928 nuclei derived from 8 post-mortem human adult brains. The average sequencing depth is $2.63x10^6$ reads, comparable to Darmanis *et al.* but only about a tenth of the depth of our samples. Accordingly, far fewer genes were detected, ranging from 6,186 to 9,937, depending on the cell subclass (GABAergic, glutamatergic, unassigned, non-neuronal).

5.6.2 Zhang et al. deconvolution

I first utilised the Zhang *et al.* dataset, optimising the deconvolution by deconvoluting Zhang pseudosamples, the Zhang Two dataset, and the Dorsal Root Ganglion neurons. The t(1;11) samples were then deconvoluted using the settings these datasets suggested were optimal.

For the Zhang *et al.* deconvolution, pseudosamples were deconvoluted using a variety of marker gene numbers (25 to 500 in increments of 25) per cell line, and HK values ranging from 1-10, and then 15 to 50 in increments of 5. It should be noted that these pseudosamples and marker genes are identical to those used in the mouse cortical deconvolution; the difference is that different housekeeping genes are being utilised, and that the marker genes have been filtered beforehand so that only genes orthologous between human and mouse have been retained. I chose to filter for orthology before selecting the genes; therefore the deconvolution with 50 genes utilises the top 50 orthologues for each sample rather than the top 50 overall. I chose to do this as I had observed that having the same number of marker genes per cell line had given good deconvolution in the mouse deconvolutions.

MAD ranged from 0.07 (marker=100, HK=2) to 0.23 (marker=425, HK=50). Increasing marker numbers or HK numbers appeared to cause a gradual and gentle increase in MAD. The results of the two deconvolutions with the lowest and highest MADs are shown in Figure 64. We can see the clear contrast between the two results. The best deconvolution has narrow peaks for all cell types, five of which peak within 0.1 of the optimal value of 1. The worst has broad graphs indicating a wide variety of estimations, and none of these have peaked within 0.1 of the optimum of 1. It is

notable that the lowest MAD is lower than that in either of the mouse deconvolutions with the Zhang *et al.* dataset.

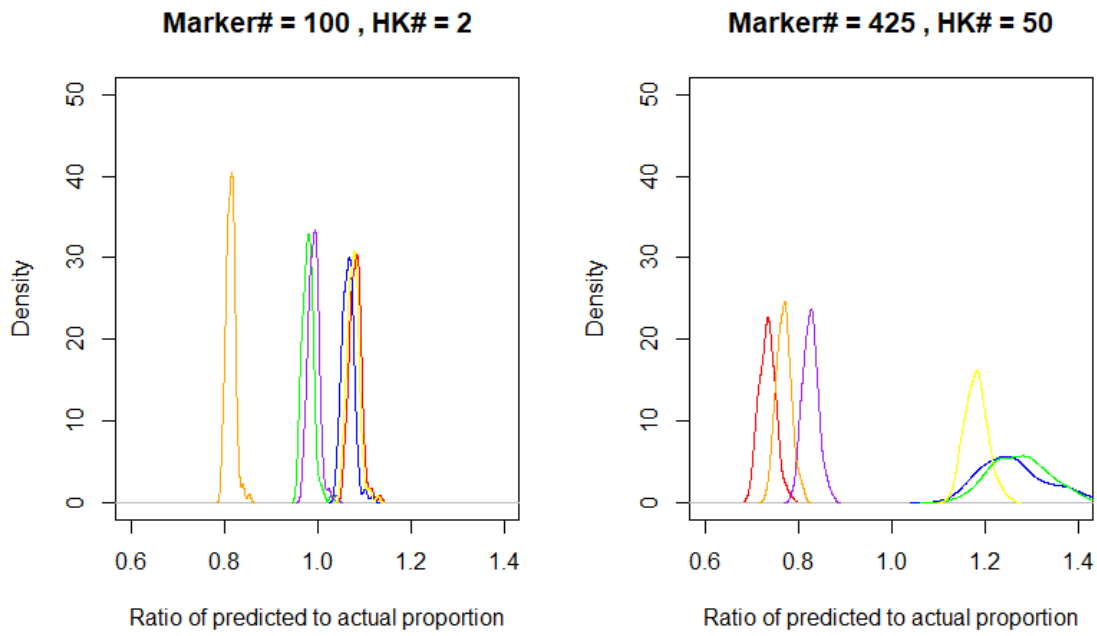


Figure 64. Set of ratio of predicted to actual proportions for 100 pseudosamples against density of estimates. On the left is the best deconvolution, while on the right is the worst. Each cell type is indicated by a different colour. Astrocytes=Red, Neurons=Blue, Oligodendrocyte Precursor Cells=Green, Myelinating Oligodendrocytes=Purple, Microglia=Orange, Endothelial Cells=Yellow. The ideal scenario would be a straight line at 1, indicating perfect invariant deconvolution.

5.6.2.1 Deconvolution of comparison samples

5.6.2.1.1 Zhang Two

I deconvoluted the Zhang Two dataset using all settings. When calculating MAD from this deconvolution, for 16/20 marker gene numbers the optimal MAD was seen if HK was 3 or less. The lowest MAD was 0.0413 at HK=1, marker=150, while the largest was 0.291 at HK=40, marker=425. MAD increased sharply at HK=25 or more and Marker Gene=250, where it roughly doubles across HK, or 350, where it roughly increases by 50%. Overall the results as marker gene varies are similar to the deconvolution of the Zhang pseudosamples, as expected. Examples are shown in Figure 65. The graphs are highly similar to the mouse cortical Zhang Two deconvolution. However, the results when marker=100, or 125, are not very similar and show a different prediction of the mouse whole cortex samples.

Deconvolution of the RNA-Seq data using Zhang et al. Cell type enriched datasets

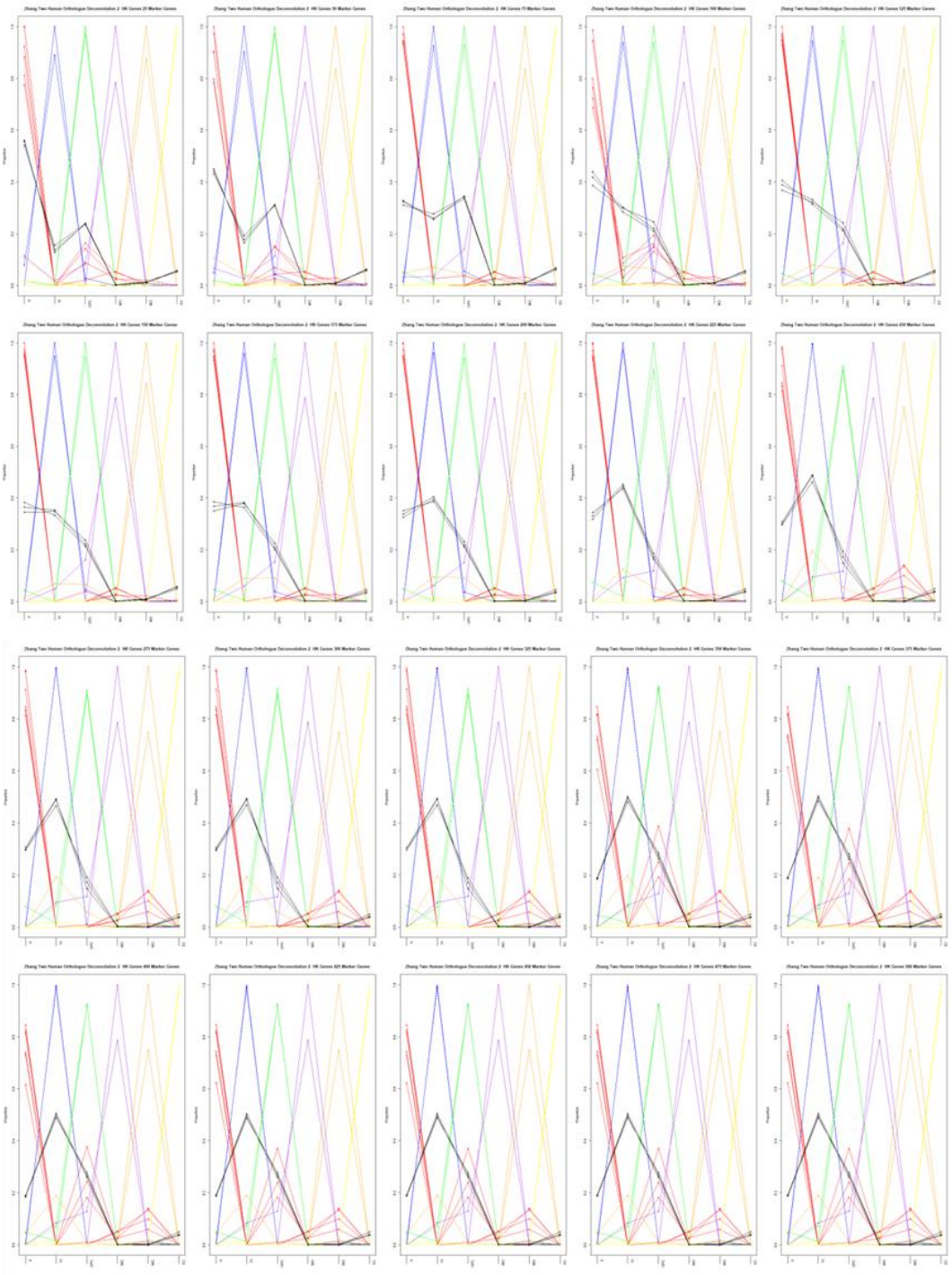


Figure 65. Set of predictions of Zhang Two dataset for the deconvolutions where HK#=2 and Marker gene # varies from 25 to 500. Optimum is 100. Each cell type is indicated by a different colour. Astrocytes=Red, Neurons=Blue, Oligodendrocyte Precursor Cells=Green, Myelinating Oligodendrocytes=Purple, Microglia=Orange, Endothelial Cells=Yellow, Whole Cortex=Black. Predicted cell types are on X axis and proportions are on the Y.

Using the optimum Zhang pseudosample settings of marker=100 and HK=2, I deconvoluted the Zhang Two dataset. The deconvolution of the "Zhang Two" dataset

is shown in Table 28. We can see that all cell types are quite well predicted; most are at >90% predicted as being what they are, with a mixed result for newly formed oligodendrocytes as being between oligodendrocyte precursor cells and myelinating oligodendrocytes.

Sample Cell Type	Predicted proportion of matching cell type
FACS sorted astrocytes	0.722672906
FACS sorted astrocytes	0.986248185
Immunopanned astrocytes 1 month	0.945890431
Immunopanned astrocytes 4 month	0.763129896
Immunopanned astrocytes 7 month	0.687663741
Immunopanned astrocytes 9 month	0.798494882
Neuron	0.937592812
Neuron	1
Oligodendrocyte Precursor Cell	0.936229031
Oligodendrocyte Precursor Cell	0.999879788
Newly Formed Oligodendrocyte*	0.546020484
Newly Formed Oligodendrocyte*	0.572069741
Myelinating Oligodendrocyte	1
Myelinating Oligodendrocyte	0.78304833
Microglia	0.834180859
Microglia	1
Endothelial Cell	0.986062832
Endothelial Cell	1

Table 28. Results of the deconvolution of the Zhang Two dataset using human optimum settings. *Newly Formed Oligodendrocytes were matched to Oligodendrocyte Precursor cells. They also were predicted as 0.205 and 0.241 proportions of Myelinating Oligodendrocyte. Each row represents a different Zhang Two sample.

A full discussion of what settings were chosen is in the corresponding t(1:11) deconvolution section.

5.6.2.1.2 Dorsal Root Ganglion Neurons

As in the mouse deconvolution testing, I carried out deconvolution of the roughly 200 neurons described in the Li *et al.* 2016 paper[278] using a spread of HK and marker gene numbers. Since some samples (usually 10-20%) did not have all the housekeeping genes expressed, some samples could not be included in the analysis. As with the deconvolutions of this dataset using the mouse cortical and hippocampal optimum settings, I have displayed the average estimations (Figure **66**) and standard

deviations (Figure 67) of the neuron content of the >150 neurons using all combinations of HK and marker gene numbers.

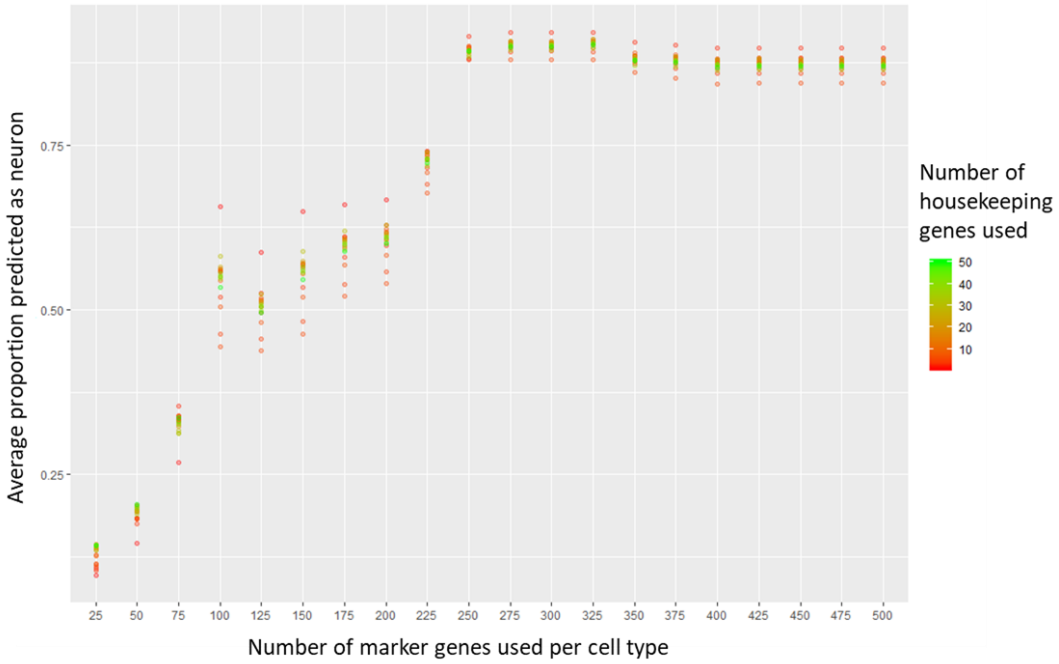


Figure 66**. Graph displaying how neuronal prediction of >150 dorsal root ganglion RNA-Seq profiles using Zhang *et al.* cell types varies with housekeeping and marker gene number. The Y axis indicates the average proportion predicted to be neuron across >150 dorsal root ganglion RNA-Seq profiles; if identification is always perfect, it would be 1. The X-axis indicates the quantity of marker genes (per cell type) used in the deconvolution, while colour indicates the number of housekeeping genes utilised. Since six cell types were used, the total marker gene number varies from 150 to 3,000. Housekeeping genes chosen using Zhang and human datasets.**

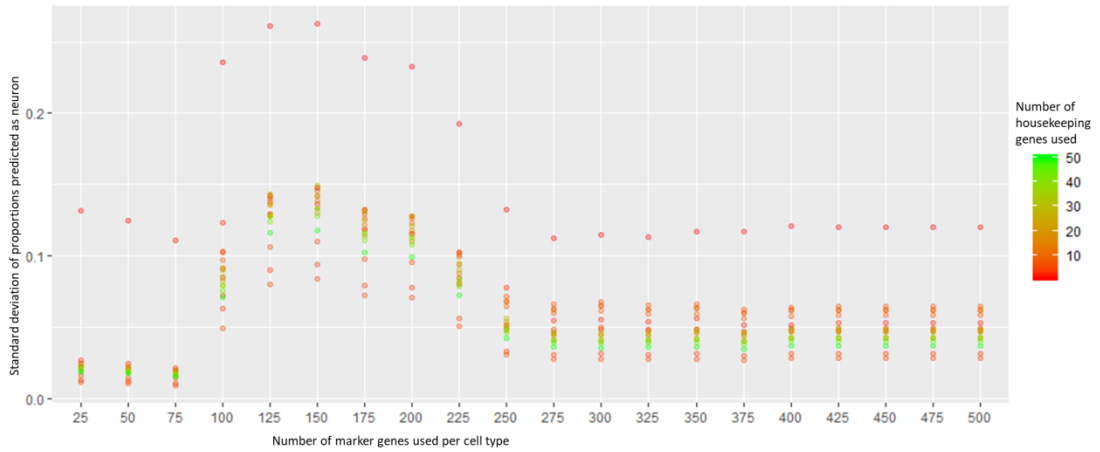


Figure 67. Graph displaying how neuronal prediction of >150 dorsal root ganglion RNA-Seq profiles using Zhang *et al.* cell types varies with housekeeping and marker gene number. The Y axis indicates the standard deviation in the proportion predicted to be neuron across >150 dorsal root ganglion RNA-Seq profiles; if identification is always perfect, it would be 1. The X-axis indicates the quantity of marker genes (per cell type) used in the deconvolution, while colour indicates the number of housekeeping genes utilised. Since six cell types were used, the total marker gene number varies from 150 to 3,000. Housekeeping genes chosen using Zhang and human datasets.

The maximum predicted proportions start at 250 marker genes and from then on there is no increase in predicted proportion with increased marker gene numbers. There is a similar pattern with standard deviation of estimates; the variation increases from 100-150 and then decreases until 250, when it steadies. We also see that one deconvolution has a particularly high standard deviation across all marker gene options; this corresponds to the deconvolution with 1 housekeeping gene. The pattern is similar to the deconvolution using the mouse cortical and hippocampal housekeeping sets, although the change is at 250 markers instead of 300.

Although the optimal for pseudosample deconvolution was marker=100, HK=2, it appears that these settings do not give the lowest standard deviation in neuron prediction rates. I examined the MAD of the deconvolutions of the pseudosamples in more detail, as initially described in 5.6.2.1. The optimal MAD was 0.07376, but at marker=250, HK=2, it increased merely to 0.0768, a 4% increase in what was already a very small MAD. A comparison of the two is shown in Figure 68.

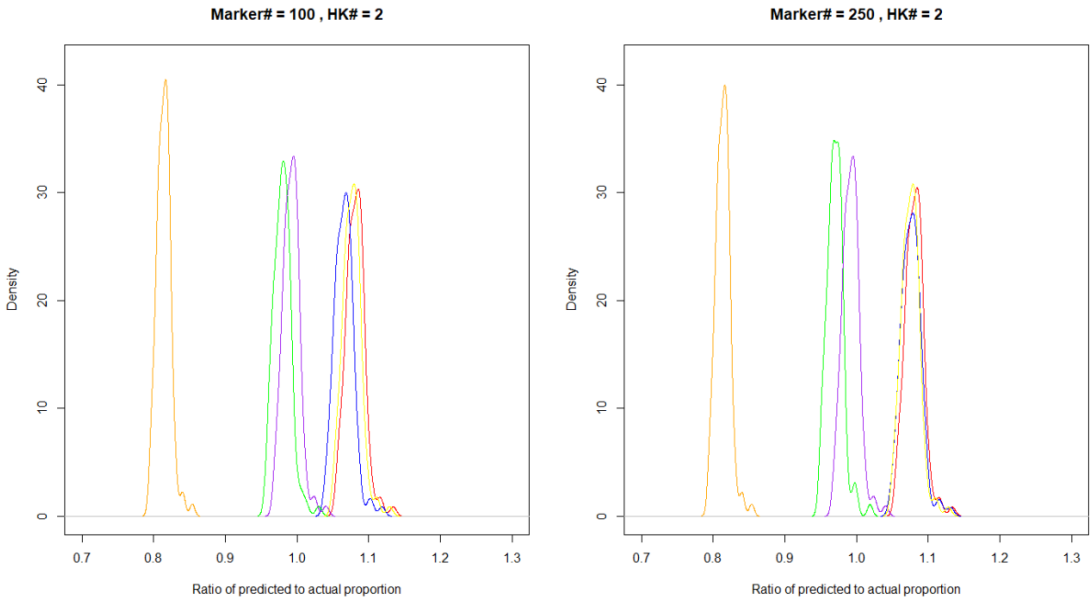


Figure 68. Set of ratio of predicted to actual proportions for 100 pseudosamples against density of estimates. On the left is the best deconvolution, while on the right is that with the same HK but more markers so as to minimise standard deviation. Each cell type is indicated by a different colour. Red=Astrocytes, Blue=Neurons, OligodendrocytePrecursorCells=Green, Myelinating Oligodendrocytes=Purple, Microglia=Orange, Endothelial Cells=Yellow. The ideal scenario would be a straight line at 1, indicating perfect invariant deconvolution.

5.6.2.1.3 Deconvolution of human t(1;11) samples using Zhang et al. datasets

As with the mouse *Der1* samples, it is important to use the settings that give accurate deconvolution of whole cortical samples and have high accuracy in Zhang pseudosample deconvolution and Zhang Two identification. In all cases, the optimal settings were suggested by low HKs, but not HK=1.

The Dorsal root ganglion proportion estimates suggest that 250 marker genes would be ideal to avoid underestimation of neurons. However, it is at this change, from 225 to 250 marker genes, where inaccuracy most sharply increases across deconvolution of the Zhang Two dataset. This is particularly at lower HKs and is mainly driven by poor astrocyte prediction. Marker gene changes have a far less dramatic effect on the Zhang pseudosamples, especially at HK 1-10 where the difference between the optimal and least optimal marker gene number for each HK changes MAD by ~1%.

The mouse whole cortex samples are deconvoluted differently across changing marker gene numbers, as expected. As numbers over 350 show exceptionally poor

astrocyte prediction, these can be ruled out. Astrocyte prediction is also poor over 250. Numbers up to 175 show very low neuron prediction in the whole cortical samples, which should be predicted as at least 40%, with isotropic fractionation estimates being much higher. This leaves numbers 200-250 as being plausible choices, with 200 having larger sensory neuron estimate deviation than 225 or 250. Both 225 and 250 have similar estimates for whole cortex; 30/33% astrocyte, 45/48% neuron, 17%/17% oligodendrocyte precursor cell, and minor amounts of the other cell types. The increase in MAD is drastic at the 250 marker for Zhang Two, but not for the Zhang dataset. I chose to utilise the three marker gene sets 200, 225, and 250, with HK=2 which is the optimum in each case.

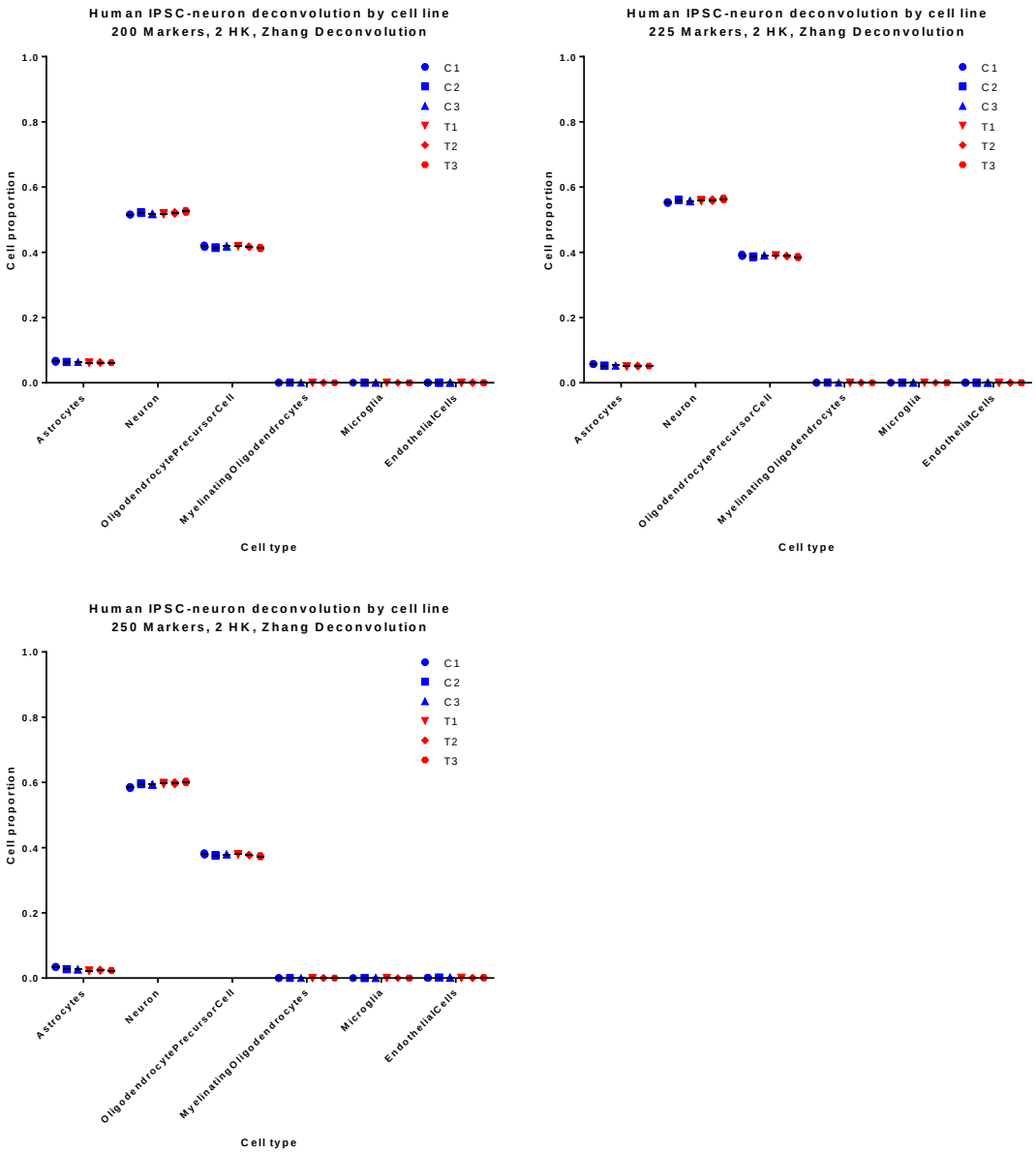


Figure 69. Deconvolution of human t(1;11) samples using Zhang *et al.* datasets, with data for separate cell lines shown. Controls are in blue and translocations in red, with different shapes for each cell line. The mean for each cell line is indicated by a black line and the three samples for each of the six lines are displayed..

The deconvolution predicts a mixed neuronal-immature oligodendrocyte culture with a minority of astrocytes. In addition, t-tests detected a significant change in astrocyte levels between the C and T genotypes for all deconvolutions, and a change in neurons in the HK=2, Marker Gene=250 genotype. The astrocytic change was a decrease in 5%/6.5%/20% with p values of $1.1x10^{-4}$/$4.5x10^{-4}$/$3.6x10^{-4}$, depending on the deconvolution. In all cases Sidak-Bonferroni correction for multiple testing was applied.

5.6.3 Darmanis et al. deconvolution

Next, I deconvoluted the t(1;11) samples using the Darmanis datasets. Darmanis pseudosamples, and the comparison Allen dataset were deconvoluted, then the t(1;11) samples.

As before, housekeeping genes were selected based on less than a fourfold difference between the maximum and minimum values of a sample, then the coefficients of variation in both our t(1;11) samples and the pure cell profiles of Darmanis *et al.* were used to rank genes by minimal variation. Only 18 genes met the fourfold variation criteria in both groups and were expressed in all samples and profiles. I therefore varied the number of housekeeping genes from 1 to 18 and the marker gene numbers from 25 to 500, in increments of 25.

The minimum MAD was found for the HK=16, Marker Gene number=225 deconvolution at 0.069, while the largest was found at HK=1, Marker Gene number=500 at 0.209. The optimal MAD is quite good, comparable to the Zhang *et al.* deconvolutions of the mouse samples (0.059 and 0.049 for cortical and hippocampal settings).

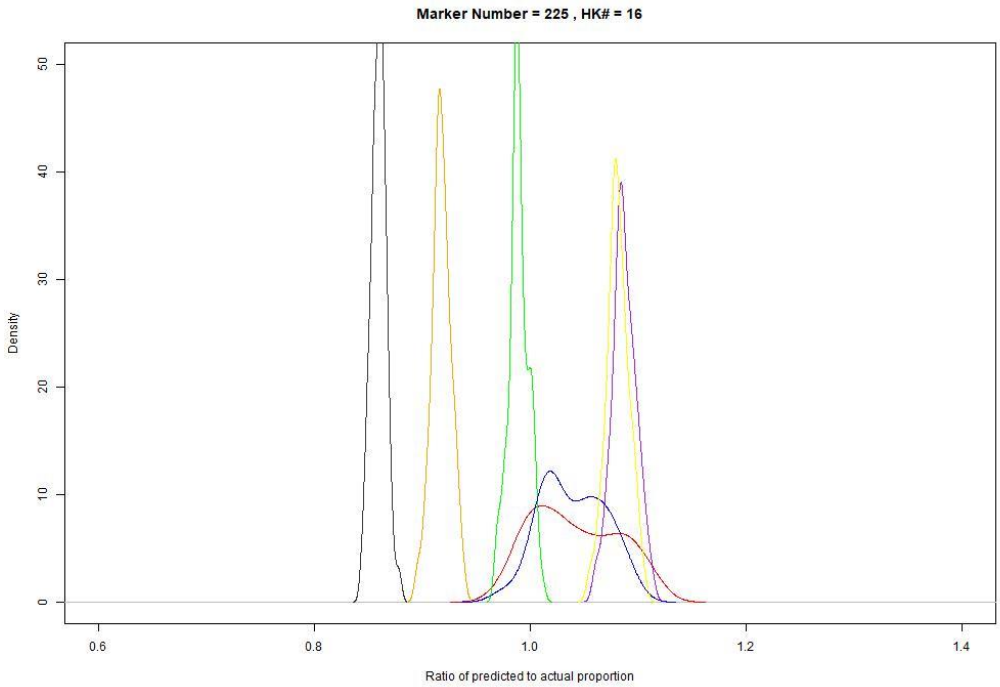


Figure 70. Set of ratio of predicted to actual proportions for 100 pseudosamples against density of estimates for the best deconvolution. Each cell type is indicated by a different colour. Astrocytes=red, Neurons=Blue, Oligodendrocyte Precursor Cells=Green, Myelinating Oligodendrocytes=Purple, Microglia=Orange, Endothelial Cells=Yellow, Grey/Black=Hybrid. The ideal scenario would be a straight line at 1, indicating perfect invariant deconvolutions

5.6.3.1 Deconvolution of comparison samples

5.6.3.1.1 Allen dataset

I carried out deconvolution of Allen samples using the same variety of HK and marker gene numbers and the Darmanis dataset, so as to test the deconvolution. Due to constraints on computing power, I was unable to utilise the entirety of the Allen dataset. I therefore randomly selected 4,000 cells from the dataset and carried out the deconvolution on these. 2,571 were “Glutamatergic”, 1,083 “GABAergic”, 234 “Non-Neuronal”, while 112 were described as “No Class”.

The first observation I made was of an extremely high dropout rate of cells with increasing HK, due to lack of expression of one or more of these housekeeping genes. The “Non-Neuronal” cells were particularly badly affected; even at HK=2 only 23 of these remained. In general oligodendrocytes were predicted accurately with HK=2, Marker Gene=225 predicting them as being 67% oligodendrocyte, changing to 69% at Marker Gene=250. Oligodendrocyte precursor cells were

predicted as a hybrid cell type rather than as a oligodendrocyte; 62% and 55% at the above settings. Astrocytes were less well predicted, as were microglia. Regardless of setting, several observations could be made. These are primarily regarding the GABAergic and Glutamatergic cell types of the Allen dataset; as described earlier, most other cells quickly dropped out due to lack of housekeeping gene expression. At no setting were these cells predicted as having an average of >~5% of either the endothelial or microglia cell types. A minority did have a larger predicted cell proportion of endothelial cells at high marker and housekeeping gene numbers, up to about 20%. I observed that at higher marker gene numbers the cells were predicted consistently and mostly as neurons; this appeared to plateau at 250 markers for HK=16, as might be expected from the previous deconvolutions. Low marker gene numbers predicted neuron proportion poorly, again reaffirming that these were not ideal settings. At these lower HKs, where non-neuronal cells were still included, it could be seen these cells had increasingly predicted neuronal proportion at high marker gene numbers, >400.

To conclude, these deconvolutions reaffirm that high marker gene numbers are suboptimal for non-neuronal cell types, but neurons are relatively well predicted at lower marker gene numbers of 250. The jump from 225 to 250 marker genes appears to mark the last step in increasing neuron prediction. At HK=16, there were 20 GABAergic neurons and 139 glutamatergic. The 20 GABAergic neurons were predicted as 78% neuron and 10% hybrid at marker gene=225, while at 250 this changed to 89% and 5%. The 139 glutamatergic neurons also changed from 83% neuron, 11% astrocyte to 88% neuron 7% astrocyte. Clearly, there is an increase in accurate prediction from what was already a relatively high number. This trend is also seen at other HK numbers, e.g. increases of 64% and 70% for the two cell types to 74% and 79% at HK=2. AT HK=16, much lower marker gene numbers result in large predictions of hybrid cell type, clearly inaccurately. This deconvolution therefore reaffirms that the choice is between 225 and 250 marker genes, where the increase at HK=16 in the Darmanis pseudosample deconvolution brings MAD from 0.069 to 0.085. I therefore deconvoluted using both settings.

Figure 71 displays the results of the deconvolution using HK=16, Marker Gene=225 and 250. T tests for differences in cell proportion across genotype revealed no significant differences in any cell. However, we can see that there is evidently much more variation in these iPSC-derived neuron samples than there was in the mouse *Der1* samples, as might be expected in non-genetically homogenous samples that may differentiate to slightly different extents, as is typical of iPSC-derived neurons. This makes the outcome somewhat unreliable; it is possible that the samples are just too variable to accurately assess the proportions. The high similarity of the two deconvolutions highlights that the variability is likely a product of the cells rather than of the deconvolution.

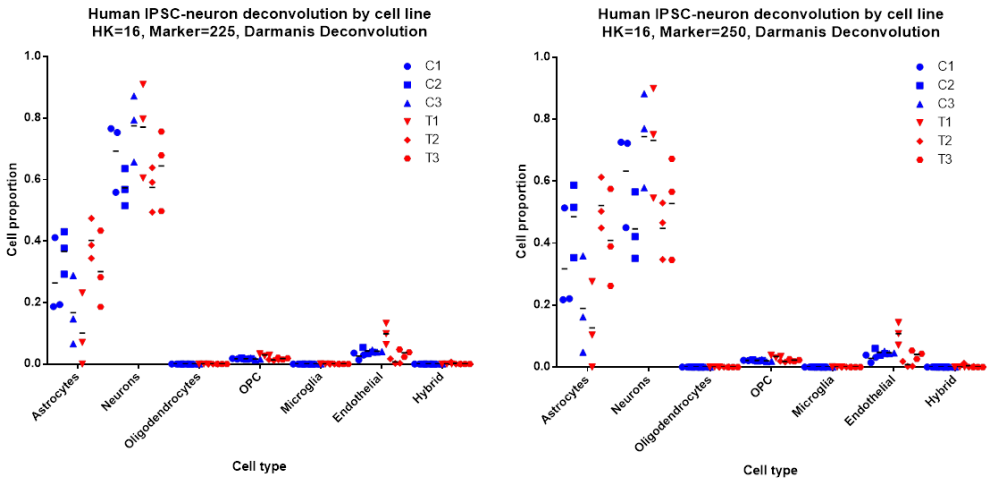


Figure 71. Deconvolution of human t(1;11) samples using Darmanis *et al.* datasets, with data for separate cell lines shown. Controls are in blue and translocations in red, with different shapes for each cell line. The mean for each cell line is indicated by a black line. Settings used were HK=16, Marker=225.

5.7 Summary of findings and discussion

There is little evidence from this investigation to suggest that the t(1;11) has an effect on cell proportions, and the same applies with the mouse hippocampus *Der1*. No pairwise effects were determined in the cortical analysis. Overall, the results suggest that the t(1;11)/*Der1* do not exert their effects via altering cell proportions. However, caveats do apply. For a start, the accuracy and power of the deconvolution is difficult to quantify. It can certainly be said from the pseudosample analysis that pseudosamples of appropriate RNA-Seq depth were accurately deconvoluted with

MADs at or below 7% in all cases. We can also confidently state that actual RNA-Seq profiles of similar depth (the Zhang Two datasets) were accurately identified in cases of pure cell types, and plausible cell proportions were given in the case of whole cortical samples. Should changes in magnitude be lower than 7%, or if they are in cell types which make up less than 10% of the total cell mixture (as pseudosamples had >10% of all cell types due to intrinsic limitations), then the power to detect these would be very low. I therefore cannot make any comment on whether such changes exist. It must be stated that a better method of quantifying cells would be to count them via staining or perhaps FACS sorting. Nevertheless, this limited analysis indicates that large changes in cell proportions are unlikely to be caused by the t(1;11) or *Der1*.

Deconvolution of the RNA-Seq data using Zhang et al. Cell type enriched datasets

6 DECONVOLUTION OF THE RNA-SEQ DATA USING ZEISEL *ET AL.* SCRNA-SEQ DATASETS

6.1 Selection of a single cell RNA-Seq dataset to examine specific cell subclasses

6.1.1 Introduction

It is accepted that the brain is comprised of a large number of distinctive cell types, including subsets of various neuronal types. As discussed in the Introduction, some psychiatric illnesses may be caused by disturbances in only a particular subset of cells. A database described by Zeisel *et al.*[157] is available at http://linnarssonlab.org/ and contains single-cell RNA-Seq data for about 3,000 cells, taken from both the somatosensory cortex and hippocampus of wild-type mice. Given that the data are generated from cells from both the cortex and hippocampus, this dataset can be used to deconvolute both my cortical and hippocampal RNA-Seq samples. I therefore used the cortical cell profiles described by Zeisel *et al.* in the deconvolution of my cortical *Der1* and t(1;11) datasets, and likewise used the hippocampal cell profiles in the deconvolution of my hippocampal *Der1* dataset.

The cortical subset of the dataset described by Zeisel *et al.* consists of 1,515 single-cell RNA-Seq profiles. There are a total of 8 major classes (interneuron, astrocyte/ependymal, oligodendrocyte, 3 classes of pyramidal neuron, microglia, and pericytes/vascular smooth muscle cells together known as mural cells) which further divided into 41 subclasses, all distinguished by the clustering method used by Zeisel *et al.*. The hippocampal subset is similar in size and character, consisting of 1,390 cells of 6 major classes (as above, but only one class of pyramidal neuron) divided into 38 subclasses. Subclasses are typically distinguished from one another by the expression of combinations of transcription factors. I used these subclasses for my deconvolution. Every subclass was found in at least two mice, and the total number of cells for each subclass ranged from 2 (for Interneuron 3) to 337 (for Oligodendrocyte 6) within each dataset, although some cell subclasses such as Int13 were present in only one dataset.

The data utilised by Zeisel *et al.* are not in the standard form of counts. Rather, they used a form of transcript tagging where each sequenced read has a unique molecular identifier (UMI) attached to its 5’ end. This means that they eliminate some sources

of technical error, important when sequencing at a low depth. Given that my samples were normalised to total sequencing depth, I consider that the analogous normalisation for the data of Zeisel *et al.* is to divide by total number of molecules sequenced. The median number was 24,287 and the average was 27,154.62. Each cell was normalised in this manner before being averaged with other cells of the same subclass to give the "pure" class or subclass profile.

6.1.1.1 Caveats

There are two issues with using the dataset described by Zeisel *et al.*. The first is intrinsic to using a large number of cell types in deconvolution; the low accuracy of DeConRNASEQ if cell proportions are low. Gong *et al.* have reported reasonable accuracy in deconvolution with cell proportions over 5%[135]. We can clearly see that the majority of cell types will be under 5% if there are 41 cell types. Since all cell proportions must sum to 1, severe over or underestimations of low cell types will result in "knock on" effects and will cause inaccuracies even in cell types which have reasonable prevalence. For this reason the deconvolution of a large number of cell types is intrinsically inaccurate.

The second possible issue is the low sequencing depth of the samples used by Zeisel *et al.* . I was unable to find high depth datasets with the same number of cell types as Zeisel *et al.*, which is to be expected given that they sequenced over 3,000 cells to retrieve these cell types. As discussed in the previous chapter, differences in sequencing depth cannot be accounted for just by normalising to sequence depth, especially for genes which are not among the most highly expressed. The issue here is that the internal structure of the dataset will differ between two datasets of different sequencing depth. However, there is one advantage in that since all the cell profiles are low-depth, they should be affected equally. In addition, I am looking for differences between samples rather than absolute proportions of cells. Therefore, I believe this issue should be relatively minor but warranted a discussion. More importantly, the low sequencing depth of the samples used by Zeisel *et al.* may also lead to problems with marker gene identification, as a gene may be moderately expressed in many cell types, yet only appear in the ones it is more highly expressed in due to lack of sequencing depth. For example, in the data described by Zeisel *et al.*

Disc1 is only found expressed in the cortical cell type S1PyrL4, a pyramidal neuron found in layer 4 of the cortex. However, *Disc1* is known to be expressed in a wide variety of cortical cell types[268]. It is likely that only in S1PyrL4 is expression high enough to result in some *Disc1* transcripts being sequenced. *Disc1* therefore appears to be a marker gene, but we know that it is probably not specific to that cell type at the higher sequencing depth. Therefore its power to accurately predict that cell type's prevalence in the deconvoluted datasets may be low, unless of course it is genuinely a marker gene and there is substantial enrichment in S1PyrL4. My solution to this will be to use a wide variety of settings, as before, and attempt to maximise deconvolution accuracy, particularly by using markers settings with high enrichment.

6.1.1.2 Markers

I experimented with another method of marker gene selection which was of use with the deconvolution of the Zeisel dataset. I hypothesised that one source of error was the expression of marker genes in cell types other than the line that they were supposed to be a marker for. To illustrate this, see Figure 42. Using Oligodendrocyte precursor cells as an example, we can see that many markers have quite good expression in neurons, and a minority have reasonable expression in astrocytes and endothelial cells. I was aware from the data described by Zhang *et al.* that fold changes for markers between the marked cell type and other cell types typically vary greatly. There was not a paucity of good marker genes. Most of the top marker genes have an expression level in their marked line several hundredfold larger than the average expression level of the other cell types, and even the least differentially expressed marker gene (*Piga3*, the 500th myelinating oligodendrocyte marker) has an expression at least threefold above the average of the other cell types. I saw that the problem might be that a marker had a large fold change between its marked line and all other lines *on average*, but might have reasonable expression in one of those other lines. High prevalence of a marker would then indicate either moderate levels of the primary cell type, or high levels of the secondary one. The marker would no longer function as a unique delineator for a single cell type. A better marker would even be one with high expression in one cell type and moderate but invariant expression in all others-but this gene would not be a sensitive detector, only a reliable one. Using

highly enriched markers would likely solve this problem, taking highly enriched as meaning not only much higher than in the next cell type but very high compared to the summed expression in all other cell types.

There is a second issue with the Zeisel dataset and markers. Unlike the Zhang dataset, where each of the six cell classes has several hundred markers with moderate fold changes >3, and several dozen with much larger fold changes, the cell subclasses are not similarly positioned here. The 41 cortical subclasses identified by Zeisel *et al.* are different in nature. Some, like Int3, Int4, and Int5, have one entirely specific marker each, while Oligo6 has 241. Therefore, picking any arbitrary number of markers and using this for each cell line either includes inferior markers, or leaves a large number of perfectly good ones out. I wanted to use a metric for maker selection which could flexibly integrate large numbers of markers if they were of an acceptable quality but exclude poor markers for other cell lines which did not have a similar number. Using a known metric rather than just picking numbers also means that all the marker genes are held to a certain standard.

I calculated a factor for all genes that I refer to as the specificity index (SI). The SI for a marker gene is equivalent to the maximum expression (the expression in the marked line) divided by the total expression of all lines. I decided to use markers with a specific SI, rather than ranking and taking the same number of genes for each cell subclasses. By filtering for markers above a certain SI, I could be sure that all cell subclasses had a number of markers above a threshold of specificity, and that there would not be issues due to one subclass having a handful of excellent, near-entirely specific markers. I thought this appropriate for the Zeisel dataset, where there were a large number of cell subclasses which had different numbers of markers. SI is functionally equivalent to "enrichment" as calculated by many papers; an SI=0.75 is equivalent to a fold enrichment of 3, while an SI=0.8 is equivalent to 4, SI=0.875 equivalent to 7, and SI=0.9 equivalent to 9. Using the same SI means that all the markers are at a certain standard of enrichment, but different lines have different numbers of markers if they are a suitable standard. The validity of this approach is shown by the reasonable MADs I found produced by the method.

6.1.2 Selection of comparison datasets to verify deconvolution

6.1.2.1 Introduction

As in the previous chapter, it was important to have datasets that could verify the accuracy of my deconvolution. For this, they needed to be similar in composition to the Zeisel *et al.* dataset in terms of RNA-Seq depth, and they had to have identified the cells in a trustworthy manner.

6.1.2.2 Allen Brain Atlas

The Allen Brain Atlas consists of several types of data including a database of single cell RNA-Seq for regions of the mouse, human, and macaque brains. Currently, cell type calls for the human middle temporal gyrus, mouse primary visual cortex and mouse anterior lateral motor area are available, along with the RNA-Seq results themselves. Several other datasets are soon to be released with cell type calls, but as it stands the datasets provide a large number of comparison cells to test my deconvolution on. I decided to use the human middle temporal gyrus and mouse primary visual cortex datasets.

The Allen Brain Atlas datasets are described in detail in their white paper as well as at the web address http://celltypes.brain-map.org/rnaseq (accessed on 16/10/2018). Mouse cells were harvested from P51-P59 animals for the most part, and then subjected to FACS sorting with neurons identified by expression of NeuN. The mouse dataset consists of over 15,000 cells, of 117 neuronal types and 16 non-neuronal types. The human middle temporal gyrus dataset was produced from post-mortem samples and has over 15,000 cells, which have been FACS sorted into neurons (90%) and non-neurons by the presence of NeuN. RNA was then harvested, converted to cDNA, and sequenced. All datasets are described as using 50bp paired end reads and were sequenced on a HiSeq 2500 instrument, so the technical details are somewhat different from our dataset. There are also some issues in that the brain regions in the Allen datasets are not hippocampus and cortex. The mouse dataset is cortical, but there is not a hippocampal dataset to match. There are many cell types which are only found in certain brain regions, and it must be noted that the cell types used in deconvolution are defined as such in the paper they originate from. This

means that the datasets in Zeisel *et al.* and Allen *et al*. will differ on the number of interneurons, pyramidal, etc cell subtypes they identify, depending on how different their clustering methods are. It will not be the case that each Allen subclass will easily map to a Zeisel subclass. It therefore seems likely that the Zeisel *et al.* deconvolution will misidentify many of the Allen categories in terms of *subclass*, although it should not misidentify them *within* the class of cells that they belong to. For example, an Allen cell being identified as a mixture of Interneuron subclasses 2 and 3 may not be biologically worrying (as these cell types may indeed be highly similar), but being identified as 40% Astrocyte 1 and 60% Pyramidal Cell 3 is worrying. Having access to so many cell profiles will be helpful.

6.1.3 Housekeeping gene normalisation to compare multiple datasets

6.1.3.1 Introduction

As it had successfully improved deconvolution using the Zhang dataset, and the rationales were the same, I decided to also use housekeeping gene normalisation to compare datasets here.

6.1.3.2 Selection of housekeeping genes

The rationale for housekeeping gene (HK) selection is the same as in the Zhang *et al.* deconvolution. They should be as uniformly expressed as possible, as well as being expressed in all cell subclasses.

In order to gauge the suitability of genes as housekeeping genes, I calculated the coefficient of variation (standard deviation divided by mean, referred to as CV) in the following manner for the datasets from the Zeisel *et al.* paper; I calculated the CV of the averages of each cell subclass (AvCV). In this case each cell profile was represented by the average of the cells within that profile, so that the 380 cells within the CA1Pyr1 group, for example, were averaged to give the CA1Pyr1 profile. The same was repeated with all cell subclasses, so that I had approximately 40 profiles. The AvCV of a gene is the CV of the gene across these profiles. Next, I determined the mouseCV, which is the CV for the gene across the *Der1* mouse samples. Since both CVs are directly comparable, I then determined the geometric mean of the

AvCV and mouseCV and ranked the genes by this measure as putative housekeeping genes from lowest geometric mean of the CVs to highest. As in the Zhang *et al.* dataset, I also restricted housekeeping gene selection to those genes which had a no greater than fourfold difference between the maximum and minimum values in both the mouseCV and the AvCV. Genes which were not universally expressed were also discarded as potential housekeeping genes. I also used GOrilla to check the gene ontologies of my putative housekeeping genes.

For the Cortical vs Zeisel comparison, the top 10 ontologies overrepresented at the top of the list of ranked potential housekeeping genes can be seen in the Appendix. It appears that none of the ontologies for process, function, or component are unique to any cell type in particular but rather denote general cell maintenance, exactly as expected for housekeeping genes.

I generated pseudosamples of the averaged cell subclasses described by Zeisel *et al.* in the same manner as in the Zhang deconvolution. I selected various numbers of housekeeping genes to use for normalisation, and I also varied the number of marker genes I utilised by SI. This was used to determine the optimal HK/SI settings for pseudosample deconvolution, and subsequently for comparison dataset deconvolution.

6.1.4 Measures of error

MAD and RMSE were calculated as in the Zhang deconvolution and utilised to determine pseudosample deconvolution efficiency.

6.2 Mouse Cortical RNA-Seq Deconvolution

6.2.1 Initial deconvolution of pseudosamples

After generation of 100 pseudosamples using the 41 cell subclasses found in the cortical dataset, I normalised both pure cell subclasses and the pseudosamples to a number of different housekeeping genes. The normalisation factor was equivalent to the geometric mean of all the utilized housekeeping gene values, selecting genes by the least variable genes first. I carried out the deconvolution using 1-10 housekeeping

genes, then 15, 20, 25…50. I also utilised marker genes, selecting these by SI value. I utilised marker genes with the specificity values of 0.75-0.95, at intervals of 0.025, as well as a final comparison with genes of values >0.999. There are therefore 10 different marker gene specificity indexes (SIs), and 18 different numbers of housekeeping genes (HKs), so this deconvolution of 100 pseudosamples has been carried out 198 times.

Examination of the deconvolution results indicated that the lowest MAD value was 0.22 (SI=0.925, HK=45) and the largest was 0.33 (SI=0.75, HK=3), indicating relatively good deconvolution, with estimates usually being less than 30% wrong. This is better than expected given the difficulty with predicting rarer cell types. In general increasing the SI appeared to increase the accuracy of the deconvolution. 15 of 18 HK deconvolutions had their maximum accuracy if SI=0.925, a pattern also seen in the RMSE (14/18 deconvolutions most accurate at SI=0.925). 8 of 10 SI deconvolutions had their maximum accuracy if HK=45. Since RMSE values seemed to reflect the MAD patterns, I therefore just looked at MAD for subsequent deconvolutions. I looked at the SI=0.925, HK=45 deconvolution in more detail.

Across the 100 pseudosamples a source of particular error stood out. Each pseudosample is comprised of 41 cell subclass and has ratios of estimated:actual proportions for each cell subclass. Across these 100 pseudosamples and 41 subclasses, the average ratio per cell subclass is 1.09, the average minimum is 0.946, and the average maximum is 5.46. Given that the optimum is 1, this is a surprisingly accurate deconvolution, although it appears to be prone to overestimation. Looking in greater detail, I saw that the Interneuron 5 cell subclass was consistently overestimated. The average ratio across 100 pseudosamples was 3, meaning this cell subclass is usually overestimated by 300%. The maximum ratio was 26, and the minimum 1.5. In total, there are 101 out of 4100 cell estimation ratios which are greater than 2, 49 of which are of Interneuron 5. The issue was clearly with this cell subclass. If I could cause it to be deconvoluted correctly, I would have an accurate deconvolution of the cell subclasses, despite their low proportions. Why this particular celltype is prone to overestimation is difficult to assess. However it may be due to a poor number of markers for the cell. Int5 has only one marker, *Ltb*, which is

solely expressed in this cell type. It does not express any markers of other cell types. Int3 has the next lowest number of markers, 2, and also does not express any markers of other cell types. It is likely this combination of poor marker number and lack of expression of other genes which combine to make it extremely difficult to assess the proportions of Int5, as it intrinsically has a low amount of data points that can inform its possible prevalence. Int3 is not characterised by similar overestimation; it is estimated at being on average 0.93 of the true proportion across 100 pseudosamples.

6.2.2 Removal of Interneuron 5

My first potential solution was crude; I removed the Interneuron 5 line from the cell subclasses, re-generated new pseudosamples without Interneuron 5, and deconvoluted them using the same spread of HKs and SIs as before.

The results were relatively similar to those with Interneuron 5 retained. The lowest MAD was 0.28 (SI=0.999, HK=15) and the largest 0.41 (SI=0.725, HK=3). I looked at the best deconvolution in more detail.

Across these 100 pseudosamples and 40 subclasses, the average ratio per cell subclass is 1.16, the average minimum is 0.95, and the average maximum is 9.59. It is evident that this deconvolution is more error prone than the previous one, as reflected in the larger MAD. The major difference between this deconvolution and the previous is that the error is not concentrated in any particular cell subclass. The cell subclass with the largest number of estimations which are more than twice or less than half the true proportion is Oligo6. It has 15 out of 100 samples which meet either of these criteria. In the previous deconvolution, Int5 had 49 of 100 samples meeting either of these criteria.

I also looked at the deconvolution with SI=0.925 and HK=45, as this was optimal when Int5 was retained. The average ratio per cell subclass is 1.19, the average minimum is 0.92, and the average maximum is 11.8, brought up by a handful of extremely large overestimations.

6.2.3 Merging Interneuron 5, 6 ,7, 8 cell types

Since removing Interneuron 5 had clearly exerted an effect on the other deconvolutions, changing not only the optimum settings but also causing other cell subclasses to be incorrectly predicted, I looked for another option to minimise the error introduced by Int5. One option would be to merge the cell subclass with other, similar cell subclasses, and see if this amalgamated subclass would be accurately deconvoluted. Looking at Zeisel *et al.*, we can see that their clustering analysis places Interneuron 5 with another group of interneurons. These are distinguished by the expression of certain genes and Interneurons 5, 6, 7 and 8 form a clade. I therefore decided to average the profiles of all cells from the subclasses Interneuron 5, 6, 7 and 8, to make a new subclass Interneuron 5678. I then generated 100 pseudosamples using this subclass, and the other 37, and deconvoluted these pseudosamples using the sample range of SIs and HKs as before.

MAD values were broadly similar to those of the initial deconvolution. The smallest was 0.286 (SI=0.99, HK=10) and the largest 0.398. 7 of 10 SI values gave their lowest MAD if the HK was 10, while the best SI value across HKs was 0.95 (8/18 times) or 0.999 (10/18 times). The optimum SI and HK was similar to that of the previous section, where Interneuron 5 was just removed. I therefore looked at the SI=0.99, HK=10 deconvolution in more detail.

Across the 100 pseudosamples and 38 cell subclasses, the average ratio of estimated:actual proportion was 1.16, slightly worse than the unmerged interneuron deconvolution. The average minimum estimate was 0.95 and the average maximum 8.98 (brought up by one particularly egregious overestimation by a factor of 65). Although these figures appear to show that the deconvolution is inferior to the unmerged interneuron deconvolution, there are some advantages.

Firstly, the source of error is not concentrated in any particular cell subclass. 88 out of 3800 cell proportion estimations are greater than 2. This is almost exactly the same proportion with this degree of error as in the unmerged deconvolution (0.023 vs 0.024). However the error is less concentrated. 18 of these estimates are of the Oligo6 subclass, 11 are in the Interneuron 15 subclass, and the rest are dispersed

throughout the various subclasses, with 1 in the merged Interneuron 5678 subclass. We can conclude that the merged deconvolution is less accurate overall, but the unmerged deconvolution is extremely inaccurate in the prediction of Interneuron 5. The results are quite similar to those where Interneuron 5 was merely removed.

I concluded that the best approach would be to carry out deconvolution of my mouse cortical samples using all options at their respective optimum settings; the unmerged, removed Int5, and merged Int5 profiles of the cell subclasses. I thought it likely but not a certainty that there would be errors in the prediction of Interneuron 5, so the other deconvolutions are likely to be more reliable but less informative due to less detailed information on the interneurons.

6.2.4 Deconvolution of Allen comparison dataset

I first tested the deconvolution using the cortical subset of the Allen Brain Atlas dataset. I utilised the same range of HK values to normalise as well as the same range of SI values to select marker genes. Although there are 15,000 RNA-Seq single cell profiles, not all of these were retained during the deconvolution process. Since the housekeeping genes I utilised were selected by expression in the Zeisel and heterozygous *Der1* cortex datasets, they are not all expressed in every Allen subclass. I noted that non-neuronal Allen subclasses particularly tended to lack expression of some of the housekeeping genes, although many of these cells did express all housekeeping genes and were correspondingly retained for the deconvolution. Cells not expressing all the required housekeeping genes were discarded. A number of cells were also removed on the basis of intermediate subclass characterisation (not being part of the “Core” cluster for each subclass). Although it would be interesting to see how these intermediate cells were deconvoluted, the Allen dataset does not note what they are intermediate to, only what they were primarily identified as being. They are therefore not useful for deconvolution.

I thought it likely that there will be Allen cells which are predicted to be a mixture of several Zeisel subclasses. This could be due to a lack of one to one relationship between Allen categories and Zeisel subclasses. However, we should expect at least that Allen cells will not present as a mixture of *classes*, i.e., Astrocytic and

Interneuronal, Oligodendrocytic and Interneuronal, etc. To analyse the efficacy of the deconvolution, I took all estimations for all cell subclasses across the thousands of Allen cells, then summed the Zeisel subclass estimates for each Zeisel class ("Oligo", "Interneuron", etc). I then produced violin plots for all classes. This allowed me to observe the Allen identification against the spread of predicted proportions for each cell class.

For the several thousand cells which have been deconvoluted, there are several different classes and subclasses according to Allen *et al.*. The classes are "Glutamatergic", "GABAergic", "Non-Neuronal", "Endothelial", "No Class", and "Non-Neuronal". The Allen subclasses are typically named according to the expression of certain markers such as parvalbumin, or are given a designation if the cell is well characterised (e.g., "Astro" and "Oligo" for astrocyte and oligodendrocyte, both subclasses in the "Non-Neuronal" class). In the results below, cells are divided by Allen subclass, coloured by Allen class, and have the spread of predicted proportions for each Zeisel class graphed (with each Zeisel class proportion defined as the sum of Zeisel subclass proportions within that Zeisel class). The question is how well the optimum setting performs. If it shows exceptionally poor identification of each cell type compared to other settings, this is strong evidence that this line of inquiry should be abandoned.

6.2.4.1 Results

It is not possible to assign a single measure of accuracy, such as MAD, to each deconvolution. As stated before, the list of housekeeping genes is generated from the Zeisel and *Der1* datasets, and therefore these are not necessarily housekeeping genes across the Allen cells. The different conditions will in some cases mean they are not expressed. I found that non-neuronal cells particularly lacked expression of the housekeeping genes when higher numbers of housekeepers were required. Different deconvolutions might optimise identification for a subset of cells but predict the majority very poorly. I elected to examine all deconvolutions by eye and particularly note how well the optimum settings performed. An example result is given in Figure 72, which contains the violin plots of the predicted proportions of the Zeisel class "Interneuron", which is the sum of the Zeisel subclasses "Int1", "Int2"…."Int16".

This is for the SI=0.7, HK=7 deconvolution. Violin plots were used as this allows the display of a large amount of data simultaneously.

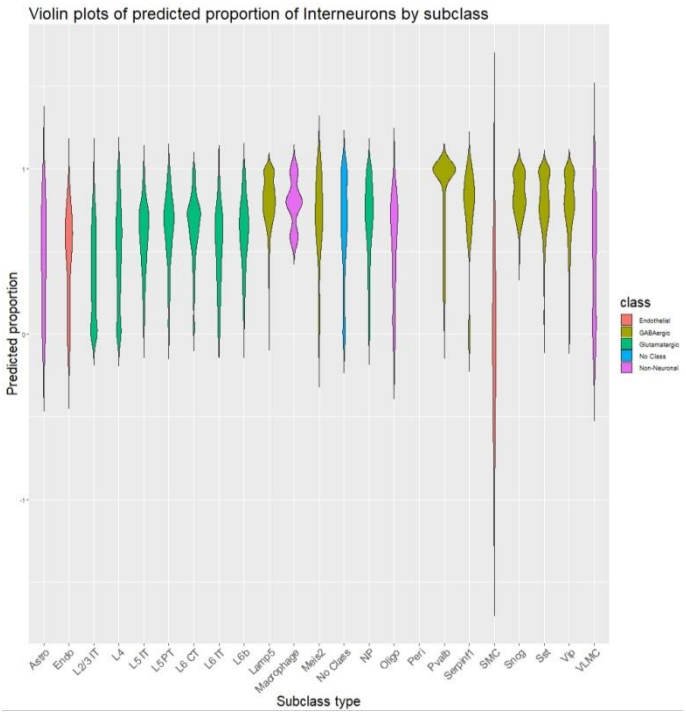


Figure 72. Violin plots of predicted Interneuron proportion for 23 subclasses of Allen cell, colour coded by class. Red=Endothelial, Yellow=GABAergic, Green=Glutamatergic, Blue=No class, Purple=Non-Neuronal. These results are of the SI=0.7, HK=7 deconvolution. Subclass types are as in the Allen dataset. The thickness of the violin plot at each predicted proportion represents its frequency according to the kernel density estimation of the deconvolution results for each subclass type.

We can see that there is a large spread of results regardless of Allen class type. Non-neuronal cells appear to be randomly distributed in their predicted "Interneuron" proportion, with the exception of macrophages which are mostly predicted as having a high level of interneurons. All GABAergic cells have a high probability of being predicted as mostly interneuron, while it appears glutamatergic cells have a broad spread. The perfect result would of course be that all cells except GABAergic ones had all results at 0 for the GABAergic cell proportion.

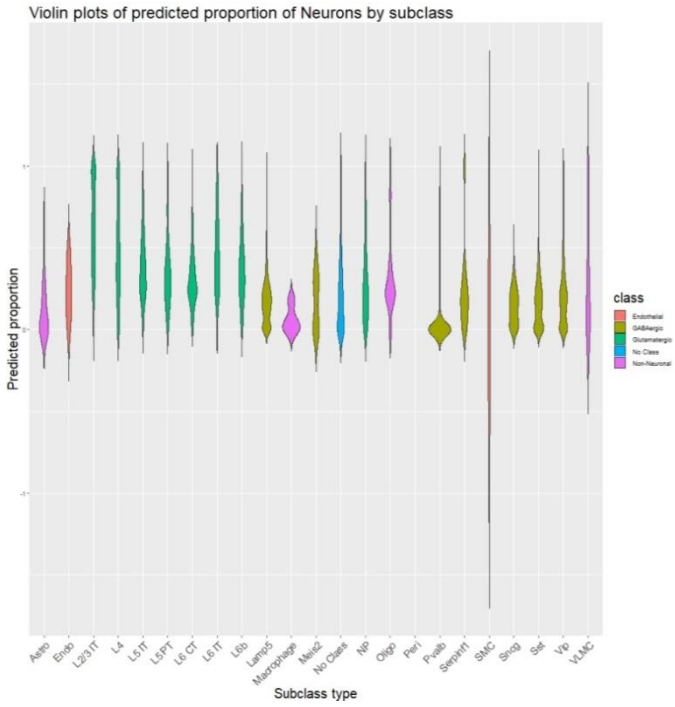


Figure 73. Violin plots of predicted Neuron proportion for 23 subclasses of Allen cell, colour coded by class. Red=Endothelial, Yellow=GABAergic, Green=Glutamatergic, Blue=No class, Purple=Non-Neuronal. These results are of the SI=0.7, HK=7 deconvolution. Subclass types are as in the Allen dataset.

Figure 73 is of the same deconvolution result but displays the predicted sums for the Zeisel class "Neuron". Here we see that most oligodendrocytes are predicted to be about 20% neuron, while most astrocytes are predicted as near 0%. The neuron classes have varying degrees of predicted neuron proportion but for most Allen subclasses (which correspond to cortical layers) cells are highly likely to have a predicted proportion below 50%. In general these deconvolution settings are not great but do display that when looking at any cell class, the cells which are of that class will have higher predicted proportions of it than cells of other classes do.

I noticed several trends across deconvolutions. The first was that the largest proportion of predicted cell type was typically Interneuron. Initially this was regardless of cell type, but with increasing SI the levels of predicted Interneuron in non-GABAergic cell type typically fell, as did the levels in Interneuron in GABAergic cell types (but to a lesser degree and from a higher initial proportion). I noted this at the HK values 20, 25, 30, 35, and 45. Low HK values typically gave erratic results with a large scatter of predicted proportions in each Zeisel class

regardless of actual cell identity. At SI<0.9, this large scatter was also generally seen in predictions except those of Neuron and Interneuron.

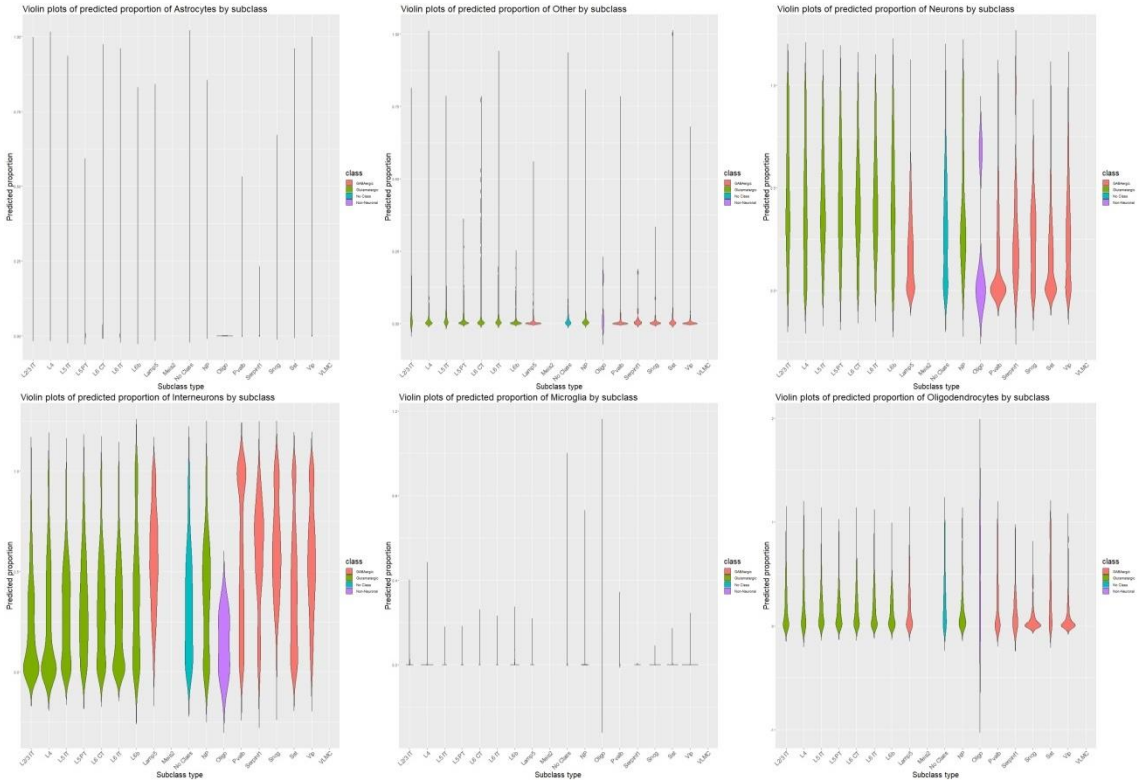


Figure 74. Violin plots of all 6 Zeisel class proportions for 18 subclasses of Allen cell, colour coded by Allen class. Red=GABAergic, Yellow=Glutamatergic, Blue=No class, Purple=Non-Neuronal. These results are of the SI=0.925, HK=45 deconvolution. Note different colours to Figure 72 and Figure 73

The SI=0.925, HK=45 deconvolution is shown in Figure 74. Note the decreased number of Allen subclasses compared to Figure 72 and Figure 73 due to the lesser number of cells and cell types which express all housekeeping genes. There are several observations to be made, in conjunction with the summarised non-graphic results as shown in Table 29. First, we can see that astrocytes and microglia are predicted as being essentially absent from the Allen cells. This is reasonable; there are neither astrocytes nor microglia within the cells which express all 45 housekeeping genes. Oligodendrocytes are most likely to be predicted as 0 in all glutamatergic and GABAergic cell types, but many of these cells are predicted as containing reasonable proportions of oligodendrocytes, or even as being 100% oligodendrocyte. Due to the low number of actual oligodendrocytes (four) the efficiency of prediction is hard to gauge; the average for these four is 42%, but for

two of them it is 0% and the other two are from the same subclass of oligodendrocytes (Oligo Synpyr) and have an average of 85%.

Predicted as being (on average):	**Number of cells**	**Astrocytes**	**Other**	**Neuron**	**Interneuron**	**Microglia**	**Oligodendrocyte**
GABAergic	3,709	3.5%	3.9%	18.8%	51.6%	0.49%	23.7%
Glutamatergic	5,340	4.8%	3.5%	42.9%	28.7%	0.5%	19.1%
Non Neuronal	5	~	3.5%	*	*	*	*

Table 29. Table of the average predicted proportions of each Zeisel class within each Allen class with SI=0.925, HK=45 the best deconvolution for the pseudosamples. Due to the low number of non-neuronal cells, these have been described verbally. The "No class" group has been removed as this has no information as to the quality of the deconvolution. *=1 oligodendrocyte predicted as 68% neuron,1 oligodendrocyte predicted as 30% interneuron, 1 cell identified as 100% microglia. 1 oligodendrocyte predicted as 84% microglia.,2 oligodendrocytes predicted as 85% oligodendrocyte.

To further understand the quality of the deconvolution of the Allen samples, I took a look at the deconvolution which produced the highest MAD in the original deconvolution of pseudosamples. This was HK=3, SI=0.75 and the results are summarised in Table 30.

Predicted as being (on average):	Number of cells	Astrocytes	Other	Neuron	Interneuron	Microglia	Oligodendrocyte
GABAergic	5,178	0.73%	3.4%	12%	82%	0.5%	4.8%
Glutamatergic	6,709	0.8%	2.6%	41.3%	58.1%	0.47%	1.2%
Non Neuronal	182	9.8%	8.1%	22.3%	51.4%	5.9%	2.2%
Endothelial	44	~	11%	21.9%	30.7%	1.2%	1.3%

Table 30. Table of the average predicted proportions of each Zeisel class within each Allen class with SI=0.75, HK=3. the worst deconvolution for the pseudosamples. The "No class" group has been removed as this has no information as to the quality of the deconvolution.

We can see that the average proportion of interneurons has been inflated across all cell types. Endothelial cells have also made an appearance as they now fit the criteria for inclusion by expression of all 3 of the housekeeping genes, as opposed to the 45 they would have needed in the previous deconvolution. None of the cells (except GABAergic cells which will contain interneurons) are on average predicted to be mostly the cell type that they are.

The overall accuracy is low, but there is at least a trend towards increasing accuracy with the better deconvolutions. The housekeeping gene selection in particular influences the deconvolution outcome, and given that the selection is based upon the variation within our mouse samples and the Zeisel *et al.* deconvolution profiles, the accuracy will be suboptimal for other datasets. It is extremely difficult to draw recommendations from the Allen dataset, and given that the deconvolution has not been accurate in identifying Allen cells it is difficult to trust its findings when applied to the mouse and t(1;11) samples, particularly given their different sample depth. I therefore suggest that the results found should be viewed as preliminary-although it is gratifying that the pseudosample deconvolution was relatively accurate. It should also be noted that housekeeping genes were selected based on validity across the samples and the Zeisel profiles; the deconvolution will be correspondingly more accurate. The methods used by the Allen institute to generate scRNA-Seq

should also be borne in mind. Neurons were distinguished by NeuN, which is a neuronal specific but not a neuronal universal marker[279,280]. However, if some neurons are being incorrectly identified as non-neuronal cell types, this doesn't explain the poor recognition of non-neuronal cells by my analysis. Cells were also sampled to a reasonable degree; at least 16 cells per type (133 types), and many cell types were specifically selected for using Cre recombinase mice. I therefore do not have doubts about the validity of the Allen dataset in this regard, and although the sequencing depth is low, so is that of the Zeisel dataset.

6.2.5 Deconvolution of mouse t(1:11) cortical samples

Deconvolution of the cortical samples was carried out in three analyses, once with marker genes from 41 pure cell subclasses, once with markers from 40 pure subclasses (lacking Int5), and once with marker genes from 38 subclasses, one of which represents the merged Interneuron subclasses 5, 6, 7, and 8. Since the HK genes were chosen for minimal variation in both Zeisel *et al.* datasets and those of the mouse cortical Der1 carriers, they were the same in both the pseudosample deconvolution and in the mouse cortical *Der1* deconvolution. Deconvolution was performed on 8 heterozygous, 8 homozygous, and 6 wild-type mouse cortical samples. For each analysis I used a variety of settings, as the Allen deconvolution had been quite poor. I utilised HK and SI numbers which were optimal for each of the three analyses in turn, but also applied them in every combination across all analyses, giving a total of 9 deconvolutions per analysis. As discussed previously, the homozygous samples appear to have an unusual interior structure; therefore ANOVA was carried out with both the homozygotes kept together, and with them split into two groups.

6.2.5.1 Unmerged deconvolution results

Although an issue with overestimation of Interneuron 5 was expected, this does not appear to have materialised at the HK=45 settings, which were optimal for this deconvolution. There is a high degree of variation in cell estimations for many cell subclasses, particularly interneurons. One-way ANOVAs were carried out for each cell subclass to examine the effects of genotype on proportion. It should be noted that

there are a total of 41x9=369 ANOVAs and results are therefore expected by chance. There were statistically significant differences between group means as determined by one-way ANOVA for several cell types in various deconvolutions. S1Pyrl5a was significant in both the SI=0.9, HK=10 and SI=0.925, HK=45 deconvolutions, and Vend2 appears three times, in every deconvolution where SI=0.925. However, Tukey's posthoc test for difference between groups did not in any case find a significant group difference for any comparison. In total, 7 cell findings were significant in the ANOVA across all 9 deconvolutions.

Given the findings, and given that there is a known difference between the homozygous samples, which divide into two groups, I carried out one-way ANOVAs in which the homozygous samples were split into their two groups, Group One and Group Two. A total of 29 ANOVAs were found significant across the 41 cell subclasses and 9 deconvolutions. Of these, Tukey's posthoc test for difference between groups found a significant difference in cell type S1PyrL6 between Group One of the homozygotes and Group Two ($p=0.029$), and between Group One and the heterozygous samples ($p=0.045$). This was in the SI=0.9, HK=10 deconvolution. The results can be viewed in Figure 75.

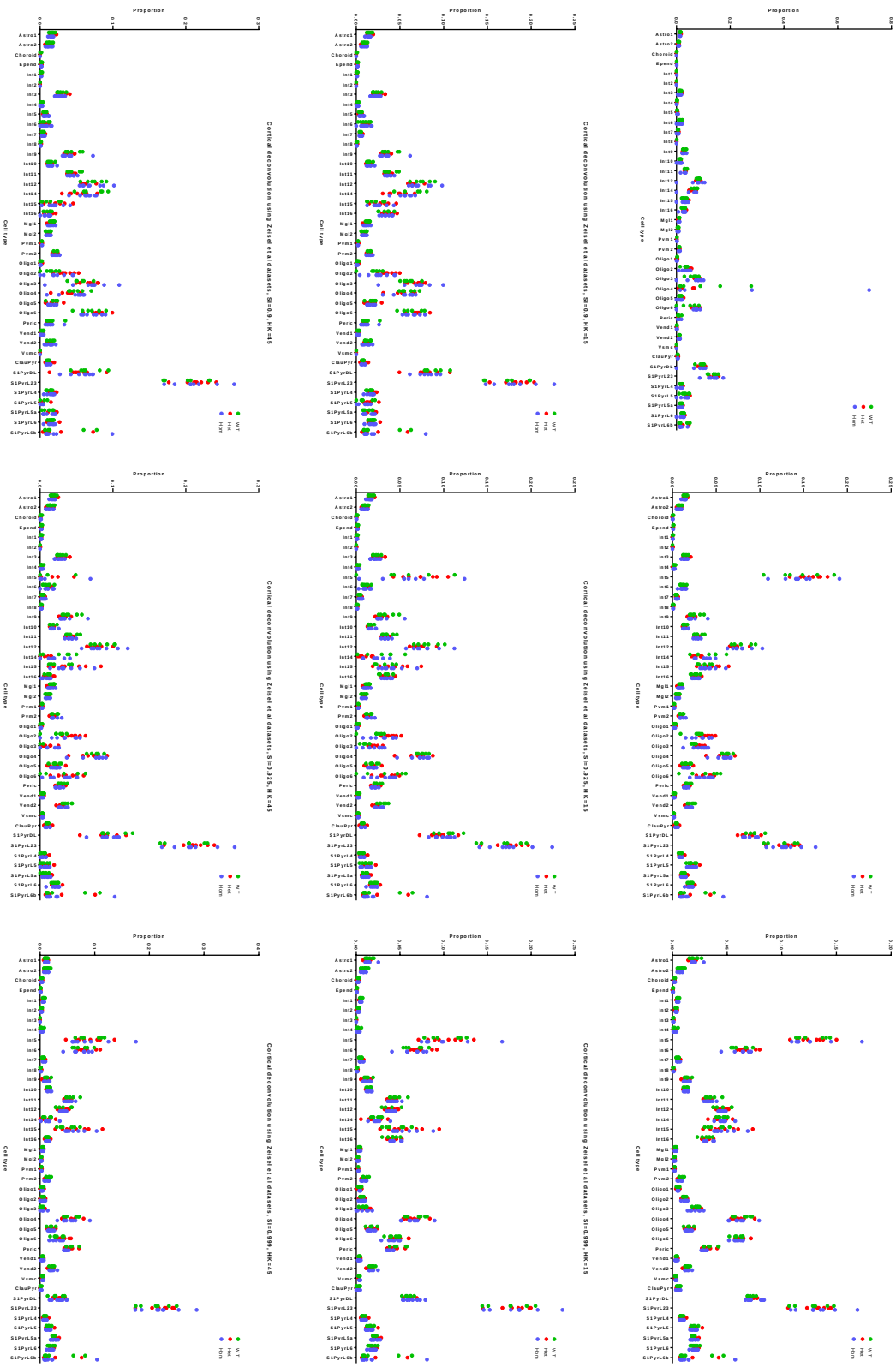


Figure 75. Deconvolution of mouse cortical samples using profiles from Zeisel *et al.*, with all cell profiles retained. WT=Wild-type, Het=Heterozygous, Hom=Homozygous, colours are green red and blue respectively.

6.2.5.2 Interneuron 5 removed results

The results can be viewed in Figure 76. One-way ANOVAs were carried out for each cell subclass to examine the effects of genotype on proportion. There were statistically significant differences between 4 group means as determined by one-way ANOVA for several subclasses, bearing a high degree of similarity to the previous deconvolution. Vend2 was again significant in every SI=0.925 deconvolution. As before, Tukey's posthoc test found no significant combinations.

I then carried out one-way ANOVAs in which the homozygous samples were split into Group One and Group Two. There were significant differences between 20 group means, which were strikingly similar to the previous deconvolution and are therefore discussed in length in the discussion section. No pairwise comparisons were significant.

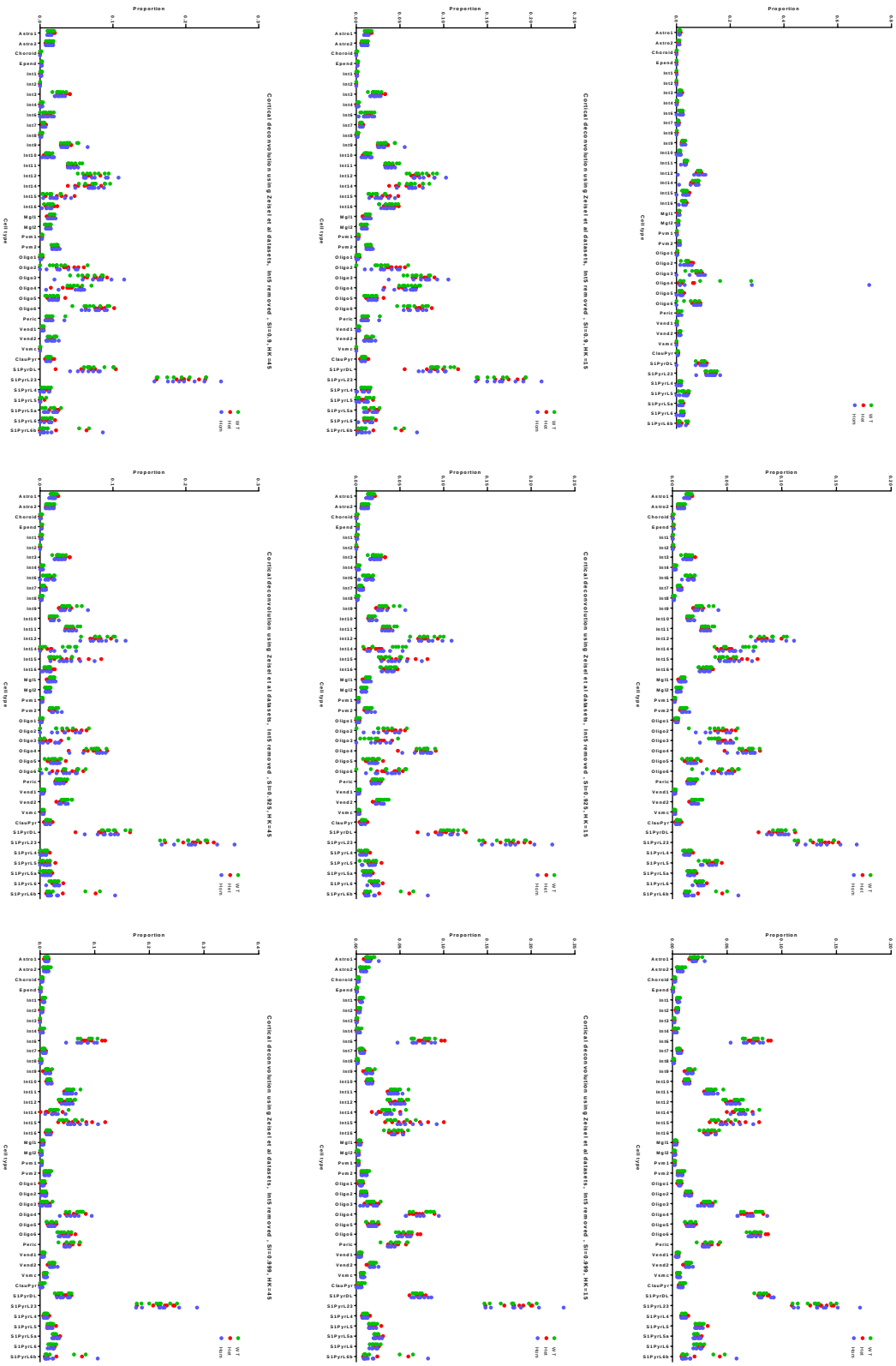


Figure 76. Deconvolution of mouse cortical samples using profiles from Zeisel *et al.*, with Int5 cell profile removed. WT=Wild-type, Het=Heterozygous, Hom=Homozygous, colours are green red and blue respectively.

6.2.5.3 Merged Interneuron 5-6-7-8 results

The results are displayed in **Figure 77**. The results are highly similar to the deconvolution without Interneuron 5. One-way ANOVAs were carried out for each cell subclass to examine the effects of genotype on proportion. There were statistically significant differences between 5 group means as determined by one-way ANOVA the same subclasses of Vend2 and S1PyrL5a, in the same deconvolutions as before. Once again, Tukey's posthoc test found no significant differences in pairwise comparisons.

I then carried out one-way ANOVAs in which the homozygous samples were split into Group One and Group Two. There were 15 statistically significant differences between group means, which again bore a close resemblance to previous analyses.

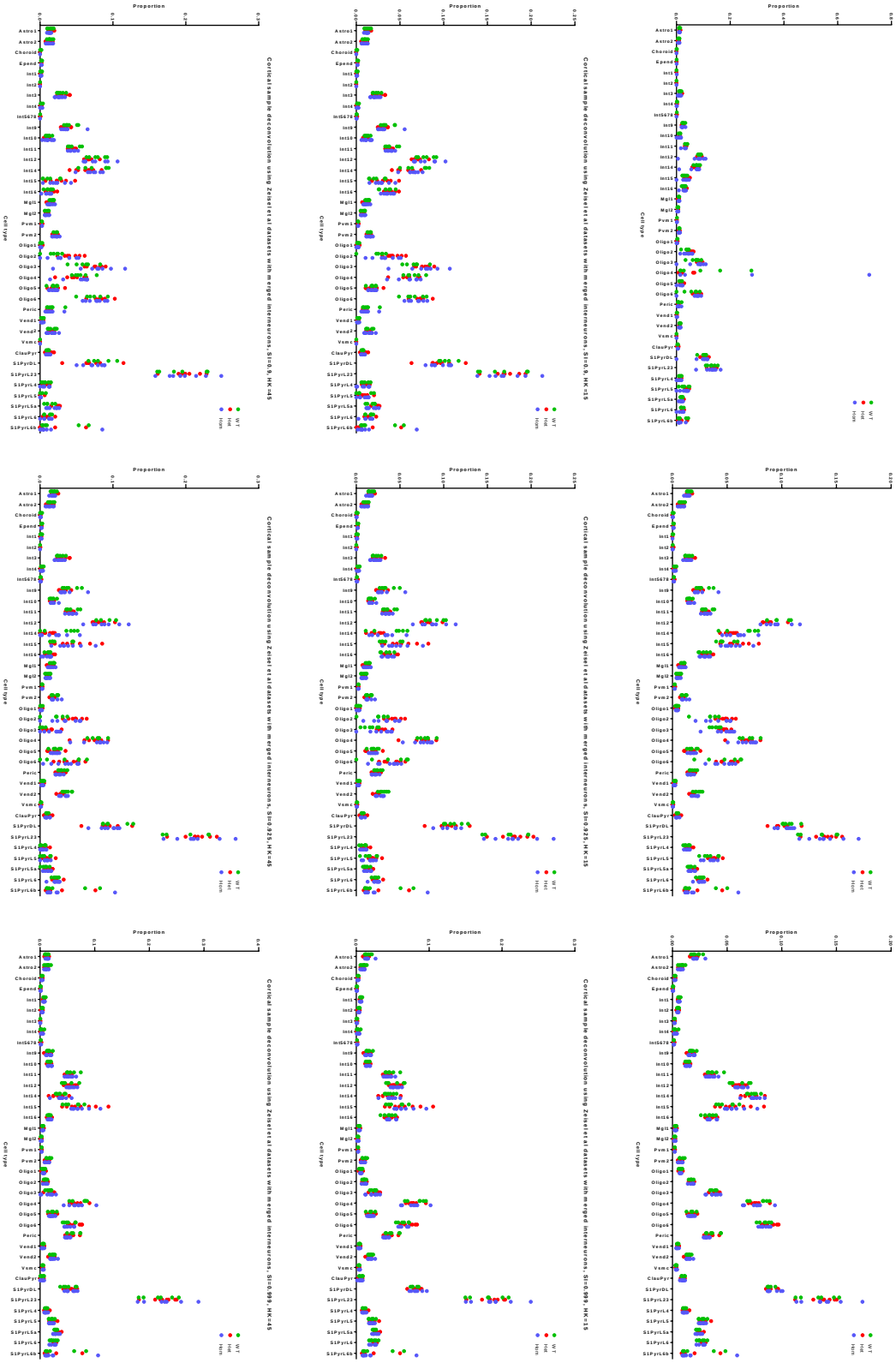


Figure 77. Deconvolution of mouse cortical samples using profiles from Zeisel *et al.*, with Interneurons 5, 6 , 7 and 8 merged. WT=Wild-type, Het=Heterozygous, Hom=Homozygous, colours are green red and blue respectively.

6.2.5.4 Conclusion

The results of the three deconvolutions are typically in agreement across one another, but across only some deconvolutions. Statistically significant differences were found in the comparisons and appeared in all three of the deconvolution approaches, summarised in Table 31. Note that only one deconvolution showed pairwise significance for genotype comparisons, and that in only one cell subclass.

	ANOVA			**ANOVA split**		
Cell subclass	Unaltered	No Int5	Int5678	Unaltered	No Int5	Int5678
Int3				1	1	1
Int4						
Int5				3		
Int10				6	6	6
Int12						
Int13						
Int14	1			3	1	
Int15				5	1	1
Oligo2				5	5	
Peric						
Vend2	3	3	3	3	2	3
S1PyrDL				2	2	1
S1PyrL5						
S1PyrL5a	2	1	2	1	1	2
S1PyrL6				1	1	1

Table 31. Condensed summary of deconvolution results. #=Number of ANOVA significances with p<0.05 in that set of 9 deconvolutions. Results with the homozygotes treated as one group are on the left, those in which they were split into two groups are on the right under "ANOVA split".

6.3 Mouse Hippocampal RNA-Seq deconvolution

6.3.1 Initial deconvolution of pseudosamples

The hippocampal dataset described in Zeisel *et al.* 1,301 single cell RNA-Seq profiles from 6 classes of cell and 38 subclasses. The data were treated in the same manner as the cortical subset of data; normalised to total transcripts sequenced, then averaged by subclass to give 38 subclass profiles.

For housekeeping normalisation, only 51 genes me*et al.*l the criteria laid out in 5.4.1.2. Therefore, when carrying out my deconvolution, the SI ranged from 0.7,0.725,0.75….0.975 and 0.999, and the HK ranged from 1 to 10, 15, 20, 25….50, 51. Pseudosamples were generated in the same manner as previously and error was measured in MAD.

The minimum MAD was 0.24 (at SI=0.975, HK=1) and the maximum was 0.98 (SI=0.999, HK=50). Increasing HK resulted in a roughly linear increase in MAD, and increasing SI resulted in small decreases, except for SI=0.999 which had the largest MAD for all HKs. I looked at the SI=0.975, HK=1 deconvolution in more detail, although using one housekeeping gene for normalisation is problematic. Increasing the number of housekeeping genes increased the number of cell types for which the average estimate was more than twofold greater than the actual proportion (at HK=1 one cell type met this criterion, at HK=3 three did, at HK=5 four did).

In the SI=0.975, HK=1 deconvolution the error seemed to be concentrated in one subclass, Choroid. Every single pseudosample underestimated its proportions by at least 50%. The only other subclass which was occasionally mis-estimated by a factor of 2 was Int12, with 16 proportions mis-estimated by this factor. It appears the issue may be with sufficient cell numbers as the hippocampal subset of the Zeisel *et al.* dataset only contains one Choroid cell.

6.3.2 Removal of Choroid

I removed the Choroid cell from the dataset, recalculated the pseudosamples, and deconvoluted with the same spread of SI and HK values.

As before, the minimum MAD was at HK=1, where it equalled 0.23 (SI=0.85). Maximum was 0.475, indicating that deconvolution is overall more accurate if Choroid is removed. I then looked at the HK=1 SI=0.85 deconvolution in more detail.

The mean of the average of the 100 estimation ratios per cell subclass was 1.17, while the average of the minimums was 0.87, and the maximum was 8.64. This is relatively poor compared to the unaltered deconvolution, where the figures were 1.002, 0.94, and 1.088. The main difference appeared to be that the altered deconvolution occasionally produces estimates which are highly inaccurate (ranging as high as x227 fold overestimated) while the unaltered deconvolution does not (it's highest overestimation is x2 fold). On this basis, it does not appear that removing the Choroid subclass is a satisfactory solution.

6.3.3 Merging Choroid and Ependymal cell types

Since the results had been favourable in the cortical deconvolution, I merged the Choroid subclass with the other subclass it clustered with, Epend, and recalculated pseudosamples.

The minimum MAD was 0.235 (SI=0.925, HK=1) and the maximum was 0.495 (SI=0.725, HK=50). I examined the best deconvolution in more detail. Several subclasses, including CA1PyrInt, Choroid/Epend and Vsmc, were consistently underestimated with maximum estimates of 0.56, 0.75, and 0.74 respectively. However, the mean of the average estimation ratios per cell subclass was 1.0009, while redoing the deconvolution with more housekeeping genes resulted in a higher "average of averages". It appeared that averaging the Choroid and Epend subclasses had barely improved Choroid deconvolution, but had introduced an equally severe problem with the CA1PyrInt subclass and a lesser one with Vsmc. I therefore elected not to deconvolute by merging these lines.

6.3.4 Deconvolution of Allen comparison dataset

I deconvoluted the Allen dataset using the same spread of HK and SI values as used in the Zeisel deconvolution. These are cortical cells, so the deconvolution is not

likely to be as accurate as it would otherwise be. The greatly decreased number and quality of housekeeping genes may also result in inferior deconvolution compared to the cortical Allen deconvolution. As with that deconvolution, the goal here is to ensure that a minimum level of accuracy is retained. The discussion in 6.2.4 is relevant to this deconvolution as well, and the results are presented in the same manner.

6.3.4.1 Results

I deconvoluted the Allen dataset with the same housekeeping genes, and range of HK and SI values as in the pseudosample deconvolution. The minimum MAD there was at SI=0.975, HK=1, while the maximum was SI=0.999, HK=50. A summary of the results of these Allen deconvolutions using these settings are shown in Table 32 and Table 33 respectively.

Predicted as being (on average):	**Number of cells**	**Astrocytes**	**Other**	**Neuron**	**Interneuron**	**Microglia**	**Oligodendrocyte**
GABAergic	5,288	11.60%	23.20%	33.70%	42.50%	0.37%	5.80%
Glutamatergic	6,736	8.60%	33.30%	48.70%	20.70%	0.97%	11%
Non Neuronal	431	8.90%	20%	39.90%	19.80%	9.30%	9.60%
Endothelial	144	2.50%	47.90%	14.90%	6.90%	0.99%	16.70%

Table 32. Results of deconvolution of SI=0.975, HK=1 deconvolution, the most accurate settings for pseudosample deconvolution.

We can see that at the minimum MAD settings the GABAergic and glutamatergic cells have a relatively high average predicted proportion. The Non-Neuronal cells are heavily predicted as being neurons; this is concerning as these are mostly astrocytes and oligodendrocytes.

Predicted as being (on average):	Number of cells	Astrocytes	Other	Neuron	Interneuron	Microglia	Oligodendrocyte
GABAergic	2,626	25.1%	16.3%	31%	38%	2.2 %	2.8%
Glutamatergic	5,340	14.4%	22.5%	57.2%	19.6%	1.3%	6.6%
Non Neuronal	2	~	1	*	~	~	~

Table 33. Results of deconvolution of SI=0.999, HK=50 deconvolution, the least accurate settings for pseudosample deconvolution. *=1 VLMC cell predicted as 89% due to 89% predicted as VSMC. Both are smooth muscle cells.*= 1 oligodendrocyte predicted as 71%

At the maximum MAD settings, the settings are if anything slightly better at predicting each cell type identity, although it is possible that this is due to the higher HK settings removing more cells, leaving only those which are closer in character to those of the Zeisel dataset.

I thought it might be of use to look at the standard deviation of each guess, as this would indicate whether similar cells (by Class) gave similar values. The results of this are given in Table 34 and Table 35 for the most and least optimal deconvolution settings in terms of MAD, respectively.

Predicted as being (on average):	Number of cells	Astrocytes	Other	Neuron	Interneuron	Microglia	Oligodendrocyte
GABAergic	5,288	0.172	0.311	0.338	0.386	0.022	0.15
Glutamatergic	6,736	0.148	0.351	0.346	0.311	0.026	0.239
Non Neuronal	431	0.261	0.329	0.404	0.328	0.246	0.249
Endothelial	144	0.135	0.415	0.289	0.188	0.046	0.345

Table 34. Results of deconvolution of SI=0.975, HK=1 deconvolution, the most accurate settings for pseudosample deconvolution.

Standard deviation for:	Number of cells	Astrocytes	Other	Neuron	Interneuron	Microglia	Oligodendrocyte
GABAergic	2,626	0.231	0.238	0.305	0.338	0.044	0.126
Glutamatergic	5,340	0.154	0.262	0.323	0.285	0.037	0.186
Non Neuronal	2	0	0.266	0.357	0.082	0.021	0.05

Table 35. Standard deviation of deconvolution of SI=0.999, HK=50 deconvolution, the least accurate settings for pseudosample deconvolution.

There does not appear to be any particular pattern of standard deviation; the better settings do not give any noticeable decrease in standard deviation and occasionally have a higher value, indicating more variation. The same findings as in the cortical deconvolution apply- the analysis must be viewed with caution.

6.3.5 Deconvolution of mouse Der1 hippocampal samples

6.3.5.1 Unmerged deconvolution results

As before, I deconvoluted using a variety of settings. I had shown that increasing HK increased MAD in pseudosample deconvolution, and that SI=0.99 also had very high MAD. The lowest MAD was at SI=0.975. I therefore utilised the settings of SI=0.925, 0.95, 0.975, and HK=1, 2, and 5. The results of the deconvolution can be seen in Figure 78. There is also large variation in the proportion of the CA1Pyr1 subclasses in the HK=1 deconvolutions. Both these issues are possibly due to using just one housekeeping gene in normalisation. One-way ANOVAs were carried out for each cell type to examine the effects of genotype on proportion. There were 15 statistically significant differences between group means as determined by one-way ANOVA. Tukey's posthoc test for difference between groups found that no pairwise differences were significant. Int4 and Int14 were each significant 5 and 4 times respectively, while Peric was 3 times, and Int3 was twice. Int12 was significant in a single deconvolution.

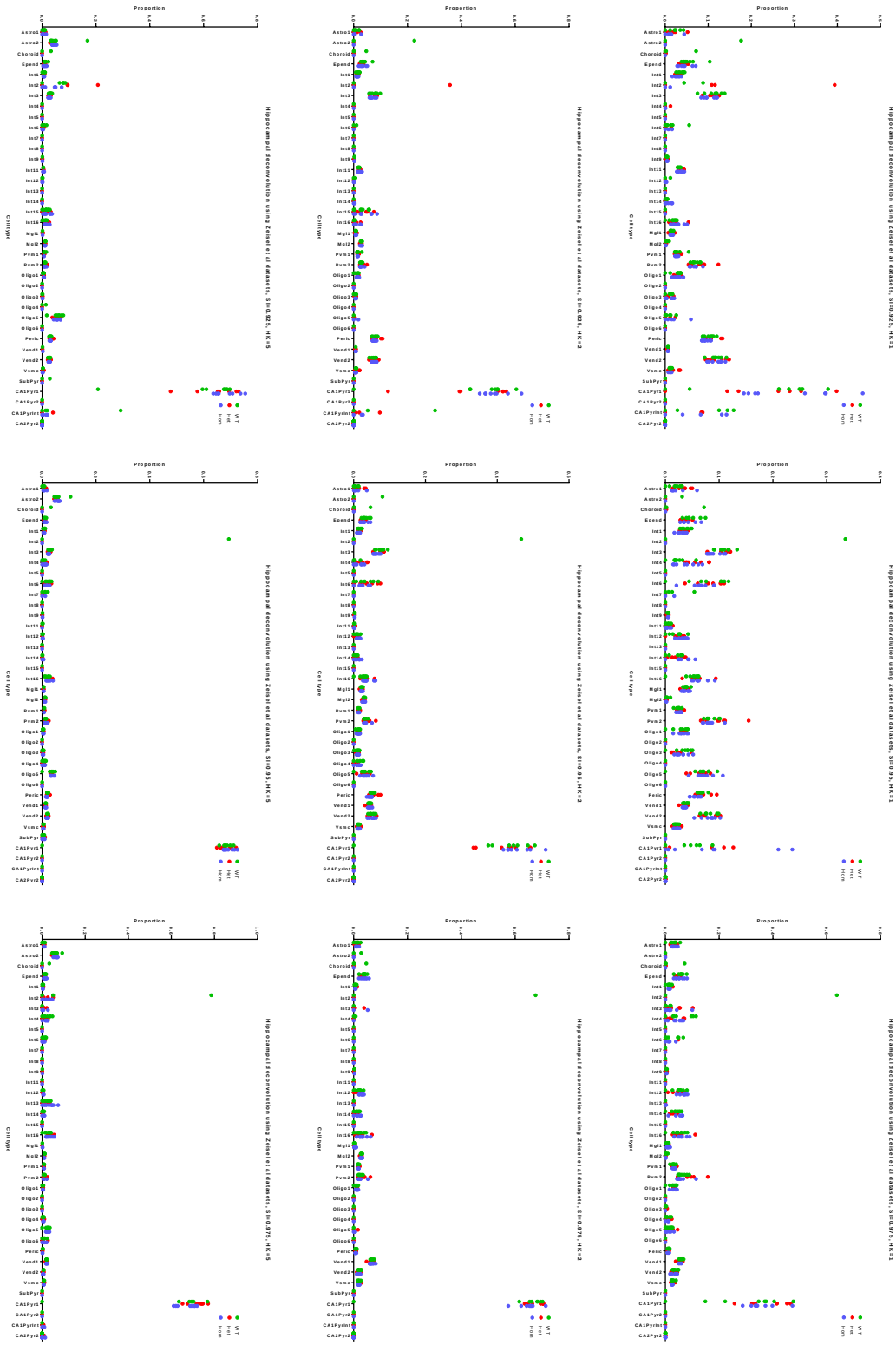


Figure 78. Deconvolution of mouse hippocampal samples using profiles from Zeisel *et al.*. WT=Wild-type, Het=Heterozygous, Hom=Homozygous, colours are green red and blue respectively.

6.4 Human iPSC-derived neuron deconvolution

As it had shown itself to be an acceptable method of comparing datasets, and gave superior deconvolution, I carried out housekeeping gene normalisation of the datasets prior to deconvolution. This was carried out exactly as before; excluding genes with a fourfold difference between the maximum and minimum samples values, taking the geomean of the coefficient of variations for both datasets, and taking a number of genes to take the geomean of as a normalisation quotient. In addition, if the deconvolution dataset was mouse, then only orthologous genes were utilised.

6.4.1 Selection of appropriate datasets for human deconvolution

6.4.1.1 Zeisel *et al.* dataset

The Zeisel *et al.* dataset has been previously described and was utilised in addition to the Zhang *et al.* dataset. The cortical subset of this dataset was utilised as this is the most appropriate comparison for the human neurons.

6.4.1.2 Allen *et al.* datasets

The Allen Brain Atlas datasets are described in detail in their white paper as well as at the web address http://celltypes.brain-map.org/rnaseq (accessed on 16/10/2018). The human dataset I utilised is comprised of single nuclei RNA-Seq of the middle temporal gyrus. Single nuclei RNA-Seq is somewhat less than ideal as many transcripts in human neurons are locally translated at dendrites and other non-nuclear locations. These locally translated transcripts number as high as 2,550, although of course many of these may also be nuclearly translated and will be partially represented in the Allen datasets[277]. It is clear that single nuclei sequencing, as opposed to single cell, will not capture all the information available. There are 15,928 nuclei derived from 8 post-mortem human adult brains. The average sequencing depth is 2.63x106 reads, comparable to Darmanis *et al.* but only about a tenth of the depth of our samples. Accordingly, far fewer genes were detected, ranging from 6,186 to 9,937, depending on the cell subclass (GABAergic, glutamatergic, unassigned, non-neuronal). The dataset therefore suffers from many of the same issues as the Zeisel *et al.* dataset used in the examination of detailed cell

subclasses in the mouse deconvolutions, but it has the advantage of being of human origin.

6.4.2 Deconvolution of pseudosamples

I carried out deconvolution of pseudosamples using a variety of SI and HK gene values. The values for SI ranged from 0.725, 0.75, 0.775… up to 0.975 and 0.999. HK values ranged from 1 to 10, then 15, 20, 25….50. A total of 216 deconvolutions were therefore carried out.

As with the analogous mouse deconvolutions, increasing SI tended to increase the accuracy of the deconvolution. This could be because genes involved in specialised subclass functions have tended to be well conserved, so comparisons to a new species do not heavily alter the pattern of deconvolution. Deconvolution was actually superior in terms of MAD; the minimum was 0.14 (SI=0.925, HK=2) and the maximum was 0.26 (SI=0.725, HK=50). This compares to a low of 0.22 up to a high of 0.33 in the mouse cortical deconvolution. The optimum SI was always 0.925, while the next was always 0.95 or 0.975. It is noteworthy that the largest number of housekeeping genes and the most loosely defined markers gave the worst deconvolution; a result which makes sense. Although high SIs are typically superior in other deconvolutions, high HK numbers are not always.

I examined the optimum deconvolution in more detail. There were no cell subclasses with over 10 estimations which were more than twofold incorrect, although as with the mouse deconvolution the “Int5” subclass was the worst predicted with a maximum overestimation of 7.49 fold and an average overestimation of 1.75 fold. On average though, cell subclasses were well predicted. The average of the average estimated:actual proportion for each cell subclass across 100 pseudosamples was 1.02, while the average of the minimum and maximums were 0.96 and 2.16. All in all, these are reasonable deconvolutions. I decided to not attempt to remove “Int5”, as this had not helped in the mouse deconvolution to any great degree. I also looked at the deconvolution with SI=0.925 and HK=15, which has a MAD of 0.199. I was interested in whether this would be a useful alternative setting due to concerns about using two housekeeping genes to normalise. Unfortunately the “Int5” subclass was

badly overestimated here with an average estimation of 2.75 times the correct proportion and with one third of samples having an estimation which was at least twofold incorrect. I therefore elected to continue with the SI=0.925 HK=2 settings, although I would first need to see how these deconvoluted another dataset.

6.4.3 Deconvolution of Allen comparison dataset

I used the same range of settings for SI and HK, and the dataset described in 5.6.1.3. As described in the Allen section in the mouse cortical and hippocampal deconvolutions, the data is presented as violin plots where appropriate.

Although I expected some cells to be dropped as HK increased (due to lack of expression for some of these genes) the dropout rate was extremely high. By HK=20, only 43 cells were left in the analysis, all of them Glutamatergic. It is possible that this is reflected in the fact that the human optimal HK number is frequently low regardless of the deconvoluting dataset, while the mouse deconvolution optimal varies but is 45 in the cortical Zeisel deconvolution and 40 or 25 in the cortical and hippocampal Zhang deconvolutions, respectively. Another likely factor is the relatively restricted number of cell types (primarily astrocytic and neuronal) in an iPSC-derived neuronal culture compared to an adult mouse brain; selection of housekeepers from this will likely discriminate against non-neuronal cell types. By HK=10, 3806 cells remain, and at HK=15, it is 500.

It was very difficult to discern a pattern across deconvolutions at the lower HK numbers, especially if I looked at Allen subclasses. To resolve this I looked at class level rather than subclass level. See results of the optimal deconvolution in terms of MAD, SI=0.925, HK=2, in Figure 79. Cell types had what appeared to be a nearly binary split between cells that were 100% of that type and those that were 0%.

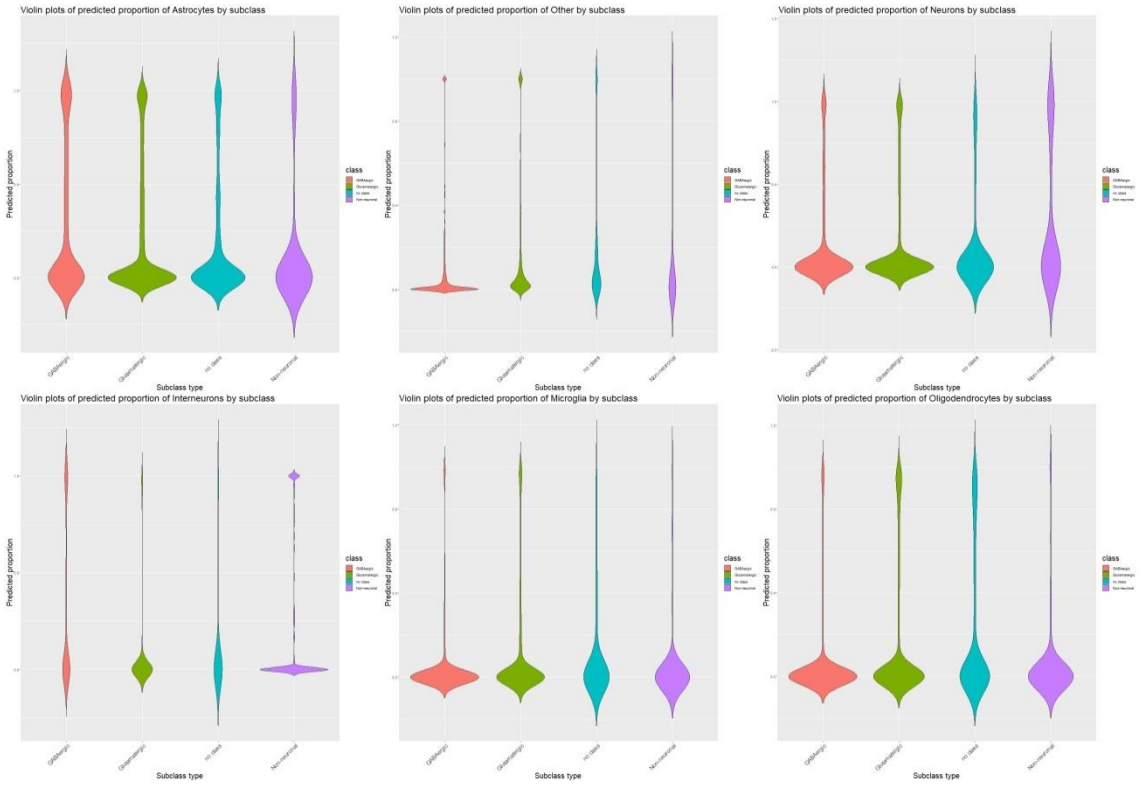


Figure 79. Violin plots of all 6 Zeisel class proportions for 4 classes of Allen cell, colour coded by Allen class. Red=GABAergic, Yellow=Glutamatergic, Blue=No class, Purple=Non-Neuronal. These results are of the SI=0.925 HK=2 deconvolution. Allen classes are shown rather than classes for ease of viewing.

We can see that all classes show a clear split between these binary options, regardless of actual cell identity. This is most pronounced in the "Interneuron" proportion and least in the "Astrocyte". I also observed that in general, very low SI values (0.725 chiefly) had a high propensity for higher levels of Interneuron prediction across all cell types, although a substantial proportion of cells were still reporting 0 proportion of this. In addition, there was little change in deconvolution proportions with increasing SI after SI=0.925, across almost all HK values. Finally, there was no deconvolution which had a majority of cells of any type predicted as mostly being of that cell type.

To conclude; it does appear that the deconvolution as applied to other datasets gives very poor results. This is probably partially due to the choice of "housekeeping" genes, as these were chosen specifically for each dataset and the human t(1;11) samples. I believe this is the case as the majority of Allen cells do not even express all of these housekeeping genes. There are some notable facts; firstly, that the

deconvolution giving the optimal pseudosample deconvolution is in no way inferior to other deconvolutions, and that no deconvolution gives even majority correct proportions for each cell type. Secondly, the deconvolution appears to "settle" at SI=0.925, probably indicating that this is at least a stable if not optimal setting for gene specificity. Given this, I decided to press ahead and deconvolute the human t(1;11) samples.

6.4.4 Deconvolution of human t(1;11) samples

I carried out the deconvolution of the human t(1;11) samples using a variety of settings, normalising to total count depth for all samples. The results are displayed in Figure 80.

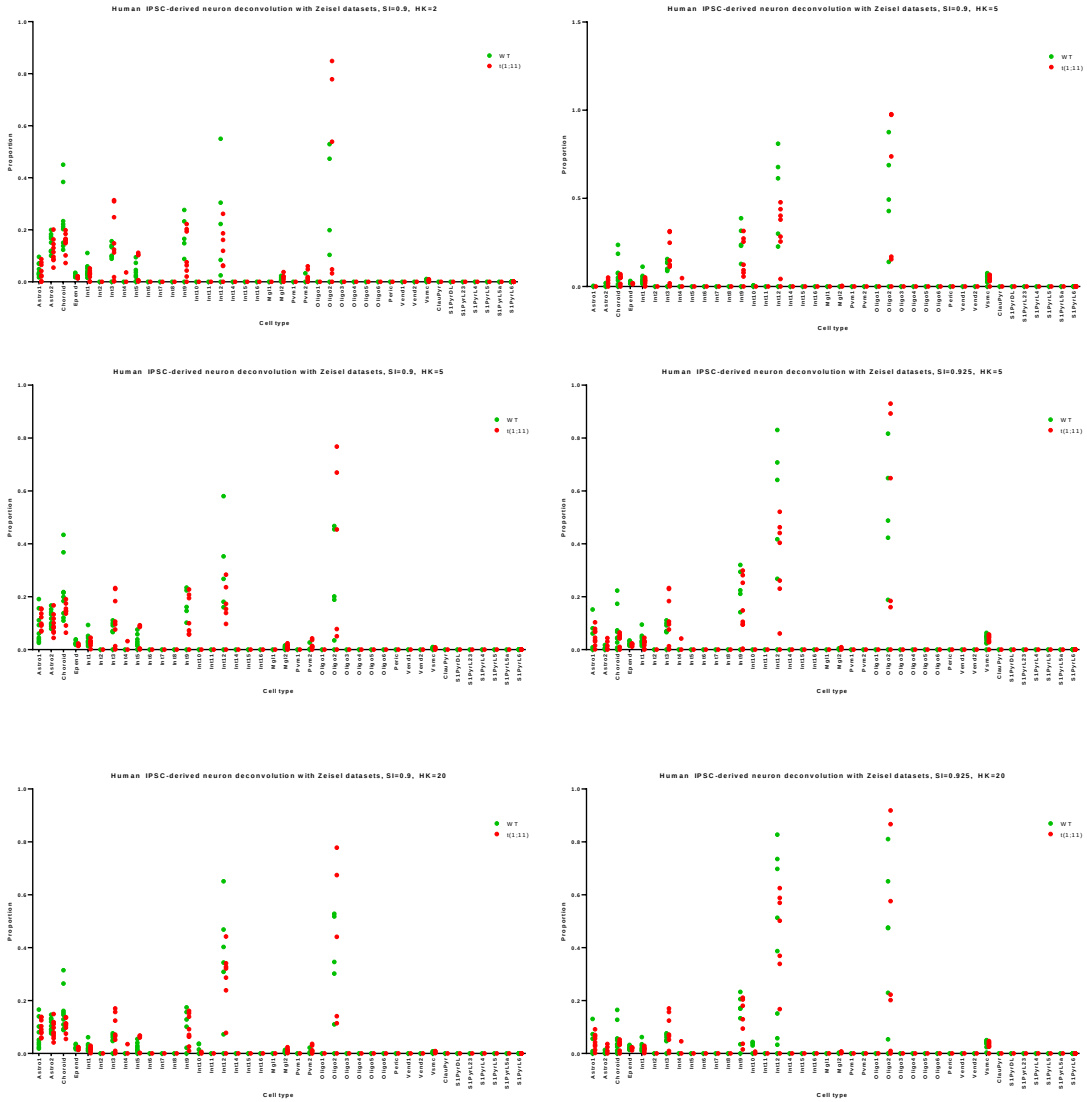


Figure 80. Deconvolution of human t(1;11) samples using profiles from Zeisel *et al.*. WT=Wild-type, t(1;11)=carrier, colours are green and red respectively.

T-test revealed that no cell type was significantly different in proportion across the two genotypes once multiple testing correction (Sidak-Bonferroni) was applied. Discussion of the three deconvolutions is in the Discussion chapter, but the overall findings are of low reliability.

Deconvolution of the RNA-Seq data using Zeisel et al. scRNA-Seq datasets

7 INVESTIGATION OF DIFFERENTIALLY EXPRESSED GENES PERTAINING TO CELL TYPES

7.1 Introduction

Expression weighted cell type enrichment (EWCE) is a method developed by Skene *et al.* to determine whether a list of genes, or SNPs associated with genes, have greater expression in a particular cell type than expected by chance[281]. In addition to a list of genes of interest, the approach requires a database of expression profiles for reliably identified cell types such as that of Zeisel *et al.* Skene *et al.* have applied this method with success to investigate whether genomic variants associated with schizophrenia converge on particular cell types[133]. The rationale behind this approach is to determine which cell types are of particular relevance to disease, which will be of use in drug design and disease modelling. EWCE essentially sums for a list of genes the specificity values (similar to SI) in each cell type to give a set of specificity scores, then selects a large number of lists of equal length from the background list of expressed genes to give a probability distribution of scores for each cell type. One can then see how whether a list of interest contains more highly specific genes for a particular type cell than the background rate of specificity in a large number of lists. Skene *et al.* recommend utilising at least 10,000 background lists to give a fair sampling of specificity.

Skene *et al.* used a large database from the Karolinska Institute, of which the Zeisel *et al.* dataset is a subset. This encompassed mouse scRNA-Seq datasets generated from nearly 10,000 cells, identified by hierarchical clustering, originating from the midbrain, hypothalamus, striatum, cortical interneurons, an oligodendrocyte dataset, and the somatosensory cortex and hippocampal datasets described by Zeisel *et al.* Data were generated in the same manner as Zeisel *et al.* They found that hits from the CLOZUK and PGC GWAS (both schizophrenia GWAS) were significantly enriched in the broader cell classes of striatal medium spiny neurons, cortical interneurons, neocortical somatosensory pyramidal cells, and CA1 hippocampal pyramidal cells, and looked at more specifically defined subclasses as well[29,60]. For depression they found enrichment for genes specific to GABAergic and dopaminergic interneurons. It should be noted that they didn't impose a minimum value for what was considered a "specific gene", as the values are summed across all genes for each cell type, although they did exclude genes with low expression. They

also found that genes that encoded proteins which are i) Antipsychotic drug targets ii) loss of function mutation intolerant iii) part of the postsynaptic density and iv) RBFOX binding were significantly enriched for specificity in the aforementioned medium spiny neurons and pyramidal cell types. Finally, it should be noted that they carried out further analyses controlling for the enrichments found in significant cell types- a way to see if the significance signals were overlapping. This analysis showed that the significance of the hippocampal and cortical pyramidal cells were not independent, i.e., were caused by the same genes. This implies a common pyramidal cell dysfunction in schizophrenia.

I wanted to investigate whether the lists of differentially expressed genes from my mouse and human RNA-Seq t(1;11) and controls showed any enrichment for genes specific to particular cell types. My question was simpler to answer than that posed by Skene *et al.*. Since I do not have GWAS level data, but only lists of differentially expressed genes from particular brain regions or tissue cultures, I do not need to consider factors such as LD, ambiguously placed SNPs, or gene length. I was particularly interested in whether there would be any convergence between my results looking at a unique and highly penetrant translocation, and those of Skene *et al.*, who used information from GWAS looking for common variation associated with mental illness. Given that pyramidal cells seemed to be of particular importance in the Skene *et al.* analysis, I was especially curious to see if this would re-emerge. If there were overlaps, this would imply that the aetiology of the t(1;11) related psychiatric illness is similar to that of psychiatric illness related to common variation. If not, this could imply that a different route to the disease state is found in each scenario, meaning that the same phenotype is reached through different biological processes. There are some drawbacks; Skene *et al.* were able to look at schizophrenia and depression associated variants separately, while in the Scottish pedigree the t(1;11) predisposes to both and accounts for most of the genetic risk for psychiatric disease. The genetic aetiology across family members with the translocation is likely to be very similar and it will not be possible to parse out risk to the different disorders, as Skene *et al.* did.

It should also be made explicit what this study is and what it is not. EWCE does not mean that the alterations necessarily exert their effects in the cell types indicated by the analysis; just that those cell types highly express those altered genes. It is of course likely that these genes would have important functions for that cell type, and it would be the logical starting point when searching for altered physiology. There are similarities to the deconvolution approach in that both look for altered genes associated with particular cell types; however there are key differences. First is that input and output of deconvolution is quantitative, even if only roughly so. Second is that the downregulation or upregulation of a gene associated with a particular cell type both implicate it in EWCE; while these signals have opposing effects in deconvolution approaches. Thirdly, the comparison dataset is important. It should be as close as possible in nature to the sample the gene list originated from. A test may indeed highlight certain cell types as significant; but if these cell types don't exist in the original sample it is questionable how useful this is. Since the approach utilised by Skene *et al.* uses gene enrichments; including new cell types can alter the relative strength of each cell type signal.

Skene *et al.* utilised the full Karolinska Institute Superset of data, of which the Zeisel *et al.* cortex and hippocampus datasets constitute a subset. This superset has 24 cell classes and 149 cell subclasses from diverse regions and cell types of the brain. All cells were sequenced in the same manner, using UMI tags as described earlier in this thesis so as to allow accurate quantification. The cells were then clustered in the same hierarchical clustering method as Zeisel *et al.*, which generated classes and subclasses. Skene *et al.* then identified these using known marker expression, histology, or molecular studies[133]. It also contains some embryonic cell types.

It is key to note that the samples I am using for the EWCE are cortical-like in the case of the t(1;11) cells, and are hippocampal and cortical (including most of the whole brain except hippocampus) for the mouse samples. Skene *et al.* are asking a broader question relating to many regions of the brain, whereas I have lists of differentially expressed genes from the hippocampal region of the brain or brain minus hippocampus, or a neuronal cortical like culture. Skene *et al.* used lists of genes associated with major mental illness by GWAS or MAGMA studies as they

were seeking in which brain regions and which cells these mutations exerted their effects (assuming a cell with high expression of the gene would be affected by the mutation). Hits from GWAS or MAGMA are not definitively associated with any brain region (indeed, this was what Skene *et al.* were interested in), so it was appropriate for Skene *et al.* to use datasets from all parts of the brain. In contrast my lists of differentially expressed genes are tied to the culture or brain region their samples originate from. In this sense it is more sensible to compare them to a similar cell dataset if possible. A positive result for a gene list from a culture or a brain region highly similar to the region the cells are from is more informative; it means that this perturbed gene list from a particular type of culture corresponds to a particular cell type. Given that I could not rebuild the Karolinska dataset to remove certain brain areas, I was left with three choices for each gene list; whether to compare to the KI superset (with or without embryonic cells), to the Zeisel hippocampal dataset, or to the Zeisel somatosensory cortex dataset.

I also investigated whether it would be possible to rebuild a new dataset from the datasets indicated in Skene *et al.*. This would contain the majority of the cell types across many brain regions, but not the cells from the hippocampus (contributing both to hippocampal-unique classes and common classes), and would therefore be a better comparison for the cortical gene lists than the KI superset. However, not all of the data is released to the public yet and only the cell profiles, not the individual data for each of the thousands of cells used to make those profiles, are available. I had considered removing the hippocampal cells from the superset as well, but without the individual cell data from all papers, not just Zeisel *et al.*, there is no way to remove the contribution of each brain region to the cell profiles or much more importantly know whether any of the cell profiles are brain region specific. Of the 149 cell subclasses, I do not know how many cells were used to produce each cell subclass, or where in the brain these cells were derived from, or whether the subclasses are present in all parts of the brain or only some. I therefore did not alter the KI superset in removing the hippocampal specific lines. Given that common cell types clustered together by group this likely means that they are largely similar across brain regions and are not pressing problems. I did remove embryonic cell types from the analysis

as these are not relevant to adult cells and the relevant cell types are clearly designated; the specificity values for each gene were then recalculated. The results for the cortical lists are also highly similar if hippocampal cells are removed (data not shown).

I also argue that two issues are particularly important in considering the appropriate comparison. The first concerns risks of false positives, the second risks of false negatives.

- Genes randomly selected from the list of expressed genes in brain region X are more likely to be specific to/enriched in brain region X, compared to genes randomly selected from the list of expressed genes in all brain regions. This is because the list of X-expressed genes is a subset of all expressed genes. It is by definition enriched for X-specific genes if they exist. As EWCE utilises gene specificity, this means that cell types which are highly enriched for these X-specific genes will appear significant. These cell types will also by definition be X-specific. As differentially expressed genes between sample A and B of region X can only include genes expressed in region X, this list of differentially expressed genes therefore will appear enriched for genes and cell types specific to X if compared to similar length lists drawn from all across the brain. There is therefore an unknown risk of false positives in comparing to a larger dataset. A larger datas*et al.*so requires a greater multiple testing correction; therefore only the correct comparison should be utilised, as over-comparing will dilute significance due to including spurious cell types, while under-comparing will inflate significance.
- If most cell types are genuinely targeted by the mutation, then it will be difficult to assess cell type significance as there is no “non-significant” baseline to compare to. One would in this case just say that specific genes in general are targeted, but this could be due to other, unknown factors.

As the cortical samples we utilised consisted of large tracts of the mouse brain, a comparison to the entire KI superset seemed most appropriate rather than using the somatosensory cortex dataset described by Zeisel *et al.*. I removed embryonic cell

types. Although psychiatric illnesses are neurodevelopmental conditions, the corresponding cell types are no longer represented in the mouse RNA-Seq profiles as they were produced from adult mice. I also used the KI superset, similarly without embryonic cell types, to analyse the hippocampal gene lists. However I urge caution in interpreting the results due to the issue of p value inflation I raise above. However, convergence between this and other studies of the *Der1* hippocampus mean that there is supporting evidence for what I eventually found, and as the more informative comparison it made sense to utilise the larger dataset. I therefore used the KI superset without embryonic cell types to analyse the mouse derived gene lists.

I used the Zeisel *et al.* cortical profiles to analyse the t(1;11) gene list, as the cells have been described as being "foetal cortical" in nature. I considered using the KI superset either in its totality, or just utilising embryonic cell types. There are eight embryonic-unique cell types in total. They are limited in their scope; there are three dopaminergic cell types, three GABAergic cell types, oculomotor and trochlear nucleus embryonic neuorns, and red nucleus embryonic neurons. As described above, the astrocytes, oligodendrocytes, etc. profiles are generated from all regions of the brain and all samples, and it is not possible to separate out the contributions from embryonic cell types. It is therefore not possible to create new cell profiles using just embryonic cell types for embryonic astrocytes, embryonic oligodendrocytes, etc. from the KI superset, although these "embryonic" versions of adult cell types surely exist. An embryonic dataset is therefore extremely limited in scope. I therefore had to choose between using the entire KI superset or the Zeisel cortical dataset. The former gives inaccurate comparisons as the cell types are from the whole brain, and only a few are embryonic. The Zeisel dataset is non-embryonic but does include a variety of cell types. I opted to use it.

It is important to recall DISC1's involvement in a diverse array of processes including cellular migration, development, and mitochondrial activity (see Introduction). DISC1 immuno-reactive neurons have been found throughout all layers of the human cortex, and in rat cortical astrocytes, neurons, oligodendrocytes, and microglia[69,268]. Hence, there is a potential for the t(1;11)/*Der1* to alter the activities of a wide variety of cell types.

Figure 81 is taken from Zeisel *et al.* describing their cell profiles from both hippocampal and cortical samples[157]. It shows some of the transcription factor genes highly expressed in each of their cell subtypes, as well as some pan-type transcription factors. As many of these, including the orthologues of *Dlx6*, *Dlx1*, *Lhx9*, *Epas1* are differentially expressed in the t(1;11) samples it is likely some of these cell types will be highlighted by the analysis.

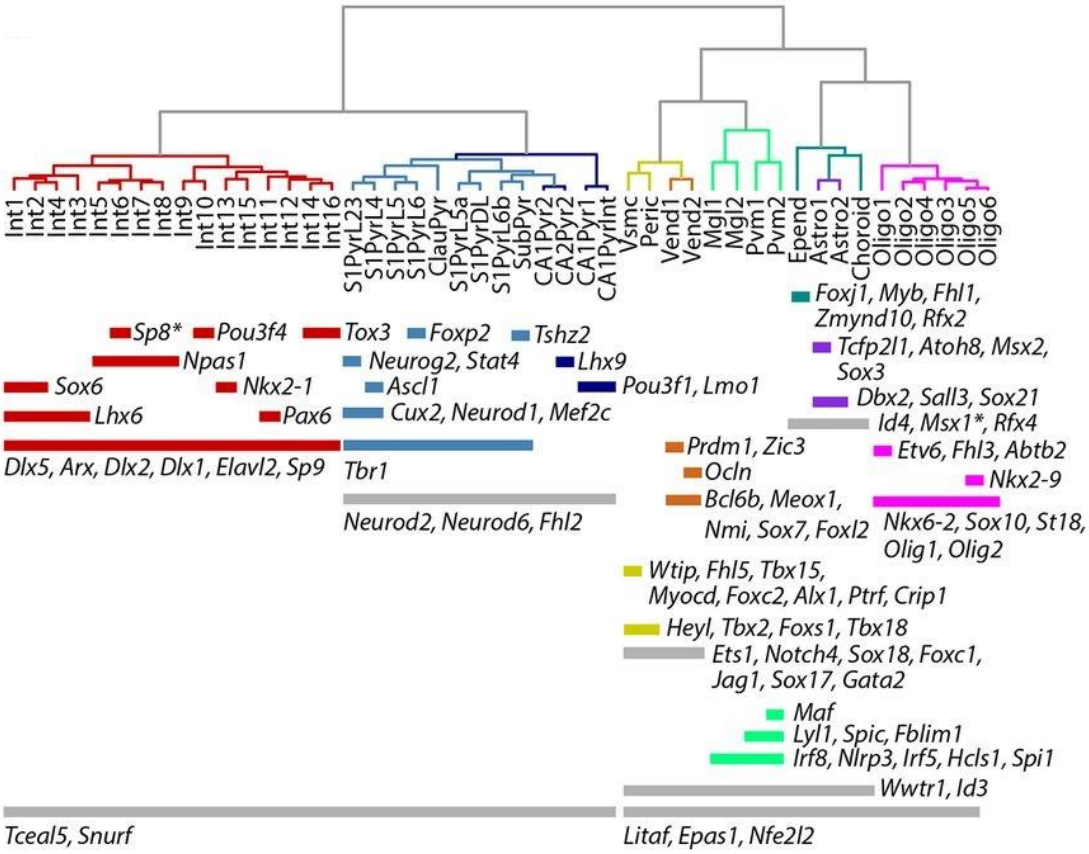


Figure 81. Figure from Zeisel *et al.*[157]. Each cell type is colour coded and subtypes are labelled. Inteneuron=Red, Neuron=Blue, Orange/Yellow=Endothelial/Mural, Microglia=Green, Astrocyte/Ependymal=Purple/Dark Green, Oligodendrocyte=Pink. Bars indicate how widespread a gene is, with narrow bars being specific to only a few subtypes.

7.2 Results

In general I found that the class level identifications closely tracked the subclass results in terms of significance. All human derived gene lists were compared against the Zeisel cortical dataset. The Der1 cortical and hippocampal mouse data were compared against the KI superset minus embryonic samples. In all cases I used genes

implicated by DESeq2; those which are differentially expressed at the whole gene level with a BaseMean greater than half that of *DISC1/Disc1* and an adjusted p value<0.05. This is the appropriate list of genes to use as the reference datasets use gene level not exon level expression. Genes were used regardless of direction or magnitude of fold change, as the goal of EWCE is just to see what cells might have transcriptional alterations. All comparisons used $1x10^5$ lists selected from the background to build up a background distribution of gene specificity. Bonferroni corrections were used for multiple testing. It should be noted in some cases that the actual list of altered gene was more enriched for cell types than any of $1x10^5$ lists selected from the background. A Monte-Carlo based analysis I developed also closely concurred with the significance of all cell types (data not shown); this was initially developed as there were issues with utilising the "EWCE" package. These issues were eventually overcome.

7.3 Human iPSC-derived neuron data

I looked for enrichments of specific genes in both the 8 class level identifications of the data described in Zeisel *et al.*(interneuron, astrocyte/ependymal, etc) and the 41 subclass level identifications. Given that Skene *et al.* have used more datasets, their class and subclass identifications are different.

The results of this analysis are displayed in Figure 82. The sample list contained 1,252 genes. We can see that a variety of cell types have significant enrichment for the genes differentially expressed in iPSC-derived neurons carrying the t(1;11). It is of course not necessarily the case that these cells actually exist in the iPSC-derived culture. Notably, only a single type of pyramidal cell showed significant enrichment. This was SS_S1PyrL23. These cells are distinguished by mainly being expressed in Layers 2 and 3 of the cortex and for having high staining of the *Rasgrf2* gene compared to other cells. They lack expression of deeper cell markers such as *Synpr2, Foxp2, Cplx3,* or *Rorb*.

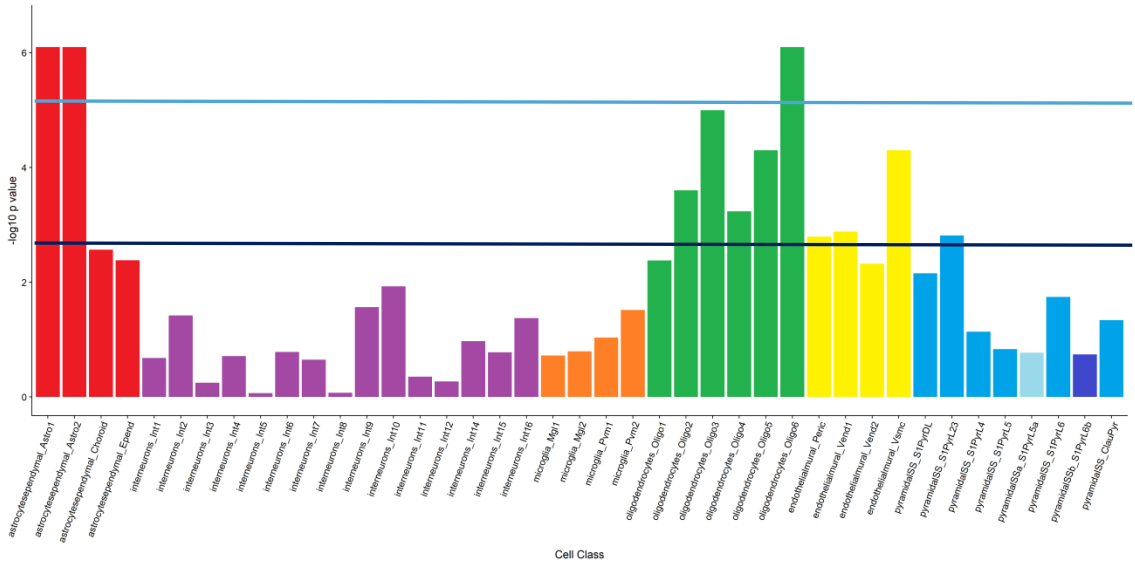


Figure 82. Results of EWCE for list of human iPSC-derived neuron differentially expressed genes in t(1;11) cells compared to cortical dataset. Cell subclasses are coloured by class; Red=Astrocyte/Ependymal, Purple=Interneuron, Orange=Microglia, Green=Oligodendrocyte, Yeloow=Endothelial/Mural, Blue=Pyramidal Neuron (all classes). The dark blue line indicates the threshold for significance , Bonferroni corrected p value<0.05. The light blue line indicates p values of <1x10^{-5} as there were no background lists with as much specificity as the differentially expressed gene list. Results above this line cannot be graphed on a log scale as the –log10 of 0 is not defined.

The changes in Oligodendrocyte expressed genes are also interesting. Zeisel *et al.* hypothesised that the 6 oligodendrocyte subclasses represented different stages in oligodendrocyte maturation. Oligo1 does not express the genes associated with oligodendrocyte precursor cells, so they hypothesised it is the first post-mitotic state for oligodendrocytes. Figure 83 displays the differential expression of the Oligo subclasses according to Zeisel *et al.*. However the significance appears to be driven by genes expressed in all oligodendrocyte subclasses, rather than by genes which are highly significant in a single cell type. Genes such as these typically have a low SI in several subclasses and a high SI in a single class. The top two markers for the 5 significant subclasses are *ATP8B1*, *GNA12* (Oligo2, SIs 0.2, 0.19), *GSTP1*, *METRN* (Oligo3, SIs 0.07, 0.08), *HGBZ*, *COL11A2* (Oligo4, SIs 0.21, 0.18), *S100B*, *APOD* (Oligo5 SIs 0.32, 0.23), *CAR2*, *APOD* (Oligo6, SIs 0.18, 0.16). Overall these are low scores, and the presence of *APOD* twice offers the clue that markers driving significance are oligodendrocyte-wide as opposed to subclass specific. Oligodendrocyte-wide markers which are differentially expressed include *DCT* (SI=0.87), *CNP* (SI=0.83), *GPR37* (SI=0.76), *SEMA3D* (SI=0.75), *SLC44A1*

(SI=0.75), *CAR2* (SI=0.7), *APOD* (SI=0.68). As can be seen in Figure 83, most of the subclass markers show non-zero expression in other Oligodendrocyte cell types too. *DCT* and *CNP* show similar oligodendrocyte class specificity to classic genes such as *MBP*, *MOG*, and *MOPB*, which have SIs in the mid-80s, for the broad Oligodendrocyte class but are not differentially expressed.

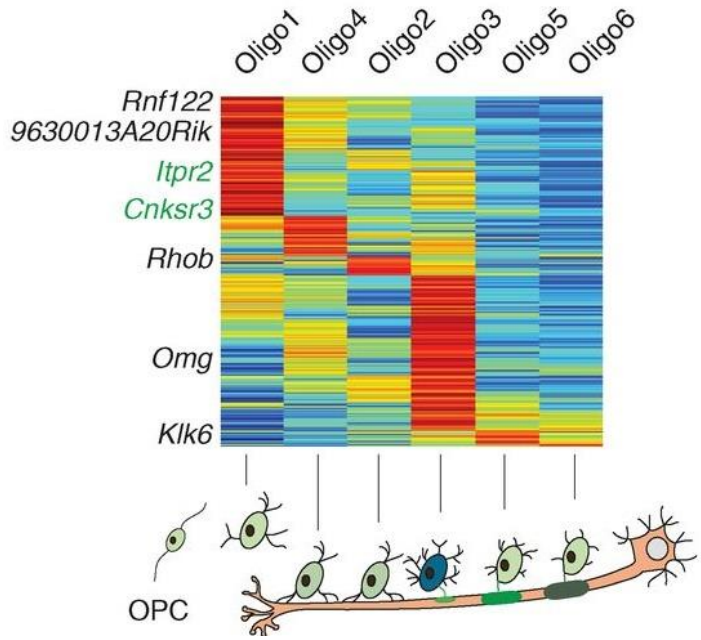


Figure 83. Adapted from Zeisel *et al.* 2015 figure 3[157]. A heatmap of genes showing differential and progressive expression across the Oligo1 to Oligo6 classes. Red=high expression, blue=low (scale not given by Zeisel *et al.*).

In addition, Astro1 and Astro2 subclasses are noted as having enrichment. These subclasses were distinguished by Zeisel *et al.* by differential expression of a number of markers, and also show different localisation, as displayed in Figure 84.

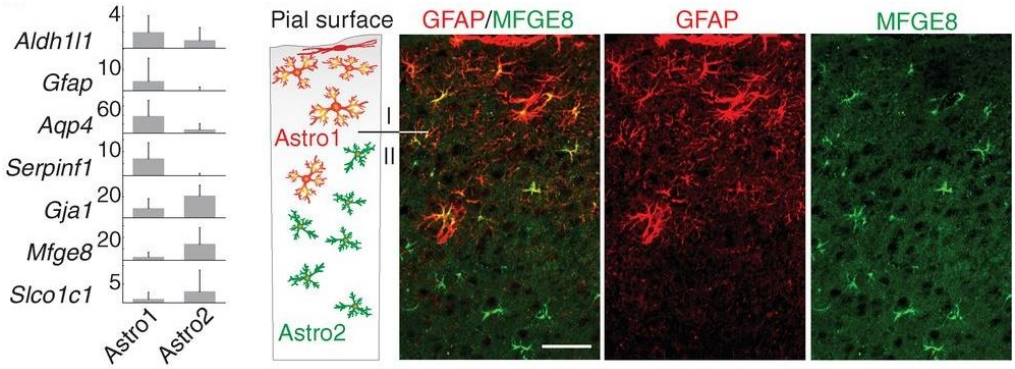


Figure 84. Adapted from Zeisel *et al.* 2015 figure 3[157]. Left is differential expression of a number of cell markers for Astro1 and Astro2. Right shows immunohistochemistry for glial fibrillary acidic protein (red, Astro1) and MFGE8 (green, Astro2), heavily associated with Astro1 and Astro2 respectively.

7.3.1 Gene ontology analysis

I also carried out a gene ontology analysis to see if the genes associated with particular cell types highlighted by the EWCE analysis were associated with specific functions, components, or processes. I used GOrilla to carry out this analysis. Gene lists were retrieved by filtering the list of KI Superset genes for those differentially expressed in t(1;11) iPSC-derived neurons. For each significant cell class, the list of genes which have their maximum expression in that class were extracted and used for GOrilla. The approach has some drawbacks in that some genes contribute to multiple cell classes but as they can only have their maximum expression in one line, they only appear in one. The analysis is therefore less than optimal. However a flat approach of using a certain number of genes will not work due to the differing number of specific genes per cell line.

The concept was to see if the differentially expressed genes known to be highly specific to certain cell types converged on any distinctive processes in those cells. Although the functions of the cell types are known, it is possible that disturbances caused by t(1;11)/*Der1* affect only a subset of cell activities. This could lead to potential pathways or functions to investigate in future experiments. Hook *et al.* had shown, for example, that catecholamine release appeared to be highly abnormal in iPSC-derived neurons of patients with schizophrenia[137]. I hoped to indicate if particular pathways in particular cell types might be disturbed. It is beneficial to know not only what cell types differentially expressed genes are specific to, but also

what activities within those cell types they are involved in. Although the approach is inferior to actually culturing those cell types and assumes that the genes are relevant solely to those cells they are most specific to, it is informative particularly if converges are observed.

The PyrSS class has three significant Processes including “negative regulation of gonadotropin secretion”, “opioid receptor signaling pathway” driven by genes *INHBA, SIGMAR1*, *OPRK1*, and a single significant related Function “opioid receptor activity”. The only significant component is “synapse part” with p value 3.9×10^{-4} driven by genes *KCNIP3, ERC2, IGSF21, CAP2, DDN, WFS1, SLITRK1, SIGMAR1*. However in all these cases the FDRs are quite high, ranging from 0.74 to 1, due to the relatively low number of 46 genes. Nevertheless this might indicate some kind of opioid related dysfunction in the t(1;11) neurons. There also appear to some metabolic dysfunctions in other cell types. The AstrocyteEpendymal class has significance for the terms “fatty acid beta-oxidation”, $p=1.62 \times 10^{-4}$, and “neurotransmitter metabolic process”, $p=7.33 \times 10^{-4}$, which has genes such as *GLDC, NQO1, SLC1A3*. Oligodendrocytes also appear particularly affected, with 28 Process terms significant. 13 of these contain the word “metabolic” and 8 the word “biosynthetic”. Oligodendrocytes are of course responsible for myelination of axonal sheaths. Although myelination is not a significant GO term it is possible that some of the metabolic GO terms relate to dysfunctions either in the biosynthetic process or in the cells in general, which would impair their ability to carry out this function.

7.4 Mouse cortical data

7.4.1 Mouse cortex Heterozygous

A number of classes were found significant for enrichment in the list of genes differentially expressed in the *Der1* heterozygous cortex, as shown in Figure 85. The sample list contained 2,112 genes. Given that both *Avp* and *Oxt* were found differentially expressed, it is no surprise that their corresponding neuron class of "Oxytocin and Vasopressin Expressing Neurons" is implicated. *Cxcl14*, *Sema3c*, are genes highly expressed in the Interneuron class, while the Hypothalamic Dopaminergic Neurons class is of particular relevance to schizophrenia and is associated with the genes *Hap1*, *Gabrq*, *Pomc*.

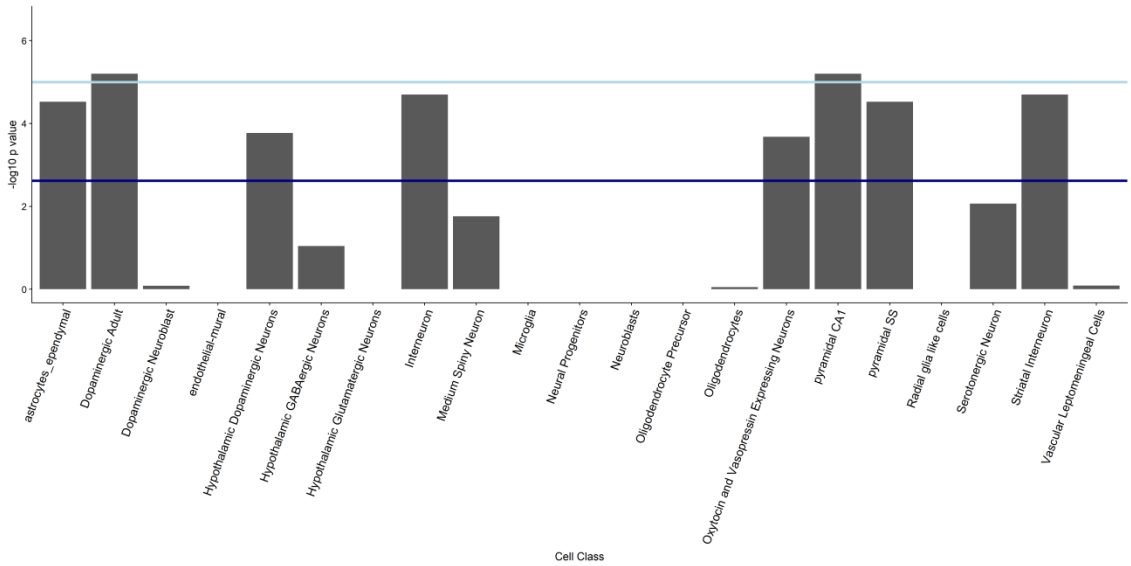


Figure 85. Results of EWCE for list of differentially expressed genes in *Der1* heterozygous cortex compared to WT cortical dataset. The dark blue line indicates the threshold for significance using the KI superset analysis, Bonferroni corrected p value<0.05. All of the indicated cell types are also significant if the hippocampal-specific classes are removed (except pyramidal CA1 which is removed as a class). The light blue line indicates p values of <1x10^{-5} as there were no background lists with as much specificity as the differentially expressed gene list. Results above this line cannot be graphed on a log scale as the –log10 of 0 is not defined.

The subclass results are displayed in Figure 86.

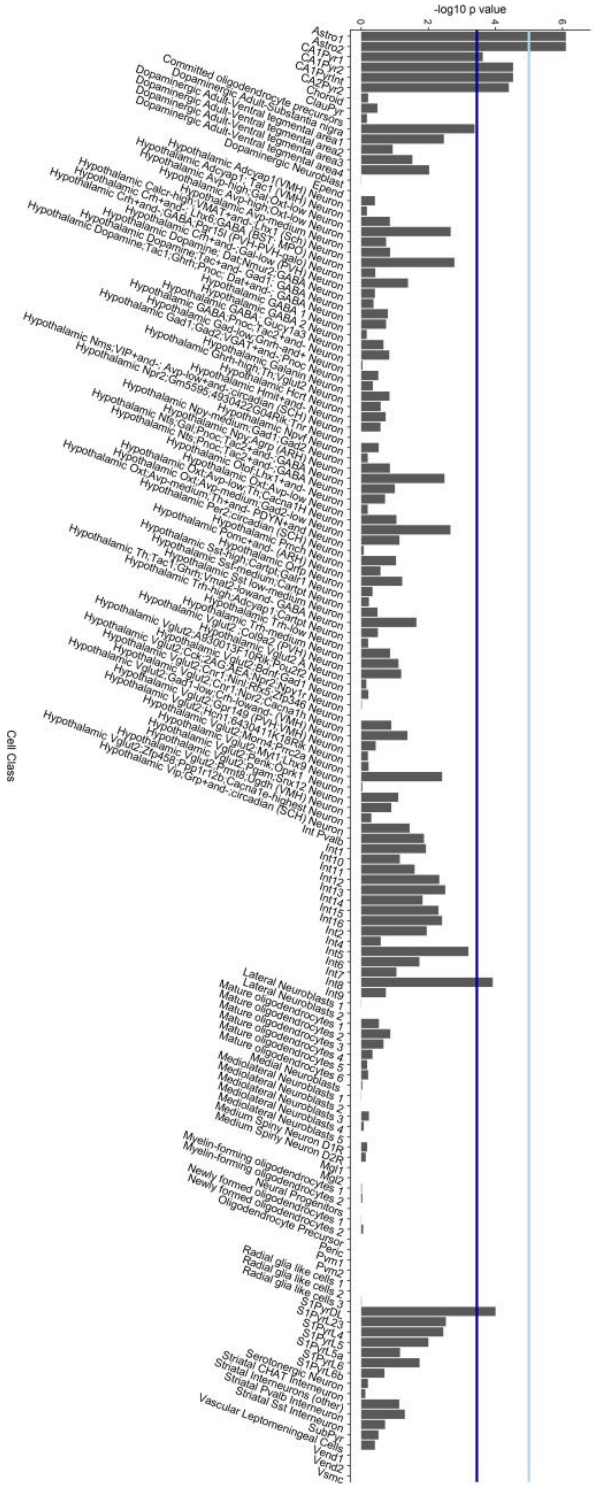


Figure 86. Results of EWCE for list of differentially expressed genes in Der1 heterozygous cortex compared to WT cortical dataset at subclass level. The dark blue line indicates the threshold for significance using the KI superset analysis, Bonferroni corrected p value<0.05. Note it is higher than in class analysis as there are more cell types to correct for. The light blue line indicates p values of $<1x10^{-5}$ as there were no background lists with as much specificity as the differentially expressed gene list. Results above this line cannot be graphed on a log scale as the –log10 of 0 is not defined.

7.4.1.1 Gene ontology analysis

I carried out a gene ontology analysis as above on the classes. Some of the results were predictable; the "Oxytocin and Vasopressin expressing Neuron" class had processes all of relevance to the functions of these proteins and containing both *Oxt* and *Avp*. Others were more novel. "Hypothalamic dopaminergic neurons" and "Striatal Interneurons" had processes linked to the GABAergic signalling pathway (two subunits of $GABA_AR$) and butyrate metabolism (only one gene, *Acads*) respectively. It should be noted that due to the low number of genes in the above processes, the FDRs are quite high. "Dopaminergic Adult" also has terms of high relevance to the cell type, indicating dysfunction of this specialised cell type as well.

The "AstrocyteEpendymal" class has some differences and some similarities compared to the human t(1;11) GO terms of the same class. Many of the terms relate to fatty acid metabolism, and transport also emerges as a theme amongst less significant GO terms. "Regulation of transport" has p value=$1.3x10^{-4}$ and contains the transport genes *Slc1a2, Slc9a3rl, Slc38a3*, and the gene *Slc25a18* is also dysregulated. The "PyrSS" class shows a very clear convergence on terms relating to protein trafficking and localisation within the cell, with genes such as *Dlg4* (related to the gene encoding PSD-95) and *Bsn* (a synaptic release protein) altered. *Cacnb3* and *Cacn*g*3* were also altered.

7.4.2 Mouse cortex Group One

No cell types were associated with the list of genes differentially expressed between mouse homozygous cortical and WT samples. However, as discussed in the appropriate chapter, the homozygous cortical *Der1* samples separate into two groups. I examined the associations with the list of genes differentially expressed in each of these groups. The results of the Group One analysis are shown in Figure 87, while the results for the Group Two analysis are shown in Figure 89.

The class level results are shown in Figure 87. The sample list contained 692 genes. The strongest signal is coming from the "Medium Spiny Neuron" and "pyramidal CA1" classes. A number of interesting genes are differentially expressed and associated with the "Medium Spiny Neuron", including *Drd1*. Also associated and

differentially expressed are a large number of phosphodiesterases associated with signal transduction; *Pde7b, Pde1b, Pde10a*, which range in specificity from 0.45 to 0.67. *Gpr88* and *Gpr6* are also seen; the former appears to control Medium Spiny Neuron firing, with knockout mice having decreased GABAergic and increased glutamatergic signalling efficiency. Interestingly, this knockout mouse had differential expression of Rgs4 protein, a possible regulator of synaptic plasticity[282]. *Rgs4* maximum specificity across the 24 classes is 0.29 in "Pyramidal SS" neurons, the second highest is 0.17 in "Medium Spiny Neurons". It too is differentially expressed here.

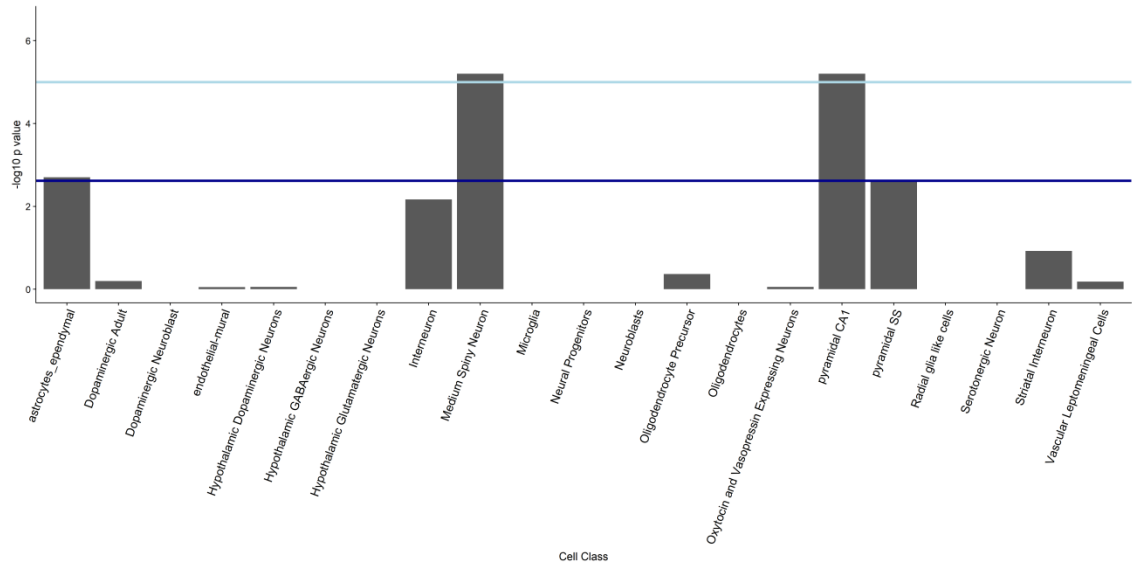


Figure 87. Results of EWCE for list of differentially expressed genes in Der1 homozygous cortex Group One compared to WT cortical dataset. The dark blue line indicates the threshold for significance using KI superset analysis, Bonferroni corrected p value<0.05. All of the above cell types are also significant if the hippocampal-specific classes are removed (except pyramidal CA1 which is removed as a class). The light blue line indicates p values of $<1x10^{-5}$ as there were no background lists with as much specificity as the differentially expressed gene list. Results above this line cannot be graphed on a log scale as the –log10 of 0 is not defined. Note that pyramidal SS is significant.

Both subclasses of the "medium spiny neuron" class are significant as shown in Figure 88, which displays the results of the subclass analysis.

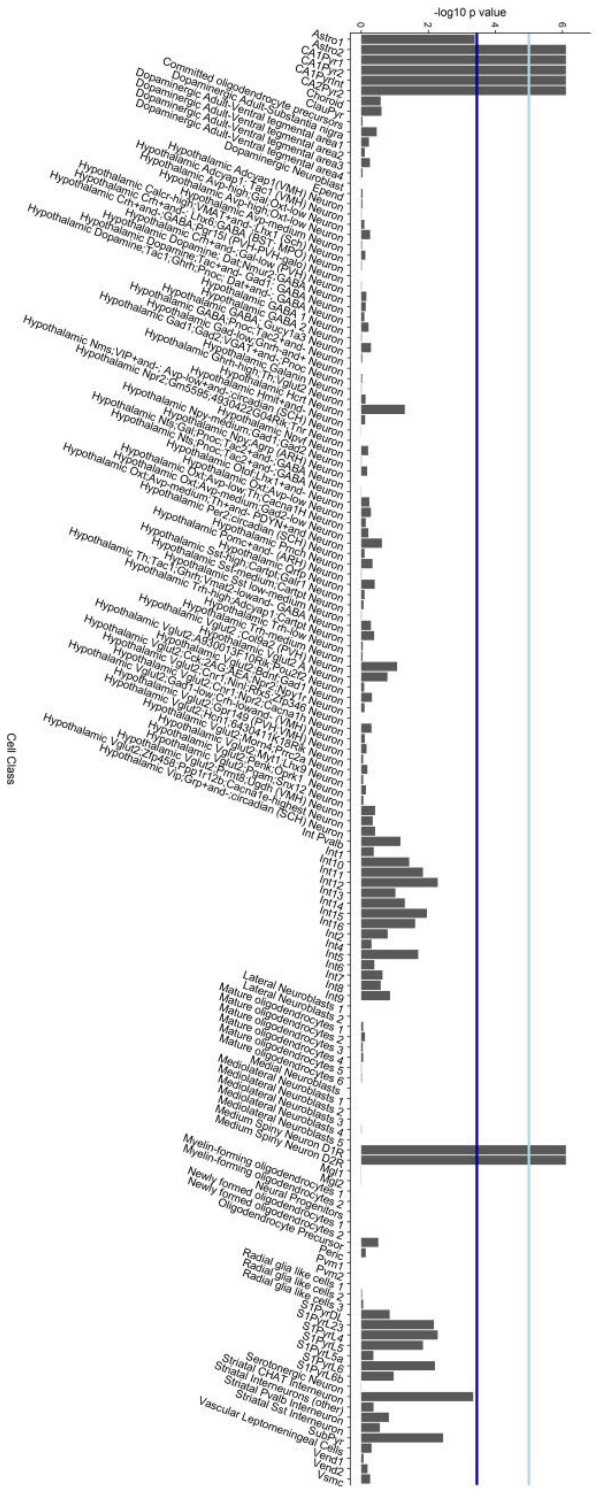


Figure 88. Results of EWCE for list of differentially expressed genes in Der1 homozygous cortex Group One compared to WT cortical dataset at subclass level. The dark blue line indicates the threshold for significance using the KI superset analysis, Bonferroni corrected p value<0.05. Note it is higher as there are more cell types to correct for. The light blue line indicates p values of $<1x10^{-5}$ as there were no background lists with as much specificity as the differentially expressed gene list. Results above this line cannot be graphed on a log scale as the –log10 of 0 is not defined.

These two subclasses correspond to D1R and D2R, but many of the genes are reasonably expressed in both cell subclasses with a few diverging genes like *Drd1* and *Adora2a* showing preferences for each subclass. The Astro2, CA1Pyr1, CA1Pyr2, CA1PyrInt, and CA2Pyr2 subclasses are also significantly enriched. This is probably due to similar processes being highlighted between the CA1 pyramidal neurons and the SS pyramidal neurons.

7.4.2.1 Gene ontology analysis

I carried out a gene ontology analysis as above on the classes. The "AstrocyteEpendymal" class has some differences and some similarities compared to the human t(1;11) GO terms of the same class. Notably, "Transport" is the top significant term at $p=5.25x10^{-6}$ with genes such as *Apoe*, *Sdc4* (syndecan-4, a heparan sulfate proteoglycan), and *Dbi*, a lipid metaboliser which acts on Diazepam/Valium. Fatty acid metabolism also appears, with *Cpt2*, the gene encoding carnitine palmitoyltransferase II, differentially expressed. The human homologue is also differentially expressed and associated with a similar GO term. The two genes *Apoe, Agt* (angiotensin), both relate to the GO term "cholesterol esterification". Finally, the PyrSS class has many terms driven by potassium and calcium receptor subunits (*Cacnb4, Kcnb1, Kcns2, Kcna1*)., and the ATPases which power them. The 12^{th} term in this cell subclass is "transport", $p=6.23x10^{-4}$, which includes not only these metal ion transporters but also a kinesin *Kif3c* and the neurotransmitter transporter *Slc6a17*. The human homologue of this gene is found at dendritic spines[283]. PyrSS also includes the *Arc* gene.

7.4.3 Mouse cortex Group Two

The class results are displayed in Figure 89. The sample list contained 2,619 genes. Many classes appear significant in both Group One and Group Two. Many of the genes driving the Medium Spiny Neuron significance are the same as in Group One, namely *Adora2a, Gpr88, Scn4b, Nexn, Actn2, Pde10a, Drd1*. This observation led me to carry out an analysis using just the overlapping genes, which is described later. Astrocyte_Ependymal significance is notably due to *Gfap*, the marker for these cell types.

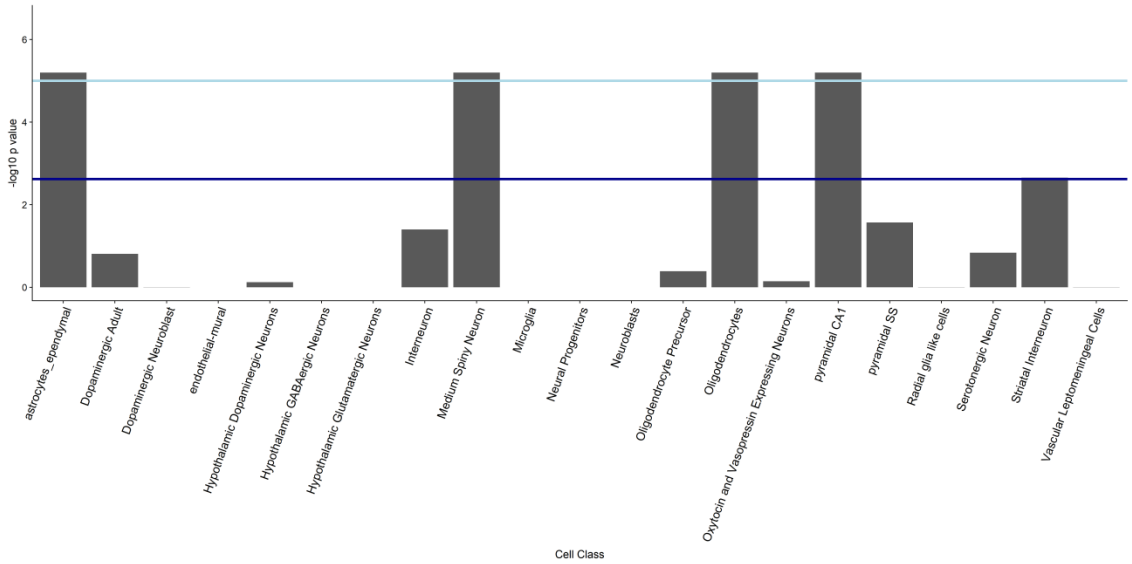


Figure 89. Results of EWCE for list of differentially expressed genes in Der1 homozygous cortex Group Two compared to WT cortical dataset. The dark blue line indicates the threshold for significance the KI superset analysis., Bonferroni corrected p value<0.05. All of the above cell types are also significant if the hippocampal-specific classes are removed, as is the class "Striatal Interneruon", except pyramidal CA1 which is removed as a class. The light blue line indicates p values of <1x10^{-5} as there were no background lists with as much specificity as the differentially expressed gene list. Results above this line cannot be graphed on a log scale as the –log10 of 0 is not defined. Striatal Interneuron is significant.

Subclass results are shown in Figure 90. In addition to the subclasses enriched in Group One, Group Two also has the Astro1 subclass significantly enriched, along with 8 types of oligodendrocyte.

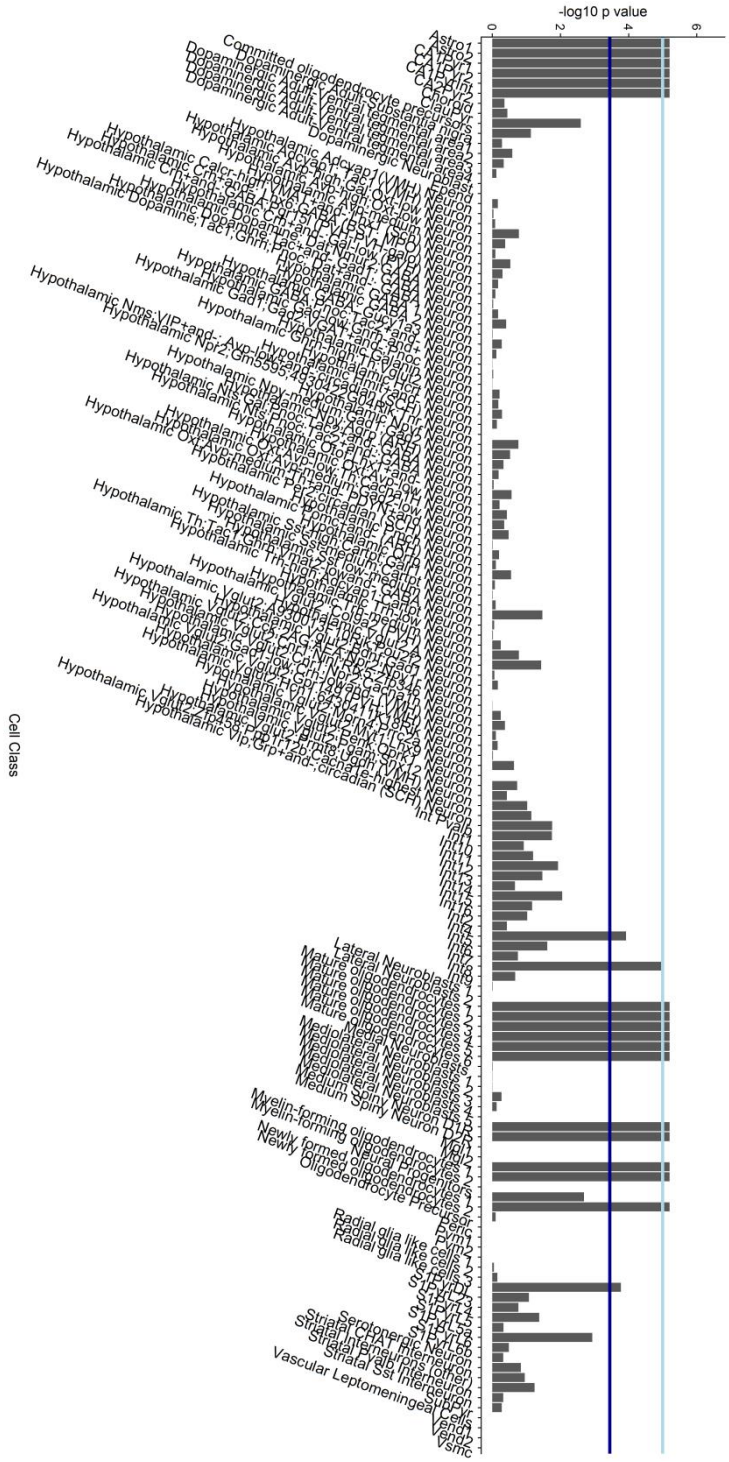


Figure 90. Results of EWCE for list of differentially expressed genes in Der1 homozygous cortex Group Two compared to WT cortical dataset at subclass level. The dark blue line indicates the threshold for significance using the KI superset analysis, Bonferroni corrected p value<0.05. Note it is higher as there are more cell types to correct for. The light blue line indicates p values of $<1x10^{-5}$ as there were no background lists with as much specificity as the differentially expressed gene list. Results above this line cannot be graphed on a log scale as the –log10 of 0 is not defined.

Significant oligodendrocytes include 6 types of "Mature Oligodendrocyte" and two types of "Myelin-forming Oligodendrocyte". This appears to be due to the classic markers *Mog, Mobp,* with the appearance of the genes *Selenoi, Selenop, Selenok* as well. These selenium related proteins are not well studied, but one paper has shown that male mice deficient in Selenop and in another selenium metabolism protein Scly show neurodegeneration, apparently due to a failure of GABAergic inhibition development[284]. This shows some similarities to the observed differences in GABAergic-maturation related genes described elsewhere. As in the Group One comparison, the subclasses Astro2, CA1Pyr1, CA1Pyr2, CA1PyrInt, and CA2Pyr2 subclasses are also significantly enriched along with both "Medium Spiny Neuron" subclasses.

7.4.3.1 Gene ontology analysis

I carried out a gene ontology analysis as above on the classes. The same classes are involved as previously; however, the number of involved genes is far greater and the GO terms have correspondingly smaller associated p-values. The terms themselves bear similarities in the "AstrocyteEpendymal" class, and many of the genes overlap.

However, other classes have GO terms diverging from those they have in other sample sets. "Striatal Interneuron" has the GO term "response to estradiol" with genes *Pam, Ramp3, Socs2,* a contrast to the related terms in the heterozygous cortex. The activation of estradiol receptors appears to be necessary for long-term potentiation in striatal interneurons, so despite the low number of associated genes this is a highly intersting finding[285].

The "PyrCA1" class also has entirely new GO terms, relating to metabolism, particularly that of RNA, and gene expression. 116 genes relate to gene expression. Two are bromodomain proteins, while the previously mentioned *Ntrk3* appears too, as does *Bbs7* in the RNA metabolism GO term. Camk2a and Dendrin are also encoded by differentially expressed genes related to this cell type, and they have roles in plasticity.

"Oligodendrocytes" appears as a significant class, in agreement with the human (t1:11). Unlike there however, its GO terms are highly specific to the function of myelination. Dysregulation of *Mag* (myelin-associated protein), *Mbp* (myelin basic protein), *Plp1* (a type of myelin protein) and a claudin *Cldn11* associated with myelin all point towards a serious defect in the myelinating propetries of these cells.

7.4.4 Mouse cortex overlapping Group One and Group Two

This list of genes is those differentially expressed in both groups. As described previously, they show an unusual pattern in expression. The sample list contained 249 genes. The class results are shown in Figure 91, while subclass results are in Figure **92**.

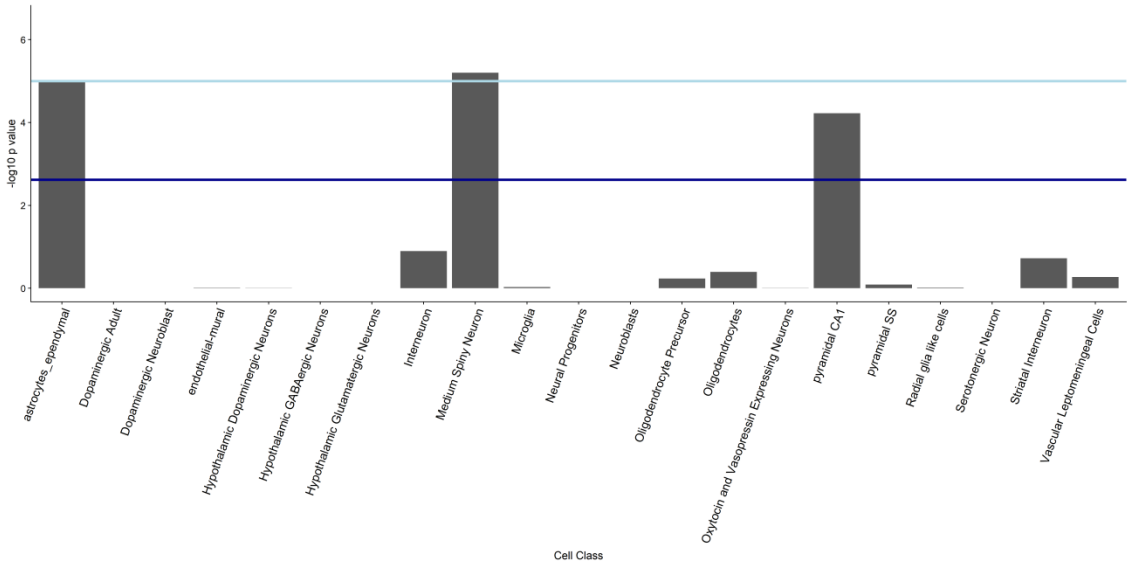


Figure 91. Results of EWCE for list of differentially expressed genes overlapping between Group One and Group Two of homozygous Der1 cortex compared to WT cortical dataset. The dark blue line indicates the threshold for significance using the KI superset analysis, Bonferroni corrected p value<0.05. All of the above cell types are also significant if the hippocampal-specific classes are removed (except pyramidal CA1 which is removed as a class). The light blue line indicates p values of $<1x10^{-5}$ as there were no background lists with as much specificity as the differentially expressed gene list. Results above this line cannot be graphed on a log scale as the –log10 of 0 is not defined.

In addition to the subclasses enriched in Group One, the list overlapping with Group Two also has the Astro1 subclass significantly enriched. As described in the previous section, a large number of genes show reasonable expression in both Medium Spiny Neuron subclasses, with further genes being primarily expressed in one.

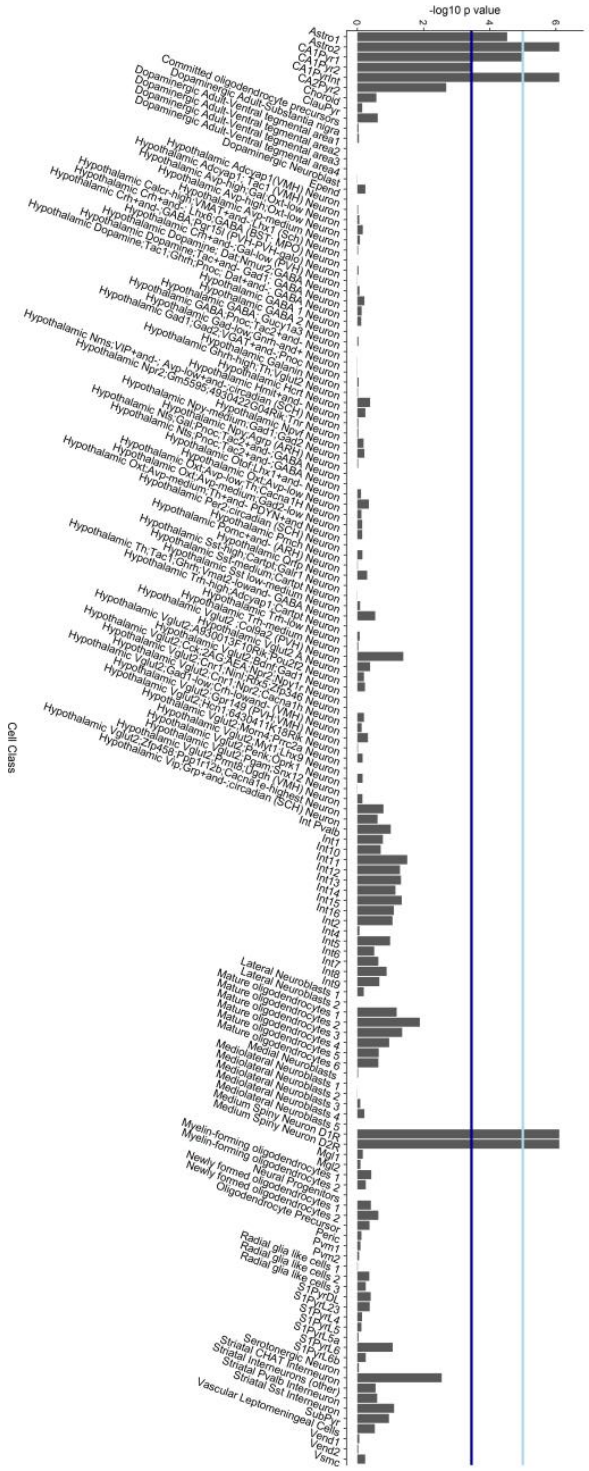


Figure 92**. Results of EWCE for list of differentially expressed genes in both Der1 homozygous cortex Group One and Group Two, when each is separately compared to WT cortical dataset at subclass level. The dark blue line indicates the threshold for significance using the KI superset analysis, Bonferroni corrected p value<0.05. The light blue line indicates p values of <1x10^{-5} as there were no background lists with as much specificity as the differentially expressed gene list. Results above this line cannot be graphed on a log scale as the –log10 of 0 is not defined.**

7.4.5 Mouse cortex overlapping Group One, Group Two, and cortex heterozygotes

This list is a subset of the previous; also overlapping with the cortex heterozygous list. The sample list contained 120 genes. Class results are in Figure 93, while subclass results are in Figure 94. Only one class is significantly enriched in the subset of genes differentially expressed in all models; "Astrocytes_ependymal". Genes highly enriched in this class include *Apoe, S100a1, Mlc1, Fxyd1, Slc25a18*. At the subclass level, both subclasses of Astrocyte are significantly associated with the gene list, with the aforementioned genes showing good expression in both classes. In addition, CA1 pyramidal neurons have emerged as significant, although given that the SS neurons are not also altered this may not point to identical pyramidal dysfunction across all *Der1* groups.

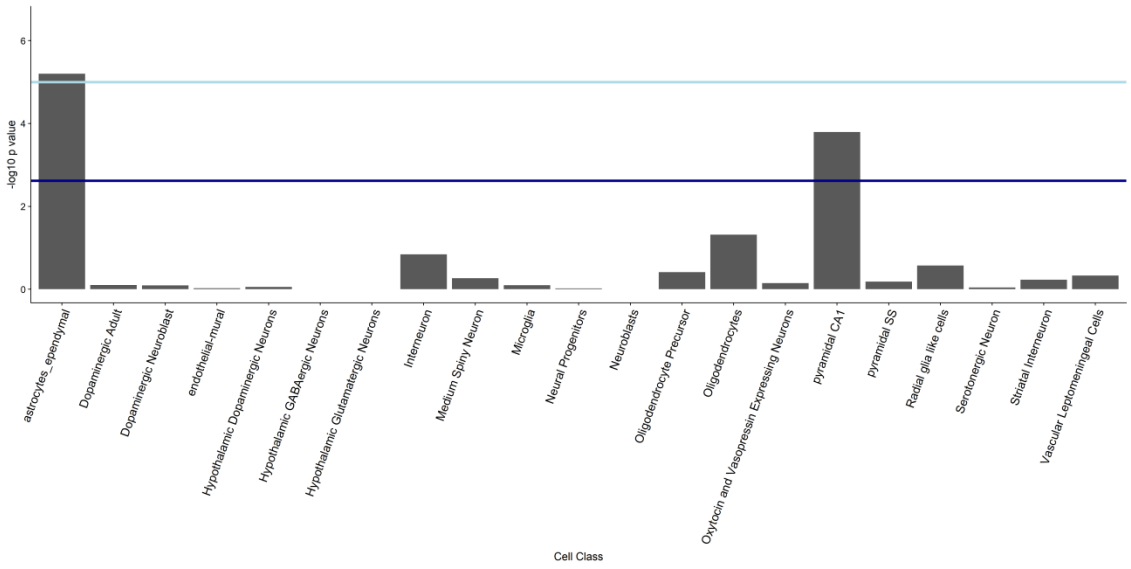


Figure 93. Results of EWCE for list of differentially expressed genes overlapping between Group One and Group Two of homozygous Der1 cortex, as well as with heterozygous Der1 cortex compared to WT cortical dataset. The dark blue line indicates the threshold for significance the KI superset analysis., Bonferroni corrected p value<0.05. The light blue line indicates p values of <1x10^{-5} as there were no background lists with as much specificity as the differentially expressed gene list. Results above this line cannot be graphed on a log scale as the –log10 of 0 is not defined.

Less subclasses are significant in this analysis; with only Astro1, Astro2, and CA1PyrInt appearing significant. This makes Astro1 and Astro2 among the most consistent subclasses to have enriched genes overrepresented among differentially expressed genes.

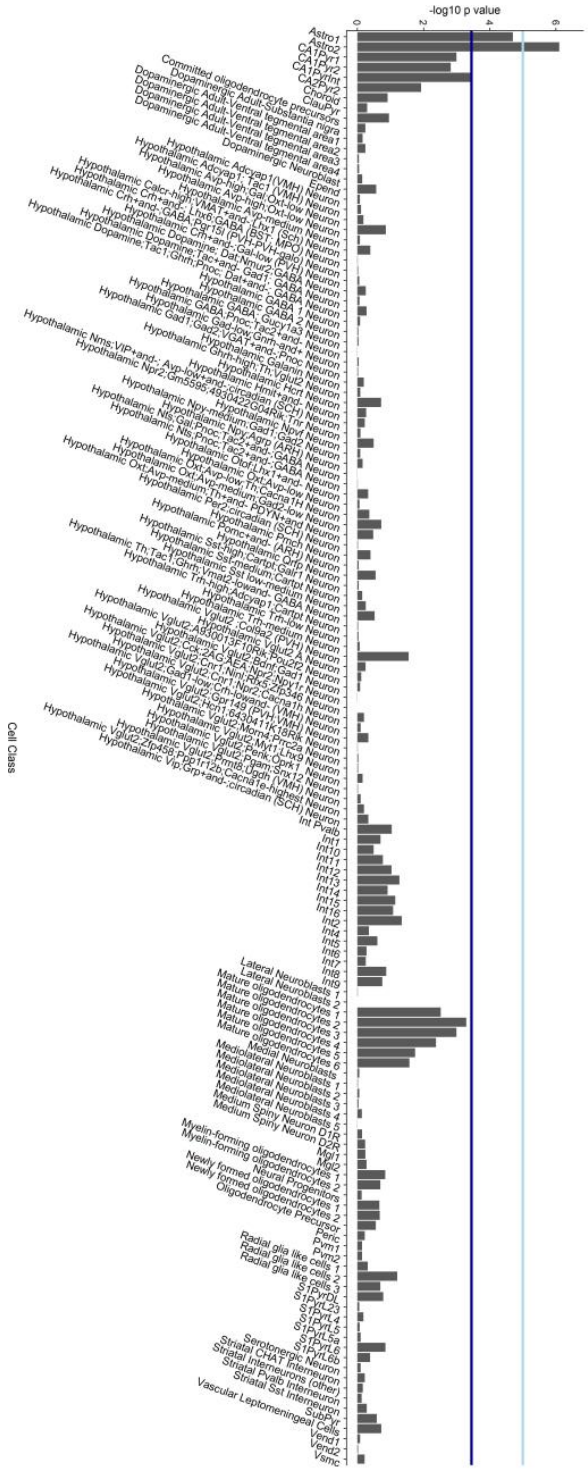


Figure 94. Results of EWCE for list of differentially expressed genes overlapping between Group One, Group Two, and Cortex Heterozygotes when each is compared to WT cortical dataset at subclass level. The dark blue line indicates the threshold for significance using the KI superset analysis, Bonferroni corrected p value<0.05. Note it is higher as there are more cell types to correct for. The light blue line indicates p values of <1x10^{-5} as there were no background lists with as much specificity as the differentially expressed gene list. Results above this line cannot be graphed on a log scale as the –log10 of 0 is not defined.

7.5 Mouse hippocampal data

The number of differentially expressed genes between mouse heterozygous t(1;11) carriers and wild type controls is 175. The results of the class analysis are displayed in Figure 95.

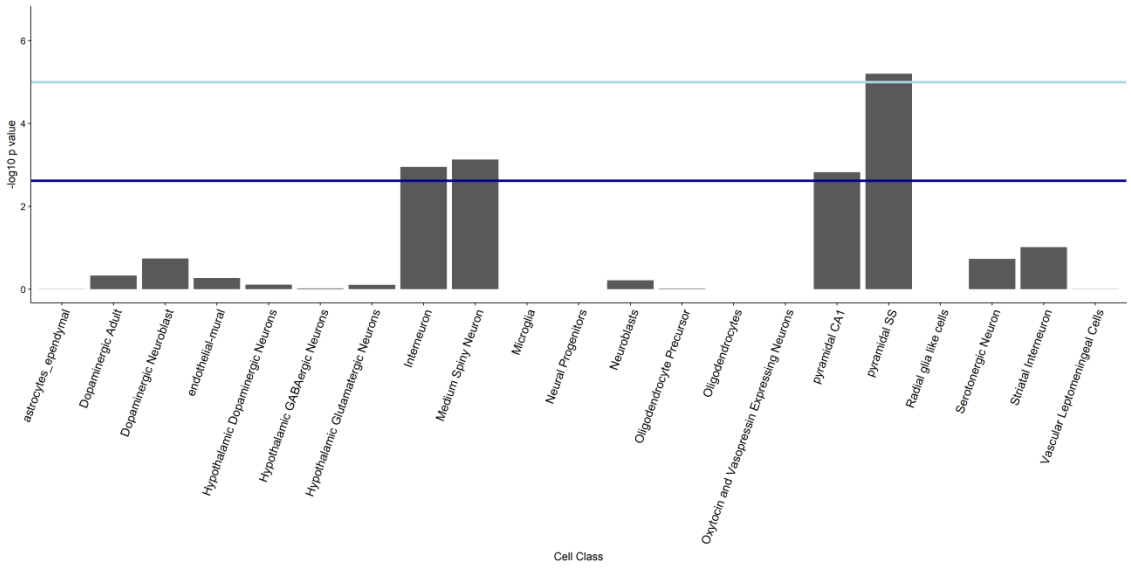


Figure 95. Results of EWCE for list of mouse hippocampal differentially expressed genes compared to KI superset. The dark blue line indicates the threshold for significance using the KI superset analysis, Bonferroni corrected p value<0.05. The light blue line indicates p values of $<1x10^{-5}$ as there were no background lists with as much specificity as the differentially expressed gene list. Results above this line cannot be graphed on a log scale as the –log10 of 0 is not defined.

The same classes appear to be affected as in the cortex; “Medium Spiny Neurons” and pyramidal cells. In addition, “Interneuron” has emerged as a significant class. The subclass results are displayed in Figure 96. Only a single subclass, S1PyrL5a, is significant.

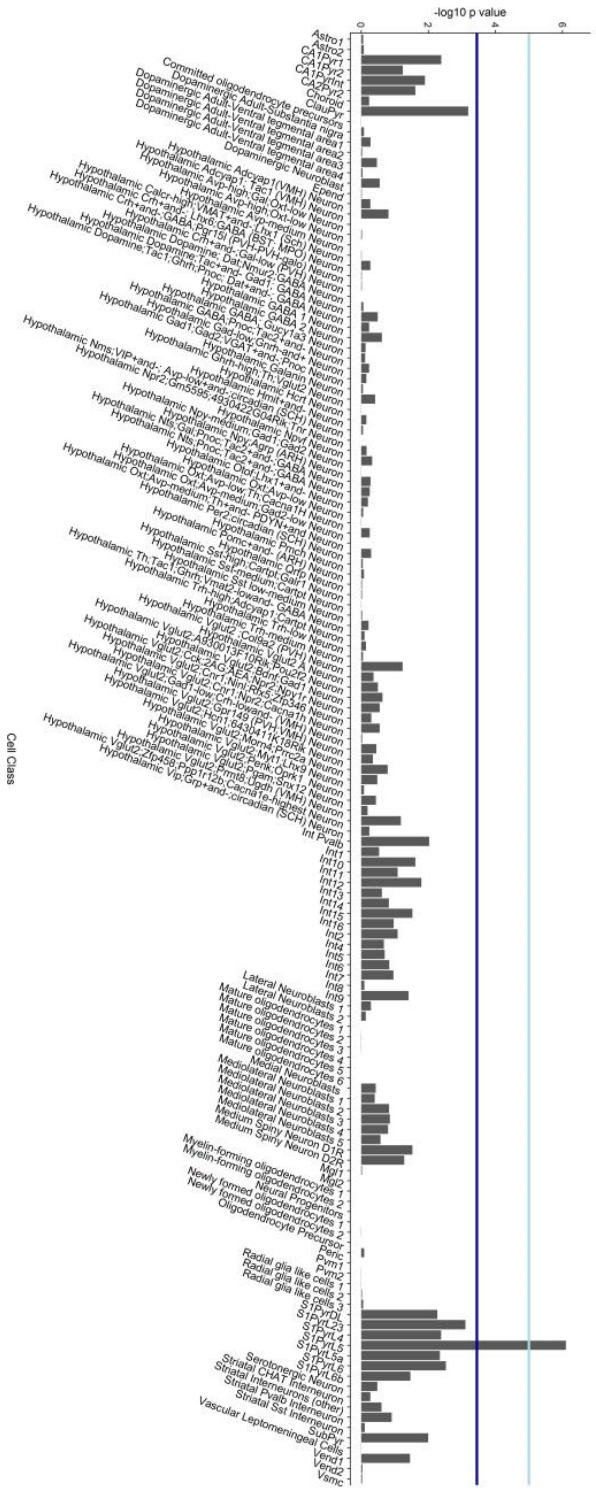


Figure 96. Results of EWCE for list of differentially expressed genes in Der1 heterozygous hippocampus compared to WT hippocampal dataset at subclass level. The dark blue line indicates the threshold for significance using the KI superset analysis, Bonferroni corrected p value<0.05. Note it is higher as there are more cell types to correct for. The light blue line indicates p values of $<1x10^{-5}$ as there were no background lists with as much specificity as the differentially expressed gene list. Results above this line cannot be graphed on a log scale as the –log10 of 0 is not defined.

7.5.1 Gene ontology analysis

I carried out a gene ontology analysis as above on the classes. The classes have very distinctive and obvious patterns. “Medium Spiny Neuron” related GO terms are all of relation to synaptic transmission, although the number of genes involved is low, only *Syt2* and *Unc13c*. “PyrSS” has two stronger patterns, each with more related genes. The first is related to neuronal apoptosis, driven by *Thrb, Scn2a1, Cit, Bok. Cit* and *Thrb* arte also related to the second process with multiple GOterms; development. These terms are also driven by the genes *Sox5, Plxnd1* (a plexin, the interacting partner of semaphorins which aid neuronal direction) and *Cask. Cask*, as a member of the superfamily which includes *Dlg* genes, encodes a MAGUK protein. Cask-null mice are embryonic lethal and Cask-deficient neurons appear to have abnormal levels of neurexins and neuroligins (important synaptic molecules), although the experiment showing this had a relatively small sample size[286].

No cell types were associated with the list of genes differentially expressed between mouse homozygous hippocampal and WT samples. However, the number of differentially expressed genes in this analysis was extremely low, and *Der1* status did not even register as the first or second component in principal component analysis. Therefore, a lack of results is not surprising here.

7.6 Differentially expressed genes from published papers

GO term analyses were not possible as neither Wen *et al.* nor Srikanth *et al.* provided a list of all expressed genes, just lists of differentially expressed genes[132,287].

7.6.1 Wen et al.

Wen *et al.* looked at iPSC-derived neurons from a family carrying a *DISC1* frameshift mutation[132]. The family is small however, and the frameshift is not unambiguously linked with psychiatric illness, as in addition members of the pedigree have psychiatric illness but no *DISC1* frameshift. A total of 3,697 genes were differentially expressed in their study, although they did not use as large a number of RNA samples as our study (one control and two mutants, all in triplicate). A further description is given in the Introduction.

The results of this analysis are displayed in Figure 97. It is clear that genes highly associated with a broad range of cells are affected, similar to the mouse cortical analysis. In addition, both the Astrocyte subclasses have re-emerged as significant, as in the t(1;11) analysis. Three subclasses of the endothelial/mural cell class are also implicated; this class is distinguished from other cell groups chiefly by high expression of *Cldn5*. Many cell types are apparently affected.

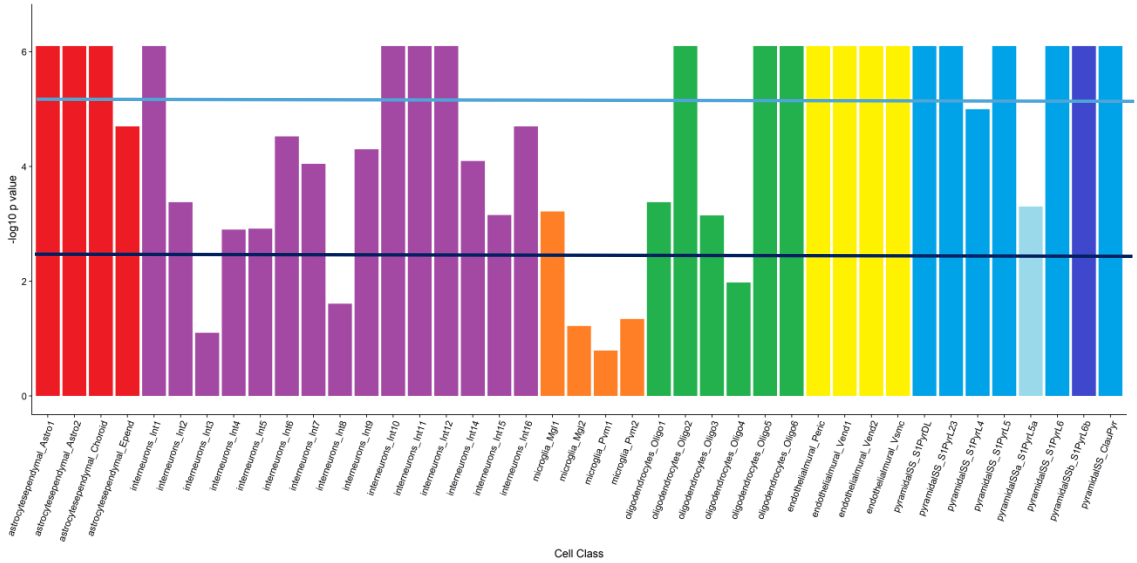


Figure 97. Results of EWCE for list of differentially expressed genes found by Wen *et al.* compared to cortical dataset. Cell subclasses are coloured by class; Red=Astrocyte/Ependymal, Purple=Interneuron, Orange=Microglia, Green=Oligodendrocyte, Yeloow=Endothelial/Mural, Blue=Pyramidal Neuron (all classes). The dark blue line indicates the threshold for significance , Bonferroni corrected p value<0.05. The light blue line indicates p values of <1x10^{-5}.

7.6.2 Brennand et al.

The study of Brennand *et al.* was the first to look at differential expression in iPSC-derived neurons from psychiatric patients[124]. Lines were established from patients with idiopathic schizophrenia, although the sample numbers were small. It utilised a microarray approach to look at expression differences and found 596 genes differentially expressed at p<0.05 and fold-change >1.3, which I searched for association with here. A further description is given in the Introduction.

The results of this analysis are displayed in Figure 98. As with the t(1;11) samples, the list of genes shows enrichment for both Astrocyte types (also significant in my analysis), and almost all oligodendrocyte maturation stages. Two Interneuron

subclasses are implicated; Int1 and Int14. Both subclasses are characterised by high expression of Neuropeptide Y, with Int14 also having high expression of *Gda* compared to other cell types. *Gda* encodes Cypin, a PSD-95 interactor with homology to DPYSL1, a protein related to DPYSL2 and DPYSL3[288]. Int14 also has relatively high expression of the chromatin modifiers *Tox* and *Tox3*, which are differentially expressed. Int1 meanwhile is characterised among the interneuron subclasses by the highest expression of *Sst*, which encodes somatostatin. This cell type is highly restricted to the somatosensory cortex. All subclasses of the endothelial/mural cell class are also implicated; these are distinguished from other cell groups chiefly by high expression of *Cldn5*. This is not a differentially expressed gene however; although genes that are include *IFI44*, *CYYR1*, *SLC16A9* and *LEF1*, the homologues of which show high expression in both "Vend1" and "Vend2", as well as pericytes in the case of *CYYR1*.

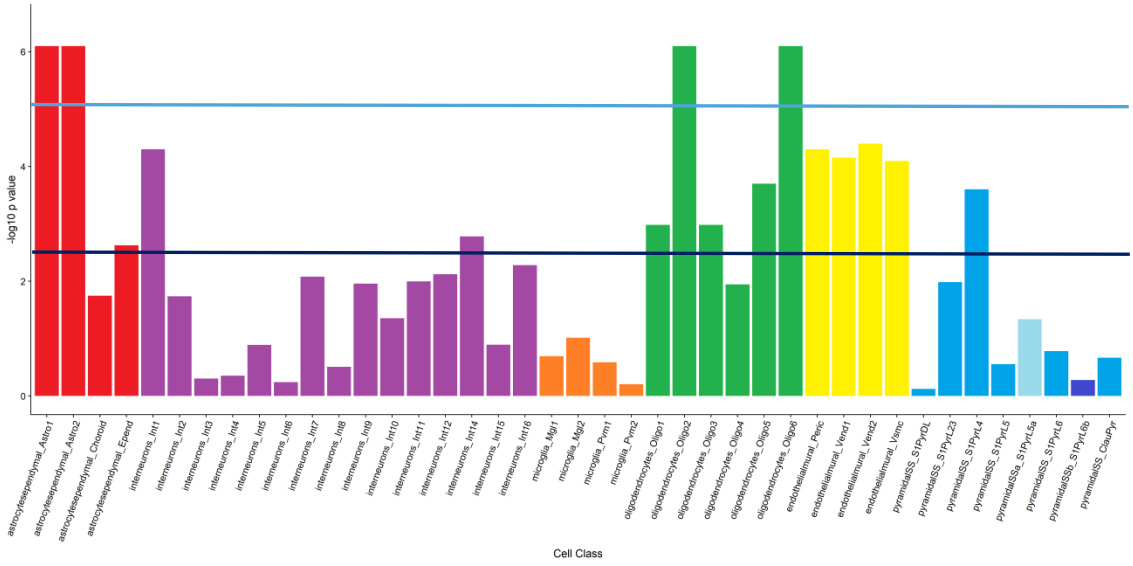


Figure 98. Results of EWCE for list of differentially expressed genes found by Brennend *et al.* compared to cortical dataset. Cell subclasses are coloured by class; Red=Astrocyte/Ependymal, Purple=Interneuron, Orange=Microglia, Green=Oligodendrocyte, Yeloow=Endothelial/Mural, Blue=Pyramidal Neuron (all classes). The dark blue line indicates the threshold for significance, Bonferroni corrected p value<0.05. The light blue line indicates p values of $<1x10^{-5}$.

Differential expression of genes associated with both subclasses of astrocytes is also observed; genes contributing to this include *Gfap* and *Aqp4*, known astrocytic genes.

7.6.3 Srikanth et al.

Srikanth *et al.* looked at two different timepoints in the production of neurons directly from iPSCs[138], a protocol that differs from ours which utilises neural precursor intermediates. They induced mutations in the iPSC lines prior to neuron differentiation, in exon 2 and exon 8 of *DISC1*. The first mutation should remove all *DISC1* isoforms while the second is closer in its effect on *DISC1* by inducing a truncation close to the breakpoint. They also looked at heterozygous and homozygous carriers of these mutations. Sx2 and Sx8 refer to the mutations, w/m to wild type/mutant status, and 18 or 50 to the timepoints. For example, Sx8wm50 refers to the heterozygous carriers of the exon 8 truncation at the day 50 stage. I used these lists to generate EWCE data. The Sx2mmd50, Sx8mmd50, and Sx8wmd50 results are displayed in turn.

7.6.3.1 Sx2mmd50

The results of this analysis are displayed in Figure 99. This mutation, which would presumably have the greatest effect on *DISC1* expression, appears to have caused the dysregulation of genes highly expressed in a very broad variety of cell types. A total of 1,393 genes were dysregulated. Most pyramidal cell subclasses are significant along with most interneuron cell subclasses, possibly indicating dysregulation of a number of broadly expressed markers for each cell class.

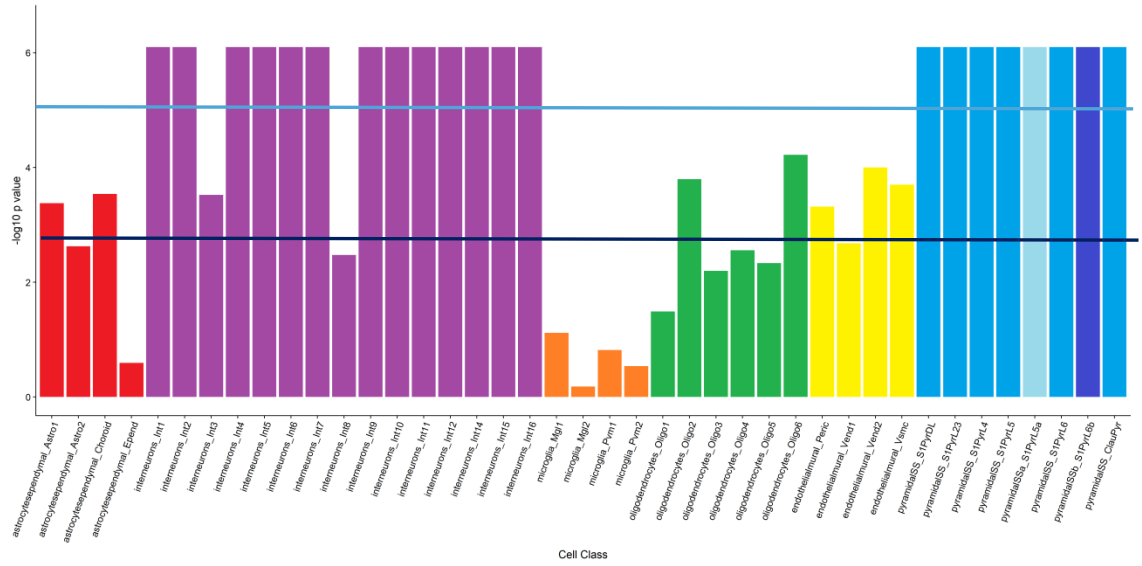


Figure 99. Results of EWCE for list of differentially expressed genes found by Srikanth *et al.* for the Sx2mmd50 model compared to cortical dataset. Cell subclasses are coloured by class; Red=Astrocyte/Ependymal, Purple=Interneuron, Orange=Microglia, Green=Oligodendrocyte, Yeloow=Endothelial/Mural, Blue=Pyramidal Neuron (all classes). The dark blue line indicates the threshold for significance, Bonferroni corrected p value<0.05. The light blue line indicates p values of <1x10^{-5}.

7.6.3.2 Sx8mmd50

The results of this analysis are displayed in Figure 100. The effects are quite interesting. The list of affected genes, totalling 116 genes, is only highly enriched in pyramidal cells, but is enriched in all of these cell subclasses. This is essentially a subset of the cell subclasses enriched in the Sx2mmd50 dataset. I looked to see if the same genes might be responsible; of the 27 genes differentially expressed in both experiments and expressed in the Zeisel dataset; a handful have high expression confined to the pyramidal neuron class. These include *Neurod2*, *Neurod6*, *Kcnip3*, which have an SI of 0.6 or more for this class. Both of the *Neurod* genes encode bHLH neurogenic transcription factors.

It is difficult to work out the contribution of these 27 overlapping genes. However, I did note that the total SI for the neuronal subclasses of the 114 genes differentially expressed in the Sx8mmd50 dataset was 21.2, while the corresponding figure for the 1,391 genes of the Sx2mmd50 dataset was 189.8. The 27 genes contribute 6.7 in each case. This corresponds to 31% of contribution for 23.6% of genes in the Sx8mmd50 dataset, and 3.5% of contribution for 1.94% of genes in the Sx2mmd50

dataset, which is not particularly strong evidence in favour of these overlapping genes being key.

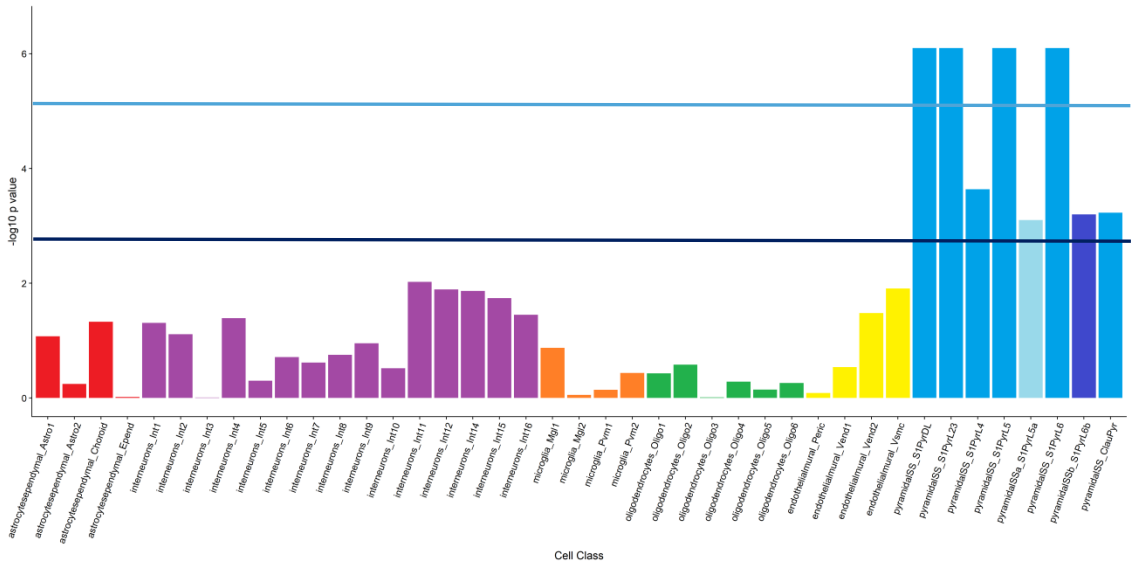


Figure 100. Results of EWCE for list of differentially expressed genes found by Srikanth *et al.* for the Sx8mmd50 model compared to cortical dataset. Cell subclasses are coloured by class; Red=Astrocyte/Ependymal, Purple=Interneuron, Orange=Microglia, Green=Oligodendrocyte, Yeloow=Endothelial/Mural, Blue=Pyramidal Neuron (all classes). The dark blue line indicates the threshold for significance, Bonferroni corrected p value<0.05. The light blue line indicates p values of <1x10^{-5}.

7.6.3.3 Sx8wmd50

The results of this analysis are displayed in Figure 101. This final Srikanth gene list is generated from a model which has the effect on full length DISC1 closest to that of the t(1;11), although it does not result in gene fusion. 487 genes were dysregulated. We can see that Astrocyte subclass 1, Oligo subclass 6 (corresponding to mature myelinating oligodendrocytes), and two endothelial cell types are among those subclasses which are enriched for genes showing differential expression.

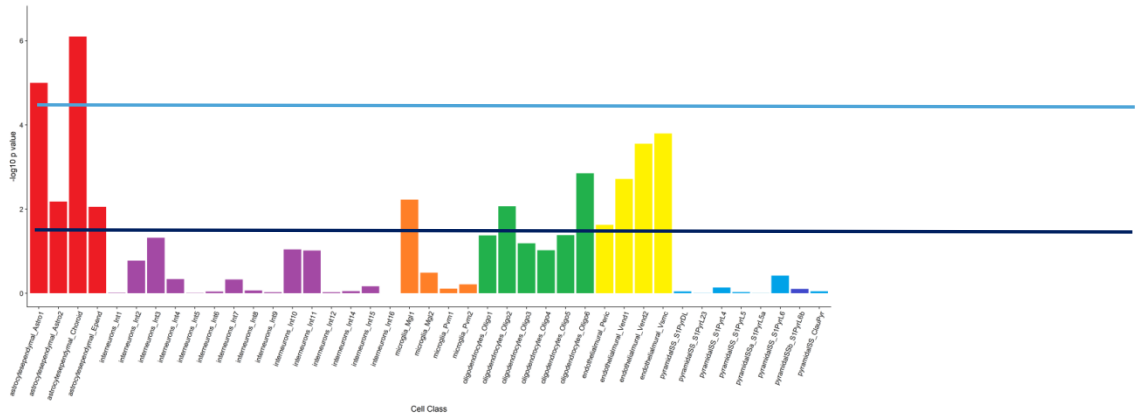


Figure 101. Results of EWCE for list of differentially expressed genes found by Srikanth *et al.* for the Sx8wmd50 model compared to cortical dataset. Cell subclasses are coloured by class; Red=Astrocyte/Ependymal, Purple=Interneuron, Orange=Microglia, Green=Oligodendrocyte, Yeloow=Endothelial/Mural, Blue=Pyramidal Neuron (both classes). The dark blue line indicates the threshold for significance, Bonferroni corrected p value<0.05. The light blue line indicates p values of <1x10^{-5}.

7.7 Discussion

7.7.1 IPSC-derived neuron data

Comparisons between EWCE data for the t(1;11) neurons and neurons carrying DISC1 mutations reveal that many of the cell types implicated by one analysis reappear in others. A full overview of the enrichments using the Zeisel *et al.* cortical dataset is given in Table 36 and a view of the overlaps between gene sets for enrichment is given in the Appendix.

	Astrocyte/Ependymal	Interneuron	Microglia	Oligodendrocyte	Endothelial/Mural	Pyramidal
t(1;11)	Astro1, Astro2			Oligo2-6	Peric, Vend1, Vsmc	Pyr2/3
Wen	All	All except Int2, Int8	Mgl1	All except Oligo4	All	All
Brennend	Astro1, Astro2	Int1, Int14		All except Oligo4	All	Pyr4
Sx2mmd50	Astro1, Choroid	All except Int8		Oligo2, Oligo6	Peric, Vend2, Vsmc	All
Sx8mmd50						All
Sx8wmd50	Astro1, Choroid			Oligo6	Vend2, Vsmc	

Table 36. Evaluation of cell types with associated genes significantly enriched among differentially expressed genes in each list according to the EWCE analysis using the Zeisel dataset. Classes are as in the Zeisel *et al.* dataset.

These cross comparisons reveal that there are substantial overlaps and some differences between the cells highlighted by the analyses of each human neuron model. The cells that most appear are Astro1, Oligo6, Vsmc (5 occurrences), Oligo2, Pyr2/3, Pyr4, (4 occurrences), and several with 3 occurrences including Astro2, Choroid, Int1, Int14, and other Pyramidal and Oligodendrocytic cell types. It is

worthwhile to note which cells did not appear in any analysis; these are the microglial subtypes Mgl2, Pvm1, and Pvm2. The disturbance of genes associated with the astrocyte subclasses Astro1 and Astro2 are consistent findings and appear to be genuine; indicating possible dysregulation of astrocytic activity in the presumably early developmental stage the iPSC-derived neurons represent. As discussed below, these also appear in several mouse cortex lists as well, so the disturbances appear to be present in both relatively immature human iPSC-neuron models and the mature mutant mouse cortex model. It is also compelling that the Srikanth homozygous *DISC1* mutation at exon 8 has effects which are a subset of those caused by the homozygous exon 2 mutation, which should affect all isoforms.

There are some conclusions to be drawn from the data and it is possible to a limited extent to discuss these changes in the context of cortical architecture. It must be stressed that as the iPSC-derived cells are highly immature and are in an essentially 2D culture it is inaccurate to speak of them as having any kind of cortical architecture. However I initially found it notable that frequent dysregulation of genes associated with Astro1 and Astro2 was found, which Zeisel *et al.* described as astrocytes associated with different sides of the layer I/II cortical boundaries. I also noted that the Pyramidal II/III subclass was frequently dysregulated as well, and thought this might be a significant co-occurrence. Yet in the Sx8wm50 neuron cultures, the Astro1 significance does not concur with a II/III significance. Similarly, these two subclasses are not both seen significant in any mouse analysis.

7.7.2 Mutant mouse data

In the *Der1* mouse cortex the cell classes “astroyctes_ependymal”, “medium spiny neuron”, and pyramidal cells are significantly enriched in every gene list. At the subclass level, the Astro1, Astro2, CA1Pyr1, CA1Pyr2, CA2Pyr2, and CA1PyrInt subclasses are all significant in the heterozygous *Der1* comparison. In addition, both subclasses of the medium spiny neuron, Astro2, CA1Pyr1, CA1Pyr2, and CA1PyrInt subclasses are also significantly enriched in both homozygote groups, as well as the list of genes overlapping between both homozygote groups and the heterozygote group. We see that the lists appear to converge on the same few cell types across the

mouse brain, perhaps indicating a consistent pathological effect of the *Der1* in the cortex. Astro1 or Astro2 significance does not co-present with layer-specific pyramidal dysfunction but only with general pyramidal cells dysfunction at the class level, so it does not appear that there has been a layer specific dysfunction as I hypothesised in the previous section.

The hippocampal data shows similarity to the cortical; with the pyramidal classes and medium spiny neurons implicated. In contrast the Interneuron class is newly significant, while there is no evidence of astrocyte_ependymal significance.

These results should be discussed in the context of the findings of my colleague, Marion Bonneau, who has studied the anatomical morphology of the Der1 cortex and hippocampus. Bonneau found that there were no gross changes in hippocampal size in heterozygous Der1 mice. There was a trend towards cortical thinness and significantly increased lateral ventricle size in the heterozygotes. However, these findings were not apparent in an MRI study which used a larger sample size. Bonneau also found that there was no difference in the staining of a particular cell type, parvalbumin positive interneurons, in the prefrontal cortex of either the heterozygous or homozygous Der1 mice. However, there was an overall increase in these interneurons in the hippocampus. These cells possibly correspond to the cell type Int2 (a hippocampal cell type) or IntPvalb; both of which highly express parvalbumin. Neither appeared significant in my mouse hippocampal heterozygous analysis; although the class of Interneuron did. The fact that some cell types have emerged as significant in the EWCE approach suggests that changes in expression of specific markers are both up and down; indicating general dysfunction rather than cellular absence. This is in agreement with Bonneau's data as well.

7.7.3 Conclusion

Dysfunction may not affect all cells equally. The persistent significance of Astrocytes is highly interesting; Astro1 is significant in every single analysis with the exception of the mouse heterozygous *Der1* hippocampus and the iPSC-derived neuron with two exon8 altered *DISC1* alleles. Astro2 is nearly as broadly disturbed. Astrocytes release neurotransmitters and have been associated with synapses for

some time; it was suggested two decades ago that they should be considered part of a "tripartite synapse"[289]. More recently, mouse studies have shown that the release of D-serine (an NMDAR co-agonist) by hippocampal astrocytes increases during wakefulness, and mice further into wakefulness are more adept at learning contextual fear memory. Given the link between NMDAR and learning/memory, the author's conclusion was that astrocytes can modulate neuronal sensitivity to memory formation[290]. The release of D-serine was mediated by α7 nicotinic acetylcholine receptors. Their disruption here is potentially important in this regard and has relevance to the phenotypes of major mental illness.

The Srikanth cell line with two mutations in Exon2 had particularly broad disruption across interneuron and pyramidal cell lines, while many other iPSC-neuron models, including our own, had pyramidal disruption. The evidence also suggests that in adult life the mouse cortex continues to exhibit some cellular abnormalities. It is possible that misplacement of developing cells, indicated by the mouse dysregulation of Hox genes including *Dlx2*, *Dlx6*, *Vax1*, *Vax2*, as well as guidance genes like *Ntn5*, *Slit1* and *Slit3*, could result in a phenotype without gross pathology, which Bonneau did not detect, but with subtle wiring and transcriptomic differences. This would make sense in the context of psychiatric disease being developmental but is a speculative explanation currently. We also saw some emergence of significance in the t(1;11) neurons of many oligodendrocyte classes, particularly relating to metabolic processes, whereas these did not appear altered at all in the *Der1* mouse models. This also points towards a developmental role for astrocytes and oligodendrocytes in the aetiology of t(1;11) pathology. As a final comment, this investigation has not revealed a single "smoking gun" cell type responsible for the effects of the t(1;11), or for those of other *DISC1/Der1* mutations. This is not entirely surprising given *DISC1*'s expression in a variety of cell types. Genes associated with many cell types, particularly many neuronal subclasses, appear to be dysregulated and developmental aspects will almost certainly be involved.

It is highly notable that the mouse cortical gene lists consistently implicated pyramidal cells, and in the homozygous *Der1* groups only, Medium Spiny Neurons. It is these cell groups that were highlighted by Skene *et al.* in their original analysis

using common variation, although they also implicated cortical interneurons (with weaker signal), which this analysis did not indicate as clearly. This different signal could be due to simply different aetiologies between common disease-mutation and our unique mutation. Alternatively developmental timing may be responsible, i.e., we are looking at particular cultures and mouse samples, which can only correspond (even if roughly so) to a single time point. In contrast, mutations predisposing to schizophrenia might exert their effects at a different time point, and indeed the KI superset is comprised of datasets of various ages[157,291–293].

The GO term analysis offers some clues as to what functions might be disrupted in the various cell types, and whether these functions, and the genes driving them, differ across models. In the "AstrocyteEpendymal" cell class, we can see that many terms bear high similarity between the *Der1* heterozygous cortex (CHet) and the *Der1* homozygous cortex Group Two (G2), with the first group (G1) and the human t(1;11) neurons (HTN) being more divergent. 23 and 39 genes drive the term "carboxylic acid catabolic process" in CHet and G2, respectively, and 15 are overlapping. These include *Aldoc, Glul, Gldc,* and *Sardh*. Two are also altered in the same G1 GO term. These are *Acaa2* and *Sardh.* However, particularly of interest are the "Medium Spiny Neuron" and pyramidal classes. In G1 and G2, the GO terms of the MSNs are near-identical and gene numbers are large enough that the FDRs are quite low. The term "regulation of metal ion transport" has 14 and 13 genes associated respectively; 9 of these are identical and include *Drd1a, Rgs9, Scn4b,* and *Adora2a*. In addition, all 14 genes of G1 are upregulated, while all 13 of G2 are downregulated in comparison to the WT. The term "regulation of long term synaptic potentiation" is significant in both groups, with $p=1.73x10^{-4}$ and $p=5.22x10^{-5}$, and contains the same genes, *Ptpn5 Mme, Adora2a, Drd1a*. As before, all these genes are upregulated in G1 and downregulated in G2. The human homologue of *Ptpn5* encodes STEP, a brain specific phosphatase which acts on both AMPAR and NMDAR subunits to oppose synaptic strengthening. It has been proposed that abnormally high and abnormally low expression of STEP is harmful. Stimulation of α7 receptors by Aβ appears to result in STEP over-activity, and STEP is inactivated by antipsychotic agents. Conversely, sufficient expression of STEP appears

necessary for stress resilience[294]. STEP's relevance to synaptic plasticity is obvious and the fact that it is disturbed in both homozygotes, albeit in opposite directions, is interesting.

Finally, pyramidal cells were notably disturbed in several models. In the CA1 pyramidal neuron class, the Chet model displays terms related to cell metabolism and basic activities. Although these are vague terms, more specific terms such as "negative regulation of cell morphogenesis involved in differentiation" ($p=7.91x10^{-4}$) point towards development as being abnormal in these mice. Genes here include *Slit1, Nrp1,* and *Sema3e,* while another term relating to development is dendrite morphogenesis ($p=1.68x10^{-5}$), with genes *Chrna7, Nrp1, Camk2a, Slitrk5, Rock2.* The top 10 Process terms in the G1refer to synaptic activity and ion transport, and include some of the same genes such as *Chrna7* and *Sema3e*. In contrast, the G2 and hippocampus have diverging changes. The G2 changes revolve around transcription factor disruption and fundamental RNA-related activity, while the hippocampus has only 3 genes driving its GO terms; *Cnih2, Greb1l, Crlf1.* The SS changes show a similar pattern, although this class is not significant in G2. Chet terms relate to protein trafficking and localisation, while the hippocampal terms relate to neuronal death, development, and differentiation. The G1 top term is "regulation of transmembrane transport", with 11 genes driving the term including *Arc, Cacnb4, Rgs4,* which have been discussed elsewhere in this thesis, and three potassium receptor subunits *Kcna1, Kcnb1, Kcns2.* Transport evidently emerges as a theme in the cortical samples. In addition to widespread pyramidal cell disruption, the facts remain that astrocytes have emerged as being implicated in schizophrenia pathology by this analysis. We also see clear convergences of our *Der1* homozygous model and common variation on Medium Spiny Neurons, a cell type of relevance.

8 DISCUSSION AND CONCLUSIONS

8.1 Thesis

This thesis describes the analysis of RNA-Seq data from two models relating to psychiatric illness. The first is the iPSC-derived neuronal model of the t(1;11), a translocation which segregates with highly increased risk of psychiatric illness in a Scottish pedigree. The translocation disrupts three genes, *DISC1* and *DISC2* on chromosome 1, and *DISC1FP1* on chromosome 11. Only *DISC1* is known to encode a protein[66]. The second model is a corresponding mouse model, referred to as *Der1* and described in detail by Malavasi *et al.* 2018[70]. Creation of the model involved the insertion of 150kb of human genetic material 3' to the t(1;11) chromosome 1 breakpoint downstream of the analogous location in the mouse *Disc1* gene, removing 98kb in the process[70]. Both models were subjected to RNA sequencing of comparable sequencing depth. Neurons carrying the t(1;11), and two brain regions of the mouse, were harvested and sequenced. RNA-Seq data were analysed for regional effects where appropriate, for differential expression, and some selected differentially expressed genes were verified by RT-qPCR. It also describes the overlaps seen with studies of interest and overrepresentation of specific GO terms, especially those relevant to psychiatric illness. Data from both the iPSC-derived neuronal and mouse models were also used in two separate analyses, the first to detect whether the relative proportions of cells were altered in the samples, and the second to examine if the differentially expressed genes were found associated with any particular cell types, and if so, whether these cell associated genes were associated with particular functions.

8.2 DESeq2 and DEXSeq analysis

The DESeq2 analysis of the human and mouse datasets has confirmed a number of interesting changes, while rejecting some hypotheses. Both iPSC-derived neurons and *Der1* mice displayed the expected effects on *DISC1/Disc1* expression, and were clearly affected by the translocation/*Der1*. Over 1,200 genes were found differentially expressed in the human neurons, with over 2,000 in the heterozygous *Der1* mouse cortex, and a significantly lower number of 175 in the hippocampus. The effect on the homozygotes was less clear, but they differed from the WT samples. The cortex samples showed a surprising splitting into two groups which is

discussed in detail in the relevant chapter. It can also be stated that there is no evidence for the t(1;11) causing regional expression effects, or for significant transcriptional alterations of potential *DISC1* interactors. There is ample evidence for extensive overlap with analogous experiments utilising iPSC-derived neurons with *DISC1* mutations, and with genes implicated by studies searching for common mutation predisposing to schizophrenia. This was seen in both the human and mouse cortex datasets. Potential *Disc1* interactors in the mouse were altered at a rate greater than expected by chance[73].

qPCR results confirmed changes in a number of genes, while also confirming the reliability of the human RNA-Seq analysis. The qPCR results in the mouse cortex showed a looser relationship between qPCR score and gene counts, suggesting that the mouse analysis is less reliable. This apparent imbalance may simply be due to the larger variation in total sequencing depth across mouse samples, however. In both datasets the changes show a general linear trend between the overall log2fold change of the qPCR score and counts between genotypes, so this is reassuring as to the validity of the data. The changes in the mouse were also more subtle and rarely resulted in the doubling or halving of a gene's expression. Finally it must be noted that PCAs showed that mouse sex was a significant factor in differentiating the samples; correlating with the second largest factor after genotype. It is possible that this is a factor in non-significance of genes and the appropriate solution would be to utilise a larger sample set. As it stands if the samples were split by sex, 4 Hets vs 3 WTs would not be an adequate sample size for multiple qPCRs. As to the genes which were actually confirmed, a number of highly relevant genes were shown to be differentially expressed. These findings and their possible relevance are discussed below, along with other differentially expressed genes and the gene ontologies overrepresented among them. In order to give a full and complete discussion, reference is made to the results of the EWCE analysis, a summary of which is given in the relevant chapter. This discussion therefore covers i) the altered genes, ii) which cells and activities these appear to be related to, and iii) how this might result in deleterious phenotypes.

Human genes I confirmed by qPCR included *BBS1, CALB1, DRD2, GPC1,METRN, NTRK2, QKI,* with confirmed exon level changes in *DPYSL3, NTRK2,* and *SLC12A2*. All of these, except for *DRD2*, were downregulated. Mouse confirmed cortical changes included *Arc* and *Avp*, both in the heterozygote, both downregulated. Summaries of all genes can be found in the relevant sections of Chapters 3 and 4.

8.3 Deconvolution analysis

The purpose of the deconvolution analysis, utilising DeconRNASeq developed by Gong *et al.*, was to search for changes in cell specific marker expression indicating differences in prevalence of these cell types[135]. Unlike the EWCE analysis of Skene *et al.*, this is roughly quantitative and takes account of the direction of gene expression changes, and makes the assumption that consistent changes in expression of sets of cell specific markers indicate changes in cell proportions. Two separate sets of analyses were carried out for the deconvolution analysis for both mouse datasets as well as the human dataset. The first utilised the RNA-Seq data of enriched profiles described by Zhang *et al.* and looked at broad class level differences in the prevalence of each cell type[203]. The second looked at scRNA-Seq data generated by Zeisel *et al.*, and looked at subclass level differences[157]. A third, minor analysis involved the use of data generated by Darmanis *et al.*, specifically to give a human to human comparison for class level data[295]. In all cases, different numbers of marker genes and housekeeping normalising genes were utilised, and these settings were verified using pseudosamples.

The Zhang cortical analysis was characterised by quite high accuracy in the deconvolution of pseudosamples, with mean absolute difference in proportion of the 6 cell types being an average of around 6% between predicted and actual proportions. The comparison dataset, Zhang Two, was also highly well predicted and these enriched samples were identified as being over 95% of the correct cell type, except at high marker numbers which tend to incorporate markers of reduced specificity. The cortical samples in the Zhang Two dataset bear high similarity in cell proportions, to the proportions found in our sample deconvolutions. The hippocampal deconvolution was highly similar, with even better accuracy of

pseudosample deconvolution, as were the human Zhang and Darmanis deconvolutions. These all also noted that higher marker numbers were suboptimal. Overall these are reassuring proofs as to the validity of the deconvolution. There was no genotype effect on any cell type in the hippocampal or cortical *Der1* samples that showed pairwise significance. The human results however bore some significances; the Zhang analysis showed a mild but universally significant effect on astrocyte proportions (decreased in t(1;11)) which survived multiple testing correction (Sidak-Bonferroni). Astrocytes had also been shown to be frequently significant in the mouse cortical analysis, though never in pairwise comparisons. In contrast, the highly variable Darmanis deconvolution had no clear effect.

To conclude, there are no changes which survived post-hoc significance testing in the *Der1* mouse samples. However, it is possible that astrocytic proportions or transcriptional activity are altered in the human cells. The Group One and Group Two of the cortical homozygotes also were more divergent from one another when t-tests were carried out, indicating greater dysfunction. This ties in with the results of the EWCE implicating astrocytes universally, and indicates that the dysfunction thereof is due to decreased cell numbers or activity.

The Zeisel analysis had diverging results. In all cases, confidence in these results must be lower, as the comparative dataset deconvolution showed very poor prediction, and in any case poor housekeeping expression across the comparative Allen dataset means there were few cells of some types to compare to. Mean absolute difference for pseudosamples was also worse than in the Zhang analysis; minimums of 20% for cortical, 24% for hippocampal, 14% for human. These are, in real terms, quite large margins of error. This is to be expected given the larger number of cell types. The cortical analyses were highly dependent on whether Der1 homozygous cortical samples were split into two groups; this is an indication of the difference of these two sample sets, as also shown by the Zhang analysis. Results were also quite dependant on deconvolution settings; there were no cell types which displayed universal significance according to an ANOVA test, and only one pairwise difference between genotypes was ever reported significant (out of a large number). The cell type exhibiting the most common agreement on significance was Int10, with

a total of two thirds of the deconvolutions regarding it as significant when cortex homozygotes were split into two groups. Three variations were used for the cortical deconvolution, leaving out or merging certain cell types, and the Der1 sample proportions were quite similar regardless of this, with most cell types not changing a great deal if the deconvolution settings were the same. In contrast, changing deconvolution settings resulted in large changes in overall predicted cellular proportions. The hippocampal analysis was characterised by similar variation, particularly if the number of normalising housekeeping genes was altered, and no cell type displayed universal significance or posthoc testing significance. Int4 and Int14 showed the highest degree of support. The human analysis displayed no evidence for any cell types being significant.

To conclude, it would be irresponsible to not be suspicious of the Zeisel deconvolution results. The high variation in pseudosample deconvolution and poor Allen prediction means they should not be relied upon. In any case, no cell type possessed universal significance across the deconvolution settings in any of the three analyses. As stated earlier, this poor prediction is an inevitable consequence of large numbers of cell types. In this light it is unsurprising that I could not carry out a reliable analysis. I do note that the cortical analyses changed drastically when the homozygote samples were split into groups or kept together; highlighting the divergence between these two groups.

8.4 EWCE analysis

The EWCE analysis revealed that differentially expressed genes were associated with a variety of cell classes, “AstrocyteEpendymal”, “PyrSS”, “Oligodendrocytes”, and “EndothelialMural” being significantly implicated in the human cells. The challenge is now to explore and explain the functional relevance of these implications.

Pyramidal cells were significant in most analyses, although the exact GO terms and genes implicated tended to differ across the models. The human PyrSS genes were few in number and related primarily to gonadotrophin signalling. Those of the hippocampal heterozygous Der1 were primarily cell death related. The Der1 mouse

het cortex had a number of highly interesting and related genes altered, which in many cases converge on calcium influx. In addition to the subunits encoded by Cacnb3 and Cacng3, I noticed a co-dysregulation of *Chnra7* (down), *Chrnb2* (down), *Lynx1* (up), and *Lypd6* (up). There is a wealth of research into the latter two genes, which are related to the toxins in snake venoms and function as endogenous modulators of nicotinic acetylcholinergic receptors[296]. *Lypd6* is highly enriched in neuronal tissue, and overexpression enhances nicotine-evoked calcium influx through nAChRs, while knockdown decreases this influx[296]. In contrast, knockdown of *Lynx1* appears to enhance this influx. The authors also suggested that Lynx1 could exert a neuroprotective effect by preventing nAChRs-mediated excitotoxicity [297] Lynx1 also controls dendritic spine dynamics; knockdown appears to increase the rate of dendritic spine formation and removal[298], It has been referred to as a "cholinergic brake", which rises in expression in adult life to block plasticity in the mouse visual cortex[299].

The nAchRs are described as having quite divergent effects on neuronal activity; expression of them in interneurons in cortical layers 2/3 means they have an inhibitory effect on pyramidal cells, whereas they appear to have an different role in layer 6 as they are expressed directly by the pyramidal cells themselves[300]. This appears to be via receptors containing the subunit encoded by *Chrnb2*, and results in the strengthening of glutamatergic synapses. Nicotine can stimulate LTP in L6 neurons, but not the shallower layers, and the subunit composition of the receptor is important too. For example α7 encoded by *Chrna7* is not necessary for nicotine-LTP in L6[300]. However, it is the case that the *Der1* cortical samples and the t(1;11) neurons are not distinctly associated with any layer, and it is therefore premature to draw conclusions about the effect of disrupted cholinergic transmission on cells, given its layer specific effects. It is also the case that differential expression of *Lynx1* and *Lypd6* in various cell types may modulate the effects of endogenous acetylcholine. One research group found these two "proto-toxins" are expressed in parvalbumin positive and somatostatin positive interneurons, respectively, and not co-expressed. Mice lacking *Chrna7* also show deficits in parvalbumin positive interneurons, and that gene is downregulated in the *Der1* heterozygous cortex[198]. The

Karolinska Institute dataset shows that *Lynx1* is broadly and near equally expressed in pyramidal cells from multiple layers of the cortex, while *Lypd6* is much more associated with interneurons (especially Sst^+ ones).

However it is highly interesting that *Lynx1* is upregulated in the *Der1* cortex while *Chrnb2* is downregulated, implying diminished cholinergic receptor activity. A possible outcome from this would be diminished capacity for LTP in the mouse heterozygous cortex in the appropriate layers. Is the same situation occurring in the mouse heterozygous hippocampus? The pyramidal cells show clear convergences on cell death here, and surprisingly both *Chrna4* and *Lypd6b* are upregulated. α4 (upregulated in the het hippocampus) and β2 (downregulated in the het cortex) subunits together form a nicotinic receptor. Unfortunately there is little information available on *Lypd6b* at present, but if it is similar to *Lypd6* this is evidence that the nicotinic situation is reversed in the hippocampus compared to the cortex. Excitotoxicity could be the cause of the dysregulated cell death genes we observe. It must be stressed that this is a preliminary theory; there are of course many processes at work in neurons, and the dysregulations could be by chance. It must also be noted that in each brain region only one of the pair of subunits is dysregulated, but there are some interesting convergences on the theme of diminished synaptic activity in the cortex from other cell types.

As mentioned previously, astrocytes are of immense importance to neuronal function, and have been designated as part of a "tripartite synapse" by some researchers. It has been shown that fluctuations in Ca^{2+} cytosolic concentration alter astrocytic activity, and that *in vivo* elevations of this concentration have been observed in response to synaptic release of norepinephrine and glutamate in the cortex[301]. Of direct relevance might be the role of astrocytes in clearing neurotransmitters, preventing excitotoxicity, and maintaining homeostasis. A number of differentially expressed genes which are most highly expressed in the "AstrocyteEpendymal" class are related to this very function. Glutamate is a particularly potent excitotoxic agent which is removed from the synaptic cleft by glutamate transporters EAAT1-4, with the first two being primary involved in this function and expressed in astrocytes[302]. *SLC1A3*, encoding EAAT1, is differentially

expressed, as is *SLC1A6*, encoding EAAT4, although this is described as a neuronal gene and correspondingly does not have its maximum expression in "AstrocyteEpendymal" cells[302]. Nevertheless downregulation of the gene in neurons would have a similar effect of excess synaptic glutamate. A similar scenario is seen with glycine, another neurotransmitter. Three of the four genes involved in the glycine cleavage/synthesis system are differentially expressed, as is the receptor *GLRA1*. Of the three, *GLDC* has maximum expression in astrocytes, while the other subunits *DLD, GCSH*, as well as *GLRA1* are expressed across a variety of neurons. All the genes mentioned above are downregulated in the t(1;11) samples, with the exception of *GLRA1*. The functional implications of these changes would mean excess glutamate and glycine; possibly at the "trisynapse" given the role of astrocytes in clearing these neurotransmitters here and the fact that most of the genes are astrocyte-associated. The imbalance might have the effect of unusual synaptic plasticity, as synaptic NMDARs must be activated for both LTP/LTD and NMDAR-mediated cell death[303]. However, it has been shown that these synaptic NMDARs exhibit a preference for D-serine rather than glycine as a co-agonist[303], and in addition that high levels of glycine can stimulate LTD in opposition to LTP[304]. Astrocytes themselves do release neurotransmitters, including D-serine and glutamate in response to neuronal activity[301]. This has been shown to modulate plasticity via NMDARs, and it is the case that impaired clearance of neurotransmitters, regardless of neuronal or astrocytic source, could result in aberrant plasticity or even cell death. The dysregulation of both the glycine and glutamate neurotransmitters is therefore some evidence for aberrant activity of NMDARs. The findings of our research group, published as Malavasi *et al.* 2018, showed evidence for potential weaker synaptic activity in *Der1* mice; as suggested by total PSD-95 prevalence being unchanged but its distribution shifted towards less nanodomains per PSD-95 cluster[70]. In this context, the human t(1;11) alterations fit perfectly and give a potential explanation for this finding; aberrant LTD. The calcium/nicotinergic transcriptional alterations seen in the mouse *Der1* cortical heterozygote also fit well with the phenotypes observed by Malavasi *et al.*.

A second theme in the universally altered "AstrocyteEpendymal" cell set was fatty acid and lipid metabolism. Human genes included *CPT2, ECI2* (fatty acid oxidisers), *GPC5,* and *IMPA2* (lipid phosphatase). *Cpt2* was also changed in the mouse homozygous G1 group, as was *Cpt1a* in the mouse heterozygote. *Apoe* was altered in all three mouse groups, although I could not confirm this with qPCR in the mouse cortical heterozygote. Myelin is an unusually lipid-heavy construct; 70% of the dry weight is composed of lipid. Astrocytes can promote the myelination of neurons in oligodendrocyte-neuron culture, and it has been shown that they are a source of the lipids needed to form myelin. SCAP, the sterol sensor which activates cholesterol and fatty acid related transcription factors in astrocytes, is important for myelination formation in mice. Mice with $SCAP^{-}$ oligodendrocytes have a neurological phenotype (microcephaly, tremors, increased lethality) and hypomyelination. Mice lacking the same protein in astrocytes also displayed microcephaly, hypomyelination, and downregulated Mag and Mbp[305]. This has crucial implications for proper neuronal functioning. Cholesterol has been shown to have effects on synapse formation too. Both glial-derived media containing cholesterol, and cholesterol itself, increase electrophysiological activity and synapsin/glutamate receptor staining in neurites[306]. It has been suggested that the carrying agent for this cholesterol is Apoe-positive lipoproteins[307]. More recent papers have shown that increased cholesterol elimination boosts dendritic output, which suggests that less cholesterol is better for synaptic activity[308]. These authors also reported increased phosphorylation of Trk compared to TrkB, although total levels of TrkB (encoded by *Ntrk2*, homologue confirmed differentially expressed in t(1;11) neurons by qPCR) were not reported. The authors suggested that this might be related to the distribution of lipids and TrkB in the cell membrane[308]. This also suggests that fatty acid metabolism in neurons alone, even without accompanying glia, is important. Therefore whether in neurons or astrocytes (which should be present in some proportion in our samples) the alterations in fatty acid metabolism matter. Given the deficit in lipid metabolism by astrocyte-associated genes, and the dysregulation of *Apoe* in all mouse models examined here, one might expect it to be the case that the t(1;11) and *Der1* models display myelination deficits. It must be noted that as the *Apoe* changes were not significant at qPCR level, the changes are either minor, not genuine, or the small

sample size makes discovery difficult. Expression was also not investigated in the cortical homozygotes.

The human cells, as well as the *Der1* homozygote cortex Group Two, had the cell type "Oligodendrocytes" significantly associated with differentially expressed genes. The *Der1* cortex Group Two oligodendrocyte genes have a clear association with myelination, with *Mbp*, *Mag*, and *Plpl1* being differentially expressed and associated with this term. Similarly, this thesis shows that the human gene *QKI*, a regulator of these crucial myelination genes, is differentially expressed in the t(1;11) cultures[216]. There is no evidence of differential proportions of oligodendrocytes as shown by my deconvolution analysis. The t(1;11) cultures should not contain a large number of oligodendrocytes, but as described earlier in the *QKI* section of Chapter 3 reasonable expression of immature markers is observed. *Qk* (the *QKI* homologue) is also reasonably expressed across cell types in the Karolinska Institute superset; mostly in astrocytes and oligodendrocytes, but with some expression in pyramidal and neural progenitor cells. Also dysregulated in the human neurons are *NRG1* and *ERBB4*[70]. Interestingly, multiple papers have shown that inadequate signalling of this receptor-ligand pair results in myelination deficits. NRG1 promotes the survival, migration, and proliferation of Schwann cells, while mouse lines with dominant-negative Erbb receptors in oligodendrocytes and myelinating cells have thinner neurons, abnormal myelination forming, and abnormal expression of myelination proteins including Mbp (but not Mag)[309]. The group went on to show these mice had hypersensitivity to amphetamine, greater dopamine-induced signalling, and increased dopamine receptor expression (type 1 significant, no stated distinction between receptors within this type)[310]. *Drd1* and *Drd2* are both dysregulated in the cortical *Der1* homozygotes (both up in Group One, down in Group Two), and *DRD2* is qPCR confirmed as differentially expressed in the t(1;11) neurons. It is, to reiterate, a target of anti-psychotic medication.

Relevant to this cell type is the publication of a recent paper examining iPSC-derived t(1;11) carrying oligodendrocytes, as well as myelination phenotypes of the *Der1* mice, and members of the Scottish pedigree[311]. They found t(1;11) carriers of the pedigree, 8 in total, all with a psychiatric diagnosis, had altered white matter

connectivity compared to 13 controls (12 without a diagnosis, one with). The white matter tracts connecting grey matter nodes were decreased in strength and in number, by a mean of 1.81% and 1.64% respectively. iPSC-derived oligodendrocytes carrying the t(1;11) had a lower proportion of KI-67^{+} differentiating cells after three weeks, as well as drastically reduced *DISC1*, expressed at 30% of the WT level. This implies erroneous early development; as also suggested in t(1;11) neurons by my findings that the differentiation stage expressed genes *GPC1*, *METRN*, *SLC12A2* (exon level) are downregulated. RNA-Seq showed differential expression of 228 genes, with GO terms such as "nervous system development" and "myelination" overrepresented. Oligodendrocytes were smaller in t(1;11) carrying lines, and the *Der1* mouse heterozygous cortex showed unusual myelination, with more myelin sheaths and shorter myelin internode lengths. The authors also noted that oligodendrocytes have been associated with schizophrenia in particular previously[311].

To summarise these papers, EWCE findings, and qPCR results;

i. Other researchers have shown abnormalities of white matter in the t(1;11) family, and that the *Der1* mouse cortex and t(1;11) carrying oligodendrocytes are abnormal. The mice have deficits in myelination, while the cells are smaller and appear to differentiate abnormally, apparently earlier.

ii. Researchers have also shown that myelination deficits result from impaired Nrg1/Erbb4 signalling, and the consequences include amphetamine sensitization caused by upregulated dopamine receptors.

iii. Other researchers have shown that astrocytes are a crucial source of fatty acids for the synthesis of myelin in oligodendrocytes.

iv. Astrocytes are also important for the buffering of glutamate, glycine, and other neurotransmitters, which aids in synaptic plasticity. They also produce cholesterol, carried by Apoe, which can cause TrkB phosphorylation ratio changes and also impacts on synaptic plasticity.

v. I have demonstrated here the dysregulation of functions relating to these factors. Downregulation of *QKI* (point i, myelination), of *SLC12A2*/*GPC1*/*METRN* (point i, development), of *NTRK2* (point iv, both gene and exon) were proven by qPCR. Notably, *Apoe* could not be confirmed in the mouse, but *Arc* downregulation may have relevance to synaptic plasticity.

vi. The RNA-Seq implicates astrocyte neurotransmitter homeostasis in both mouse and human datasets (point iv), as well as astrocyte fatty acid metabolism (point iii, leading to point i and ii).

vii. Previously described were the downregulation of *ERBB4* and *NRG1*, with the co-occurring upregulation of *DRD2* (point ii)[70].

These papers, EWCE findings, and qPCR results are summarised in image form in Figure 102.

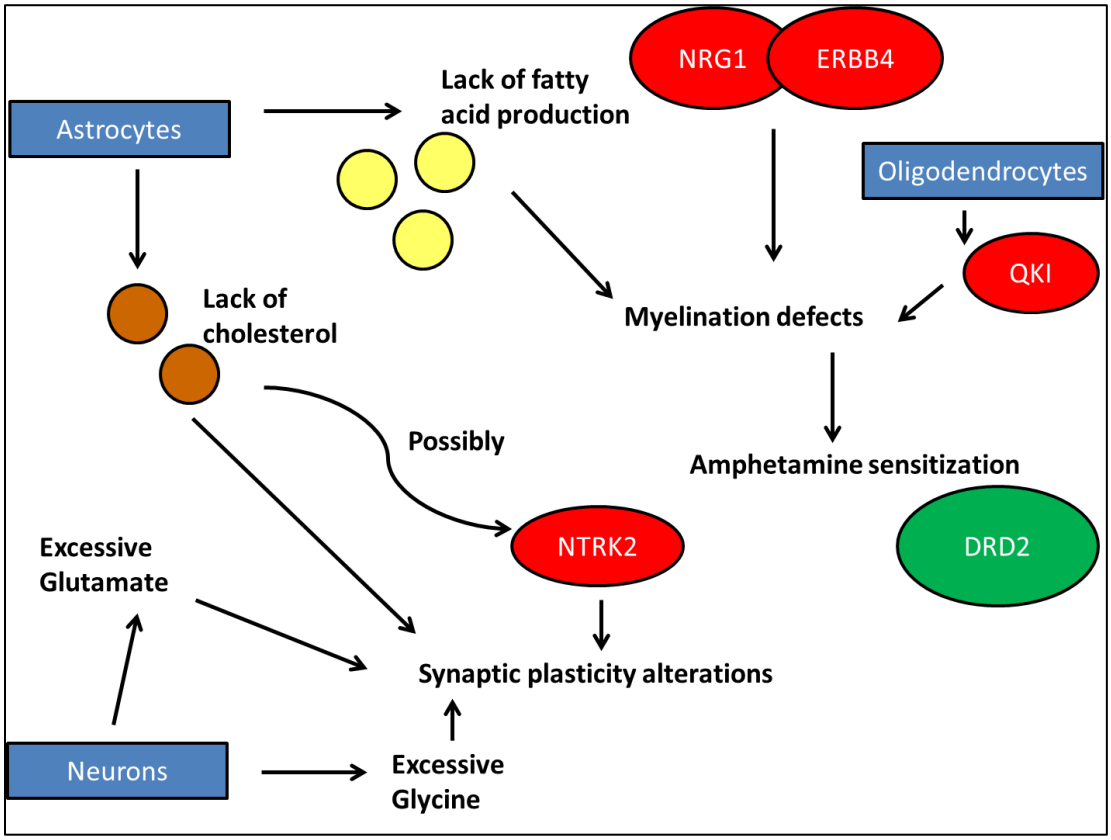


Figure 102. Conclusions of discussion. Cell types indicated by EWEC are represented by blue squares. Genes which are qPCR verified as differentially expressed in the t(1;11) iPSC-derived neurons are represented by circles; red indicating downregulated in t(1;11), green indicating upregulated in t(1;11).

A logical conclusion can be developed from these findings and the links between them. The recent finding that myelin and white matter integrity is altered in oligodendrocytes and carriers of the t(1;11) fits cleanly with the convergence of two separate pathways on myelin production. The first is ERBB4/NRG1 signalling, leading to upregulation of dopamine receptor signalling and impaired myelination, while the second is astrocytic malfunction. This not only can alter the supply of fatty acids needed to produce myelin but also has effects on synaptic plasticity. There is evidence for altered synaptic activity in the *Der1* mice and verified qPCR changes in related genes in the t(1;11) and *Der1* samples. There is therefore a strong case for myelination and synaptogenesis being altered by t(1;11)/*Der1*, likely via one or both of the pathways of astrocytic malfunction and ERBB4/NRG1 signalling causing hyperdopaminergic signalling.

Exactly how myelination and synaptogenesis, and hypomyelination and synaptic weakening specifically, are related is difficult to say. However, a review by O'Rourke *et al.* notes that "OPCs are the only glial cell type to receive direct synaptic input from neurons…and glutamatergic synaptic signalling can direct the local translation of a myelin protein (myelin basic protein) at the site of axon-OPC contact in vitro". There are clearly links between the two processes which enhance the communicative aspects of neuronal activity, and the findings of this thesis and recent papers point towards these links as potentially being important in major mental illness.

8.5 Future directions

This thesis has highlighted the disruption of many functions in the t(1;11) neuron cultures, and described a plausible dysregulation in the activity or proportion of astrocytes or related functions. The high level of overlap between the t(1;11) neuron cultures and *Der1* mice, especially as regards the EWCE analyses, implies that the *Der1* mouse accurately models some aspects of the t(;11). As described above, there are extensive links between synaptic dysregulation (notably dopaminergic hyperfunctioning and ERBB4/NRG1 signalling), myelination, and astrocyte homeostasis and metabolic support. Future approaches involving the t(1;11) could be biochemical and investigate these disturbed processes, although the relative immaturity of the iPSC-derived neurons make this difficult. Responses of the cells to dopamine, glutamate, and other neurotransmitters, as well as drugs, would be highly interesting and could utilise the *Der1* if the t(1;11) neurons are not sufficiently mature. Other potential avenues of experimentation could involve the BBS subunits, which are important for early cellular migration and division. Dendritic outgrowth is also a target for future work. One drawback of the "cells-in-a-dish" approach is the difficulty in accessing the cell-specific effects of receptors such as the nAchRs; as described above, these have diverging effects in different layers of the cortex.

Now that an oligodendrocyte t(1;11) model has been made which displays abnormal phenotypes, the obvious next move is to examine an astrocytic t(1;11) model. This would be of high interest; interactions between t(1;11) astrocytes and other cell types

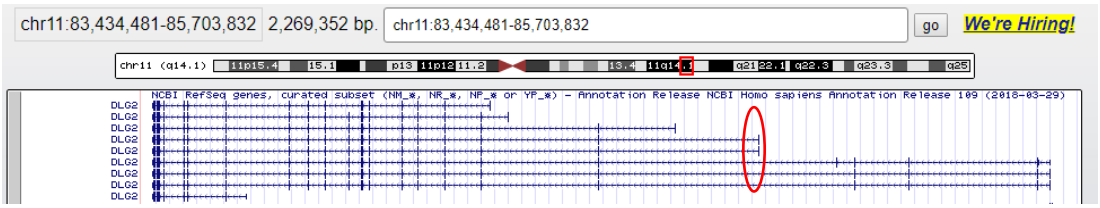
(neurons, oligodendrocytes) could be examined by co-culture and compared to WT astrocytes. Indeed, t(1;11) neurons and oligodendrocytes would allow the examining of emergent phenotypes only evident when several cell types carry a mutation. In nature, all these cells have the t(1;11), and the most accurate insights will come from an experimental design that acknowledges this. Astrocytes have already been generated from iPSCs with protocols that report near universal expression of astrocytic markers and efficient generation[312]. Some of these papers, which mainly focus on neurodegenerative diseases, have specifically sought to examine the non-cell autonomous nature of those diseases. They therefore utilised cell co-cultures, examining phenotypes of relevance to neurodegeneration[313]. Could this be extended to RNA-Seq? There exist bioinformatics methods to extract the individual RNA-Seq profiles of mixed cultures, where each cell type originates from a different species[314]. In theory, this could be used to examine the status of each cell type, e.g., t(1;11) neurons and *Der1* mouse astrocytes cultured together vs t(1;11) neurons and WT mouse astrocytes. This would sacrifice some accuracy (as mouse astrocytes are not human astrocytes) but would allow RNA-Seq analysis of individual cell types. The added advantage is that both of these models already exist, and as this thesis shows, astrocytes appear to be highly important to t(1;11) pathology.

9 Appendix

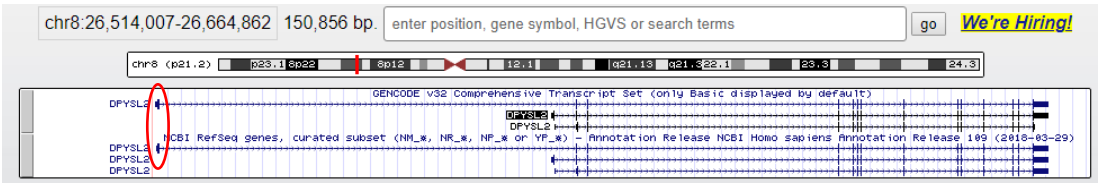
9.1 Exon illustrations in selected qPCR DEXSeq candidates

Differentially expressed exons which were targeted with primer pairs are highlighted with a red circle. Images taken from UCSC Genome Browser.

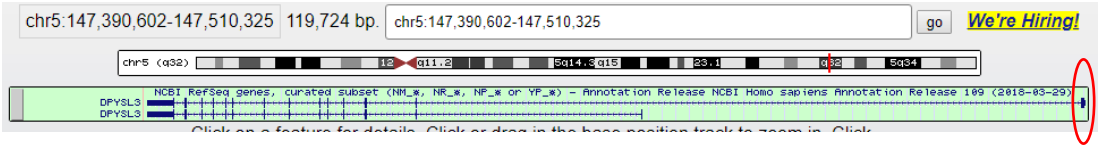
9.1.1 DLG2



9.1.2 DPYSL2

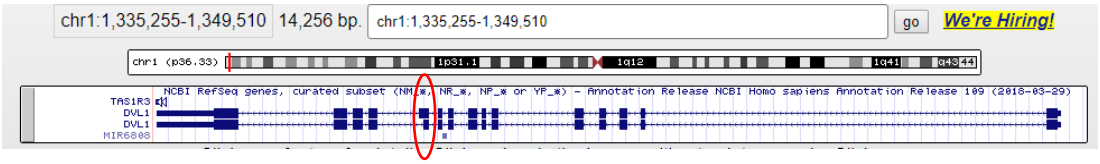


9.1.3 DPYSL3

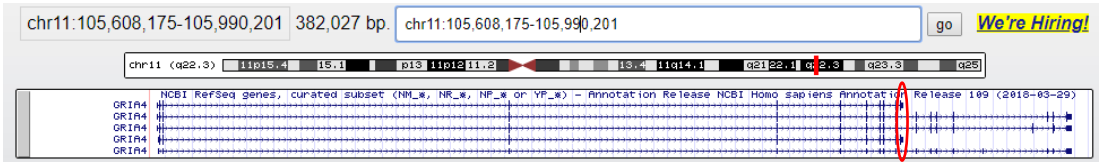


9.1.4 DVL1

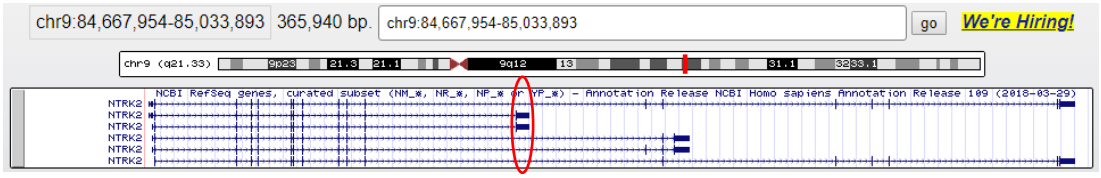
Note that two related exons are within the circle. The non-overlapping part of the larger exon is where one primer was located, while the second was in an adjacent exon.



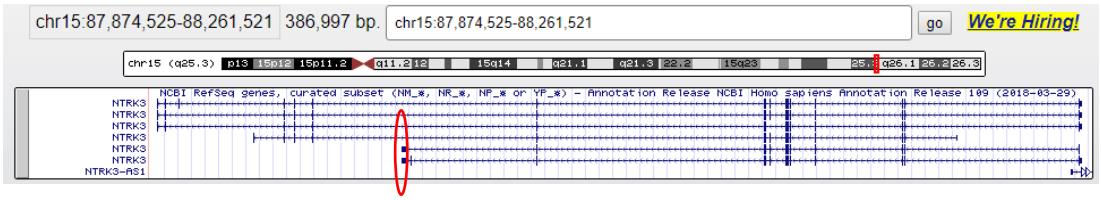
9.1.5 GRIA4



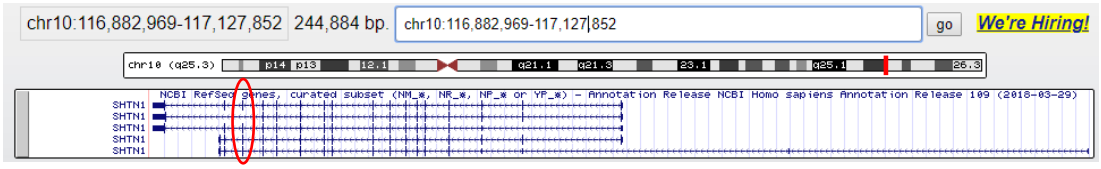
9.1.6 NTRK2



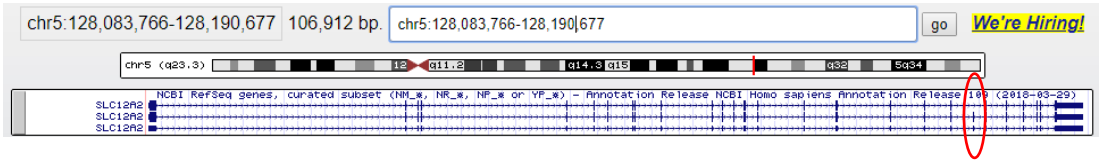
9.1.7 NTRK3



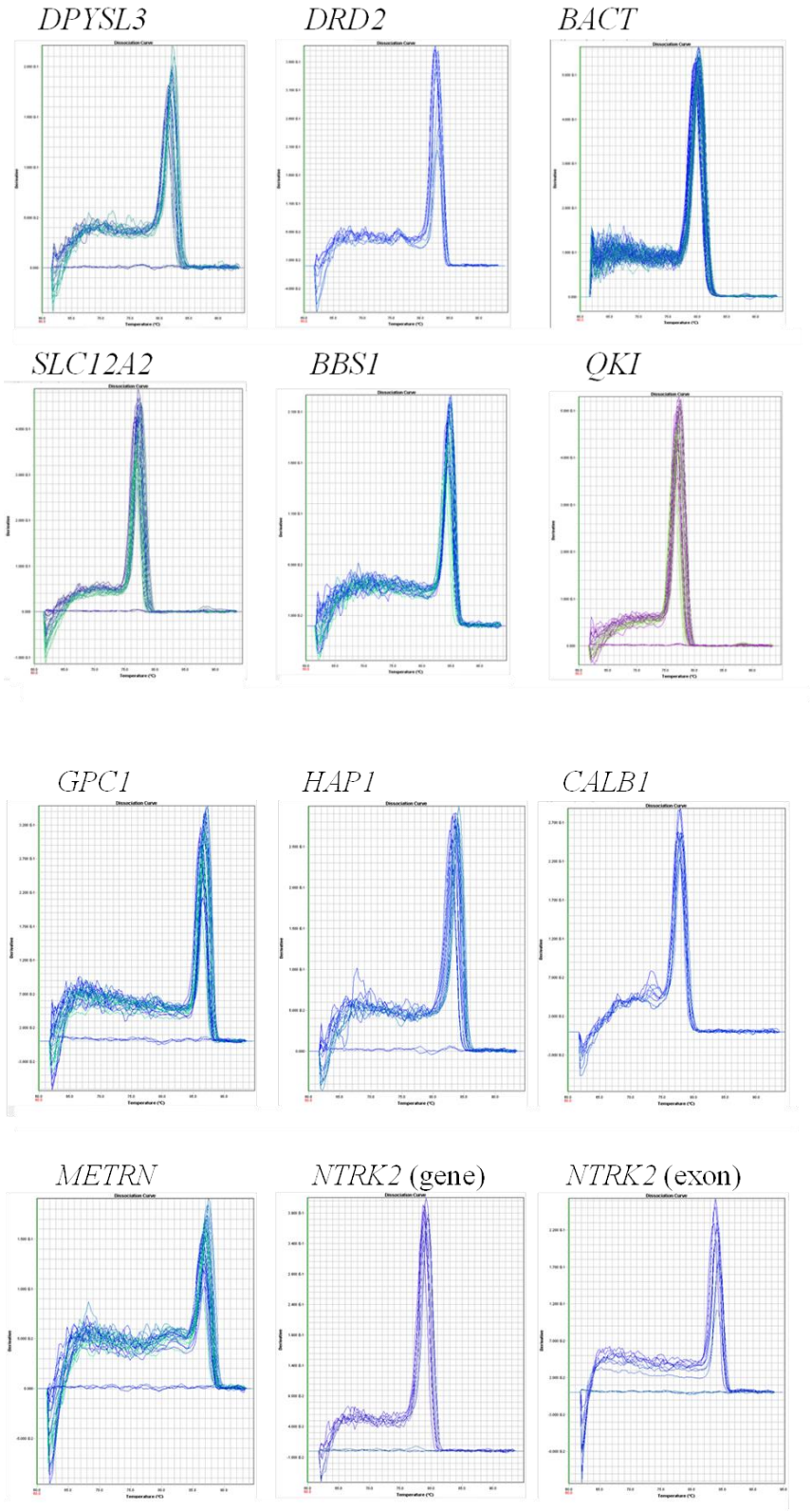
9.1.8 SHTN1



9.1.9 SLC12A2

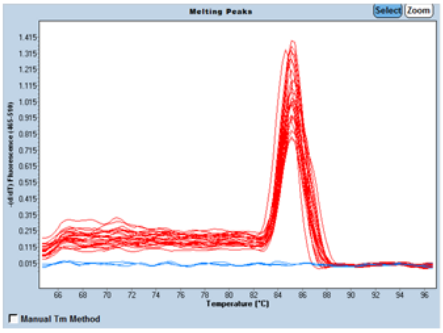


9.2 Dissociation curves of human qRT-PCR products, showing a single product (flat lines are negative controls).

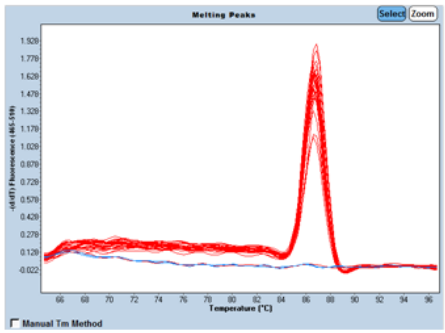


9.3 Dissociation curves of mouse qRT-PCR products, showing a single product in all qPCRs (and blue flat lines in negative controls).

Arc



Avp



9.4 Top 25 GO terms for housekeeping genes chosen in deconvolution, by Process, Function, and Component

9.4.1 Cortex Zhang

Process	P-value	FDR q-value	Enrichment	Function	P-value	FDR q-value	Enrichment	Component	P-value	FDR q-value	Enrichment
cellular metabolic process	5.51E-20	8.45E-16	1.61	binding	3.49E-11	1.58E-07	1.21	intracellular organelle part	4.53E-34	8.79E-31	1.73
macromolecule metabolic process	6.58E-19	5.05E-15	1.7	ubiquitin-like protein binding	2.11E-09	4.76E-06	6.59	intracellular part	7.28E-33	7.05E-30	1.33
metabolic process	6.28E-18	3.21E-14	1.51	organic cyclic compound binding	9.16E-09	1.38E-05	1.44	organelle part	2.26E-31	1.46E-28	1.67
nitrogen compound metabolic process	8.58E-18	3.29E-14	1.61	RNA binding	1.20E-08	1.36E-05	2.19	membrane-bounded organelle	4.95E-29	2.40E-26	1.46
primary metabolic process	2.10E-16	6.44E-13	1.54	heterocyclic compound binding	2.41E-08	2.18E-05	1.43	intracellular membrane-bounded organelle	1.06E-28	4.11E-26	1.5
macromolecule catabolic process	2.28E-16	5.83E-13	3.2	ubiquitin binding	3.33E-08	2.51E-05	6.98	protein-containing complex	7.75E-27	2.50E-24	1.81
organic substance metabolic process	2.76E-15	6.05E-12	1.5	nucleoside phosphate binding	1.30E-07	8.36E-05	1.73	intracellular organelle	4.96E-25	1.37E-22	1.39
cellular macromolecule metabolic process	1.88E-14	3.61E-11	1.72	nucleotide binding	1.30E-07	7.31E-05	1.73	organelle	5.04E-24	1.22E-21	1.37
proteolysis involved in cellular protein catabolic process	2.69E-14	4.59E-11	3.54	catalytic activity	1.57E-07	7.86E-05	1.37	nuclear part	6.27E-22	1.35E-19	1.92
cellular macromolecule catabolic process	9.32E-14	1.43E-10	3.18	small molecule binding	6.79E-07	3.06E-04	1.62	cytoplasmic part	1.15E-20	2.23E-18	1.49
modification-dependent macromolecule catabolic process	1.36E-13	1.90E-10	3.63	ubiquitin-like protein ligase binding	1.07E-06	4.39E-04	3.01	catalytic complex	1.09E-18	1.92E-16	2.7
protein metabolic process	1.69E-13	2.16E-10	1.76	ubiquitin protein ligase binding	1.32E-06	4.98E-04	3.06	cell part	3.05E-18	4.93E-16	1.19
modification-dependent protein catabolic process	2.88E-13	3.40E-10	3.62	protein binding	2.50E-06	8.67E-04	1.23	proteasome accessory complex	5.27E-14	7.85E-12	23.99
ubiquitin-dependent protein catabolic process	4.82E-13	5.28E-10	3.64	modification-dependent protein binding	2.81E-06	9.05E-04	4.21	mitochondrial part	1.66E-13	2.30E-11	2.87
cellular localization	7.50E-13	7.66E-10	2.09	purine ribonucleoside triphosphate binding	3.49E-06	1.05E-03	1.7	mitochondrial protein complex	6.75E-13	8.72E-11	4.73
proteolysis	6.92E-12	6.63E-09	2.48	ribonucleotide binding	5.34E-06	1.51E-03	1.67	mitochondrion	8.14E-13	9.86E-11	2.14
cellular catabolic process	9.71E-12	8.76E-09	2.24	purine ribonucleotide binding	7.39E-06	1.96E-03	1.66	peptidase complex	8.96E-13	1.02E-10	8.29
macromolecule localization	1.51E-11	1.29E-08	2.04	proteasome-activating ATPase activity	7.51E-06	1.88E-03	24.72	nucleoplasm	1.82E-12	1.96E-10	2.02
catabolic process	1.76E-11	1.42E-08	2.12	nucleoside-triphosphatase activity	7.82E-06	1.86E-03	2.04	nucleus	2.25E-12	2.29E-10	1.48
intracellular transport	3.84E-11	2.94E-08	2.31	purine nucleotide binding	9.30E-06	2.10E-03	1.65	proteasome regulatory particle	1.14E-11	1.10E-09	29.66
organonitrogen compound metabolic process	5.26E-11	3.84E-08	1.57	pyrophosphatase activity	1.48E-05	3.17E-03	1.97	organelle membrane	2.10E-11	1.94E-09	2.17
protein localization	1.01E-10	7.04E-08	1.99	hydrolase activity, acting on acid anhydrides	1.61E-05	3.31E-03	1.96	proteasome complex	5.32E-11	4.68E-09	9.12
protein catabolic process	1.37E-10	9.15E-08	3.5	hydrolase activity, acting on acid anhydrides, in phosphorus-containing anhydrides	1.61E-05	3.16E-03	1.96	endopeptidase complex	6.84E-11	5.77E-09	8.97
establishment of protein localization	1.39E-10	8.85E-08	2.23	ubiquitin-specific protease binding	2.29E-05	4.31E-03	13.24	ribonucleoprotein complex	1.06E-10	8.58E-09	2.51
organic substance catabolic process	2.65E-10	1.63E-07	2.15	ATP binding	2.31E-05	4.18E-03	1.71	proteasome regulatory particle, base subcomp	1.19E-10	9.24E-09	24.72

9.4.2 Cortex Zeisel

Process	P-value	FDR q-value	Enrichment	Function	P-value	FDR q-value	Enrichment	Component	P-value	FDR q-value	Enrichment
translation	3.59E-51	5.50E-47	9.44	structural constituent of ribosome	2.19E-47	9.91E-44	12.25	protein-containing complex	1.11E-67	2.16E-64	2.31
peptide biosynthetic process	2.80E-49	2.14E-45	8.93	RNA binding	2.89E-36	6.54E-33	3.9	intracellular organelle part	8.91E-63	8.69E-60	1.98
peptide metabolic process	1.26E-44	6.44E-41	6.98	structural molecule activity	5.66E-27	8.53E-24	4.26	organelle part	1.25E-59	8.16E-57	1.91
amide biosynthetic process	7.68E-43	2.94E-39	7.09	rRNA binding	1.54E-17	1.74E-14	11.06	intracellular part	3.00E-58	1.46E-55	1.41
cellular amide metabolic process	1.71E-36	5.25E-33	5.04	mRNA binding	4.73E-17	4.28E-14	5.39	ribonucleoprotein complex	2.79E-55	1.09E-52	5.18
cellular nitrogen compound biosynthetic process	7.06E-35	1.80E-31	3.71	nucleic acid binding	6.03E-17	4.55E-14	1.92	ribosome	5.02E-54	1.63E-51	12.61
cellular macromolecule biosynthetic process	7.64E-34	1.67E-30	4.17	heterocyclic compound binding	4.86E-14	3.14E-11	1.59	cytoplasmic part	1.01E-49	2.80E-47	1.75
cellular nitrogen compound metabolic process	6.76E-33	1.29E-29	2.42	organic cyclic compound binding	1.35E-13	7.61E-11	1.57	intracellular organelle	6.18E-48	1.51E-45	1.52
macromolecule biosynthetic process	8.26E-33	1.41E-29	3.9	ribonucleoprotein complex binding	1.17E-12	5.88E-10	6.08	organelle	2.71E-45	5.88E-43	1.49
organonitrogen compound biosynthetic process	1.35E-32	2.07E-29	4.11	unfolded protein binding	1.21E-12	5.46E-10	7.6	cytosolic part	1.69E-42	3.29E-40	8.51
cellular metabolic process	2.36E-31	3.29E-28	1.77	protein-containing complex binding	1.24E-12	5.10E-10	2.36	ribosomal subunit	5.78E-42	1.03E-39	9.84
metabolic process	9.57E-29	1.22E-25	1.65	enzyme binding	1.31E-11	4.92E-09	1.86	intracellular membrane-bounded organelle	3.11E-36	5.06E-34	1.55
cellular biosynthetic process	5.82E-25	6.86E-22	2.64	translation factor activity, RNA binding	2.61E-10	9.08E-08	7.52	cell part	1.52E-35	2.29E-33	1.25
macromolecule metabolic process	9.04E-25	9.90E-22	1.8	translation initiation factor activity	3.30E-10	1.07E-07	9.88	membrane-bounded organelle	2.79E-32	3.89E-30	1.48
nitrogen compound metabolic process	2.85E-24	2.92E-21	1.71	threonine-type endopeptidase activity	2.15E-09	6.49E-07	15.3	cytosolic large ribosomal subunit	1.96E-30	2.54E-28	14.73
cellular macromolecule metabolic process	8.23E-24	7.89E-21	1.95	threonine-type peptidase activity	2.15E-09	6.08E-07	15.3	mitochondrial part	4.10E-27	4.99E-25	3.86
cellular process	2.47E-23	2.23E-20	1.3	ubiquitin-like protein ligase binding	3.28E-09	8.73E-07	3.48	intracellular non-membrane-bounded organelle	4.74E-27	5.44E-25	2.23
organic substance biosynthetic process	1.57E-22	1.34E-19	2.48	ubiquitin protein ligase binding	3.87E-09	9.71E-07	3.55	non-membrane-bounded organelle	1.20E-26	1.30E-24	2.21
biosynthetic process	8.40E-22	6.78E-19	2.42	protein tag	3.88E-09	9.24E-07	22.72	inner mitochondrial membrane protein complex	8.28E-26	8.51E-24	10.65
organic substance metabolic process	1.45E-20	1.11E-17	1.58	NADH dehydrogenase (ubiquinone) activity	2.06E-08	4.67E-06	15.04	large ribosomal subunit	3.94E-25	3.84E-23	9.38
cellular protein metabolic process	1.61E-20	1.18E-17	2.14	NADH dehydrogenase (quinone) activity	2.06E-08	4.44E-06	15.04	mitochondrial membrane part	3.99E-24	3.71E-22	7.39
primary metabolic process	1.95E-19	1.36E-16	1.59	electron transfer activity	2.95E-08	6.07E-06	7.03	mitochondrial protein complex	2.59E-23	2.29E-21	6.47
protein metabolic process	5.00E-19	3.33E-16	1.91	large ribosomal subunit rRNA binding	3.66E-08	7.20E-06	23.81	catalytic complex	8.49E-23	7.20E-21	2.88
organonitrogen compound metabolic process	1.58E-17	1.01E-14	1.74	NADH dehydrogenase activity	8.11E-08	1.53E-05	12.98	mitochondrial inner membrane	9.00E-23	7.32E-21	5.09
cellular protein-containing complex assembly	2.33E-17	1.43E-14	3.36	proton transmembrane transporter activity	9.07E-08	1.64E-05	5.1	mitochondrial membrane	1.14E-21	8.88E-20	4.14

9.4.3 Hippocampus Zhang

Process	P-value	FDR q-value	Enrichment	Function	P-value	FDR q-value	Enrichment	Component	P-value	FDR q-value	Enrichmer
cellular metabolic process	4.29E-26	6.57E-22	1.71	binding	3.49E-11	1.58E-07	1.21	intracellular organelle part	4.53E-34	8.79E-31	1.73
macromolecule metabolic process	9.43E-25	7.23E-21	1.82	ubiquitin-like protein binding	2.11E-09	4.76E-06	6.59	intracellular part	7.28E-33	7.05E-30	1.33
nitrogen compound metabolic process	9.06E-24	4.63E-20	1.72	organic cyclic compound binding	9.16E-09	1.38E-05	1.44	organelle part	2.26E-31	1.46E-28	1.67
metabolic process	2.89E-22	1.11E-18	1.58	RNA binding	1.20E-08	1.36E-05	2.19	membrane-bounded organelle	4.95E-29	2.40E-26	1.46
primary metabolic process	9.90E-22	3.03E-18	1.64	heterocyclic compound binding	2.41E-08	2.18E-05	1.43	intracellular membrane-bounded organelle	1.06E-28	4.11E-26	1.5
organic substance metabolic process	3.71E-21	9.47E-18	1.6	ubiquitin binding	3.33E-08	2.51E-05	6.98	protein-containing complex	7.75E-27	2.50E-24	1.81
cellular macromolecule metabolic process	3.34E-18	7.30E-15	1.83	nucleoside phosphate binding	1.30E-07	8.36E-05	1.73	intracellular organelle	4.96E-25	1.37E-22	1.39
protein metabolic process	2.43E-17	4.66E-14	1.88	nucleotide binding	1.30E-07	7.31E-05	1.73	organelle	5.04E-24	1.22E-21	1.37
macromolecule catabolic process	6.09E-17	1.04E-13	3.25	catalytic activity	1.57E-07	7.86E-05	1.37	nuclear part	6.27E-22	1.35E-19	1.92
proteolysis involved in cellular protein catabolic process	6.16E-17	9.44E-14	3.85	small molecule binding	6.79E-07	3.06E-04	1.62	cytoplasmic part	1.15E-20	2.23E-18	1.49
modification-dependent macromolecule catabolic process	6.21E-17	8.65E-14	4.05	ubiquitin-like protein ligase binding	1.07E-06	4.39E-04	3.01	catalytic complex	1.09E-18	1.92E-16	2.7
modification-dependent protein catabolic process	1.33E-16	1.70E-13	4.05	ubiquitin protein ligase binding	1.32E-06	4.98E-04	3.06	cell part	3.05E-18	4.93E-16	1.19
ubiquitin-dependent protein catabolic process	2.17E-16	2.56E-13	4.08	protein binding	2.50E-06	8.67E-04	1.23	proteasome accessory complex	5.27E-14	7.85E-12	23.99
proteolysis	5.06E-16	5.54E-13	2.81	modification-dependent protein binding	2.81E-06	9.05E-04	4.21	mitochondrial part	1.66E-13	2.30E-11	2.87
cellular macromolecule catabolic process	1.55E-15	1.58E-12	3.36	purine ribonucleoside triphosphate binding	3.49E-06	1.05E-03	1.7	mitochondrial protein complex	6.75E-13	8.72E-11	4.73
organonitrogen compound metabolic process	2.75E-15	2.64E-12	1.7	ribonucleotide binding	5.34E-06	1.51E-03	1.67	mitochondrion	8.14E-13	9.86E-11	2.14
cellular localization	7.20E-15	6.49E-12	2.19	purine ribonucleotide binding	7.39E-06	1.96E-03	1.66	peptidase complex	8.96E-13	1.02E-10	8.29
macromolecule localization	1.62E-13	1.38E-10	2.14	proteasome-activating ATPase activity	7.51E-06	1.88E-03	24.72	nucleoplasm	1.82E-12	1.96E-10	2.02
protein catabolic process	4.58E-13	3.69E-10	3.88	nucleoside-triphosphatase activity	7.82E-06	1.86E-03	2.04	nucleus	2.25E-12	2.29E-10	1.48
intracellular transport	6.91E-13	5.29E-10	2.44	purine nucleotide binding	9.30E-06	2.10E-03	1.65	proteasome regulatory particle	1.14E-11	1.10E-09	29.66
cellular protein metabolic process	2.40E-12	1.75E-09	1.86	pyrophosphatase activity	1.48E-05	3.17E-03	1.97	organelle membrane	2.10E-11	1.94E-09	2.17
protein localization	3.14E-12	2.19E-09	2.08	hydrolase activity, acting on acid anhydrides	1.61E-05	3.31E-03	1.96	proteasome complex	5.32E-11	4.68E-09	9.12
proteasomal protein catabolic process	1.48E-11	9.87E-09	4.09	as above, in phosphorus-containing anhydrides	1.61E-05	3.16E-03	1.96	endopeptidase complex	6.84E-11	5.77E-09	8.97
cellular nitrogen compound metabolic process	2.93E-11	1.87E-08	1.76	ubiquitin-specific protease binding	2.29E-05	4.31E-03	13.24	ribonucleoprotein complex	1.06E-10	8.58E-09	2.51
catabolic process	4.59E-11	2.82E-08	2.09	ATP binding	2.31E-05	4.18E-03	1.71	proteasome regulatory particle, base subcomplex	1.19E-10	9.24E-09	24.72

9.4.4 Hippocampus Zeisel

Process	P-value	FDR q-value	Enrichment	Function	P-value	FDR q-value	Enrichment	Component	P-value	FDR q-value	Enrichment
translation	1.55E-11	2.38E-07	15.09	RNA binding	1.23E-10	5.54E-07	6.32	protein-containing complex	8.90E-11	1.72E-07	2.57
peptide biosynthetic process	2.97E-11	2.28E-07	14.27	mRNA binding	2.64E-08	5.96E-05	12.94	ribonucleoprotein complex	1.07E-10	1.04E-07	6.95
cellular nitrogen compound biosynthetic process	4.20E-11	2.15E-07	6.22	heterocyclic compound binding	4.31E-08	6.48E-05	2.41	ribosome	2.25E-09	1.45E-06	17.16
cellular macromolecule biosynthetic process	4.91E-11	1.88E-07	7.31	nucleic acid binding	4.71E-08	5.31E-05	3.05	myelin sheath	5.16E-08	2.50E-05	15.09
macromolecule biosynthetic process	2.67E-10	8.18E-07	6.55	organic cyclic compound binding	6.48E-08	5.85E-05	2.37	intracellular part	6.16E-08	2.39E-05	1.44
amide biosynthetic process	5.65E-10	1.44E-06	11.05	translation factor activity, RNA binding	7.48E-08	5.63E-05	27.69	intracellular non-membrane-bounded organelle	1.67E-07	5.38E-05	2.98
peptide metabolic process	1.20E-09	2.62E-06	10.34	nucleoside-triphosphatase activity	1.99E-07	1.28E-04	5.84	non-membrane-bounded organelle	1.93E-07	5.36E-05	2.96
cellular biosynthetic process	1.44E-08	2.76E-05	4.13	pyrophosphatase activity	3.88E-07	2.19E-04	5.51	cytoplasmic part	3.73E-07	9.03E-05	1.82
organonitrogen compound biosynthetic process	4.11E-08	7.00E-05	6.03	hydrolase activity, acting on acid anhydrides	4.04E-07	2.03E-04	5.49	intracellular organelle	6.56E-07	1.41E-04	1.56
organic substance biosynthetic process	4.24E-08	6.50E-05	3.88	as above, in phosphorus-containing anhydrides	4.04E-07	1.83E-04	5.49	organelle	1.49E-06	2.90E-04	1.53
cellular nitrogen compound metabolic process	6.45E-08	8.98E-05	3.13	structural constituent of ribosome	4.50E-07	1.85E-04	14.7	intracellular organelle part	6.30E-06	1.11E-03	1.84
biosynthetic process	7.19E-08	9.18E-05	3.76	unfolded protein binding	7.01E-06	2.64E-03	18.66	organelle part	6.38E-06	1.03E-03	1.8
cellular amide metabolic process	9.75E-08	1.15E-04	6.96	purine ribonucleoside triphosphate binding	7.45E-06	2.59E-03	3.43	cytosolic part	1.26E-05	1.88E-03	8.9
nitrogen compound metabolic process	1.88E-06	2.06E-03	2.07	CTP binding	7.97E-06	2.57E-03	350.77	cell part	2.87E-05	3.97E-03	1.26
cellular macromolecule metabolic process	4.91E-06	5.01E-03	2.41	purine ribonucleotide binding	1.21E-05	3.65E-03	3.3	intracellular membrane-bounded organelle	2.87E-05	3.71E-03	1.58
protein localization	1.14E-05	1.09E-02	3.32	purine nucleotide binding	1.32E-05	3.72E-03	3.28	ribosomal subunit	2.99E-05	3.63E-03	9.93
cellular metabolic process	1.21E-05	1.09E-02	1.92	ribonucleotide binding	1.35E-05	3.58E-03	3.27	polysomal ribosome	6.81E-05	7.76E-03	37.58
macromolecule localization	1.33E-05	1.13E-02	3.28	translation elongation factor activity	1.44E-05	3.62E-03	61.9	cytosolic large ribosomal subunit	7.17E-05	7.72E-03	17.99
primary metabolic process	1.34E-05	1.08E-02	1.91	ATPase activity	2.18E-05	5.17E-03	6.73	smooth endoplasmic reticulum	9.28E-05	9.47E-03	33.95
establishment of localization in cell	1.54E-05	1.18E-02	3.67	sulfonylurea receptor binding	2.39E-05	5.39E-03	233.85	membrane-bounded organelle	2.24E-04	2.17E-02	1.47
macromolecule metabolic process	1.67E-05	1.22E-02	2.08	GTP binding	3.39E-05	7.28E-03	7.63	ribonucleoprotein granule	2.64E-04	2.44E-02	8.73
metabolic process	2.17E-05	1.51E-02	1.79	purine ribonucleoside binding	3.74E-05	7.67E-03	7.51	perinuclear region of cytoplasm	2.87E-04	2.53E-02	4.65
cellular protein metabolic process	2.34E-05	1.56E-02	2.72	purine nucleoside binding	4.04E-05	7.92E-03	7.42	mitochondrial proton-transporting ATP synthase complex, coupling factor F(o)	3.54E-04	2.98E-02	70.15
protein metabolic process	2.72E-05	1.74E-02	2.4	ribonucleoside binding	4.11E-05	7.74E-03	7.4	proton-transporting ATP synthase complex, coupling factor F(o)	4.31E-04	3.48E-02	63.78
organonitrogen compound metabolic process	3.29E-05	2.02E-02	2.17	rRNA binding	4.69E-05	8.46E-03	20.04	proteasome regulatory particle, base subcomplex	5.17E-04	4.00E-02	58.46

9.4.5 t(1;11) neurons Zhang

Process	P-value	FDR q-value	Enrichment	Function	P-value	FDR q-value	Enrichment	Component	P-value	FDR q-value	Enrichment
regulation of mRNA metabolic process	1.47E-15	2.27E-11	10.09	RNA binding	1.56E-12	7.09E-09	3.44	protein-containing complex	3.83E-16	7.47E-13	2.22
mRNA processing	1.59E-14	1.22E-10	8.32	protein binding	1.01E-09	2.30E-06	1.41	catalytic complex	4.59E-12	4.48E-09	3.92
RNA splicing	7.83E-14	4.02E-10	8.91	threonine-type endopeptidase activity	1.87E-07	2.84E-04	36.88	intracellular organelle part	5.17E-11	3.37E-08	1.57
non-canonical Wnt signaling pathway	3.52E-12	1.35E-08	14.98	threonine-type peptidase activity	1.87E-07	2.13E-04	36.88	proteasome complex	2.08E-10	1.02E-07	22.5
regulation of cellular amino acid metabolic process	7.81E-12	2.40E-08	24.59	heterocyclic compound binding	4.37E-07	3.97E-04	1.69	endopeptidase complex	2.44E-10	9.51E-08	22.13
mRNA metabolic process	6.62E-11	1.70E-07	5.47	organic cyclic compound binding	7.43E-07	5.62E-04	1.67	organelle part	4.19E-10	1.36E-07	1.52
regulation of cellular amine metabolic process	1.18E-10	2.59E-07	18.91	nucleic acid binding	1.42E-06	9.22E-04	1.88	nuclear speck	4.74E-10	1.32E-07	6.8
regulation of transcription from RNA polymerase II promoter in response to stress	1.88E-10	3.61E-07	14.89	cytochrome-c oxidase activity	1.10E-05	6.23E-03	28.1	nucleoplasm	1.10E-09	2.69E-07	2.37
negative regulation of G2/M transition of mitotic cell cycle	2.81E-10	4.81E-07	17.36	oxidoreductase activity, acting on a heme group of donors, oxygen as acceptor	1.10E-05	5.54E-03	28.1	intracellular part	2.73E-09	5.98E-07	1.25
regulation of DNA-templated transcription in response to stress	3.06E-10	4.71E-07	14.23	heme-copper terminal oxidase activity	1.10E-05	4.98E-03	28.1	peptidase complex	3.82E-09	7.45E-07	16.39
regulation of G2/M transition of mitotic cell cycle	4.28E-10	5.98E-07	10.26	oxidoreductase activity, acting on a heme group of donors	1.33E-05	5.51E-03	26.82	nuclear part	3.89E-09	6.91E-07	1.97
regulation of mRNA catabolic process	4.57E-10	5.86E-07	10.2	enzyme binding	9.21E-05	3.49E-02	2.07	ribonucleoprotein complex	8.64E-09	1.41E-06	3.59
Wnt signaling pathway, planar cell polarity pathway	5.59E-10	6.61E-07	16.21	ATPase activity, coupled	1.11E-04	3.89E-02	4.31	cytosol	1.15E-08	1.73E-06	1.92
regulation of hematopoietic stem cell differentiation	6.77E-10	7.45E-07	19.82	proton transmembrane transporter activity	1.44E-04	4.69E-02	7.57	spliceosomal complex	1.38E-08	1.93E-06	9.96
NIK/NF-kappaB signaling	6.77E-10	6.95E-07	19.82	molecular_function	2.46E-04	7.45E-02	1.1	respiratory chain complex	1.63E-08	2.12E-06	17.61
positive regulation of Wnt signaling pathway	9.38E-10	9.02E-07	11	copper chaperone activity	2.71E-04	7.69E-02	73.76	membrane-bounded organelle	2.94E-08	3.59E-06	1.46
Fc receptor signaling pathway	9.38E-10	8.49E-07	11	ATP binding	3.77E-04	1.01E-01	2.24	nuclear body	7.22E-08	8.29E-06	4.1
negative regulation of cell cycle G2/M phase transition	9.53E-10	8.15E-07	15.37	cuprous ion binding	4.49E-04	1.13E-01	59.01	nucleoplasm part	8.89E-08	9.64E-06	3.42
antigen processing and presentation of exogenous peptide antigen via MHC class I, TAP-dependent	1.15E-09	9.34E-07	18.7	electron transfer activity	5.92E-04	1.41E-01	7.38	catalytic step 2 spliceosome	9.90E-08	1.02E-05	14.05
regulation of cell cycle G2/M phase transition	1.18E-09	9.05E-07	9.45	ATPase activity	6.17E-04	1.40E-01	3.48	intracellular organelle	1.27E-07	1.24E-05	1.43
positive regulation of canonical Wnt signaling pathway	1.37E-09	1.00E-06	12.39	adenyl ribonucleotide binding	6.30E-04	1.36E-01	2.15	proteasome core complex	1.41E-07	1.31E-05	38.82
regulation of transcription from RNA polymerase II promoter in response to hypoxia	1.68E-09	1.18E-06	17.94	mRNA binding	6.30E-04	1.30E-01	3.79	nucleus	2.85E-07	2.53E-05	1.77
antigen processing and presentation of exogenous peptide antigen via MHC class I	1.90E-09	1.27E-06	17.7	nucleotide binding	6.55E-04	1.29E-01	1.94	organelle	4.34E-07	3.68E-05	1.35
Fc-epsilon receptor signaling pathway	1.92E-09	1.23E-06	14.32	nucleoside phosphate binding	6.60E-04	1.25E-01	1.93	intracellular membrane-bounded organelle	4.55E-07	3.70E-05	1.52
regulation of establishment of planar polarity	2.11E-09	1.30E-06	14.18	proteasome-activating ATPase activity	6.71E-04	1.22E-01	49.17	cytochrome complex	5.03E-07	3.93E-05	30.73

9.4.6 t(1;11) neurons Zeisel

Process	P-value	FDR q-value	Enrichment	Function	P-value	FDR q-value	Enrichment	Component	P-value	FDR q-valı	Enrichment
translational initiation	4.69E-38	7.22E-34	13.48	RNA binding	8.50E-50	3.86E-46	3.6	protein-containing complex	1.19E-53	2.31E-50	2.16
mRNA metabolic process	1.94E-37	1.49E-33	5.36	structural constituent of ribosome	2.66E-29	6.03E-26	10.21	intracellular organelle part	5.14E-45	5.02E-42	1.61
protein targeting to ER	1.87E-33	9.60E-30	14.46	nucleic acid binding	2.09E-19	3.16E-16	1.83	organelle part	7.37E-42	4.80E-39	1.57
establishment of protein localization to endoplasmic reticulum	1.07E-32	4.11E-29	13.89	protein binding	1.69E-17	1.92E-14	1.29	ribonucleoprotein complex	5.02E-35	2.45E-32	3.84
protein localization to endoplasmic reticulum	2.80E-32	8.62E-29	12.99	structural molecule activity	2.39E-16	2.18E-13	3.31	cytoplasmic part	6.54E-34	2.55E-31	1.53
cotranslational protein targeting to membrane	5.00E-31	1.28E-27	14.63	heterocyclic compound binding	3.34E-15	2.53E-12	1.54	intracellular part	1.68E-28	5.46E-26	1.23
SRP-dependent cotranslational protein targeting to membrane	1.27E-30	2.79E-27	15.04	NADH dehydrogenase (ubiquinone) activity	1.26E-14	8.20E-12	15.12	ribosomal subunit	1.77E-28	4.93E-26	9.09
intracellular transport	1.16E-28	2.23E-25	3.13	NADH dehydrogenase (quinone) activity	1.26E-14	7.18E-12	15.12	cytosol	1.82E-28	4.44E-26	1.9
nuclear-transcribed mRNA catabolic process, nonsense-mediated decay	1.51E-28	2.58E-25	12.08	organic cyclic compound binding	1.69E-14	8.52E-12	1.52	cytosolic part	7.03E-26	1.53E-23	7.42
establishment of localization in cell	1.28E-27	1.97E-24	2.88	NADH dehydrogenase activity	2.01E-14	9.11E-12	14.73	catalytic complex	6.43E-25	1.26E-22	3.08
translation	1.52E-27	2.13E-24	8.65	oxidoreductase activity, acting on NAD(P)H, quinone or similar compound as acceptor	1.56E-13	6.46E-11	11.78	membrane-bounded organelle	3.10E-24	5.50E-22	1.43
cellular metabolic process	3.04E-27	3.90E-24	1.6	rRNA binding	2.54E-10	9.63E-08	9.09	ribosome	4.24E-23	6.90E-21	11.45
viral transcription	2.50E-26	2.96E-23	11.59	translation initiation factor activity	6.38E-10	2.23E-07	10.69	organelle	5.08E-23	7.63E-21	1.35
cellular macromolecule catabolic process	2.77E-26	3.05E-23	3.98	molecular_function	6.43E-10	2.09E-07	1.09	intracellular organelle	6.00E-22	8.37E-20	1.4
intracellular protein transport	5.09E-26	5.22E-23	3.95	threonine-type endopeptidase activity	6.78E-10	2.05E-07	17.23	extracellular exosome	2.73E-21	3.55E-19	2.42
peptide biosynthetic process	6.31E-26	6.07E-23	7.92	threonine-type peptidase activity	6.78E-10	1.92E-07	17.23	extracellular vesicle	6.24E-21	7.62E-19	2.39
nuclear-transcribed mRNA catabolic process	1.06E-25	9.63E-23	8.37	translation factor activity, RNA binding	7.69E-10	2.05E-07	7.66	extracellular organelle	6.52E-21	7.48E-19	2.39
protein targeting to membrane	1.70E-24	1.45E-21	9.44	mRNA binding	3.19E-09	8.06E-07	3.39	respiratory chain complex	4.05E-20	4.39E-18	13.15
peptide metabolic process	2.34E-24	1.89E-21	6	oxidoreductase activity, acting on NAD(P)H	3.48E-09	8.33E-07	6.38	vesicle	5.35E-20	5.49E-18	2.04
mRNA catabolic process	3.10E-24	2.39E-21	7.7	enzyme binding	9.51E-09	2.16E-06	1.78	intracellular membrane-bounded organelle	1.13E-18	1.10E-16	1.46
viral process	3.77E-24	2.76E-21	4.37	ubiquitin-like protein ligase binding	1.37E-08	2.96E-06	3.54	mitochondrial membrane	4.15E-17	3.86E-15	3.75
symbiont process	3.77E-24	2.63E-21	4.37	ubiquitin protein ligase binding	1.50E-08	3.10E-06	3.63	cytosolic large ribosomal subunit	9.91E-17	8.79E-15	12.99
macromolecule catabolic process	8.22E-24	5.50E-21	3.49	electron transfer activity	4.70E-08	9.29E-06	5.74	mitochondrial inner membrane	1.69E-16	1.44E-14	4.56
metabolic process	1.11E-23	7.14E-21	1.51	protein tag	1.27E-07	2.40E-05	20.89	nuclear part	2.64E-16	2.15E-14	1.67
protein targeting	1.55E-23	9.52E-21	6.04	binding	1.32E-07	2.41E-05	1.12	proteasome complex	3.03E-16	2.37E-14	12.33

9.4.7 t(1;11) neurons Darmanis

Process	P-value	FDR q-value	Enrichment		Function	P-value	FDR q-value	Enrichment		Component	P-value	FDR q-value	Enrichment
RNA splicing	8.03E-07	1.24E-02	17.08		RNA binding	1.54E-04	7.00E-01	4.51		intracellular organelle part	6.49E-06	1.27E-02	1.94
mRNA processing	2.09E-06	1.61E-02	14.5		pre-mRNA binding	7.70E-04	1.00E+00	48.33		organelle part	1.11E-05	1.09E-02	1.88
regulation of RNA splicing	7.93E-06	4.07E-02	30.16							intracellular organelle	1.97E-05	1.28E-02	1.82
mRNA metabolic process	2.30E-05	8.87E-02	9.54							catalytic step 2 spliceosome	9.06E-05	4.42E-02	33.66
establishment of Golgi localization	2.97E-05	9.13E-02	235.62							organelle	1.37E-04	5.35E-02	1.64
RNA splicing, via transesterification reactions	9.12E-05	2.34E-01	16.18							nucleoplasm part	4.97E-04	1.62E-01	5.47
with bulged adenosine as nucleophile										nuclear speck	5.29E-04	1.48E-01	10.22
mRNA splicing, via spliceosome	9.12E-05	2.01E-01	16.18							ribonucleoprotein complex	5.91E-04	1.44E-01	5.29
Golgi localization	9.60E-05	1.85E-01	134.64							spliceosomal complex	6.39E-04	1.39E-01	17.35
RNA splicing, via transesterification reactions	9.90E-05	1.69E-01	15.84							nuclear body	7.33E-04	1.43E-01	6.55
establishment of localization in cell	1.10E-04	1.70E-01	4.73							intracellular membrane-bounded organelle	9.19E-04	1.63E-01	1.86
regulation of mRNA splicing, via spliceosome	1.19E-04	1.66E-01	30.73										
cellular component assembly	1.75E-04	2.24E-01	3.83										
regulation of mRNA metabolic process	2.63E-04	3.12E-01	12.28										
negative regulation of RNA splicing	2.65E-04	2.92E-01	81.96										
Golgi organization	2.94E-04	3.02E-01	22.62										
intracellular transport	3.20E-04	3.07E-01	4.81										
regulation of mRNA processing	3.22E-04	2.92E-01	21.92										
cellular localization	5.41E-04	4.63E-01	3.77										
RNA processing	6.62E-04	5.36E-01	5.18										
regulation of viral process	9.23E-04	7.10E-01	15.28										
interleukin-12-mediated signaling pathway	9.36E-04	6.86E-01	43.84										

9.5 Full GO Process terms for cell types, as in EWCE analysis.

BIBLIOGRAPHY

Word template by Friedman and Morgan, 2014.
https://neuraldischarge.wordpress.com

1. World Health Organization. Preventing suicide. A global imperative. *CMAJ* **143**, 609–610 (2014).

2. Murray, C. J. L. *et al.* Disability-adjusted life years (DALYs) for 291 diseases and injuries in 21 regions, 1990-2010: A systematic analysis for the Global Burden of Disease Study 2010. *Lancet* **380**, 2197–2223 (2012).

3. Whiteford, H. A. *et al.* Global burden of disease attributable to mental and substance use disorders: Findings from the Global Burden of Disease Study 2010. *Lancet* **382**, 1575–1586 (2013).

4. Vigo, D., Thornicroft, G. & Atun, R. Estimating the true global burden of mental illness. *The Lancet Psychiatry* **3**, 171–178 (2016).

5. Fleishman, M. Issues in psychopharmacosocioeconomics. *Psychiatr. Serv.* **53(12) Dec**, US http://psychservices (2002).

6. McCrone, P., Dhanasiri, S., Patel, A., Knapp, M. & Lawton-Smith, S. Schizophrenic Disorders. *Paying Price Cost Ment. Heal. Care Engl. to 2026* 51–67 (2008). doi:10.1093/obo/9780199828340-0105

7. van Os, J. & Kapur, S. Schizophrenia. *Lancet* **374**, 635–645 (2009).

8. Freedman, R. *et al.* The Initial Field Trials of DSM-5: New Blooms and Old Thorns. *Am. J. Psychiatry* **170**, 1–5 (2013).

9. Tandon, R. *et al.* Definition and description of schizophrenia in the DSM-5. *Schizophr. Res.* **150**, 3–10 (2013).

10. Zipursky, R. B., Reilly, T. J. & Murray, R. M. The myth of schizophrenia as a progressive brain disease. *Schizophr. Bull.* **39**, 1363–1372 (2013).

11. Phillips, M. L. & Kupfer, D. J. Bipolar Disorder 2 - Bipolar disorder diagnosis: Challenges and future directions. *Lancet* **381**, 1663–1671 (2013).

12. Craddock, N. & Owen, M. J. The Kraepelinian dichotomy - Going, going... but still not gone. *Br. J. Psychiatry* **196**, 92–95 (2010).

13. Anderson, I. M., Haddad, P. M. & Scott, J. Clinical Review: Bipolar Disorder. *Br. Med. J.* **345**, 1–10 (2012).

14. Fava, M. & Kendler, K. S. Major depressive disorder. *Neuron* **28**, 335–341 (2000).

15. Eaton, W. W. *et al.* The burden of mental disorders. *Epidemiol. Rev.* **30**, 1–14 (2008).

16. Hjorthøj, C., Stürup, A. E., Mcgrath, J. J. & Nordentoft, M. Years of potential life lost and life expectancy in schizophrenia: a systematic review and meta-analysis. (2017). doi:10.1016/S2215-0366(17)30078-0

17. Power, R. A. *et al.* Fecundity of patients with schizophrenia, autism, bipolar disorder, depression, anorexia nervosa, or substance abuse vs their unaffected siblings. *Arch. Gen. Psychiatry* **70**, 22–30 (2013).

18. Tandon, R., Nasrallah, H. A. & Keshavan, M. S. Schizophrenia, ‘just the facts’ 4. Clinical features and conceptualization. *Schizophr. Res.* **110**, 1–23 (2009).

19. McGrath, J. *et al.* A systematic review of the incidence of schizophrenia: The distribution of rates and the influence of sex, urbanicity, migrant status and methodology. *BMC Med.* **2**, 13 (2004).

20. Fombonne, E. Epidemiology of pervasive developmental disorders. *Pediatr. Res.* **65**, 591–598 (2009).

21. Van Dongen, J. & Boomsma, D. I. The evolutionary paradox and the missing heritability of schizophrenia. *Am. J. Med. Genet. Part B Neuropsychiatr. Genet.* **162**, 122–136 (2013).

22. Crow, T. J. Schizophrenia as the price that Homo sapiens pays for language: A resolution of the central paradox in the origin of the species. *Brain Res. Rev.* **31**, 118–129 (2000).

23. Keller, M. C. & Miller, G. Resolving the paradox of common, harmful, heritable mental disorders: Which evolutionary genetic models work best? *Behav. Brain Sci.* **29**, 385–404 (2006).

24. Mitchell, K. J. The Genetics of Neurodevelopmental Disorders. *Genet. Neurodev. Disord.* **21**, 1–356 (2015).

25. Sullivan, P. F., Daly, M. J. & O'Donovan, M. Genetic architectures of psychiatric disorders: The emerging picture and its implications. *Nat. Rev. Genet.* **13**, 537–551 (2012).

26. Cardno, A. *et al.* Heritability Estimates for Psychotic Disorders. *Arch. Gen. Psychiatry* **56**, 162 (1999).

27. Lichtenstein, P. *et al.* Common genetic determinants of schizophrenia and bipolar disorder in Swedish families: a population-based study. *Lancet* **373**, 234–239 (2009).

28. Purcell, S. M. *et al.* Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. *Nature* **460**, 748–752 (2009).

29. Ripke, S. *et al.* Biological insights from 108 schizophrenia-associated genetic loci. *Nature* **511**, 421–427 (2014).

30. Mitchell, K. J. & Porteous, D. J. Rethinking the genetic architecture of schizophrenia. *Psychol. Med.* **41**, 19–32 (2011).

31. McClellan, J. M., Susser, E. & King, M. C. Schizophrenia: A common disease caused by multiple rare alleles. *Br. J. Psychiatry* **190**, 194–199 (2007).

32. Xu, B. *et al.* Strong association of de novo copy number mutations with sporadic schizophrenia. *Nat. Genet.* **40**, 880–885 (2008).

33. Sebat, J. *et al.* Strong association of de novo copy number mutations with autism. *Science (80-.).* **316**, 445–449 (2007).

34. Fromer, M. *et al.* De novo mutations in schizophrenia implicate synaptic networks. *Nature* **506**, 179–184 (2014).

35. Kirov, G. *et al.* De novo CNV analysis implicates specific abnormalities of postsynaptic signalling complexes in the pathogenesis of schizophrenia. *Mol. Psychiatry* **17**, 142–153

(2012).

36. Walsh, T. *et al.* Rare structural variants disrupt multiple genes in neurodevelopmental pathways in schizophrenia. *Science (80-.).* **320**, 539–543 (2008).

37. Malaspina, D. *et al.* Advancing Paternal Age and the Risk of Schizophrenia. *Arch. Gen. Psychiatry* **58**, 361 (2001).

38. Buckley, P. M. Advancing Paternal Age and Bipolar Disorder. [Miscellaneous]. *Year B. Psychiatry Appl. Ment. Heal. 2009;2009 Suppl. C326-327* **2009 Suppl**, 326–327 (2013).

39. Hultman, C. M., Sandin, S., Levine, S. Z., Lichtenstein, P. & Reichenberg, A. Advancing paternal age and risk of autism: New evidence from a population-based study and a meta-analysis of epidemiological studies. *Mol. Psychiatry* **16**, 1203–1212 (2011).

40. Kong, A. *et al.* Rate of de novo mutations and the importance of father-s age to disease risk. *Nature* **488**, 471–475 (2012).

41. Gratten, J. *et al.* Risk of psychiatric illness from advanced paternal age is not predominantly from de novo mutations. *Nat. Genet.* **48**, 718–724 (2016).

42. Schork, N. J., Murray, S. S., Frazer, K. A. & Topol, E. J. Common vs. rare allele hypotheses for complex diseases. *Curr. Opin. Genet. Dev.* **19**, 212–219 (2009).

43. Gibson, G. Rare and common variants: Twenty arguments. *Nat. Rev. Genet.* **13**, 135–145 (2012).

44. Sud, A., Kinnersley, B. & Houlston, R. S. Genome-wide association studies of cancer: Current insights and future perspectives. *Nature Reviews Cancer* **17**, 692–704 (2017).

45. Schmitt, A., Malchow, B., Hasan, A. & Falkai, P. The impact of environmental factors in severe psychiatric disorders. *Frontiers in Neuroscience* **8**, 19 (2014).

46. Marangoni, C., Hernandez, M. & Faedda, G. L. The role of environmental exposures as risk factors for bipolar disorder: A systematic review of longitudinal studies. *Journal of Affective Disorders* **193**, 165–174 (2016).

47. Tsuang, M. Schizophrenia: Genes and environment. *Biol. Psychiatry* **47**, 210–220 (2000).

48. Guilmatre, A. *et al.* Recurrent rearrangements in synaptic and neurodevelopmental genes and shared biologic pathways in schizophrenia, autism, and mental retardation. *Arch. Gen. Psychiatry* **66**, 947–956 (2009).

49. Smoller, J. W. *et al.* Identification of risk loci with shared effects on five major psychiatric disorders: A genome-wide analysis. *Lancet* **381**, 1371–1379 (2013).

50. Meyer, U., Feldon, J. & Dammann, O. Schizophrenia and autism: Both shared and disorder-specific pathogenesis via perinatal inflammation? *Pediatr. Res.* **69**, 26–33 (2011).

51. Murray, R. M. *et al.* A developmental model for similarities and dissimilarities between schizophrenia and bipolar disorder. *Schizophr. Res.* **71**, 405–416 (2004).

52. Weinberger, D. R. From neuropathology to neurodevelopment. *Lancet* **346**, 552–557 (1995).

53. Howes, O. D. & Kapur, S. The dopamine hypothesis of schizophrenia: Version III - The final common pathway. *Schizophr. Bull.* **35**, 549–562 (2009).

54. Cowen, P. J. & Browning, M. What has serotonin to do with depression? *World Psychiatry* **14**, 158–160 (2015).

55. Javitt, D. C. & Zukin, S. R. Recent advances in the phencyclidine model of schizophrenia. *Am. J. Psychiatry* **148**, 1301–8 (1991).

56. Chan, S. Y., Matthews, E. & Burnet, P. W. J. ON or OFF?: Modulating the N-Methyl-D-Aspartate Receptor in Major Depression. *Front. Mol. Neurosci.* **9**, 169 (2016).

57. Bramham, C. R. *et al.* The Arc of synaptic memory. *Experimental Brain Research* **200**, 125–140 (2010).

58. Südhof, T. C. Neuroligins and neurexins link synaptic function to cognitive disease. *Nature* **455**, 903–911 (2008).

59. Hall, J., Trent, S., Thomas, K. L., O'Donovan, M. C. & Owen, M. J. Genetic risk for schizophrenia: Convergence on synaptic pathways involved in plasticity. *Biol. Psychiatry* **77**, 52–58 (2015).

60. Pardiñas, A. F. *et al.* Common schizophrenia alleles are enriched in mutation-intolerant genes and maintained by background selection. *bioRxiv* 068593 (2016). doi:10.1101/068593

61. Stahl, E. *et al.* Genomewide association study identifies 30 loci associated with bipolar disorder. *bioRxiv* 173062 (2018). doi:10.1101/173062

62. Cassano, G. B. *et al.* L-type calcium channels and psychiatric disorders: A brief review. *Am. J. Med. Genet. Part B Neuropsychiatr. Genet.* **153B**, 1373–1390 (2010).

63. Anderson, S. M. *et al.* CaMKII: A biochemical bridge linking accumbens dopamine and glutamate systems in cocaine seeking. *Nat. Neurosci.* **11**, 344–353 (2008).

64. Moosmang, S., Lenhardt, P., Haider, N., Hofmann, F. & Wegener, J. W. Mouse models to study L-type calcium channel function. *Pharmacol. Ther.* **106**, 347–355 (2005).

65. Blackwood, D. H. *et al.* Schizophrenia and affective disorders--cosegregation with a translocation at chromosome 1q42 that directly disrupts brain-expressed genes: clinical and P300 findings in a family. *Am. J. Hum. Genet.* **69**, 428–33 (2001).

66. Millar, J. K. Disruption of two novel genes by a translocation co-segregating with schizophrenia. *Hum. Mol. Genet.* **9**, 1415–1423 (2000).

67. James, R. *et al.* Disrupted in Schizophrenia 1 (DISC1) is a multicompartmentalized protein that predominantly localizes to mitochondria. *Mol. Cell. Neurosci.* **26**, 112–122 (2004).

68. Ishizuka, K. *et al.* DISC1-dependent switch from progenitor proliferation to migration in the developing cortex. *Nature* **473**, 92–96 (2011).

69. Kirkpatrick, B., Cascella, N., Ozeki, Y., Sawa, A. & Roberts, R. C. DISC1 Immunoreactivity at the Light and Ultrastructural Level in the Human Neocortex. *J. Comp. Neurol* **497**, 436–450 (2006).

70. Malavasi, E. L. V. *et al.* DISC1 regulates N-methyl-D-aspartate receptor dynamics: abnormalities induced by a Disc1 mutation modelling a translocation linked to major mental illness. *Transl. Psychiatry* **8**, 184 (2018).

71. Schurov, I. L., Handford, E. J., Brandon, N. J. & Whiting, P. J. Expression of disrupted in Schizophrenia 1 (DISC1) protein in the adult and developing mouse brain indicates its role in neurodevelopment. *Mol. Psychiatry* **9**, 1100–1110 (2004).

72. Eykelenboom, J. E. *et al.* A t(1;11) translocation linked to schizophrenia and affective disorders gives rise to aberrant chimeric DISC1 transcripts that encode structurally altered, deleterious mitochondrial proteins. *Hum. Mol. Genet.* **21**, 3374–3386 (2012).

73. Camargo, L. M. *et al.* Disrupted in Schizophrenia 1 interactome: Evidence for the close connectivity of risk genes and a potential synaptic basis for schizophrenia. *Mol. Psychiatry* **12**, 74–86 (2007).

74. Millar, J. K. *et al.* Genetics: DISC1 and PDE4B are interacting genetic factors in schizoprenia that regulate cAMP signaling. *Science (80-.).* **310**, 1187–1191 (2005).

75. Kamiya, A., Sedlak, T. W. & Pletnikov, M. V. DISC1 pathway in brain development: Exploring therapeutic targets for major psychiatric disorders. *Frontiers in Psychiatry* **3**, 1–7 (2012).

76. Klein, P. S. & Melton, D. A. A molecular mechanism for the effect of lithium on development. *Proc. Natl. Acad. Sci.* **93**, 8455–8459 (1996).

77. Mao, Y. *et al.* Disrupted in Schizophrenia 1 Regulates Neuronal Progenitor Proliferation via Modulation of GSK3β/β-Catenin Signaling. *Cell* **136**, 1017–1031 (2009).

78. Thornton, G. K. & Woods, C. G. Primary microcephaly: do all roads lead to Rome? *Trends Genet.* **25**, 501–510 (2009).

79. Brandon, N. J. *et al.* Understanding the Role of DISC1 in Psychiatric Disease and during Normal Development. *J. Neurosci.* **29**, 12768–12775 (2009).

80. Mykytyn, K. *et al.* Identification of the gene that, when mutated, causes the human obesity syndrome BBS4. *Nat. Genet.* **28**, 188–191 (2001).

81. Ishizuka, K. *et al.* DISC1-dependent switch from progenitor proliferation to migration in the developing cortex. *Nature* **473**, 92–96 (2011).

82. Lambert de Rouvroit, C. & Goffinet, A. M. Neuronal migration. *Mech. Dev.* **105**, 47–56 (2001).

83. Hirotsune, S. *et al.* Graded reduction of Pafah1b1 (Lis1) activity results in neuronal migration defects and early embryonic lethality. *Nat. Genet.* **19**, 333–339 (1998).

84. Bradshaw, N. J. *et al.* PKA Phosphorylation of NDE1 Is DISC1/PDE4 Dependent and Modulates Its Interaction with LIS1 and NDEL1. *J. Neurosci.* **31**, 9043–9054 (2011).

85. Cedar, H., Kandel, E. R. & Schwartz, J. H. Cyclic adenosine monophosphate in the nervous system of Aplysia californica. I. Increased synthesis in response to synaptic stimulation. *J. Gen. Physiol.* **60**, 558–69 (1972).

86. Mancuso, M. *et al.* Autosomal dominant psychiatric disorders and mitochondrial DNA multiple deletions: Report of a family. *J. Affect. Disord.* **106**, 173–177 (2008).

87. Campos, Y. *et al.* Mitochondrial myopathy, cardiomyopathy and psychiatric illness in a Spanish family harbouring the mtDNA 3303C>T mutation. *J. Inherit. Metab. Dis.* **24**, 685–687 (2001).

88. Clay, H. B., Sillivan, S. & Konradi, C. Mitochondrial dysfunction and pathology in bipolar disorder and schizophrenia. *Int. J. Dev. Neurosci.* **29**, 311–324 (2011).

89. Murphy, L. C. & Millar, J. K. Regulation of mitochondrial dynamics by DISC1, a putative risk factor for major mental illness. *Schizophrenia Research* **187**, 55–61 (2017).

90. Atkin, T. A., MacAskill, A. F., Brandon, N. J. & Kittler, J. T. Disrupted in Schizophrenia-1 regulates intracellular trafficking of mitochondria in neurons. *Molecular Psychiatry* **16**, 122–124 (2011).

91. Lipsky, R. H. *et al.* Disrupted in Schizophrenia 1 (DISC1): Association with Schizophrenia, Schizoaffective Disorder, and Bipolar Disorder. *Am. J. Hum. Genet.* **75**, 862–872 (2004).

92. Reis, K., Fransson, Å. & Aspenström, P. The Miro GTPases: At the heart of the mitochondrial transport machinery. *FEBS Lett.* **583**, 1391–1398 (2009).

93. Ogawa, F. *et al.* DISC1 complexes with TRAK1 and Miro1 to modulate anterograde axonal mitochondrial trafficking. *Hum. Mol. Genet.* **23**, 906–919 (2014).

94. Taylor, M. S., Devon, R. S., Millar, J. K. & Porteous, D. J. Evolutionary constraints on the Disrupted in Schizophrenia locus. *Genomics* **81**, 67–77 (2003).

95. Ogawa, F. *et al.* NDE1 and GSK3β Associate with TRAK1 and Regulate Axonal Mitochondrial Motility: Identification of Cyclic AMP as a Novel Modulator of Axonal Mitochondrial Trafficking. *ACS Chem. Neurosci.* **7**, 553–564 (2016).

96. Murphy, L. C. & Millar, J. K. Regulation of mitochondrial dynamics by DISC1, a putative risk factor for major mental illness. *Schizophr. Res.* **187**, 55–61 (2017).

97. Pandey, J. P. & Smith, D. S. A Cdk5-Dependent Switch Regulates Lis1/Ndel1/Dynein-Driven Organelle Transport in Adult Axons. *J. Neurosci.* **31**, 17207–17219 (2011).

98. Wenderski, W. *et al.* ERK regulation of phosphodiesterase 4 enhances dopamine-stimulated AMPA receptor membrane insertion. *Proc. Natl. Acad. Sci.* **110**, 15437–15442 (2013).

99. Sachs, N. A. *et al.* A frameshift mutation in Disrupted in Schizophrenia 1 in an American family with schizophrenia and schizoaffective disorder. *Mol. Psychiatry* **10**, 758–764 (2005).

100. Green, E. K. *et al.* Evidence that a DISC1 frame-shift deletion associated with psychosis in a single family may not be a pathogenic mutation [1]. *Mol. Psychiatry* **11**, 798–799 (2006).

101. Crowley, J. J. *et al.* Deep resequencing and association analysis of schizophrenia candidate

genes. *Molecular Psychiatry* **18**, 138–140 (2013).

102. Kockelkorn, T. T. J. P. *et al.* Association study of polymorphisms in the 5′ upstream region of human DISC1 gene with schizophrenia. *Neurosci. Lett.* **368**, 41–45 (2004).

103. Devon, R. S. *et al.* Identification of polymorphisms within Disrupted in Schizophrenia 1 and Disrupted in Schizophrenia 2, and an investigation of their association with schizophrenia and bipolar affective disorder. *Psychiatr. Genet.* **11**, 71–78 (2001).

104. Song, W. *et al.* Identification of high risk DISC1 structural variants with a 2% attributable risk for schizophrenia. *Biochem. Biophys. Res. Commun.* **367**, 700–706 (2008).

105. Thomson, P. A. *et al.* 708 Common and 2010 rare DISC1 locus variants identified in 1542 subjects: Analysis for association with psychiatric disorder and cognitive traits. *Mol. Psychiatry* **19**, 668–675 (2014).

106. Hodgkinson, C. A. *et al.* Disrupted in schizophrenia 1 (DISC1): association with schizophrenia, schizoaffective disorder, and bipolar disorder. *Am J Hum Genet* **75**, 862–872 (2004).

107. Malavasi, E. L. V., Ogawa, F., Porteous, D. J. & Millar, J. K. DISC1 variants 37W and 607F disrupt its nuclear targeting and regulatory role in ATF4-mediated transcription. *Hum. Mol. Genet.* **21**, 2779–2792 (2012).

108. Clapcote, S. J. *et al.* Behavioral Phenotypes of Disc1 Missense Mutations in Mice. *Neuron* **54**, 387–402 (2007).

109. Lipina, T. V., Zai, C., Hlousek, D., Roder, J. C. & Wong, A. H. C. Maternal Immune Activation during Gestation Interacts with Disc1 Point Mutation to Exacerbate Schizophrenia-Related Behaviors in Mice. *J. Neurosci.* **33**, 7654–7666 (2013).

110. Lipina, T. V. *et al.* Enhanced dopamine function in DISC1-L100P mutant mice: implications for schizophrenia. *Genes, Brain Behav.* **9**, 777–789 (2010).

111. Kakuda, K. *et al.* A DISC1 point mutation promotes oligomerization and impairs information processing in a mouse model of schizophrenia. *J. Biochem.* **165**, 369–378 (2019).

112. Kvajo, M. *et al.* A mutation in mouse Disc1 that models a schizophrenia risk allele leads to specific alterations in neuronal architecture and cognition. *Proc. Natl. Acad. Sci.* **105**, 7076–7081 (2008).

113. Kamiya, A. *et al.* A schizophrenia-associated mutation of DISC1 perturbs cerebral cortex

development. *Nat. Cell Biol.* **7**, 1067–1078 (2005).

114. Tomoda, T., Sumitomo, A., Jaaro-Peled, H. & Sawa, A. Utility and validity of DISC1 mouse models in biological psychiatry. *Neuroscience* **321**, 99–107 (2016).

115. Jaaro-Peled, H. *et al.* Subcortical dopaminergic deficits in a DISC1 mutant model: A study in direct reference to human molecular brain imaging. *Hum. Mol. Genet.* **22**, 1574–1580 (2013).

116. Jiang, Z., Cowell, R. M. & Nakazawa, K. Convergence of genetic and environmental factors on parvalbumin-positive interneurons in schizophrenia. *Front. Behav. Neurosci.* **7**, 116 (2013).

117. Chung, D. W., Wills, Z. P., Fish, K. N. & Lewis, D. A. Developmental pruning of excitatory synaptic inputs to parvalbumin interneurons in monkey prefrontal cortex. *Proc. Natl. Acad. Sci.* **114**, E629–E637 (2017).

118. Abazyan, B. *et al.* Prenatal interaction of mutant DISC1 and immune activation produces adult psychopathology. *Biol. Psychiatry* **68**, 1172–1181 (2010).

119. Niwa, M. *et al.* Knockdown of DISC1 by In Utero Gene Transfer Disturbs Postnatal Dopaminergic Maturation in the Frontal Cortex and Leads to Adult Behavioral Deficits. *Neuron* **65**, 480–489 (2010).

120. Takahashi, K. *et al.* Induction of Pluripotent Stem Cells from Adult Human Fibroblasts by Defined Factors. *Cell* **131**, 861–872 (2007).

121. Hou, P. *et al.* Pluripotent stem cells induced from mouse somatic cells by small-molecule compounds. *Science* **341**, 651–4 (2013).

122. Yamanaka, S. Induced pluripotent stem cells: Past, present, and future. *Cell Stem Cell* **10**, 678–684 (2012).

123. Tiscornia, G., Vivas, E. L. & Belmonte, J. C. I. Diseases in a dish: Modeling human genetic disorders using induced pluripotent cells. *Nat. Med.* **17**, 1570–1576 (2011).

124. Brennand, K. J. *et al.* Modelling schizophrenia using human induced pluripotent stem cells. *Nature* **473**, 221–225 (2011).

125. Kobayashi, N. R. *et al.* Modelling disrupted-in-schizophrenia 1 loss of function in human neural progenitor cells: Tools for molecular studies of human neurodevelopment and neuropsychiatric disorders. *Mol. Psychiatry* **15**, 672–675 (2010).

126. McCartney, D. L. *et al.* Altered DNA methylation associated with a translocation linked to major mental illness. *npj Schizophr.* **4**, 5 (2018).

127. Woo, T. U. W. Neurobiology of schizophrenia onset. *Curr. Top. Behav. Neurosci.* **16**, 267–295 (2014).

128. Takahashi, J. *et al.* Differentiation-defective phenotypes revealed by large-scale analyses of human pluripotent stem cells. *Proc. Natl. Acad. Sci.* **110**, 20569–20574 (2013).

129. Dolmetsch, R. & Geschwind, D. H. The Human Brain in a Dish: The Promise of iPSC-Derived Neurons. *Cell* **145**, 831–834 (2011).

130. Gunhanlar, N. *et al.* A simplified protocol for differentiation of electrophysiologically mature neuronal networks from human induced pluripotent stem cells. *Mol. Psychiatry* **23**, 1336–1344 (2018).

131. Santostefano, K. E. *et al.* A practical guide to induced pluripotent stem cell research using patient samples. *Lab. Investig.* **95**, 4–13 (2015).

132. Wen, Z. *et al.* Synaptic dysregulation in a human iPS cell model of mental disorders. *Nature* **515**, 414–418 (2014).

133. Skene, N. G. *et al.* Genetic identification of brain cell types underlying schizophrenia. *Nat. Genet.* **50**, 825–833 (2018).

134. Mohammadi, S., Zuckerman, N., Goldsmith, A. & Grama, A. A Critical Survey of Deconvolution Methods for Separating Cell Types in Complex Tissues. *Proc. IEEE* **105**, 340–366 (2017).

135. Gong, T. & Szustakowski, J. D. DeconRNASeq: a statistical framework for deconvolution of heterogeneous tissue samples based on mRNA-Seq data. *Bioinformatics* **29**, 1083–1085 (2013).

136. Beaulieu, J.-M., Gainetdinov, R. R. & Caron, M. G. Akt/GSK3 Signaling in the Action of Psychotropic Drugs. *Annu. Rev. Pharmacol. Toxicol.* **49**, 327–347 (2009).

137. Hook, V. *et al.* Human iPSC neurons display activity-dependent neurotransmitter secretion: Aberrant catecholamine levels in schizophrenia neurons. *Stem Cell Reports* **3**, 531–538 (2014).

138. Srikanth, P. *et al.* Genomic DISC1 Disruption in hiPSCs Alters Wnt Signaling and Neural Cell Fate. *Cell Rep.* **12**, 1414–1429 (2015).

139. Marshall, C. R. *et al.* Contribution of copy number variants to schizophrenia from a genome-wide study of 41,321 subjects. *Nat. Genet.* **49**, 27–35 (2017).

140. Skene, N. G. *et al.* Genetic identification of brain cell types underlying schizophrenia. *Nat. Genet.* **50**, 825–833 (2018).

141. Okita, K. *et al.* A more efficient method to generate integration-free human iPS cells. *Nat. Methods* **8**, 409–412 (2011).

142. Chambers, S. M. *et al.* Highly efficient neural conversion of human ES and iPS cells by dual inhibition of SMAD signaling. *Nat. Biotechnol.* **27**, 275–280 (2009).

143. Bilican, B. *et al.* Physiological normoxia and absence of EGF is required for the long-term propagation of anterior neural precursors from human pluripotent cells. *PLoS One* **9**, (2014).

144. Dechiara, T. M. *et al.* Velocimouse: Fully es cell-derived f0-generation mice obtained from the injection of es cells into eight-cell-stage embryos. *Methods Mol. Biol.* **530**, 311–324 (2009).

145. Jiang, H., Lei, R., Ding, S.-W. & Zhu, S. Skewer: a fast and accurate adapter trimmer for next-generation sequencing paired-end reads. *BMC Bioinformatics* **15**, 182 (2014).

146. Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).

147. Trapnell, C. *et al.* Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat. Biotechnol.* **28**, 511–515 (2010).

148. Roberts, A., Trapnell, C., Donaghey, J., Rinn, J. L. & Pachter, L. Improving RNA-Seq expression estimates by correcting for fragment bias. *Genome Biol.* **12**, R22 (2011).

149. Trapnell, C. *et al.* Differential analysis of gene regulation at transcript resolution with RNA-seq. *Nat. Biotechnol.* **31**, 46–53 (2013).

150. Andrews, S. FastQC - A quality control tool for high throughput sequence data. http://www.bioinformatics.babraham.ac.uk/projects/fastqc/. *2010* (2010). Available at: https://www.bioinformatics.babraham.ac.uk/projects/fastqc/.

151. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15**, 550 (2014).

152. Anders, S., Reyes, A. & Huber, W. Detecting differential usage of exons from RNA-seq data.

Genome Res. **22**, 2008–17 (2012).

153. Zhang, Y. *et al.* An RNA-Sequencing Transcriptome and Splicing Database of Glia, Neurons, and Vascular Cells of the Cerebral Cortex. *J. Neurosci.* **34**, 11929–11947 (2014).

154. Zhang, Y. *et al.* Purification and Characterization of Progenitor and Mature Human Astrocytes Reveals Transcriptional and Functional Differences with Mouse. *Neuron* **89**, 37–53 (2016).

155. Darmanis, S. *et al.* A survey of human brain transcriptome diversity at the single cell level. *Proc. Natl. Acad. Sci. U. S. A.* **112**, 7285–90 (2015).

156. Li, C.-L. *et al.* Somatosensory neuron types identified by high-coverage single-cell RNA-sequencing and functional heterogeneity. *Cell Res.* **26**, 83–102 (2016).

157. Zeisel, A. *et al.* Cell types in the mouse cortex and hippocampus revealed by single-cell RNA-seq. *Science (80-.).* **347**, 1138–1142 (2015).

158. Harewood, L. *et al.* The effect of translocation-induced nuclear reorganization on gene expression. *Genome Res.* **20**, 554–564 (2010).

159. Tiihonen, J. *et al.* Sex-specific transcriptional and proteomic signatures in schizophrenia. *Nat. Commun.* **10**, 1–11 (2019).

160. Costa-Silva, J., Domingues, D. & Lopes, F. M. RNA-Seq differential expression analysis: An extended review and a software tool. *PLoS One* **12**, e0190152 (2017).

161. Thorvaldsdottir, H., Robinson, J. T. & Mesirov, J. P. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Brief. Bioinform.* **14**, 178–192 (2013).

162. Ryan, N. M. *et al.* DNA sequence-level analyses reveal potential phenotypic modifiers in a large family with psychiatric disorders. *Mol. Psychiatry* **23**, 2254–2265 (2018).

163. Eden, E., Navon, R., Steinfeld, I., Lipson, D. & Yakhini, Z. GOrilla: a tool for discovery and visualization of enriched GO terms in ranked gene lists. *BMC Bioinformatics* **10**, 48 (2009).

164. Eden, E., Lipson, D., Yogev, S. & Yakhini, Z. Discovering Motifs in Ranked Lists of DNA Sequences. *PLoS Comput. Biol.* **3**, e39 (2007).

165. Greene, C. *et al.* Dose-dependent expression of claudin-5 is a modifying factor in schizophrenia. *Mol. Psychiatry* **23**, 2156–2166 (2018).

166. Wei, J., Graziane, N. M., Gu, Z. & Yan, Z. DISC1 Regulates GABAA Receptor Trafficking and Inhibitory Synaptic Transmission in Cortical Neurons. *J. Biol. Chem.* **290**, jbc.M115.656173 (2015).

167. Plump, A. S. *et al.* Slit1 and Slit2 Cooperate to Prevent Premature Midline Crossing of Retinal Axons in the Mouse Visual System. *Neuron* **33**, 219–232 (2002).

168. Südhof, T. C. Synaptotagmins: why so many? *J. Biol. Chem.* **277**, 7629–32 (2002).

169. Karran, E., Mercken, M. & Strooper, B. De. The amyloid cascade hypothesis for Alzheimer's disease: An appraisal for the development of therapeutics. *Nat. Rev. Drug Discov.* **10**, 698–712 (2011).

170. Hallermann, S. *et al.* Bassoon Speeds Vesicle Reloading at a Central Excitatory Synapse. *Neuron* **68**, 710–723 (2010).

171. Sjöblom, B., Salmazo, A. & Djinović-Carugo, K. α-Actinin structure and regulation. *Cellular and Molecular Life Sciences* **65**, 2688–2701 (2008).

172. Kinzfogl, J., Hangoc, G. & Broxmeyer, H. E. Neurexophilin 1 suppresses the proliferation of hematopoietic progenitor cells. *Blood* **118**, 565–575 (2011).

173. Coyle, J. T. NMDA receptor and schizophrenia: A brief history. *Schizophr. Bull.* **38**, 920–926 (2012).

174. Soto, D., Altafaj, X., Sindreu, C. & Bayés, A. Glutamate receptor mutations in psychiatric and neurodevelopmental disorders. *Commun. Integr. Biol.* **7**, e27887 (2014).

175. Beales, P. L. Lifting the lid on Pandora's box: The Bardet-Biedl syndrome. *Curr. Opin. Genet. Dev.* **15**, 315–323 (2005).

176. Thomson, P. A. *et al.* DISC1 genetics, biology and psychiatric illness. *Front. Biol. (Beijing).* **8**, 1–31 (2013).

177. Kojetin, D. J. *et al.* Structure, binding interface and hydrophobic transitions of Ca2+-loaded calbindin-D28K. *Nat. Struct. Mol. Biol.* **13**, 641–647 (2006).

178. Lledo, P. M., Somasundaram, B., Morton, A. J., Emson, P. C. & Mason, W. T. Stable transfection of calbindin-D28k into the GH3 cell line alters calcium currents and intracellular calcium homeostasis. *Neuron* **9**, 943–54 (1992).

179. Liang, C. L., Sinton, C. M. & German, D. C. Midbrain dopaminergic neurons in the mouse:

co-localization with Calbindin-D28K and calretinin. *Neuroscience* **75**, 523–33 (1996).

180. Choi, W. S., Lee, E., Lim, J. & Oh, Y. J. Calbindin-D28K prevents drug-induced dopaminergic neuronal death by inhibiting caspase and calpain activity. *Biochem. Biophys. Res. Commun.* **371**, 127–131 (2008).

181. Yu, W. & Greenberg, M. L. Inositol depletion, GSK3 inhibition and bipolar disorder. *Future Neurol.* **11**, 135–148 (2016).

182. Airaksinen, M. S., Thoenen, H. & Meyer, M. Vulnerability of Midbrain Dopaminergic Neurons in Calbindin-D $_{28k}$ -deficient Mice: Lack of Evidence for a Neuroprotective Role of Endogenous Calbindin in MPTPtreated and Weaver Mice. *Eur. J. Neurosci.* **9**, 120–127 (1997).

183. Steinlein, O. K. & Bertrand, D. Nicotinic receptor channelopathies and epilepsy. *Pflügers Arch. - Eur. J. Physiol.* **460**, 495–503 (2010).

184. Magnusson, A., Stordal, E., Brodtkorb, E. & Steinlein, O. Schizophrenia, psychotic illness and other psychiatric symptoms in families with autosomal dominant nocturnal frontal lobe epilepsy caused by different mutations. doi:10.1097/01.ypg.0000056173.32550.b0

185. Dani, J. A. & Bertrand, D. Nicotinic Acetylcholine Receptors and Nicotinic Cholinergic Mechanisms of the Central Nervous System. *Annu. Rev. Pharmacol. Toxicol* **47**, 699–729 (2007).

186. Ripoll, N., Bronnec, M. & Bourin, M. Nicotinic receptors and schizophrenia. *Curr. Med. Res. Opin.* **20**, 1057–1074 (2004).

187. Li, B., Woo, R. S., Mei, L. & Malinow, R. The Neuregulin-1 Receptor ErbB4 Controls Glutamatergic Synapse Maturation and Plasticity. *Neuron* **54**, 583–597 (2007).

188. Jaaro-Peled, H. *et al.* Neurodevelopmental mechanisms of schizophrenia: understanding disturbed postnatal brain maturation through neuregulin-1-ErbB4 and DISC1. *Trends Neurosci.* **32**, 485–495 (2009).

189. Gerlai, R., Pisacane, P. & Erickson, S. Heregulin, but not ErbB2 or ErbB3, heterozygous mutant mice exhibit hyperactivity in multiple behavioral tasks. *Behav. Brain Res.* **109**, 219–27 (2000).

190. Oikari, L. E. *et al.* Cell surface heparan sulfate proteoglycans as novel markers of human neural stem cell fate determination. *Stem Cell Res.* **16**, 92–104 (2016).

191. Jen, Y. H. L., Musacchio, M. & Lander, A. D. Glypican-1 controls brain size through regulation of fibroblast growth factor signaling in early neurogenesis. *Neural Dev.* **4**, 33 (2009).

192. Ronca, F., Andersen, J. S., Paech, V. & Margolis, R. U. Characterization of Slit protein interactions with glypican-1. *J. Biol. Chem.* **276**, 29141–7 (2001).

193. Rong, J., Li, S.-H. & Li, X.-J. Regulation of intracellular HAP1 trafficking. *J. Neurosci. Res.* **85**, 3025–3029 (2007).

194. Kittler, J. T. *et al.* Huntingtin-associated protein 1 regulates inhibitory synaptic transmission by modulating gamma-aminobutyric acid type A receptor membrane trafficking. *Proc. Natl. Acad. Sci. U. S. A.* **101**, 12736–41 (2004).

195. Gauthier, L. R. *et al.* Huntingtin Controls Neurotrophic Support and Survival of Neurons by Enhancing BDNF Vesicular Transport along Microtubules. *Cell* **118**, 127–138 (2004).

196. Sheng, G. *et al.* Huntingtin-associated protein 1 interacts with Ahi1 to regulate cerebellar and brainstem development in mice. *J. Clin. Invest.* **118**, 2785–2795 (2008).

197. Iyer, N. V *et al.* Cellular and developmental control of O2 homeostasis by hypoxia-inducible factor 1 alpha. *Genes Dev.* **12**, 149–62 (1998).

198. Demars, M. P. & Morishita, H. Cortical parvalbumin and somatostatin GABA neurons express distinct endogenous modulators of nicotinic acetylcholine receptors. *Mol. Brain* **7**, 75 (2014).

199. Schmidt-Kastner, R., van Os, J., W.M. Steinbusch, H. & Schmitz, C. Gene regulation by hypoxia and the neurodevelopmental origin of schizophrenia. *Schizophr. Res.* **84**, 253–271 (2006).

200. Nicodemus, K. K. *et al.* Serious obstetric complications interact with hypoxia-regulated/vascular-expression genes to influence schizophrenia risk. *Mol. Psychiatry* **13**, 873–877 (2008).

201. Koolen, D. A. *et al.* Mutations in the chromatin modifier gene KANSL1 cause the 17q21.31 microdeletion syndrome. *Nat. Genet.* **44**, 639–641 (2012).

202. Nishino, J. *et al.* Meteorin: A secreted protein that regulates glial cell differentiation and promotes axonal extension. *EMBO J.* **23**, 1998–2008 (2004).

203. Zhang, Y. *et al.* An RNA-Sequencing Transcriptome and Splicing Database of Glia, Neurons,

and Vascular Cells of the Cerebral Cortex. *J. Neurosci.* **34**, 11929–11947 (2014).

204. Marín, O. Interneuron dysfunction in psychiatric disorders. *Nat. Rev. Neurosci.* **13**, 107–120 (2012).

205. Brooks-Kayal, A. Epilepsy and autism spectrum disorders: Are there common developmental mechanisms? *Brain Dev.* **32**, 731–738 (2010).

206. Le, T. N. *et al.* Dlx homeobox genes promote cortical interneuron migration from the basal forebrain by direct repression of the semaphorin receptor neuropilin-2. *J. Biol. Chem.* **282**, 19071–81 (2007).

207. Eide, F. F. *et al.* Naturally occurring truncated trkB receptors have dominant inhibitory effects on brain-derived neurotrophic factor signaling. *J. Neurosci.* **16**, 3123–9 (1996).

208. Ohira, K. & Hayashi, M. A new aspect of the TrkB signaling pathway in neural plasticity. *Curr. Neuropharmacol.* **7**, 276–85 (2009).

209. Yacoubian, T. A. & Lo, D. C. Truncated and full-length TrkB receptors regulate distinct modes of dendritic growth. *Nat. Neurosci.* **3**, 342–349 (2000).

210. Kim, S. W. & Cho, K. J. Activity-dependent alterations in the sensitivity to BDNF-TrkB signaling may promote excessive dendritic arborization and spinogenesis in fragile X syndrome in order to compensate for compromised postsynaptic activity. *Med. Hypotheses* **83**, 429–435 (2014).

211. Quach, T. T., Honnorat, J., Kolattukudy, P. E., Khanna, R. & Duchemin, A. M. CRMPs: Critical molecules for neurite morphogenesis and neuropsychiatric diseases. *Molecular Psychiatry* **20**, 1037–1045 (2015).

212. Tejeda, H. A., Shippenberg, T. S. & Henriksson, R. The dynorphin/κ-opioid receptor system and its role in psychiatric disorders. *Cell. Mol. Life Sci.* **69**, 857–896 (2012).

213. Pfeiffer, A., Brantl, V., Herz, A. & Emrich, H. M. Psychotomimesis mediated by kappa opiate receptors. *Science* **233**, 774–6 (1986).

214. Volk, T. & Artzt, K. *Post-Transcriptional Regulation by STAR proteins*. (2010). doi:10.1007/978-1-4419-7005-3

215. Noveroske, J. K., Hardy, R., Dapper, J. D., Vogel, H. & Justice, M. J. A new ENU-induced allele of mouse quaking causes severe CNS dysmyelination. *Mamm. Genome* **16**, 672–682 (2005).

216. Aberg, K., Saetre, P., Jareborg, N. & Jazin, E. Human QKI, a potential regulator of mRNA expression of human oligodendrocyte-related genes involved in schizophrenia. *Proc. Natl. Acad. Sci. U. S. A.* **103**, 7482–7 (2006).

217. Hardy, R. J. QKI expression is regulated during neuron-glial cell fate decisions. *J. Neurosci. Res.* **54**, 46–57 (1998).

218. Irie, M. *et al.* Binding of neuroligins to PSD-95. *Science (80-.).* **277**, 1511–1515 (1997).

219. Law, K. L. E. & Wong, K. Design and performance evaluation of active bandwidth brokers. *Ann. des Telecommun. Telecommun.* **59**, 88–107 (2004).

220. Frank, R. A. W. *et al.* NMDA receptors are selectively partitioned into complexes and supercomplexes during synapse maturation. *Nat. Commun.* **7**, 11264 (2016).

221. Dakoji, S., Tomita, S., Karimzadegan, S., Nicoll, R. A. & Bredt, D. S. Interaction of transmembrane AMPA receptor regulatory proteins with multiple membrane associated guanylate kinases. *Neuropharmacology* **45**, 849–856 (2003).

222. Carlisle, H. J., Fink, A. E., Grant, S. G. N. & O'dell, T. J. Opposing effects of PSD-93 and PSD-95 on long-term potentiation and spike timing-dependent plasticity. *J. Physiol.* **586**, 5885–5900 (2008).

223. Arai, M. & Itokawa, M. A hard road in psychiatric genetics: Schizophrenia and DPYSL2. *Journal of Human Genetics* **55**, 397–399 (2010).

224. Inagaki, N. *et al.* CRMP-2 induces axons in cultured hippocampal neurons. *Nat. Neurosci.* **4**, 781–782 (2001).

225. Morimura, R., Nozawa, K., Tanaka, H. & Ohshima, T. Phosphorylation of Dpsyl2 (CRMP2) and Dpsyl3 (CRMP4) is required for positioning of caudal primary motor neurons in the zebrafish spinal cord. *Dev. Neurobiol.* **73**, 911–920 (2013).

226. Kedracka-Krok, S. *et al.* Clozapine influences cytoskeleton structure and calcium homeostasis in rat cerebral cortex and has a different proteomic profile than risperidone. *J. Neurochem.* **132**, 657–676 (2015).

227. Kowara, R., Moraleja, K. L. & Chakravarthy, B. Involvement of nitric oxide synthase and ROS-mediated activation of L-type voltage-gated Ca2+ channels in NMDA-induced DPYSL3 degradation. *Brain Res.* **1119**, 40–49 (2006).

228. Quinn, C. C., Gray, G. E. & Hockfield, S. A family of proteins implicated in axon guidance

and outgrowth. *J. Neurobiol.* (1999). doi:10.1002/(SICI)1097-4695(199910)41:1<158::AID-NEU19>3.0.CO;2-0

229. Mlodzik, M. The Dishevelled Protein Family: Still Rather a Mystery After Over 20 Years of Molecular Studies. *Curr. Top. Dev. Biol.* **117**, 75–91 (2016).

230. Kessels, H. W. & Malinow, R. Synaptic AMPA Receptor Plasticity and Behavior. *Neuron* **61**, 340–350 (2009).

231. Naito, Y., Lee, A. K. & Takahashi, H. Emerging roles of the neurotrophin receptor TrkC in synapse organization. *Neuroscience Research* **116**, 10–17 (2017).

232. Dierssen, M. *et al.* Transgenic mice overexpressing the full-length neurotrophin receptor TrkC exhibit increased catecholaminergic neuron density in specific brain areas and increased anxiety-like behavior and panic reaction. *Neurobiol. Dis.* **24**, 403–418 (2006).

233. Toriyama, M. *et al.* Shootin1: A protein involved in the organization of an asymmetric signal for neuronal polarization. *J. Cell Biol.* **175**, 147–57 (2006).

234. Nawaz, M. S. *et al.* CDKL5 and Shootin1 Interact and Concur in Regulating Neuronal Polarization. *PLoS One* **11**, e0148634 (2016).

235. Kubo, Y. *et al.* Shootin1-cortactin interaction mediates signal-force transduction for axon outgrowth. *J. Cell Biol.* **210**, 663–676 (2015).

236. Ergin, V., Erdogan, M. & Menevse, A. Regulation of Shootin1 Gene Expression Involves NGF-induced Alternative Splicing during Neuronal Differentiation of PC12 Cells. *Sci. Rep.* **5**, 17931 (2015).

237. Dzhala, V. I. *et al.* NKCC1 transporter facilitates seizures in the developing brain. *Nat. Med.* **11**, 1205–1213 (2005).

238. Kahle, K. T. *et al.* Roles of the cation–chloride cotransporters in neurological disease. *Nat. Clin. Pract. Neurol.* **4**, 490–503 (2008).

239. Morita, Y. *et al.* Characteristics of the cation cotransporter NKCC1 in human brain: alternate transcripts, expression in development, and potential relationships to brain function and schizophrenia. *J. Neurosci.* **34**, 4929–40 (2014).

240. Kim, J. Y. *et al.* Interplay between DISC1 and GABA signaling regulates neurogenesis in mice and risk for schizophrenia. *Cell* **148**, 1051–1064 (2012).

241. Devine, M. J., Norkett, R. & Kittler, J. T. DISC1 is a coordinator of intracellular trafficking to shape neuronal development and connectivity. *Journal of Physiology* **594**, 5459–5469 (2016).

242. Wang, D.-C., Chen, S.-S., Lee, Y.-C. & Chen, T.-J. Amyloid-β at sublethal level impairs BDNF-induced arc expression in cortical neurons. *Neurosci. Lett.* **398**, 78–82 (2006).

243. Hashimoto, T. *et al.* Apolipoprotein E, especially apolipoprotein E4, increases the oligomerization of amyloid β peptide. *J. Neurosci.* **32**, 15181–92 (2012).

244. Puglielli, L., Tanzi, R. E. & Kovacs, D. M. Alzheimer's disease: the cholesterol connection. *Nat. Neurosci.* **6**, 345–351 (2003).

245. Nathan, B. P. *et al.* Apolipoprotein E4 inhibits, and apolipoprotein E3 promotes neurite outgrowth in cultured adult mouse cortical neurons through the low-density lipoprotein receptor-related protein. *Brain Res.* **928**, 96–105 (2002).

246. Winslow, J. T., Hastings, N., Carter, C. S., Harbaugh, C. R. & Insel, T. R. A role for central vasopressin in pair bonding in monogamous prairie voles. *Nature* **365**, 545–548 (1993).

247. Young, L. J. & Wang, Z. The neurobiology of pair bonding. *Nat. Neurosci.* **7**, 1048–1054 (2004).

248. HOFFMAN, P. L., RITZMANN, R. F., WALTER, R. & TABAKOFF, B. Arginine vasopressin maintains ethanol tolerance. *Nature* **276**, 614–616 (1978).

249. McAuliffe, J. J., Joseph, B., Hughes, E., Miles, L. & Vorhees, C. V. Metallothionein I,II deficient mice do not exhibit significantly worse long-term behavioral outcomes following neonatal hypoxia–ischemia: MT-I,II deficient mice have inherent behavioral impairments. *Brain Res.* **1190**, 175–185 (2008).

250. Qu, W. & Waalkes, M. P. Metallothionein blocks oxidative DNA damage induced by acute inorganic arsenic exposure. *Toxicol. Appl. Pharmacol.* **282**, 267–74 (2015).

251. Carrasco, J., Penkowa, M., Hadberg, H., Molinero, A. & Hidalgo, J. Enhanced seizures and hippocampal neurodegeneration following kainic acid-induced seizures in metallothionein-I + II-deficient mice. *Eur. J. Neurosci.* **12**, 2311–22 (2000).

252. Juárez-Rebollar, D., Rios, C., Nava-Ruíz, C. & Méndez-Armenta, M. Metallothionein in Brain Disorders. *Oxid. Med. Cell. Longev.* **2017**, 1–12 (2017).

253. Nieoullon, A. *et al.* The neuronal excitatory amino acid transporter EAAC1/EAAT3: does it represent a major actor at the brain excitatory synapse? *J. Neurochem.* **98**, 1007–1018 (2006).

254. Underhill, S. M. *et al.* Amphetamine Modulates Excitatory Neurotransmission through Endocytosis of the Glutamate Transporter EAAT3 in Dopamine Neurons. *Neuron* **83**, 404–416 (2014).

255. Mortensen, M., Patel, B. & Smart, T. G. GABA Potency at GABA(A) Receptors Found in Synaptic and Extrasynaptic Zones. *Front. Cell. Neurosci.* **6**, 1 (2011).

256. Zhang, R., Lahens, N. F., Ballance, H. I., Hughes, M. E. & Hogenesch, J. B. A circadian gene expression atlas in mammals: Implications for biology and medicine. *Proc. Natl. Acad. Sci. U. S. A.* **111**, 16219–16224 (2014).

257. Li, S. *et al.* CGDB: A database of circadian genes in eukaryotes. *Nucleic Acids Res.* **45**, D397–D403 (2017).

258. Lowrey, P. L. & Takahashi, J. S. Genetics of circadian rhythms in mammalian model organisms. in *Advances in Genetics* **74**, 175–230 (Academic Press Inc., 2011).

259. Nudel, R. & Newbury, D. F. FOXP2. *Wiley Interdiscip. Rev. Cogn. Sci.* **4**, 547–560 (2013).

260. Walker, R. M. *et al.* The DISC1 promoter: Characterization and regulation by FOXP2. *Hum. Mol. Genet.* **21**, 2862–2872 (2012).

261. Lencz, T. *et al.* Genome-wide association study implicates NDST3 in schizophrenia and bipolar disorder. *Nat. Commun.* **4**, 2739 (2013).

262. De Cat, B. & David, G. Developmental roles of the glypicans. *CELL Dev. Biol.* **12**, 117–125 (2001).

263. Sarrazin, S., Lamanna, W. C. & Esko, J. D. Heparan sulfate proteoglycans. *Cold Spring Harb. Perspect. Biol.* **3**, (2011).

264. Yamagishi, S. *et al.* Netrin-5 is highly expressed in neurogenic regions of the adult brain. *Front. Cell. Neurosci.* **9**, 146 (2015).

265. Garrett, A. M. *et al.* Analysis of Expression Pattern and Genetic Deletion of Netrin5 in the Developing Mouse. *Front. Mol. Neurosci.* **9**, 3 (2016).

266. Avila Cobos, F., Vandesompele, J., Mestdagh, P. & De Preter, K. Computational deconvolution of transcriptomics data from mixed cell populations. *Bioinformatics* **34**, 1969–1979 (2018).

267. Gong, T. *et al.* Optimal Deconvolution of Transcriptional Profiling Data Using Quadratic

Programming with Application to Complex Clinical Blood Samples. *PLoS One* **6**, e27156 (2011).

268. Seshadri, S. *et al.* Disrupted-in-Schizophrenia-1 expression is regulated by -site amyloid precursor protein cleaving enzyme-1-neuregulin cascade. *Proc. Natl. Acad. Sci.* **107**, 5622–5627 (2010).

269. Tarazona, S., García-Alcalde, F., Dopazo, J., Ferrer, A. & Conesa, A. Differential expression in RNA-seq: A matter of depth. *Genome Res.* **21**, 2213–2223 (2011).

270. Mortazavi, A., Williams, B. A., McCue, K., Schaeffer, L. & Wold, B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat. Methods* **5**, 621–628 (2008).

271. Zhang, Y. *et al.* Purification and Characterization of Progenitor and Mature Human Astrocytes Reveals Transcriptional and Functional Differences with Mouse. *Neuron* **89**, 37–53 (2016).

272. Jin, H., Wan, Y. W. & Liu, Z. Comprehensive evaluation of RNA-seq quantification methods for linearity. *BMC Bioinformatics* **18**, 117 (2017).

273. Keller, D., Erö, C. & Markram, H. Cell Densities in the Mouse Brain: A Systematic Review. *Front. Neuroanat.* **12**, 83 (2018).

274. Mota, B. & Herculano-Houzel, S. All brains are made of this: a fundamental building block of brain matter with matching neuronal and glial masses. *Front. Neuroanat.* **8**, 127 (2014).

275. Mariappan, M. *et al.* A ribosome-associating factor chaperones tail-anchored membrane proteins. *Nature* **466**, 1120–1124 (2010).

276. Meylan, E. M. *et al.* Involvement of the agmatinergic system in the depressive-like phenotype of the Crtc1 knockout mouse model of depression. *Transl. Psychiatry* **6**, e852–e852 (2016).

277. Rangaraju, V., tom Dieck, S. & Schuman, E. M. Local translation in neuronal compartments: how local is local? *EMBO Rep.* **18**, 693–711 (2017).

278. Li, C.-L. *et al.* Somatosensory neuron types identified by high-coverage single-cell RNA-sequencing and functional heterogeneity. *Cell Res.* **26**, 83–102 (2016).

279. Mullen, R. J., Buck, C. R. & Smith, A. M. NeuN, a neuronal specific nuclear protein in vertebrates. *Development* **116**, 201–11 (1992).

280. Sarnat, H. B., Nochlin, D. & Born, D. E. Neuronal nuclear antigen (NeuN): a marker of

neuronal maturation in early human fetal nervous system. *Brain Dev.* **20**, 88–94 (1998).

281. Skene, N. G. & Grant, S. G. N. Identification of Vulnerable Cell Types in Major Brain Disorders Using Single Cell Transcriptomes and Expression Weighted Cell Type Enrichment. *Front. Neurosci.* **10**, 16 (2016).

282. Lovinger, D. M. New twist on orphan receptor GPR88 function. *Nat. Neurosci.* **15**, 1469–1470 (2012).

283. Iqbal, Z. *et al.* Homozygous SLC6A17 mutations cause autosomal-recessive intellectual disability with progressive tremor, speech impairment, and behavioral problems. *Am. J. Hum. Genet.* **96**, 386–396 (2015).

284. Pitts, M. W. *et al.* Competition between the Brain and Testes under Selenium-Compromised Conditions: Insight into Sex Differences in Selenium Metabolism and Risk of Neurodevelopmental Disease. *J. Neurosci.* **35**, 15326–15338 (2015).

285. Tozzi, A. *et al.* Endogenous 17β-estradiol is required for activity-dependent long-term potentiation in the striatum: interaction with the dopaminergic system. *Front. Cell. Neurosci.* **9**, 192 (2015).

286. Atasoy, D. *et al. Deletion of CASK in mice is lethal and impairs synaptic function*. (2007).

287. Srikanth, P. *et al.* Genomic DISC1 Disruption in hiPSCs Alters Wnt Signaling and Neural Cell Fate. *Cell Rep.* **12**, 1414–1429 (2015).

288. Firestein, B. L. *et al.* Cypin: a cytosolic regulator of PSD-95 postsynaptic targeting. *Neuron* **24**, 659–72 (1999).

289. Araque, A. *et al.* Tripartite synapses: glia, the unacknowledged partner. *Trends Neurosci.* **22**, 208–15 (1999).

290. Papouin, T., Dunphy, J. M., Tolman, M., Dineley, K. T. & Haydon, P. G. Septal Cholinergic Neuromodulation Tunes the Astrocyte-Dependent Gating of Hippocampal NMDA Receptors to Wakefulness. *Neuron* **94**, 840-854.e7 (2017).

291. Marques, S. *et al.* Oligodendrocyte heterogeneity in the mouse juvenile and adult central nervous system. *Science (80-.).* **352**, 1326–1329 (2016).

292. Romanov, R. A. *et al.* Molecular interrogation of hypothalamic organization reveals distinct dopamine neuronal subtypes. *Nat. Neurosci.* **20**, 176–188 (2017).

293. La Manno, G. *et al.* Molecular Diversity of Midbrain Development in Mouse, Human, and Stem Cells. *Cell* **167**, 566-580.e19 (2016).

294. Karasawa, T. & Lombroso, P. J. Disruption of striatal-enriched protein tyrosine phosphatase (STEP) function in neuropsychiatric disorders. *Neurosci. Res.* **89**, 1–9 (2014).

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

296. Darvas, M. *et al.* Modulation of the Ca2+ conductance of nicotinic acetylcholine receptors by Lypd6. *Eur. Neuropsychopharmacol.* **19**, 670–81 (2009).

297. Miwa, J. M. *et al.* The prototoxin lynx1 acts on nicotinic acetylcholine receptors to balance neuronal activity and survival in vivo. *Neuron* **51**, 587–600 (2006).

298. Sajo, M., Ellis-Davies, G. & Morishita, H. Lynx1 Limits Dendritic Spine Turnover in the Adult Visual Cortex. *J. Neurosci.* **36**, 9472–8 (2016).

299. Morishita, H., Miwa, J. M., Heintz, N. & Hensch, T. K. Lynx1, a Cholinergic Brake, Limits Plasticity in Adult Visual Cortex. *Science (80-.).* **330**, 1238–1240 (2010).

300. Verhoog, M. B. *et al.* Layer-specific cholinergic control of human and mouse cortical synaptic plasticity. *Nat. Commun.* **7**, 12826 (2016).

301. Perea, G., Navarrete, M. & Araque, A. Tripartite synapses: astrocytes process and control synaptic information. *Trends Neurosci.* **32**, 421–431 (2009).

302. Bélanger, M. & Magistretti, P. J. <DialoguesClinNeurosci-11-281.pdf>. *Dialogues Clin Neurosci* **11**, 281–295 (2009).

303. Papouin, T. *et al.* Synaptic and Extrasynaptic NMDA Receptors Are Gated by Different Endogenous Coagonists. *Cell* **150**, 633–646 (2012).

304. Zhang, X.-Y. *et al.* Glycine induces bidirectional modifications in N-methyl-D-aspartate receptor-mediated synaptic responses in hippocampal CA1 neurons. *J. Biol. Chem.* **289**, 31200–11 (2014).

305. Camargo, N. *et al.* Oligodendroglial myelination requires astrocyte-derived lipids. *PLOS Biol.* **15**, e1002605 (2017).

306. Goritz, C., Mauch, D. H. & Pfrieger, F. W. Multiple mechanisms mediate cholesterol-induced synaptogenesis in a CNS neuron. *Mol. Cell. Neurosci.* **29**, 190–201 (2005).

307. Mauch, D. H. *et al.* CNS synaptogenesis promoted by glia-derived cholesterol. *Science (80-.).* **294**, 1354–1357 (2001).

308. Moutinho, M. *et al.* Neuronal cholesterol metabolism increases dendritic outgrowth and synaptic markers via a concerted action of GGTase-I and Trk. *Sci. Rep.* **6**, 30928 (2016).

309. Chen, S. Neuregulin 1-erbB Signaling Is Necessary for Normal Myelination and Sensory Function. *J. Neurosci.* **26**, 3079–3086 (2006).

310. Roy, K. *et al.* Loss of erbB signaling in oligodendrocytes alters myelin and dopaminergic function, a potential mechanism for neuropsychiatric disorders. *Proc. Natl. Acad. Sci. U. S. A.* **104**, 8131–6 (2007).

311. Vasistha, N. A. *et al.* Familial t(1;11) translocation is associated with disruption of white matter structural integrity and oligodendrocyte-myelin dysfunction. *bioRxiv* 657163 (2019). doi:10.1101/657163

312. Jones, V. C., Atkinson-Dell, R., Verkhratsky, A. & Mohamet, L. Aberrant iPSC-derived human astrocytes in Alzheimer's disease. *Cell Death Dis.* **8**, e2696–e2696 (2017).

313. di Domenico, A. *et al.* Patient-Specific iPSC-Derived Astrocytes Contribute to Non-Cell-Autonomous Neurodegeneration in Parkinson's Disease. *Stem Cell Reports* **12**, 213–229 (2019).

314. Qiu, J. *et al.* Mixed-species RNA-seq for elucidation of non-cell-autonomous control of gene transcription. *Nat. Protoc.* **13**, 2176–2199 (2018).