Novel Schizophrenia Risk Genes and Gene Expression.

Deborah Knight

Thesis submitted for the degree of Doctor of Philosophy

2012

Department of Psychological Medicine and Clinical Neurosciences

Cardiff University

Supervisors: Prof. Michael O'Donovan,

Prof. Lesley Jones and Prof. Derek Blake

Contents Page.

Acknowledgements ……x

Summary ……xi

Chapter 1: General Introduction ……..1

1.1 Schizophrenia ……1

1.1.1 History ……1

1.1.2 Symptoms ……....1

1.1.3 Prevalence …….....2

1.1.4 Neurobiology ……...2

1.1.5 Heritability ……...3

1.1.6 Environmental Risk ……3

1.1.7 Genetic Risk ……....4

1.1.8. Genome-Wide Association Studies (GWAS) ……..…5

1.1.9 Polygenic Model of Schizophrenia ……...…5

1.2 *ZNF804A* …….…….5

1.2.1 Discovery as susceptibility gene……...5

1.2.2 Replication Studies …….……6

1.2.3 Psychosis and the overlap of Schizophrenia and Bipolar Disorder .6

1.2.4 Copy Number Variation in *ZNF804A* …….....7

1.2.5 *ZNF804A* mRNA Expression ……..……7

1.3. Intermediate Phenotypes and *ZNF804A* risk Variant ……8

1.4 *ZNF804A* and Drug Response ……..…10

1.5 *ZNF804A* Risk Variant …….……11

1.5.1 Putative Function of ZNF804A …….11

1.5.2. ZNF804A Protein …….12

1.6 Mouse models in the Understanding of the Human Brain ……..……12

1.7 The Mouse Orthologue of ZNF804A ……..13

1.8 Aims and Objectives ……13

Chapter 2: Materials and Methods ……15

2.1 Samples ……15
2.1.1 Mouse Mutant ……15
2.1.2 ENU Random mutagenesis and Speed Congenics ……15
2.1.3 Brain Tissue Collection ……16
2.2 DNA/RNA extraction ……17
2.2.1 DNA Extraction ……17
2.2.2 RNA Extraction ……17
2.2.2.1 RNA Clean Up – Column Purification ……17
2.2.2.2 DNase Treatment ……18
2.2.3 RNA quality assessment ……18
2.2.3.1 Agilent 2100 Bioanalyser ……18
2.2.3.2. 28s/18s rRNA Ratio ……19
2.2.3.3 RNA Integrity Number (RIN)……19
2.3 Reverse Transcription ……19
2.4 Polymerase Chain Reaction ……19
2.4.1 Primer design ……20
2.4.2 PCR Optimisation ……20
2.4.3 Agarose Gel Electrophoresis ……21
2.5 Genotyping ……21
2.6 Sequencing ……22
2.6.1 PCR Clean up ……22
2.6.2 Sequencing Reaction ……22
2.6.3 Post Sequencing Clean up ……23
2.6.4 Sequencing Analysis of C59X in Zfp804a ……23
2.7 High Resolution Melting Analysis (HRMA) ……24
2.7.1. HRMA PCR Conditions ……25
2.8 Global Analysis of Gene Expression ……25
2.8.1 Exon Arrays ……25
2.8.2 RNA Labelling, Hybridisation and Scanning of Exon Arrays ……26
2.8.2.1 RNA Labelling ……26

2.8.2.2 Exogenous Spike-in Controls ……..……27
2.8.2.3 Hybridisation and Scanning of the Exon Array ……..……27
2.8.3 Expression Analysis ……..….28
2.8.3.1 Partek Genomics Suite……..…28
2.8.3.2 Data upload ……28
2.8.3.3 Preprocessing ……..29
2.8.3.3.1. Robust Multichip Averaging (RMA) …….29
2.8.3.3.2 Background Correction ……29
2.8.3.3.3 Normalisation ……..30
2.8.3.3.4 Probe Summarisation ……...30
2.8.3.4 Annotating the Dataset ……..…..31
2.8.3.5 Quality Control Measures ……..…..32
2.8.3.5.1 Affymetrix Expression Console (v1.1.2) ……..…32
2.8.3.5.2 Principle Component Analysis (PCA) ……..…32
2.8.3.5.3 Exogenous Spike-in Controls ……..32
2.8.3.5.4 All Probeset (and Positive Control) RLE Mean……...33
2.8.3.5.5 Perfect Match (PM) Mean ……33
2.8.3.5.6 All Probeset (and Positive Control) Mean Absolute Deviation (Mad) Residual Mean ……33
2.8.3.5.7 Positive vs. Negative AUC ……..33
2.8.3.6 Statistical Analysis ……34
2.8.3.6.1 Differential Expression Algorithm ……..34
2.8.3.6.2 Filtering ……...34
2.8.3.6.3 Alternative Splicing Algorithm ……...35
2.8.3.6.4 Multiple Test Correction (MTC) and Non-independence of data ……36
2.8.3.7 Visualisation of Alternative Splice Events ……37
2.8.4 Alternative Algorithms for Detecting Differential Splicing ……..…37
2.8.4.1 easyExon ……37
2.8.4.2 AltAnalyze (v.2.0.7) ……...38

2.8.4.3 Microarray detection of Alternative Splicing (MiDAS) ……..…39

2.8.4.4 FIRMA ……40

Chapter 3: Expression analysis in Zfp804a ENU mutant mice ……41

3.1 Introduction ……..41

3.2 Methods ……48

3.2.1 Sample ……...48

3.2.1.1 The Zfp804a C59X Mice ……48

3.2.2 Female C59X Expression Study ……50

3.2.3 Male C59X Expression Study ……...50

3.2.4 Combined Analyses ……...50

3.2.5 Zfp804a mRNA Levels ……51

3.2.6 Sequencing the Mutation ……...51

3.2.7 Sample preparation and quality ……51

3.2.8 The Affymetrix Genechip Mouse Exon 1.0 ST Array ……..51

3.2.9 Statistical Analysis ……52

3.2.9.1 Partek Genomics Suite (Version 6.5) ……52

3.2.9.2 Data upload ……52

3.2.9.3 Probe Filtering ……52

3.2.9.4 Preprocessing ……..52

3.2.9.4.1 Background Correction ……52

3.2.9.4.2 Quantile Normalisation ……52

3.2.9.4.3 Probe Summarisation ……...53

3.2.9.5 Annotating the Dataset ……...53

3.2.9.6 Quality Control (QC) ……..53

3.2.9.7 Filtering ……..54

3.2.9.8. Statistical Algorithm ……..54

3.2.9.9 Multiple Test Correction (MTC) ……55

3.2.10 Visualisation of Alternative Splice Events. ……55

3.2.11 Degree of Overlap ……...55

3.2.12 Alternative Algorithms for Detecting Splicing ……...56

3.3 Results ……..57

3.3.1. Abundance of Zfp804a Transcript…….57

3.3.2 Sequencing the C59X mutation …….57

3.3.3 RNA Quality ……..58

3.3.4 Affymetrix Genechip Exon 1.0 ST Array……...…59

3.3.4.1 Quality Control ……..59

3.3.4.1.1 Hybridisation Efficiency ……60

3.3.4.1.2 Examining the Global Expression Pattern with Principle Components Analysis (PCA) …….……61

3.3.4.2 Identification of Differentially Expressed Genes between Female Wildtype and C59X Mutant Mice ……..69

3.3.4.3 Identification of Differentially Spliced Genes between Female C59X Mutant and Wildtype Mice ……..69

3.3.4.4. Replication study in Male C59X mice ……..72

3.3.4.5 Multiple Test Correction ……86

3.3.4.6 Analysis to Determine the Degree of Replication ……..88

3.3.4.7 Manual inspection of Results to Determine Direction of Effect ……89

3.3.4.8 Combined Analysis …… 95

3.3.4.9 Alternative Algorithms ……...98

3.3.4.9.1 easyExon …… 98

3.3.4.9.2 AltAnalyze (Version 2.0) ……....99

3.3.4.10 Significant Results Overlap between easyExon, AltAnalyze and Partek GS …….101

3.3.4.11 C59X Mutant and Wildtype expression of Zfp804a ……..108

3.4 Discussion ……...115

Chapter 4. RNA-Sequencing …….119

4.1 Introduction ……119

4.2 Methods ……..120

4.2.1 Sample …….120

4.2.2. Sample Preparation …….120

4.2.3. cDNA Library Quality Control ……...120
4.2.4 Sequencing ……...121
4.2.5 RNAseq Quality Control ……121
4.2.6 Alignment, Assembly and Differential Expression Analysis ……...122
4.2.6.1. Sequence Alignment and Identification of Splice Junctions ...122
4.2.6.2. Transcript Assembly ……123
4.2.6.3. Differential Expression Analysis ……124
4.3 Results ……125
4.3.1 cDNA Library Quality Control …….. 125
4.3.2 Sequence Quality Control ……127
4.3.2.1 Assessing 'Per base sequence quality' ……127
4.3.2.2 Per sequence quality Scores ……128
4.3.2.3 Per base sequence content ……129
4.3.2.4 Per base GC content ……131
4.3.2.5. Per sequence GC content ……132
4.3.2.6 Per base N content ……... 133
4.3.2.7 Sequence length distribution ……134
4.3.2.8 Sequence Duplication Level ……135
4.3.2.9 Overrepresented sequences ……...137
4.3.2.10 Overrepresented K-mers ……137
4.3.2.11 RNAseq Quality ……138
4.3.3 Zfp804a C59X mutation ……139
4.3.4 Differential Expression in C59X Mutants ……...142
4.3.5 *Alpha Synuclein* (*Snca*) ……143
4.3.6 RNAseq Predictions of Differential Splicing ……..145
4.4 Discussion ……..149
Chapter 5. Expression Analysis in Embryonic C59X Mice ……153
5.1 Introduction ……153
5.2 Materials and Methods ……...156

5.2.1 Sample ……156

5.2.2 Genotyping ……..157

5.2.2.1 Gender PCR ……..157

5.2.3 RNA Processing ……..158

5.2.4 Affymetrix Exon Array ……...158

5.2.4.1 Quality Control ……158

5.2.4.2 Determining Snca Expression ……..158

5.2.4.3 Partek Genomics Suite ……158

5.2.4.4 EasyExon ……..159

5.2.4.5 AltAnalyze ……159

5.2.4.6 Overlap Analysis ……..159

5.3 Results ……160

5.3.1 Embryonic Sample ……...160

5.3.2 RNA quality ……162

5.3.3 Quality Control ……164

5.3.4 Deletion at the *Snca* Locus ……...174

5.3.6 Embryonic C59X Mutant vs Wildtype Analysis ……179

5.3.6.1 Differential Gene expression ……179

5.3.6.2 Differential Splicing ……179

5.3.7 Overlap between the findings from the Embryonic and Adult studies …180

5.3.8 Alternative Algorithms ……183

5.3.8.1 easyExon ……...183

5.3.8.2 AltAnalyze ……183

5.3.8.3. Differential Expression and Splicing Results Common to Partek GS, AltAnalyze and easyExon ……184

5.3.8.3.1. Expression: ……..……184

5.3.8.3.2. Splicing: ……187

5.4 Discussion ……...193

Chapter 6. Technical Artefacts ……196

6.1 Introduction ……196

6.2 Methods ……..198

6.2.1. Linked Sequence variants affecting cDNA-probe binding ……198

6.2.2. Exclusion of Chromosome 2 ……..199

6.2.3 Impact of Intensity Threshold……...199

6.2.4 Impact of Gene size ……200

6.2.5 Re-analysing the data with more stringent thresholds and exclusion of Chromosome 2 ……..200

6.2.6 Embryonic and Adult Dataset Replication Following Stringent Analyses ……...200

6.2.7 Expression and Splicing of *Zfp804a* ……200

6.3 Results ……201

6.3.1 Linkage with the C59X Mutation ……201

6.3.2 Exclusion of Chromosome 2 ……...208

6.3.3 Assessment of Intensity Cut-offs ……210

6.3.4 Determining the Effect of Gene Size on Differential Splicing Results ...213

6.3.5 Re-analysis With More Stringent Intensity Filters and Exclusion of Chromosome 2 ……..214

6.3.5.1 Differential Expression between C59X Mutants and Wildtypes ……...214

6.3.5.2 Differential Splicing Between Embryonic C59X Mutants and Wildtypes ……217

6.3.6 Expression and Splicing in Zfp804a Following increased Intensity Filtering ……220

6.3.6.1 Differential Expression of Zfp804a ……..220

6.3.6.2 Differential Splicing of Zfp804a ……..222

6.4 Discussion ……...227

Chapter 7. Relevance of altered Zfp804a function for Schizophrenia ……..230

7.1 Introduction ……230

7.2 Methods ……..233

7.2.1 Pathway Analysis ……233

7.2.1.1 Gene and Background lists ……...233

7.2.1.2 Metacore GeneGO ……234

7.2.1.3 DAVID ……236

7.2.3 Genetic Analysis using the PGC database ……..238

7.2.3.1 Investigating PGC Schizophrenia top hits for Differential Expression and Splicing in C59X mutants ……...238

7.2.3.2 Investigating the C59X Differentially Expressed and Spliced Genes for Association with Psychosis ……..238

7.2.3.2.1 Approximation Method using Brown's p value ……239

7.2.3.2.2 Simes' P value ……...239

7.3 Results ……240

7.3.1 Pathway Analysis based on embryonic expression data ……240

7.3.1.1 Pathways Enriched for Differentially Expressed Genes …….. 240

7.3.1.2 Pathways Enriched for Genes Differentially Spliced between C59X Mutants and wildtypes ……...247

7.3.1.3. The effect of Probeset Number on Pathway Analysis …….255

7.3.2 Genetic Analysis ……..260

7.3.2.1 Investigating PGC Schizophrenia Top Hits for Significant Differential Expression and Splicing in the Embryonic Dataset ……..260

7.3.2.2 Relevance of Genes Differentially expressed in C59X Mutants to Disease Risk ……262

7.3.2.3 Relevance of Genes Differentially Spliced in C59X Mutants to Disease Risk ……263

7.4 Discussion ……...264

Chapter 8. General Discussion ……..268

8.1 Research Findings ……..268

8.2 Limitations of the data ……272

8.3 Future Work ……273

8.4 General Conclusion ……274

Appendix ……276

Chapter 2 Appendices ……...276

Chapter 3 Appendices ……...277

Chapter 4 Appendices ……...283

Chapter 5 Appendices ……...287

Acknowledgements.

I would first like to thank my supervisors Prof. Mick O'Donovan, Prof. Lesley Jones and Prof. Derek Blake for the opportunity to undertake this PhD and for the support, encouragement and advice throughout.

Thanks to everyone in Psychological Medicine especially the girls in my office; Sarah, Amy, Denise, Evie, Didi, Jade and Irina. This whole process would have been a lot harder without the support, laughter and chocolate they have provided. I must also thank Hywel for all the advice he has given throughout and for patiently listening to me practise every one of my talks. Thanks to Dobril for helping me with the RNAseq analysis. I am also very grateful to Tamara and Jess for all their help with the mouse work.

Thanks to all my friends, especially Hannah for the constant words of encouragement and for making sure I had a break once in a while. Thanks to Jo for never complaining once about having to live in a sea of papers and to Becky for keeping my stress levels in check.

To my wonderful family, especially my Gran and Sam for helping me keep things in perspective when things got tough. Finally I would like to thank my parents whose words never fail to calm and comfort me and whose love and support has got me to where I am today, this thesis is dedicated to them.

Summary.

ZNF804A was (at the time this work started) one of only a few robustly implicated schizophrenia susceptibility genes, due to replicated genome-wide significant evidence for association between a polymorphism in the gene and schizophrenia. Determining the function of the ZNF804A protein, which is currently unknown, may provide a way of elucidating the pathophysiology of this relatively common, complex disorder. Based on the hypothesis that the ZNF804A protein regulates gene expression or splicing, the aim of this thesis was to identify genes that exhibit altered expression or splicing in brain tissue from mice in which the orthologue *Zfp804a* carries a nonsense mutation.

No robust evidence was obtained that showed the effects of the mutation on differential expression in individual genes. Although this finding does not support the hypothesis that ZNF804A acts directly to regulate gene expression, the results may reflect the possibility that effects on gene expression may be too subtle to be detected using the methods applied. Evidence was obtained to show the mutation affected the alternative splicing of a number of individual genes, which could suggest a role for ZNF804A in the direct or indirect regulation of alternative splicing.

Through RNA sequencing, I identified a novel transcript in Zfp804a with an alternative exon upstream of the Refseq exon 1. I also showed that a proportion of the significant splicing differences identified in mutants were artefacts of strain differences in gene sequences that are likely to affect the efficiency of hybridisation on the exon array.

Genes identified as differentially spliced between mutants and wildtypes were enriched in axon guidance and cell adhesion pathways, both thought to be important during development. The findings of this thesis suggest the novel hypothesis that *ZNF804A* effects risk for schizophrenia via aberrant splicing in the above pathways that are critical to normal brain development. Further studies with increased power are required to understand the effects on gene expression.

Chapter 1: General Introduction.

1.1 Schizophrenia

Schizophrenia is a severe psychiatric disorder with psychosis being a prominent feature. Due to its chronic course, early onset and poor treatment response, it contributes substantially to human morbidity, and also has a major negative impact on the social and economic functioning of affected individuals, their families, and the wider society. At present little has been established about the specifics of schizophrenia aetiology. This lack of knowledge is a clear impediment for new approaches in the design of treatments with greater efficacy. Schizophrenia is ranked within the top ten most disabling and costly disorders in society (Murray & Lopez, 1996; Freedman, 2003) and the estimated cost of schizophrenia to society in the UK in 2004-2005 was ~£6.7 billion (Mangalore & Knapp, 2007). There is a clear need to determine the underlying pathophysiology of this often devastating disorder to enable the improvement of treatment and outcome.

1.1.1 History

Schizophrenia was first described by Kraepelin (1899) whose classification of a dementia praecox described a degenerative disorder of cognitive disturbance distinct from manic depressive psychosis. The degenerative disorder Kraepelin described appeared inaccurate with many patients showing improvement in symptoms and so the disorder was re-termed in 1908 by Eugen Bleuler as schizophrenia from the Greek ‘split mind’ and refers to the disruption in thought and cognitive function that are characteristic of the disorder.

1.1.2 Symptoms

Schizophrenia is a heterogeneous disorder. Symptoms are often classified as either positive or negative (Jablensky et al., 2006). Positive symptoms, those with features that are not present in normal individuals, include hallucinations, delusions and disorganised

thoughts. The negative symptoms, functions which are present in normal individuals but often absent in schizophrenia, include social withdrawal and emotional flattening. A cognitive deficit is also frequently observed affecting memory, attention and executive function (Dikeos et al., 2006). Diagnosis is based upon the assessment of behaviour as no biomarkers have been determined for the disorder. No disorder specific neuropathology has been identified preventing confirmation of a correct diagnosis in the post mortem brain.

1.1.3 Prevalence

Schizophrenia has a worldwide prevalence of 1%. (Gottesmann, 1991). Symptoms typically present in late adolescence to early twenties with a higher lifetime risk in males (McGrath et al., 2004).

1.1.4 Neurobiology

Whilst the pathophysiology underlying schizophrenia remains unclear the occurrence of hallucinations and delusions as well as cognitive deficits have implicated brain functions associated with perception and cognition in the underlying pathophysiology of schizophrenia (Ross et al. 2006). Early hypotheses of schizophrenia aetiology centered on a hyper-dopaminergic system based on the efficacy of D2 receptor antagonists at alleviating the positive symptoms (Snyder, 2006). Findings that dopamine activity is aberrant in sub-cortical regions during psychotic periods supports this view (Howes et al., 2009). The efficacy of drug treatments which in addition target serotonin 5HT2A receptors, points to a complex aetiology of the disorder and although these drugs are generally effective at treating positive symptoms, and in the case of clozapine, negative symptoms, the pharmacological studies of how they work have not advanced understanding of the complex pathophysiology underlying schizophrenia. The varying efficacy of such drugs in schizophrenia patients emphasize the heterogeneity of the disorder. The occurrence of schizophrenia-like symptoms in healthy individuals following the use of phencyclidine (PCP) and ketamine which are both NMDA antagonists introduced the hypo-glutamatergic hypothesis (Coyle, 2006) and this has been given further credibility by some promising evidence for the treatment of

symptoms, including negative symptoms, by drugs which modulate NMDA-receptors (Coyle, 2006). A possible role for gabaergic, in addition to dopaminergic and glutamatergic input has also been supported in pharmacological studies (Javitt et al., 2008).

Gross abnormalities in the schizophrenia patient brain including enlargement of the lateral ventricles and an accompanying reduction in overall brain volume (Steen et al., 2006) are present from birth and not progressive which has led to a neurodevelopmental hypothesis for schizophrenia (Weinberger, 1986). Regional structural differences such as reduced hippocampal and pre frontal cortex (PFC) volume as well as altered cytoarchitecture (Harrison, 2000) have also been observed, although these are less replicable than the changes in total brain and ventricular volume. In addition neural distribution and spine density appear abnormal in the hippocampus and prefrontal cortex in the brains of schizophrenic patients (Wong & Van Tol., 2003). Despite numerous hypotheses, the underlying aetiology and pathophysiology of schizophrenia remains elusive.

1.1.5 Heritability

Evidence from family, twin and adoption studies have shown there to be a large genetic component to schizophrenia risk, with heritability being ~80% (Cardno. & Gottesman, 2000), yet evidence from monozygotic twins highlights the additional influence of environmental factors (Gottesman, 1991; McGrath & Murray 2003).

1.1.6 Environmental Risk

The immune response is thought to be involved in aetiology and infections such as influenza, poliovirus and *Toxoplasma gondii* (Brown & Susser, 2002) have been associated with schizophrenia. Aberrant events occurring during pregnancy and birth have been suggested to increase risk of schizophrenia such as obstetric complications like preeclampsia (Dalman et al., 1999). In addition urbanicity (Allardyce et al., 2001) and drug use (Arseneault et al., 2002) have all been suggested to contribute to risk,

although the mechanisms through which these environmental factors might act are unclear.

1.1.7 Genetic Risk

While its origins are enigmatic, it has been known for a long time that genes make a substantial contribution to population risk, the heritability of schizophrenia being ~80%, therefore genetics has been seen as an important tool in trying to understand the causes of the disorder. Despite this, finding the specific risk genes involved has, as in most complex diseases, been a major challenge and early linkage and association studies did not reliably identify any susceptibility genes operating in the wider case population at high levels of confidence. However, until recently, these studies have been subject to small sample sizes and in turn low power (Kirov et al., 2005; Owen et al., 2005; Ross et al., 2006).

The common disease-rare variant hypothesis supports the idea that substantial knowledge of disease aetiology can be gained from understanding rare variants even if observed in very low frequency (only a limited number of people). Given relatively large effect sizes, they can inform a great deal on the underlying biology of a disease as has been exemplified in AD, PD and HD (Ross et al., 2006, Ross & Margolis, 2005). A number of studies have found that specific rare copy number variants (CNVs) occur more frequently in schizophrenia (<1%) compared to controls (<0.1%) showing that rare alleles are involved in the disorder (ISC., 2008; Kirov et al., 2009; Stefansson et al., 2008). Rare *de novo* CNVs occur significantly more frequently in cases (5%) than controls (2%) indicating the involvement of *de novo* CNVs in schizophrenia pathogenesis. *De novo* CNVs identified in schizophrenia patients were enriched for genes associated with the post-synaptic density particularly those involved in NMDA and ARC post-synaptic signalling suggesting the involvement of these pathways in schizophrenia pathophysiology (Kirov et al., 2012). A proportion of genetic risk for schizophrenia is therefore attributed to rare CNVs, however at present, it is in general unclear which of the (generally many) genes within specific CNVs are relevant to schizophrenia.

1.1.8. Genome-Wide Association Studies (GWAS)

The advent of genome-wide association studies (GWAS), which like linkage but unlike candidate gene studies, requires no knowledge of disease pathophysiology, and like other association study designs can in principle detect small effect sizes, changed the way genetic studies could be carried out. In part this was because of the absence of an *a priori* requirement for selecting specific candidates based upon prevailing theories of disease origin, but another important factor was that the use of the very large sample sizes required for GWAS allowed robust evidence to emerge for a number of loci. Early GWAS of schizophrenia were small and did not strongly support any particular candidates (Lencz et al., 2007; Sullivan et al., 2008). The application of GWAS to very large discovery and replication samples, however resulted in the identification of strong evidence for association to a small number of loci.

1.1.9 Polygenic Model of Schizophrenia

The theory of a polygenic model for schizophrenia (Gottesman & Shields, 1967) remained unsubstantiated at the molecular genetic level for many years, but was recently confirmed empirically when a Genome Wide Association Study (GWAS) showed risk for schizophrenia to be conferred by very many, possibly thousands, of alleles conferring small increments to risk (OR<1.1) (ISC, 2009). As schizophrenia is a polygenic disorder that is characterised by a heterogeneous set of symptoms it can be inferred that the underlying pathophysiology will also show heterogeneity. Despite the complex nature of the disorder and difficulties determining its aetiology, the identification of genes associated with the disorder, despite their small effect size, when considered together in the context of biological pathways and molecular mechanisms may inform networks and pathways aberrant in schizophrenia.

1.2 *ZNF804A*

1.2.1 Discovery as susceptibility gene

Association of a polymorphism (rs1344706) within *zinc finger protein 804A* (*ZNF804A*) and schizophrenia was first highlighted by O'Donovan et al., in 2008 in a large

discovery and replication sample. The replication analysis provided strong evidence for association around *ZNF804A* ($P = 1.61 \times 10^{-7}$) and this association surpassed genome-wide levels of significance when the affected phenotype included bipolar disorder ($P = 9.96 \times 10^{-9}$) (O'Donovan et al., 2008). This indicated *ZNF804A* as a strongly supported candidate (O'Donovan et al., 2008) for both schizophrenia and bipolar disorder.

1.2.2 Replication Studies

The association of rs1344706 within the *ZNF804A* gene with schizophrenia has been replicated (ISC, 2009; Stefansson et al., 2009) several times, including in an Irish Case-Control Study of Schizophrenia (ICCSS) sample (N=1021 cases, 626 controls) (P=0.0113). In this study 11 SNPs in linkage disequilibrium (LD) with rs1344706 were also investigated for association. Another SNP, rs7597593 (P=0.0013) showed the most significant association with schizophrenia (Riley et al., 2010). Another study replicated the association of rs1344706 and schizophrenia (odds ratio OR = 1.08, P = 0.0029) in 5164 schizophrenia cases and 20,709 controls in addition to replicating the significant association when a bipolar disorder phenotype was added to the sample (OR = 1.09, P = 0.00065) (Steinberg et al., 2011). A meta-analysis of almost 60,000 subjects provided convincing evidence that *ZNF804A* was a susceptibility gene for schizophrenia ($P=4x10^{-11}$) and even more compellingly for a wider phenotype including bipolar disorder ($p=2x10^{-13}$) (Williams et al., 2010). In a meta-analysis undertaken by the schizophrenia Psychiatric GWAS Consortium (PGC) (PGC, 2011a) rs1344706 was not one of 7 genome-wide significant variants identified as being associated with schizophrenia. The odds ratio in this study was similar to that observed in the previous meta-analysis (OR 1.10) (Williams et al., 2010) despite the sample being almost half the size. Given that the odds ratio is similar in a sample half the size this is consistent for a true association between *ZNF804A* and schizophrenia. *ZNF804A* is therefore considered as one of the most robustly associated schizophrenia susceptibility genes.

1.2.3 Psychosis and the overlap of Schizophrenia and Bipolar Disorder

There is strong evidence for genetic overlap between schizophrenia and bipolar disorder (ISC, 2009). With regards to *ZNF804A*, joint analyses of schizophrenia and the bipolar disorder phenotype provide even stronger evidence for association ($p = 2x10^{-13}$) than

schizophrenia alone. A stronger association between rs1344706 and bipolar disorder has been reported, but not yet confirmed, when considering a sub-group of bipolar disorder patients with psychosis (Lett et al., 2011) suggesting this gene operates as a more general risk factor for psychosis. Evidence for shared genetic risk has also been shown for other genome-wide significant susceptibility genes. *CACNA1C* originally associated with bipolar disorder is also significantly associated with schizophrenia (Green et al., 2010, PGC, 2011a). *Neurogranin* (*NRGN*) and the MHC region originally identified as schizophrenia susceptibility loci were shown to have nominally significant association ($P < 0.05$) with bipolar disorder while *Polybromo-1* (*PBRM1*), originally associated with bipolar disorder is also significantly associated with schizophrenia ($P = 0.00015$) (Williams, et al., 2011). Based on this evidence it is now widely viewed that schizophrenia and bipolar disorder are not aetiologically discrete entities, and cross diagnostic approaches should be used with regards to research (Craddock & Owen, 2010).

1.2.4 Copy Number Variation in *ZNF804A*

As well as common risk associated with *ZNF804A*, two CNVs that include *ZNF804A* (at least in part) in patients with psychosis were identified relative to none in controls (Steinberg et al., 2011). The same study also identified another CNV in *ZNF804A* in a patient with anxiety (Steinberg et al., 2011). Two independent studies of autism spectrum disorder (ASD) subjects respectively identified CNVs spanning *ZNF804A* (Griswold, 2012) and balanced chromosomal abnormalities in *ZNF804A* (Talkowski et al., 2012). No rare SNPs within *ZNF804A* itself have been associated with schizophrenia (Dwyer et al., 2010), but identification of CNVs which include *ZNF804A* provide evidence for the contribution of rare variation in *ZNF804A* to the risk of schizophrenia and again emphasises the overlap between schizophrenia and a wider psychosis phenotype as well as with neurodevelopmental disorders.

1.2.5 *ZNF804A* mRNA Expression

An RNA sequencing study of differential expression of inducible pluripotent stem cells (iPSCs) and differentiated neurons found large expression differences in genes involved

in neuropsychiatric disorders, including that of *ZNF804A* (Lin et al., 2011). Another study reported that in neurons derived from human iPSCs reprogrammed from fibroblasts derived from schizophrenia patients, altered expression of *ZNF804A* occurred in those derived from some but not all of the patients **(**Brennand et al., 2011). A novel exon within the intron 2 of *ZFP804A* (denoted Exon 2.2) was recently discovered in post mortem brain tissue from occipital lobe and LCL cultured cells (Okada et al., 2012), which the authors postulate could encode a novel immature 88 amino acid protein. Compared to equal abundance of the protein coding transcript, the novel variant was found to be downregulated in schizophrenia patients relative to controls (Okada et al., 2012). Higher *ZNF804A* mRNA expression has been associated with the schizophrenia risk allele in post-mortem brain tissue from the prefrontal cortex (Riley et al. 2010; Williams et al., 2010) but the former is not thought to be attributable to the latter (Williams et al., 2010). Expression of *ZNF804A* in post mortem brain samples from healthy individuals has in one study been shown to be dependent on an interaction between genotype at another associated SNP in *ZNF804A* (rs7597593) and gender (Zhang et al., 2011a). Female carriers of the protective allele had significantly higher levels of *ZNF804A* mRNA relative to risk allele carriers and a trend for reduced levels was observed in males with the protective allele relative to risk allele carriers (Zhang et al., 2011a), but this is as yet unreplicated.

1.3. Intermediate Phenotypes and *ZNF804A* risk Variant.

Studying intermediate phenotypes can be beneficial in heterogenous disorders such as schizophrenia (Rose et al., 2012). Aberrant connectivity has been observed in the brains of healthy *ZNF804A* risk allele carriers relative to non risk allele carriers (Esslinger et al., 2009; Rassetti et al. 2011; Paulus et al., 2011). Although the neurophysiological basis for this is unknown, this abnormal connectivity between the dorso lateral prefrontal cortex (dlPFC) and hippocampus suggests a role for ZNF804A at the neural systems level.

Based on evidence of aberrant connectivity it was postulated that white matter volume or connectivity may be affected in rs1344706 risk carriers (Esslinger et al., 2009). Increased white matter volume has been observed in healthy subjects homozygous for the risk allele (Lencz et al., 2010; Wei et al., 2012; Wassink et al., 2012) which might relate to altered connectivity in the brains of those with the risk allele. In another study, no main effect of rs1344706 risk variant on total brain volume or regional brain volumes (Cousijn et al., 2012) was identified, but an interaction between genotype and disorder has been reported. In that study, reduced grey matter thickness was found in healthy individuals homozygous for rs1344706 risk variant (Lencz et al., 2010; Voineskos et al., 2011), but this contrasted with increased gray matter volume in patients who were homozygous risk carriers relative to non-risk allele carriers (Donohoe et al., 2011).

However, the increased gray matter in cases who carry the risk allele may relate to a finding that performance in working and episodic memory tasks is better in rs1344706 risk allele carriers compared to non-risk allele carriers, but this is observed specifically in patients not in controls (Walters et al., 2010). The authors of the latter study proposed that ZNF804A was related to a type of schizophrenia with relative sparing of cognitive function rather than that ZNF804A was itself increasing cognitive performance. This idea was also supported by evidence that association between rs1344706 and schizophrenia was stronger in patients with the highest IQ (Walters et al., 2010; Chen et al 2012). The observation of preserved cognitive phenotype in *ZNF804A* risk variant carriers has been replicated in schizophrenia patients in a processing speed task (Van Den Bossche et al., 2012).

However, a deletion syndrome known as "2q31.2q32.3 syndrome," which includes the deletion of *ZNF804A* as well as *NEUROD1, PDE1A* and *ITGA4* has been reported in 3 patients and is characterised by clinical features including mental retardation and developmental delay (Cocchella et al., 2010), although the causal gene(s) for this phenotype are unknown.

Theory of mind (ToM) is a concept which encompasses an individual's ability to infer the thoughts, feelings and intentions of others and marked deficits in this cognitive process are observed in schizophrenia patients (Bora et al., 2009). There is evidence of altered activity in brain areas associated with ToM (dorsomedial PFC and left tempero-parietal cortex) in healthy carriers of the risk allele, possibly indicating a role for ZNF804A in the neural networks underlying ToM processes (Walter et al., 2010). Executive control has also been reported to be altered in rs1344706 risk allele carriers (Balog et al., 2011), suggesting a broader social cognitive function for ZNF804A.

Although the evidence from endophenotype studies largely remain to be confirmed, and the statistical support for the associations is weaker than for association to the primary phenotype, imaging and cognitive studies on subjects with the rs1344706 variant have led to the hypothesis that a subtype of schizophrenia may exist defined by a preserved cognitive function and increased grey matter volume present only in risk allele carriers (Donohoe et al., 2011). This could reflect a relatively distinct pathophysiology in those with the *ZNF804A* risk variant which if confirmed could be used to inform drug targeting.

1.4 *ZNF804A* and Drug Response.

There is weak evidence to suggest association between the risk allele at ZNF804A and poor response to atypical antipsychotics. Patients with the risk allele had a poorer response to atypical antipsychotics measured using the Positive and Negative Syndrome Scale (PANSS) (Zhang et al., 2012; Mossner et al., 2012). Two additional SNPs in *ZNF804A* associated with schizophrenia, rs35676856 and rs61739288 were also correlated with a poor response to atypical antipsychotics as measured using the PANNS (Xiao et al., 2011). Again, this may point to a subtype of pathophysiology, but since spared cognitive function is usually considered a good prognostic factor, these results are not obviously compatible with those from the cognition studies.

1.5 ZNF804A Risk Variant.

The *ZNF804A* gene found on chromosome 2q32.1 has 4 exons and spans 341 kb. Despite extensive searching, rs1344706 was found to be the variant with the strongest signal for association with schizophrenia (Williams et al., 2010). As there is no evidence that rs1344706 is in strong LD with any other variants, the association may not be attributable to rs1344706 being in LD with the casual variant. The rs1344706 SNP is within 30bp of conserved mammalian sequence (Donohoe et al., 2010) hypothesised to show this degree of conservation due to the presence of transcription binding sites (Riley et al., 2010). Given the location of the associated SNP (rs1344706) within intron 2 of *ZNF804A*, the most likely predicted function is regulation of transcription or splicing leading to the hypothesis that risk is inferred via altered regulation of gene expression or RNA processing. The prediction that the rs1344706 risk allele (T) maintains binding sites for the brain expressed transcription factors MYT1l and POU3F1/OCT-655 (Riley et al., 2010), along with evidence to show that rs1344706 does form sequence-specific DNA-protein complexes (Hill & Bray, 2011), provides evidence that rs1344706 is a functional variant, but the identity of the nuclear binding protein(s) remains unknown (Hill & Bray, 2011). Another variant (rs13423388) significantly associated with schizophrenia (Zhang et al., 2011b) is found in *ZNF804A* 3kb downstream of rs1344706, this region is also highly conserved between mouse and human (Zhang et al., 2011b; using UCSC Browser, 2009) all of which supports the hypothesis that the allele may be involved in transcription factor binding or splicing.

1.5.1 Putative Function of ZNF804A.

As recently as 4 years ago there were no unequivocally implicated genes in schizophrenia (O'Donovan et al. 2009). The replication of association between *ZNF804A* and schizophrenia provides robust evidence that it is a schizophrenia susceptibility gene. At the time this study started, this was the most strongly implicated schizophrenia susceptibility gene from which insights into schizophrenia pathogenesis could be derived. However, deriving such insights was hampered by the fact that the function of the gene was unknown. The broad basis underpinning this thesis was that determining the function of the protein encoded by *ZNF804A* may provide a valuable way of elucidating the pathophysiology of this common, complex disorder.

1.5.2. ZNF804A Protein

ZNF804A is predicted to be a zinc finger protein containing a single zinc finger domain. The single domain is a cysteine2histadine2 (C2H2)-type domain. This domain is commonly found in transcription factors, leading to the hypothesis that ZNF804A may also function as a transcription factor (Williams et al, 2010), although this proposition is tentative since zinc finger domains are found in other classes of protein such as those with housekeeping functions (Pieler and Bellefroid, 1994).

Changes in gene expression have been observed in both the knockdown of ZNF804A in a neural cell line (Hill et al., 2012) and the over expression of ZNF804A in E11 rat forebrain progenitor cells (Girgenti et al., 2012) with chromatin immunoprecipitation assays suggesting this is a direct effect of ZNF804A on expression (Girgenti et al., 2012). These studies are discussed in much greater detail in chapter 3, section 1.

1.6 Mouse Models in the Understanding of the Human Brain.

There is a high degree of conservation between mouse and human with comparable biochemical pathways found in the two species. Methods of genetic manipulation are well characterised in the mouse and generating a strain of mice is efficient due to their short gestation periods (Stevens et al., 2007). In particular inbred strains of mice have been created by researchers to achieve genetic homogeneity between mice of the same strain offering valuable consistency across different mice and research carried out in different labs and in different countries when using the same strain. This valuable resource offers the chance to attribute any phenotype to the particular manipulations made by the researcher, rather than genetic heterogeneity. However, inbred strains of mice are not immune to genetic drift or to de novo mutations and so the stable strains of inbred mice may not be as homogenous as previously thought (Stevens et al. 2007). The advantages and disadvantages of mouse models with regards to the work in this thesis are outlined in more detail in chapter 3, section 1.

1.7 The Mouse Orthologue of ZNF804A.

The mouse orthologue of ZNF804A is *Zfp804a*. It is a 206, 221bp gene which encodes a 1200aa protein. A single Refseq mRNA (NCBI) has been identified. Zfp804a has been identified as a downstream target of the Hoxc8 protein both *in vitro* and in embryonic mice (Chung et al., 2010). A Hoxc8 binding site has been identified within intron one of Zfp804a and Hox binding sites were identified that were conserved across species. In response to Hoxc8 binding, Zfp804a mRNA expression has been found to be unregulated *in vitro*, suggesting the possibility that the influence of Hoxc8 on Zfp804a may be relevant to the pathophysiology underlying schizophrenia and psychosis.

In the adult mouse brain, interaction of Hoxc8 and Zfp804a has been demonstrated in the cortex and in the whole brain. The HOX protein family has been previously implicated in development, including particularly brain patterning (Tischfield et al., 2005). Members of the family have been described to have altered expression in epilepsy, mental retardation and subtypes of ASD (Zollino et al, 2011; Bosley et al., 2007). As Hoxc8 appears to regulate the expression of Zfp804a this could implicate Zfp804a as being a mediator of the effects of HOX members in developmental processes relevant to schizophrenia.

1.8 Aims and Objectives

The broad objectives of this thesis are to identify mechanisms by which the *zinc finger protein 804A* gene encoding ZNF804A might influence risk of schizophrenia and a wider psychosis phenotype. To achieve this, based upon the hypothesis that ZNF804A is a regulatory protein, I studied the consequences of altered ZNF804A expression in the brains of mice carrying truncating mutations at *ZNF804A* (in mouse, known as *Zfp804a*). It was my hypothesis that genes identified as showing altered expression will contain downstream mediators of the effects of ZNF804A on disease risk, and that identifying these genes would inform both on the function of this gene and on pathways relevant to schizophrenia. Further, under a polygenic model of disease involving large numbers of risk alleles, I also postulated that the human orthologues of downstream targets, direct and indirect, of Zfp804a will contain variants that influence risk of psychosis.

Prior to the start of this thesis, an ENU mouse line with a *Zfp804a* nonsense mutation had been bred to isolate the *Zfp804a* mutation from the ENU parental strain genomic background in order to facilitate behaviour and expression studies. My plan was to use a global transcriptomics driven approach to identify genes in which mRNA expression is altered in the brains of animals carrying the *Zfp804a* nonsense mutation, and functional pathways enriched for such genes. The specific approaches were initially based upon the Affymetrix GeneChip Mouse Exon 1.0 ST Array analysis followed by global transcriptomics approaches based upon Illumina next generation sequencing. Target genes identified as differentially expressed or otherwise regulated by *Zfp804a* were then tested for aetiological relevance in the large genome-wide association case-control datasets from subjects with schizophrenia and bipolar disorder from the Psychiatric GWAS Consortium (PGC, 2011a; PGC, 2011b) using a multilocus genetic association approach. Overall, the aim was to elucidate genes and specific pathophysiological mechanisms relevant to the aetiology of schizophrenia and other psychoses.

Chapter 2: Materials and Methods.

2.1 Samples

2.1.1 Mouse Mutant

Prior to the work described in this thesis, a premature termination codon (PTC) mutation in exon 2 of Zfp804a was identified (C59X) by another PhD student (T. AlJanabi) who had screened an ENU mouse library (ENU DNA Archive, MRC Mary Lyon Centre, Harwell) consisting of thousands of DNA samples from F1 ENU mutagenised mice (along side frozen sperm samples). The mutation resulted in a two base substitution from GT to AA replacing the wild type cysteine residue with a stop codon and is denoted from here on as C59X. The mutation is expected to either truncate the translated protein or initiate nonsense mediated decay (NMD), both of which are predicted to result in aberrant ZNF804A protein. Following this screen of the DNA archive a request to Harwell was made to recover the mutation into a mutant mouse.

2.1.2 ENU Random mutagenesis and Speed Congenics.

ENU random mutagenesis involves the use of a chemical mutagen, ENU (N-ethyl-N-nitrosourea) to induce germline mutations. The mutations are thought to occur at random throughout the genome at a rate of approximately 1.5-6 mutations per locus per 1000 mutagenised offspring but this may vary according to the mouse strain and dose of the ENU (Hitotsumachi et al., 1985; Quwailid et al., 2004). Once the ENU mutations in Zfp804a had been chosen the mutation was recovered into a mutant mouse using corresponding frozen sperm from the archive at Harwell via the *in vitro* fertilisation technique, as described in Coghill et al. (2002). The ENU strain was generated from ENU treated male Balb/c mice bred with C3H/HeJ females as part of the UK ENU mouse mutagenesis program. Mutations are then screened for in the F1 mutants (Nolan et al., 2000). A cohort was then bred from these mice in order to generate lines used in the present study. F1 heterozygote mutant males from Harwell with the C59X mutation were backcrossed onto a C57BL/6J background using a speed congenics approach (Markel et al., 1997; Visscher et al., 1999) by T. AlJanabi. Backcrossing is carried out onto a wildtype background to remove any potential confounding mutations, natural or ENU derived. Using backcrossing to achieve congenicity should ensure that any observations made are due to the mutation of interest rather than unwanted additional mutations. Mice were backcrossed onto the C57BL/6J background as this strain breed

well and are also genetically and phenotypically well characterised. Strain specific markers were used to identify mice heterozygous for the ENU mutation with the highest proportion of C57BL/6J background strain and these mice were backcrossed onto a C57BL/6 background. Using this approach, by the F3 generation the mice were predicted to have ~96% of the C57BL/6 background. F3 female and male, C59X heterozygotes were intercrossed to produce the $F3_i$ intercross generation. The $F3_i$ generation were estimated to have 96.13% C57BL/6J background. Initial expression analysis was carried out on the $F3_i$ generation. Higher levels of purity are ideal but slower progress in deriving the lines made this necessary on pragmatic grounds. Whilst there is the possibility of additional ENU mutations it is unlikely the same mutation would be found in two mice. At the start of the study, it was anticipated the impact of such mutations, if they were to arise, on expression results would be unlikely to cause group effects, although they might result in increased noise. To generate the experimental cohorts, two heterozygotes (F3) were intercrossed so that $F3_i$ homozygote, heterozygote and WT littermates were generated (T. AlJanabi).

Mice were housed in group cages of 2-5 mice in the Behavioural Neurosciences Laboratory in the School of Psychology, Cardiff University. Mice were kept under a 12 hour light dark cycle with lights on at 07:00 and lights off at 19:00 and were given food and water *ad libitum*. For breeding of the $F3_i$ generation up to four females were introduced to the male's homecage and left for one week. Females not pregnant after two weeks and infertile males were removed from the breeding programme.

2.1.3 Brain Tissue Collection

Mice were euthanized by cervical dislocation (schedule 1) following which the brain was extracted from the skull and immediately snap frozen in liquid nitrogen after which it was stored at -80°C until further processing.

2.2 DNA/RNA extraction

2.2.1 DNA Extraction.

Genomic DNA was extracted from mouse tail tips. Tail tips were immediately snap frozen in liquid nitrogen and stored at -80c until DNA extraction. Tails were lysed using a 400-500µl (dependent of tail size) mix of Protinase K and a tail lysis buffer and left overnight in a 55-60°c water bath. Samples were then spun for 10 minutes (13000rpm at 4°C) and the supernatant was transferred to a new tube. An equal volume of isopropanol was added to the supernatant, mixed and then left for 20-30 minutes at 2-8°C. Each sample was then centrifuged for 10 minutes (13000rpm at 4°C) and the supernatant was discarded. Samples were left to air dry for 1 hour and the resultant DNA was resuspended in 100µl of nuclease-free water.

2.2.2 RNA Extraction

1ml of Trizol (Sigma, St. Louis, MO) was added to brain tissue in a matrix tube (MP Biomedicals) and homogenised using a Bio-one homogenizer. Following centrifugation (12,000 x g for 5 minutes at 4°C) the supernatant was transferred to a new tube and allowed to stand at room temperature (RT) for 5 minutes. Chloroform (0.2ml/1ml of Trizol) was then added and the tube was shaken for 15 seconds before again being left to stand at RT this time for 10 minutes. Samples were centrifuged (12, 000 x g for 15 minutes at 4°C) before removing the upper phase to a fresh tube. RT isopropanol was then added at $1/10^{th}$ the volume in the tube and samples were left to stand for 5mins at RT before centrifugation (12, 000 x g for 10 minutes at 4°C). 0.5ml of isopropanol was added to the supernatant in a new tube and left to stand at RT for 10 minutes then centrifuged (12, 000 x g for 10 minutes at 4°C). The supernatant was then discarded leaving the RNA pellet in the tube. Following a wash with 1ml of 75% EtOH (made with nuclease-free water) the samples were centrifuged (12, 000 x g for 5 minutes at 4°C) and the supernatant was discarded. Pellets were air dried for ~ 10 minutes prior to resuspension in 50µl RNase-free water.

2.2.2.1 RNA Clean Up – Column Purification.

Following RNA extraction total RNA was purified to remove excess salts and contaminants using the RNeasy Mini Kit (QIAGEN, Hilden, Germany) according to the manufacturer's instructions.

2.2.2.2 DNase Treatment

Removal of contaminating DNA was facilitated by using the DNA-free kit (Ambion) which uses recombinant DNase1 for the digestion of any DNA present in the RNA sample. The procedure was carried out according to the manufacturer's instructions.

2.2.3 RNA quality assessment

RNA quality was determined using two measures; the ratio of two ribosomal RNAs (28s/18s) and the RNA integrity number (RIN) both of which were analysed using the Agilent 2100 Bioanalyser.

2.2.3.1 Agilent 2100 Bioanalyser.

A chip containing the RNA samples, a size ladder and a fluorescence marker was inserted into a Bioanalyser, microfluidic station. The principle behind the instrument is much like electrophoresis. A voltage gradient is run across the chip via an electrode and due to the negatively charged nature of RNA it migrates through a polymer matrix. Smaller molecules pass through the matrix easier, thus the matrix separates the molecules according to size. The dye molecules intercalate and migrate with the RNA and fluorescence is recorded using laser activation. An RNA 6000 ladder was run as a reference and contains 6 individual fragments which range in size from 0.2 – 6 Kb. Each sample was compared to the ladder fragments to determine its concentration and enable the identification of the rRNA peaks. Good quality, intact RNA is defined by the following features;

- Two clear and distinct ribosomal RNA peaks (28s and 18s rRNA).
- The baseline between the internal marker and the 18s peak is relatively flat. The absence of peaks in this region means there is no or very little smaller molecules in the sample which represent degraded rRNA or tRNAs. (rRNA is particularly sensitive to degradation if extracted from tissue using mechanical homogenisation or if the tissue has been frozen as both make shearing of the molecules more likely. The 28s rRNA is particularly sensitive to shearing as it is a larger molecule than the 18s rRNA).

2.2.3.2. 28s/18s rRNA Ratio.

The rRNA ratio is determined by dividing the 28s peak by the 18s peak. As 28s is the larger molecule the ratio is expected to be 2:1 (2>) if no degradation of the RNA has occurred. Due to the mechanical nature of tissue homogenisation this is rarely the case and a ratio of 1 or above is acceptable for most analyses (Ambion).

2.2.3.3 RNA Integrity Number (RIN).

The RIN metric is produced after running the RNA sample on an Agilent bioanalyser. The entire electropherogram of each sample is considered, not just the two rRNA peaks and a score from 1-10 is generated with 1 representing the most degraded RNA and 10 representing intact RNA.

RIN is generally a reliable metric, but it does not always equate that a good RIN score will mean the experiment being undertaken will be successful. It is also important to note that rRNA integrity is used to infer mRNA integrity. Both are usually comparable, but differences can occur as rRNA is considered more stable. For gene expression assays such as microarrays, the consensus is that a RIN of at least 7 is adequate, but 8 or above is preferable (Schroeder et al., 2006).

2.3 Reverse Transcription.

Using the Superscript II First-Strand synthesis system for RT-PCR (Invitrogen) and random primers (Invitrogen) 1µg of DNase treated total RNA was used as a template for cDNA synthesis according to the manufacturer's instructions.

2.4 Polymerase Chain Reaction

The polymerase chain reaction (PCR) allows the amplification of a specific DNA target. DNA is synthesised using the enzyme Taq Polymerase in addition to the 4 deoxyribonucleotide triphosphates (dNTPs), adenine (A), cytosine (C), guanine (G) and thymine (T) and standard buffer. Oligonucleotide primers are designed which are complementary to the sequence flanking the region to be amplified. The complementary strand of DNA is then synthesised in a 5'-3' direction with the two primers acting as the double stranded starting point, initiating DNA synthesis. A series of approximately 30-

45 temperature cycles facilitates the 3 steps of PCR to be carried out. The first step involves the denaturing of the double stranded DNA leaving a single stranded DNA template. The primers then anneal to their complementary sequence on the single stranded DNA. The final elongation step allows the synthesis of the complementary DNA strand by taq polymerase.

2.4.1 Primer design

Primers were designed using primer3 (v 0.4.0, http://frodo.wi.mit.edu/primer3/input.htm) and BLAST Primer http://www.ncbi.nlm.nih.gov/tools/primer-blast/), which uses a combination of the primer3 algorithm and BLAST to determine primer specificity against the mouse genome database (National Centre for Biotechnology Information). Specificity was also checked using BLAT (http://genome.ucsc.edu/cgi-bin/hgBlat) (Kent, 2002a). Default settings were used where possible so that primer annealing temperature was ~60°C and GC content was between 30 and 80%. Primers were synthesised by Sigma.

2.4.2 PCR Optimisation

PCR reactions were performed using C1000 Thermocyclers (BioRad Laboratories, Inc). HotStar Taq (Qiagen) was used in all PCR reactions. The standard PCR cycling conditions are outlined below. When an assay failed using these conditions, PCR was carried out on control samples to find the optimum temperature for primer annealing (Tm) using a temperature gradient. PCR was performed in 12µl reaction volumes using 3µl genomic DNA (4ng/µl), 0.56µl of each primer (5pmol concentration), 0.96µl dNTPs (5mM each), 1.2µl of 10X buffer (Qiagen), and 0.06µl of HotStarTaq polymerase (10units/µl, Qiagen).

For cDNA the 20µl product from the RT-PCR of 1µg RNA was diluted 1:5. Then 3µl added to the mastermix as described above.

PCR cycling Conditions.

1. 95°C for 15 minutes
2. 94°C for 20 seconds

3. ~60°C for 30 seconds
4. 72°C for 1 minute
5. Repeats steps 2-4 for 44 cycles
6. 72°C for 10 minutes
7. 15°C hold

2.4.3 Agarose Gel Electrophoresis

PCR products were separated by size and visualised using agarose gel electrophoresis, which is facilitated by the negatively charged phosphate groups of DNA. The porous nature of the agarose gel allows the negatively charged DNA to move toward the anode when an electric current is applied. The rate at which the DNA fragment moves toward the anode is a property of fragment size, smaller fragments passing with less resistance than larger ones.

Gels were 1-2% agarose, dependent upon the resolution required. A 1.5 % gel was comprised of 1.5g of agarose (AGTC Bioproducts) dissolved in 100ml of 0.5x TBE buffer (Ultra Pure electrophoresis grade, National Diagnostics). The solution was heated to dissolve the agarose and then cooled slightly once clear. 1.5µl of Ethidium Bromide solution (10mg/ml) was then added. The solution was poured into a gel cast, into which gel combs had been placed to allow well formation and allowed to cool to a solid.

Appropriate volumes of PCR product and a loading buffer were combined prior to loading into a well in the gel. 6x loading buffer was made up of 15% ficoll, 0.25% bromophenol blue, and 0.25% xylene cyanel in water. To determine the size of each fragment, 2µl of size standard (1kb plus DNA ladder, Invitrogen) was run alongside each row of samples. Gels were run in electrophoresis tanks at 100-120V for a time appropriate to separate a fragment of the expected size. DNA fragments were visualised using a UV transilluminator (UVP) and images were recorded using a Kodak Electrophoresis Gel analysis system.

2.5 Genotyping

Samples were initially genotyped by sequencing genomic DNA or mRNA using Sanger sequencing of exons 1-3 of Zfp804a (described in full in 2.6). Primers in Appendix 2.1.

An assay was then designed utilising high resolution melting analysis (HRMA) (described in full in section 2.7).

2.6 Sequencing

Sequencing was carried out using the Sanger Sequencing technique implemented using Big Dye termination chemistry (Applied Biosystems). Four fluorescently labelled dideoxy-nucleotide-triphosphates (ddATP, ddCTP, ddGTP and ddTTP) when incorporated, terminate DNA synthesis during primer extension. In doing so a series of nested fragments are produced each varying in length by a single base. By running the sample through a capillary on a 3100 capillary sequencer (Applied Biosystems, Foster City, CA) the fragments are sorted in size and a base specific fluorescent dye allows the identity of each terminal base to be identified using a fluorescence detector.

2.6.1 PCR Clean up

PCR products were purified using Ampure XP® (Agencourt). 12µl of the PCR product was mixed with 18µl of Ampure reagent using a Beckman-Coulter NX liquid handler. This enables the removal of contaminants such as salts, primers, DNA polymerases and unincorporated dNTPs. Ampure XP® consists of para-magnetic particles to which PCR amplicons bind. This facilitated the separation of PCR products from contaminants with the use of a magnet, to which the magnetic particles (bound with the DNA) adhered. Contaminants (not stuck to the magnetic beads) were then washed away using 85% Ethanol. Purified product was eluted in 195µl of nuclease free water.

2.6.2 Sequencing Reaction

5µl of clean product was added to 5µl of sequencing reaction mix. Sequencing reaction mix was made up of 1.917µl 5X BigDye sequencing buffer, 0.116µl BigDye termination mix, 1µl of forward or reverse PCR primer (4pmol/µl) and 1.917µl nuclease free water. The BigDye termination mix includes the four dNTPs which are unlabelled along with the four fluorescently labelled dideoxyribonucleotide triphosphates

(ddNTPs) and the Sequenase enzyme. The sequencing reaction consisted of the following steps:

1. 96°C for 2 minutes

2. 96°C for 10 seconds

3. 50°C for 5 seconds

4. 60°C for 4 minutes

5. Repeat steps 2-4 for 24 cycles

6. 4°C for 4 minutes

2.6.3 Post Sequencing Clean up

The clean up reaction was carried using the Beckman-Coulter NX liquid handler. Contaminants such as salts, primers, DNA polymerases and unincorporated ddNTPs were removed by mixing 10µl of PCR product with 7.5µl of CleanSeq® (Agencourt) reagent which contains magnetic beads to which the amplimeres bind. When the magnetic beads bind to a magnet, the unbound contaminants can be washed away with 85% Ethanol. Purified PCR products were then eluted in 90µl sterile water.

2.6.4 Sequencing Analysis of C59X in Zfp804a

Cleaned sequence products were passed to the School of Medicine Central Biological Services (CBS) for running on an ABI3100 36cm capillary sequencer (Applied Biosystems, Foster City, CA) with polyacrylamide POP6 (Applied Biosystems). The raw data generated were analysed using Sequence Analysis Software (Applied Biosystems) contained on the AB3100 PRISM genetic analyser. Each base was called according to its corresponding fluorescent signal. Genotype was then determined using a combination of NovoSNP (Weckx et al., 2005) and Sequencher (Gene Codes) software. In each instance the sequence traces of each sample were aligned to a reference sequence. The position of the C59X mutation was inspected in each sample to determine if the sample was a homozygous wildtype, C59X heterozygous or C59X homozygous mutant (Fig. 2.1).

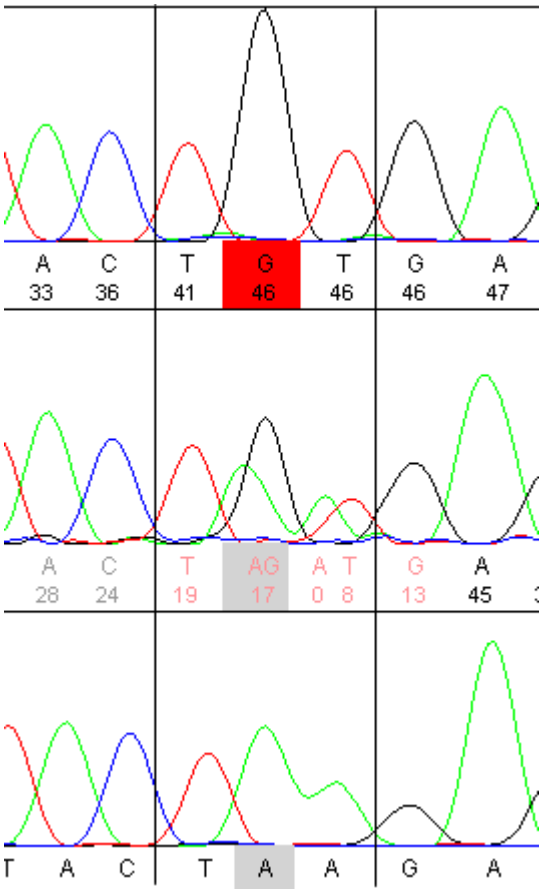


Figure 2.1. Sequencing the C59X mutation. Examples of sequencing traces from NovoSNP (Weckx et al., 2005). From top to bottom a homozygote wildtype mouse, a C59X heterozygote and a C59X homozygote mutant. Between the two vertical lines is the cysteine residue in the homozygote wildtype and the stop codon following a two base substitution from GT to AA in the C59X homozygote mutant.

2.7 High Resolution Melting Analysis (HRMA).

In order to quickly determine the genotype of each sample, an assay was designed which utilised the high resolution melting analysis technique. This analysis is based on the principle that the melting temperature of a PCR amplimere is a product of its sequence composition (Ririe et al., 1997). Utilising fluorescent dyes which bind to double stranded DNA, the melt curve of the DNA during the extension phase of PCR can be monitored in real time using changes in fluorescence to indicate the release of the fluorescent dye from single stranded DNA. Melt profiles are compared to a reference sequence and differences from the reference are observed as a change in the melting temperature (Liew et al 2004; Palais et al., 2005). Changes in the shape of the melting curve are indicative of hetero and homoduplexes, which are formed during PCR

of heterozygous loci (Graham et al., 2005). Homozygous wildtypes and C59X mutants were distinguishable from C59X heterozygotes, which meant the requirement for further homozygotes was quickly determined, however the distinction between wildtype and C59X mutant homozygotes required the sample to be further processed using Sanger sequencing to confirm genotype and so the attempt to make the genotyping process more efficient was ultimately not achieved.

2.7.1. HRMA PCR Conditions.

A 12µl PCR reaction was made up of; 4µl genomic DNA (4ng/µl), 0.56µl of each primer (5pmol concentration), 0.96µl dNTPs (5mM each), 1.2µl of 10X LCgreen Plus (Idaho Technologies), 1.2µl of 10x LCgreen Plus 20mM MgCl PCR Buffer (Idaho Technologies) and 0.06µl of HotStarTaq polymerase (10units/µl, Qiagen). Primers were designed to span the mutation in exon 2 of Zfp804a (Appendix 2.2)and the PCR cycling conditions were as described in 2.4.2.

HMRA was performed according to the manufacturer's instructions using a LightScanner (Idaho Technologies). 12µl of each sample was denatured by increasing the temperature at a rate of 0.1°C/s to a maximum temperature of 98°C. Fluorescent datapoints were collected continuously at a rate of 14 points/°C. Using a semi-automated analysis (Dwyer et al., 2009) the melting profiles were assessed. Once normalised, samples were analysed using the LightScanner software Call-IT^{TM} (Idaho Technologies) using the high sensitivity setting. The melt curve profile for each sample was plotted and then automatically called by the software by grouping samples according to similarity.

2.8 Global Analysis of Gene Expression.

2.8.1 Exon Arrays

The Affymetrix GeneChip Mouse Exon 1.0 ST array (exon array) was chosen for expression analysis as it facilitates fairly comprehensive and accurate measurement of gene expression changes as well as the identification of both known and novel splice events. The chip contains over 5 million probes targeting all known and predicted mouse exons. The difference between the exon array and the more traditional 3' *in vitro*

transcription (IVT) arrays is the removal of mismatch control probes for each of the probes on the chip, which frees considerable space for more probes to target exons throughout a transcript. Despite quite considerable differences in the design of exon chip relative to 3' arrays, gene expression performance has been shown to be comparable in several studies (Bemmo et al., 2008; Okoniewski et al., 2007) with sensitivity levels for detecting gene expression in the same range as the 3' IVT arrays (Abdueva et al. 2007). Each of the 5 million probes are 25 bases in length and have been synthesised on to the exon array which consists of a coated quartz surface. As described in detail below, fluorescently labelled RNA hybridised to the exon array and the fluorescent signal was used to infer and quantify expression levels.

2.8.2 RNA Labelling, Hybridisation and Scanning of Exon Arrays

All labelling, hybridisation and scanning steps were carried out by Central Biotechnology Services (CBS), Cardiff University.

2.8.2.1 RNA Labelling

Extracted total RNA was prepared for hybridisation to the exon chip using the Whole Transcriptome (WT) Expression Kit (Ambion) and the Affymetrix Genechip WT terminal labelling kit (Affymetrix, Santa Clara, CA, USA). RNA was reverse transcribed to first strand cDNA using specially engineered primers with a T7 promoter (Ambion). Primers were not complementary to ribosomal RNA (rRNA) sequences, removing the requirement to carry out an rRNA reduction step. Both poly-A and non poly-A mRNA (Ambion Protocol) was targeted.

Second strand synthesis was carried out using DNA polymerase followed by RNA degradation using RNase H. Using an *in vitro* transcription step complementary (antisense) RNA (cRNA) was synthesised using T7 RNA polymerase from the second strand cDNA template (Van Gelder et al., 1990). Transcribed cRNA was then purified using nucleic acid binding beads and isopropanol to remove any unwanted salts or enzymes. A second cycle of cDNA synthesis was then completed by reverse transcribing 10µg of cRNA using random primers. Fragmentation of the cDNA was carried out as part of the Affymetrix Genechip WT terminal labelling kit. To ensure reproducible and uniform fragments the kit incorporated dUTP into the DNA as part of the first-strand cDNA synthesis reaction in the second cycle. Both UDG (uracil DNA

glycosylase) and APE 1 (apurinic/apyrimidinic endonuclease 1) were used to treat the DNA and recognise and cleave at the unnatural dUTP sites during the fragmentation step. Fragments were approximately 25-200 bases in length. RNase H was again used to degrade the RNA, followed by clean up of the remaining single stranded cDNA. Clean-up was performed using nucleic acid binding beads and ethanol**.**

Samples were labelled with terminal deoxynucleotidyl transferase (TdT) using an Affymetrix DNA labelling reagent covalently linked to biotin. This method, in contrast to the traditional 3' IVT array, utilises random hexamer- linked T7 promoters to synthesise cDNA so that a DNA/DNA complex forms on hybridisation to the chip and amplification is not restricted to polyA RNA (Abdueva et al., 2007).

2.8.2.2 Exogenous Spike-in Controls.

Affymetrix kits include spike-in controls which are added to each sample prior to first strand synthesis and allow the user to assess the efficiency of the hybridisation process. These positive controls are a set of *Escherichia coli* genes; BioB, BioC, BioD and cre. The genes are not present in eukaryotic samples and so act as exogenous controls. Each is spiked-in at a known concentration and amplified simultaneously with the sample. The concentration of each control is such that a certain rank order of intensities is expected allowing the efficiency of the hybridisation process to be determined independent of sample quality (Affymetrix Protocol). Oligo B2 is also used as it specifically hybridises to probes placed at the corners of each array and is used by the Affymetrix console software to align grids to the chip (Bolstad, 2008).

2.8.2.3 Hybridisation and Scanning of the Exon Array

A hybridisation cocktail made up of the labelled cDNA and the controls was inserted into the exon chips (Fig. 2.2). Following 16hr hybridisation, the hybridisation cocktail was removed and replaced with wash buffer. Chips were then washed and stained with a series of stain cocktails (Genechip Hybridisation, Wash and Stain Kit, Affymetrix) using a fluidics station 450 (Affymetrix). Streptavidin-phycoerythrin (SAPE) was used to stain the chips, binding to the biotin label (Bolstad, 2008). Each chip was inserted into the scanner (GeneChip Scanner 3000 7G, Affymetrix) and the Affymetrix GeneChip Command Console (AGCC) was used to control the scanning process. A raw

image or .dat file was initially generated. Each exon array contains features or squares approximately 3μm in size. Raw data were aligned to the grid system using the Oligo B2 controls (2.8.2.2) allowing the intensity of each feature to be determined (Bolstad, 2008). Intensity of each feature was calculated by considering only pixels which resided within the feature, not those found on the border (Bolstad, 2008). The intensity data for each feature or probe cell is then acquired by AGCC to generate the probe cell intensity file (CEL file) which is commonly the starting point for exon array analysis. The chip was then ejected from the scanner.

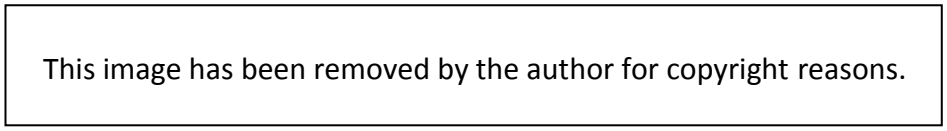


Figure 2.2 The Affymetrix GeneChip Mouse Exon 1.0 ST Array. Hybridisation cocktail, including the labelled cDNA sample is inserted into the chip via the septa. **(**Diagram Courtesy of Affymetrix from the Affymetrix GeneChip® WT terminal labelling and hybridisation user manual).

2.8.3 Expression Analysis

2.8.3.1 Partek Genomics Suite.

Partek Genomics Suite (version 6.5 and beta 6.6, St. Louis, MO) is a purpose built software suite designed to enable analysis of a number of high-throughput technologies.

2.8.3.2 Data upload

Initially samples were uploaded into Partek as CEL files. With each set of CEL files a corresponding sample sheet was prepared and uploaded. This contained information such as sample ID, age, gender and genotype and allowed the grouping of samples according to these attributes. The import process was customised to allow samples to be preprocessed with the criteria required for the specific experiment.

2.8.3.3 Preprocessing

Once raw intensity values (CEL files) are generated several steps must take place in order to produce meaningful expression data. These steps collectively are termed preprocessing. The purpose of preprocessing is to standardise the data obtained across all of the chips in a study to produce results with minimal batch differences and to identify any potential outliers. Preprocessing consists of 3 discrete steps, which I carried out in the following order; background correction, normalisation and summarization.

2.8.3.3.1. Robust Multichip Averaging (RMA).

The method chosen to carry out preprocessing was the Robust Multichip Averaging (RMA) procedure (Irrizarry et al. 2003) which does not rely on mismatch (MM) probe values. Previous Affymetrix arrays contained a corresponding MM probe for each perfect match (PM) probe on the array. The MM probes differ from the PM probes at the 13^{th} base only and as such are used to measure non-specific binding. As the exon array has no MM probes, RMA is therefore suitable for exon array analysis. RMA was carried out using both Partek Genomics Suite (Partek Incorporated, St Louis, USA) and Expression Console™ V1.1.2 (Affymetrix). In each case only the core probes were included. During preprocessing using RMA, raw data (X) undergo background correction (B), normalisation (N) and then summarisation (S) to produce expression data (E) so that $E = S(N(B(X)))$.

2.8.3.3.2 Background Correction.

RMA uses a convolution model for background correction. The idea is that the probe signal (S) will consist of both signal (x) and background (y), ($S = x + y$). The model uses a smoothed density plot to assume that x will be distributed exponentially (α) and that y is distributed normally ($N(\mu,\sigma^2)$). Background correction uses the observed signal to predict the expected signal intensity for each probe, once non-specific signal has been removed (Bolstad, 2008). At this step each array is corrected independently using values found only on that array.

To obtain the estimates a nonparametric test is used. PM probe intensities are plotted, and then a density estimator is fitted, which estimates the mode of the distribution. Anything above the mode is used to predict the exponential parameter. The normal

distribution is determined by fitting a half normal to anything below the model (Bolstad, 2008; Irrizarry et al., 2003a)

2.8.3.3.3 Normalisation.

Normalisation is required to remove obscuring or technical variation, which is not a result of true biological differences between samples. Normalisation was carried out using the quantile normalisation method. Previous studies, reviewing several different methods showed quantile normalisation worked most effectively (Bolstad et al., 2003) and allowed more sensitive and specific detection of differential expression when using GeneChips, which have a 1 sample/array design (Irrizarry et al., 2003).

The aim of quantile normalisation is to make the distribution of the probe intensities the same for each of the arrays in the experiment. Normalising the distribution of each array to the mean distribution should remove inter array variance. No reference array is used, instead each arrays' probe level values are sorted (ranked). Quantile normalisation is carried out at the probe level for every probe on each array (Bolstad et al., 2003). The non-parametric, quantile normalisation algorithm uses a matrix and essentially sorts and averages across rows and columns representing probes and arrays respectively. Initially the probe intensities are sorted from lowest to highest for each of the arrays. Then the average of each quantile is determined by averaging probe intensities across each row of the dataset. The columns are then rearranged so that each probe intensity value goes back to its original ranked position (Bolstad et al., 2003).

2.8.3.3.4 Probe Summarisation.

A summary measure called median polish was carried out on background adjusted, normalised and $\log_2$ transformed PM values, to estimate log scale expression values (Irrizarry et al., 2003b). This was performed by fitting a robust linear model at the probe level, ensuring any probe-specific affinity differences would have a minimal effect. The RMA model uses the assumption that a probe's expression level is determined by how much RNA there is available to bind to the probe (the chip effect, *e*) the affinity of the probe (*a*) and the error in measurement (ε) applied to the following formula:

$$PM_{ij} = e_i + a_j + \varepsilon_{ij}$$

Where *i* represents the array and *j* the probe. Based on the intensity value for a particular probeset (PM_{ij}) the algorithm determines the possible combinations of *e* and *a* that would result in the observed PM value (Irrizarry et al., 2003b).

Each probe in an Affymetrix probeset interrogates a different segment of a gene (Gardina & Turpaz, 2008). Summarisation acts to combine the signal intensity for probes within a probeset to obtain a single expression value for the probeset (Bemmo et al., 2008).

2.8.3.4 Annotating the Dataset.

Probes on the Affymetrix exon array are annotated as either core, extended or full. The labels denote the annotation confidence of the sequence targeted by each probe. Probes which target Refseqs or mRNAs from Genbank have the highest level of annotation confidence and are termed core probesets. Probesets with an extended annotation, target sequences defined using expressed sequence tags (ESTs). Probesets annotated as full have the lowest annotation confidence and target transcripts annotated using *ab initio* prediction software.

All analyses reported in this thesis were restricted to the core metaprobe set, unless otherwise stated which target genes which have been sequenced, cloned and curated manually (Gardina & Turpaz, 2008). By removing more speculative content the likelihood of false positive calls was reduced (Gardina & Turpaz, 2008). All core probesets have been defined by Affymetrix as unique, meaning they should not cross-hybridise. Restriction to the core probesets meant a total of ~15,000 genes, which were either RefSeq genes or full length mRNA from GenBank were included.

There is no perfect congruence between a probeset and an exon. Each probeset covers what Affymetrix define as a probeset region (PSR). Each PSR represents a region of the genome which is considered to be independent unit (Robinson & Speed, 2009) and is generally represented by 4 probes. Data were analysed using the Affymetrix annotation files, NetAffx, version na31. mm9.

2.8.3.5 Quality Control Measures.

Standard QC measures were undertaken to identify potential dataset outliers that might indicate technical problems in the analysis of individual samples. A series of metrics were generated using Affymetrix Expression Console (Version 1.1.2) and Partek GS (v6.6). All metrics were generated following RMA at the gene level on core probes. Using only the core meta-probeset removes the potential of increased variability due to the higher rate of unexpressed genes found in the extended and full annotations.

Both the mean absolute deviation (MAD) and relative log expression (RLE) metrics were chosen as they are robust against experimental conditions (Gardina & Turpaz, 2008). Present/absent calls were generated at the probeset level as detection p values. Whilst no standardised cutoffs currently exist, the QC measures were useful for identifying potential outliers, which if present were removed or closely monitored in downstream statistical analysis (Gardina & Turpaz, 2008).

2.8.3.5.1 Affymetrix Expression Console (v1.1.2)

Following download of the appropriate library file (MoEx-1_0-st-v1.) from Net affx (Affymetrix), CEL files were uploaded and summarised to produce a probe level summarisation file (CHP file). Exon arrays were analysed using the RMA-sketch workflow as this allows both gene and exon level analysis (Affymetrix).

2.8.3.5.2 Principle Component Analysis (PCA).

To visualise similarities and differences in the expression data PCA was carried out using Partek GS. PCA reduces the dimensionality of the data to a few components which between them explain most of the variance in the data.

2.8.3.5.3 Exogenous Spike-in Controls.

As described in 2.8.2.2, exogenous controls added at known concentrations were used to assess the efficiency of the hybridisation, wash and scanning procedures. Adequate efficiency was denoted by the correct rank order in the intensity of each control.

2.8.3.5.4 All Probeset (and Positive Control) RLE Mean.

The 'all probeset RLE mean' measures the mean absolute relative log expression (RLE). To determine this statistic the signal of each probeset on an array is compared to the median signal value across all arrays in the study. The metric is the mean of these differences from all the probe sets. This is the measure of how different a sample is relative to the consensus, with very high values denoting an array with different signals from others in the experiment (Affymetrix, 2007).

2.8.3.5.5 Perfect Match (PM) Mean.

This metric is based on raw intensity data and represents the mean intensity for all perfect match (PM) probes on the array before pre-processing (e.g., RMA) (Affymetrix, 2007).

2.8.3.5.6 All Probeset (and Positive Control) Mean Absolute Deviation (Mad) Residual Mean.

Different probes will give out different intensities even when a common target binds to them. RMA creates a model for these individual probe responses and arrays with multiple probes behaving differently to the model can then be identified. Differences between predicted and actual values are defined as the residual. Each probe will have a residual value from the model. An individual probe residual value that varies from the median reflects a poor fit to the model. By determining the mean absolute deviation, the overall fit to the model of every probe on the array can be established. If the residuals have a very large mean absolute deviation from the median value this is indicative of arrays with poor quality data (Affymetrix, 2007).

2.8.3.5.7 Positive vs. Negative AUC.

This metric measures the area under the curve (auc) of a receiver operating characteristic (ROC) plot. The ROC curve is a plot of the detection of positive controls against the false detection of negative controls (exon and intron probesets respectively, which target ~100 constitutively expressed genes). To generate the curve it must be determined if the probeset signals effectively separate positive control signals and negative control signals, which measure true and false positives respectively. It is a robust measurement for overall data quality often used as a first pass metric. Typical

values range between 0.8 and 0.9. In theory a value of 1 for this metric would indicate perfect separation, but a value of 0.5 or below indicates no separation and thus no difference between positive and negative controls (Affymetrix, 2007). Whilst values falling significantly below 0.8 may indicate an outlier, values above 0.8 do not necessarily indicate good quality data (Affymetrix, 2008).

2.8.3.6 Statistical Analysis

Statistical analysis was carried out using Partek Genomics Suite (v6.5 Partek Inc, St. Loius, MO, USA).

2.8.3.6.1 Differential Expression Algorithm.

Exons were first summarised to genes using the mean of probeset intensities. The core gene summary file was then used as the input file for the statistical analysis, performed to determine differentially expressed genes between wildtypes and C59X mutants. A 1-way ANOVA model was used which implemented the Method of Moments (Eisenhart, 1947). The model included:

$$Y_{ij} = \mu + \text{Genotype}_i + \varepsilon_{ij}$$

In this model Y_{ij} represents the j^{th} observation on the i^{th} Genotype. μ is the common effect for the whole experiment. ε_{ij} represents the random error present in the j^{th} observation on the i^{th} Genotype. Errors ε_{ij} are assumed to be normally and independently distributed with mean 0 and standard deviation δ for all measurements.

In addition to determining differentially expressed genes a linear contrast between 2 specific groups within the context of an ANOVA was performed to determine fold changes between wildtype and C59X mutants. Fisher's Least Significant Difference (LSD) method was used to determine fold changes between wildtypes and C59X mutants (Tamhane and Dunlop, 2000).

2.8.3.6.2 Filtering

Whilst gene level analysis does not vary from the traditional 3' IVT workflow, exon level analysis requires additional steps to control the false positive rate. These include filtering data prior to statistical analysis thus reducing the need for multiple test correction (MTC) and visual inspection of data (Affymetrix, 2006). The most common

filtering procedure is the removal of probesets based on low annotation confidence and as described previously only probesets with the highest annotation confidence (core probesets) were included. It is also common for probesets to be removed based on expression signal. This is particularly relevant to exon arrays as low expression can be misinterpreted as differential splicing. Probesets were filtered based on signal intensity values. Initially probesets with a maximum $\log_2$ signal < 3 were excluded from the statistical analysis. This threshold was then increased to 4, 5 and 6 and the group mean log2 intensity was also considered. Final analyses were based on criteria that excluded any probeset with a group mean, $\log_2$ intensity signal <6.

2.8.3.6.3 Alternative Splicing Algorithm

Alternative splice p values were generated using Partek Genomics Suite's custom alternative splice ANOVA. A one-way ANOVA was used for the individual female and male experiments, but for the combined analysis a two-way ANOVA was used, which included gender as a factor in addition to genotype. In both cases the ANOVA implemented the Method of Moments (Eisenhart, 1947). In all instances genotype was used as the alternative splice factor, allowing splice differences between C59X mutants and wildtypes to be identified. Differential splicing in C59X mutants was identified using the following model:

$$\gamma = \mu + G + E + G*E + S(G) + \varepsilon$$

Where γ is the expression of the transcript

μ is the mean expression of the transcript,

G is the gene expression differences between the two levels of genotype

E is the differential exon expression, independent of genotype

G*E is the interaction of splicing and genotype (differential exon expression between mutant and wildtype),

S is the sample effect, denoted as both a random effect (this assumes animals used are representative of the population and these exact samples would not be represented

again) and nested in genotype (meaning no one sample belonged to both genotype groups)

and ε is the error term

The ANOVA uses an interaction term between the two groups and the probeset (G*E) to see if probeset expression varies at the two levels of group (C59X mutants and wildtypes) (Bemmo et al., 2008). Both differential expression and alternative splice p values are generated from this model which is performed at the exon level. To determine fold change at the transcript and exon level a linear contrast was included in the model between C59X mutants and wildtypes using the Fisher's Least Significant Difference (LSD) method (Tamhane and Dunlop, 2000).

2.8.3.6.4 Multiple Test Correction (MTC) and Non-independence of data.

Carrying out multiple tests on the same dataset increases the chance of falsely rejecting the null hypothesis. This problem is particularly applicable to the exon array as 1.4 million tests would need to be carried out to analyse all the data on the array. It has been estimated that this could result in as many as 70,000 false positives (5%) (Okoniewski & Miller, 2008). Conventionally this type of issue is corrected using a Bonferroni Correction (Holm, 1979) in which the desired p value threshold (e.g., $p<0.05$) is divided by the number of test being carried out. This type of correction conserves the family-wise error rate (FWER) by reducing the probability that any given individual significant result represents a type I error. Using such a stringent correction does however have a direct consequence on the number of type-II errors generated (the rejection of true effects).

Whilst the application of Bonferroni MTC to the array data is easy to implement to standard array results the exon array is more complex. Due to data being generated at the exon level, one of the key assumptions is violated as there is non-independence between the probesets of a gene.

Currently, there is no consensus on how to deal with multiple testing in exome level data. Some recommend pre analysis filtering to reduce the number of tests by as much as possible to reduce the necessity for MTC (Della et al., 2008; Okoniewski and Miller 2008; Whistler et al., 2010) but this offers no general solution.

In the absence of an accepted approach, I applied both Bonferroni correction and the false discovery rate (FDR) (Benjamini and Hochberg, 1995) methods to gene expression and alternative splice data. The FDR ascribes a threshold to the significant data so that it will contain a predefined number of false positives (Okoniewski and Miller 2008). It can also give an idea of how reliable the dataset is overall (Okoniewski and Miller 2008). An FDR threshold of 0.05 was set. The Bonferroni correction is the most conservative MTC method giving increased confidence that findings surviving this stringent correction are true findings.

2.8.3.7 Visualisation of Alternative Splice Events.

Predicted splice changes were visualised using the geneview produced as part of the alternative splice output. Probeset intensity was plotted for each group. To compare the results to known gene annotations, the data were also visualised using the UCSC genome browser (Kent et al., 2002b) (http://genome.ucsc.edu/). This was done by first identifying the sequence of the probeset using the probeset ID entered into NetAffx (http://www.affymetrix.com/analysis/index.affx). Then this sequence was BLATed in the UCSC genome browser (Kent et al., 2002a).

2.8.4 Alternative Algorithms for Detecting Differential Splicing.

The following software programmes and algorithms were used in addition to Partek GS and the Alternative Splice ANOVA to establish how robust the results generated were.

2.8.4.1 easyExon

easyExon (Chang et al., 2008) is a software tool developed for the assessment of CEL files to determine differentially expressed and spliced genes. The software was used in addition to Partek GS as the different filtering criteria and statistical algorithms implemented allowed the opportunity to determine how robust events identified in Partek GS were by determining if they replicated with these different criteria. easyExon was launched from the command line and CEL files were uploaded and preprocessed using Affymetrix Power Tools (APT), which generated the summary files required for statistical analysis. The RMA-sketch preprocessing algorithm was selected using the

mm9 mouse database and the core meta-probeset. Detection above background (DABG) p values were also generated and included in the summary files. This measure allows distinction between probesets expressed above and within the background signal. A significant p value ($p \leq 0.05$) represents a probeset with signal above background signal. Probesets were excluded if DABG p value was >0.05 in at least half of the samples (4/8) (default setting). Any probesets not meeting the criteria were represented in grey in the graphical representation. A log stabilisation factor of 16 was added to the summarised signal prior to log transformation. Transcript clusters were excluded if they contained less than 4 or more than 200 probesets as visual inspection is difficult when probesets exceed these limits. The statistical algorithm chosen was MiDAS (as described in 2.8.4.3). A MiDAS p value of ≤ 0.05 was considered significant.

2.8.4.2 AltAnalyze (v.2.0.7)

AltAnalyze is open access software for alternative splice analysis (Emig et al., 2010). After specifying the use of an Affymetrix platform, the species *mus musculus* and the Exon ST array, CEL files were processed. Prior to calculations of gene expression probesets with large cross-hybridisation scores were removed. The core meta-probeset was used to determine both gene expression and alternative splicing so as to be consistent with the Partek GS analyses. This is in accordance with the Affymetrix recommendation that core probes are used to determine constitutive gene expression. It is important to note at this stage, that the way in which AltAnalyze defines the core probeset differs slightly to that in Partek. Whilst the Affymetrix annotated core probesets form the core set of probesets, any probesets which uniquely align to a single Ensembl gene are included in addition. This means the AltAnalyze defined core meta-probeset contains a larger number of probesets therefore probesets may be included in the analysis which wouldn't be present when carrying out core probeset analyses in Partek.

Probesets were required to have detection above background (DABG) p values ≤ 0.05 and an accompanying non-log expression value greater than 70 to be included in the analyses. Gene expression was first calculated and the above expression criteria had to be met in at least one of the experimental groups for the probeset to be included. Gene

expression was calculated by averaging the expression of all core probesets aligning to a particular gene. Gene expression measures are also used to normalise probeset expression when determining differential splicing. In this case the probeset had to be expressed (as defined by the aforementioned criteria) in both groups. This excludes the possibility of a probeset being predicted as differentially spliced when in fact the gene is not expressed in one of the groups. Genes were excluded from differential splice analysis if none of the probesets within the gene met the criteria. The FIRMA algorithm (Purdom et al., 2008) was chosen for the differential splice analysis again using the core probesets. Mutants were defined as the experimental group and wildtypes as the baseline group. By default genes with differential expression fold changes above 3 are excluded from the differential splice output as the differential splicing is likely to be a consequence of the differential gene expression.

Differential expression is determined using a 1-way ANOVA and adjusted for multiple-test correction using the Benjamini and Hochberg (1995) FDR. Fold changes are calculated using a geometric subtraction of the experimental group from the wildtypes. Fold change is the non log2 transformed fold value.

In AltAnlayse the FIRMA output differs slightly to the method initially proposed by Purdom et al. (2008) (described in 2.8.4.4) as only summary statistics are presented, which are generated by taking the average FIRMA score for the control group and subtracting it from the average FIRMA score of the experimental group.

2.8.4.3 Microarray detection of Alternative Splicing (MiDAS).

Microarray analysis of differential splicing (MADS or MiDAS) was developed to overcome the increased noise found on Affymetrix arrays due to having ~4 probes per exon (Xing et al., 2008). This increased noise was thought to impair the detection of true splice events. MiDAS is based on the ANOVA algorithm and determines differences in exon and gene level signals. The probe logarithmic intensity error (PLIER) algorithm is implemented to generate gene level signals from all probes within each exon of the gene (Affymetrix, 2005b). Background noise is removed by using the median intensity of GC matched antigenomic probes on the array then exon level expression is estimated using PLIER. The method assumes that under the null hypothesis, if an exon is not differentially spliced the log expression at that exon will not differ from gene level signal for all samples (Affymetrix, 2005b). The ratio between

exon and gene signal is determined using the splicing index logged. Variance is stabilised by adding a constant prior to logging.

2.8.4.4 FIRMA.

Finding isoforms using Robust Multichip Analysis (FIRMA) is a robust RMA model, fitted at the probe level (Purdom et al., 2008). FIRMA determines how consistent expression is at a probeset relative to transcript expression within a particular sample. Initially each gene's expression level is estimated using the RMA model (Irizarry et al., 2003) then alternative splicing is determined using a score generated from the estimation step. Each exon is given a score based on how much its probe's signals deviate from the expected gene expression level (Purdom et al., 2008). FIRMA is a more general additive model than the RMA model (described in 2.8.3.3.1). First the RMA model is fitted then the residuals from this are produced. A score for each exon (j) and sample (i) is generated based on the median of the four residuals (one for each probe) of that exon (j) and sample (i). This score is used to determine how much the exon signal differs from the expected transcript expression.

Chapter 3: Expression analysis in Zfp804a ENU mutant mice

3.1 Introduction.

The availability of a mouse with disrupted Zfp804a offers the opportunity to investigate consequences of that disruption. In doing so, I aim to identify possible disease mechanisms. Several approaches could be taken to achieve this, but I chose gene expression analysis, due to the availability of accurate and efficient global assays for the quantification of mRNA expression relative to global proteomics assays for mice, with the caveat that mRNA abundance would not necessarily equate to protein abundance due to post-transcriptional regulation processes. Measuring mRNA in the brains of mice with disrupted Zfp804a would allow consequential altered expression of downstream targets to be identified and if achieved this would provide evidence that Zfp804a does have a function in the regulation of the expression of other genes. Identifying the biological pathways, which subsets of these genes belong to may help identify molecular mechanisms relevant to the aetiology of schizophrenia and psychosis. By comparing gene expression in Zfp804a mutants and wildtype mice using a global exon microarray, expression or splicing changes between the mutants and wildtypes could highlight genes influenced by Zfp804a which could be implicated in schizophrenia and would imply a direct or indirect role for Zfp804a in the regulation of expression and or splicing. Determining a role for Zfp804a in splicing regulation irrespective of which genes are spliced would be informative. In addition to elucidating ZNF804A function other schizophrenia susceptibility genes, could be determined, which may allow a more comprehensive understanding of pathways affected in schizophrenia aetiology.

To investigate the consequences of disrupted Zfp804a on gene expression and splicing, studies could be carried out using mouse models or cell lines. Whilst both have certain advantages and limitations, as discussed below, I chose to carryout global expression analyses using a mouse model, however complementary work using cell lines was also carried out by another PhD student. The mouse as a model organism offers high system complexity and a high degree of conservation with humans. The similar anatomy of the mouse and human brain is another advantage of using mice as model organisms as well as the availability of a broad range of validated behavioural tests, the results of which

can be used to complement genetic studies (Eriksen & Janus, 2007). In addition extensive research carried out using mouse models has facilitated the acquisition of a large amount of data and with the use of inbred strains of mice, assurance that mice will be genetically almost identical, with the exception of naturally occurring new mutations. The C57BL/6J strain used for backcrossing (described in chapter 2.1.2) was the first mouse strain to have its genome sequenced and as such is genetically well characterised (Mouse Genome Sequencing Consortium et al., 2002). Mouse models are commonly used in expression studies to understand the function of human genes in diseases such as amyotrophic lateral sclerosis (ALS) (Chen et al., 2010) and Huntington's disease (Morton et al., 2005). The short gestation of the mouse and ability to easily manipulate and monitor their external environment makes the mouse model extremely suitable for studies investigating gene expression. Certain caveats need to be considered when translating findings from mouse models to human disorders. Results can be affected by genetic variation among mice, phenotypic differences can occur dependent on the strain of mouse used and biochemical differences can prevent the same mutation in the homologous gene creating the same phenotype in a mouse (Erickson, 1996). Differences in mouse and human life-spans are not always relative in terms of disease for example in Duchenne's muscular (DM) dystrophy the mutation of the gene *Mdx* is not symptomatic in mice and may reflect that the disease is not normally diagnosed until the second year in humans which is the entire lifespan of a mouse (Erickson, 1996). Generating mouse models can be very time consuming and costly in contrast to studies using cell lines which are relatively cost effective and cells can be readily manipulated, but the caveat with this work being that they do not necessarily provide an accurate model of the cell in vivo.

A genome-wide significant and replicated association at the *ZNF804A* locus with schizophrenia (O'Donovan et al., 2008; Stefannson et al., 2009; ISC 2009, Riley et al., 2010; Williams et al., 2010; Steinberg et al., 2011; Zhang et al 2011a 2011b) suggests that despite conferring a small effect *ZNF804A* is a schizophrenia susceptibility gene. It is important to consider that association studies point to a region not a gene, but there is no evidence for any other functional units in the region. At the time of the first published association, very little was known about the gene and its encoded protein and

four years later there is still very little published evidence pertaining to the function of ZNF804A and how it may influence disease risk.

Given the location of the associated SNP (rs1344706) within intron 2 of *ZNF804A*, the most likely predicted function is regulation of transcription or splicing leading to the hypothesis that risk is inferred via altered regulation of gene expression or RNA processing. Fine mapping and resequencing have not uncovered a more strongly associated SNP, nor was rs1344706 found to be in strong linkage disequilibrium (LD) with any other variant. This led the authors to tentatively conclude the association is unlikely to be because rs1344706 is in LD with another variant that is causal (Williams et al., 2010). rs1344706 is located within a ~30bp region of conserved mammalian sequence (Donohoe, et al. 2010), hypothesised to show this degree of conservation due to the presence of transcription binding sites (Riley et al 2010). Higher ZNF804A RNA expression is associated with the risk allele in post-mortem brain tissue (Riley et al. 2010; Williams et al., 2010).

In the presence of the rs1344706 risk allele (T), the adjacent sequence was predicted, using bioinformatics, to be a binding site of the brain expressed transcription factors Myt1L and POU3F1/Oct655. Both are predicted to have functions in the development of the CNS and in particular oligodendrocyte development (Nielsen et al,. 2004; Collarini et al., 1992). Sequence including the protective allele (G) was predicted to be a binding site for two other transcription factors, the ubiquitously expressed Homez and the CNS expressed Hmx2 (Riley et al., 2010). Evidence that rs1344706 does form sequence-specific DNA-protein complexes with nuclear binding proteins was found using electromobility shift assays (EMSA) and highlights rs1344706 as a functional variant (Hill & Bray, 2011). The prediction that the nuclear transcription factors Homez and Hmx2 bound to sequence containing the rs1344706 G allele (Riley et al., 2010) was not, however confirmed (Hill and Bray, 2011) and the identity of the nuclear binding protein(s) remains unknown. The intensity of nuclear protein binding was increased when the protective allele (G) was present in the oligonucleotide sequence relative to the risk allele (T). Allele specific alterations in DNA-protein complex formation may explain the association of the risk allele with increased ZNF804A mRNA expression (Riley et al., 2010, Williams et al., 2010). These findings do not rule out the presence

of other functional variants in *ZNF804A*, but they do provide a functional basis for a direct association involving rs1344706.

Within *ZNF804A*, also within intron 2 approximately 3kb downstream of the associated SNP is another highly conserved region between mouse and human (Zhang et al., 2011b; using UCSC Browser, 2009). The conserved nature of this region may suggest its involvement in transcription factor binding or splicing. Interestingly, one variant within this conserved region, rs13423388 showed some evidence for association with schizophrenia (Zhang et al., 2011b) but no studies to date have determined if allelic differences at this SNP affect ZNF804A expression.

The ZNF804A protein has been predicted to contain a single zinc finger domain. This domain, specifically a Cys2His2 (C2H2)-like fold group, is the best characterised of the many classes of zinc fingers. C2H2-type domains are known for their sequence specific DNA-binding properties, as well as protein-protein interactions and RNA binding (Gamsjaeger et al., 2007) and are commonly found in transcription factors. Thus a role for ZNF804A in the regulation of gene expression has been proposed. ZNF804A has two paralogues ZNF804B and GPATCH8. Like ZNF804A, little is known about these two genes affording little insight into the possible functions of ZNF804A. However, *GPATCH8* is thought to encode a protein with a zinc finger domain and an RNA processing domain (Kaneko et al., 2011).

To date, two independent studies have evaluated the effects of altered ZNF804A expression on downstream gene expression. Knockdown of ZNF804A in a neural cell line resulted in 154 consistent expression changes between mutant and wildtype in two siRNA experiments (Hill et al., 2012), more than would be expected by chance. Pathway analysis of the corresponding transcripts revealed enrichment for genes in biological adhesion pathways as well as the subsidiary cell adhesion pathway (Hill et al., 2012). Overexpression of ZNF804A in E11 rat forebrain progenitor cells resulted in expression changes in 4 of 37 previously implicated schizophrenia genes (Girgenti et al., 2012). At the cellular level there is reason to believe from two studies that ZNF804A has a direct or indirect effect on gene expression. To evaluate if ZNF804A directly or indirectly regulates expression a Chromatin immunoprecipitation assay (ChIP) was utilised. Binding of ZNF804A was observed in the promoter regions of 2 of the 4 genes, PRSS16 and COMT. Both genes contained motifs predicted to interact with

zinc fingers consistent with the CHiP binding and a direct effect of ZNF804A on expression (Girgenti et al., 2012).

Whilst there is a growing body of evidence pointing to a role for ZNF804A in transcription regulation, no study to date has considered the implications of altered ZNF804A on splicing. The proportion of splice events predicted to be conserved between human and mouse are small (Nurtdinov et al., 2003; Sorek et al., 2004). The splice events that are conserved may be indicative of isoforms with critical biological functions (Sorek et al., 2004).

Alternative splicing is now thought of as the primary mechanism responsible for generating much of the diversity observed in the human proteome. Alternative splicing involves the processing of multiple mature mRNA transcripts from a single precursor mRNA and is thought to occur in as many as 50% of human genes (Johnson et al., 2003) and ~95% of multi-exon genes (Wang et al. 2008; Pan et al. 2008). Approximately 15% of point mutations leading to Mendelian disease do so by affecting splicing (Johnson et al., 2003), with aberrant splicing being associated with a number of diseases including cancer (Venables, 2006) and cystic fibrosis (Faustino & Cooper, 2003). The omission or inclusion of functional domains via alternative splicing can alter the function of the encoded protein. Alternative splicing can also affect protein function by altering the affinity of the protein to other proteins or ligands (Yeo et al., 2005). Factors influencing alternative splicing include the secondary structure of the mRNA, the elongation rate of RNA polymerase II and perhaps most influential, RNA binding proteins, which bind to cis-elements found in both the exon and the intron of the pre-mRNA and either enhance or silence splicing. The length of the exon and intron as well as the strength of the splice site have also been implicated in splice regulation (Yeo et al., 2005) as have epigenetic mechanisms (Luco & Misteli, 2011). Changes in chromatin conformation can alter the elongation rate of RNA polymerase II. A faster elongation rate increases the likelihood that multiple splice sites will become accessible to the splicing machinery essentially simultaneously. If one splice site is weaker than the other the machinery is likely to be recruited to the stronger splice site due to the competitive nature of the binding. Studies have shown that dense nucleosome regions slow the elongation rate of RNA polymerase II (Hodges et al., 2009). This increases the time for spliceosome assembly at weaker splice sites promoting the inclusion of cassette exons. There is also thought to be an enrichment of histone marks at exons. The histone

marks are thought to recruit RNA binding proteins, which when bound to the pre-mRNA can enhance or repress splicing (Luco & Misteli, 2011).

Microarray analysis has, until recently been the primary technique used to study global gene expression. Its high throughput capabilities revolutionized the way in which the transcriptome was studied and with lowering costs, the use of arrays has enabled large amounts of expression data to be generated. However, the completion of the human genome project afforded a better understanding of the mechanisms important in generating proteome diversity namely alternative splicing, thus the shortcomings of traditional microarrays were highlighted. The placement of probes at the 3' end of the transcript meant alternative isoforms in which the 3' exon was spliced out were not assayed and that those with common 3' ends could not be distinguished. From these arrays an incomplete picture of the transcriptome was generated as well as potentially incorrect measurements of total gene expression. The necessity to incorporate and measure expression of alternatively spliced isoforms is now widely recognised and investigators can now choose arrays in which probes are placed throughout the transcript allowing known and novel splice events to be interrogated in addition to the investigation of total gene expression changes (Okoniewski & Miller, 2008).

GWAS of psychosis have identified several associations and pointed to a number of possible susceptibility genes, however understanding the function of these genes and the mechanisms responsible for disease pathophysiology requires additional experiments. In this thesis, to investigate potential effects of ZNF804A on gene expression and RNA processing, a mouse line in which the ZNF804A orthologue, Zfp804a had been identified as mutated in an ENU library at Harwell (ENU DNA Archive, MRC Mary Lyon Centre, Harwell) was utilised. The mutation generated a premature termination codon (PTC) in exon 2 of Zfp804a (chapter 2.1.2) which is likely to disrupt Zfp804a function either through an aberrant product or activation of the nonsense mediated decay mechanism.

Nonsense mediated decay (NMD) is highly conserved across species and acts as both a surveillance mechanism, distinguishing premature and normal stop codons, as well as a regulator of gene expression (Stalder & Mühlemann, 2008). NMD inhibits translation and leads to the decay of the NMD substrate (Rebbapragada & Lykke-Anderen, 2009). Both transcription and splicing are necessary for NMD, exemplified by the insensitivity

of intronless genes harbouring PTCs to NMD (Stalder & Mühlemann, 2008). The NMD mechanism is only triggered if a premature termination codon (PTC) is more than 50-54bp upstream of an exon-exon boundary (Nagy & Maquat, 1998). NMD reduces the abundance of C-terminally truncated polypeptides which could have deleterious, dominant negative or gain of function effects (Frishmeyer & Dietz, 1999).

Initiation of NMD is dependent upon the position of the PTC. If the PTC is less than 50-54 nucleotides upstream of the last exon-exon junction it is possible that the transcript will escape NMD (Bashyam, 2009). Correct definition of exon-exon boundaries is critical to the NMD mechanism, exemplifying why splicing is critical to NMD, at least in the mammalian system. The regulation of this process is enhanced by exon junction complexes (EJC) which help to define exon-exon junctions. EJCs are deposited during splicing 20-24bp upstream (5') of each exon-exon junction, in a sequence non-specific manner. A termination codon upstream of one or more EJCs triggers NMD (Lykke-Anderson, 2002). Based on the above criteria it was anticipated that the PTC in exon 2 of Zfp804a would trigger the NMD mechanism resulting in the reduced abundance of the Zfp804a mRNA transcript in the C59X mutants relative to the wildtypes.

Work Described.

The objective was to identify the unknown mechanisms by which the *ZNF804A* gene, encoding ZNF804A, influences risk of schizophrenia and bipolar disorder. The presence of a classic Cis(2)His(2) (C2H2) zinc finger domain commonly found in transcription factors, along with evidence of allelic differences in the binding intensity of nuclear binding proteins at rs1344706 (Hill & Bray, 2011; Riley et al., 2010) may predict a role for ZNF804A at the level of transcription regulation and RNA processing. To test these hypotheses, and with the aim of identifying mechanisms downstream of ZNF804A which may influence disease risk, the impact of altered *ZNF804A* function on gene expression in the brains of mice which carry a truncating mutation in the mouse orthologue of *ZNF804A*, *Zfp804a* was determined.

3.2 Methods

3.2.1 Sample

A mouse line carrying an ENU induced premature termination codon (PTC) in exon 2 of Zfp804a was rederived by in vitro fertilization at Harwell (ENU DNA Archive, MRC Mary Lyon Centre, Harwell). The mutation is denoted as C59X as it involves a two base substitution from GT to AA resulting in the replacement of a cysteine residue with a stop codon. Founder mice (F_1) with the C59X mutation were then generated from a Balb/c x C3H/Hej cross (ENU treated male Balb/c mice bred with C3H/HeJ females) at Harwell, these F_1 mice were then bred to form the experimental cohort by another PhD student at Cardiff University (T. AlJanabi). C59X mice were backcrossed onto the C57BL/6J mouse strain to obtain congenicity. This process was accelerated using the speed congenics method, in which mice were screened for strain specific markers. Mice carrying the ENU mutation which also had the highest proportion of C57BL/6J markers were used in the breeding experiments. F3 generation female C59X heterozygotes were intercrossed with F3 male C59X heterozygotes to produce the $F3_i$ intercross generation from which the brain tissue was derived for initial expression studies. The $F3_i$ generation were estimated to have 96.13% C57BL/6J background. This level of C57BL/6J background was not ideal (as discussed in 2.1.2), but provided a useful screen in which any observable expression differences could be validated in mice with >97% purity when these mice became available.

3.2.1.1 The Zfp804a C59X Mice.

All experiments were conducted on brain tissue extracted from mice with a nonsense mutation in exon 2 (C59X) of the Zfp804a gene (Fig. 1) or their wildtype littermates.

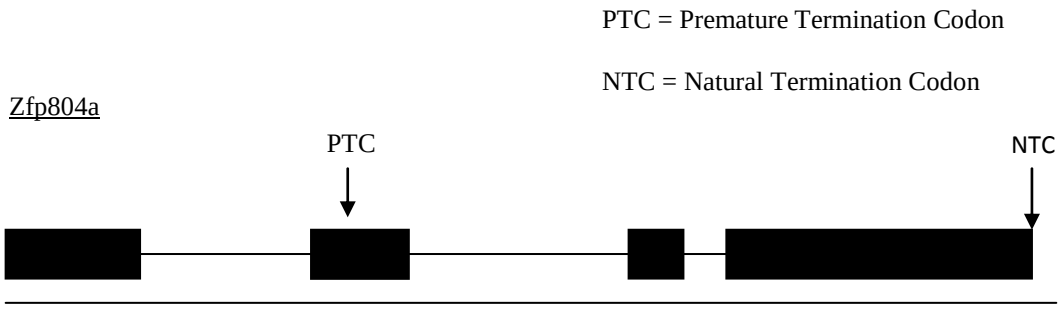


Figure 1. PTC Mutation in Zfp804a. The ENU mutation chosen to be re-derived into a line was a PTC in exon 2 of Zfp804a. The mutation is a dinucleotide polymorphism from GT to AA creating a premature stop codon (TAA) (C59X).

Third generation ($F3_i$) mice with estimated 96.13% pure C57BL/6J background (i.e., the non-ENU background) were available at the time of conducting the initial expression studies. While ideally, higher levels of non-ENU background are desirable (in case additional ENU induced mutations are present) the likelihood that any two mice would have the same undesired ENU mutation is small (probability calculations are discussed in Chapter 2.1.2).

Genotype had no obvious effect on development or health of the mice (T. AlJanabi). Prior to expression analysis male Zfp804a mutant mice were observed on a range of behavioural tests carried out by another PhD student (T. AlJanabi) to characterise behavioural phenotypes that may result from the PTC. Behavioural tests included locomotor activity, PPI open field, elevated plus maze and rotarod and a subset of the mice spent 24hrs in a phenotyper. None of the mice received drug treatments nor did they undergo any food or water restriction programs. Further details are provided in Chapter 2.1.2.

Two waves of expression studies were carried out on adult C59X brain tissue, the first in female mice and the second in males. As all the males of the $F3_i$ cohort were used for the initial battery of behavioural experiments there was a restricted number of C59X mutants available for expression studies as the availability of brain tissue from the mice was dependent upon their completion of behavioural tests. A greater availability of female C59X homozygotes dictated that the first wave of expression analyses was carried out on 3.5 month old female C59X mutant and wildtype mice. A sufficient number of male C59X mutants of the same generation ($F3_i$) became available once all

behavioural tasks had been completed, by which time these mice were 6 months old. Both time points represent sexually mature adult mice.

3.2.2 Female C59X Expression Study.

8 mice were included in the first expression experiment, 4 wild type at the *ZNF804A* locus and 4 homozygotes (for the PTC mutation in exon 2). All were 3.5 months old. Following brain dissection (Chapter 2.1.3) RNA was extracted from the left hemisphere as described (Chapter 2.1.3). The integrity of the RNA was determined using the Agilent 2100 Bioanalyser (Chapter 2.2.3.1).

3.2.3 Male C59X Expression Study.

The second wave of expression experiments consisted of 8 male mice, 5 wild type at the *ZNF804A* locus and 3 homozygotes (for the PTC mutation in exon 2). RNA was extracted from the whole brain as previously described. Male mice were 6 months old. Evaluation in male mice allowed the exclusion of expression changes occurring due to variation in stage of the female oestrous cycle.

Whilst half brains were used in the first wave, whole brains were considered in the second wave to rule out hemispheric effects on expression.

3.2.4 Combined Analyses.

To increase the power of the study all 16 samples were analysed in a combined analysis consisting of 9 wildtype and 7 C59X mutant samples. Differences in gender, age and scan date were accounted for in the study as described later.

3.2.5 Zfp804a mRNA Levels.

The nonsense mutation within exon 2 of Zfp804a was predicted to initiate the nonsense mediated decay surveillance mechanism. Activation of such a mechanism would be expected to result in the reduced abundance of the mRNA transcript. To determine if such a mechanism was occurring in the C59X mice qualitative analysis of the Zfp804a mRNA transcript was performed using RT-PCR.

3.2.6 Sequencing the Mutation.

Amplified cDNA was sequenced from the mice to ensure the mutants were expressing the mutation in the mature message. Sequencing was carried out using the primers specified in appendix 2.1. Sequencing is described in full in Chapter 2.6.

3.2.7 Sample preparation and quality.

Brains were removed and immediately snap frozen in Liquid Nitrogen. Brains were taken from storage at -80°C and RNA was isolated by myself from either the left hemisphere or whole brain (minus olfactory bulbs) and stored at -80°C. RNA was isolated and processed as described in Chapter 2.2 & 2.3.

3.2.8 The Affymetrix Genechip Mouse Exon 1.0 ST Array

The initial acquisition of brain tissue, RNA preparation and all statistical analyses were carried out by me. The labelling, hybridisation and scanning steps were carried out at Cardiff University by M. Musson within the Central Biotechnology Services (CBS) facility as described in section 2.8.

Labelling, hybridisation and scanning procedures are described in full in Chapter 2, Section 2.8.2.

3.2.9 Statistical Analysis

3.2.9.1 Partek Genomics Suite (Version 6.5)

Partek Genomics Suite (version 6.5 and beta 6.6, St. Louis, MO) is a purpose built software suite designed to enable analysis of a number of high-throughput technologies.

3.2.9.2 Data upload

CEL files were uploaded into Partek GS and sample files were produced using Excel (2007) allowing the identification of experimental groups to be recognised by the software (See 2.8.3.2).

3.2.9.3 Probe Filtering

Only the core meta-probeset was included in the analysis, unless stated otherwise.

3.2.9.4 Preprocessing

Once raw intensity values (CEL files) were generated several steps were taken in order to produce meaningful expression data. These steps collectively are termed preprocessing (Chapter 2, Section 2.8.3.3). Samples were preprocessed using robust multichip averaging (RMA) (Irrizarry et al. 2003). The algorithm consists of three discrete steps background correction, normalisation using quantile normalisation and summarisation using the median polish technique. Intensity values were Log transformed (base 2). Full details are given in Chapter 2, section 2.8.3.3.

3.2.9.4.1 Background Correction

Background correction removed non-biological variation. This was carried out on each chip independently (Chapter 2, section 2.8.3.3.2).

3.2.9.4.2 Quantile Normalisation

Following background correction the data were normalised to remove array bias or variation which is not a result of true biological differences between samples. The technique used for normalisation was quantile normalisation which was carried out across all chips on background corrected data (Chapter 2.8.3.3.3).

3.2.9.4.3 Probe Summarisation

Summarisation of probe level data to a combined probeset intensity value was achieved using the median polish algorithm. This was the final of the three RMA steps and was performed on background corrected, normalised and log transformed intensity values (Chapter 2.8.3.3.4).

3.2.9.5 Annotating the Dataset

Data were analysed using the Affymetrix annotation files, NetAffx, version na31. mm9. Using only the core meta-probeset meant targeted sequences had been sequenced, cloned and curated manually and therefore had high annotation confidence (Affymetrix). Only probes annotated by Affymetrix as unique were included in my analysis to account for the problem of probe cross-hybridisation.

3.2.9.6 Quality Control (QC)

Following preprocessing, I reviewed a number of recommended quality control metrics in order to check the quality of the data. QC is particularly important with exon array data which is prone to greater numbers of false positives as excess noise can be misinterpreted as differential splicing (Gardina & Turpaz, 2008). I generated standard QC measures, described in chapter 2.8.3.5 using Expression Console (Affymetrix, Version 1.1.2) and Partek GS (v6.6). QC was run at the Gene level on all CEL files using RMA sketch on core probes and using a log2 scale.

There are no standardized cut offs. For this reason, as recommended by Affymetrix the distribution of several metrics was assessed to ensure the microarray experiment has passed a minimum level of quality control. Within each metric, any sample with values two standard deviations away from the mean was flagged. Samples outside of this range across 3 or more metrics were either excluded or monitored in all downstream analyses, dependent on the severity of the case (Gardina & Turpaz, 2008). Removal of an outlier in only two or three metrics may be more detrimental than beneficial depending on sample size, as power may be considerably reduced. As there was no standard procedure outliers were removed in extreme cases (outlying in more than 3 metrics). Samples with values outlying in 3 or less metrics were instead monitored after statistical

analysis to ensure they were not behaving differently to other samples within the experimental group. Samples were considered as one single group and also within their experimental group (wildtype or C59X mutant). Samples were assessed within their experimental groups as they would be expected to behave more similarly and so outliers may be more apparent.

3.2.9.7 Filtering

Probesets were filtered prior to statistical analysis based on annotation confidence and intensity, to optimise the likelihood of identifying true expression differences. Probesets were initially excluded if they had a maximum log2 intensity value <3, but this was increased for more stringent analysis.

3.2.9.8. Statistical Algorithm.

Details of the algorithms used to calculate differential expression and splicing p values are given in Chapter 2, section 8.3.6.1 & 3. Briefly I determined differential expression and splicing using the custom alternative splice ANOVA in Partek Genomics Suite. A one-way ANOVA was performed for individual female and male experiments with a two-way ANOVA used in the combined male/female analyses in order to covary for gender differences. In all instances genotype was added to the model and was specified as the alternative splice factor, meaning differential splicing was determined based on differential probeset expression across the 2 levels of genotype (wildtype and C59X mutant). A total of 15,808 and 15,813 transcript clusters were included in the female and male analyses respectively which equates to ~194,000 probesets. Gender was added to the model for all combined analyses to account for gender differences. This was also perfectly correlated with age, scan date, and hemisphere versus whole brain differences between samples. Thus all variance attributable to these factors are embraced by a single factor. For combined analyses 15,833 probesets were tested following probeset filtering. Both differential expression and differential splicing p values were generated from this model. Genes showing significant differential expression and splicing were then filtered and prioritised based on statistical significance. To determine fold change at the transcript level a linear contrast was included in the model between C59X

mutants and wildtypes using the Fisher's Least Significant Difference (LSD) method (Tamhane and Dunlop, 2000).

3.2.9.9 Multiple Test Correction (MTC)

Data were corrected for multiple testing using both the step up False Discovery Rate (FDR) (Benjamini & Hochberg, 1995) and the more stringent Bonferroni correction (Holm, 1979). An FDR threshold of 0.05 was set, meaning 5% of the results were expected to be false positives. The Bonferroni correction is the most conservative MTC method as p values are corrected by the number of tests carried out.

3.2.10 Visualisation of Alternative Splice Events.

Visual inspection was carried out for differentially expressed and spliced genes using the geneview produced as part of the alternative splice output. The geneview plots the expression at each probeset across the transcript for each experimental group and allows the position of the probeset to be visualised relative to the exons of the transcript using a RefSeq track from the UCSC genome browser (Kent et al., 2002b) (http://genome.ucsc.edu/).

3.2.11 Degree of Overlap.

To determine how robust the differential expression and splice results were, the degree of overlap or replication between female and male experiments was assessed. Simply considering whether compared with a null distribution, there is an excess of transcripts significant in the female experiment that are also significant in the male experiment does not account for the possibility of non-null distributions in the two datasets (i.e. if in a replication dataset, 50% of all genes show nominally significant effects, then by chance, not 5% but 50% of genes significant in a discovery sample should show effects that replicate at the P=0.05 level). To account for this a 2x2 contingency table (Table 1) was constructed and statistical analysis performed using the chi-square test (χ^2) to determine whether the proportion of genes that replicate differ contingent on whether they were or were not significant in the other dataset. The Pearson's chi-square test was

performed using SPSS (v16), except for instances in which the expected cell frequency was less than 5, in which case the Fisher's exact test was used.

	Male Experiment - Significant Genes	**Male Experiment – Non-Significant Genes**
Female Experiment - Significant Genes	n = number of genes significant in both female and male analyses	n = number of genes significant in the female experiment that are not significant in the male experiment
Female Experiment - Non-Significant Genes	n = number of genes not significant in Female experiment that are significant in the male experiment	n = number of genes that are not significant in either analysis

Table 1. 2x2 Contingency table used for Chi-square test.

3.2.12 Alternative Algorithms for Detecting Splicing.

Multiple statistical algorithms are available for the analysis of differential splicing from exon array data. The underlying structure of each model has the same fundamental assumptions in that probeset expression is predicted by the model and compared to the null hypothesis that the expression of the probeset is proportional to other probesets within a gene across all samples considered (Affymetrix, 2005b). Each algorithm determines how much the probeset expression diverges from the model and a p value is generated based on this. Affymetrix recommended considering several different tests on the data for robust identification of alternative splice events. Within this thesis I used the ANOVA (Chapter 2.8.3.6.3) FIRMA (Chapter 2.8.4.4) and MiDAS (Chapter 2.8.4.3) statistical algorithms each described in full in Chapter 2.

3.3 Results.

3.3.1. Abundance of Zfp804a Transcript.

To establish if the NMD mechanism was operating in the C59X homozygote mice, levels of Zfp804a mRNA were qualitatively assessed using RT-PCR. Zfp804a mRNA was expressed in homozygote C59X mice as well as in WT mice (Figure 3.3.1). This result provides evidence to strongly argue against the NMD mechanism and could indicate that a compensation mechanism is being utilised (Chapter 4, section 4.3.3).

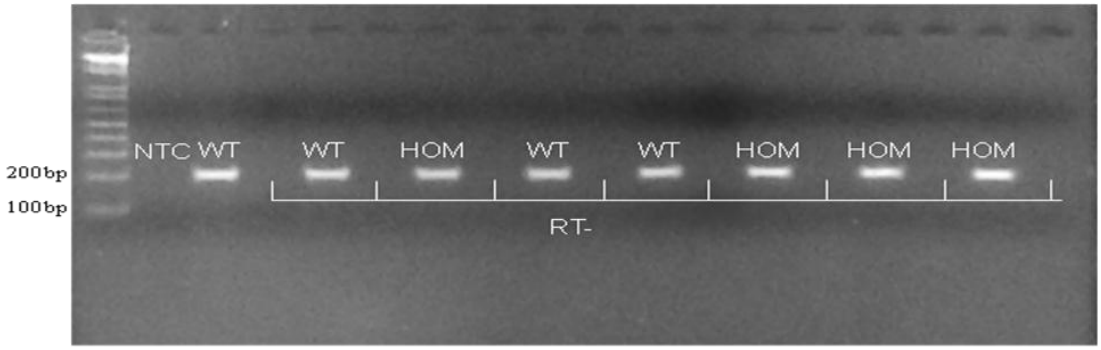


Figure 3.3.1 Abundance of Zfp804a in WT and ENU Mutants. The abundance of the Zfp804a transcript in $F3_i$ mice was determined using an RT-PCR assay. PCR products were run out on a gel. An amplimer of 209bp corresponds to Zfp804a. **NTC** No template control **WT** Wildtype **Hom** Homozygote C59X mutant **RT-** Reverse Transcriptase negative control.

3.3.2 Sequencing the C59X mutation.

To ensure the mutation was present in the mRNA transcript of the mutant mice, sequencing of the mRNA transcript was carried out by myself. Sequence traces from both homozygote mutants and WT controls can be seen in Figure 3.3.2. I confirmed genotype from the sequencing traces to ensure mRNA matched previous genotype assignment using gDNA from tail tips (carried out by T. AlJanabi).

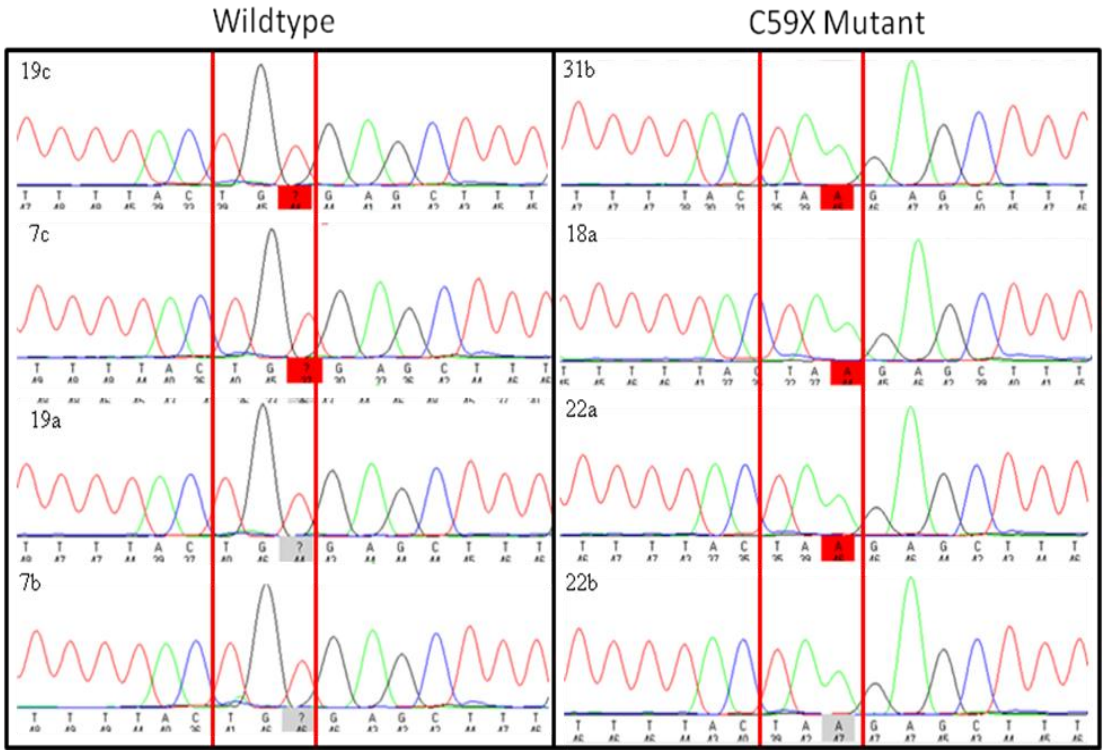


Figure 3.3.2. Zfp804a Exon 1-3 Sequence Results. Sequencing from exon 1-3 of Zfp804a viewed in NovoSNP (version 3.0.1) enabled the site of the ENU mutation to be observed in each of the four female mice used in the expression study (right hand side, red box) rather than the wildtype cysteine residue (left hand side, red box). 19c, 7c, 19a and 7b = 4 wildtype mice. 31b, 18a, 22a and 22b = 4 mutant mice.

3.3.3 RNA Quality.

Prior to establishing mRNA expression or splice differences which may be present between mutant and wt, RNA samples were analysed for quality. One of the most important determinants of the quality of an expression study is the quality of the RNA. Standard thresholds for quality RNA include a 28s/18s ratio above 1.0, ideally close to 2 (Ch 2.2.3.2) and an RIN of 7 or above, ideally 8 (Ch 2.2.3.3) (Shroeder et al., 2006). An example of one of the Agilent Bioanalyser (Ch 2.2.3.1) traces is presented in figure 3.3.3. All 16 samples (male & female) had a RIN above 8 and 28s/18s ratios at 1 or above.

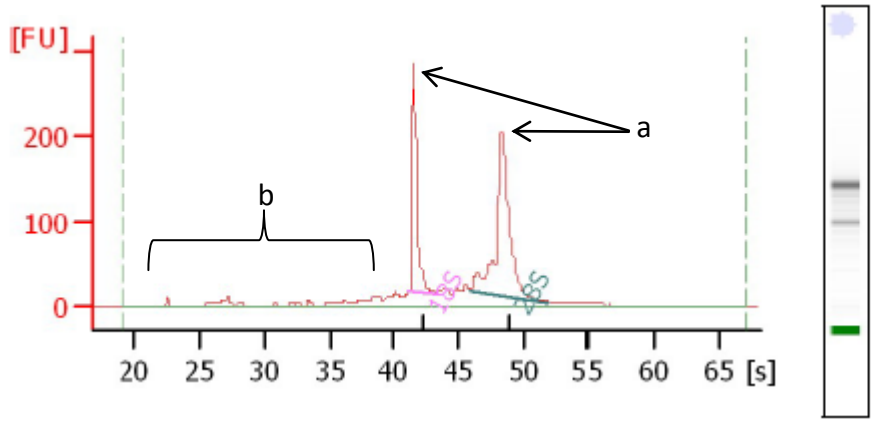


Figure 3.3.3. Determination of RNA quality. Each sample was run on an Agilent 2100 Bioanalyser to determine RNA quality using the RIN and rRNA ratio. The above example displays the results from sample 22b, a female mutant. The bioanalyser uses electrophoretic separation of RNA and both an electropherogram (left) and a gel image (right) is produced for each sample. Fluorescence (Fu) is plotted on the Y axis and time in seconds (s) on the X axis. The above example had a RIN of 9.1 and a 28s/18s ratio of 1.7. This high quality RNA is observable from the graph by the two clearly defined 18s and 28s peaks (2 clear bands on the gel) (a) as well as low levels of smaller RNA molecules which are the products of degraded 18s and 28s rRNA and tRNAs (b).

3.3.4 Affymetrix Genechip Exon 1.0 ST Array.

To determine the expression profiles of C59X mutants relative to wildtypes I used the Affymetrix exon array. Exon arrays not only facilitate more accurate determination of gene expression levels, but also allow processes such as alternative splicing to be interrogated. As the study was performed in two stages the results are presented accordingly. The initial analysis performed in female mice and the second analysis in male mice. The degree of overlap in the results across the two studies was then identified.

3.3.4.1 Quality Control.

The experiment can be assessed for quality using a number of variables. Each experiment was assessed first using a hybridisation efficiency metric, then by qualitative assessment using principle component analysis (PCA), signal distribution and relative log expression signal. Following this 6 quality metrics that I generated using expression console (Affymetrix Version 1.1.2) and Partek GS were used to determine if any sample met outlier criteria as outlined by Affymetrix as lying more

than 2 standard deviations from the mean. This was done both as a whole group and within biological replicates (experimental groups).

3.3.4.1.1 Hybridisation Efficiency.

The efficiency of sample hybridisation to the chip was assessed using 4 *E.coli* internal controls. Each is hybridised to the chip at a known concentration and based on this predefined concentration the signal intensities of each should follow an order from lowest expression in BioB hybridised at the lowest concentration up to Cre, hybridised at the highest concentration. From the graph (Fig 3.3.4) it is clear that each of the 4 controls has a Log 2 expression in each of the 8 female samples in the expected rank order and from this it can be inferred that the hybridisation of each of the samples to the exon chip was efficient and should have no detrimental effects on expression.

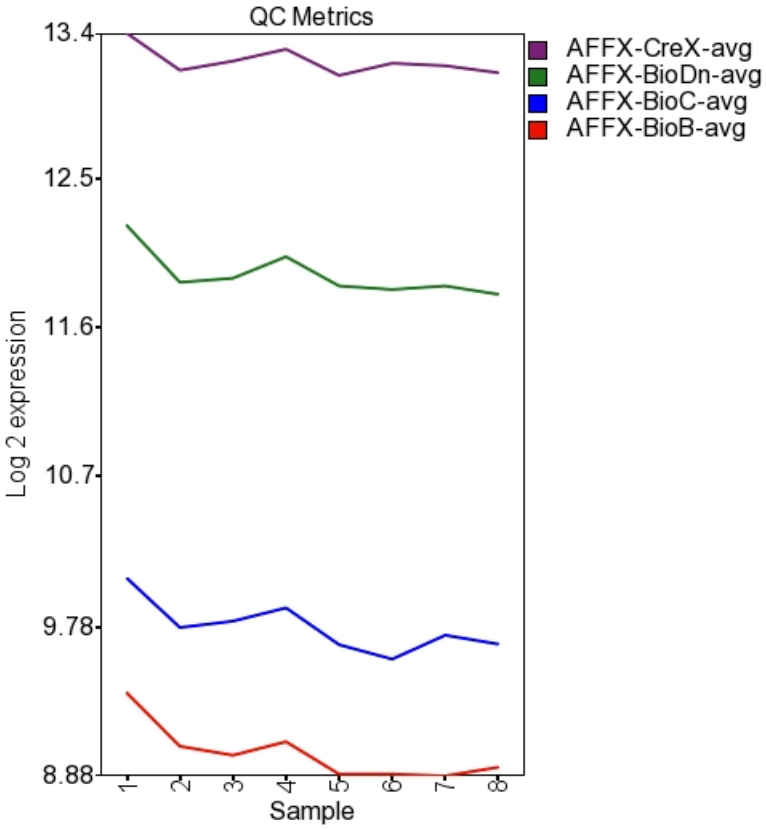


Figure 3.3.4 Quality Control. Hybridisation Efficiency of Affymetrix GeneChip Mouse Exon 1.0 ST Arrays. Hybridisation Efficiency is determined by the ranked signal intensities of four *E.Coli* controls (BioB, BioC, BioD and Cre). The expression signal of the hybridisation for each of the 4 controls (y-axis) is plotted for each of the 8 female samples (x-axis). Efficient hybridisation is denoted by a rank order of BioB<BioC<BioD<Cre and is observed in all 8 female samples.

3.3.4.1.2 Examining the Global Expression Pattern with Principle Components Analysis (PCA).

To determine the global expression profile in C59X mutants I performed PCA on the normalised intensity of all core probesets. Visual inspection of the PCA plot (Fig 3.3.5) showed no obvious clustering of samples suggesting the gene expression profiles of each samples was different. The spread of the samples suggests there is a considerable degree of variation in the expression profiles of the female mice and this variation occurs across all samples not between Wt and C59X mutants. Whilst this makes observing outliers difficult there were no egregious outliers observed by looking at the plot, although more thorough quantitative QC analyses would be needed to determine if

this is correct. There is a very slight degree of separation of mutants and wildtypes along the second principle component although there is no clustering of biological replicates.

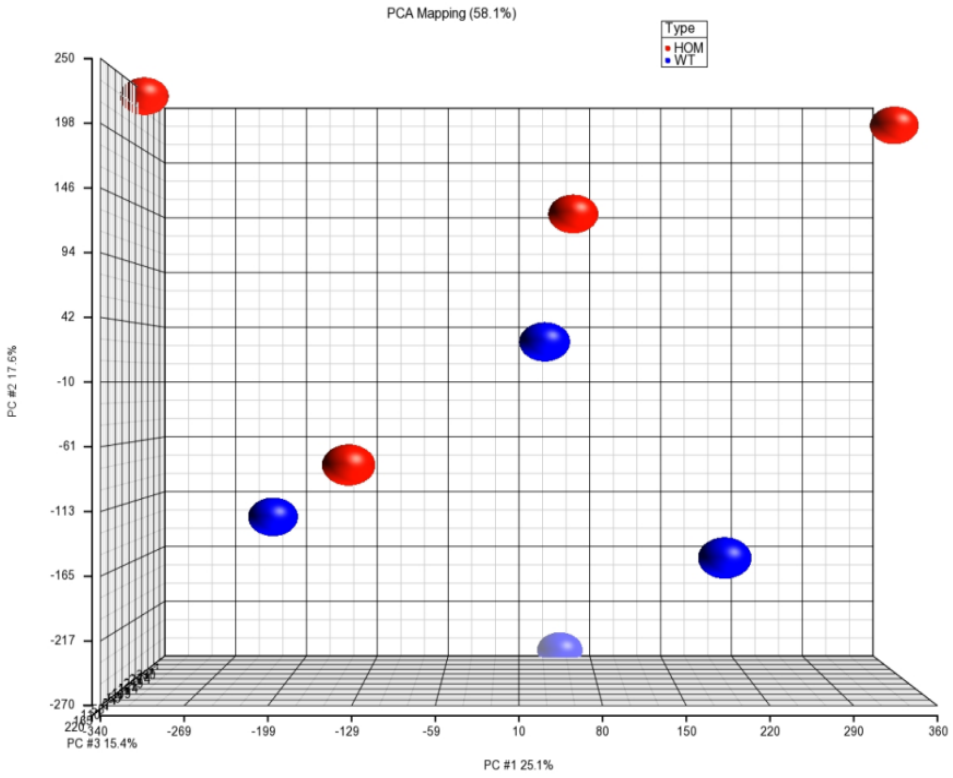


Figure 3.3.5. Principal Components Analysis (PCA) of Female C59X mutant and wildtype samples. The first and second principal components are displayed in the plot. Samples do not appear to cluster by genotype and a considerable amount of variation between biological replicates is observed. **Hom** Mutant C59X (red); **WT** wildtype (blue).

Each sample will have an expression profile which can be plotted as the range in signal intensities and the frequency of each of these signal intensities. From this plot the distribution of the signal intensity for each sample can be observed (Fig 3.3.6). The plot shows that all samples followed a normal distribution. There was little variation observed in this distribution for all 8 samples therefore data appeared to be of good quality. A number of other QC metrics were assessed using data I generated from expression console (Affymetrix) and Partek GS.

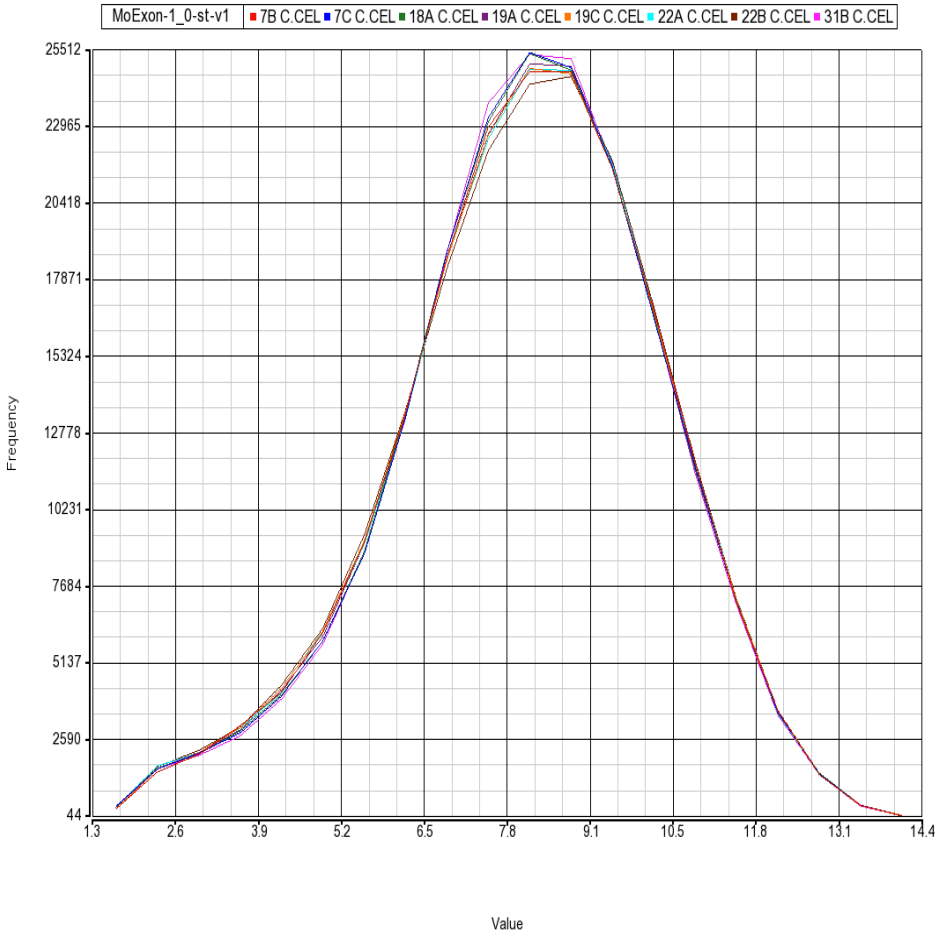


Figure 3.3.6. Histogram of Signal Intensity. The range of signal intensities were plotted on the x-axis with the frequency of each of these intensities for each of the 8 samples plotted on the y axis to create the distribution of signal intensities. Each of the 8 samples follows a normal pattern of distribution with the distribution of all 8 samples tightly clustered.

Box plots were also generated to assess the degree of variation in signal distribution in the 8 samples prior to and after preprocessing steps (Fig. 3.3.7). The mean, interquartile range and spread of the data were comparable across all samples even prior to normalisation and summarisation (Ch2.8.3.3.3 & 2.8.3.3.4). Any observable differences were corrected for by the preprocessing procedures. The relative log expression signal boxplot was also plotted which was used to determine, overall how differently one sample behaves relative to all other samples. This metric was also quantitatively assessed in tables 1 and 2.

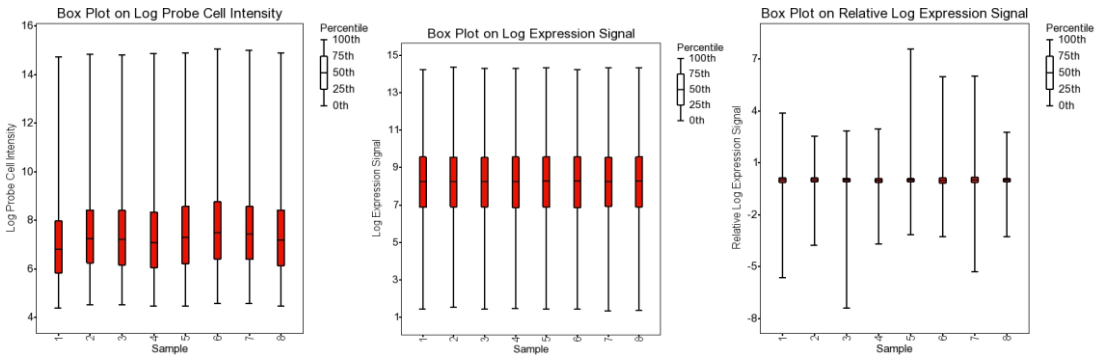


Figure 3.3.7. Quality Control. Boxplots of Expression Signals. The left box plot shows the log probe cell intensity for each sample prior to any normalisation or summarisation, with the central box plot showing the log expression signals of probsets following normalisation and summarisation. The small amount of variation prior to preprossecing is corrected by the normalisation and summarisation procedures. The right boxplot is generated by taking the expression signal at a particular probeset on a particular chip and comparing it to the median signal across all the chips, this is done for every probeset on the chip and so gives an indication if one sample is behaving very differently to the other samples, this was not apparent for the 8 female samples.

The quality of the data generated for each sample were also quantitatively assessed by generating values for 6 quality metrics in both Partek GS and Expression Console (Affymetrix). A detailed description of each of these metrics can be found in (Ch2.8.3.5.4 - 2.8.3.5.7) but briefly the overall brightness of the chip was assessed (PM mean), comparison of residuals to the median (Mad residual mean using all probesets and just the positive controls), the overall performance of each chip relative to the other chips (RLE mean, using all probesets and just the positive controls) and finally how well the probesets signals separated positive and negative signals (pos vs. neg auc). The samples were considered both within experimental group (C59X mutant or wildtype) (Table 1) and as a whole group (Table 2). This is because it may be expected that mutants and wildtype would behave differently from each other, but within the wildtype or mutant group the samples would be expected to behave similarly. Calculating the values which were 2 standard deviations either side of the mean (Bold values in each table) allowed thresholds to be established to determine if any sample was defined as an outlier (2> standard deviations from the mean). The results presented, were metrics I generated using Partek GS. Whilst differences in the absolute numbers were found between Partek and Expression Console the same outcome was observed in both QC analyses. No outliers were identified in any of the metrics generated when using either

software, even when considering the samples within their experimental groups. All 8 samples were considered good quality and were taken forward for statistical analysis.

C59 Mutants

Sample ID	PM Mean	All probeset mad residual mean	Pos control mad residual mean	All probeset rle mean	Pos control rle mean	Pos vs neg auc
18A C.CEL	375.32	0.15	0.1	0.16	0.13	0.87
22A C.CEL	422.82	0.13	0.09	0.15	0.11	0.88
22B C.CEL	512.25	0.15	0.11	0.21	0.18	0.87
31B C.CEL	409.68	0.16	0.11	0.21	0.18	0.87
Mean -2*SD	**313.29**	**0.12**	**0.08**	**0.12**	**0.08**	**0.86**
Mean +2*SD	**546.75**	**0.17**	**0.12**	**0.25**	**0.22**	**0.88**
Within Threshold	**YES**	**YES**	**YES**	**YES**	**YES**	**YES**

Wildtypes

Sample ID	PM Mean	All probeset mad residual mean	Pos control mad residual mean	All probeset rle mean	Pos control rle mean	Pos vs neg auc
7B C.CEL	297.56	0.17	0.13	0.19	0.15	0.87
7C C.CEL	372.48	0.16	0.11	0.18	0.14	0.87
19C C.CEL	379.77	0.16	0.11	0.18	0.14	0.87
19A.CEL	388.97	0.14	0.09	0.14	0.1	0.87
Mean -2*SD	**275.75**	**0.13**	**0.08**	**0.13**	**0.08**	**0.87**
Mean +2*SD	**443.64**	**0.18**	**0.14**	**0.21**	**0.18**	**0.88**
Within Threshold	**YES**	**YES**	**YES**	**YES**	**YES**	**YES**

Table 1 Quality Control Metrics. Six quality control metrics were generated following sample pre-processing. Each metric is described in full in chapter 2 and can be used to determine the quality of the data for each sample. The mean and standard deviation are given for both experimental groups (Wt and C59X Mutants). Affymetrix recommend that any sample more than 2 standard deviations from the mean be flagged as an outlier and potentially excluded from further analyses. The thresholds for values more than 2 standard deviations from the mean, in either direction are reported in bold. The final row states whether or not each of the biological replicates had a value within this threshold, with YES denoting that all samples are fine and NO denoting that an outlier(s) has been flagged up. **N.B:** Figures are displayed to 2 decimal places. To determine if the value was within threshold

the figures were considered to more decimal places therefore each instance where the value displayed for a sample matches a threshold, the value is within threshold when considering to more decimal places.

Sample ID	pm_mean	All probeset mad residual mean	Positive control mad residual mean	All probeset rle mean	Positive control rle mean	Pos vs neg auc
7B C.CEL	297.56	0.17	0.13	0.19	0.15	0.87
7C C.CEL	372.48	0.16	0.11	0.18	0.14	0.87
18A C.CEL	375.32	0.15	0.10	0.16	0.13	0.87
19C C.CEL	379.77	0.16	0.11	0.18	0.14	0.87
22A C.CEL	422.82	0.13	0.09	0.15	0.11	0.88
22B C.CEL	512.25	0.15	0.11	0.21	0.18	0.87
31B C.CEL	409.68	0.16	0.11	0.21	0.18	0.87
19A.CEL	388.97	0.14	0.09	0.14	0.10	0.87
Mean -2xSD	**274.39**	**0.12**	**0.08**	**0.12**	**0.08**	**0.86**
Mean +2xSD	**515.32**	**0.18**	**0.13**	**0.23**	**0.20**	**0.88**
Within Threshold	**Yes**	**Yes**	**Yes**	**Yes**	**Yes**	**Yes**

Table 2. Quality Control Assessment of all 8 female Samples Combined. When considering the 8 female samples as one group there is no indication that any sample is an outlier and behaving differently to any of the other samples.

3.3.4.2 Identification of Differentially Expressed Genes between Female Wildtype and C59X Mutant Mice.

Following QC, gene expression analysis was conducted using an alternative splicing ANOVA (Partek GS) to identify genes differentially expressed between C59X mutant and wildtypes. Due to the increased numbers of probes and their placement throughout the transcript the exon array should facilitate more accurate measurement of gene expression. As described in the methods, gene expression p values can be generated using the differential expression and the alternative splice algorithms (Ch2.8.3.6.1&3). Whilst both the workflows were used to determine gene expression differences for clarity only the alternative splice algorithm data will be presented in the next section. I set the algorithm so that genotype was the splicing factor as described in Ch2.8.3.6.3 to determine changes in gene expression between wildtype and C59X mutants. Only probesets annotated by Affymetrix as unique (and not known to cross-hybridise) were included. Cross-hybridising probesets have been previously described to greatly increase the numbers of false positives (Xing et al., 2008). Any probeset with a mean $\log_2$ intensity <3 was excluded leaving a total of 15,808 transcripts (equating to 194,293 probesets). The output of the experiment was a list of genes with both differential expression and alternative splice p values. I then determined the number of genes significantly differentially expressed and spliced using a number of p value stringencies and following multiple test corrections using the false discovery rate (FDR) and the Bonferroni correction (Ch2.8.3.6.4).

3.3.4.3 Identification of Differentially Spliced Genes between Female C59X Mutant and wildtype Mice.

The identification of both known and novel splice events is facilitated with the exon array due to probeset placement within the exons and not across exon boundaries. It is important to note that when using chips such as the exon array the idea is to determine differential splicing rather than identify known or novel alternative splice events *per se*. Either type of alternative splice event can be detected by the platform but only if the expression of each isoform differs between the groups in the experiment, thus it is a relative measure. It is only when the expression of an isoform/exon in one group diverges from that seen in the other group that differential splicing is called and significant p values are observed (Robinson and Speed, 2009).

The numbers of significant differentially expressed and spliced genes at different significance thresholds are displayed in table 3. The hypothesis that ZNF804A may function as a transcription factor was addressed by looking to see if there were any genes significantly differently expressed between the WT controls and the Zfp804a mutant mice. 6% of genes were significantly differentially expressed ($p \leq 0.05$). Given that 15808 transcripts were included in the analysis the family-wise error rate needed to be controlled to limit the number of false positives. The false discovery rate (Benjamini & Hochberg, 1995) was set at 0.05 at which threshold it is expected that 5% of the significant changes are false positives. The more stringent Bonferroni correction was used to set a threshold adjusted for number of tests such that this P value is expected to be attained by chance by any transcript only once in 20 complete experiments (Holm, 1979). Using an FDR threshold of 0.05 and a Bonferroni correction for 15,808 tests, in no genes were the expression levels significantly different between the groups. The results provide no evidence for a direct link between the mutation in Zfp804a and regulation of gene expression.

As the evidence for a role in the regulation of transcription was not apparent, differential splicing was assessed between C59X mutants and Wildtypes to determine if the mutation in Zfp804a was affecting RNA processing. The alternative splicing algorithm predicts splicing based on expression at a particular probeset (or exon) relative to the pattern of expression observed throughout the whole transcript. Significant differential splicing is called when probeset expression diverges from the transcript expression pattern (Ch2.8.3.6.3). Results showed that a similar proportion (~6%) of the total number of transcripts were differentially spliced between wildtype and C59X mutants as those predicted to be differentially expressed at the nominal p value $p \leq 0.05$ (Table 4). The splicing results were however more robust than in the differential expression as evidenced by the number of transcripts remaining significant at more stringent p values and following multiple test correction. A total of 79 and 21 genes remained statistically significant following multiple test correction with an FDR threshold of 0.05 and the Bonferroni correction for 15808 tests, respectively.

P Value Threshold	No. of Significant Genes	% of Genes
Unadjusted p≤0.05	968	6.12
p≤0.01	226	1.43
p≤0.001	23	0.15
p≤0.0001	2	0.01
Following Multiple Test Correction		
With FDR 0.05	0	0
With Bonferroni Correction (15808 tests)	0	0

Table 3. The number of genes differentially Expressed between Female Wildtype and Mutant C59X mice. Whilst a considerable number of genes are found to be differentially expressed at a nominal p value following adjustments for multiple testing the number of significant genes reduces to 0, suggesting no individual genes show strong expression differences between the wildtypes and the mutants.

From the initial exploration of the data, there is more evidence for the hypothesis that aberrant Zfp804a effects RNA processing than that proposing an effect on differential expression.

Following the acquisition of male C59X brain tissue the experiment was repeated, using the same methodological and analytical procedures (Ch2). Replication in male mice was carried out to ascertain if the results observed in the female mice could be recapitulated and to remove any possibility that the results were attributable to variation in the stage of the oestrous cycle among the female mice.

Alternative Splicing	No. of Significant Genes	% of Genes
Unadjusted p<0.05	1010	6.39
p<0.01	400	2.53
p<0.001	136	0.86
p<0.0001	54	0.34
Following Multiple Test Correction		
With FDR 0.05	79	0.50
With Bonferroni Correction (15808 tests)	21	0.13

Table 4. The number of differentially spliced genes between wildtype and C59X mutant female mice. Whilst a similar proportion of genes are shown to be differentially spliced to those differentially expressed at a nominal p value of $p<0.05$, when considering the more stringent p values and in particular the number of genes significant following multiple test correction there is a greater proportion of spliced genes.

3.3.4.4. Replication study in Male C59X mice.

The experiment was repeated when there were sufficient numbers of male homozygote C59X mice available for expression analysis. The same QC measures were generated and the same criteria used for outlier analysis. The hybridisation efficiency of all 9 male (4 C59X mutants, 5 wildtypes) samples was sufficient, as denoted by the correct rank order in expression values of 4 internal controls (Fig 3.3.8). Whilst observing the expression patterns of these 4 *E. Coli* controls it was noted that sample 6, a C59X mutant, had a divergent $\log_2$ expression pattern of the four controls relative to the other 8 samples. Following the generation of a PCA plot (Fig 3.3.9) the overall pattern of gene expression for each of the samples was assessed. There was no indication of samples clustering by genotype. A single cluster is observed in the top right of the plot where the majority of samples lie. Two samples were not found within this cluster one of which is separated from the other samples along the first principle component (Indicated with arrow A). This sample corresponds to sample that appeared divergent in the hybridisation efficiency assessment. The other sample, a wildtype (17c) is separated from the other samples along principle component 2 (Indicated with arrow B). A normal distribution of signal intensity was observed for all male mice (Fig. 3.3.10) however both 17b and 17c, highlighted in the PCA, were at the extremities of the distribution toward the mean frequency range of the plot. I then generated boxplots of the log

expression signal to establish if the degree of variation in 17b and 17c was consistent in these plots and if so whether the 2 samples should be excluded (Fig. 3.3.11).

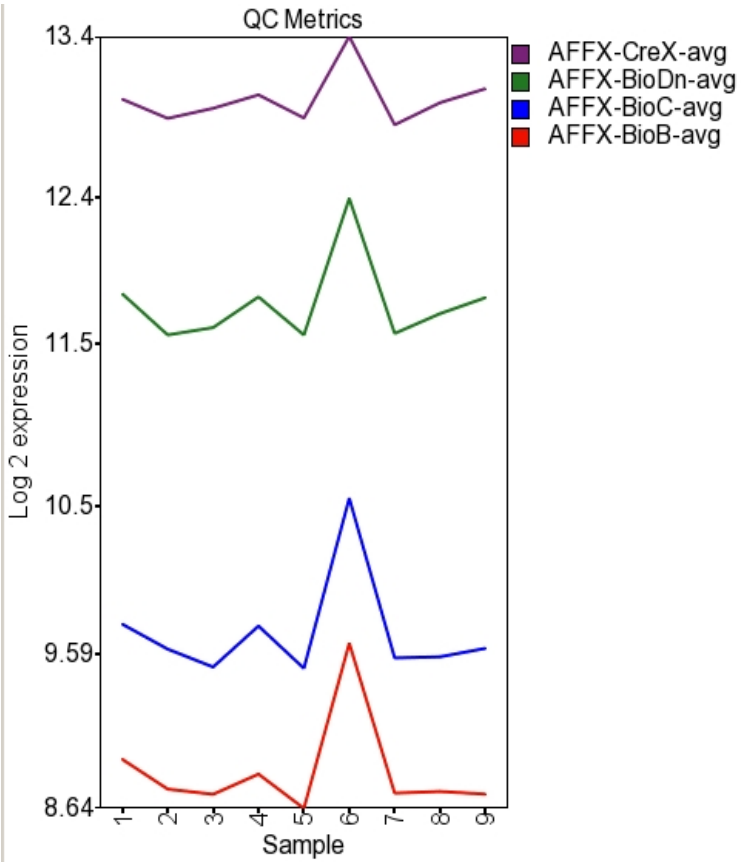


Figure 3.3.8. Quality Control. Hybridisation Efficiency of Affymetrix GeneChip Mouse Exon 1.0 ST Arrays. Efficient hybridisation is denoted by a rank order of BioB<BioC<BioD<Cre, four *E.Coli* internal controls that are hybridised to the chip at known, staggered concentrations. This rank order is observed in all of the Male wildtype and C59X samples (n = 9). In assessing the quality of the hybridisation process it was also noted that the log 2 expression in sample 6 (a C59X mutant) was quite different to the other samples.

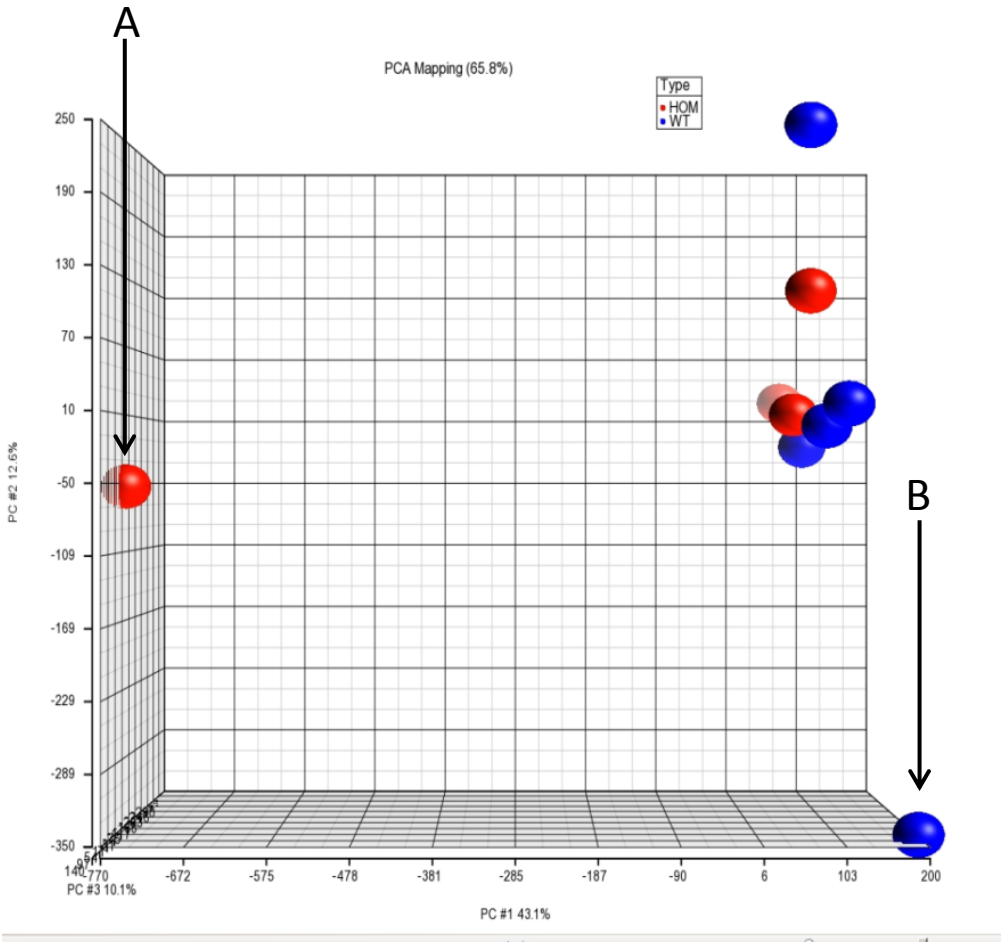


Figure 3.3.9. Quality Control. Principal Component Analysis of Male samples. The samples do not appear to cluster by genotype along the first and second principle components. Instead a cluster of samples is observed in the top right of the plot with two samples (labelled with arrows) separated from this cluster suggesting a different overall pattern of gene expression in these two samples.

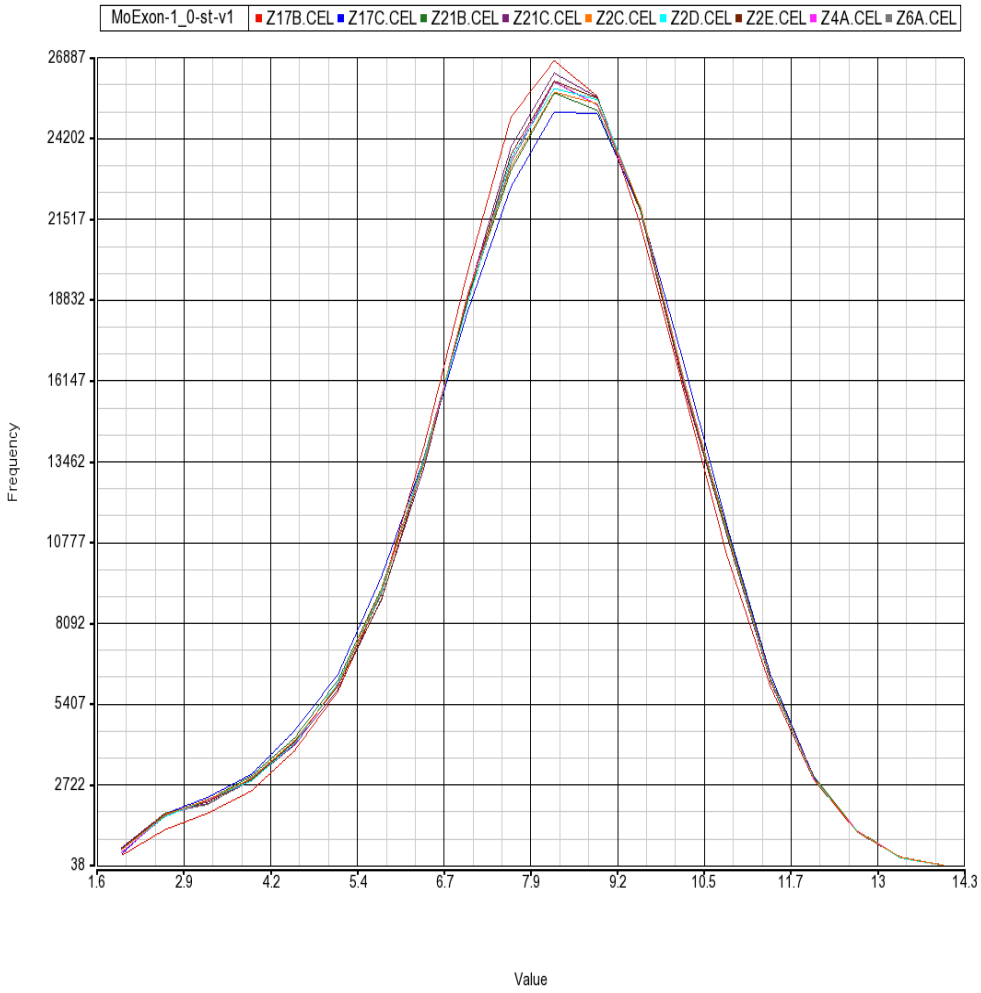


Figure 3.3.10. Quality Control. Histogram of Signal Intensity. The distribution of signal intensities. Intensity is plotted along the x-axis and frequency of intensity plotted on the y-axis. Distribution is normal across all male wildtype and C59X mice. At the peak of the curve the differences in the distribution of the samples becomes slightly evident with both the mutant, 17b and the wildtype, 17c furthest from the cluster of the other samples.

Boxplots generated to examine the log expression signals of each male sample are shown in Figure 3.3.11. Whilst the normalisation and summarisation of the data appeared to correct the variance in expression, the relative log expression signal boxplot (right plot) clearly shows that sample 17b (6), the same C59X mutant as observed previously, is behaving differently relative to the other 8 samples. The inter-quartile range (IQR) of this sample is much larger than that observed in the other samples and this characteristic is often present in samples of poor quality. Sample 17c (7), also separate from the PCA cluster, had a slightly larger IQR but was not as distinguishable as sample 17b. In the four qualitative and subjective measures analysed so far, the male

C59X mutant (17b) appeared to be a consistent outlier, whilst the male wildtype (17c), demonstrated outlier characteristics in several of the metrics but not all. To more quantitatively establish if either of these samples should be excluded I generated the same 6 metrics as described previously (Ch2.8.3.5.4-7) to determine as accurately as possible if 17b and 17c were outliers, defined as metric values consistently greater than 2 standard deviations away from the mean. First considering all male samples together (Table 5), the mutant 17b (sample 6) that appeared to be an outlier in the qualitative metrics, had values greater than 2 standard deviations from the mean in 5 of the 6 metrics generated. Again the results presented are those I generated in Partek GS, but analysing the QC results in Expression Console also resulted in sample 17B being an outlier in 5 of the 6 metrics. In both instances the metric in which sample 17b was within 2 standard deviations of the mean was the positive versus negative auc metric. 17b has a value (0.88) above which Affymetrix guidelines suggest may be an outlier, Affymetrix stipulate that values above 0.8 do not guarantee good quality data (Chapter 2.8.3.5.7) Samples were then considered by experimental group, as a better indication of an outlier may be how samples behave relative to their biological replicates (Affymetrix, 2008) (Table 6). Within experimental group comparisons showed no evidence of outlier behaviour again this was consistent between analyses performed in Partek GS and Expression Console. As the wildtype sample, 17c satisfied criteria both within and across groups it was deemed of adequate quality for gene expression analyses. Due to the ambiguous outlier nature of sample 6, I plotted several MA plots to establish the extent to which the intensity of this sample varied when directly compared with each of the other samples and to enable the decision of whether the sample should be excluded (Fig. 3.3.12).

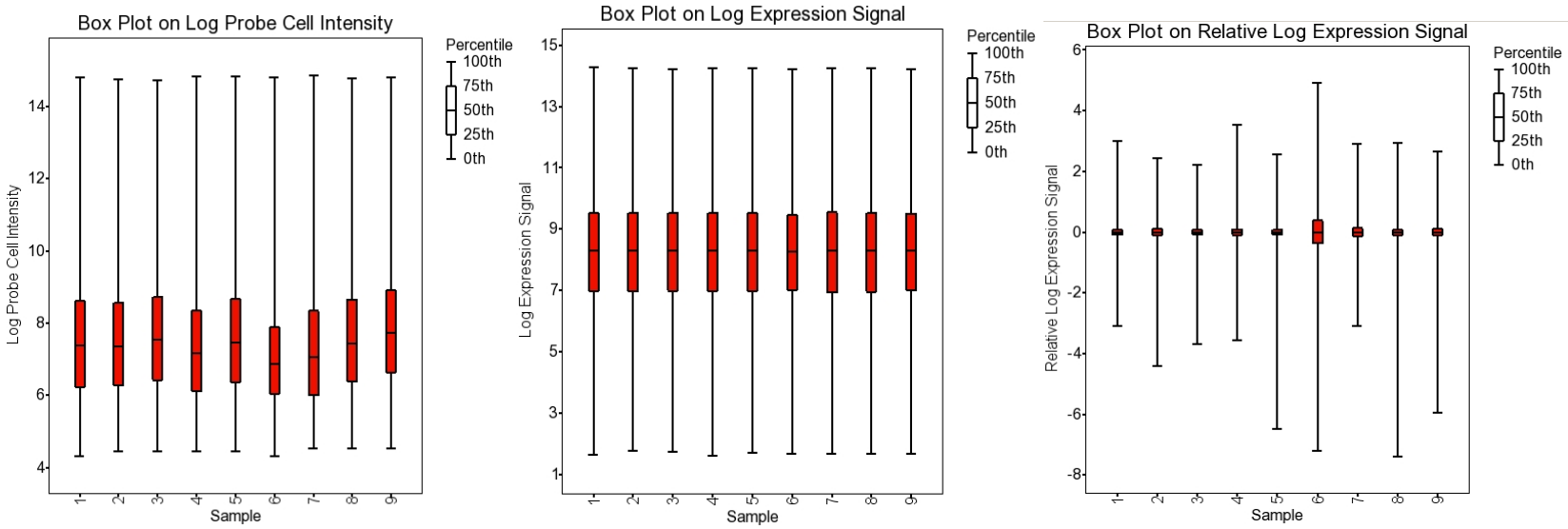


Figure 3.3.11. Quality Control. Boxplots of Expression Signals. The variation observed in the probe cell intenisty before normalisation and summarisation (left) is corrected following normalisation and summarisation (centre). The relative log expression signal boxplot (right), which is used to give an indication of samples behaving differently to the consensus, demonstrates sample 6 has a much wider IQR. Samples with a larger IQR may represent a low quality array.

Sample ID	PM Mean	All Probeset Mad Residual Mean	Pos Control Mad Residual Mean	All Probeset RLE Mean	Pos Control RLE Mean	Pos vs. Neg AUC
Z2C.CEL	414.85	0.12	0.09	0.13	0.10	0.86
Z2D.CEL	396.00	0.13	0.10	0.15	0.13	0.87
Z2E.CEL	430.13	0.12	0.09	0.13	0.10	0.86
Z4A.CEL	356.77	0.14	0.09	0.14	0.11	0.86
Z6A.CEL	435.08	0.12	0.09	0.12	0.09	0.86
Z17B.CEL	**276.81**	**0.23**	**0.16**	**0.48**	**0.49**	0.88
Z17C.CEL	381.87	0.15	0.10	0.19	0.14	0.86
Z21B.CEL	437.45	0.12	0.09	0.14	0.11	0.87
Z21C.CEL	484.40	0.13	0.10	0.15	0.13	0.86
Mean -2xSD	**282.63**	**0.07**	**0.05**	**-0.04**	**-0.10**	**0.86**
Mean +2xSD	**520.34**	**0.21**	**0.15**	**0.41**	**0.41**	**0.88**
Within Threshold	**No**	**No**	**No**	**No**	**No**	**No**

Table 5. Quality Control. 6 Quality Control Metrics Generated for Male wildtype and C59X mutants. In 5 of the 6 metrics the values generated for sample Z17B were outside of the thresholds recommended by Affymetrix (red).

C59X Mutants						
Sample ID	**PM Mean**	**All Probeset MAD Residual Mean**	**Pos Control MAD Residual Mean**	**All Probeset RLE Mean**	**Pos Control RLE Mean**	**Pos vs. Neg AUC**
Z2D.CEL	396.00	0.13	0.10	0.15	0.13	0.87
Z2E.CEL	430.13	0.12	0.09	0.13	0.10	0.86
Z4A.CEL	356.77	0.14	0.09	0.14	0.11	0.86
Z17B.CEL	276.81	0.23	0.16	0.48	0.49	0.88
Mean -2*SD	**233.03**	**0.05**	**0.04**	**-0.11**	**-0.18**	**0.86**
Mean +2*SD	**496.83**	**0.26**	**0.18**	**0.56**	**0.59**	**0.88**
Within Threshold	**Yes**	**Yes**	**Yes**	**Yes**	**Yes**	**Yes**
Wildtypes						
Sample ID	**PM Mean**	**All Probeset MAD Residual Mean**	**Pos Control MAD Residual Mean**	**All Probeset RLE Mean**	**Pos Control RLE Mean**	**Pos vs. Neg AUC**
Z2C.CEL	414.85	0.12	0.09	0.13	0.10	0.86
Z6A.CEL	435.08	0.12	0.09	0.12	0.09	0.86
Z17C.CEL	381.87	0.15	0.10	0.19	0.14	0.86
Z21B.CEL	437.45	0.12	0.09	0.14	0.11	0.87
Z21C.CEL	484.40	0.13	0.10	0.15	0.13	0.86
Mean -2*SD	**356.01**	**0.11**	**0.08**	**0.10**	**0.08**	**0.86**
Mean +2*SD	**505.45**	**0.15**	**0.10**	**0.20**	**0.15**	**0.88**
Within Threshold	**Yes**	**Yes**	**Yes**	**Yes**	**Yes**	**Yes**

Table 6. Quality Control. Considering the 6 quality control metrics by experimental groups. When considering the samples as biological replicates within their experimental groups none of the samples had values which surpassed the threshold in any of the 6 metrics (When considering more than 2 decimal places).

To evaluate the exclusion of sample 17b, MA plots (3.3.12) were generated. The MA plot is used to compare the intensity of two samples. Each point in the plot represents a probe intensity value. The average for each probe intensity value is plotted on the x-axis with the difference between the two samples plotted on the y-axis. I first generated plots comparing sample 17b to all other samples (Fig. 3.3.12A & B) and then by exploring all sample combinations with the exception of those including 17b (Fig. 3.3.12C & D). To further evaluate the decision to retain sample 17c this was also plotted against sample 21b. The variation in intensity was much greater in sample 17b and so it was decided that this samples would be excluded and omitted from all downstream analyses. MA plots of sample 17c were consistent with the other samples and so confirmed the decision to retain the sample in the statistical analysis. QC following the removal of 17b (Table 3.7& 3.8) did not indicate any outliers and therefore following the exclusion of sample 17b the remaining 8 male samples (3 homozygote C59X mutants and 5 wildtypes) were pre-processed (Ch2.8.3.3). The Partek GS and Expression console QC analyses varied only in that sample 17c was outside the 2 standard deviation rule for one metric in Partek but two in expression console. This still does not qualify 17c as an outlier, but highlights the slight variation in metrics generated by the two programmes.

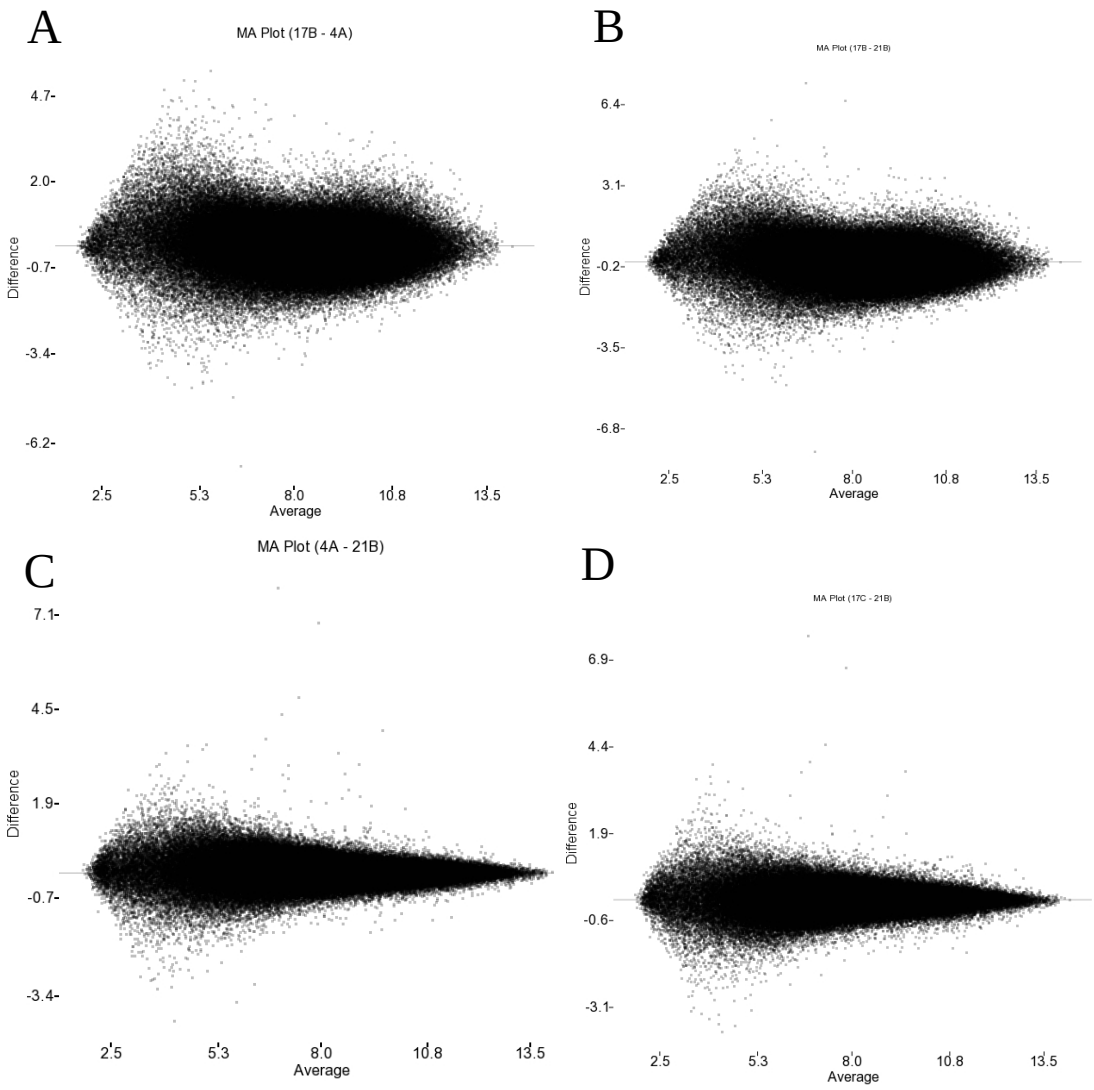


Figure 3.3.12. MA plots of signal Intensity. Signal intensities at all probesets on the array were directly compared between two samples at a time and the difference in intensity at each probeset plotted. **A** Comparison of potential outlier 17b and another male mutant sample (4a). This plot was representative of all plots generated when comparing each male sample in turn with 17b as exemplified in **B** where 17b was compared to 21B, a male wildtype. There was considerable variation between the two samples, in each case, particularly when considering probesets with higher average intensity values as indicated by the wider spread of the data at the right of the plot (A &B). When the same two samples (4a and 21b) were compared to each other **C** and when 21b was compared to 17c **D** the plots all had a reduced variation toward the right of the graph representing probesets with higher average intensity values. Both plots were representative of those generated when comparing any 2 male samples with the exception of 17b.

Sample ID	PM Mean	All Probeset Mad Residual Mean	Pos Control MAD Residual Mean	All Probeset RLE Mean	Pos Control RLE Mean	Pos vs Neg AUC
Z2C.CEL	414.85	0.12	0.09	0.13	0.11	0.86
Z2D.CEL	396.00	0.13	0.10	0.16	0.13	0.87
Z2E.CEL	430.13	0.12	0.09	0.13	0.10	0.86
Z4A.CEL	356.77	0.14	0.09	0.15	0.11	0.86
Z6A.CEL	435.08	0.12	0.09	0.12	0.09	0.86
Z17C.CEL	381.87	0.15	0.10	**0.19**	0.13	0.86
Z21B.CEL	437.45	0.12	0.09	0.14	0.11	**0.87**
Z21C.CEL	484.40	0.13	0.10	0.15	0.13	0.86
Mean -2xSD	**338.63**	**0.11**	**0.08**	**0.11**	**0.09**	**0.86**
Mean +2xSD	**495.51**	**0.15**	**0.10**	**0.19**	**0.14**	**0.87**
Within Threshold	**Yes**	**Yes**	**Yes**	**No**	**Yes**	**No**

Table 3.7. QC measure following the removal of 17b showed the data to be consistant and no samples were determined to be outliers.

C59X Mutants						
Sample ID	**PM Mean**	**All Probeset MAD Residual Mean**	**Pos Control MAD Residual Mean**	**All Probeset RLE Mean**	**Pos Control RLE Mean**	**Pos vs Neg AUC**
Z2D.CEL	396.00	0.13	0.10	0.16	0.13	0.87
Z2E.CEL	430.13	0.12	0.09	0.13	0.10	0.86
Z4A.CEL	356.77	0.14	0.09	0.15	0.11	0.86
Mean -2xSD	**320.88**	**0.12**	**0.09**	**0.12**	**0.09**	**0.86**
Mean +2xSD	**467.73**	**0.14**	**0.10**	**0.17**	**0.14**	**0.87**
Within Threshold	**Yes**	**Yes**	**Yes**	**Yes**	**Yes**	**Yes**
Wildtypes						
Sample ID	**PM Mean**	**All Probeset MAD Residual Mean**	**Pos Control MAD Residual Mean**	**All Probeset RLE Mean**	**Pos Control RLE Mean**	**Pos vs. Neg AUC**
Z2C.CEL	414.85	0.12	0.09	0.13	0.11	0.86
Z6A.CEL	435.08	0.12	0.09	0.12	0.09	0.86
Z17C.CEL	381.87	0.15	0.10	0.19	0.13	0.86
Z21B.CEL	437.45	0.12	0.09	0.14	0.11	0.87
Z21C.CEL	484.40	0.13	0.10	0.15	0.13	0.86
Mean -2xSD	**356.01**	**0.11**	**0.08**	**0.10**	**0.08**	**0.86**
Mean +2xSD	**505.45**	**0.15**	**0.10**	**0.20**	**0.15**	**0.88**
Within Threshold	**Yes**	**Yes**	**Yes**	**Yes**	**Yes**	**Yes**

Table 3.8 Male C59X mice Quality Control Metrics. Following the removal of sample 17b normalisation and summarisation were again carried out in Partek Genomics Suite and the same metrics were generated to analyse the quality of each array in order to identify any outliers. Samples were separated by genotype prior to determining if any individual arrays were outside of a pre-defined threshold of 2 standard deviations from the mean for each of the metrics generated.

The 8 samples were preprocessed simultaneously using quantile normalisation and median polish summarisation (Ch2.8.3.3). As with the female samples, I next determined the number of transcripts predicted to be differentially expressed and spliced. The same criteria as used in the female analysis enabled comparisons to be made across the two experiments and the degree of overlap to be established.

Differential Expression	**No. of Significant Genes**	**% of Genes**
Unadjusted p≤0.05	744	4.7
p≤0.01	152	0.96
p≤0.001	25	0.16
p≤0.0001	6	0.04
Following Multiple Test Correction		
FDR 0.05	0	0
Bonferroni Correction (15813 tests)	0	0

Table 3.9. Total number of transcripts included in the analysis was 15,813.

Approximately 5% of the total number of transcripts were significantly differentially expressed at a nominal p value (p≤0.05) (Table 3.9). Following correction for multiple testing no genes were significant using an FDR threshold of 0.05 and Bonferroni correction for 15,813 tests respectively. A slightly larger proportion of the total number of transcripts (15,813) was predicted to be differentially spliced. As seen in the female mice the numbers remaining significant were greater for differentially spliced genes than differentially expressed genes, with a small proportion of genes remaining statistically significant following multiple test correction (Table 3.10).

Alternative Splicing	No. of Significant Genes	% of Genes
Unadjusted p≤0.05	1240	7.84
p≤0.01	462	2.92
p≤0.001	143	0.9
p≤0.0001	68	0.43
Following Multiple Test Correction		
FDR 0.05	99	0.63
Bonferroni Correction (15813 tests)	34	0.22

Table 3.10. Differentially Spliced Genes in Male C59X Mutant Mice. ~7% of the total number of transcripts were found to be differentially spliced in the C59X mutants (p≤0.05) with ~0.6% still significant following correction for multiple testing using the FDR at a threshold of 0.05.

3.3.4.5 Multiple Test Correction.

Pre-analytical filtration steps, including probeset filtering by intensity, were applied to the data to reduce the impact of multiple testing (Della et al., 2008). To further reduce the number of potential false positives in the data the Bonferroni correction and the Benjamini and Hochberg FDR correction were applied to the data. Whether or not MTC should be applied to data at the exon level is still a matter of contention. In alternative splice analysis the data are considered at the level of the probeset (exon) relative to the expression across a transcript. Thus far more data points are included in the analysis and the exons belonging to a given transcript are dependent (correlated) and thus violate the assumption of independence of data points in many multiple test correction approaches, including the Benjamini and Hochberg FDR. Whilst the application of the correction may not be statistically accurate, the exemption of such a correction would most likely result in a high proportion of false positives in the dataset.

One way to establish the contribution of false positives to the number of significant genes would be to permute the data and observe if the number of significant genes varies considerably to those originally observed. A similar proportion of significant

genes in the permuted data set would be suggestive of a high degree of false positives. This option could not be applied to the data as random permutations and subsequent p values could not be generated on such a small data set. As a compromise and in the absence of adult heterozygote samples, cases were compared with other cases and controls with other controls in order to understand the contribution of false positives to the results. The total of 16 male and female samples were split by genotype comprising of 7 mutants and 9 wildtypes and then within in each genotype group samples were split balancing litter and gender as best as possible. First the wildtype controls were divided into either group 1 or 2, with 4 and 5 samples in each respectively. The preprocessing was carried out in the same way as described previously the only difference in the analysis was that differential expression and alternative splicing were determined between the two arbitrary groups (1 and 2) as only wildtype samples were included. This enabled an estimate to be determined of the number of genes predicted to be differentially expressed and spliced by chance alone. The same was repeated for the mutant samples comparing group 1 (3 mutants) to group 2 (4 mutants) again balanced by gender and litter. Gender was covaried in each instance, as the PCA in both sets of data showed clustering of samples by gender. A max intensity filter of Log_2 3 was used with probesets included if they were below this threshold but showed significant results $p \leq 0.05$. A considerable number of genes are significant (Table 3.11) even when comparing samples of the same genotype split into arbitrary groups, but the numbers are less than those observed in the actual analysis comparing mutants to wildtypes in each instance. This suggests there are true positives in the data.

Differential Expression	p≤0.05	FDR 0.05	Bonferroni Correction
Wildtype vs. Wildtype	559	0	0
Mutants vs. Mutants	712	0	0
Females Wildtype vs. Mutant	968	0	0
Males Wildtype vs. Mutant	744	0	0
Differential Splicing			
Wildtype vs. Wildtype	352	7	5
Mutants vs. Mutants	580	22	9
Females Wildtype vs. Mutant	1010	79	21
Males Wildtype vs. Mutant	1240	99	34

Table 3.11. False Positives Rate. The number of significant differentially expressed or spliced genes are represented in the table either at an unadjusted p value of p≤0.05 or following multiple test correction with an FDR threshold of 0.05 or with the Bonferroni correction for the number of tests performed. The data generated by comparing within genotype groups is shown in the table with the original data, generated by comparing C59X mutants and wildtypes below (blue). The number of significant differentially expressed genes was slightly greater in the original comparison relative to the within genotype dataset suggesting the results are more than would be expected by chance artefacts. The number of predicted differentially spliced genes observed between C59X mutants and wildtypes was approximately double that observed in the within genotype data. This suggests that the predictions of differential splicing are not just chance findings produced by artefacts of the statistical algorithm.

3.3.4.6 Analysis to Determine the Degree of Replication.

Significant results are usually prioritised for validation based on significance, fold change and functional relevance to the disease being studied. An advantage of having both female and male datasets was that it enabled the degree of overlap across the two studies to be assessed. Genes found to overlap, despite the differences in gender, brain preparation and age would therefore be considered more robust and make good candidates for validation by quantitative PCR (qPCR). Overlap was measured by comparing the observed number of overlapping genes between males and females to the number that would be expected to overlap by chance. If the number of significant changes that overlap were greater than the number expected by chance the data would be considered more robust. To precisely calculate the overlap, 2x2 contingency tables were produced and the significance of the overlap was tested using the Chi squared test (Methods 3.2.11). A summary of the results is shown in Table 3.12. A significant overlap was found for differentially spliced transcripts at all stringency thresholds

investigated. Overlap for differentially expressed transcripts was less convincing with a significant overlap observed only for transcripts found to be significantly differentially expressed in the females at $p \leq 0.01$. This suggests the splicing data were more robust than the differential expression data.

3.3.4.7 Manual inspection of Results to Determine Direction of Effect.

Due to the significant overlap of differentially spliced transcripts across the two experiments a manual inspection was undertaken to establish if the replicating transcripts were predicted to show splice events in the same exon and have the same direction of effect. Starting with the most stringent set of overlapping transcripts (those significant in the females at $p<0.0001$ and significant in the males at $p<0.05$) the geneviews (Fig. 3.3.13), which I generated in Partek GS, were manually inspected to establish the exon(s) generating the significant splice signal in the female dataset and whether the same event was evident in the males. If the same exon(s) was differentially spliced in the males I then determined if the direction of effect was concordant.

N.B: The geneview displays the probeset expression of all probesets within a transcript cluster. Varying levels of intensity are evident for each probeset, however this does not represent true intensity differences across the transcript and is often attributable to sequence specific probe effects. Therefore when assessing the geneview for a real splice event, the differences in intensity between probesets should be ignored and the parallel nature of the lines joining the probesets should be considered. If parallel lines are apparent across the length of the transcript this suggests that the transcript is differentially expressed between the two groups. Where there is a divergence from the parallel nature of the lines this is indicative of differential splicing of an exon between the two groups.

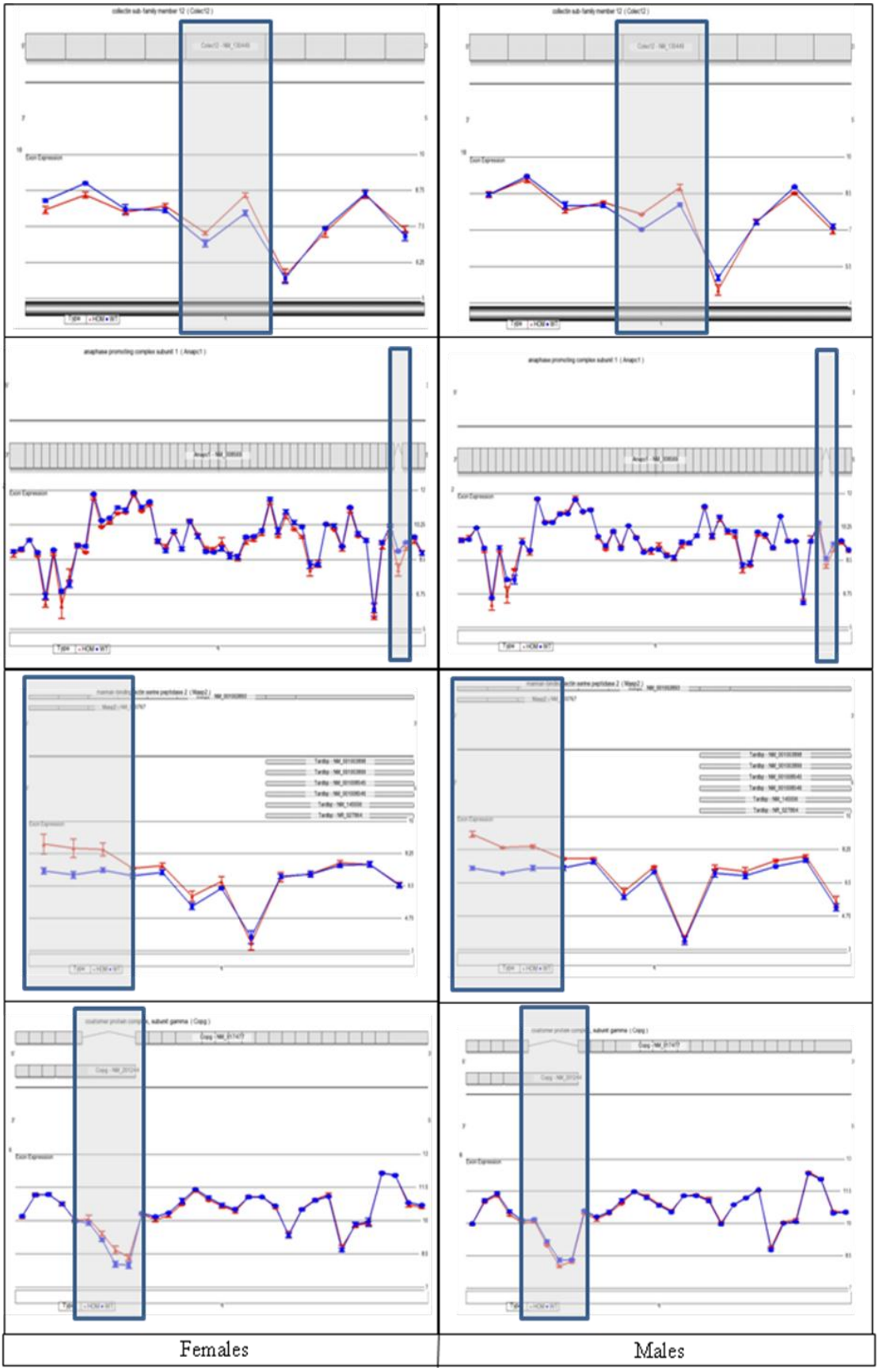
Differential Expression

Significance Threshold	Observed	Expected	Exact sig 2-sided	Exact sig 1-sided	Replication
P≤0.05	54	45.4	0.18	0.11	Not significantly more than expected by chance
P≤0.01	17	10.6	0.06	0.04	Significantly more than expected by chance (p<0.05)
P≤0.001	2	1.1	0.30	0.30	Not significantly more than expected by chance (Fisher's Exact test)
P≤0.0001	1	0.1	0.09	0.09	Not significantly more than expected by chance (Fisher's Exact test)

Alternative Splicing

Significance Threshold	Observed	Expected	Exact sig 2-sided	Exact sig 1-sided	Replication
P≤0.05	111	79.2	0.00	0.00	Significantly more than expected by chance
P≤0.01	61	31.4	0.00	0.00	Significantly more than expected by chance
P≤0.001	34	10.7	0.00	0.00	Significantly more than expected by chance alone
P≤0.0001	23	4.2	0.00	0.00	Significantly more than expected by chance (Fisher's Exact test)

Table 3.12. Replication Analysis using Chi Square Test. Significant replication in the male study was determined by establishing whether the number of genes that overlapped was greater than the number expected by chance using Pearson's Chi Square test. Results were determined for both one and two-tailed hypotheses, however as the null hypothesis was that results would not replicate in the male study the one-tailed significance value was used to determine replication. Only when considering the 226 genes significant at p≤0.01 in the females is an overlap observed in the males which is significantly more than would be expected by chance (χ^2 (1) = 4.061, p<0.05) for differentially expressed genes. In contrast at every significance threshold the genes significantly differentially spliced in the females overlap with numbers significantly more than would be expected by chance. Due to an expected count less than 5 the Fisher's Exact test significance value is reported for differential expression p≤0.001, p≤0.0001 and alternative splicing p≤0.0001.



Females	Males

Figure 3.3.13 Replication of Potential Splice Events. After I generated geneviews in Partek GS, I manually inspected plots of all genes showing significant overlap to determine if the position and direction of the splice events was consistent. At the top of each geneview the known Refseq transcript(s) is plotted, with the $\log_2$ expression values for each of the probesets (within the defined transcript cluster) plotted underneath. Probeset expression is plotted for the C59X mutants and wildtypes in red and blue respectively. In this figure, the geneviews of four genes are displayed with the female and male plots on the left and right hand side respectively. The splice event thought to be generating the significant splice p value is highlighted within the shaded box. The top 3 plots show genes in which the same splice event is observed in both males and females. The bottom plot shows potential differential isoform expression between mutants and wildtypes in the females, however only a subset of the probesets are also differentially expressed in the males and in one probeset the effect is in the opposite direction.

Geneviews were inspected (Fig. 3.3.14) to establish how many of the 17 significantly, overlapping differential expression results were consistently regulated in the same direction in both males and females (Table 3.13). 14 genes (82%) were consistently up or downregulated in the females and males. Differential spliced transcripts were inspected to ensure the position of the predicted spliced exon(s) and the direction of effect was consistent in both experiments (Table 3.13). A significant overlap between males and females was observed at each of the p value stringencies tested. Each set of genes was inspected starting with the most stringent group ($p \leq 0.0001$). Of the 111 genes found to overlap between male and female experiments, 70% replicated in at least one of the exons predicted to be differentially spliced and in the same direction. With increased p value stringency the percentage of splice events found to be consistent across the two experiments increased, with 91% of genes overlapping at $p \leq 0.0001$ showing concordant changes. The observed degree of overlap not only increased the confidence in the data, but suggests that the predicted splice events are not resultant of technical artefacts such as probe bias.

Differential Expression

Significance threshold in Females	No. Genes Significant	Overlap in Males (p≤0.05)	No. of Genes with Same direction of Effect	%
p≤0.001	226	17	14	82

Alternative Splicing

Significance threshold in Females	No. Genes Significant	Overlap in Males (p≤0.05)	No. of Genes in which the Differentially spliced Exon and Direction of Effect is Consistent	%
p≤0.0001	54	23	21	91
p≤0.001	136	34	29	85
p≤0.01	400	61	48	79
p≤0.05	1009	111	78	70

Table 3.13 Determining the consistency in the overlap in Male and Female Results. Manual inspection of both females and male geneviews allowed the percentage of overlapping genes in which the differential expression (top) or splice (bottom) event was consistent to be determined. Differentially expressed genes had to have altered expression in the same direction in both females and males to be counted. Consistent overlap in a differentially spliced transcript required that the splice event, thought to be contributing to the splice signal in females, was replicated in the males and that the change was in the same direction.

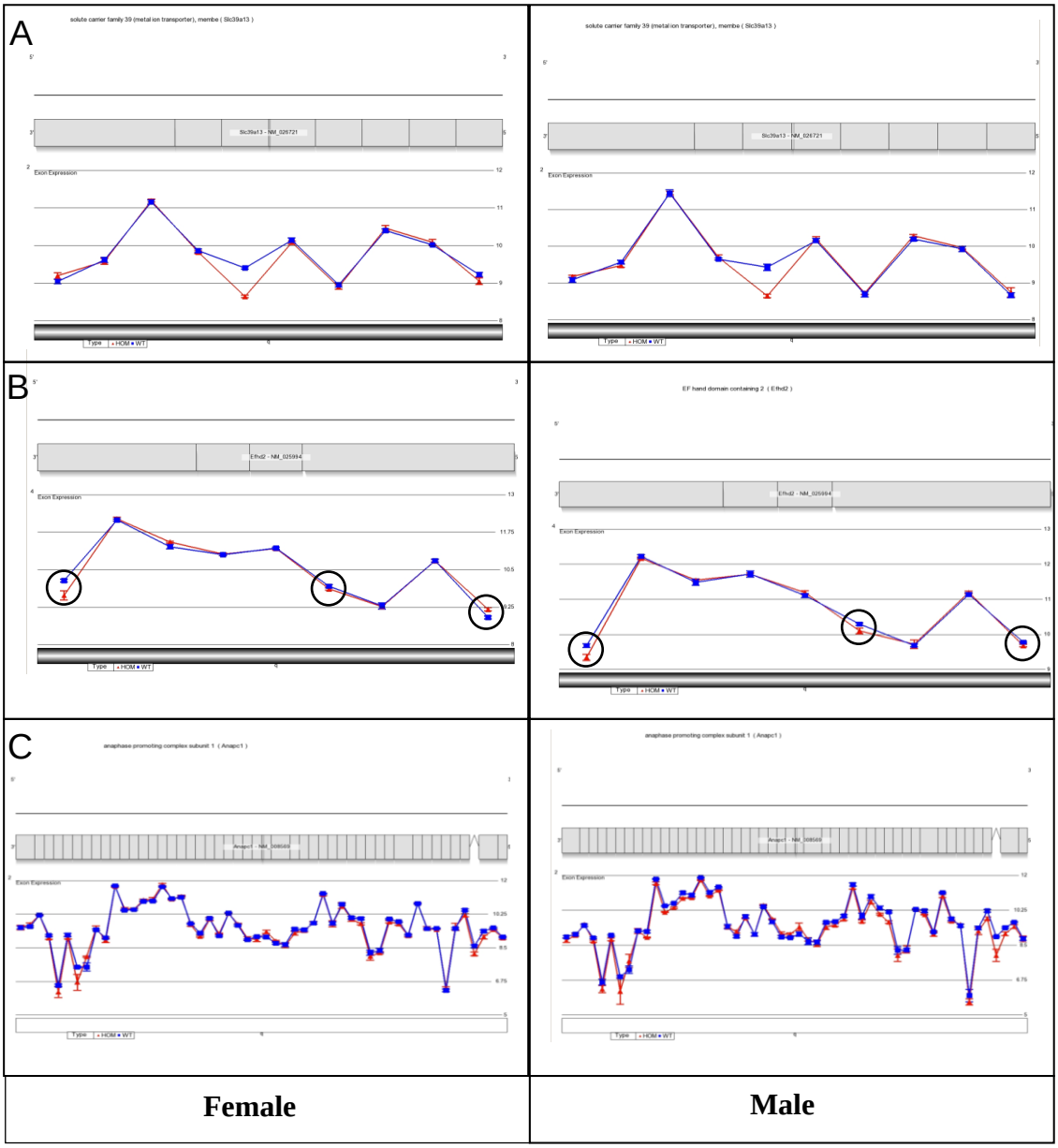


Figure 3.3.14 Inspection of Alternative Splicing Results. The geneviews displayed represent 3 genes (1 in each row) with the expression results for both females (left column) and males (right column). **A.** The geneview for Slc39a13 is an example of a result in which the splice event was easy to identify, demonstrated by a similar expression pattern in wildtypes and mutants across all probesets in the transcript except one, where a clear divergence in the expression between mutants and wildtypes was observed. **B.** An example of a transcript which only partially replicates across the male and female experiments. 3 differentially expressed probesets (circled) were identified in females, but only two of these are differentially expressed in the same direction in the males. **C.** The geneview is representative of transcripts in which the large number of probesets made it difficult to determine where the alternative splice signal was coming from, which made determining if there was an overlap between males and females difficult.

3.3.4.8 Combined Analysis.

Due to the observed overlap in the two individual experiments I felt it was appropriate to combine the samples to increase the power and then repeat the analysis. This meant the expression and splicing was compared between 9 wildtypes (4 female, 5 male) and 7 C59X mutants (4 female, 3 male) using a two-way ANOVA model, which included gender as a covariate. A total of 15,833 transcripts were tested and the proportion of these predicted to be significantly differentially expressed or spliced is shown in Table 3.14. The number of transcripts significantly differentially expressed at a nominal p value ($p \leq 0.05$) was 7% of the total number of transcripts, more than observed in the individual female and male experiments (6% and 4% respectively). There were also a small number of genes in the combined experiment which survived correction for multiple testing even after using the very stringent Bonferroni correction. The impact of combining the samples did not increase the number of differentially spliced as expected. With the increase in power a decreased number of predicted splice events was observed. From the 7% and 6% observed previously the proportion of spliced transcripts decreased to 4%, almost half the number of differentially expressed transcripts.

Differential Expression	No. of Significant Genes	% Of Genes	Expected
Unadjusted p≤0.05	1043	7	792
p≤0.01	242	2	158
p≤0.001	30	0.19	16
p≤0.0001	11	0.07	2
FDR 0.05	9	0.06	
Bonferroni Correction (15,833 tests)	3	0.02	

Alternative Splicing	No. of Significant Genes	% Of Genes	Expected
Unadjusted p<0.05	617	4	792
p≤0.01	226	1	158
p≤0.001	92	1	16
p≤0.0001	54	0.34	2
FDR 0.05	65	0.41	
Bonferroni Correction (15,833 tests)	29	0.18	

Table 3.14. Combined Analysis Results. The top table shows the proportion of genes differentially expressed between C59X mutants and wildtypes. A small number of genes remained significant following multiple test correction (FDR 0.05 and Bonferroni), something which was not observed when considering the female and male experiments separately.

The combined analysis showed 3 genes to be significant following a Bonferroni correction of 15,833 tests and 9 using the less stringent FDR correction at a threshold of 0.05. Whilst this number is small it is still important to consider these genes, as one or more of them may have downstream effects on the expression or splicing of other genes. For this reason these 9 genes were viewed to determine how likely the expression change was (Appendix 3.1). Of the 9 only 2 looked like differential gene expression (Mettl5 and Nfe212) with two others possibly being differential expression of known alternative transcripts. The remaining 5 showed differential expression in only a subset of the probesets within the transcripts, which could suggest differential expression of novel isoforms. Of the 3 Bonferroni significant transcripts only Mettl5 displays attributes of a differentially expressed gene, with all probesets showing differential expression in the geneview. None of the 9 transcripts had differential expression fold

changes greater than or equal to 1.5 which is a standardised cut off for differential expression, reducing the reliability of the results.

The top 6 results from the combined analysis for differential splicing are shown in Fig 3.3.15. In each case the predicted splice event was inspected to determine the likelihood of such an event. Of the six events 2 were potential false positives as only one of multiple probesets in the exon showed differential expression. From this Frzb, Slc39a13, Fam171b and Ssfa2 remain potential differential splice candidates.

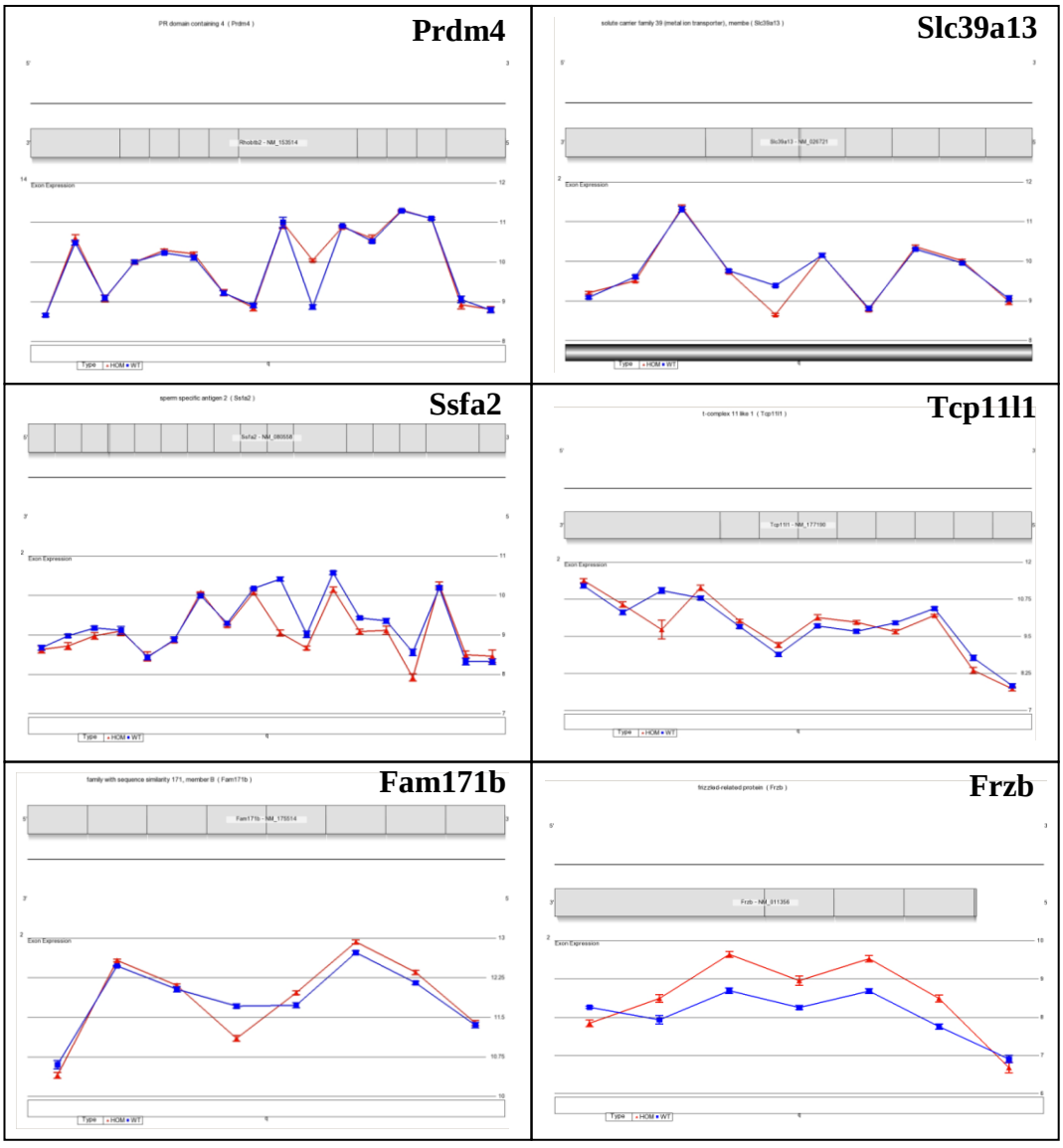


Figure 3.3.15. Geneviews of Significantly Differentially Spliced Genes (Combined Analysis). The majority of cases appear to be called due to alternative exon usage. This type of splice event is the most common and also the easiest to detect in this type of analysis. The event in Tcp11l1 is suggestive of a false positive as not all probesets within the exon follow the same expression pattern, as is this case in Prdm4. Frzb's

geneview is suggestive of an unknown alternative isoform, or 3' and 5' edge effects masking a differentially expressed rather than spliced transcript.

3.3.4.9 Alternative Algorithms.

Assessment of differential expression and splicing were carried out using alternative algorithms in different software packages to establish the robustness of the data. Both easyExon and AltAnalyze are open source, freely available applications. In easyExon the MiDAS algorithm was used to determine differential splicing and the FIRMA algorithm was applied in AltAnalyze, the methods for each are described in full in Chapter 2.8.4. Identification of the same differentially expressed and spliced genes when using different algorithms and filtering criteria would add validity to the data.

3.3.4.9.1 easyExon.

The false positive rate was controlled by first removing probesets which did not have a DABG p value ≤0.05 in at least 50% of the samples. With a significance threshold of p≤0.05 and a fold change cut off of greater than or equal to 1.5, 34 genes were predicted to be differentially expressed in the female experiment of which 10 had RefSeq IDs. 16 genes were significant in Partek when applying the same thresholds. Of the 10 easyExon results with Refseq IDs, 8 genes were found to overlap between Partek GS and easyExon (Table 3.15). Using EasyExon to assess the male experiment, 26 genes met the criteria for differential expression, but only 1 had a gene symbol; *Frizzled related protein* (*Frzb*). Applying the same significance and fold change cutoffs in Partek GS resulted in three differentially expressed genes, one of which was Frzb.

741 and 764 genes were significantly differentially spliced in female and male C59X mice respectively. This corresponds to the 1010 and 1240 genes identified when using Partek GS. Genelists were cross referenced with significant differentially spliced genes in Partek GS using the Affymetrix transcript cluster ID annotation. 146 and 197 genes overlapped in the female and male datasets respectively. Whilst the nominally significant numbers vary quite considerably this probably reflects the different algorithms and stringency cut offs used. The number of genes which do overlap are

similar to the number significant in Partek GS at p≤0.001 and perhaps represents more reliable calls of alternative splicing by both algorithms.

Gene Symbol	RefSeq ID	Gene Description
Snca	NM_001042451	Synuclein, alpha
Egr2	NM_010118	early growth response 2
Arc	NM_018790	activity regulated cytoskeletal-associated protein
Scg5	NM_009162	secretogranin V
Mela	D10049	melanoma antigen
Npas4	NM_153553	Neuronal PAS domain protein 4
Dusp1	NM_013642	dual specificity phosphatase 1
Nr4a1	NM_010444	nuclear receptor subfamily 4, group A, member 1

Table 3.15 Significantly Differentially Expressed Genes. Using a p value cut off of p≤0.05 (nominal) and a fold change threshold of greater than or equal to 1.5, 8 genes were significant using both Partek GS and EasyExon software tools on the female dataset.

3.3.4.9.2. AltAnalyze (Version 2.0.7).

AltAnalyze is an integrated software workflow (Emig et al., 2010) used to preprocess samples and carry out statistical analysis. This software has previously been shown (AltAnalyze manual) to predict splicing with high specificity and reasonable sensitivity in a previously published dataset (Xing et al., 2008). Both the adult female and male datasets were analysed using AltAnalyze to detect differential gene expression and splicing as described in Chapter 2.8.4.2.

The results of the analyses are shown in table 3.16. Following software specific filtering, as described in chapter 2 which included the core meta-probeset and excluded probesets with a DABG p value ≥0.05 and a non log expression value ≤70, the number of genes predicted to show differential expression and splicing were calculated (Table 3.16). Overlap with Partek was then determined using the transcript cluster ID. First considering the female data the differential expression overlap showed that of the 1464 genes significant in AltAnalyze and the 968 in Partek, 373 overlapped when using the transcript cluster ID. It is important to note that due to the different filtering methods

and different annotation criteria (chapter 2.8.4.2) that the probesets included in the two statistical analyses will have varied. 860 transcripts were predicted to show differential splicing in AltAnalyze compared to the 1010 predicted by Partek GS. The overlap between Partek and AltAnalyze for differential splicing in female C59X mutants and wildtype was 167 transcripts. Next considering the male experiment, 1272 genes were significantly differentially expressed in AltAnalyze and 744 in Partek, of which 234 overlapped between the two analyses. 966 genes were predicted to be differentially spliced in AltAnalyze and 1240 in Partek, of which 212 transcripts were significant in both.

There is a degree of overlap observed considering the differences in the initial annotation of genes and the differences in the algorithms used. To determine if any of the predicted expression or splice differences between the C59X mutants and wildtypes were common to all 3 of the software programmes the overlap was established again using the Transcript cluster ID as the common identifier.

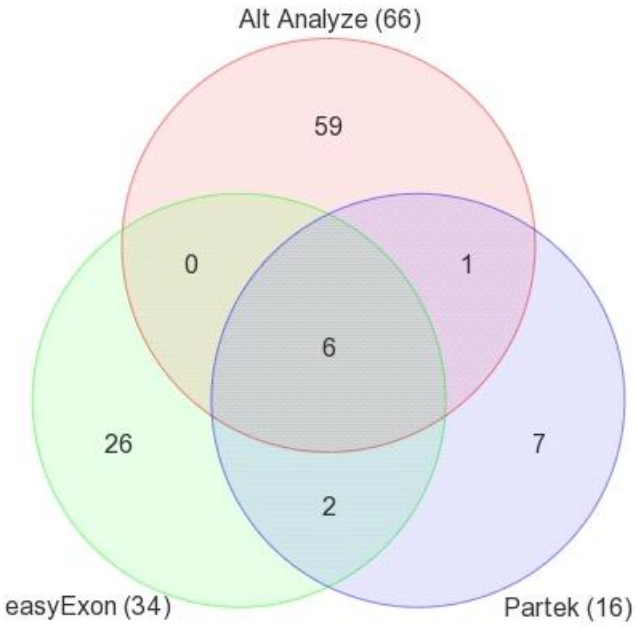
Differential Expression	**AltAnalyze**		
Female	1464		
Male	1272		
Differential Splicing			
Female	860		
Male	966		
Differential Expression	**AltAnalyze**	**Partek**	**Overlap**
Female	1464	968	373
Male	1272	744	234
Differential Splicing	**AltAnalyze**	**Partek**	**Overlap**
Female	860	1010	167
Male	966	1240	212

Table 3.16. AltAnalyze Results and Overlap with Partek GS. The number of differentially expressed and spliced genes predicted when using the algorithms in AltAnalyze.

3.3.4.10 Significant Results Overlap between easyExon, AltAnalyze and Partek GS.

When using easyExon, by default significance and fold change criteria are applied so that only genes with a significance p value ≤0.05 and an accompanying fold change of greater than or equal to 1.5 are considered differentially expressed between C59X mutants and wildtypes. Due to the additional fold change threshold the number of genes in the easyExon output predicted to be differentially expressed is considerably lower than that predicted using Partek and AltAnalyze. For consistency the same fold change cut offs (greater than or equal to 1.5) were applied to the Partek and AltAnalyze results prior to determining any overlap.

Caveats to determining the overlap lie in the different ways in which the data were annotated, filtered and the statistical approach as well as the different ways in which cross-hybridising probesets are dealt with. First considering genes predicted to be differentially expressed between female C59X mutants and wildtype only 6 genes were found to be common to all 3 different methods (Fig. 3.3.16). These genes were *Arc, Dusp1, Egr2, Npas4, Nr4a1* and *Snca*. When considering that only 16 genes in Partek GS meet both the significance (p≤0.05) and fold change (greater than or equal to 1.5) criteria then over a third of these are significant in the other two analyses. None of the 6 overlap with the 9 FDR (0.05) significant genes from the combined analysis but this may reflect the small fold changes (less than or equal to 1.5) observed in the 9 genes and reduced the likelihood of them being true expression changes. In contrast to the 9 genes from the combined analysis, all 6 of these overlapping expression candidates looked like true expression changes from the manual inspection of the geneview. Due to the robust nature of these 6 candidates, in terms of significance from multiple algorithms and fold changes above a suggested 1.5 cut-off each one is a candidate for validation using an independent assay.



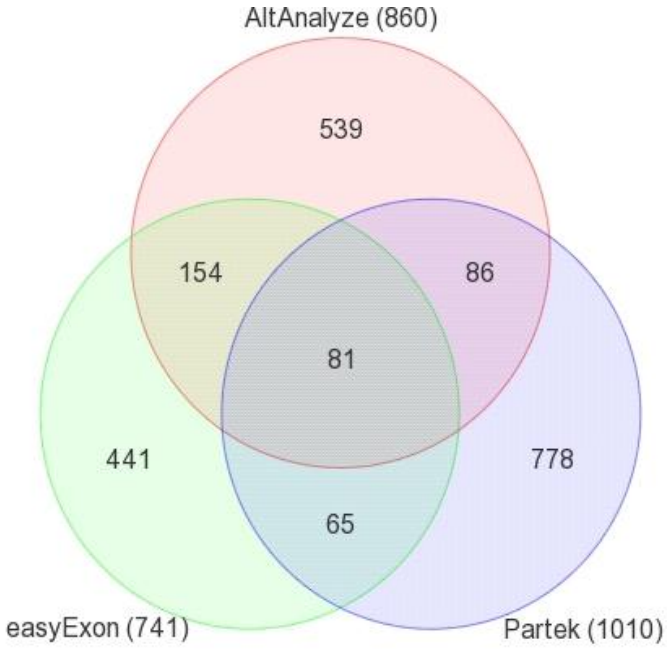


Figure 3.3.16. Genes Overlap. Genes predicted to be significantly differentially expressed (top) between Partek GS, AltAnalyze and easyExon. Only 6 genes were found to be common to all 3 significant gene lists, these genes were Arc, Dusp1, Egr2, Npas4, Nr4a1 and *Snca*. 81 genes were predicted to be differentially spliced when using Partek GS, AltAnalyze and easyExon for analyses.

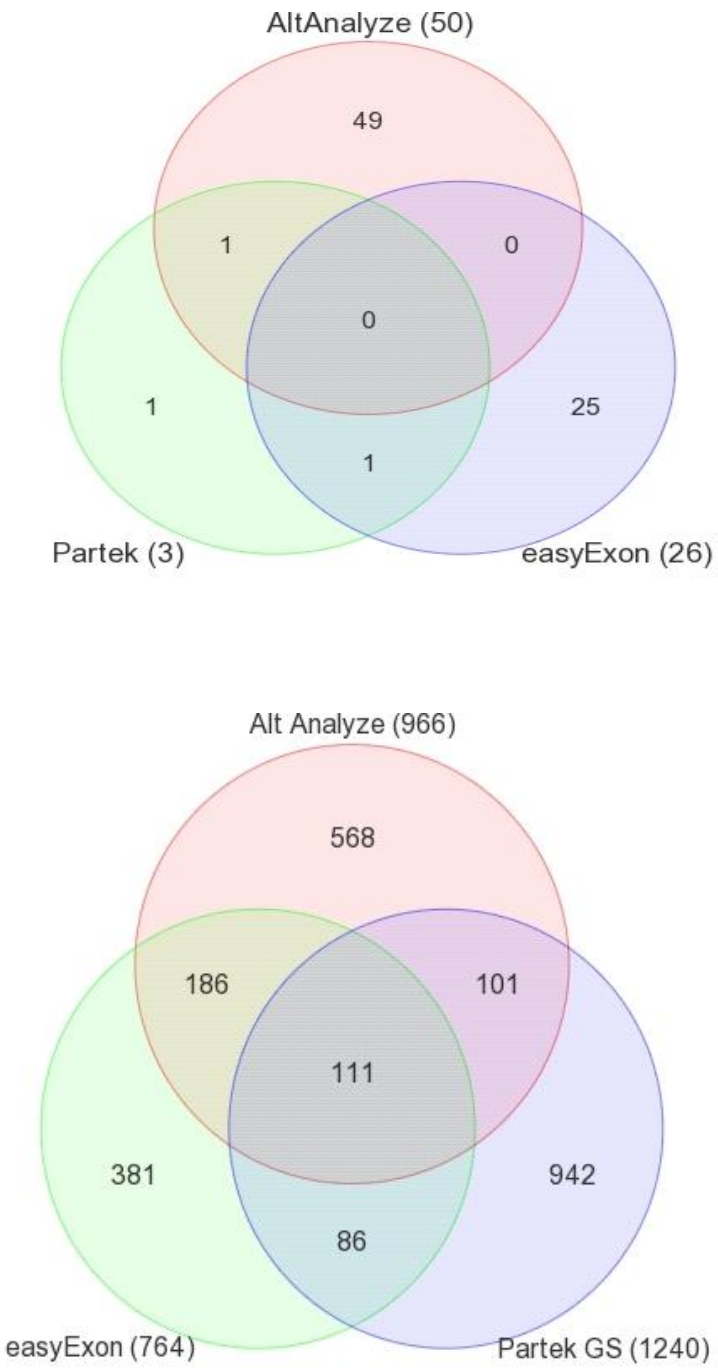


Figure 3.3.17. Male Differential Expression and Splicing Overlap. No genes were predicted to be significantly differentially expressed between male C59X mutants and wildtypes (top) when combining the results from the 3 statistical algorithms. 111 genes were predicted to be differentially spliced and common to the 3 statistical algorithms applied using Partek GS, AltAnalyze and easyExon.

For differential splicing 81 genes were found to be common to all 3 analyses. For male results no significantly differentially expressed genes were common to all 3 analyses (Fig. 3.3.17). The splicing results showed 111 genes to be significant in all 3. An overlap of the 81 and 111 genes revealed a splice overlap of 13 genes common to all statistical algorithms and programs and male and females (Table 3.17).

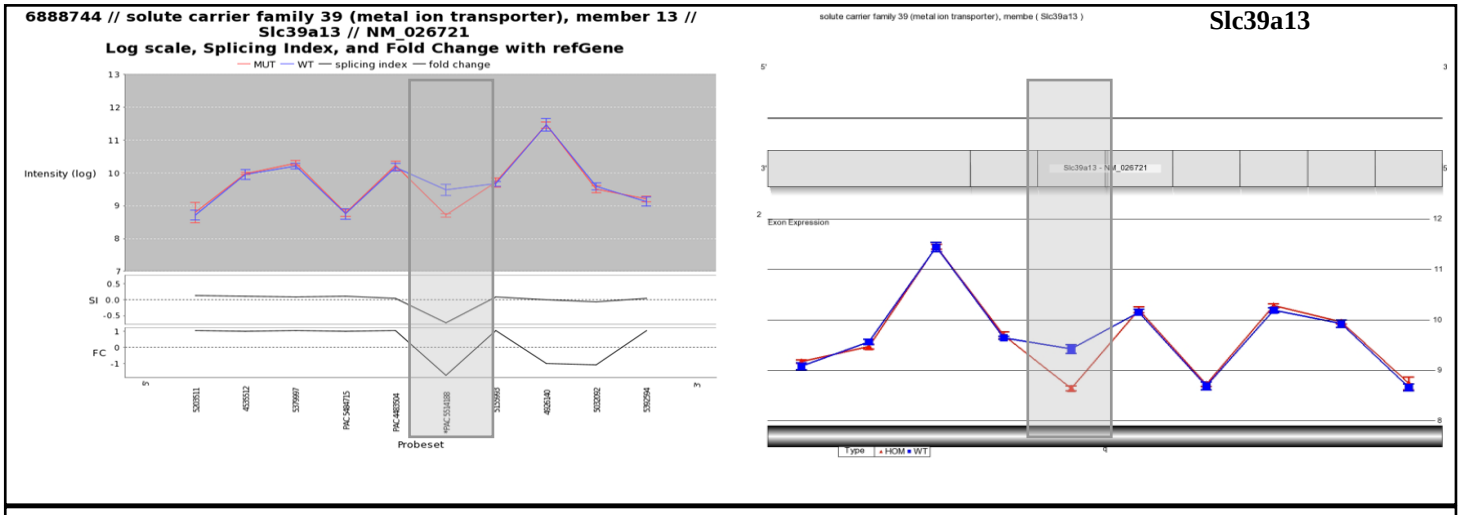
Of these 13, 9 were found to have at least one predicted differentially spliced probeset common to female and male experiments, whether analysed in Partek GS, easyExon or AltAnalyze. 9 were significant following multiple test correction with an FDR threshold of 0.05 and 5 were significant following a Bonferroni correction for the 15,800 tests in both experiments using Partek GS. Examples of these splice events are displayed in Figure 3.3.18. Adding to this the results from the combined analysis *Fam171b, Slc39a13*, and *Ssfa2* are strong differential splice candidates.

From the data generated in easyExon and AltAnalyze and the overlap with Partek GS, a manageable list of potential expression and splice candidates for validation could be generated. Prior to any validation using RT-PCR the results of the planned RNAseq data were awaited to further confirm the results and add to their validity.

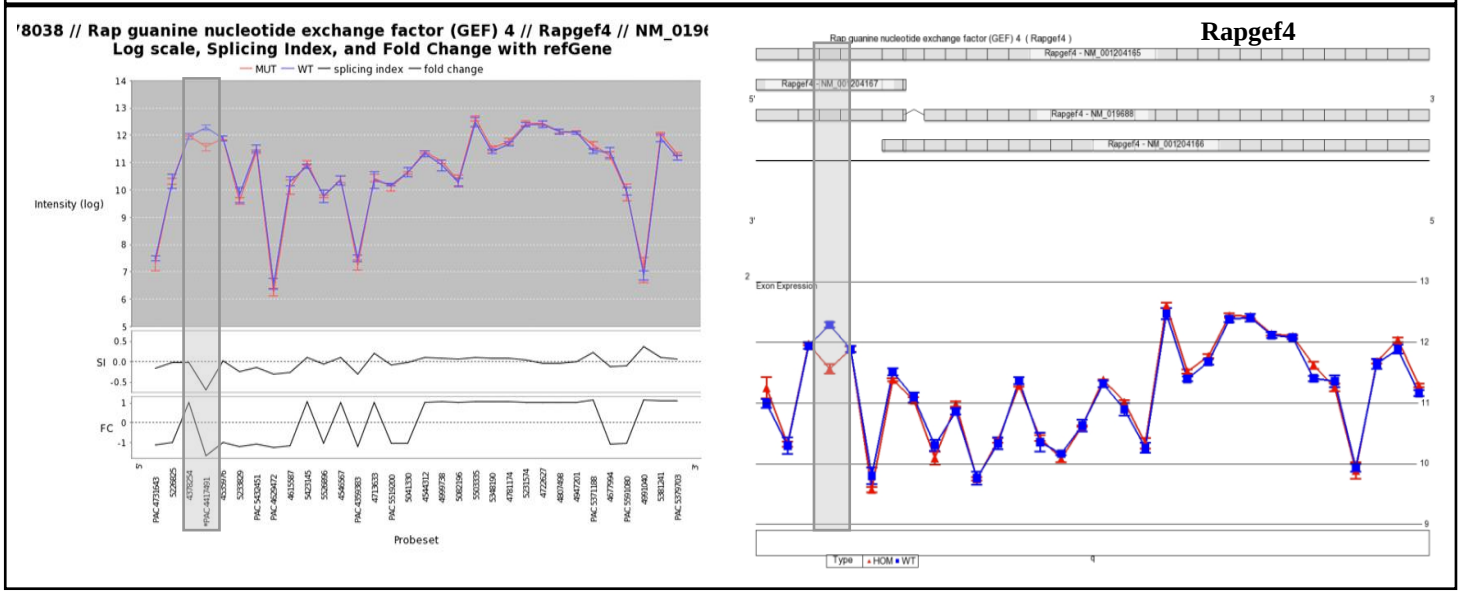
Transcript ID	Gene Symbol
6890648	*2010106G01Rik**
6919195	*Centb5/Acap3**
6858773	*Colec12*
6878657	*D4300389N05Rik/Fam171b***
6918763	*Dffa*
6888151	*Frzb**
6878031	*Itga6*
6878655	*Itgav**
6879054	*Lrp4*
6878038	*Rapgef4***
6825657	*Rhobtb2/Prdm4***
6888744	*Slc39a13***
6878469	*Ssfa2***

Table 3.17. Predicted Differentially Spliced Probesets. Manual inspection of the geneview (Partek GS) and indicated probesets in AltAnalyze and easyExon allowed the probesets predicted to be differentially spliced to be identified in the 13 genes significant in all analyses. In 9 (grey) of the 13 genes, at least one of the differentially spliced probesets was common to both female and male experiments using each of the statistical algorithms. * Significant following FDR threshold of 0.05 in Partek GS. ** Significant following Bonferroni correction of 15,808 and 15,813 tests in female and male analyses respectively.

Females



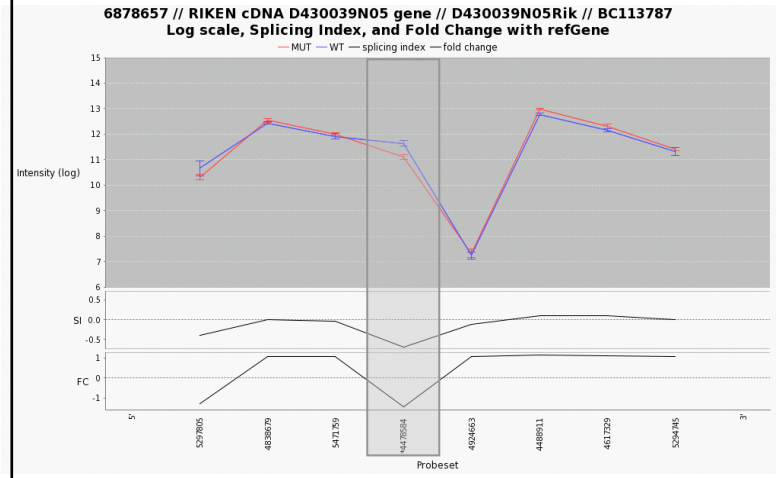
Slc39a13

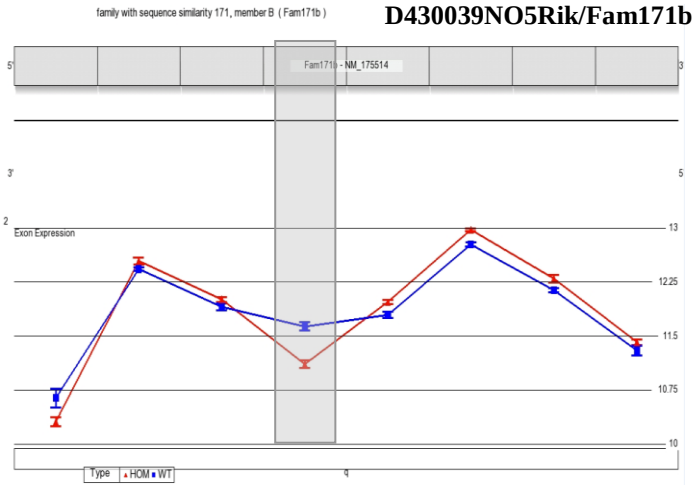


Rapgef4

Males

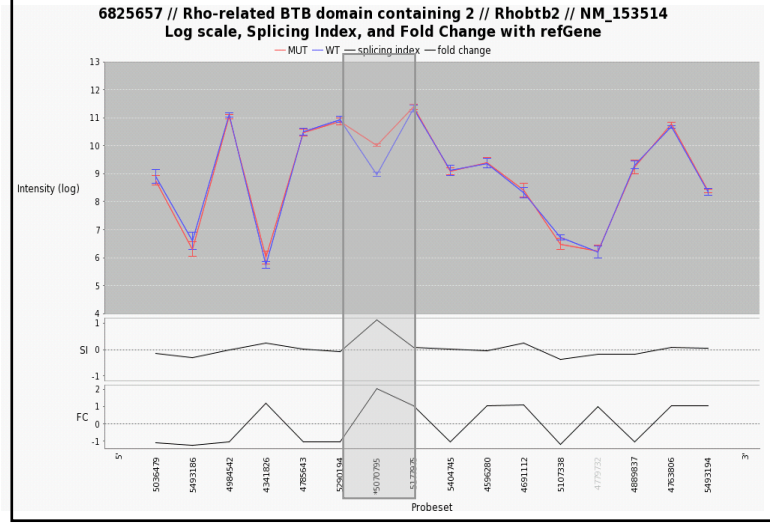
D430039NO5Rik/Fam171b

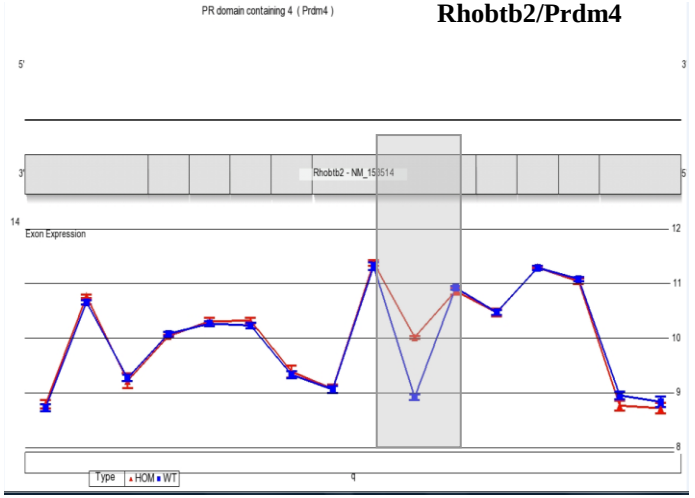




D430039NO5Rik/Fam171b

Rhobtb2/Prdm4





Rhobtb2/Prdm4

Figure 3.3.18. Differentially Spliced Transcripts Common to AltAnalyze, easyExon and Partek GS. Predicted splice events in *Slc39a13*(top) and *Rapgef4* (bottom) (Females) and Fam171b (top) and *Prdm4/Rhobtb2* (bottom) in males. The geneviews from easyExon (left) and Partek GS (right) are displayed. The probeset predicted to be differentially spliced between C59X mutants and wildtypes is highlighted by a grey box. In each instance the predicted probeset was identical in Partek and easyExon with the same probeset being identified as significantly differentially spliced by AltAnalyze (no geneview is produced in the output) for both female and male experiments. The easyExon trace shows the mean normalised, log transformed and variance-stabilised intensities for each probeset. The fold change values are plotted below the intensity plot and a significant MiDAS p value is indicated by (*) next to the corresponding probeset ID, all of which helped identify the splice event. The splice index diverges from 0 when the intensity signal ratio at a probeset is different from the average intensity signal ratio across the transcript, which in each of the examples above was consistent with the significant splice p value.

N.B: *Prdm4/Rhobtb2* is found on the antisense strand and so is displayed from right to left in easyExon. In Partek GS the Refseq is displayed in a 3' to 5' direction but from left to right, which is why the geneviews appear different at first inspection.

3.3.4.11 C59X Mutant and Wildtype expression of Zfp804a.

All results presented were carried out using the core meta-probeset, due to the increased annotation confidence in this set. One of the disadvantages of using only the core meta-probeset, specific to this study, was that Zfp804a was not targeted by any core probes. Instead the two Zfp804a transcripts (one protein coding, the other a processed transcript) were targeted by 11 extended probesets and 17 full probesets. As the primary objective of this study was to determine the effects of altered Zfp804a on gene expression and splicing, and as a qualitative difference in Zfp804a between C59X mutants and wildtype was not obvious from the RT-PCR (3.3.1), an analysis using the extended meta-probeset (which includes all core probes in addition to extended probes) was carried out to determine if any differences in Zfp804a expression could be observed between C59X mutants and wildtypes (Figure 3.3.19).

No significant differential expression was observed, as predicted from the RT-PCR (3.3.1), but a trend for upregulation in the mutants was observed in all 3 analyses. Using the FIRMA algorithm in AltAnalyze resulted in significant differential expression of Zfp804a in male and female C59X mutants (unadjusted FIRMA p value = 0.02) with upregulation of Zfp804a in the mutants. The results did not remain after applying the 1.5 fold change cut-off, as fold change was 1.26 and 1.24 in female and male

respectively. Following a t-test in easyExon to determine differential expression at Zfp804a the results showed that no significant difference was present between C59X mutants and wildtypes when using a fold change cut-off of 1.5. Without this cut-off significant differential expression was observed in female C59X mutants ($p = 0.008$) and males ($p = 0.01$) with upregualtion of Zfp804a observed in mutants as seen in Partek GS and AltAnalyze.

Significant differential splicing was observed in Zfp804a between C59X mutants and wildtypes (Females $p = 5.16 \times 10^{-9}$; Males $p = 3.07 \times 10^{-8}$; combined $p = 3.31 \times 10^{-16}$). From manual inspection of each geneview it appeared that the splicing signal was being generated from the central of three probesets targeting intron 1 of the protein coding transcript (Refseq), which corresponds to the alternative exon 1 in the Ensemble processed transcript (Figure 3.3.19). The processed transcript is predicted to have an alternative promoter and so the first exon overlaps with intron 1 of the RefSeq transcript.

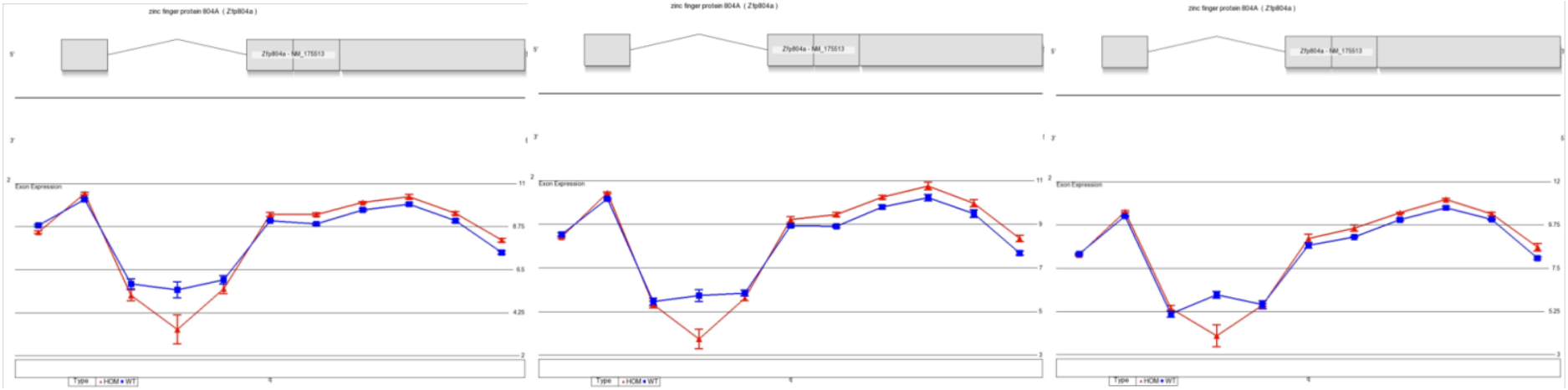
Three of the probesets are unique to the processed transcript, the central one of which is predicted to show differential expression between the wildtypes and C59X mutants. The possibility of differential expression of the two Zfp804a transcripts between C59X mutants and wildtypes is unlikely due to the absence of differential expression in the two flanking probesets. The predicted differentially spliced probeset and the adjacent downstream probeset are separated by only 3bp, the differences in expression at only one of these probesets is suggestive of a false positive prediction.

The observed pattern of expression could be explained by a novel variant, but this would be a small fragment of less than 82bp or a technical artefact such as probe cross-hybridisation to another part of the genome. The probeset sequence was checked for any potential cross-hybridisation using both BLAST and BLAT and was found to be a perfect match to Zfp804a only. The probeset does contain a dbSNP (rs28042740), if however this SNP was present in the mice, then it would be expected to affect hybridisation in both groups equally and so should not cause the differential expression observed. Next the sequence surrounding the probeset was checked for the presence of potential splice sites. The consensus 5'and 3'splice sites are GT and AG respectively. Within the 3 bases separating the spliced probeset and adjacent 3' probeset the 5' consensus splice site of GT is present in the sequence, upstream of the spliced probeset

also contains the conserved AG 3' splice site (Fig. 3.3.20). If an alternative cassette exon were present and incorporated more frequently in wildtypes, then it would be expected that a separate band with an additional ~80bp would be present when carrying out RT-PCR between exon one and three and this is not present in any of the 16 samples analysed. The splice event could also be indicative of an independent transcript, albeit a short transcript based on the lack of differential expression in the two flanking probesets. Primers specific to such a transcript could not be designed due to the very small distance between this probeset and the downstream probeset.

The extended analysis in easyExon gave a significant result for differential splicing in Zfp804a in both the female and male analyses (MiDAS p value = 0.04 in both) (Figure 3.3.21). The probeset predicted to be differentially spliced, however was different. In female C59X mutants the probeset predicted to be differentially spliced was in agreement with the Partek GS results. In the male analysis the probeset targeting the third exon of Zfp804a (4929502) was predicted to be differentially spliced in male C59X mutants.

In AltAnalyze (as explained in Chapter 2.8.4.2) Zfp804a was actually included in the core analysis. No differential splicing was found in female C59X mice, but two probesets were predicted to be differentially spliced in male C59X mutants. One of these probesets (4929502) replicates the finding from the easyExon analysis, again found specifically in male C59X mutants. The other differentially expressed probeset (5424882) is upstream of exon 1 and is predicted to be due to alternative promoter usage. As no splicing event is found consistently across the different statistical algorithms and as there appears to be inconsistencies between female and male C59X mutants this suggests the finding may be a false positive.



	Female	Male	Combined
Differential Expression p value	0.19	0.75	0.99
Differential Splice p value	5.16x10-9	3.07x10-8	3.31x10-16
Fold Change	1.05	-1.04	1.02

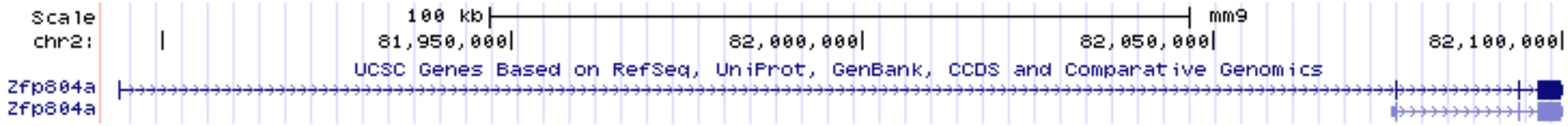


Figure 3.3.19 Zfp804a Expression and Splicing. No differential expression was observed in Zfp804a. Significant differential splicing was present in female, male and combined experiments and appeared to be due to the differential splicing observed in a single probeset targeting the first intron of the processed transcript. Both the Refseq, protein coding transcript and the Ensembl, processed transcript are displayed at the bottom of the figure.

agctgtagtgccatctctctctctctctctctctctctctctctctcatctctttagatgaagaacttaaataatttgcagaaaagtaacttag
ggtgcaaaaattcaatttttaaaatttaataaaatttatcatggaatcaattacaatgacaaaataaatgaagtttaaattaacaatata
tttatgtaaatatttatgaaaagaatgaaaatatccatttgcatatataattggcagtacatatctcttcttaaaggccacatttcagatgt
tcaccaaagaggattttatagtagctaagaattagctgtttacagcagaaccaaactacttttggaagctgtattttaatgttggcaaag
attggatctatgttcacttaatgcctctttctatatgatttactgtctttcttgtacttctgctgttttctttttttttcatttatattggcttaatctg
tgaattctttgctttttttgactcatgttttcaaacttgttttgctgatttttttatttgtctttgttctttttctagttctaattgttatat**taaagt**
tgacaacagttgggtcaaagtaaactcaagggagtgctcattgtatgtttcccttgaaattttatgtttcag

Figure 3.3.20. The sequence represents the alternative exon 1 from the Zfp804a processed transcript. 3 Affymetrix extended probesets are found in this sequence (shaded) with the middle one predicted to show differential expression between C59X mutants and wildtypes (bold). No differential expression is observed in the two flanking probesets suggestive of an alternative exon, specifically within this region. The consensus 5' splice site (GT) is observed downstream of the probeset in the 3 nucleotides residing between this probeset and the downstream probeset. A potential 3' splice site is also present (AG) upstream of the differentially expressed probeset. Variable and less conserved sequences surrounding the consensus 5' and 3' splice sites have been proposed, which are thought to be important for accurate recognition of splice sites by the spliceosome. The above sequence does resemble these sequences in part. These sequences are known to be particularly variable in alternative exons.

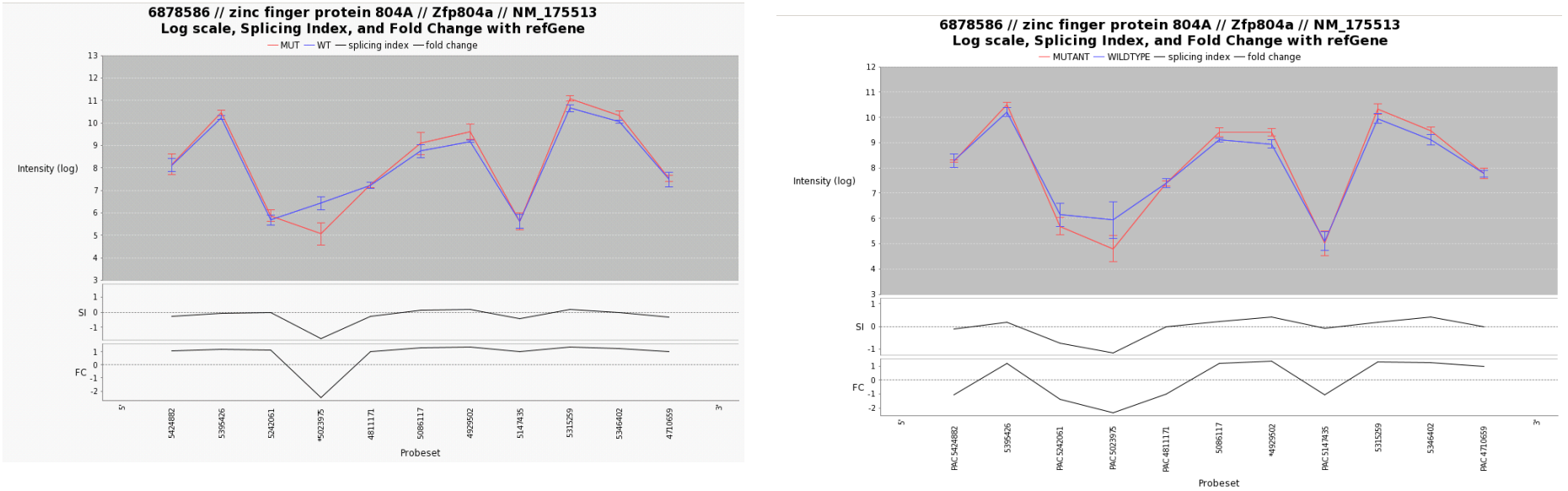


Figure 3.3.21. easyExon Extended Analysis C59X Mutants and Wildtypes. Whilst significant differential splicing between C59X mutants and wildtypes is observed in both females (Left, p = 0.04) and males (right, p = 0.04) the probeset predicted to be differentially spliced is different. In females the probeset is the same as that observed when carrying out the analysis in Partek GS. Differential splicing in male C59X mutants is observed at the 7th probeset, 4929502 which targets exon 3 of Zfp804a.

3.4 Discussion.

Sequencing confirmed the presence of a nonsense or PTC mutation in exon 2 of the Zfp804a transcript carried by the mutant mouse line. Qualitative assessment by RT-PCR and quantitative assessment using the exon array (Affymetrix) revealed no significant differential expression of Zfp804a between the C59X mutant and wildtype mice. This finding was unexpected based on the prediction that the PTC would initiate the nonsense mediated decay surveillance mechanism and thus reduce the abundance of the Zfp804a mRNA specifically in the mutants.

The C59X mutant transcript with the PTC is not a substrate for NMD, however this does not mean that protein levels are unaffected. The only way to determine the effects of the PTC on Zfp804a protein is to measure the protein itself using a suitable antibody, unfortunately such an antibody is not currently available. There are examples in which transcripts with PTCs in the first exon escape NMD, for example β-globin (Neu-Yilik et al, 2011). PTCs in exon 1 of β-globin are bypassed with re-initiation of translation occurring at an upstream, in-frame start codon (Neu-Yilik et al, 2011), although the N-terminally truncated protein is not functional (Neu-Yilik et al., 2011). I have not found any known examples where a PTC in exon 2 of a gene escapes NMD.

Read through can occur when a PTC is actually read as a normal codon by the tRNA, due to the fidelity of the 3rd base of the codon. The transcript escapes NMD and translation continues in the normal reading frame which generally results in full length protein being translated but at reduced levels.

tRNA has been described to temporarily detach from the mRNA being translated, particularly at stop codons, where the ribosome is thought to pause. This can result in tRNA reattachment at an in frame nucleotide to the +1 frame, a process known as frameshifting (Farabaugh et al., 1993). When the distance between detaching and re-attachment of the ribosome is greater and reattachment occurs at an out of frame codon this is called bypassing (Weiss et al., 1987). Bypassing and frameshifting are examples of ways in which a PTC containing mRNA transcript can escape the NMD mechanism but may result in an aberrant protein (Herr et al 2000; Neu-Yilik et al., 2011). If a shift in the reading frame occurs, this could result in a protein which is mis-folded which could result in the removal of the protein via the endoplasmic reticulem-associated protein degradation (ERAD) system (Sommer & Wolf, 1997). The system is responsible

for the detection of the incorrectly or incompletely folded proteins and their subsequent degradation (Hamptom, 2002).

Literature on the stop codon as a tetranucleotide suggests the 4th base is important to the fidelity of a stop codon with UAAG and UAAA being much more efficient at causing termination than UAAC and UAAU (McCaughan et al., 1995). The 4th base in C59X mutants is a G, therefore the ability of the stop codon to terminate translation would be expected to be high. In the absence of nonsense mediated decay, a truncated Zfp804a protein may be translated.

The absence of reduced Zfp804a mRNA abundance in the C59X mutants was unexpected, however I postulate that it is likely that the resulting translated protein will be truncated with a loss of function in homozygotes potentially accompanied by a gain of function if there is a truncated protein. In the absence of an antibody, I am unable to test these hypotheses.

A detailed look at the expression across the Zfp804a transcript in both C59X mutants and wildtype revealed if anything a trend for upregulation of the protein coding transcript in mutants. Significant differential splicing of *Zfp804a* mRNA between C59X mutants and wildtypes was also observed which replicated across the female and male experiments, and the significance of which increased when combining all samples into one group. Differential splicing between C59X mutants and wildtypes appeared to be occurring in the processed transcript (Ensembl) rather than the Refseq protein coding transcript at a single probeset, suggestive of differential cassette exon usage. The idea that a novel smaller transcript may be present was explored but validation was impeded due to the proximity of the predicted spliced probeset from the adjacent probeset which showed no differential expression between C59X mutants and wildtypes. When considering splicing of Zfp804a in other software packages this splice event failed to replicate across algorithms and experiments. This event may therefore reflect a technical artefact.

The number of genes differentially expressed between the C59X mutants and wildtypes were small. In the female and male experiments no genes were significantly differentially expressed following multiple test corrections. Whilst the role of Zfp804a

in transcription regulation can not be dismissed, these data do not support the hypothesis that Zfp804a regulates gene expression. When combining male and female data together to increase power, several genes were significantly differentially expressed following multiple test correction, but when inspecting the geneviews the transcripts did not show differential expression across the entire transcript.

When the data were processed using alternative filtering and statistical methods, differential expression was seen for only 6 genes across all methods. These 6 genes are relatively robust candidates for altered expression as they were identified in the female experiment in each of the three statistical packages. The geneviews for all 6 generated in Partek GS are also indicative of true expression differences across the whole of the gene. Each gene displayed expression differences between C59X mutant and wildtypes of more than or equal to 1.5 fold. Despite the robust nature of these findings, identifying only 6 genes from ~15,000 does not provide conclusive evidence that Zfp804a regulates gene expression. Although these 6 changes are less likely to be technical errors, the statistical support is not strong enough to exclude these being chance positives. It is important not to dismiss these genes though or rule out a role for Zfp804a in gene expression regulation, particularly in light of the results of previous studies in which evidence for such regulation has been found in cellular models (Hill et al., 2012; Gigenti et al., 2012).

The results of the alternative splicing analysis are somewhat more robust with ~7% of transcripts tested showing differential splicing between C59X mutants and wildtypes, many remaining significant following multiple test correction. When assessing the replication across the female and male experiments, significantly more genes than expected by chance replicated. Detailed inspection of the specific splice events, which is important as splicing results are known to be subject to more false positives (Bemmo et al., 2008), showed that in a high proportion of the transcripts the differentially spliced probeset(s) was the same in female and male experiments. Altered splicing between the C59X mutants and wildtype also occurred in the same direction across experiments.

Combining the male and female data in addition to carrying out the independent analyses with alternative filtering criteria and algorithms produced a list of 9 strong differential splice candidates. The lists used for the overlap had not been corrected for multiple testing and so differences in the lists may reflect the different false positive calls arising from the different annotation, filtering and statistical procedures used. To

allow for the fact that transcripts are included/excluded in each method according to different criteria, I included only transcripts that were included under each approach. The genes remaining are less likely to result from artefacts of a particular type of analysis. The overlap also produced a more manageable list size from which manual inspection of the data was carried out.

Whilst the use of female brain tissue and $F3_i$ generation mice was not ideal this was all that was initially available. A major limitation of using mice from this generation is the potential for mice to harbour a number of additional ENU mutations, although at the outset, it was thought unlikely that the same additional ENU mutation would be found in enough mice to impact upon the experiment. Nevertheless, clearly it was important to carry out the study on mice from a later generation in which the likelihood of such additional mutations is reduced further (This analysis is presented in Chapter 5).

Despite gender differences, there is a robust overlap of splicing changes replicated across the male and female mouse studies. In addition when combining the samples in a larger analysis and covarying for gender, splice differences are still apparent between C59X mutants and wildtypes. These are apparent despite the use of half and whole brain in female and male mice respectively, and the fact that the male animals were twice the age of the female mice, adding to the robustness of the data. In mice of different genders and age groups common expression and splice variation is found between C59X mutants and wildtypes.

Of the differential splice candidates, Rap guanine nucleotide exchange factor (*RapGef4*) is highly enriched in the brain. Rare non-synonymous variants in this gene have been associated with autism (Bacchelli et al., 2003). Two further interesting genes are the integrins Itga6 and Itgav which form part of a family of cell adhesion proteins. Itga6 is expressed in neuronal tissue and there is evidence of its role in development processes (De Arcangelis et al., 1999). Itgav is expressed in the striatum and all cortical layers (Pinkstaff et al., 1999) and is thought to have a role in neuronal migration during cortical development (Anton et al., 1999).

To determine if these genes, and the others identified as altered in the exon array, could be validated in an independent assay, exploratory expression and splicing analysis was carried out using next generation RNA sequencing (RNAseq).

Chapter 4. RNA-Sequencing

4.1 Introduction.

Having looked at expression and splice changes in the C59X mouse strain using the exon array (Affymetrix) I found that the $F3_i$ generation mice homozygous for the mutation exhibited altered expression and splicing relative to wildtype controls. Replication across 2 independent experiments and 3 different software programs highlighted a number of consistent changes. To support the exon array work I proceeded by monitoring expression and splicing in the same mice using next generation RNA sequencing (RNAseq).

The implementation of next-generation sequencing has enabled global RNA sequencing in a relatively unbiased and quantitative way. I sought to apply this technology to confirm the expression and splicing changes I had observed in the array data. Unlike microarray analysis in which gene expression is determined indirectly from the degree of hybridisation to probes, in RNAseq a direct measurement can be taken based on the number of sequence reads at specific transcripts. This is therefore an unbiased approach, with no restriction to the transcripts represented on the chip one happens to be using. Given the unbiased nature of RNAseq, novel transcripts can be identified, and the application of RNAseq has led to the realisation that the transcriptome is far from complete, novel transcripts being discovered in each successive study (Trapnell et al 2012).

Given the expense of RNAseq, the initial plan was to carry out a pilot study on a subset of the samples from $F3_i$ mice used in the exon array analysis, and, if technically successful, followed up with a larger independent, later generation sample. The intention was that the initial pilot would allow the validity of the exon array findings to be determined whilst acting as a technical validation of the exon array procedure and analysis. The larger follow-up study would then allow a biological validation of any consistent expression or splice changes in the C59X mutants.

4.2 Methods.

4.2.1 Sample.

Brain RNA samples from 4 of the males used in the previous chapter were chosen for the pilot based on quality and volume of RNA. 2 C59X mutants (ZBDEZ2e & ZBDEZ4a) and 2 wildtypes (ZBDEZ6a & ZBDEZ21c) were chosen. All samples had an RNA integrity number (RIN) above 8.

4.2.2. Sample Preparation.

RNA was prepared using the low-throughput (LT) protocol with the Illumina TruSeq® RNA sample prep kit (Illumina, San Diego, CA, USA). Samples were initially re-quantified using 1µl of sample on the Nanodrop 1000 spectrophotometer and then prepared as per the manufacturer's instructions by myself and K. Matripriganda (Lab manager responsible for setting up new technologies in the host lab) with assistance from laboratory technicians. In addition to the 4 samples a Universal Human reference RNA (Stratagene) sample was used as a positive control and a no template sample as a negative control. 4µg RNA in 50µl H_2O was purified using oligo dT magnetic beads to capture poly-A RNA which was then fragmented. Based on the published protocol, to ensure optimal coverage of the transcriptome in conjunction with efficient library production, fragments were expected to range from 120-200bp with an average size of 150bp. Double stranded cDNA synthesis was achieved using reverse transcriptase and random hexamer primers. Blunt ends were formed by removing overhangs with a proprietary mix either of exonuclease (degrade 3' overhangs) or polymerase (to fill in 5' overhangs). Prior to ligating indexing adapters to the fragments, a single A nucleotide was added to the 3' blunt end of the fragment which is complementary to the T nucleotide at the 3' end of the adapter and so aids the ligation of fragment and adapter. This is done using an A-tailing mix. PCR was used to amplify the DNA library and only fragments with adapters at both ends were amplified. Each step was carried out in accordance with the manufacturer's instructions.

4.2.3. cDNA Library Quality Control.

To determine the fragment sizes within, and quality of each cDNA library, 1µl of sample was loaded into a DNA chip and run on the Agilent 2100 Bioanalyser (Agilent

Technologies, Palo Alto, CA). Distribution must be within the expected range for enrichment to be successful. Fragment size and purity were assessed with a product size of 200-300bp expected for paired-end libraries, according to the manufacturer's instructions.

4.2.4 Sequencing.

Samples were sent to the core sequencing facility at Bristol University. Paired end (~100bp) reads were generated on an Illumina Genome Analyzer IIx. Raw sequence in the form of fastq files were acquired from the Bristol facility for analysis.

4.2.5 RNAseq Quality Control.

Raw files (FASTQ) files were imported into the FastQC program (v.0.10.1) (Babraham Bioinformatics) (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/) to determine the quality of the sequencing. Basic statistics as well as 10 QC statistics are generated. For each metric FastQC will mark the sample as either acceptable in quality, warn that sample may be verging on poor quality or state that the sequence has an error or has failed the metric. If a sample fails a certain metric this simply means the results differ from that expected in FastQC and should not be taken that the sample is poor quality and should be discarded. FastQC is used to determine areas in the data which may have quality issues and allows you to determine if the source is one that was expected given the type of experiment or if a genuine quality issue has arisen. Errors made during cluster generation and sequencing are easier to detect than errors made during library preparation. This is because the errors made in the former two affect the detected signal and therefore the quality score. Basic statistics such as the total number of sequences processed, sequence length and the number of sequences filtered are produced for each input sequence. An explanation of the 10 QC metrics is described in detail alongside the presentation of the results (4.3.2.1-4.3.2.10) (Further information can be found at: http://www.bioinformatics.babraham.ac.uk/projects/fastqc/Help/3%20Analysis%20Modules/).

4.2.6 Alignment, Assembly and Differential Expression Analysis.

To align reads to the genome, assemble transcripts and quantify expression changes the Tuxedo protocol was followed (Trapnell et al., 2009; Trapnell et al., 2010). This protocol is in wide use (Lin et al., 2011, Trapnell et al., 2012). It is particularly compatible with sequencing carried out using the Illumina sequencing platform and works best with sequencing generated on model organisms for which a reference sequence is available. The following procedure was carried out using Unix.

4.2.6.1. Sequence Alignment and Identification of Splice Junctions

Raw sample files were aligned to the genome (*mus musculus* NCBI build 37.1) and splice sites identified using TopHat (v.1.3.2) (Trapnell et al., 2009) (http://tophat.cbcb.umd.edu/), which is part of the software suite collectively known as the tuxedo protocol/suite. TopHat runs on the Linux operating system and was developed specifically for reads generated using the Illumina genome analyser. Reads that are 75bp or greater are optimal for analyses. TopHat utilises Bowtie (http://bowtie-bio.sourceforge.net/index.shtml) (Langmead et al., 2009) for initial sequence alignment to the genome. Bowtie alone is not sufficient as the software cannot align reads to reference sequences which differ as the result of large mismatches such as might occur if the read spans an intron. TopHat takes initially unaligned reads and separates them into shorter reads called segments which are then re-aligned to the reference genome. Tophat then infers that segments within a read that map more than 100bp apart most likely span a splice junction and in this way a list of splice sites is generated. As no reference splice site annotations are used in this process, novel splice sites can be identified (Trapnell et al., 2012). Fastq files were used as the input for TopHat. The script used for alignment, alongside a description of the components can be found in Appendix 4.1

The output from TopHat is a BAM file (binary version of sequence alignment map (SAM)). Alignments were viewed by indexing the BAM file for each sample using SAMtools (http://samtools.sourceforge.net/) (For script see Appendix 4.2) and then uploading them into the Integrative Genomics Viewer (IGV) (http://www.broadinstitute.org/igv/) (Robinson et al 2011; Thorvaldsdottir et al., 2012)

4.2.6.2. Transcript Assembly.

Cufflinks (v.1.2.1) (Trapnell et al., 2010) (http://cufflinks.cbcb.umd.edu/) was then used to assemble transcripts from splice site information from TopHat. Cufflinks also forms part of the tuxedo protocol. The aligned BAM file is used as input into Cufflinks. Due to many genes having multiple isoforms, the assignment of reads to a particular isoform rather than another is not straightforward. Cufflinks assembles the data in the most parsimonious way possible to explain the data, assembling the minimum number of transcript fragments which can explain the splice sites identified. Removal of pre-mRNA transcripts and artefact transcripts is also carried out at this point. Each sample is aligned individually. When sequencing is of insufficient depth (less than 10 million reads) the amount of partial transcript fragments is greater which can increase the risk of false isoform calls. To overcome this, identified novel isoforms and transcripts of high interest should be validated using an alternative method such as RT-PCR, or Rapid Amplification of cDNA Ends (RACE) can also be used to ensure transcripts ends are more accurately defined as recommended by Trapnell et al. (2012) (Scripts used for transcript assembly can be viewed in Appendix 4.3).

A normalisation procedure is required in RNAseq, as longer transcripts produce more fragments compared to shorter transcripts. If two transcripts had equal abundance but one was twice as long as the other, the longer transcript would appear to have twice the (reads) abundance of the shorter transcript. Therefore the length of a transcript must be taken into account in abundance calculations. In Cufflinks this is done by considering the number of reads per transcript and then normalising this value by the length of the transcript. In addition the reads need to be normalised to the total number of sequenced fragments from the sequencing machine to account for run to run variability (Trapnell et al., 2012).

The normalisation method used in Cufflinks for paired-end reads is FPKM (fragments per kilobase of transcripts per million mapped fragments) (Mortazavi et al., 2008). FPKM is comparable to the RPKM (reads per kilobase of transcripts per million mapped reads) used for single-end sequences. The FPKM is proportional to the abundance of a transcript (Trapnell et al 2010).

When the data contain a small proportion of genes that are highly expressed, FPKM is subject to a known bias that can skew the differential expression analysis and results. This is dealt with by using the –N option (described in Appendix 4.3) which normalises

to the upper quartile of expressed genes rather than the total number of mapped fragments as the former is more robust when there are less abundant genes or transcripts (http://cufflinks.cbcb.umd.edu/).

The assemblies of each sample were then merged together using CuffMerge (part of the Cufflinks software package). This helps to compensate for relatively poor sequencing depth. Correct reconstruction of a gene in a single sample is difficult when there is insufficient sequencing depth. By merging all samples together the gene is more likely to be reconstructed accurately. The reference (NCBI build 37.1) was also merged with the samples forming a comprehensive annotation (Trapnell et al., 2012). To merge the assemblies I used gedit (a text editor) to create a file called 'assemblies.txt' within which the assembly files for each sample were listed and then CuffMerge (Appendix 4.4).

4.2.6.3. Differential Expression Analysis.

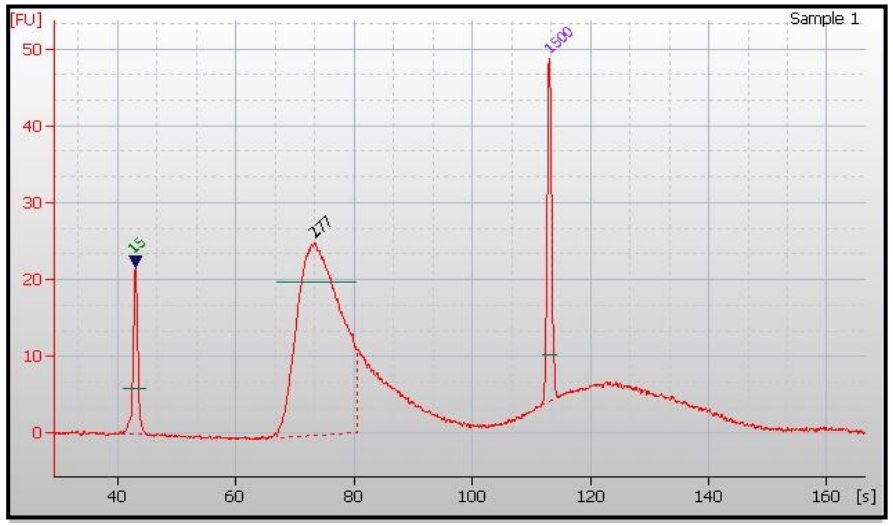
To compare the relative amounts of assembled transcripts represented in each sample and determine if the difference was statistically significant Cuffdiff (http://cufflinks.cbcb.umd.edu/) was used (Appendix 4.5). Samples were compared according to experimental group, therefore samples 1 and 2 representing the 2 C59X mutants were compared to samples 3 and 4 representing the 2 wildtypes. The expression of a transcript is calculated based on the number of reads as described above. To increase the accuracy of expression measurements, Cuffdiff models the technical variation in the data, which can arise due to artefacts in the library preparation or sequencing procedure and this is then adjusted for. Cuffdiff tries to estimate this variation using a likelihood based approach (Roberts et al., 2011) and uses this information when determining expression differences between experimental groups (Trapnell et al., 2012). The relative abundances of each transcript are compared using the Cuffdiff command. If a transcript or gene has several isoforms it can be problematic to determine which isoform a read is from. A linear model is used in Cuffdiff to work out, using maximum likelihood, which way of assigning abundance to each transcript best explains the reads generated (Trapnell et al., 2012). Total gene expression equals the combined expression of the relevant isoforms. Following this a results file is produced containing results of differential expression and differential splicing analyses.

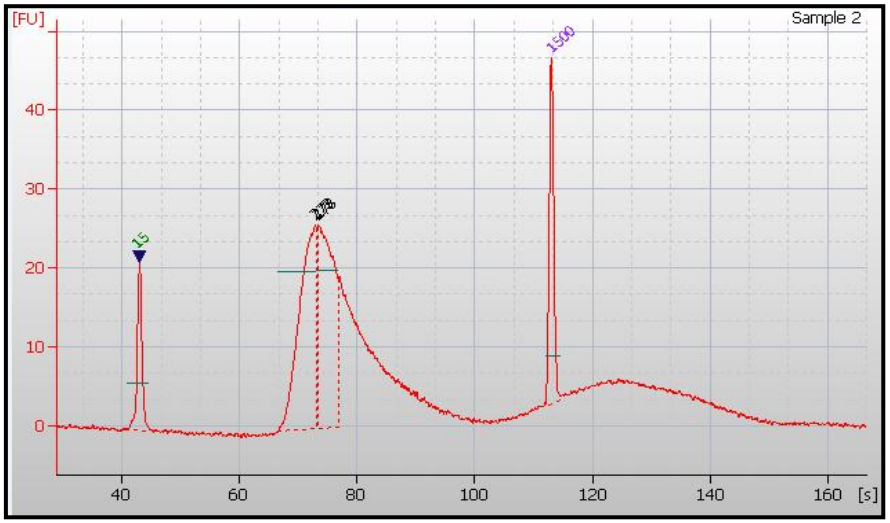
4.3 Results.

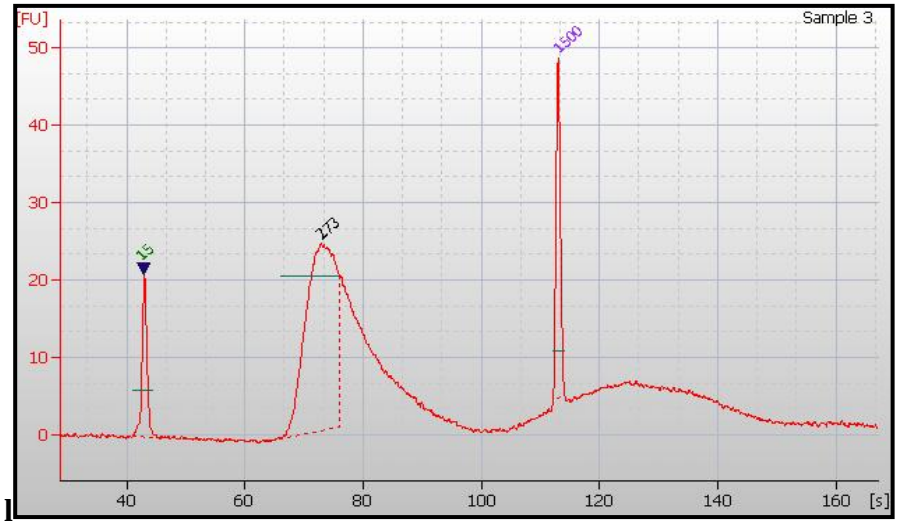
4.3.1 cDNA Library Quality Control.

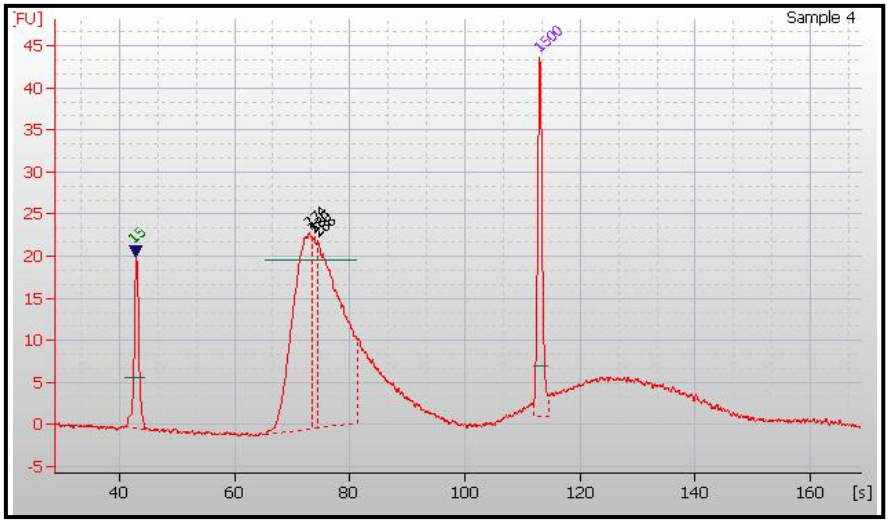
To ensure the cDNA library preparation was comprised of uniform sized fragments each sample was run on the Agilent Bioanalyser (Fig. 4.1).

Figure 4.1. cDNA library Quality Assessment. 1µl of each of the 4 samples was loaded into a chip and run on the Agilent 2100 Bioanalyser. In each sample the majority of fragments were around ~270bp, which is within the expected size range. The second peak at ~1500bp is likely to represent concatenated adapter sequences. As this represents only a small proportion of the sample this is considered adequate quality (Illumina) and I was able to proceed to the cluster generation step with these samples.


[FU]
Sample 1
15
277
1500
[s]


[FU]
Sample 2
15
1500
[s]


[FU]
Sample 3
15
273
1500
[s]


[FU]
Sample 4
15
1500
[s]

4.3.2 Sequence Quality Control.

From the bioanalyser trace the fragments in each sample appeared to have a similar distribution. Therefore the cDNA library preparation stage passed this quality control measure. The sequencing data were assessed for quality using FastQC software. Initially basic statistics were generated which showed all samples to contain mate-paired end reads of 110bp sequences. The total number of sequences was 41M, 61M, 61M and 54M for samples 1-4 respectively.

4.3.2.1 Assessing 'Per base sequence quality'

The distribution of quality scores at each position in the read is plotted. Each base in a read is assigned a quality score using a Phred-Like algorithm where $Q_{phred} = -10\log_{10}(p)$, where p represents the estimated probability that a base call is incorrect (Ewing et al., 1998; Ewing & Green, 1998). For example a quality score of 30 would represent a 1 in 1000 chance that the base had been called incorrectly (Ewing & Green, 1998). At each position a box and whisker plot is drawn (Fig. 4.2). A higher quality score represents a more accurate base call. A warning is issued if the lower quartile of any base is less than 10 or if the median is less than 25 at any base position. Sequencing is deemed to have failed quality control if the lower quartile of any base is less than 5 or if the median is less than 20 at any base position (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/). The plot in Figure 4.2 was generated from the sequencing of sample 1 (a C59X mutant) which had the lowest per base quality score. The median quality in each sample went below a median quality score of 20 in at least one base toward the end of the read. The low quality scores in this metric for each of the 4 samples indicate poor quality sequencing. The very low quality scores at the end of the sequences most likely reflect the decrease in quality calls on most sequencing platforms further into the sequencing process.

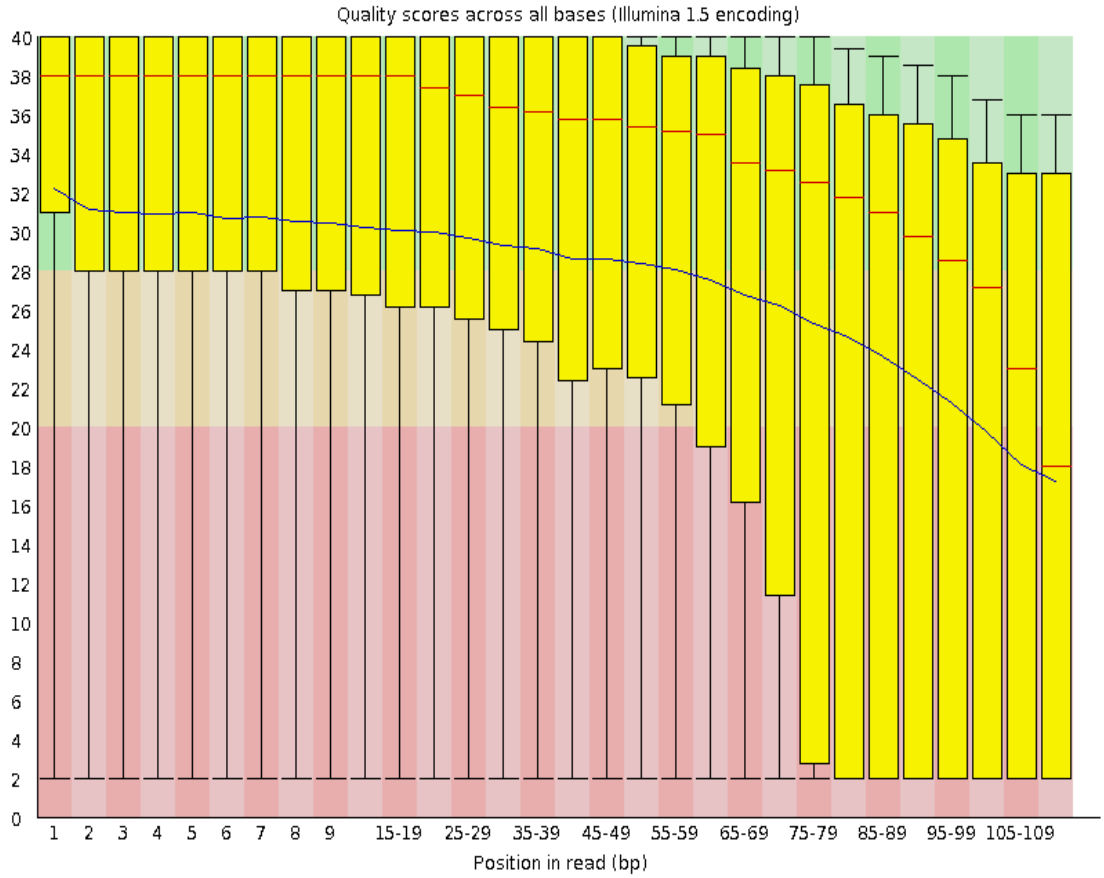


Figure 4.2. Per Base Sequence Quality. The position of the base within the read is plotted on the x-axis with quality (phred) score on the y-axis. A higher quality score represents a more accurate base call. Three zones are present on the y axis, depicted by a green, orange and red background colour, which depict good, reasonable and poor quality base calls respectively (as defined in section 4.3.2.1). The yellow boxes represent the inter quartile range (IQR), with the upper and lower whiskers representing the 10 and 90% points. The red line represents the median and the blue line the mean quality. The above is an example showing sample 1, a C59X mutant. All the samples failed this metric due to the median quality being less than 20 at, at least one base position.

4.3.2.2 Per sequence quality Scores. This metric is generated to determine if a subset of sequences from each sample have universally low quality scores. Low quality scores can arise due to poor imaging based on position on the flow cell. Sequences with low quality scores should only represent a small percentage of the total number of sequences. FastQC denote sequences with modal mean quality scores below 27 (0.2% error rate) as quality of that sample is poor (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/). Samples 2 (Fig. 4.3) and 4 had modal mean quality scores of greater than 27 (Fig. 4.3), with the quality score being

38 in both. The modal mean quality score in samples 1 and 3 was only 2, in one of the pair-end reads. Both samples were considered to have poor quality based on this quality score. As more than a small proportion of the data had low mean low quality score this could reflect a systematic error where perhaps one end of the flow cell was incorrectly read, however this is merely speculative.

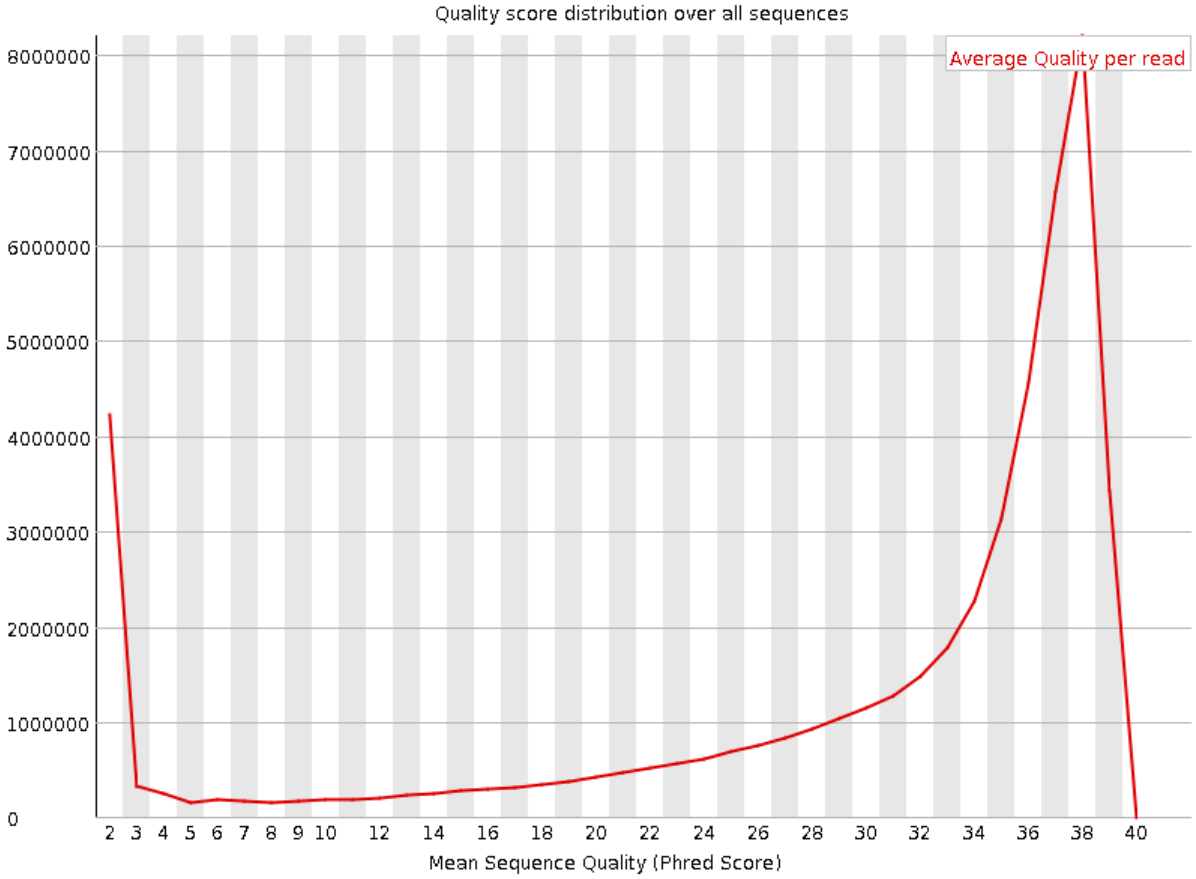


Figure 4.3. Per sequence quality Scores. Along the x-axis the mean phred quality score was plotted with frequency on the y-axis. In this example of sample 2 (C59X mutant) the modal mean quality score is 38 which is above the FastQC cut-off of 27 and represents good sequence quality scores. The majority of sequences have good quality scores and only a small subset have low quality scores, which most likely reflects their position on the edge of the field of view when imaging the sequences.

4.3.2.3 Per base sequence content.

This metric determines the proportion of G, T, A and C bases as a function of the position along the read throughout the sequence run for a given sample (Fig. 4.4). Across the read you would expect the line representing %G, T, A and C to be constant.

This demonstrates the relative amount of each base and should reflect the overall amount of each base in the transcriptome. If the cDNA library generated was random then the lines would be expected to run parallel as an equal proportion of each base would be expected. A bias in one particular base is indicated by a divergence form the parallel lines. If this divergence occurs in different bases according to the position within the read this may reflect a contaminating overrepresented sequence. If a divergence from parallel lines (and therefore unequal representation of each of the four bases) is seen across all positions of the read it indicates that the library was biased for particular sequences or a systematic error may have occurred during sequencing, such as difficulty in sequencing AT rich repetitive sequences (Harismendy et al., 2009). A difference of more than 10% between A and T or C and G at any position in the read would indicate questionable quality, likewise a difference of more than 20% between A and T or C and G in any position is representative of an unequal proportion of each of the 4 bases throughout a read which is indicative of low quality sequencing (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/). This however was not the case in any of the 4 samples. Each had a difference between A and T or C and G that was >20% at one or more position in the sequence. In each case this occurred in the first ~10bp of the read. As this bias is not present across the whole sequence it is unlikely due to a systematic error. That fact it occurs in the first ~10 bases is most likely accounted for by the random primers used in the cDNA library preparation, which bias the nucleotide composition of a sequence specifically at the start of a read (Hansen et al., 2010).

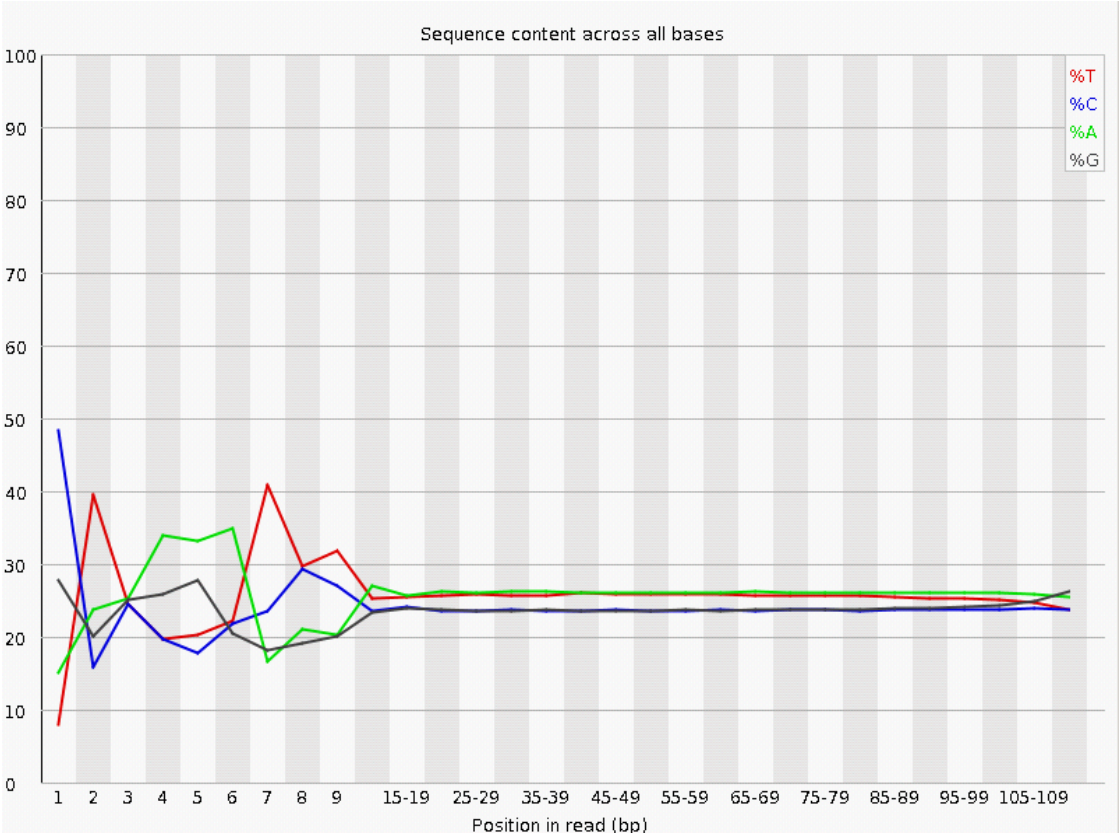


Figure 4.4 Per base Sequence Content. The proportion of G, T, A and C bases as a function of the position along the read throughout the sequence run is plotted above for sample 3 (this plot is representative of the plot observed for all 4 samples). The difference between A and T or C and G that was greater than 20% at one or more positions in the read occurred in all 4 samples which may indicate poor quality sequencing as defined in (4.3.2.3), however this was only observed in the first ~10bp of the read in all 4 of the samples.

4.3.2.4 Per base GC content.

This metric determines the GC content as a function of the position along the read throughout the sequence run for a given sample (Fig. 4. 5). Variable GC content across the positions in a read may indicate a strongly overrepresented sequence that is contaminating the library. A greater than 10% fluctuation from the mean GC content at any base position indicates that the sequencing is of poor quality (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/). The GC content at one or more base positions varied from the mean by more than 10% in each sample which therefore failed this metric. This always occurred in the first 15 bases of the read and again most likely reflects the random primers used in library preparation (Hansen et al., 2010)

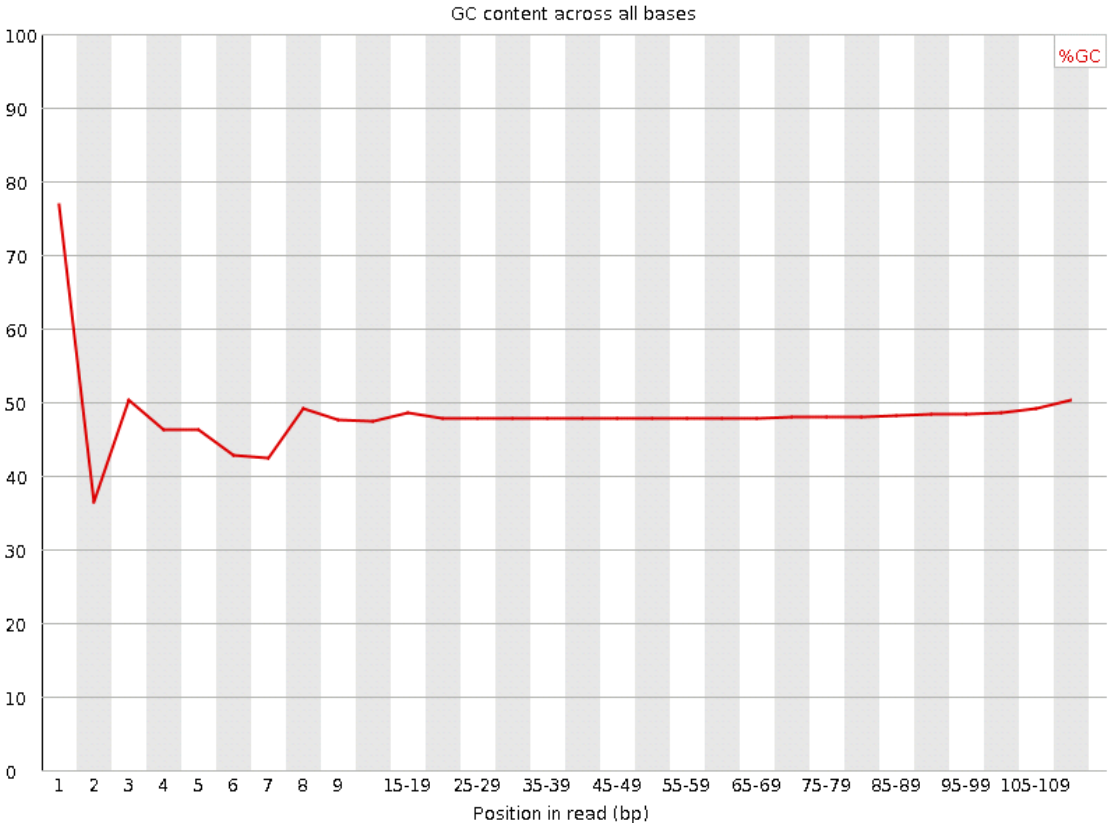


Figure 4.5 Per base GC content The GC content as a function of the position along the read, throughout the sequence run for sample 4 was plotted. A horizontal line is expected as the GC content should not vary when using a random primed RNAseq library. The GC content at one or more positions in the read varied from the mean by more than 10%. This was the case in samples 1, 2 and 3 as well. The variation in GC content from the mean was observed only at positions 1-15 of the read.

4.3.2.5. Per sequence GC content

The distribution of GC content across all sequences for a given sample was compared to a reference distribution predicted from the modal GC content in the actual data. A normal distribution is expected. The central peak should reflect the GC content of the underlying transcriptome. Contamination, e.g., from an adapter sequence, can cause an unusual distribution. If more than 15% of the reads deviate from the normal distribution this could be indicative of poor quality sequencing and should be looked into further. Distribution of GC content did not deviate from a normal distribution in samples 1 and 3. GC content deviated by more than 15% in samples 2 and 4 (Fig. 4.6) and may indicate contamination from an overrepresented sequence such as an adapter. FastQC recommends looking into fluctuations occurring in more that 15% of reads, but that

fluctuations in over 30% of reads is a real cause for concern and neither sample 2 or 4 passed this threshold (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/).

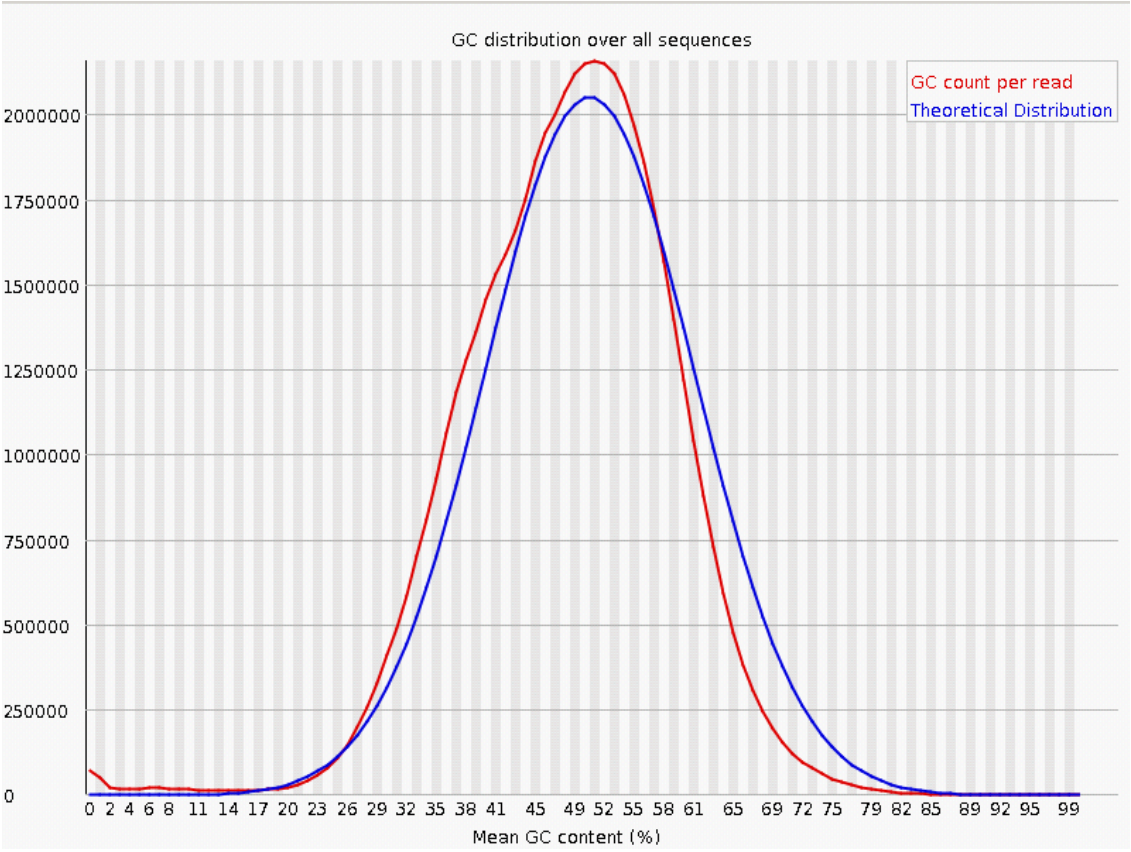


Figure 4.6 The distribution of GC content across all sequences for sample 4 was plotted. More than 15% of the reads deviated from the normal distribution which could be indicative of poor quality sequencing and should be looked into further. The central peak should reflect the GC content of the underlying transcriptome.

4.3.2.6 Per base N content.

When there is not enough confidence to make a base call, a base is designated 'N' rather than one of ACTG. The 'per base N content' metric is used to determine the percentage of base calls that were called as N's over each position of a read. A small proportion of N calls are expected, particularly toward the ends of reads but if greater than 5% of base calls were 'N', this indicates that there was insufficient evidence to make a sequencing call at that position most likely due to poor quality sequencing and this is even more likely if more that 20% of the calls at a particular position are given as 'N'(http://www.bioinformatics.babraham.ac.uk/projects/fastqc/). The results are

displayed for sample 1 (Fig. 4.7) and reflect that at no position along the read did the % N call exceed 5%. This was true of the results for samples 2, 3, and 4 as well.

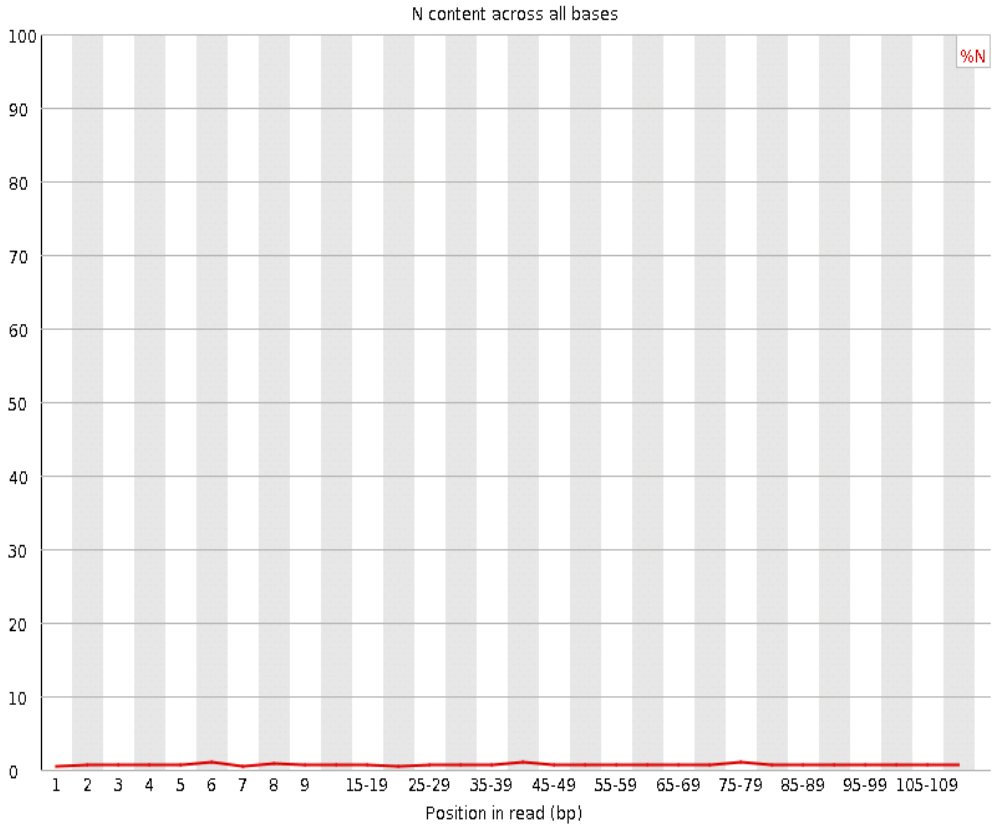


Figure 4.7. The percentage of 'N' base calls over each position of a read for sequences from Sample 1. The position in the read was plotted along the x-axis and the % of N's on the y-axis, to determine the percentage of N base calls at each position of a read. In this and the other 3 samples the % of N base calls at any of the positions of the reads did not exceed 5% which reflects adequate quality.

4.3.2.7 Sequence length distribution. Some sequencing machines will remove poor quality bases from the ends of each fragment resulting in sequences of varying length despite the library containing uniform sized fragments. If more than one sized sequence is observed in your sample or if any sequences have a length of zero then FastQC will highlight that sample as having potential quality issues. Sequence size distribution was plotted to ensure sequence length did not vary due to the removal of poor quality bases from the ends of fragments by sequencers. The graph should display a peak at a single size and this was the case for all 4 samples (Fig 4. 8) (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/).

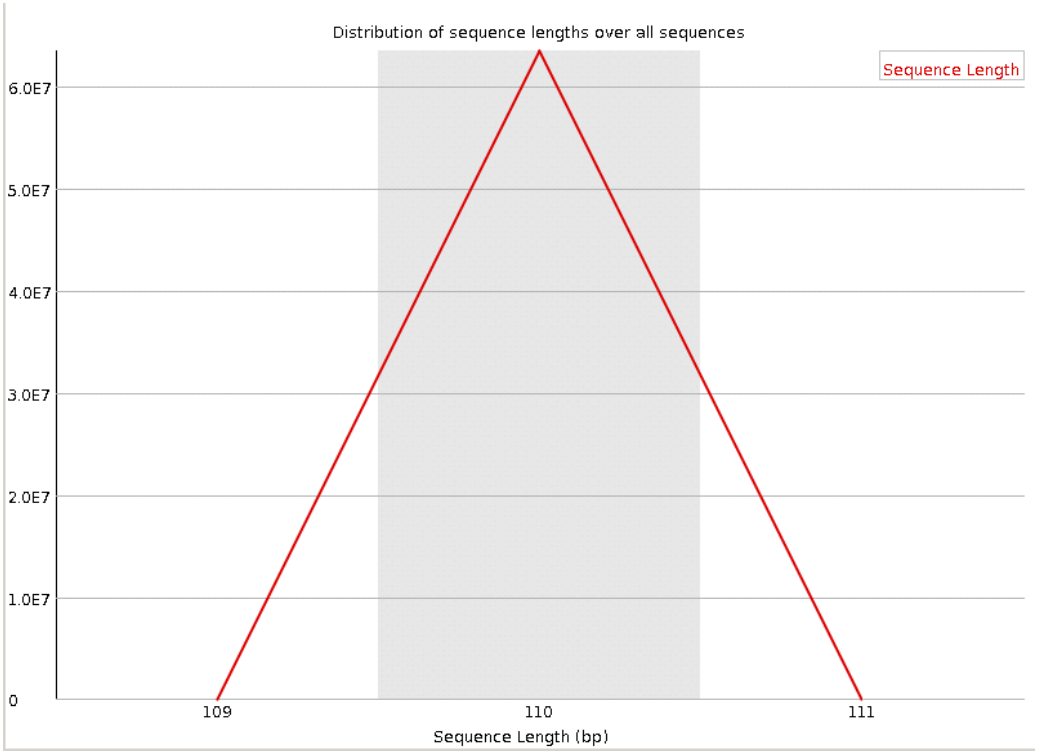


Figure 4.8 The distribution of Sequence lengths for all sequences in Sample 1. All sequences in Sample one were a uniform length of 110 bases. Exactly the same plot was generated for Samples 2, 3 and 4 with all sequences having a length of 110bp.

4.3.2.8 Sequence Duplication Level. Most sequences will be represented once (except in the final output). Low levels of sequence duplication are expected and indicate a high level of coverage of the target. High duplication levels may indicate an enrichment bias possibly during PCR amplification. How many times a sequence is duplicated is determined. Each sequence is then put in 1 of 10 categories based on how many times it is duplicated ranging from 1 (it is unique) to 10+ which indicates the sequence is present in the output 10 or more times. The proportion of duplicate to singleton (unique sequences represented only once) sequences in each of these 10 groups is then plotted (Fig. 4.9). A slight increase in the proportion of duplicates to singletons is expected in the 10+ category due to it including all sequences duplicated 10 or more times not just 10. FastQC truncates each sequence to 50bp because an exact sequence match is required to define a duplicate. FastQC only considers the first 200,000 sequences in a file to reduce processing time, and proposes that this is an adequate number to gauge the % of duplicate sequences in the whole sample (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/).

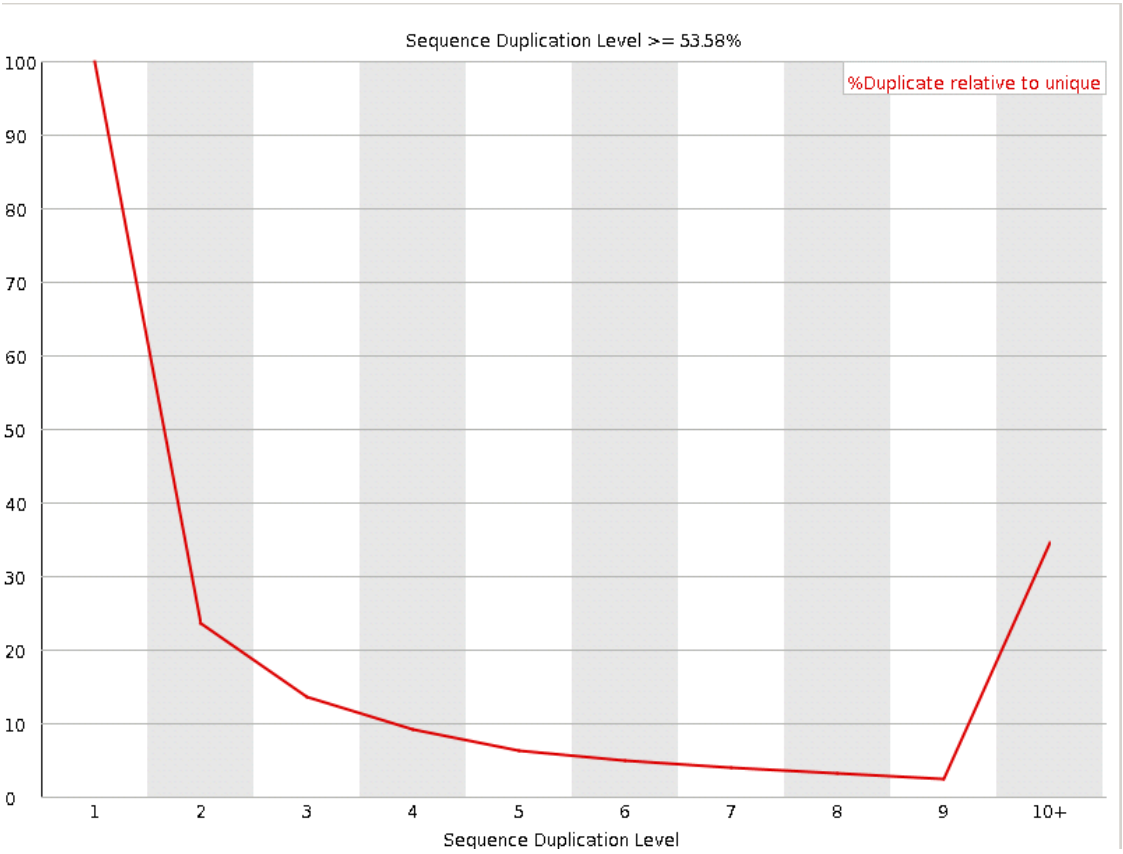


Figure 4.9. The Percentage of Duplicates in the Total Sample. A large increase in the 10+ category indicates that ~35% of samples appear in the final dataset 10 or more times. More than 50% of the total number of sequences in sample 2 (above) were estimated to be duplicates which is indicative of poor quality sequencing The percentage of duplicates relative to singleton sequences was plotted for each level of duplication from unique (1) to 10+ (duplicated 10 or more times in the sequencing output).

In Samples 1 and 3 more than 20% of the total sequences were duplicates. In samples 2 and 4, 54% and 52% of the total sequences were duplicates respectively. In FastQC if more than 50% of the total number of sequences were duplicates then this is indicative of low quality sequencing (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/). The large number of duplicates is comprised mainly of samples represented 10 or more times in the output. The plot shown in figure 4.9 was representative of each sample and although FastQC guidelines suggest that a plot like this represents poor quality sequencing it is commonly observed in RNA-sequencing. The reason for this is that with RNAseq a certain proportion of sequences will occur very frequently (e.g., housekeeping genes) whilst others will be very rare. In order to sequence the rare transcripts the common transcripts are over-sequenced, resulting in a high level of duplicates present in the data. Therefore some duplication is unavoidable for RNAseq data.

4.3.2.9 Overrepresented sequences. No single sequence is expected to contribute a significant proportion of the total sequences. If this does occur, it might reflect contamination, maybe of adapter sequences, or it could be true biological variation. Any sequence contributing more than 0.1% of the total is identified in FastQC. Again only the first 200,000 sequences are analysed so there is the possibility of missing an overrepresented sequence that occurs later in the file. The flagged overrepresented sequences are then compared to a list of common contaminants from a database accessed by FastQC and if a match is found, which is at least 20bp long and has a maximum of one mismatch, the identity is displayed. This is not robust but should be a good indicator of the type of contaminant. Often adapters are the cause and as they have similar sequences the precise adapter may not be flagged up, but another adapter will. Again reads are trimmed to 50bp. A sequence which makes up more than 0.1% of the total is considered overrepresented in FastQC and depending on the source may require further processing of the data to remove it (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/).

No overrepresented sequences were identified in the 4 forward reactions. In the reverse a single sequence was identified in all 4 samples and contributed to 0.23-0.28% of the total number of sequences and so was identified as overrepresented by FastQC. The sequence was not identified as a common contaminant, rather it was a string of N's. Given that a sequence consisting of N's is overrepresented in the sample it is surprising that the 'per base N content' was not a cause for concern. Although greater than 0.1% this is not a great enough proportion of the total number of sequences to have an impact upon the other QC results, as the problems appear to arise in the first 1 to 10 bases and the overrepresented sequence would effect more bases than this if it were contributing to low quality scores. No adapter contamination was observed in the sequences.

4.3.2.10 Overrepresented K-mers.

Smaller overrepresented sequences (less than 20bp) will be overlooked in the 'overrepresented sequence' metric (4.3.2.9) due to the applied stringency (the match must be at least 20bp with only 1 mismatch). Overrepresented 5-mers are analysed to compensate for this. The expected proportion of a k-mer is estimated using the whole library base content and then the observed proportion of the k-mer is determined from the actual count. An observed-expected ratio is then calculated. The top 6

overrepresented k-mers were plotted in a graph to demonstrate their relative enrichment across the read length. This plot can be used to determine if the enrichment in each k-mer occurs at the same position within the read each time or if it is random. To be reported as overrepresented the k-mer must show a 3-fold enrichment in the observed proportion relative to the expected proportion or there must be more than 5-fold enrichment in observed relative to expected proportion at a particular base. Greater than 10-fold enrichment of a specific k-mer indicates potential poor quality sequencing. 20% of the library is actually tested (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/).

k-mers with relative 3-fold enrichment were identified in all 4 samples. A list of the overrepresented k-mers was generated and the top 6 were plotted in a graph to demonstrate the enrichment across the read length. The most overrepresented k-mer was either TTTTT or GGGGG occurring from position 2 or 105 within the read respectively. With the exception of the GGGGG at position 105 all overrepresented k-mers were in the first 10bp of the read. Each sample had ~30 overrepresented k-mers. As the majority of overrepresented k-mers occur from the start of the read this is most likely attributable to the use of random primers in the generation of the cDNA library.

4.3.2.11 RNAseq Quality

The metrics I generated using FastQC were not designed to be used to dismiss samples based on the passing or failing of a metric, but instead used to identify areas in the sequencing where there could be a problem (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/). Whilst the sequencing in this experiment does diverge from this expected range in some metrics this is due to the expected high proportion of duplicates in an RNAseq experiment and the variation at base position 1-10 in a read due to the use of random hexamers in the library preparation. These are well documented and are not a cause for concern with regards to the quality of the data. The low ‘per base sequence content’ quality scores were however a cause for concern as this metric generated particularly low quality scores. The RNAseq data were analysed to determine genes differentially expressed and spliced in the C59X mutants, but with the caveat the any results obtained could be artefacts of the low quality sequencing and would therefore require validation with a complementary technique.

4.3.3 Zfp804a C59X mutation.

Each BAM file was visualised in the Integrative Genomics Viewer (IGV) to confirm the presence of the C59X mutation in the mutants (Fig. 4.10). The mutation was present in both mutants and the constitutive GT present in both wildtypes.

When viewing the coverage of Zfp804a I also identified two additional reads 5' of the constitutive exon 1 in one of C59X mutants that mapped across to constitutive exon 2. These two reads could indicate an alternative isoform with an alternative exon 1 that skips the constitutive exon 1 in its entirety and maps to the start of exon 2 (Fig. 4.11).

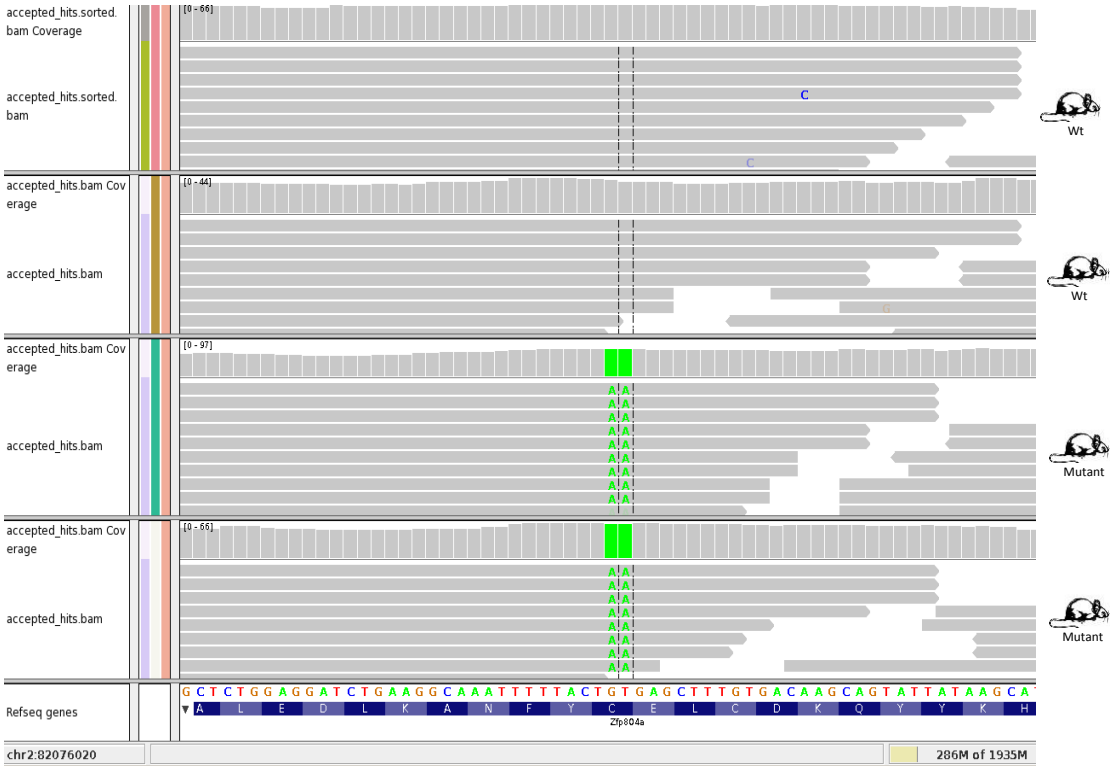


Figure 4.10. Zfp804a C59X mutation. Output from the integrative genomics viewer (IGV) shows the sequencing data of four samples following alignment. The viewer is zoomed onto exon 2 of the Zfp804a Refseq. Each of the grey bars represents a read. The C59X mutation can clearly be seen in the two mutants. The two base substitution GT>AA results in a STOP rather than the constitutive cysteine residue.

In order to confirm the predicted alternative isoform I designed primers (Appendix 4.6) to amplify between the alternative exon 1 and exon 2 which spanned the C59X mutation

in exon 2. The read representing the alternative exon is 150bp long, 94bp different to constitutive exon 1 and lies ~1300bp upstream of exon 1 with no obvious sequence similarities. These were used to amplify cDNA for the same samples used in the RNAseq experiment (Fig. 4.12). In all samples, a PCR product was detected of the size expected if the predicted novel isoform exists.

Figure 4.11. Predicted Alternative Zfp804a Isoform. The BAM alignment files (displayed using the Integrative Genomics Viewer (IGV)) are shown for all 4 samples with reads from the RNA derived from two C59X mutants (top) and two wildtypes (below). Specifically the region between Chr2 81,891638-81,895808 is shown. At the bottom is the RefSeq track. Zfp804a constitutive exon 1 is in blue. In the second track the black circle highlights 2 reads from which a novel exon upstream of Zfp804a constitutive exon 1 was identified. From the plot the reads can be seen to skip the constitutive exon 1 and map to exon 2 (out of view). This was present in only one sample, a C59X mutant.

The PCR product was next sequenced, confirming that the alternative exon links to the start of constitutive exon 2 (Fig 4.13). Sequencing also showed that the C59X mutation was present in this isoform in both the mutants. This is a novel isoform not yet documented in any of the genome browsers.

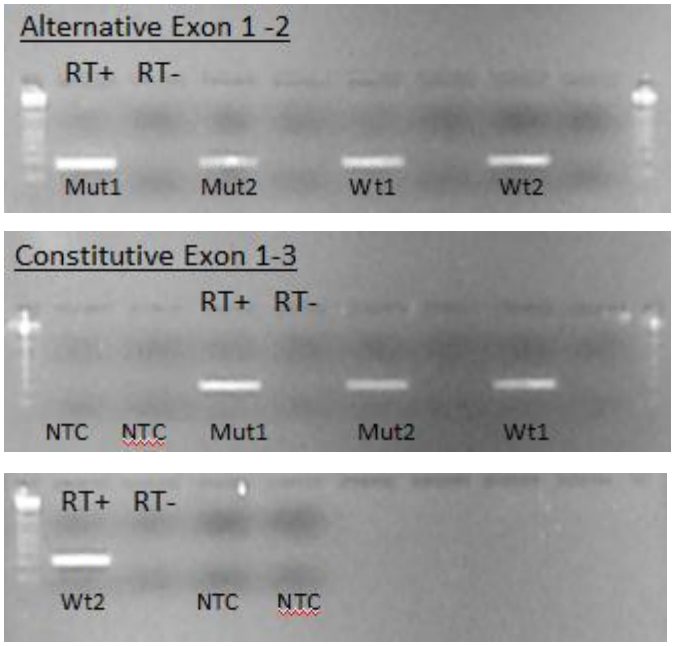


Figure 4.12. RT-PCR of Zfp804a Alternative Exon1 – Exon 2. The 4 samples run in the RNAseq experiment were assessed to determine if an isoform containing an alternative exon 1 was expressed in the mRNA. Not only was this isoform expressed (band at 173bp) in the C59X mutant from which the presence of this exon was inferred from the RNAseq output, it was expressed in all of the samples. **Mut** C59X mutant 1 and 2. **Wt** C59X Wildtype 1 and 2. **RT+** Reverse transcriptase positive **RT-** Reverse transcriptase negative control **NTC** no template control.

Following visualisation of the alignment files I used Cufflinks and Cuffdiff to determine if there were differences in expression or splicing at Zfp804a between C59X mutants and wildtypes as suggested by the exon array data. Based upon Cuffdiff metrics, both alignment and the depth of sequencing was adequate to test differential expression. Zfp804a mRNA was significantly upregulated in mutants (fold change = 1.56 upregualtion in C59X mutants relative to wildtypes, p = 2.26x10-7) and this remained significant following Benjamini and Hochberg (1995) FDR correction for multiple testing, replicating the findings from the exon array analysis. However, there were not enough alignments for splice analysis. Despite the apparent poor quality of the RNAseq data the validation of the predicted novel Zfp804a exon suggested some insights could

be gained and therefore I evaluated the most robust differential expression and splicing events predicted from the array in the sequencing data.

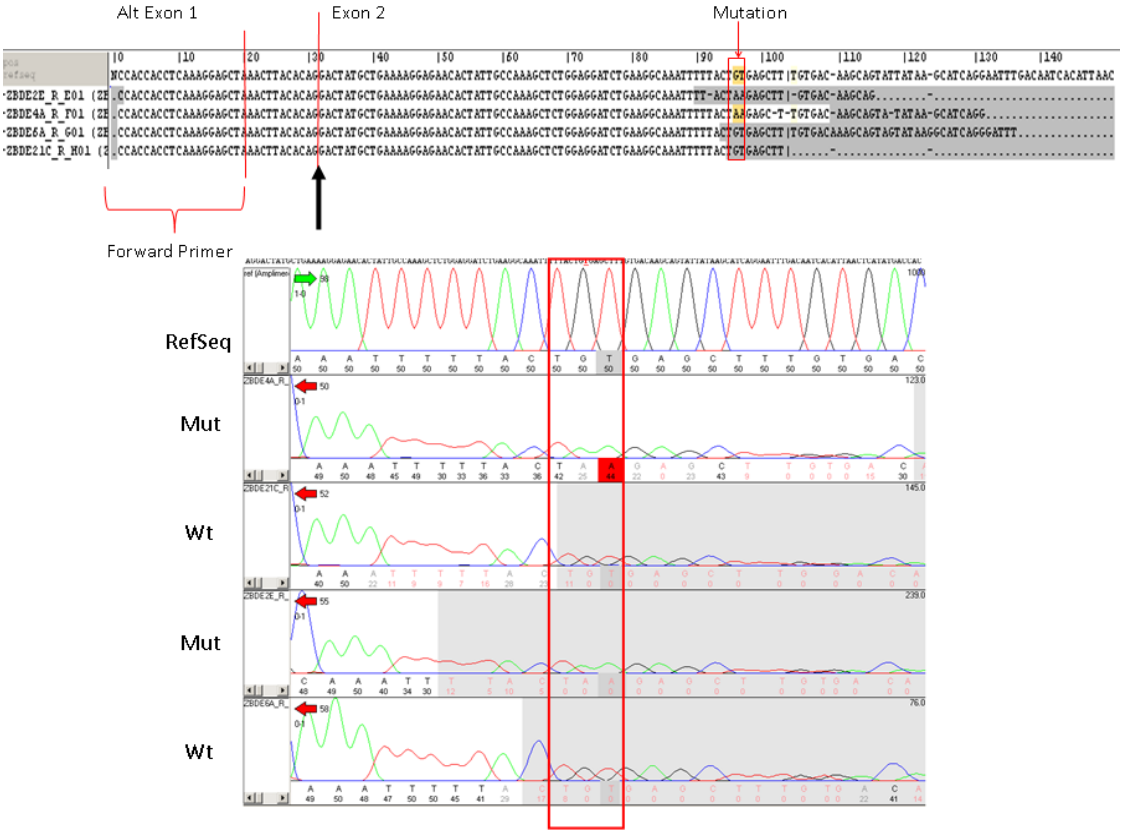


Figure 4.13 Sequencing of Zfp804a Alternative Isoform. Sequencing of the RT-PCR product showed that the constitutive exon 1 was skipped in this isoform and the alternative exon 1 sequence runs straight into constitutive exon 2 sequence (indicated by the arrow). As the C59X mutation position was spanned by the primers the traces could be analysed to see if the C59X mutation was present in this isoform. As the traces show the mutation was present in both mutants and the cysteine residue in both wildtypes (Shown in the sequence within red rectangle). Grey shaded areas indicate where sequence quality decreased toward the end of the amplicon near the reverse primer.

4.3.4 Differential Expression in C59X Mutants.

The differential expression analysis was run on 31,974 transcripts of which 30,199 had suitable quality sequencing for the tests. Of the 30,199, 2571 were significantly differentially expressed between C59X mutant and wildtype with 489 genes remaining significant following Benjamini and Hochberg FDR correction. Of the 6 genes found to be differentially expressed in females in all 3 exon array software analyses, 5 were

highly significantly differentially expressed in the RNAseq data, with the remaining gene unable to be identified in the output from cuffdiff (Table 1).

Gene Symbol	Status	Fold Change	Direction of Change	P value	FDR Sig
Arc	OK	2.76	Mutant up vs Wildtype	0	yes
Dusp1	OK	2.41	Mutant up vs Wildtype	$1.53x10^{-13}$	yes
Npas4	OK	1.92	Mutant up vs Wildtype	$4.77x10^{-09}$	yes
Nr4a1	OK	2.42	Mutant up vs Wildtype	0	yes
Snca	OK	>154.77	Mutant up vs Wildtype	$4.13x10^{-30}$	yes
Egr2	No Data				

Table 1. Replication of Top candidates from Exon Array in RNAseq Data. Of the 6 most robust differential expression findings from the exon array, 5 replicated in the RNAseq data. Egr2 was not identified in the output from Cuffdiff.

4.3.5 *Alpha Synuclein* (*Snca*)

One of the most significant differentially expressed genes was the alpha synuclein gene (*Snca*). The fold change in Snca was greater than 154.77 upregulated in C59X mutants relative to wildtypes due to the normalised expression value in the wildtype groups being 0, i.e., no expression of *Snca* in either of the 2 wildtype samples. To determine if any sequence reads were present in the wildtypes, I viewed the alignment files in IGV (Fig. 4.14). This confirmed the absence of any *Snca* reads in the two wildtypes. Reads for both *Snca* transcripts were present in both the C59X mutants.

Given this surprising result, I undertook a literature review of the *Snca* gene in mice. This revealed a paper describing a C57BL/6J substrain that have a deletion spanning the *Snca* locus (Specht & Schoepfer, 2001). This deletion was originally identified by chance in a similar gene expression study using a transgenic mouse model that had been backcrossed onto the C57BL/6JOlaHsd substrain from Harlan (Bicester, UK). Since C57BL/6J mice were used in the backcross breeding experiments in this study, it seemed possible the substrain with this deletion had been used. By determining the source of the C57BL/6J mice used for backcrossing I confirmed that the substrain used for backcrossing the C59X ENU mutants was indeed the C57BL/6JOlaHsd substrain from Harlan (Bicester, UK). To determine how widespread this effect was the average

Snca expression, based on the expression values from the array, was determined for all 16 mice used in the exon array (Fig. 4.15).

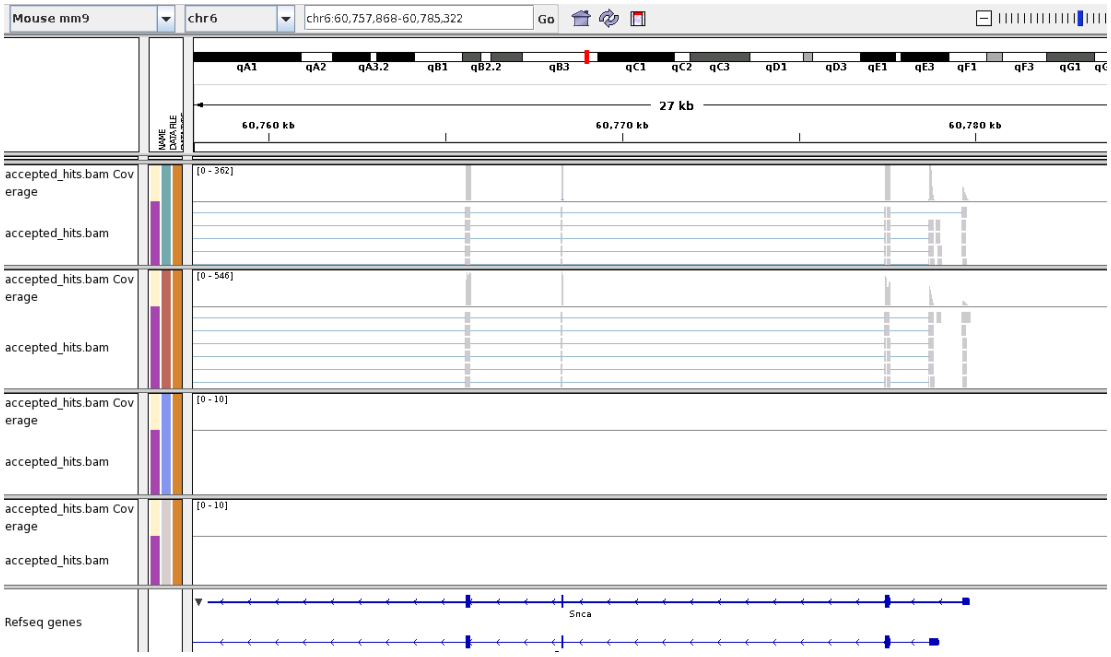


Figure 4.14. *Alpha Synuclein* (*Snca*) Expression. Viewing the alignment files of all 4 samples in IGV confirmed no reads of *Snca* transcripts (Displayed in the Refseq genes track) in the two wildtypes (bottom) where as reads were present in both C59X mutants.

All but one of the 7 C59X mutants expressed both Snca transcripts. Of the 9 wildtypes, one expressed both transcripts and two others expressed only one of the transcripts. These results therefore confirmed that the expression of the Snca gene by chance correlated with the C59X mutation in Zfp804a. Given that Snca expression correlates with Zfp804a genotype, differences in expression between C59X mutants and wildtypes could be attributable to deletion of the Snca gene.

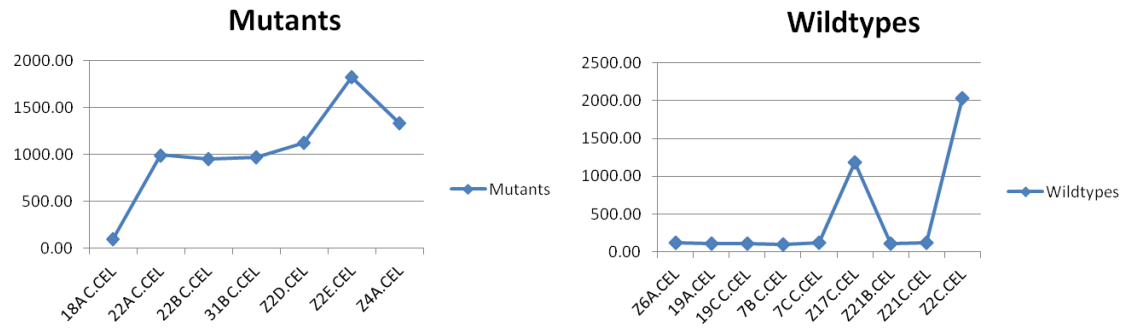
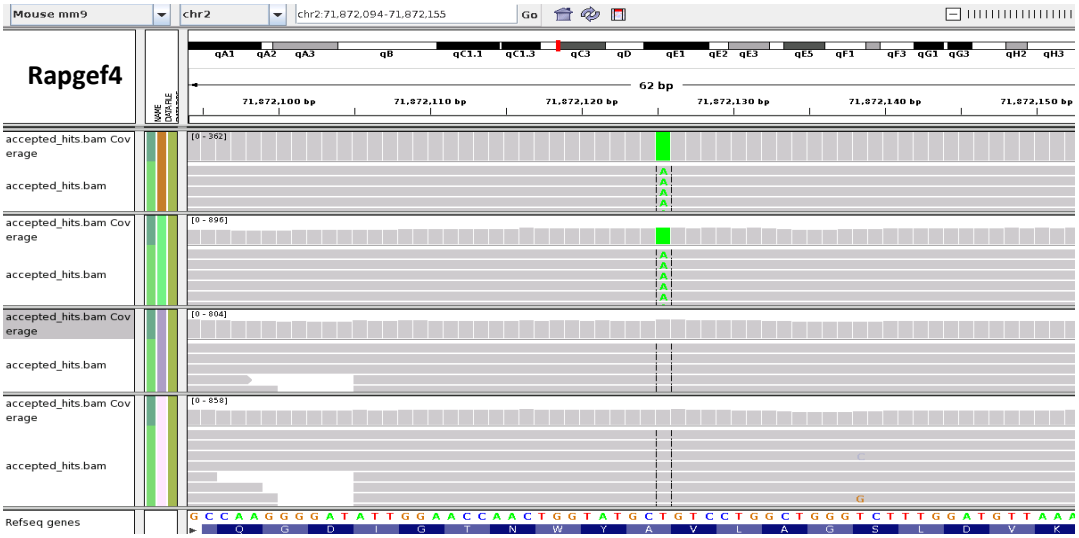


Figure 4.15. *Alpha Synuclein* Expression. Plotting the average expression of alpha synuclein (y-axis) for each sample (x-axis) showed that 6 of the 7 C59X mutants still expressed *Snca* and 2 of the 9 wildtypes expressed *Snca*. Therefore *Snca* expression appears to correlate with Zfp804a genotype.

4.3.6 RNAseq Predictions of Differential Splicing.

Next taking the splicing data, only 2957 transcripts passed read depth and quality for statistical testing. 1211 transcripts showed significant differential spicing between the 2 C59X mutants and the 2 wildtypes and 1044 remained significant following FDR multiple test correction (Benjamini and Hochberg, 1995). The numbers of nominally significant differentially spliced genes were similar to the numbers in the individual male and female analyses in the exon array study which is a surprisingly high number of genes in the RNAseq data considering only 2957 transcripts were statistically tested and may represent a large number of false positives. From the exon array 13 genes were consistently significantly differentially spliced based on 3 different software packages and in both male and female analyses of which only 5 have sufficient data to generate a test statistic in the RNAseq data. Of these, two were significantly differentially spliced, *Itga6* ($p = 0.0068$) and *Dffa* ($p = 0.01$) both of which remained significant following FDR correction. The three genes that were not significant were some of the most robust findings from the exon array with Bonferroni corrected significance values. This was surprising considering the exon array data showed them to replicate across gender, experiment, different age, different filtering criteria and different statistical algorithms. The issue of power must be considered given only two samples were in each group. To determine if there was an alternative reason the sequences targeted by probesets identified to be differentially spliced in the exon array were searched in the alignment files in IGV (Fig. 4.16).


Mouse mm9
chr2
chr2:71,872,094-71,872,155
Go
Rapgef4
62 bp
71,872,100 bp
71,872,110 bp
71,872,120 bp
71,872,130 bp
71,872,140 bp
71,872,150 bp
accepted_hits.bam Coverage
accepted_hits.bam
Refseq genes


Mouse mm9
chr2
chr2:90,903,524-90,903,588
Go
Slc39a13
65 bp
90,903,530 bp
90,903,540 bp
90,903,550 bp
90,903,560 bp
90,903,570 bp
90,903,580 bp
ccepted_hits.bam
efseq genes


Mouse mm9
chr2
chr2:79,500,418-79,500,486
Go
Ssfa2
69 bp
79,500,430 bp
79,500,440 bp
79,500,450 bp
79,500,460 bp
79,500,470 bp
79,500,480 bp
accepted_hits.bam
efseq genes

Figure 4.16. C59X Mutant Specific Single Nucleotide Polymorphisms in Probe Target Regions. Consistently significant differential splicing at 3 genes (*Rapgef4*, *Slc39a13* and *Ssfa2*) observed in the exon array experiment failed to replicate in the RNAseq analysis. Visualisation of the sequences targeted by the probesets in IGV highlighted the presence of C59X mutant specific SNPs which might have had an impact upon hybridisation efficiency on the Exon array. In each of the three examples the 2 C59X mutant sequencing reads are displayed as the first two tracks of the IGV display and the 2 wildtypes the second two. The SNPs lie between the tram tracks and is presented as the letter which represents the base change. In all 3 the probeset sequence matches the wildtype sequence. The corresponding exons were downregulated in the C59X mutants relative to the wildtypes. In *Ssfa2* and *Slc39a13* the SNPs were dbSNPs (dbSNP build 128) in *Rapgef4* the SNP was novel.

In each instance the alignment files for the corresponding probeset region contained a SNP which was present in the majority of all reads of the 2 C59X mutants and not present at all in both wildtypes. Each of these SNPs corresponded to the location of a probe and so the identified mutant specific SNP may have affected the hybridisation efficiency of the probe binding to the target. Revisiting the geneviews of these same three genes from the exon array data showed that in each instance it was the C59X mutants which showed down regulation relative to the wildtypes and this would fit with a reduced hybridisation efficiency effect since the probe sequence corresponds to that of the wildtype and not the C59X background (Fig. 4.17). Each of these genes lies on Chromosome 2 and the SNPs may potentially represent mutations linked to the C59X mutation.

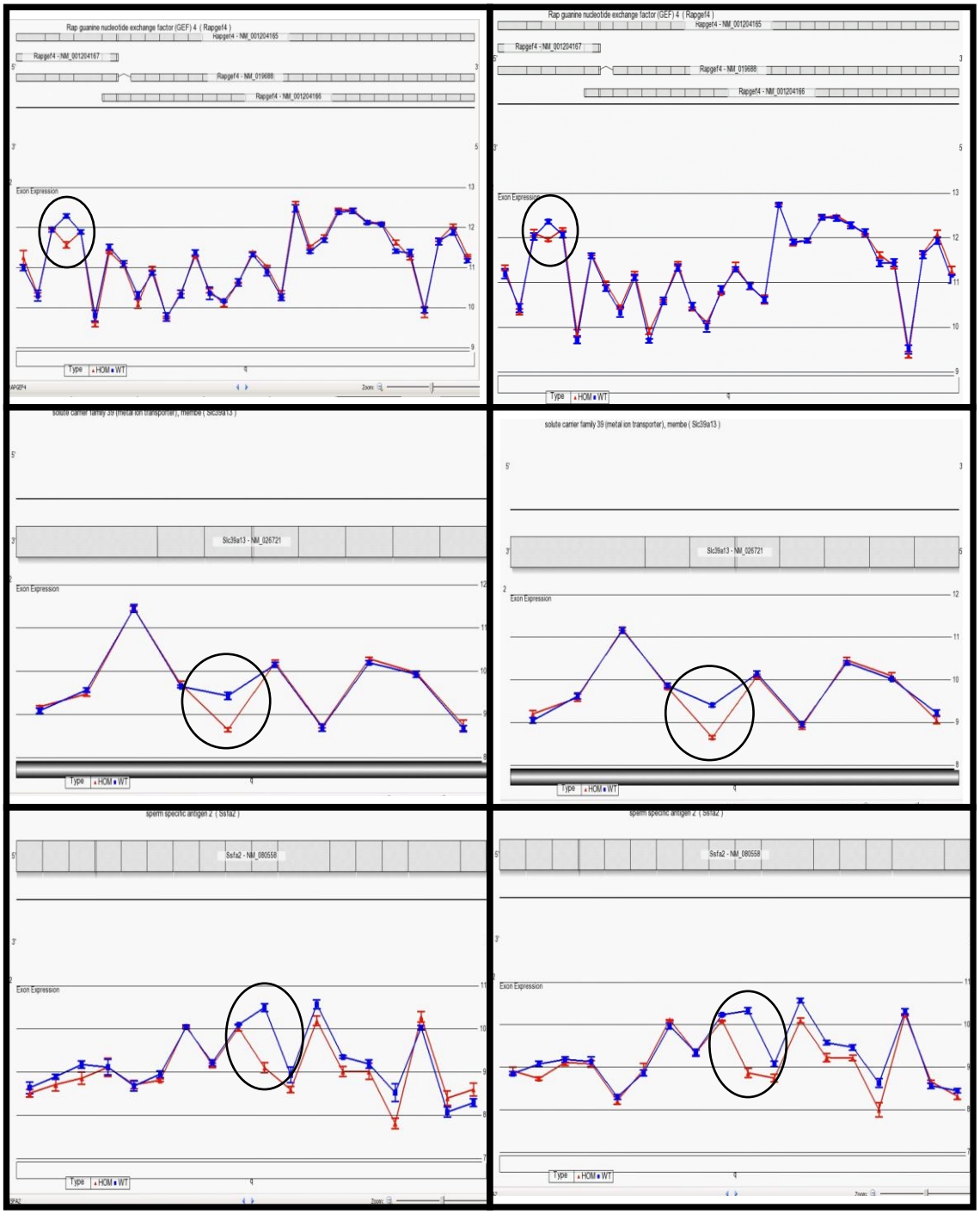


Figure 4.17. Direction of Effect in Probesets with Predicted Differential Splicing. The exon array results predicted differential splicing in the circled probesets in each of the 3 genes *Rapgef* 4 (top), *Slc39a13* (middle) and *Ssfa2* (bottom). The C59X mutants differ from the wildtype at a base, and that base does not match the sequence of the probe which is perfectly complimentary to that of the WT. This may cause reduced hybridisation efficiency specifically in the mutants resulting in a false call of downregulation in the relevant probeset. The probesets are downregulated in C59X mutants (Hom) in all 3 genes for females (left) and males (right).

4.4 Discussion.

Next generation RNA sequencing was carried out on a pilot sample of 2 male C59X mutants and 2 male wildtypes. Initial RNA and library preparation quality control showed good quality RNA and fragments respectively, yet QC on the sequencing data highlighted numerous poor quality attributes of the data. Most notably the reliability of base calling based on phred quality scores was poor particularly towards the end of each read. It was evident that there was variable GC content and disproportionate representation of each of the four bases both of which occurred within the first 10 bases of the read. Based on common causes for such results I hypothesised that the most likely explanation was that a sequence such as an adapter was contaminating the sample, however no adapter or any other contaminating sequence was identified frequently enough in the total number of sequences to confirm this. Duplicate levels were high but this is often observed in RNAseq data when highly expressed sequences are over-sequenced in order to sequence the rare transcripts. Despite a high number of duplicate sequences present in the data no reads were excluded as this is actually more detrimental to the results due to the considerable loss in data and the ambiguity of the source of such duplicates. Whilst PCR amplified duplicates can result in false positive results, duplicates from highly expressed genes represent true positive findings (Bainbridge et al., 2010).

Following the Tuxedo Protocol alignment files were initially assessed in the Integrative Genomics Viewer (IGV) (Robinson et al., 2011; Thorvaldsdóttir et al., 2012) with the intention of determining the presence of the C59X mutation in the mutants specifically. This was confirmed and led to the identification of a novel exon in *Zfp804a* upstream to the Refseq exon 1 in 1 C59X mutant. The alternative exon 1 appeared to skip the constitutive exon 1 and splice to exon 2 of *Zfp804a*. RT-PCR followed by sequencing validated this transcript which was found in all 4 samples, and confirmed the C59X mutation was also retained in exon 2 of both C59X mutants. Determining the open reading frame as well as the start and end positions of the alternative exon and transcript is necessary to fully understand *Zfp804a* expression in the C59X mice, but was not able to be carried out within the timeframe of this PhD.

Although the RNAseq was generally of poor quality, I undertook an assessment where possible of the most robust results from the exon array experiment. It is important to consider when comparing the exon array and RNAseq data that each protocol used

slightly varying amplification methods. The RNAseq protocol targeted polyA RNA whereas the exon array samples were random primed. Despite this, concordance between the two platforms has been demonstrated (Raghavachari et al., 2012). 5 of the 6 most robust differential expression results replicated in the RNAseq data with significance surviving multiple test correction using the FDR correction (Benjamini and Hochberg, 1995). This therefore provides strong evidence that these are true expression changes. I discovered that both wildtype samples had no reads in *Snca*. This was found to be the result of a deletion which spans the *Snca* locus present in the C57BL/6JHsdOla substrain which was used for backcrossing the C59X mutation. Expression results from the array of all 16 $F3_i$ mice found the issue to be widespread and most worryingly to correlate with genotype.

Silent mutations are a common problem in inbred strains of mice and have been documented on numerous occasions. Examples include the discovery that the 129S6/SvEv inbred substrain has a 25bp deletion in the *Disc1* gene resulting in a frameshift and a premature termination codon (Koike et al., 2006). Silent mutations, which have no observable phenotype, are particularly problematic as the mice are continually used for breeding and the *de novo* silent mutation becomes fixed in the strain. Stable silent mutations in an inbred strain can confound studies when attempting to determine the effects of a particular mutation, when in fact two mutations are present. In the C57BL/6JHsdOla strain the deletion occurred following transfer of the strain to a breeding centre in Harlan. Due to the lack of an observable phenotype the mice are continually used in breeding and the deletion became stable in the substrain. It was only when expression studies began that used the C57BL/6JHsdOla strain for backcrossing that the deletion which spans the *Snca* locus was discovered (Specht & Schoepfer, 2001).

Specifically the deletion affects 365kb of chromosome 6, which includes *Snca* and one other gene; *Multimerin* (Specht & Schoepfer, 2004), but this gene is not thought to be brain expressed (Leimeister et al., 2002). Neither compensatory up regulation of β or γ *synuclein* has been observed in the strain nor changes in the expression of other genes in an array study in these mice (Specht and Shoepfer., 2001). These observations suggest that mice with the *Snca* deletion would not have altered expression in other genes as a result of this mutation alone and therefore differential expression observed between C59X mutants and wildtypes is most likely the result of disrupted Zfp804a, however

this is not definite and there is always the caveat the *Snca* deletion could confound the results.

De novo mutations which arise and have no apparent phenotype become problematic when the mutation is in such close physical proximity that it tends to be co-inherited with the mutation that is the target of the study, so distinguishing the effects of each is difficult (Gajovic et al., 2006). This cannot have occurred in the C59X mice as *Zfp804a* lies on Chromosome 2 and *Snca* on Chromosome 6. The problem lies with the use of the $F3_i$ generation in expression studies as the mice only had ~96% C57BL/6JOlaHsd purity and you would expect ~7 additional mutations from the ENU derived founder strain.

The evidence for a deletion at the *Snca* locus puts into question the validity of the exon array and RNAseq experiments using tissue derived from the $F3_i$ generation. Unfortunately, delays in the availability of mice meant the $F3_i$ generation had to be used for the initial experiments, although the intention was to repeat the experiments in later (F7) generations which are expected to have at least 98% of the C57BL/6JHsdOla genome.

Poor sequence quality meant 8 out of the 13 differential splice candidates from the exon array were not able to be tested in the RNAseq data. Of those that had adequate read quality, replication of significant differential splicing was observed in only two of the 5. Considering only 2957 genes were statistically tested and 1211 were significant this is no greater than would be expected by chance. The lack of replication is possibly attributable to false positives in the array data or not enough power in the study given only an n of 2 in each group. The variants observed in the C59X mutant sequence may have influenced the efficiency of probe hybridisation on the exon array, and consequently a false positive splice prediction due to misinterpreted downregulation of the exon in the mutants is observed. A large source of errors in array results are generated by SNPs in sequences targeted by probes. Humans are outbred and genetically heterogeneous and there is an abundance of literature on the best way to deal with probes that target SNPs on assays designed for humans (Duan et al., 2008; Gamazon et al., 2010). Isogenic samples should not produce such false positives. The mice in this experiment are heterogeneous and differences correlate with the Zfp804a genotype. The most likely explanation for C59X mutant specific polymorphisms is that at $F3_i$, polymorphisms that distinguish the C57BL/6JHsdOla and Balb/c or C3H/HeJ

strains are still present, as indicated with the *Snca* finding. There is also the possibility that the C59X mutant specific SNPs found on chromosome 2 are linked to the C59X mutation and being inherited together. A major strength of the RNA sequencing data was the ability to assess the region surrounding the C59X mutation in both C59X mutant and wildtypes at single base resolution. Whilst only two mice from each experimental group were sequenced, there was an unambiguous finding of sequence variants in the C59X mutants specifically. It is possible that these strain specific sequence variants will not be separated from the C59X mutation by recombination by the F7 generation and therefore the same genes would be predicted to be differentially spliced.

Whilst the data were not of the most robust quality, the RNAseq experiment enabled me to make several critical findings. The identification of a novel *Zfp804a* isoform when fully characterised is likely to be informative. Without the RNAseq data the identification of a deletion at the *Snca* locus and several polymorphisms in the C59X mutants would not have been possible. These findings are critical to the interpretation of the data generated with $F3_i$ generation mice. The occurrence of these strain specific sequence variants was investigated in the F7 generation to enable the correct interpretation of the results (Chapter 6).

Chapter 5. Expression Analysis in Embryonic C59X Mice.

5.1 Introduction.

One of the leading hypotheses of schizophrenia is that it has its early origins in disruption of brain development *in utero*, but that the major symptoms remain latent until adolescence or adulthood (Weinberger, 1986). Although the evidence is not definitive, a number of lines of evidence are broadly in favour of this hypothesis (Owen et al., 2011).

Follow up of births occurring during or just after the 1957 influenza epidemic showed those exposed to the epidemic *in utero* had an increased rate of schizophrenia diagnosis in adulthood compared to those not exposed to the epidemic (Mednick et al., 1988; O'Callaghan et al., 1991a; Cooper, 1992) leading to the hypothesis that the influenza virus might be an *in utero* insult increasing risk of schizophrenia, but these conclusions have been widely challenged (Kendall and Kemp, 1989; Bowler & Torrey, 1990; Crow et al., 1991, Crow et al., 1994). Increased rates of influenza in pregnant women with young children have been observed (Hennessy et al., 1964), an observation considered by some to support a contribution to schizophrenia from in utero infection since an increased rate of schizophrenia has been suggested in younger siblings (Farina et al., 1963). An enrichment of winter and spring births in those with schizophrenia has being widely reported (O'Callaghan et al., 1991b) and may be more common in patients who have no family history of the disorder, again possibly implicating environmental factors such as viral infection which are more prevalent at that time of year (Sham et al., 1992). Malnutrition during the first trimester (Susser et al., 96), toxoplasmosis (Brown et al., 2005), respiratory infection (Brown et al., 2000) and bacterial infections (Sorenson et al., 2009) have also been suggested to cause damage *in utero* resulting in aberrant development as well as neonatal adversity, including obstetric complications (Murray et al., 1985) and CNS infection (Rantakallio et al., 1997). While the nature of the insult is disputed there appears to be a body of evidence that pre and perinatal insults may contribute to schizophrenia. Early studies indicated the second trimester of pregnancy as the temporal window when increased susceptibility occurred (Mednick et al., 1988) more recently studies have suggested insults occurring as early as the first trimester and as late as the neonatal period increased risk, with the Prenatal Determinants of Schizophrenia (PSD) study reporting risk is incurred during either conception or the first few weeks of pregnancy (Meyer et al., 2007).

How the putative insults lead to pathophysiology which predisposes schizophrenia is less clear. The immune response, which is common to the different types of infections, may play a role. Increased levels of IL-8 in mothers of schizophrenia spectrum disorder patients have been reported (Brown et al., 2004). IL-8 is thought to have an important role in brain development and thus increased levels may cause aberrant development (Gilmore, 1997), but IL-8 protein levels may themselves be mediated by genetic risk. Insults during development of the CNS may disrupt important processes such as cell proliferation and migration leading to aberrant axonal connectivity (Murray and Lewis, 1987).

Structural differences in the schizophrenic brain such as enlarged ventricles and reduced cortical volume appear to be present at the onset of disease (Roberts, 1991) and are often described as non-progressive, again suggesting important brain changes have already occurred prior to the onset of clinical symptoms. Other features which favour a neurodevelopmental hypothesis of schizophrenia include observations of motor, behavioural and cognitive impairment in children suggestive of a 'pre-schizophrenic' brain. These observations are not apparent in every schizophrenia patient and it maybe that factors occurring during development are responsible for increased risk in a subset of schizophrenia cases (Sham et al., 1992). The neurodevelopmental hypothesis of schizophrenia is also strengthened by the finding that copy number variants (CNVs) have been identified that are common to schizophrenia and neurodevelopmental disorders such as ADHD, autism and mental retardation (Mefford et al 2008; Wassink et al., 2001).

Infectious agents administered to pregnant rats have consequences on neurodevelopment and the timing of the insult affects the extent of damage, with earlier insults associated with more widespread damage (Meyer et al., 2007). Administration of infectious agents in rats at E18 caused brain atrophy and whiter matter thinning as well as gene expression changes in the PFC and hippocampus at P35 (equating to adolescence) (Fatemi et al., 2008). Lesions in the rat hippocampus at P7 have been reported to impair the pre-pulse inhibition (PPI) startle response (a putative schizophrenia endophenotype or biomarker) following puberty (P35) relative to rats with sham lesions. This fits with a neurodevelopmental model of schizophrenia where by an insult occurs during development but a phenotype is not observed until early adulthood (Lipska et al., 1995). Prenatal exposure of rats to viruses (Piontkewitz et al. 2012), bacteria or cytokines (Samualsson et al., 2006) have elicited behaviours in

offspring akin to certain cognitive and behavioural symptoms observed in schizophrenia and perhaps more convincingly some of these studies have shown amelioration of these symptoms following administration of atypical antipsychotics (Borrell et al., 2002). Whilst a number of studies report findings in favour of a neurodevelopmetal hypothesis it is best to view the animal literature with caution particularly when inferring similarities between animal models and infection in humans (Samuelsson et al., 2006).

A developmental hypothesis implies a latent period prior to the onset of clinical symptoms, but understanding what is happening during this latent period has proven difficult. Extensive maturation processes occurring during adolescence may initiate the requirement for systems in the brain that up until that point were not utilised (Weinberger, 1986). Synaptic development and pruning could be affected by genetics or the hormonal imbalances and stress common during adolescence (Benes et al 1994; Walker, 1994).

There are a number of studies showing genes which switch from foetal to adult expression patterns in the first few postnatal weeks. This switch is observable during this period in *NRXN1* and *NRXN3* (Lijima et al., 2011). The *SNAP25* gene is thought to switch between isoforms a and b between postnatal day 25 and 35 which is thought to alter the efficacy of synaptic transmission to allow the stabilisation of the developed neuronal circuit (Bark et al., 2004). Similar postnatal switches in isoform expression are seen in NUMB (Bani-Yaghoub et al., 2007) CELF and MBNL (Kalsotra et al., 2008). The latter two regulate a number of alternative splice events themselves and misregulation of their targets has been implicated in Myotonic Dystrophy (DM) (Charlet et al., 2002). DM is characterised by muscle wasting and myotonia and is caused by the misregulation of splicing. Several of the genes involved fail to switch from foetal to adult isoforms resulting in the isoforms being expressed at inappropriate times (Charlet et al., 2002).

A number of splice regulatory factors are themselves developmentally regulated for example PTB (Boutz et al., 2007), which has implications for the splicing of downstream targets. In this way it is thought that a splicing network exists initiated from regulation of splice factors which in turn regulate their downstream targets resulting in extensive developmental regulation of alternative spliced isoforms (Revil et al., 2010).

If Zfp804a does have a role in regulation of transcription and or splicing the genes it regulates may be developmentally regulated. In support of this hypothesis an RNA

sequencing study looking at differential expression during neurogenesis of human neurons from iPS cells showed one of the genes with large expression changes to be ZNF804A (Lin et al., 2011). Based on a neurodevelopmental hypothesis of schizophrenia, splicing differences observed between C59X mutants and wildtypes that occur during development may point to genes fundamental to aberrant development relevant to the aetiology of schizophrenia. To determine if developmentally regulated splice or expression changes occurred as a result of the C59X mutation, I undertook an exon array experiment using brain tissue from embryonic day 18.5 mice as this time point has been likened to the second trimester of pregnancy in humans (Fatemi et al., 2008), which although controversial, is the period when the embryo may be most susceptible to environmental factors which could predispose schizophrenia risk in later life.

5.2 Materials and Methods

5.2.1 Sample

To study embryonic expression in later generation mice (~98.5% C57BL/6HsdOla background), I set up heterozygote intercrosses using mice with ~98.5% C57BL/6 HsdOla genome from either the 7^{th} or 8^{th} generation (Appendix 5.1). Mice used for breeding were caged individually. Female mice were placed in the male home cage in the evening and the observation of a vaginal plug the following morning was recorded as embryonic day 0.5 (E0.5). Female mice were then returned to their home cage. 18 days later whole brain and tail tip samples were collected from each embryo on embryonic day 18.5 (E18.5). The pregnant female was killed by cervical dislocation and then placed ventral side up and 70% ethanol was applied to the body. An incision was made and then a cut down the midline using scissors to expose the uterus from which the embryos were carefully removed. The embryo was decapitated and the tip of the tail collected. The head was then placed in ice cold 1x PBS solution if necessary and the brain extracted from the skull. Brain and tail samples were immediately snap frozen in liquid nitrogen then stored at -80°c until RNA extraction.

5.2.2 Genotyping

DNA was extracted from tail tips (Chapter 2.2.1) and quantified using the Nanodrop 1000 spectrophotometer (Thermo Scientific) before being amplified in a PCR reaction (Chapter 2.4) using primers spanning the C59X mutation (Appendix 2.1). Sequencing was carried out as previously described (Chapter 2.6) and genotype assessed in NovoSNP version 3.0.1 (Weckx et al., 2005).

5.2.2.1 Gender PCR.

The gender of the embryonic samples was assessed using a multiplex PCR with gDNA extracted from tail tips. Two primer sets were used, one for the Y-linked gene, *Ssty* and one for an autosomal control gene, *Om1a* (Myogenin) (Appendix 5.2). Running the product on a 1.5% gel enabled males, with a band representing both *Ssty* and *Om1a*, to be distinguished from females with the single band corresponding to the *Om1a* product.

PCR Reaction.

The PCR reaction was performed using C1000 Thermocyclers (BioRad Laboratories, Inc). The PCR cycling conditions are outlined below. PCR was performed in a 25µl reaction volume using 1µl genomic DNA, 15.75µl Sterile water, 2.5µl 10X buffer, 2.0µl 25mM $MgCl_2$, 1.0µl 5mM dNTPs, 1.0µl Ssty (forward) 10µM and 1.0µl Ssty (reverse) 10µM, 0.25µl Om1a (forward) 500ng/µl and 0.25µl Om1a (reverse) 500ng/µl and 0.25µl HotStar Taq (Qiagen).

The PCR cycling conditions.

1. 94°C for 4 minutes
2. 94°C for 45 seconds
3. 61^{o}C for 45 seconds
4. 72^{o}C for 45 seconds
5. Repeat steps 2-4 for 35 cycles
6. 72^{o}C for 5 minutes

5.2.3 RNA Processing

RNA was extracted from whole brain and purified using an RNeasy column (Qiagen) as described in Chapter 2.2.2. Samples were then quantified and assessed for quality using the NanoDrop 1000 spectrophotometer (Thermo Scientific) and the Agilent 2100 Bioanalyser respectively (Chapter 2.2.3.1).

5.2.4 Affymetrix Exon Array

Total RNA was labelled and prepared for hybridisation to the Affymetrix Genechip® Mouse Exon 1.0 ST array as described in full in Chapter 2.8. Sample preparation and chip scanning protocols did not differ to those used on adult mice (Chapter 2.8).

5.2.4.1 Quality Control.

This was assessed as described in Chapter 2.8.3.5.

5.2.4.2 Determining Snca Expression

To establish how many of the embryonic mice had the C57BL/6JHsdOla specific deletion, which spanned the *Snca* locus I calculated the average raw expression values across all *Snca* probesets for each sample.

5.2.4.3 Partek Genomics Suite

The same filtering criteria and statistical algorithms were applied to the embryonic dataset as described for the adult data in chapter 3. As before, an initial analysis focused on C59X mutant and wildtype samples homozygotes but an additional analysis that included the heterozygote samples was also performed.

5.2.4.4 EasyExon.

I analysed the embryonic data in easyExon as described in chapter 2. Briefly CEL files were uploaded for the 12 mutant and wildtype samples. For a probeset to be included the DABG p value had to be ≤0.05 in 6 of the 12 samples. For differential expression analysis genes required a fold change of greater than 1.5 and a MiDAS p value ≤0.05 to be significant. Significant differential splicing was defined by a MiDAS p value ≤0.05. As only two groups can be included in the analysis only the C59X mutant and wildtype samples were compared.

5.2.4.5 AltAnalyze.

I uploaded the 12 CEL files into AltAnalyze and compared expression and splicing in C59X mutants and wildtypes using the MiDAS and FIRMA algorithms respectively as described in chapter 2.

5.2.4.6 Overlap Analysis

Using the method outlined in Chapter 3, section 2 the overlap between differentially expressed and spliced genes in embryonic samples at different p value thresholds ($p \leq 0.05$, $p \leq 0.01$, $p \leq 0.001$ and $p \leq 0.0001$) was determined in the adult dataset at a nominal $p \leq 0.05$. The adult data set used was that of all 16 samples (male and female) combined with gender covaried.

The degree of overlap for the embryonic results in each of the 3 software packages (Partek GS, easyExon and AltAnalyze) was then assessed. Differentially spliced genes from this overlap found to be significant following an FDR correction (threshold 0.05) in Partek GS were assessed to determine how many overlapped in the adult dataset at $p \leq 0.05$. Genes that were significantly differentially spliced following Bonferroni correction in adult and embryonic samples were also assessed to determine the genes common to both analyses at this stringent p value threshold.

5.3 Results

5.3.1 Embryonic Sample.

5 successful heterozygote intercrosses were set up between mice with 98.5% C57BL/6JHsdOla background (Appendix 5.2). Brain and tail tips were collected from the embryos on E18.5. DNA was extracted from the tail tips and processed using PCR (Fig. 5.1) and sequencing (Fig. 5.2) to determine gender and genotype of each embryonic sample (Table 5.1). The results of each revealed there to be a total of 7 C59X homozygote mutants and 6 wildtypes. 12 C59X heterozygotes were also selected for inclusion on the exon array. Of the 25 samples 12 were female and 13 male.

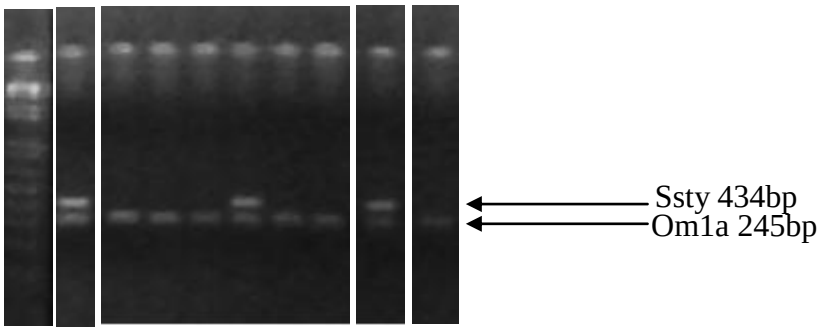


Figure 5.1 Determining Gender in E18.5 Mice. A multiplex PCR was used to determine the gender of each embryonic sample by using a Y-linked gene (*Ssty*) and an autosomal control gene (*Om1a*). *Ssty* had a product of 434bp and *Om1a* 245bp. Samples with both bands represented males and samples with only the smaller band (*Om1a*) represented female samples.

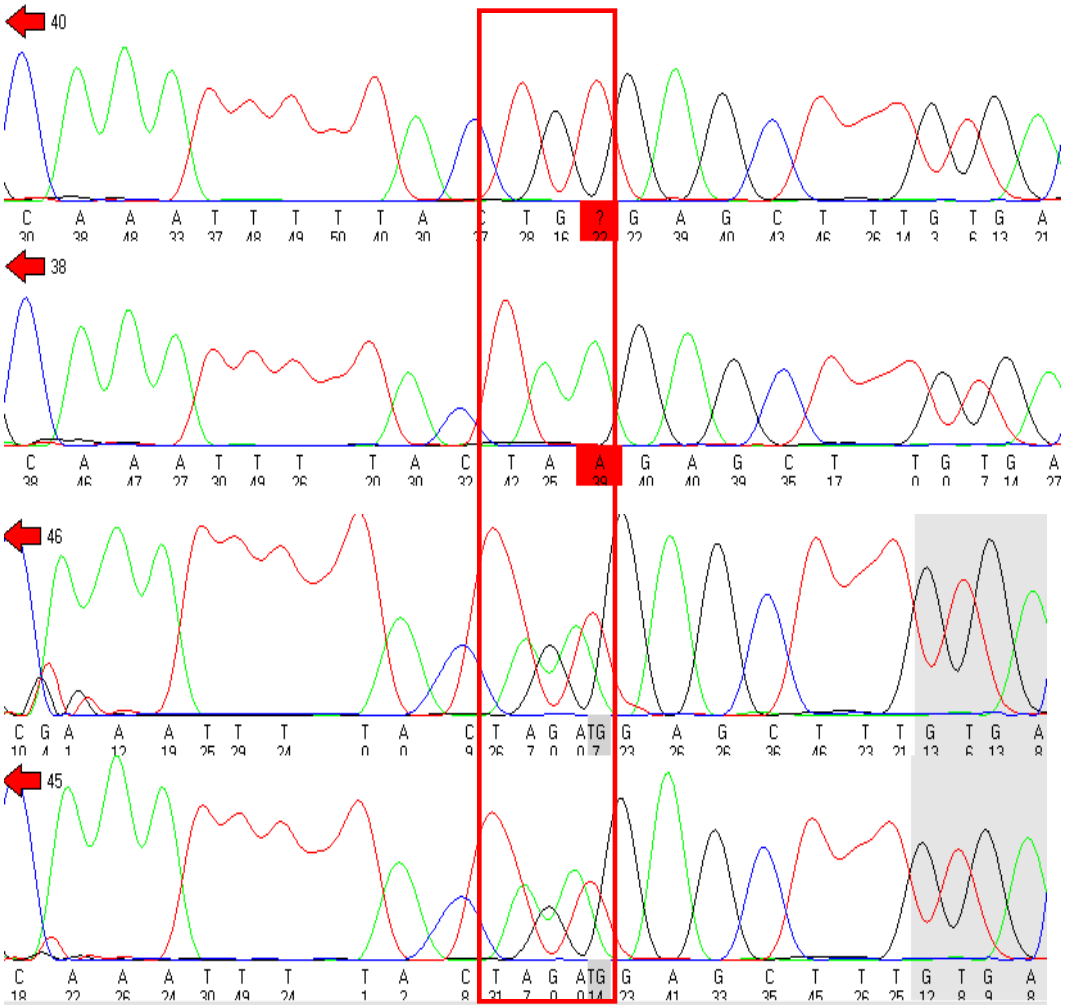


Figure 5.2 Sequencing of the C59X mutation in Embryonic Samples. PCR was carried out on gDNA using primers which spanned the C59X mutation in Zfp804a. The PCR product was then sequenced to determine the genotype of each embryo. The example above shows sequencing traces from 4 of the embryonic samples. The red rectangle highlights the position of the normal cysteine codon (TGT) in exon 2 of Zfp804a. The top trace represents a homozygous wildtype. The second trace is from a C59X homozygote mutant in which both alleles have the ENU C59X mutation. The bottom 2 traces are C59X heterozygotes each with one C59X mutant allele.

Sample ID	Genotype	Gender	Age
B8A2P1	Het	Male	E18.5
B8A2P4	Het	Female	E18.5
B8A2P5	Het	Female	E18.5
B8E1P1	Het	Male	E18.5
B8E1P7	Het	Male	E18.5
B8E1P8	Het	Male	E18.5
E27AD2P3	Het	Male	E18.5
E27AD2P4	Het	Female	E18.5
E27AD2P5	Het	Female	E18.5
E27AM2P2	Het	Male	E18.5
E27AO4P3	Het	Female	E18.5
E27AO4P4	Het	Male	E18.5
B8A2P2	Mut	Male	E18.5
B8E1P2	Mut	Female	E18.5
E27AD2P6	Mut	Male	E18.5
E27AD2P2	Mut	Male	E18.5
E27AM2P3	Mut	Female	E18.5
B8E1P3	Mut	Female	E18.5
B8E1P4	Mut	Male	E18.5
B8A2P3	Wt	Female	E18.5
B8E1P5	Wt	Female	E18.5
E27AD2P7	Wt	Female	E18.5
B8E1P6	Wt	Male	E18.5
E27AO4P2	Wt	Female	E18.5
E27AM2P1	Wt	Male	E18.5

Table 5.1. The C59X Embryonic Sample. Tail tip gDNA was extracted and used for assessment of genotype and gender in the embryonic mice. The sample was comprised of 12 C59X heterozygotes (Het), 7 C59X mutants (Mut) and 6 wildtypes (Wt), with 12 females and 13 males.

5.3.2 RNA quality.

The whole brains from these 25 samples were then processed to extract RNA and the quality was determined (Table 5.2). With the exception of E27AM2P3, RNA was of good quality with RIN above 8 and 28s/18s ratios above 1. RNA was hybridised to the exon array chips and then assessed on a number of quality metrics using Partek GS and Expression Console (Affymetrix) as described previously (Chapter 2.8.3.5).

Sample ID	Genotype	Gender	rRNA Ratio [28s/18s]	RIN
B8A2P1	Het	Male	1.8	10
B8A2P2	Hom	Male	1.8	10
B8A2P3	Wt	Female	1.8	10
B8A2P4	Het	Female	1.8	10
B8A2P5	Het	Female	1.9	10
B8E1P1	Het	Male	1.7	10
B8E1P2	Hom	Female	1.7	9.8
B8E1P3	Hom	Female	1.6	9.8
B8E1P4	Hom	Male	1.8	9.7
B8E1P5	Wt	Female	2	10
B8E1P6	Wt	Male	1.7	9.8
B8E1P7	Het	Male	1.8	10
B8E1P8	Het	Male	1.9	10
E27AD2P1	Hom	Male	1.5	9.9
E27AD2P3	Het	Male	1.9	10
E27AD2P4	Het	Female	1.9	10
E27AD2P5	Het	Female	2	10
E27AD2P6	Hom	Male	1.8	9.8
E27AD2P7	Wt	Female	1.8	10
E27AM2P1	Wt	Male	1.6	9.5
E27AM2P2	Het	Male	1.8	10
E27AM2P3	Hom	Female	0.1	6.7
E27AO4P2	Wt	Female	1.8	10
E27AO4P3	Het	Female	1.8	10
E27AO4P4	Het	Male	1.7	10

Table 5.2. RNA quality Scores. RNA samples were run on the 2100 Bioanalyser (Agilent). The resulting 28s/18s ratio and RIN scores are presented in the table. Sample E27AM2P3 (highlighted in red) has values less that the accepted (chapter 2.2.3) thresholds of 1 and 7 for the 28s/18s ratio and RIN scores respectively.

Despite its low RNA quality scores, given the difficulty obtaining samples, I decided to process E27AM2P3 along with the other 24 samples to determine if the resultant quality of the array data was sufficient for inclusion.

5.3.3 Quality Control.

I generated the same quality control metrics described in Chapter 2.8.3.5 in both Partek GS and Expression Console for all 25 samples. Initially a PCA plot was generated including all samples (Fig. 5.3). Visualisation of the PCA plot shows one sample to have very different patterns of gene expression relative to the other samples. This sample is E27AM2P3, the C59X mutant which had low RNA quality scores. The other samples do not appear to separate along the first principle component but there is indication there is some separation along the second principle component, although importantly, they do not cluster by C59X genotype.

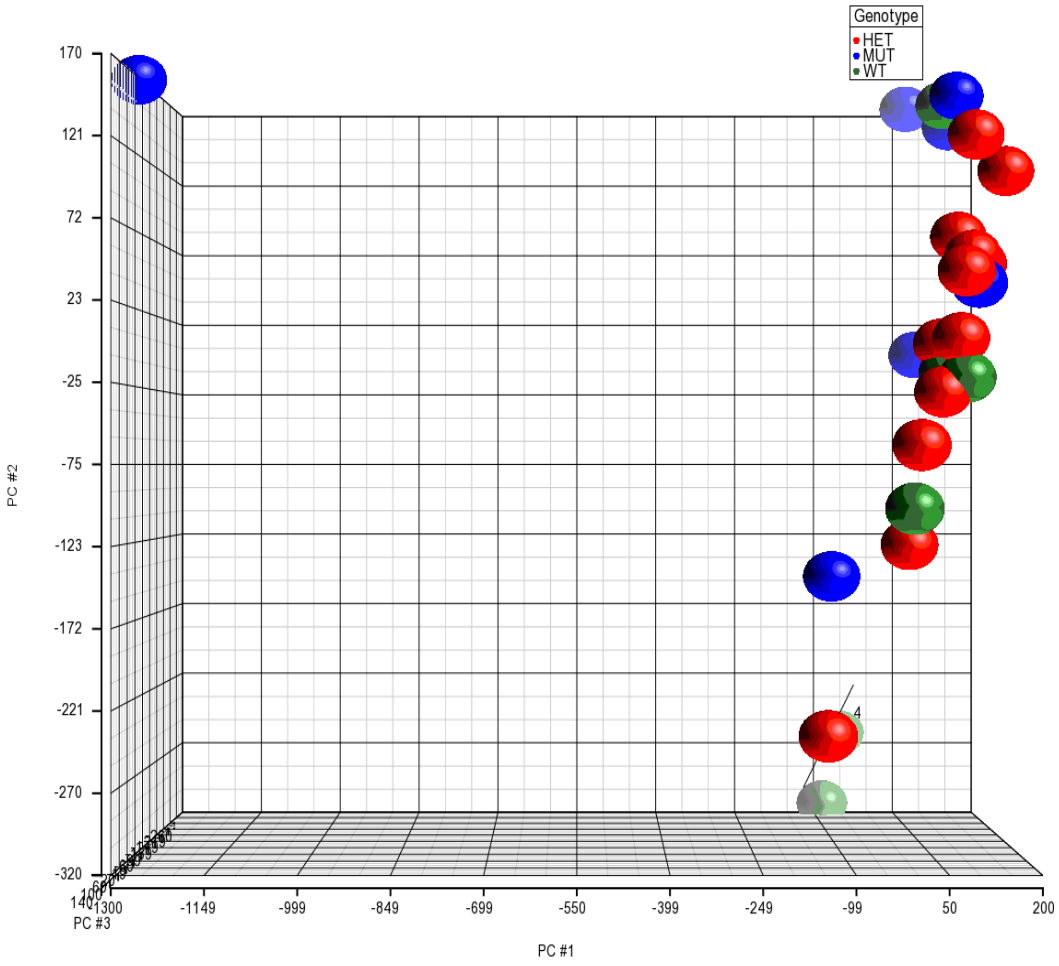


Figure 5.3. PCA of 25 Embryonic Samples. The gene expression patterns of the 25 samples is divided along the first principle component with 1 sample, found at the top left of the plot (Female, mutant), having very different expression patterns to the other 24 samples. The other samples appear to have similar patterns of global gene expression with the majority clustering in the top right of the plot.

The expression profile of each sample was plotted next (Fig. 5.4). Again, E27AM2P3 did not cluster with the other samples. To determine if the difference in expression pattern and distribution of signal intensity in E27AM2P3 may have arisen due to a reduced hybridisation efficiency, this metric was assessed next (Fig. 5.5). The expected rank order of signal intensities of 4 exogenous control genes was observed in all samples, but the expression pattern in sample E27AM2P3 is clearly different to that of the other samples. I then generated box plots of the log expression signal distribution (Fig. 5.6) to determine if preprocessing procedures corrected any of the variation observed in E27AM2P3. The mean and interquartile range of E27AM2P3 was lower relative to the other samples and the difference was not corrected by normalisation and summarisation procedures. A good way to determine if the general behaviour of a sample is different to the other samples is to plot the relative log expression signal (Chapter 2.8.3.5.4). From this plot it was apparent that the data arising from sample E27AM2P3 are different from those of all other samples.

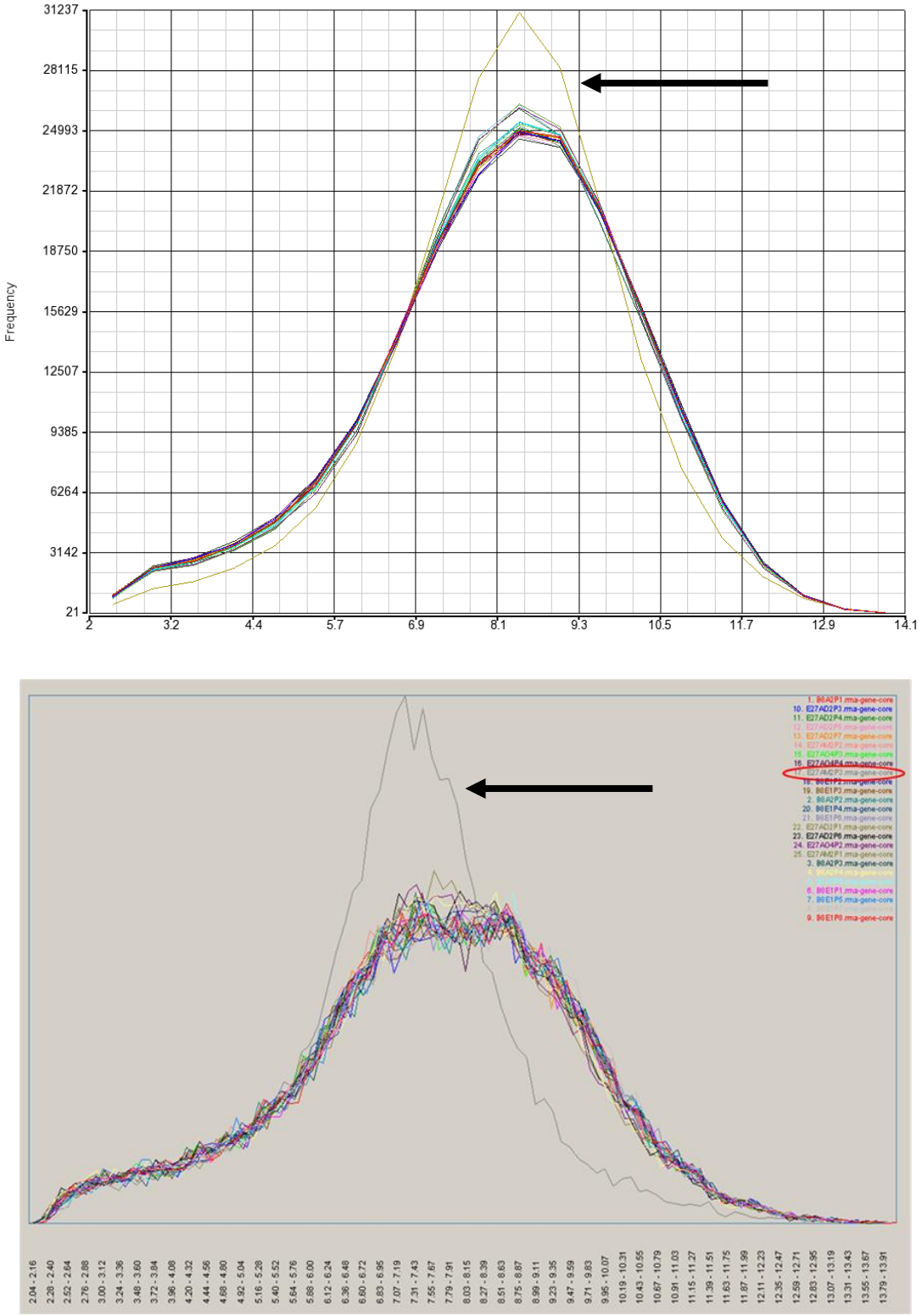


Figure 5.4. Distribution of Signal Intensity. The expression profile of all 25 samples was plotted in Partek GS (top) and Expression Console (below). The range of signal intensities are plotted along the x-axis with the frequency of each signal intensity plotted on the y-axis. One sample can be clearly distinguished from the others due to the difference in the distribution of signal intensities observable in both plots, this sample is E27AM2P3 (indicated by the arrow).

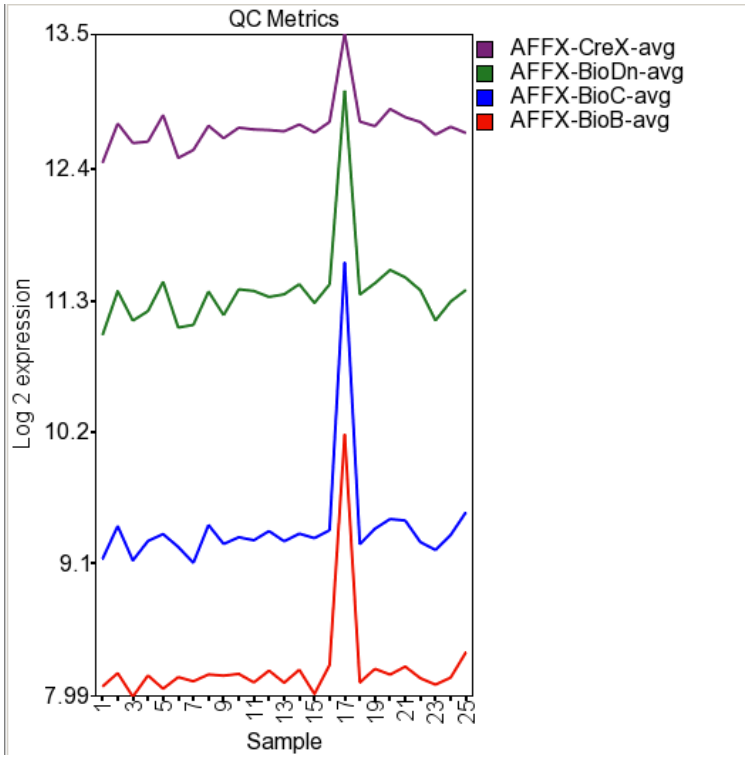


Figure 5.5 Hybridisation Efficiency in Embryonic C59X mice. 4 exogenous *Escherichia coli* genes included on the exon array were used to assess hybridisation efficiency. As each of the 4 genes are added at known concentrations, adequate hybridisation efficiency is observed from the following rank order of signal intensities from lowest to highest; BioB, BioC, BioD and cre. Whilst this rank order is observed in all samples, the log2 expression pattern of each of the 4 control genes is very different in one sample relative to the others and this sample (17) corresponds to sample E27AM2P3.

To formally assess if the sample should be removed, quantitative quality assessment was next performed using the 6 metrics described in Chapter 2.8.3.5. Samples were first considered all together (Table 5.3) and then after in groups separated by genotype (Table 5.4). Values greater than 2 standard deviations from the mean in any metric are considered outlying values, while samples with values greater than 2 standard deviations from the mean in 3 or more metrics are considered outlier samples. E27AM2P3 had values greater than 2 standard deviations from the mean in each of the 6 metrics and therefore was an egregious outlier. Based on these results there is a clear argument in favour of removing this sample from the study. Considering the samples stratified by genotype, two other samples were identified with quality values greater than 2 standard deviations from the mean; B8A2P4 and B8A2P5. B8A2P5 was only highlighted in a single metric so was not defined as an outlier. B8A2P4 had more than 3 metrics with values outside of the defined threshold and was therefore regarded as an

outlier. The same metrics were also generated in Affymetrix Expression Console (data not shown) which gave the same pattern of results.

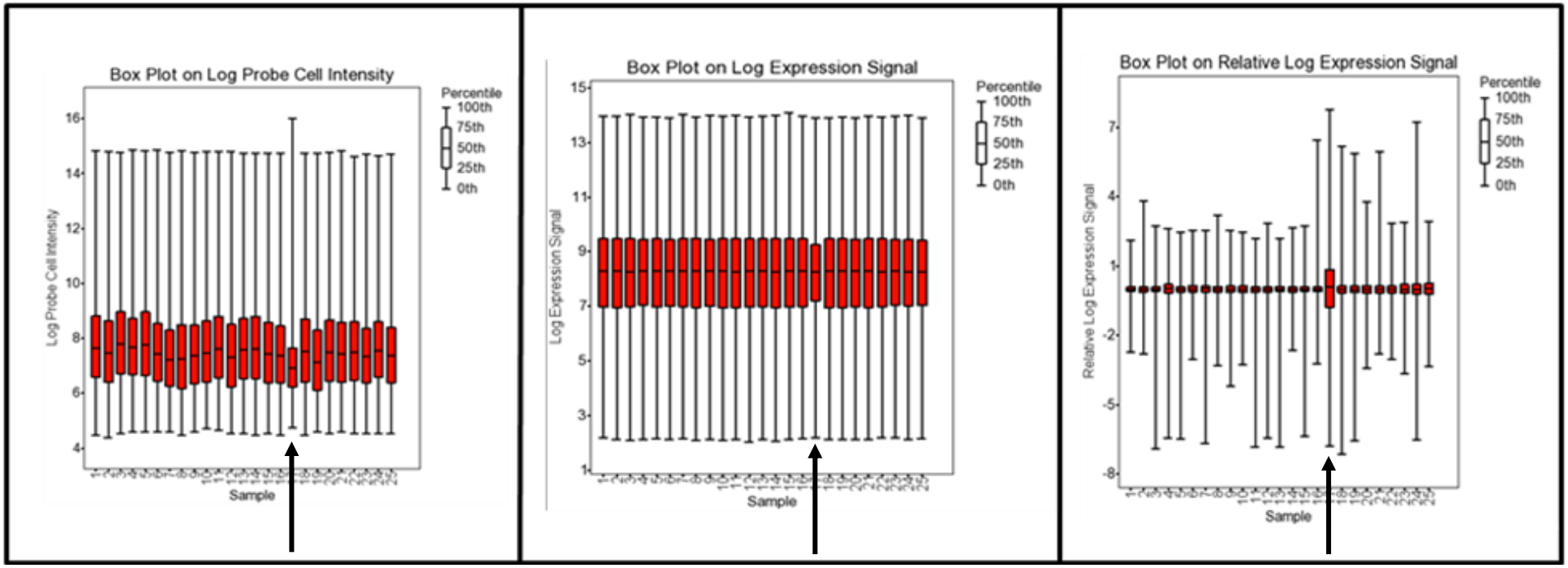


Figure 5.6 Distribution of Log Expression Signals. Sample E27AM2P3 has a reduced log probe cell intensity relative to the other samples (left) as indicated by the lower mean and inter quartile range (arrow) and this is not corrected by normalisation and summarisation procedures (central). This is indicative of a dim array, which would correlate with the lower RNA quality observed for this sample. From the relative log expression plot (right) it is clear that E27AM2P3 is behaving very differently to the other 24 samples.

Sample ID	Genotype	PM Mean	All Probeset MAD Residual Mean	Pos Control MAD Residual Mean	All Probeset RLE Mean	Pos Control RLE Mean	Pos vs. Neg AUC
1. B8A2P1.CEL	Het	453.28	0.13	0.10	0.13	0.10	0.87
2. B8A2P2.CEL	Mutant	424.61	0.14	0.10	0.15	0.11	0.87
3. B8A2P3.CEL	Wt	490.42	0.13	0.09	0.13	0.10	0.87
4. B8A2P4.CEL	Het	407.82	0.19	0.16	0.26	0.22	0.86
5. B8A2P5.CEL	Het	504.31	0.14	0.11	0.15	0.12	0.87
6. B8E1P1.CEL	Het	382.91	0.16	0.12	0.18	0.17	0.86
7. B8E1P5.CEL	Wt	333.17	0.18	0.13	0.19	0.16	0.87
8. B8E1P7.CEL	Het	392.05	0.17	0.12	0.18	0.13	0.87
9. B8E1P8.CEL	Het	368.41	0.16	0.12	0.18	0.15	0.87
10. E27AD2P3.CEL	Het	425.85	0.15	0.11	0.19	0.15	0.86
11. E27AD2P4.CEL	Het	457.78	0.14	0.10	0.15	0.12	0.86
12. E27AD2P5.CEL	Het	387.87	0.16	0.11	0.16	0.12	0.87
13. E27AD2P7.CEL	Wt	432.60	0.14	0.10	0.15	0.12	0.86
14. E27AM2P2.CEL	Het	454.74	0.14	0.11	0.16	0.12	0.87
15. E27AO4P3.CEL	Het	401.01	0.14	0.11	0.14	0.11	0.87
16. E27AO4P4.CEL	Het	366.23	0.14	0.10	0.14	0.10	0.87
17. E27AM2P3.CEL	Mutant	**214.34**	**0.43**	**0.36**	**0.99**	**0.98**	**0.79**
18. B8E1P2.CEL	Mutant	433.18	0.16	0.12	0.20	0.17	0.87
19. B8E1P3.CEL	Mutant	352.13	0.17	0.12	0.19	0.16	0.87
20. B8E1P4.CEL	Mutant	409.25	0.16	0.12	0.23	0.18	0.87
21. B8E1P6.CEL	Wt	401.05	0.15	0.11	0.20	0.17	0.87
22. E27AD2P1.CEL	Mutant	395.13	0.16	0.12	0.19	0.17	0.87
23. E27AD2P6.CEL	Mutant	329.21	0.20	0.15	0.25	0.21	0.87
24. E27AO4P2.CEL	Wt	383.60	0.18	0.14	0.26	0.24	0.85

25. E27AM2P1.CEL	Wt	343.60	0.18	0.15	0.28	0.30	0.85
Mean -SD*2		**279.43**	**0.05**	**0.02**	**-0.11**	**-0.16**	**0.83**
Mean +SD*2		**516.13**	**0.28**	**0.23**	**0.55**	**0.53**	**0.90**

Table 5.3. Quantitative Outlier Analysis in all E18.5 Samples combined. I generated values for 6 quality metrics in Partek GS which together assess the general performance of each of the chips. Following Affymetrix recommended guidelines, samples with values more than 2 standard deviations from the mean in any metric were highlighted (yellow). If any one sample was consistently highlighted in more than 3 metrics it was considered to be an outlier. From the above table it is clear that sample E27AM2P3 is an outlier as it exceeds the threshold in all 6 of the metrics analysed.

Sample ID	Genotype	PM Mean	All Probeset MAD Residual Mean	Pos Control MAD Residual Mean	All Probeset RLE Mean	Pos Control RLE Mean	Pos vs. Neg AUC
17. E27AM2P3.CEL	Mutant	214.34	**0.43**	**0.36**	**0.99**	**0.98**	**0.79**
18. B8E1P2.CEL	Mutant	433.18	0.16	0.12	0.20	0.17	0.87
19. B8E1P3.CEL	Mutant	352.13	0.17	0.12	0.19	0.16	0.87
2. B8A2P2.CEL	Mutant	424.61	0.14	0.10	0.15	0.11	0.87
20. B8E1P4.CEL	Mutant	409.25	0.16	0.12	0.23	0.18	0.87
22. E27AD2P1.CEL	Mutant	395.13	0.16	0.12	0.19	0.17	0.87
23. E27AD2P6.CEL	Mutant	329.21	0.20	0.15	0.25	0.21	0.87
Mean -SD*2		**212.19**	**0.00**	**-0.03**	**-0.28**	**-0.33**	**0.80**
Mean +SD*2		**518.62**	**0.40**	**0.33**	**0.91**	**0.90**	**0.92**

Sample ID	Genotype	PM Mean	All Probeset MAD Residual Mean	Pos Control MAD Residual Mean	All Probeset RLE Mean	Pos Control RLE Mean	Pos vs. Neg AUC
13. E27AD2P7.CEL	Wt	432.60	0.14	0.10	0.15	0.12	0.86
21. B8E1P6.CEL	Wt	401.05	0.15	0.11	0.20	0.17	0.87
24. E27AO4P2.CEL	Wt	383.60	0.18	0.14	0.26	0.24	0.85
25. E27AM2P1.CEL	Wt	343.60	0.18	0.15	0.28	0.30	0.85
3. B8A2P3.CEL	Wt	490.42	0.13	0.09	0.13	0.10	0.87
7. B8E1P5.CEL	Wt	333.17	0.18	0.13	0.19	0.16	0.87
Mean -SD*2		**280.39**	**0.11**	**0.08**	**0.09**	**0.03**	**0.84**
Mean +SD*2		**514.42**	**0.21**	**0.16**	**0.32**	**0.34**	**0.88**

Sample ID	Genotype	PM Mean	All Probeset MAD Residual Mean	Pos Control MAD Residual Mean	All Probeset RLE Mean	Pos Control RLE Mean	Pos vs. Neg AUC
1. B8A2P1.CEL	Het	453.28	0.13	0.10	0.13	0.10	0.87
10. E27AD2P3.CEL	Het	425.85	0.15	0.11	0.19	0.15	0.86
11. E27AD2P4.CEL	Het	457.78	0.14	0.10	0.15	0.12	0.86
12. E27AD2P5.CEL	Het	387.87	0.16	0.11	0.16	0.12	0.87
14. E27AM2P2.CEL	Het	454.74	0.14	0.11	0.16	0.12	0.87
15. E27AO4P3.CEL	Het	401.01	0.14	0.11	0.14	0.11	0.87
16. E27AO4P4.CEL	Het	366.23	0.14	0.10	0.14	0.10	0.87
4. B8A2P4.CEL	Het	407.82	**0.19**	**0.16**	**0.26**	**0.22**	0.86
5. B8A2P5.CEL	Het	**504.31**	0.14	0.11	0.15	0.12	0.87
6. B8E1P1.CEL	Het	382.91	0.16	0.12	0.18	0.17	0.86
8. B8E1P7.CEL	Het	392.05	0.17	0.12	0.18	0.13	0.87
9. B8E1P8.CEL	Het	368.41	0.16	0.12	0.18	0.15	0.87
Mean -SD*2		**331.54**	**0.12**	**0.08**	**0.10**	**0.06**	**0.86**
Mean +SD*2		**502.17**	**0.18**	**0.15**	**0.24**	**0.20**	**0.87**

Table 5.4. Quantitative Outlier Analysis of groups stratified by Genotype. When considering the same 25 samples by C59X genotype group (in Partek GS) E27AM2P3 was still an outlier. All wildtype samples had quality values within the specified thresholds. 2 heterozygote samples were highlighted as having at least one measure greater than two standard deviations from the mean, B8A2P4 and B8A2P5 (marked in yellow).

Regardless of the type of analysis, E27AM2P3 was a clear outlier a finding consistent with its relatively poor RNA quality. B8A2P4 was acceptable in the whole group analysis but had 4 outlier values in the stratified by genotype analysis. As it was not such an obvious outlier, and as the main analysis is focussed on comparing homozygotes, this sample was retained but was monitored in the downstream output as described later.

5.3.4 Deletion at the *Snca* Locus

That the mice used for backcrossing (C57BL/6JOlaHsd, Harlan) had a deletion at the *Snca* locus was established in Chapter 4. Whilst this may have had a confounding effect on the results generated on the $F3_i$ adult mouse data it was hypothesised that by F7 the mice would have 98.5% C57BL/6JOlaHsd background and so the deletion was likely to present in all embryonic samples. The exon array data were used to determine the expression of *Snca* in the embryonic mice. The raw expression at each probeset targeting the *Snca* gene was averaged across the genes and plotted for each of the 25 samples (Fig. 5.7). This revealed that two of the 25 mice did not have the deletion at the *Snca* locus. Both were excluded from further analysis to avoid confounding by *Snca* genotype. In addition to the outlier E27AM2P3, the exclusion of these two sample left 22 embryonic samples, 6 C59X homozygous mutants, 5 wildtypes and 11 C59X heterozygotes. The QC was repeated following the removal of the 3 samples. The Partek GS analysis is reported in Table 5.6. All samples considered together showed no outliers. Within C59X experimental group only B8A2P4 behaved slightly differently to other C59X heterozygote in 4 metrics. This sample was flagged previously and was retained and monitored for the same reasons described in 5.3.3.

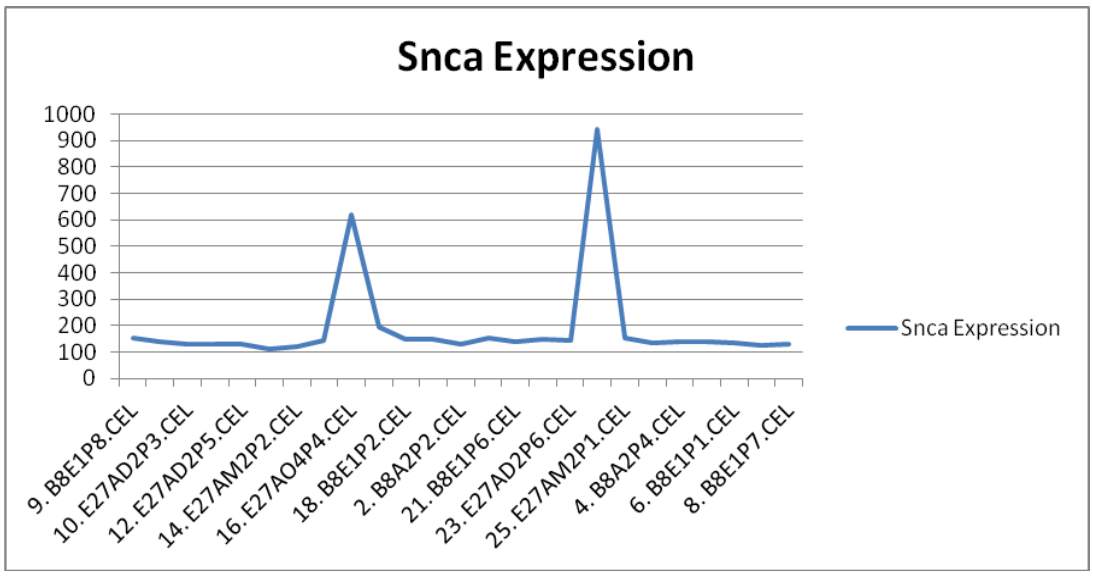


Figure 5.7 *Snca* Expression in E18.5 C59X Samples. Average raw expression values from the exon array across all probesets targeting *Snca* were plotted on the y-axis. The two peaks clearly indicate expression of the *Snca* gene in two of the samples, E27AO4P2 (female, wildtype) and E27AO4P4 (male, C59X heterozygote).

Sample ID	Genotype	PM Mean	All Probeset MAD Residual Mean	Pos Control MAD Residual Mean	All Probeset RLE Mean	Pos Control RLE Mean	Pos vs. Neg AUC
B8A2P1.CEL	Het	453.28	0.13	0.10	0.13	0.10	0.87
B8A2P2.CEL	Mutant	424.61	0.14	0.10	0.15	0.11	0.87
B8A2P3.CEL	Wt	490.42	0.13	0.09	0.13	0.10	0.87
B8A2P4.CEL	Het	407.82	0.19	0.16	0.26	0.22	0.86
B8A2P5.CEL	Het	504.31	0.14	0.11	0.15	0.12	0.87
B8E1P1.CEL	Het	382.91	0.16	0.12	0.18	0.17	0.86
B8E1P5.CEL	Wt	333.17	0.18	0.13	0.19	0.16	0.87
B8E1P7.CEL	Het	392.05	0.17	0.12	0.18	0.13	0.87
B8E1P8.CEL	Het	368.41	0.16	0.12	0.18	0.15	0.87
E27AD2P3.CEL	Het	425.85	0.15	0.11	0.19	0.15	0.86
E27AD2P4.CEL	Het	457.78	0.14	0.10	0.15	0.12	0.86
E27AD2P5.CEL	Het	387.87	0.16	0.11	0.16	0.12	0.87
E27AD2P7.CEL	Wt	432.60	0.14	0.10	0.15	0.12	0.86
E27AM2P2.CEL	Het	454.74	0.14	0.11	0.16	0.12	0.87
E27AO4P3.CEL	Het	401.01	0.14	0.11	0.14	0.11	0.87
B8E1P2.CEL	Mutant	433.18	0.16	0.12	0.20	0.17	0.87
B8E1P3.CEL	Mutant	352.13	0.17	0.12	0.19	0.16	0.87
B8E1P4.CEL	Mutant	409.25	0.16	0.12	0.23	0.18	0.87
B8E1P6.CEL	Wt	401.05	0.15	0.11	0.20	0.17	0.87
E27AD2P1.CEL	Mutant	395.13	0.16	0.12	0.19	0.17	0.87
E27AD2P6.CEL	Mutant	329.21	0.20	0.15	0.25	0.21	0.87
E27AM2P1.CEL	Wt	343.60	0.18	0.15	0.28	0.30	0.85
Mean-SD*2		**313.77**	**0.12**	**0.08**	**0.10**	**0.06**	**0.86**
Mean+SD*2		**502.62**	**0.19**	**0.15**	**0.27**	**0.25**	**0.88**

Sample ID	Genotype	PM Mean	All Probeset MAD Residual Mean	Pos Control MAD Residual Mean	All Probeset RLE Mean	Pos Control RLE Mean	Pos vs. Neg AUC
B8A2P1.CEL	Het	453.28	0.13	0.10	0.13	0.10	0.87
B8A2P4.CEL	Het	407.82	0.19	0.16	0.26	0.22	0.86
B8A2P5.CEL	Het	504.31	0.14	0.11	0.15	0.12	0.87
B8E1P1.CEL	Het	382.91	0.16	0.12	0.18	0.17	0.86
B8E1P7.CEL	Het	392.05	0.17	0.12	0.18	0.13	0.87
B8E1P8.CEL	Het	368.41	0.16	0.12	0.18	0.15	0.87
E27AD2P3.CEL	Het	425.85	0.15	0.11	0.19	0.15	0.86
E27AD2P4.CEL	Het	457.78	0.14	0.10	0.15	0.12	0.86
E27AD2P5.CEL	Het	387.87	0.16	0.11	0.16	0.12	0.87
E27AM2P2.CEL	Het	454.74	0.14	0.11	0.16	0.12	0.87
E27AO4P3.CEL	Het	401.01	0.14	0.11	0.14	0.11	0.87
Mean-SD*2		**338.46**	**0.12**	**0.08**	**0.10**	**0.07**	**0.86**
Mean+SD*2		**504.45**	**0.19**	**0.15**	**0.24**	**0.21**	**0.87**

Sample ID	Genotype	PM Mean	All Probeset MAD Residual Mean	Pos Control MAD Residual Mean	All Probeset RLE Mean	Pos Control RLE Mean	Pos vs. Neg AUC
B8E1P6.CEL	Wt	401.05	0.15	0.11	0.20	0.17	0.87
E27AM2P1.CEL	Wt	343.60	0.18	0.15	0.28	0.30	0.85
E27AD2P7.CEL	Wt	432.60	0.14	0.10	0.15	0.12	0.86
B8E1P5.CEL	Wt	333.17	0.18	0.13	0.19	0.16	0.87
B8A2P3.CEL	Wt	490.42	0.13	0.09	0.13	0.10	0.87
Mean-SD*2		**270.22**	**0.11**	**0.07**	**0.08**	**0.01**	**0.85**
Mean+SD*2		**530.12**	**0.20**	**0.16**	**0.31**	**0.33**	**0.88**

Sample ID	Genotype	PM Mean	All Probeset MAD Residual Mean	Pos Control MAD Residual Mean	All Probeset RLE Mean	Pos Control RLE Mean	Pos vs. Neg AUC
B8A2P2.CEL	Mutant	424.61	0.14	0.10	0.15	0.11	0.87
B8E1P2.CEL	Mutant	433.18	0.16	0.12	0.20	0.17	0.87
B8E1P3.CEL	Mutant	352.13	0.17	0.12	0.19	0.16	0.87
B8E1P4.CEL	Mutant	409.25	0.16	0.12	0.23	0.18	0.87
E27AD2P1.CEL	Mutant	395.13	0.16	0.12	0.19	0.17	0.87
E27AD2P6.CEL	Mutant	329.21	0.20	0.15	0.25	0.21	0.87
Mean-SD*2		**307.70**	**0.13**	**0.09**	**0.13**	**0.10**	**0.86**
Mean+SD*2		**473.46**	**0.20**	**0.15**	**0.27**	**0.23**	**0.88**

Table 5.6. QC following the removal of an outlier and *Snca* expressing Samples. When considering all 22 samples no outliers were observed. When considering the samples by C59X genotype only B8A2P4 had more than 3 metrics with values more than 2 standard deviations from the mean. This sample had been flagged previously for this reason.

5.3.6 Embryonic C59X Mutant vs Wildtype Analysis.

After exclusions, samples from 6 C59X homozygous mutants and from 5 wildtypes were analysed for differential expression and splicing.

5.3.6.1 Differential Gene expression.

I evaluated gene expression in the 6 embryonic C59X mutants relative to the 5 wildtypes using a 3-way ANOVA in Partek GS. Genotype, gender and scan date were all included as ANOVA factors. A total of 15,830 transcripts passed the filtering thresholds and were included in the analysis. The number of genes found to be significantly differentially ($P \leq 0.05$) expressed in the C59X mutants was 1346 (9%), almost double that expected by chance alone (Table 5.7) but none survived correction for multiple testing (as was observed in the adult data). Of the 6 robust differentially expressed genes between C59X mutants and wildtype in the adult data only *Npas4* was nominally significantly differentially expressed ($p = 0.046$) between embryonic C59X mutants and wildtype.

Unadjusted p value	Significant Genes	% of Genes	Expected by Chance
0.05	1346	8.50	792
0.01	225	1.42	158
0.001	15	0.09	16
0.0001	0	0	2
FDR 0.05 threshold	0	0	
Bonferroni Correction for 15830 tests	0	0	

Table 5.7 Genes differentially expressed in C59X embryonic mutants. Prior to correction for multiple testing, 1346 genes were significantly differentially expressed at $p \leq 0.05$ in the C59X mutants, however following either FDR or Bonferroni correction no genes were significant.

5.3.6.2 Differential Splicing.

I evaluated differential splicing using the alternative splice ANOVA in Partek GS. Genotype was added as the alternative splice factor (as in the adult C59X dataset) with both gender and scan date added as covariates. As before, probesets with a maximum log2 intensity <3 were excluded. The default option to retain probesets if significant differential splicing between the two groups at a p value ≤0.05 was observed, despite an

intensity below 3, was included. As observed previously in the adult data the number of significant differentially spliced genes at a nominal p value ≤0.05 was greater than the number differentially expressed (Table 5.8) with 11% (1808) of genes predicted to be differentially spliced in the C59X mutants. Using the FDR step up (Benjamini & Hochberg, 1995) set at 0.05, 172 genes were found to be significantly differentially spliced, of which 5% (~9 genes) would be predicted to be false positives. Using the very conservative Bonferroni correction, 30 genes (0.2%) were predicted to be differentially spliced in the C59X embryonic mutants. This appears to replicate the general pattern of the data observed in the adult C59X mutants, in that the number of differential splice predictions is greater and more robust than the differential expression predictions. When considering the splicing data the numbers of predicted differential splice events are greater than expected by chance alone at the more stringent statistical thresholds.

Unadjusted p value	Significant Genes	% of Genes	Expected by chance
P≤0.05	1808	11.42	792
P≤0.01	721	4.55	158
P≤0.001	214	1.35	16
P≤0.0001	104	0.66	2
FDR (0.05 threshold)	172	1.09	
Bonferroni Correction for 15830 tests	30	0.19	

Table 5.8. Differentially Spliced Genes. A total of 1808 genes were significantly differentially spliced in the embryonic C59X mutants at p≤0.05. 172 of these genes were still significant following an FDR correction using the 0.05 threshold. With a stringent Bonferroni correction for the 15,830 tests, 30 genes were significantly differentially spliced.

5.3.7 Overlap between the findings from the Embryonic and Adult studies.

The validity of the adult data came into question following the discovery that the C57BL/6JHsdOla strain harboured a mutation at the *Snca* locus which at $F3_i$ correlated with C59X genotype, *Snca* being expressed in all but one C59X mutant and deleted in the majority of the wildtypes. Prior to the discovery of the *Snca* deletion, my intention had been to investigate if there was developmental specificity between mutant and wild type animals with respect to differences in expression and splicing. However, due to the

confounding *Snca* deletion, differences between the results observed in the adult and embryonic might simply reflect the confounding effect of *Snca*. The primary use therefore of the comparison between adult and embryonic data was therefore to test the validity of the adult data by determining if any of the embryonic expression or splice changes replicated in the adult results. This is clearly not an ideal replication experiment given splice events are known to be developmentally regulated, and therefore true differences might be expected between the results obtained in the embryonic and adult studies. Nevertheless, results consistent across the studies would suggest the adult data may contain true positives unrelated to the *Scna* confound. I did consider whether in the adult data, the *Snca* deletion could be statistically adjusted for but due to the high degree of correlation between *Snca* deletion and the C59X genotype, covarying for *Snca* would essentially remove any differences between the strains.

To assess the degree of overlap between the embryonic and adult data, I compared if the observed number of genes attaining significance thresholds for each of the expression and splicing analyses in both datasets was greater than the number expected by chance (Table 5.9 & 5.10). The adult data set used was that of all 16 samples (male and female) combined with gender covaried.

In the embryonic data, significant genes were defined as those attaining a range of p value thresholds ($p \leq 0.05$, $p \leq 0.01$, $p \leq 0.001$ and $p \leq 0.0001$). Based on the number of genes surpassing each threshold in the embryonic data, the number of genes expected by chance to attain significance at $p \leq 0.05$ in the adult brain analysis were estimated using 2 x 2 contingency tables and the χ^2 test (method described in Chapter 3.2). The results for differential expression and splicing are summarised in tables 5.9 and 5.10 respectively.

For both differential expression and splice changes, among genes attaining each of the thresholds in the embryonic experiment, more also showed significant differences between mutant and wildtype in the analysis of the samples from adult brain than expected by chance. This data suggests that at least some of the differential expression and splice predictions in the adult data are neither due to chance nor to differences in *Snca* expression in the C59X mutant and wildtype groups.

Significance Threshold	OR	Significance of Overlap
P≤0.05	1.25	p=0.021
p≤0.01	1.79	p=0.007
p≤0.001	7.13	p=0.002
p≤0.0001	No genes significant	

Table 5.9 Differential Expression changes common to Embryonic and Adult C59X Mutants. OR is the odds ratio that genes nominally (≤0.05) significantly differentially expressed in the embryonic mutants will be nominally significant in the adult mutants conditional on the gene being significantly differentially expressed at the indicated P value threshold in the embryonic expression data.

Significance Threshold	OR	Significance of Overlap
p≤0.05	1.51	p<0.01
p≤0.01	2.00	p<0.01
p≤0.001	4.00	p<0.01
p≤0.0001	5.28	p<0.01

Table 5.10 Differential Splicing changes common to Embryonic and Adult C59X Mutants. OR is the odds ratio that genes nominally (≤0.05) significantly differentially spliced in the embryonic mutants will be nominally significant in the adult mutants conditional on the gene being significantly differentially spliced at the indicated P value threshold in the embryonic splicing data.

5.3.8 Alternative Algorithms.

To determine how robust the embryonic data were the data were analysed using alternative software packages which offer different filtering and statistical methods.

5.3.8.1 easyExon.

Using the easyExon, which unlike the other approaches, requires an accompanying fold change of greater than or equal to 1.5 fold only 18 genes were nominally significantly differentially expressed (MiDAS $p \leq 0.05$) and 345 genes nominally differentially spliced (MiDAS $p \leq 0.05$).

5.3.8.2 AltAnalyze.

Using the MiDAS algorithm in AltAnalyze, 1407 genes were differentially expressed (unadjusted $p<0.05$) which is similar to that in Partek GS (1346). Including a fold change filter of greater than or equal to 1.5 to make the analysis analogous to easyExon reduced this number to 38 genes. Thus many of the significant expression changes have small fold changes, and this explains the major numerical differences between easyExon and Partek GS. Using the FIRMA algorithm, 489 genes were predicted to be differentially spliced in AltAnalyze, similar to that observed in easyExon (345), but considerably less than the number observed in Partek GS (1808 genes). This may reflect the more stringent intensity filters used in easyExon and AltAnalyze to remove signals close to the background. Both use the DABG algorithm. The DABG p value represents the likelihood that the intensity value of a particular probe set is part of the background (null) distribution (Della et al., 2008). Probesets with an average DABG p value less than or equal to 0.05 in a biological group were excluded in AltAnalyze and probesets with a DABG p value less than or equal to 0.05 in half the samples were excluded in easyExon. In Partek GS probesets with a log2 intensity <3 were excluded. The larger number of results in Partek GS may therefore reflect false positives due to low expressing probesets close to the background signal being included in splicing calculations or that the smaller number of results in AltAnalyze and easyExon reflect a larger number of false negatives.

5.3.8.3. Differential Expression and Splicing Results Common to Partek GS, AltAnalyze and easyExon.

To determine how robust the expression and splicing data were the overlap between Partek GS, AltAnalyze and easyExon was established.

5.3.8.3.1. Expression:

To be included as significant in each of the methods, genes were required to have significant differential expression in C59X mutants at an unadjusted p value ≤0.05 and with an accompanying fold change of greater than or equal to 1.5. (Fig 5.8) as used in the adult analyses.

With these criteria in place a single gene showed consistent expression changes in AltAnalyze, easyExon and Partek GS (Fig. 5.8). That gene was *Osteoglycin* (*Ogn*), (Partek p = 0.0025; AltAnalyze p = 0.0032; easyExon p = 0.0032). *Ogn* was not significant in the adult data (Fig. 5.9). *Ogn* is a leucine-rich proteoglycan (Iozzo., 1999; Kukita et al., 1990) forming part of the brain exctracellular matrix. The observation of only a single gene showing consistent expression changes is not suggestive of a role for *Zfp804a* in the direct regulation of gene expression. The fact that only a single gene surpasses a threshold of $p \leq 0.05$ and a fold change greater than or equal to 1.5 suggests this requirement is excessively stringent. Under a null model 1 in 20 genes would be expected to be differentially expressed by chance (~800 genes) and under this same model finding only 1 that overlaps suggests the approach is very insensitive. When considering the results of Partek and AltAnalyze using just a p value threshold of less than or equal to 0.05 ($p \leq 0.05$) then 281 genes overlap, one of which is *Npas4* which was a robust differential expression candidate in the adult dataset. When reducing the stringency of the overlap threshold the proportion of genes is greater than the proportion you would expect by chance.

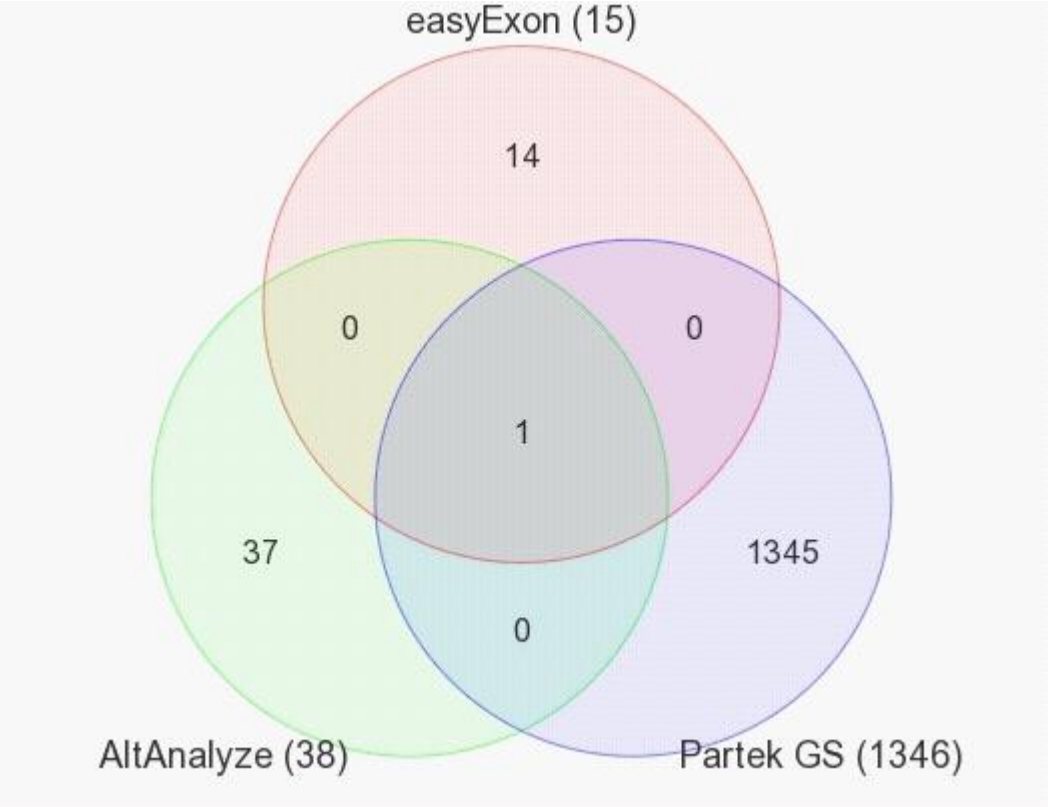


Figure 5.8 The number of Differential Expression results that replicate across Partek GS, easyExon and AltAnalyze. Only one gene (*Ogn*) is predicted to be differentially expressed in C59X mutants in all 3 programmes.

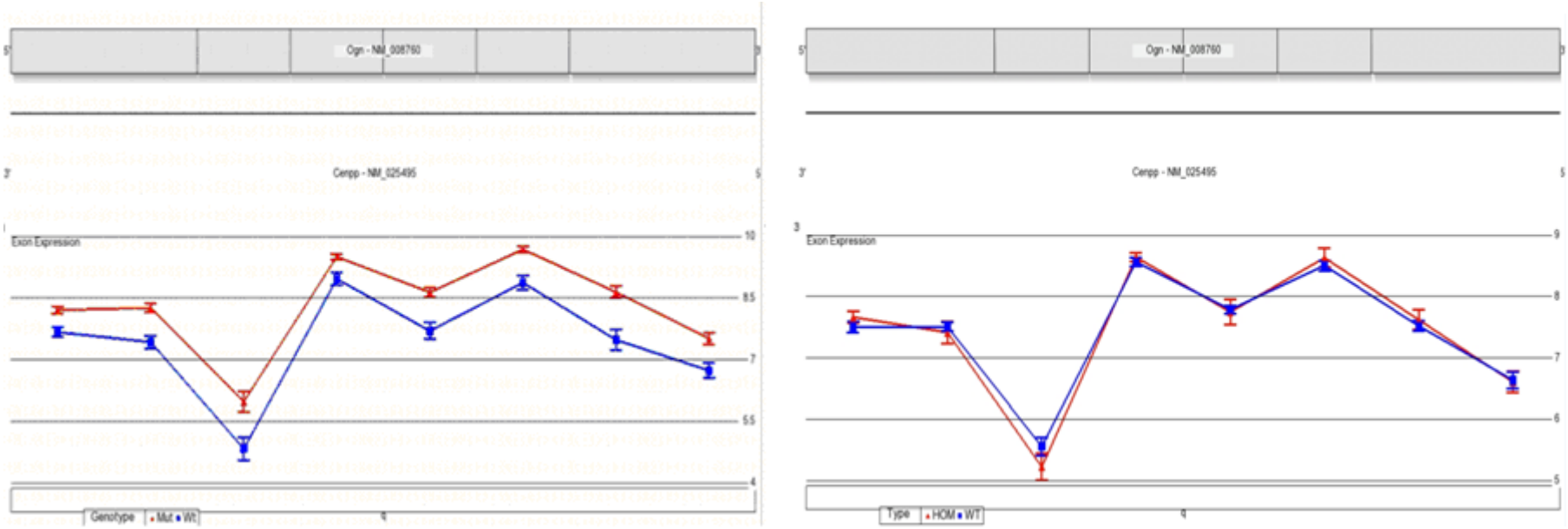


Figure 5.9 Differential Expression of *Osteoglycin* in Embryonic C59X Mutants. *Osteoglycin* was upregulated in embryonic C59X mutants (left). This expression change was not apparent in adult C59X mutants (combined sample) ($p = 0.64$) (right) and may reflect the developmental regulation of this gene. **N.B.** C59X homozygotes are represented in the above plots by a red line and wildtypes by a blue line.

5.3.8.3.2. Splicing:

Considering the alternative splicing results, 62 genes were predicted to be significantly differentially spliced across all three algorithms (Fig. 5.10). Of the genes significantly differentially spliced in C59X mutants relative to wildtypes in easyExon and AltAnalyze a large proportion were also significant in at least one other software package (185/350 for easyExon and 216/467 in AltAnalyze). Of these 62 genes, 20 had been significant after using a FDR threshold of 0.05 in Partek GS (Table 5.11).

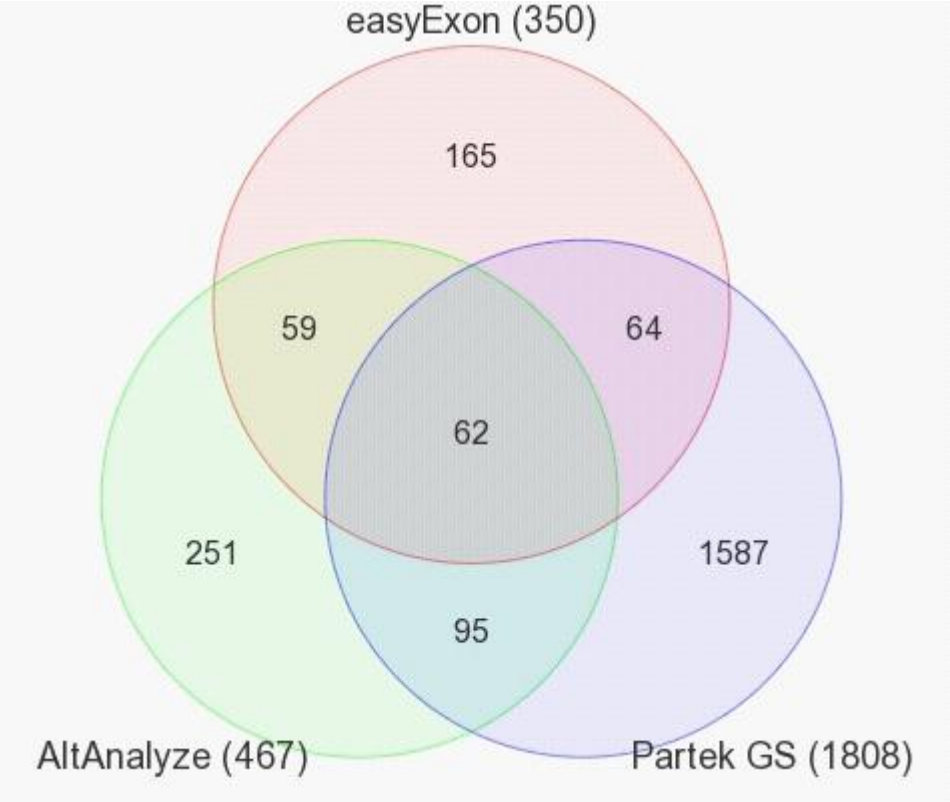


Figure 5.10 The number of Differential Splice results that replicate across Partek GS, easyExon and AltAnalyze. 62 genes are predicted to be differentially spliced by all 3 algorithms.

Of these 20 genes, 6 showed significant differences (at $P \leq 0.05$) in the analysis of adult samples of which no less than 4 were among the 13 most robust differential splice predictions in the adult data (defined as being significant in male and female datasets separately and being significant with all three software algorithms, see chapter 3 section 3) However, as shown in chapter 4, section 3.6, the apparent differential splicing at two

of these genes, *Slc39a13* and *Ssfa2* arises from probesets which span sequences at which the two strains, C3H/HeJ or Balb/c (the ENU strain), and C57BL/6JHsdOla (the strain used for backcrossing) differ by a single base with the potential to impact on probe binding and generate a false prediction of differential splicing. Also of note, of these top 20 genes, 7 mapped to chromosome 2, the same chromosome as *Zfp804a*, of which 4 mapped within 15 Mb of *Zfp804a* (Table 5.11).

This led me to the hypothesis that like *Slc39a13* and *Ssfa2*, the other robust splicing differences at genes on chromosome 2 might also reflect background strain specific sequence variants affecting probe binding which due to genetic linkage to the *Zfp804a* locus, still had not been randomly segregated with respect to the C59X mutation among offspring. To test this hypothesis the sequences targeted by the other 2 probesets predicted to be differentially spliced (*Fam171b* and *Prdm4)* out of the 4 genes that had shown robust effects in both adult and embryonic analyses were checked for dbSNPs. The sequence corresponding to the probeset predicted to be differentially spliced in *Fam171b* had relatively lower expression in the C59X mutants and had sequence variants that were present in adult mutant sequences specifically. The sequence corresponding to the probeset predicted to be differentially spliced in *Prdm4* was highly expressed in the mutant relative to the wildtype and there was no evidence of polymorphism in the sequence. Interestingly *Slc39a13, Ssfa2* and *Fam171b* are all on chromosome 2 where as *Prdm4* is on chromosome 10.

In the data I generated using Partek GS, 29 genes had differential splicing effects that were Bonferroni significant in the adult (combined male female) analysis and 30 genes in the embryonic dataset. Although the proportion of genes with these effects was similar in each, only 6 genes were common to both analyses (Table 5.12). Three of the six (*Fam171b, Slc29a13* and *Ssfa2*) were suspected to be attributable to C59X mutant strain specific variants and are found on chromosome 2. *Tcp11l1* and *Zc3h15* are also on chromosome 2 and so there was the possibility that a strain specific variant in the target sequence was causing a false positive result and it was these false positives which were responsible for the common results between embryonic and adult C59X mutants. I found both *Zc3h15* and *Tcp11l1* to contain dbSNPs in the sequence targeted by the probeset that was predicted to be differentially spliced. The exon targeted by the probeset was downregulated in the C59X mutants in both cases. A bias for genes

predicted to be highly significantly differentially spliced in both embryonic and adult C59X mutants was found on chromosome 2. This could indicate a region of chromosome 2 that was being inherited with the C59X mutation which resulted in additional C59X mutant strain specific variants (This is assessed in detail in chapter 6 section 3.1).

Gene Symbol	Chromosome	Start Position	Stop Position	Differential Splice P value (Nominal p<0.05)	FDR (0.05 threshold)	Adult Splice p value
Inpp4a	1	37356703	37476203	8.13x10-5	0.01	0.89
Cds2	2	132088919	132137786	7.23x10-6	0.003	0.13
ext2	2	93535349	93662754	5.79x10-5	0.01	0.66
Fam171b	2	83652803	83723677	1.34x10-14	1.06x10-10	5.16x10-18*
Itih2	2	10016224	10089270	1.10x10-6	8.74x10-4	1.00
Pla2g4e	2	119992148	120071314	3.33x10-6	0.002	4.23x10-4
Slc39a13	2	90901953	90928948	1.95x10-10	4.40x10-7	3.74x10-17*
Ssfa2	2	79475519	79513499	2.03x10-6	1.27x10-3	2.54x10-27*
Nomo1	7	53289086	53344037	7.79x10-9	1.23x10-5	0.72
Xab2	8	3608421	3621314	2.36x10-4	0.03	0.06
Baz2a	10	127528233	127567216	5.97x10-5	0.01	0.99
Myh10	11	68505007	68630180	8.21x10-7	6.84x10-4	0.90
Trim37	11	86940579	87064356	2.24x10-5	0.006	0.90
Dync1h1	12	111839631	111905126	6.57x10-9	1.16x10-5	0.04
Atp8a2	14	60266382	60816016	1.03x10-4	0.02	0.99
Prdm4	14	70162232	70237257	2.83x10-6	1.55x10-3	8.67x10-20*
Arf3	15	98565314	98593635	4.12x10-6	0.00186	0.86
Serpind1	16	17331484	17343665	5.91x10-5	0.01	0.51
Fbn2	18	58168277	58392191	5.84x10-5	0.01	0.98
Huwe1	X	148235370	148369956	6.46x10-6	0.002	1.00

Table 5.11 Overlapping Genes Significant After Multiple Test Correction. Of the 62 genes predicted by AltAnalyze, easyExon and Partek GS to be differentially spliced in the embryonic C59X mutants 20 were significant in Partek GS following an FDR correction set at 0.05. 6 Genes were significant in the embryonic data in Partek GS, easyExon and AltAnalyze and were also significant in the adult data at a nominal p value of <0.05. *4 genes were among the 13 most robust differential splice predictions in the adult data (defined as being significant in male and female datasets separately and being significant with all three software algorithms) and were also significant in all 3 programmes in the embryonic dataset.

Gene Symbol	Cytoband	Adult Splice p value (p<0.05)	Bonferroni Correction for 15833 Tests	E18.5 Splice p value (p<0.05)	Bonferroni Correction for 15830 Tests
Ssfa2	2qC3	2.53x10-27	4.01x10-23	2.03x10-6	0.03
Prdm4	14qD2	8.67x10-20	1.37x10-15	2.83x10-6	0.04
Tcp11l1	2qE2	1.80x10-19	2.85x10-15	1.48x10-10	2.34x10-6
Fam171b	2qD	5.16x10-18	8.18x10-14	1.34x10-14	2.13x10-10
Slc39a13	2qE1	3.75x10-17	5.93x10-13	1.94x10-10	3.08x10-6
Zc3h15	2qD	1.28x10-14	2.03x10	1.40x10-8	2.22x10-4

Table 5.12 Genes with Bonferroni Significant Differential Splicing in embryonic and Adult C59X mutants. The 6 genes in the table above were all predicted to be differentially spliced in embryonic and adult C59X mutants following a Bonferroni correction. 5 of the 6 genes are found on chromosome 2.

5.4 Discussion.

I determined the effects of the C59X mutation on expression and splicing in embryonic mice from the F7 generation which were expected to have at least 98.5% of the C57BL/6HsdOla genome. Whole transcriptome expression and splicing were investigated in brain tissue derived from E18.5 embryos. Embryonic tissue was chosen due to evidence for a neurodevelopmental origin of schizophrenia as well as the growing evidence in support for substantial alternative splicing prevalent during development. The aim was to identify expression and splice changes that occur during this developmental period which in mouse, is thought to be roughly comparable to the second trimester of human pregnancy when it is thought impaired or altered development influences predisposition to schizophrenia.

Expression of *Snca* in the ENU background strain was still observed in 2 of the 25 embryonic samples. Although deletion of *Snca* did not strongly correlate with C59X genotype, I excluded these 2 samples to rule out confounding effects expression of this gene might have on the results. The *Snca* deletion has been reported to have some phenotypic effects (Oksman et al., 2006) and may alter the expression of other genes, though evidence of this is as yet lacking (Specht & Shoepfer, 2001).

Expression and splice differences between mutant and wild type mice followed the same general pattern as observed in adult results. The proportion of differentially expressed genes was for adult and embryonic samples respectively 7% and 8%. Only one gene, *Ogn* was found to be differentially expressed in the embryonic C59X mutants using 3 different software tools. Leucine-rich proteoglycans such as *Ogn* are thought to have roles in neurite outgrowth, ECM assembly and cell adhesion (Ruoslahti, 1996). Extracellular matrix proteins such as *Ogn* (Jung et al., 2012), whilst having structurally supportive functions in the brain, are also thought to be important to the architecture of the brain and contribute to plasticity (Bonneh-Barkay & Wiley, 2009). The expression of *Ogn* is reduced in the amygdala of chronic immobilisation stress (CIS)-induced depressed mice (Jung et al., 2012). The authors suggest this may affect plasticity in the amygdala and neural circuits involved in stress which could have implications in psychiatric disorders (Jung et al., 2012). As the expression difference between C59X mutant and wildtype is not found in the equivalent analysis of the adult samples it may represent a developmentally specific expression difference between embryonic and adult C59X mutants, which could reflect the importance of this gene in development.

Interestingly, known functions of Ogn include the assembly of the extracellular matrix and cell adhesion. Extracellular matrix and cell adhesion pathways are enriched for developmentally regulated differentially spliced genes in embryonic mice (Revil et al., 2010). The absence of robust expression differences between C59X mutants and wildtypes does not appear to support a role for Zfp804a in the direct regulation of gene expression. The identification of *Ogn* is however interesting with regards to the developmental hypothesis of schizophrenia. *Ogn* was identified in all software at a nominally significant p value (p=0.003). Despite it not being one of the most significant expression differences the geneview shows consistent upregulation across the transcript in the C59X mutants with a fold change of greater than or equal to 1.5. This indicates that the observed differential expression of the *Ogn* gene between C59X mutants and wildtypes is a real difference in the sample not attributable to the software.

The percentage of genes predicted to be significantly differential spliced was slightly higher in the analysis of embryonic samples than adult (11% compared to ~6% in the adult data) but when applying a Bonferroni correction ~30 genes remained significant in both datasets. However, when considering the specific overlap of genes in this Bonferroni significant set and also from the overlap between Partek GS, easyExon and AltAnalyze it would appear that the consensus between the adult and embryonic results may, at least in part, be due to the effects of C59X mutant strain specific variants which are in genetic linkage with the C59X mutation. As this observation has been made in both adult and embryonic datasets, it is difficult to establish the proportion of true splice events which may be functionally interesting until the linked region has been determined and the false positives excluded. Further investigation of the alternative splicing results on chromosome 2, and indeed other chromosomes is therefore necessary to establish the extent of this issue. This analysis is presented in detail in the next chapter. As the observation of mutant specific variants was in genes on chromosome 2 it is likely these variants were being inherited along with the C59X mutation and therefore sequence variants specific to the mutants would not be expected in genes on other chromosomes.

When considering the splicing results in embryonic C59X mutants the number of significant results was much greater in Partek GS compared to easyExon and AltAnalyze and this may reflect the intensity filter being used in Partek GS. The default intensity filter in Partek GS may result in background signal being used to calculate differential splicing; resulting in a large number of false positives or alternatively the

number of false negatives in the easyExon and AltAnalyze may be greater due to very stringent filters being used. To ensure the distinction between true expression and background signal in Partek GS it may be necessary to undertake the analysis with more stringent intensity filters.

More in depth analysis of differential splicing taking into account inherited SNPs and intensity thresholds is necessary to ensure more accurate determination of differential splicing results which when taken forward to pathway and other downstream analyses will provide insight into sets of genes and their functions which are relevant to the pathophysiology of schizophrenia. Downstream analysis carried out on these results prior to assessing the apparent enrichment of significant splice results on chromosome 2 and the intensity cut offs would result in the identification of pathways that would not truly represent the effects of the C59X mutation and the function of Zfp804a. Therefore before conducting any further functional analyses the technical issues arising from using the C59X mouse model were further explored.

Chapter 6. Technical Artefacts

6.1 Introduction

Microarrays have been in wide use for many years in expression studies. Certain issues which have arisen from this study appear to be common pitfalls of such technology particularly when using mouse models in which backcrossing of a specific mutation is carried out on a different strain (Gajovic et al., 2006). In an attempt to address some of these issues in this chapter, I explore the technical artefacts in such experiments which can lead to false positives and where possible, address these with further analyses.

Sequence variants affecting cDNA-probe binding: The RNAseq experiment (chapter 4) identified the potential issue of linked mutations in sequences targeted by probesets predicted to be differentially spliced between C59X mutants and wildtypes. This problem is not well documented in mouse literature. The microarray technique is based on cDNA hybridising to a probe on the array with its complementary sequence. A polymorphism in the cDNA sequence that influences complementarity may reduce the hybridisation efficiency, resulting in an altered expression measure at the probeset. This is a problem if the polymorphism affecting hybridisation is specific to one of the experimental groups being studied as this could lead to false positive splicing results.

The failure of some of the most significant splice changes from the adult exon array to replicate in the RNA sequencing data brought up the possibility of a technical artefact leading to false positive splice calls in the array experiment. After the $F3_i$ experiment, I assumed this to be attributable to the mice having ~96% congenicity to the C57BL/6JHsdOla background strain, leaving the possibility that additional ENU induced mutations might still be present. However since the same genes showed highly significant changes in the embryonic data as well, this raised the issue that these mutations were in genetic linkage (i.e. these mutations had not been separated from the C59X mutation by a recombination event) with the C59X mutation since they were retained despite the now 98.5% congenicity to the C57BL/6JHsdOla strain. Based on preliminary analyses, significant differentially spliced genes around the *Zfp804a* locus and on chromosome 2 were investigated to determine if differential splicing correlated with the presence of dbSNPs in the associated probesets.

Impact of intensity filter: The use of relatively stringent intensity filters removes probesets corresponding to sequences with low expression. Low intensity signals in

probesets when normalised to total gene expression have the potential to be misinterpreted as differential splicing, so aggressive filtering for intensity can be expected to reduce the false positive rate. The optimal choice of intensity filters is dependent upon the aim of the experiment (Whistler et al., 2010). If the aim is to minimize false negative findings, or identify novel splicing events, the filter needs to be less stringent but the trade off for higher sensitivity is lower specificity, that is a higher false positive rate. The first intensity filter used was the Partek GS default which is a maximum log2 intensity <3 (except when that probeset has significant differential expression $p \leq 0.05$). This threshold is recommended in the Partek Manual to prevent the exclusion of true positives and is used elsewhere (Tian et al., 2010). This threshold corresponds to fairly low expression, and as a result, probesets corresponding to sequences that are not expressed won't be included. Others report that a more stringent log2 intensity > 6 represents reliable expression (Zhang et al., 2008), so I investigated the effects of increasing stringency up to this level.

The use of a log2 intensity filter of 3 in Partek GS resulted in the prediction of many more differential splicing events than easyExon and AltAnalyze which both filter probesets using the DABG p value. The detection above p value represents the likelihood that the intensity value of a particular probe set is part of the background (null) distribution (Della et al., 2008). In a paper by Whistler et al., (2010) they observed a similar finding where the percentage of probesets removed from the analysis was double when using the DABG ($p \leq 0.05$) over a log2 intensity of 3. With the number of genes predicted to be differentially spliced almost 3 times as many when using the log2 intensity <3 filter (Whistler et al., 2010).

Impact of gene size: Alternative splicing is thought to affect ~ 70% (Johnson et al., 2003) of mammalian genes, but this is known to increase to ~97% in multi-exon genes, with a linear increase in the number of alternative splicing events per gene as the number of exons increases (Pan et al., 2008). When considering the results of exon arrays, increased splicing in large, multi-exon genes may reflect a genuine biological finding, but the number of spurious apparent differential splice events could also increase in genes with a larger number of exons as more probes implies a greater potential for an impact of chance fluctuations. Both true biological and chance effects would imply a relationship between the probability of finding a significant differential splicing effect and gene size. To determine if this relationship was found in the C59X mice the embryonic splicing data were assessed.

After the above potential artefacts were assessed, I reappraised all the previously presented data from Chapters 3 and 5.

6.2 Methods

6.2.1. Linked Sequence variants affecting cDNA-probe binding.

To investigate potential effects of strain specific alleles in genetic linkage with the C59X mutation, I concentrated on the splicing data. Due to the way in which total expression and splicing changes are calculated, alleles affecting hybridization are expected to have a greater impact on splicing data. This is because splicing is determined by comparing expression at a single probeset relative to expression across the transcript whereas expression changes are based on average expression across all probesets in a transcript and are therefore less likely to be effected by a SNP in a single probe.

The hypothesis investigated is that C3H/HeJ or Balb/c strain alleles within sequences representing probesets that generate apparent differential splicing effects remain associated with the C59X mutation on chromosome 2 because of genetic linkage. To test this hypothesis, using the embryonic data, I determined the 50 most significant splicing events to identify if there was an enrichment of results around the *Zfp804a* locus on chromosome 2. I then took the 17 most significantly differentially spliced genes on chromosome 2 and the 50 most significantly differentially spliced genes elsewhere in the genome (17 on chromosome 2 as these had an equivalent p value level ($2x10^{-5}$) to the 50 non chromosome 2 genes) and in each identified the probeset indicative of differential splicing (Fig. 6.1). The probeset sequences were then identified using NetAffx (an Affymetrix tool for finding annotation and design information for GeneChip® arrays. http://www.affymetrix.com/analysis/index.affx) and the presence of a dbSNP (dbSNP 128) using UCSC. If a dbSNP was found within the sequence targeted by a probeset, the corresponding region was viewed in the adult mice RNAseq data (Chapter 4) using the Integrative Genomics Viewer (IGV) to confirm it distinguished C59X mutants from wildtypes.

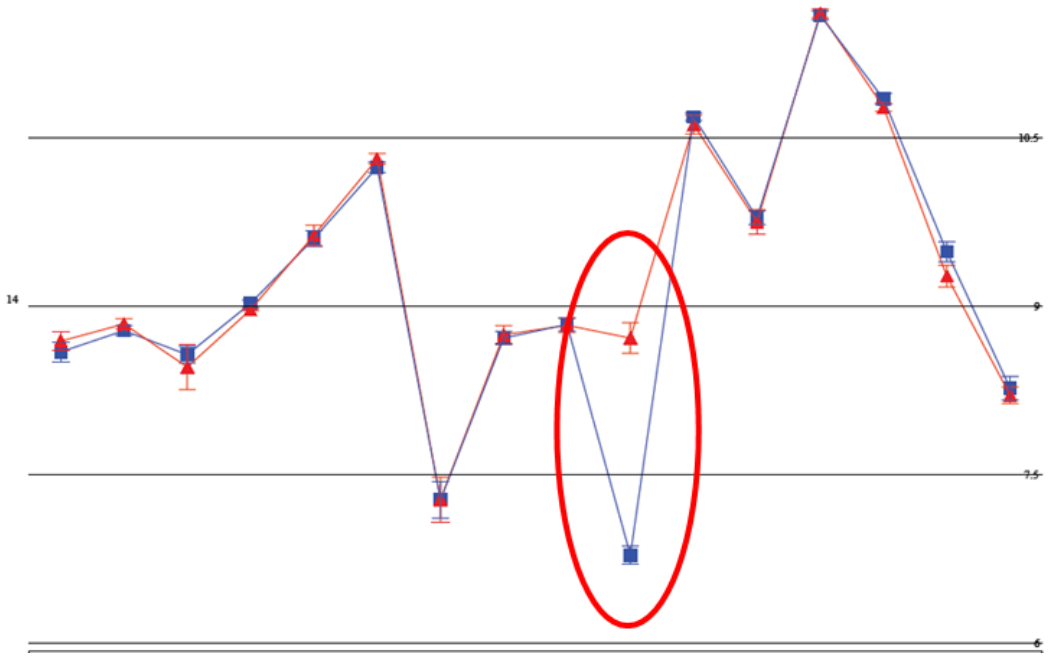


Figure 6.1 To establish the number of genes that contained dbSNPs linked to the C59X mutation the sequences corresponding to the probeset (red circle) indicative of differential splicing was checked in UCSC to determine known dbSNPs (dbSNP 128) within the targeted sequence.

6.2.2. Exclusion of Chromosome 2.

The core meta-probeset file consisted of 194,293 probesets, following the removal of the 15,713 probesets found on chromosome 2, 178,580 remained. A custom probeset file was made (Excel, 2007) with these 178,580 probeset IDs and the analysis was re-run with all other criteria as described in chapter 2.8.3.1

6.2.3 Impact of Intensity Threshold.

To determine the effects of altering the stringency of intensity filter, the alternative splice ANOVA (Partek GS) was run with both the embryonic and adult data (excluding probesets targeting chromosome 2). The original analyses, which excluded probesets with a maximum (in a sample) log2 intensity <3 unless there was differential expression of the probeset ($p \leq 0.05$) were compared with those using filters at log2 intensity <4, <5 and <6, retaining as before probesets that significantly differed ($p \leq 0.05$) between mutant and wildtype. I also investigated using a filter of *mean* log2 intensity <3, <4, <5 and <6, both retaining (as before) and excluding differentially expressed ($p \leq 0.05$) probesets.

6.2.4 Impact of Gene size.

To determine if there was a bias of significant splicing events in genes with more exons, I tested for correlation between probeset number (used as a proxy for exon number, as on average each exon is represented by one probeset) and –log alternative splice p value using the embryonic data. Both probeset number and alternative splice p value were not normally distributed and therefore a bivariate correlation using the non parametric Kendall's tau b statistic was calculated using SPSS (v16.0).

6.2.5 Re-analysing the data with more stringent thresholds and exclusion of Chromosome 2.

To obtain a more conservative analysis, I reanalyzed the embryonic data without probesets targeting chromosome 2 and with the more stringent intensity filters indicated by the above evaluation.

6.2.6 Embryonic and Adult Dataset Replication Following Stringent Analyses.

To obtain a more conservative analysis of the extent of overlap between adult and embryonic datasets, I re-examined both datasets after exclusion of chromosome 2 probesets and with a mean log2 intensity filter of 6. The embryonic data had gender and scan date covaried and the adult data were the combined data set of 16 samples with gender covaried. Overlap was determined as described in chapter 3.2.11.

6.2.7 Expression and Splicing of *Zfp804a*.

To establish if more stringent intensity cut-offs affected the previously generated expression and splicing results for *Zfp804a*, the more conservative analysis was run for both adult and embryonic datasets as described in chapter 2.8.3.6 but using the extended meta-probeset file and a mean log2 intensity = 6 filter.

6.3 Results

6.3.1 Linkage with the C59X Mutation.

6 genes (*Fam171b, Prdm4, Slc39a13, Ssfa2, Tcp11l1* and *Zc3h15*) were predicted to have differential splicing in embryonic and adult C59X mutants following a Bonferroni correction for multiple testing, 5 of which (all but *Prdm4*) had reduced expression in the exon targeted by the probeset in the C59X mutants. All 5 were identified as having C59X mutant specific alleles in the sequence targeted by the probeset predicted to be differentially spliced. These observations had led to the hypothesis that reduced hybridisation efficiency caused by the C59X specific alleles was resulting in false differential splicing predictions. All 5 were on chromosome 2 and all were in dbSNP, thus unlikely to be ENU generated, suggesting that a region of chromosome 2 with ENU strain dbSNPs was co-segregating with the C59X mutation. When assessing the distribution of the 50 most significant differential splice results from the embryonic data, there was an obvious enrichment of significant events on chromosome 2 (Fig. 6.2).

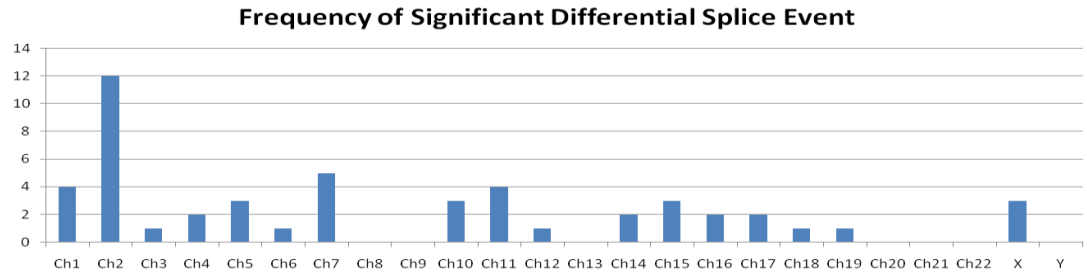


Figure 6.2 Genomic Distribution of the most Significant Differential Splice Events. Plotting the frequency of the 50 most significant differentially spliced genes from the embryonic analysis across the genome showed enrichment on chromosome 2.

In the embryonic data, the top 17 significant differentially spliced genes on chromosome 2 (Table 6.1) and the top 50 elsewhere in the genome (Table 6.2) were investigated to determine how frequently the predicted spliced probeset targeted a sequence with a dbSNP that when investigated in the RNAseq data was found to distinguish C59X mutants and wildtypes. Whilst none of the top 50 results found elsewhere in the genome had dbSNPs in the predicted spliced probeset that could be identified in the RNAseq to distinguish C59X mutants from wildtypes, 7 of the 17 on chromosome 2 did.

For each of the 7 dbSNPs in sequences targeted by probesets on chromosome 2, the RNAseq revealed the allele in the wildtype C57BL6J/HsdOla background was a perfect match for that of the probeset, the C3H/Hej or Balb/c ENU strain having a base that was non complementary to the probe. This is expected since the probesets were designed by Affymetrix based upon the NCBI build 37/mm9 sequence which itself is derived from the C57BL/6J strain. The 7 genes affected were also in proximity to *Zfp804a*. Consistent with an adverse effect of the non-reference allele on hybridisation, 6 of the 7 probesets indicated downregulation in the mutants. The exception is the gene *Calcrl*. This might point to a genuine upregulation of splicing of the gene, or simply a chance finding. This suggests that the mice are inheriting a region (under linkage) rather than just the C59X mutation which is not being broken up by recombination.

Gene Symbol	Start Position	Alt Splice P value	dbSNP	Strain carrying non-probe allele	Direction of Change
Itih2	10016220	1.10E-06	None		Mutant up
Mllt10	17986021	2.37E-05	None		Mutant up
Traf2	25373502	2.73E-05	3 dbSNPs	Variants not observed in Mutant or Wt	Mutant down
Bat2l	32079483	2.53E-06	None		Mutant down
Ssfa2	79475509	2.03E-06	rs13471129	Mutant	Mutant down
Zc3h15	83484592	1.40E-08	rs33018080	Mutant	Mutant down
Fam171b	83652793	1.34E-14	rs28032403	Mutant	Mutant down
Calcrl	84170783	1.26E-05	rs28028711	Mutant	Mutant up
Slc39a13	90901948	1.95E-10	rs13463033	Mutant	Mutant down
Caprin1	103603098	5.65E-06	rs33587735	Mutant	Mutant down
Tcp11l1	104497445	1.48E-10	rs27414619	Mutant	Mutant down
Eif2ak4	118214354	4.29E-06	2 dbSNPs	rs27424057 in both	Mutant down
Pla2g4e	119992148	3.33E-06	None		Mutant up
Zfp106	120332556	6.26E-09	None		Mutant down
Ubr1	120686005	9.62E-14	rs27438911	Both	Mutant down
Ccndbp1	120834139	2.77E-05	2 dbSNPs	Both	Mutant down
Cds2	132088884	7.23E-06	None		Mutant up

Table 6.1 Frequency of dbSNPs in Sequences of Differentially Spliced Genes on Chromosome 2. The top 17 significant differentially spliced genes found on chromosome 2 are ordered from top to bottom by genomic location. Those highlighted in yellow were genes in which the differentially spliced probeset contained a dbSNP specifically found in the mutant samples. This occurred in 7 of the genes and in 6 of the 7 the probeset was downregulated in the C59X mutants. *Zfp804a* is located between *Ssfa2* and *Zc3h15*.

Gene Symbol	Chromosome	Start Coordinate	Alternative Splice p value	dbSNP	Strain carrying non-probe allele	Direction of Change
Cul3	1	80261498	4.66E-06	None		Mutant up
Kif1a	1	94912033	6.03E-06	None		Mutant down
R3hdm1	1	129999883	9.10E-06	None		Mutant up
Cp	3	19857054	7.08E-06	None		Mutant up
Tnfsf15	4	63388118	8.52E-06	None		Mutant down
Elavl2	4	90917397	2.17E-05	None		Mutant up
Ift74	4	94281182	3.31E-06	None		Mutant up
Agrn	4	155539407	1.10E-05	None		Mutant down
Akap9	5	3928054	7.91E-06	None		Mutant up
0610007C21Rik	5	31350685	1.87E-06	None		Mutant up
Vps33a	5	123978773	3.44E-06	None		Mutant up
Eefsec	6	88173756	1.07E-05	rs37659956	Variants not observed in Mutant or Wt	Mutant down
Plxna1	6	89265692	2.27E-05	None		Mutant down
Emp1	6	135312949	3.15E-06	None		Mutant down
Saps1	7	4,583,196	7.64E-06	None		Mutant down
Supt5h	7	29099917	1.30E-06	rs36348871	Variants not observed in Mutant or Wt*	Mutant down

Table 6.2

Gene Symbol	Chromosome	Start Coordinate	Alternative Splice p value	dbSNP	Strain carrying non-probe allele	Direction of Change
Nomo1	7	53289066	7.79E-09	2 nonsynonymous; 3 synonymous	Variants not observed in Mutant or Wt	Mutant Down
Alg8	7	104520116	2.41E-05	None		Mutant up
Eif4g2	7	118214082	1.04E-05	None		Mutant down
Spon1	7	120909512	2.21E-11	None		Mutant up
Odz3	8	49285946	1.95E-05	None		Mutant up
Robo3	9	37223264	1.13E-05	None		Mutant down
Anxa2	9	69301447	1.18E-05	None		Mutant up
Clstn2	9	97344814	1.85E-05	None		Mutant down
Gpr126	10	14122391	4.75E-06	None		Mutant up
Prdm4	10	85354711	2.83E-06	rs30113788	Variants not observed in Mutant or Wt	Mutant up
Zdhhc17	10	110381449	1.79E-05	None		Mutant up
Lrp1	10	126975217	3.41E-06	None		Mutant down
Myh10	11	68505061	8.21E-07	None		Mutant up
Trim37	11	86940579	2.24E-05	None		Mutant up
Med24	11	98565905	2.02E-06	rs27041085 (C/T)	Variants not observed in Mutant or Wt*	Mutant up
Nol11	11	107027977	3.97E-07	None		Mutant up

Table 6.2

Gene Symbol	Chromosome	Start Coordinate	Alternative Splice p value	dbSNP	Strain carrying non-probe allele	Direction of Change
Abca9	11	109962063	2.09E-06	rs27036976 (C/T)	Variants not observed in Mutant or Wt*	Mutant up
Dync1h1	12	111839662	6.57E-09	None		Mutant up
Bap1	14	32064675	1.21E-06	None		Mutant up
Pcca	14	122933546	8.53E-08	None		Mutant up
Zfr	15	12047586	2.53E-06	None		Mutant up
Cdh6	15	12963955	1.04E-10	None		Mutant up
Arf3	15	98568052	4.12E-06	None		Mutant up
Ktelc1	16	38,525,264	3.12E-15	None		Mutant up
Grik1	16	87896441	1.70E-07	None		Mutant up
Dopey2	16	93712152	2.35E-05	None		Mutant down
Gtpbp2	17	46297981	8.69E-09	None		Mutant up
Ptprs	17	56551854	1.76E-08	None		Mutant down
Dpysl3	18	43480633	4.69E-06	None		Mutant up
Mus81	19	5482355	1.50E-05	None		Mutant up
Cybasc3	19	10652213	7.43E-08	None		Mutant down
Hcfc1	X	71188131	2.00E-07	None		Mutant down
Eda	X	97170945	7.91E-06	None		Mutant up
Huwe1	X	148235350	6.46E-06	None		Mutant up

Table 6.2 Frequency of dbSNPs in Sequences of Differentially Spliced Genes not on Chromosome 2. The top 50 significant differentially spliced genes found in the rest of the genome were investigated to determine if a dbSNP was present in the sequence targeted by differentially spliced probesets and if so whether that variant was specifically in the C59X mutants (identified using the RNAseq alignment files). None of the dbSNPs identified were found in the RNAseq data although the coverage was poor in the RNAseq data for 3 of the genes (*). Gene listed from top to bottom by genomic location.

Strain information could only be obtained for 3 of the 7 dbSNPs (using the SNP query form from MGI based on dbSNP build 128, http://www.informatics.jax.org/javawi2/servlet/WIFetch?page=snpQF) (Table 6.3). All 3 of the sequence variants were specific to one or both of the strains used to generate the ENU mouse line and suggests that each of these mutations was being inherited from the ENU strain along with the C59X mutation.

Gene Symbol	Strain Info
Ssfa2	Specific to C3Hej and Balb/cByJ
Zc3h15	Not there
Fam171b	Specific to C3Hej and Balb/cByJ
Calcrl	Not there
Slc39a13	Not there
Caprin1	Not there
Tcp11l1	Specific to C3Hej and Balb/cByJ

Table 6.3. Inbred Mouse Strain Specificity of C59X Mutant Specific dbSNPs. Of the 7 dbSNPs identified, strain information could only be found for 3. All 3 were specific to one or both of the mouse strains used to generate the ENU mouse line.

The data above strongly support the hypothesis that a proportion of the splice events on chromosome 2 are artefacts of ENU strain specific alleles co-segregating with the C59X mutation.

6.3.2 Exclusion of Chromosome 2.

Given the impact of linkage between polymorphisms on chromosome 2 and the C59X mutation on the analyses, I repeated the analysis after excluding all probesets that targeted chromosome 2. The general pattern of the data was broadly similar to that observed previously (chapter 5) in that more significant results were observed for splicing following multiple test correction, although there were fewer differentially expressed (Table 6.4) and spliced genes (Table 6.5).

Expression (embryonic)	**Including Chr. 2 Total Transcript Number: 15830**	**Excluding Chr. 2 Total Transcript Number: 14512**
p≤0.05	1346	**1228**
p≤0.01	225	**204**
p≤0.001	15	**11**
p≤0.0001	0	**0**
FDR 0.05	0	**0**

Expression (adult)	**Including Chr. 2 Total Transcript Number: 15833**	**Excluding Chr. 2 Total Transcript Number: 14515**
p≤0.05	1043	**902**
p≤0.01	242	**194**
p≤0.001	30	**13**
p≤0.0001	11	**3**
FDR 0.05	9	**2**
Bonferroni	3	**1**

Table 6.4 Differentially Expressed Genes following the removal of Chromosome 2 Probesets. After the ~15,700 probesets on chromosome 2 had been removed the number of differentially expressed genes in the C59X mutants was slightly lower at each p value threshold for both embryonic (top table) and adult (bottom table) results.

The number of apparently differentially spliced genes also decreased following the removal of chromosome 2 probesets from the analysis, with the adult analysis proportionately affected more than the embryonic. Following the removal of chromosome 2 probesets, only *Prdm4* (see chapter 5) remained significant in both datasets after a Bonferroni correction.

Splicing (Embyonic)	**Including Chr. 2 Total Transcript Number: 15830**	**Excluding Chr. 2 Total Transcript Number: 14512**
p≤0.05	1808	**1627**
p≤0.01	721	**629**
p≤0.001	214	**170**
p≤0.0001	104	**80**
FDR 0.05	172	**125**
Bonferroni	30	**23**

Splicing (adult)	**Including Chr. 2 Total Transcript Number: 15833**	**Excluding Chr. 2 Total Transcript Number: 14515**
p≤0.05	617	**514**
p≤0.01	226	**162**
p≤0.001	92	**51**
p≤0.0001	54	**22**
FDR 0.05	65	**22**
Bonferroni	29	**9**

Table 6.5 Differential splicing in C59X embryonic (top) and adult (bottom) mutants following the exclusion of all probesets on chromosome 2.

Due to the decreased numbers of significant differentially expressed and spliced genes it does suggest an unusual proportion of the significant results were genes found on chromosome 2 and favours the removal of these probesets as it is probable the majority are artefacts and in the absence of the ability to distinguish between true and false positives it is best to remove them all.

6.3.3 Assessment of Intensity Cut-offs.

All analyses presented thus far used the default intensity filter in Partek GS which excludes any probeset with a maximum log2 intensity less than 3 unless it is significantly differentially expressed ($p \leq 0.05$). Given the substantially higher numbers of significant differentially spliced genes with Partek GS relative to others based on the DABG algorithm (chapter 3 and 5) more stringent thresholds were tested. Analyses excluded chromosome 2. As expected, all increases in stringency (increasing the intensity threshold, using the mean threshold rather than requiring at least one sample to attain that threshold) reduced the total number of transcripts in the analyses (Table 6.6). However, when probesets are excluded based on intensity alone irrespective of differential expression p value (column called no exception), as the intensity threshold is increased, the numbers of differentially expressed and spliced genes decreases. The reverse is true when those probesets below the threshold are retained if they show significant differences between groups (columns called Max and Mean). The reduction in the former is easily explained by the exclusion of probesets with significant differential expression ($p \leq 0.05$) that have low intensity. However, it is less obvious why there should be an increase in the number of significant observations when the more stringent thresholds are applied in the latter two analyses. It maybe that a probeset with low expression is not differentially expressed between the two groups, however the probesets across the remainder of the transcript are differentially expressed. When the filters are less stringent the low expressing probeset is included so no significant differential expression is observed. When the more stringent filters are applied the low expressing probeset is excluded and the differential expression observed throughout the remained of the transcript becomes significant.

Adult Dataset	Log2 Intensity		
	Max<3	Mean<3	Mean<3 Without Exceptions
Total Transcripts	14515	14440	14422
Expression (P≤0.05)	902	911	902
Splicing (P≤0.05)	514	517	510
	Log2 Intensity		
	Max<4	Mean<4	Mean<4 Without Exceptions
Total Transcripts	14426	14302	14258
Expression (P≤0.05)	911	918	896
Splicing (P≤0.05)	516	537	502
	Log2 Intensity		
	Max<5	Mean<5	Mean<5 Without Exceptions
Total Transcripts	14288	14102	14006
Expression (P≤0.05)	913	935	876
Splicing (P≤0.05)	539	562	462
	Log2 Intensity		
	Max<6	Mean<6	Mean<6 Without Exceptions
Total Transcripts	13998	13712	13498
Expression (P≤0.05)	936	967	849
Splicing (P≤0.05)	584	639	425

Embryonic Dataset	Log2 Intensity		
	Max<3	Mean<3	Mean<3 Without Exceptions
Total Transcripts	14512	14459	14444
Expression (P≤0.05)	1228	1229	1227
Splicing (P≤0.05)	1627	1631	1628
	Log2 Intensity		
	Max<4	Mean<4	Mean<4 Without Exceptions
Total Transcripts	14410	14304	14261
Expression (P≤0.05)	1223	1239	1219
Splicing (P≤0.05)	1639	1637	1607
	Log2 Intensity		
	Max<5	Mean<5	Mean<5 Without Exceptions
Total Transcripts	14253	14080	14002
Expression (P≤0.05)	1239	1253	1208
Splicing (P≤0.05)	1625	1631	1534
	Log2 Intensity		
	Max<6	Mean<6	Mean<6 Without Exceptions
Total Transcripts	13982	13693	13535
Expression (P≤0.05)	1250	1280	1184
Splicing (P≤0.05)	1650	1682	1440

Table 6.6 The Effect of Intensity Filters on Determining Differentially Expressed and Spliced genes in Embryonic and Adult analyses of mutants versus wildtype. 'Without Exceptions' refers to the analyses which excluded probesets based on intensity alone and does not retain differentially expressed probesets ($p \leq 0.05$).

The analysis at the most stringent threshold which also excluded differentially expressed probesets with intensity values below the designated cut-off is the most conservative, and one would assume is the most robust. Previous analyses were aimed at capturing as many expression and splice differences as possible and therefore lower intensity stringencies were used to reduce the number of false negatives. When carrying out pathway analyses (presented in chapter 7) I chose what I felt to be the more robust intensity filter, where probesets are removed based on intensity, irrespective of significance. As when a probeset has low expression (log2 intensity <6) it is far more likely that the differential expression prediction is a false positive due to the probeset being expressed at an intensity which is not distinguishable from the background distribution. The caveat being real data could also be excluded. When taking this analysis into consideration and the exclusion of chromosome 2 reduction in the total number of transcripts is observed, but this time an accompanying reduction in the number of significant differentially expressed and spliced genes is also observed.

6.3.4 Determining the Effect of Gene Size on Differential Splicing Results.

Larger genes have a greater number of exons, and as a result a larger number of probesets and a greater degree of multiple testing may introduce a bias towards such genes appearing to be significant in the differential splicing assay. Larger genes may also have a true biological increase in alternative splicing. Significant differential splicing results would therefore contain a disproportionate number of larger genes. When carrying out pathway analysis on such data the analysis would be bias in identifying pathways that happen to be comprised of larger genes irrespective of the cause. To determine if there is a relationship between the probability of observing a significant splicing event in a gene and the number of probesets in the gene, I sought correlation between the number of probesets targeting a gene and the –log alternative splice p value in the embryonic data.

I hypothesised that as probeset number increased, the likelihood of a significant alternative splice event would also increase. As variables were not normally distributed, I undertook bivariate correlation with the Kendall's tau b statistic. A two-tailed test was carried out. The number of markers in a gene and the –log of the differential splice p value were weakly ($r = 0.061$) but significantly ($P<0.01$) correlated (Fig. 6.3). This supports the hypothesis that the more probesets that target a gene, the more likely that

gene is to have a significant differential splice p value. Again it is important to note this increased differential splicing may be a result of a real biological effect or the impact of multiple testing. Whilst a significant correlation is observed, less than 1% of the variance in differential splice p value is explained by the number of probesets in a transcript, which is very small.

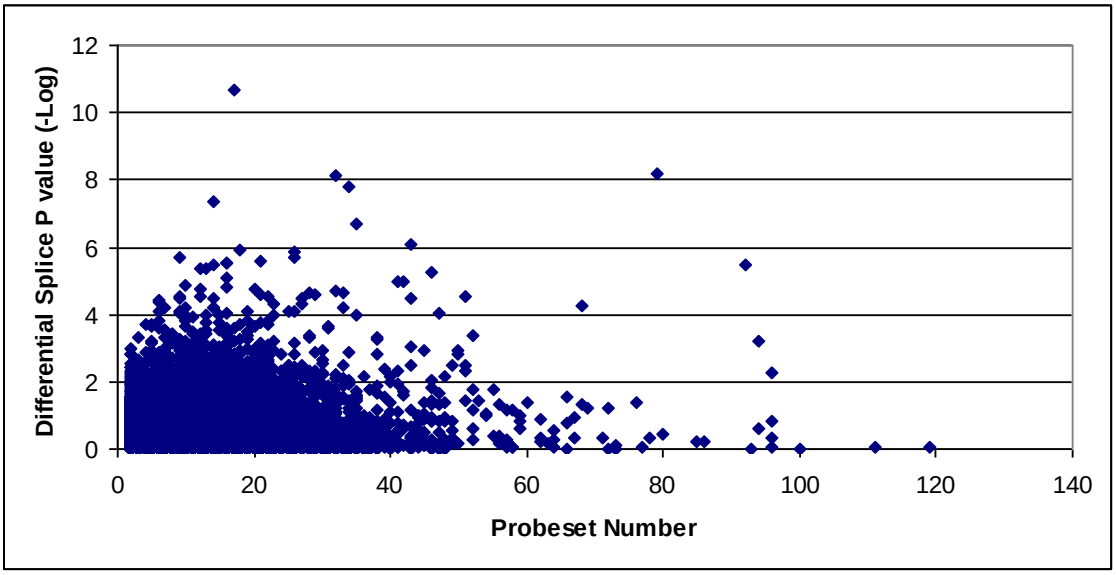


Figure 6.3 Correlation between the Number of probesets Targeting a Gene and the Signficance of Differential Splicing (-log P value).

6.3.5 Re-analysis With More Stringent Intensity Filters and Exclusion of Chromosome 2.

Having determined that the conservative analysis would be to exclude probesets that target chromosome 2 or that are expressed at a mean log2 intensity <6, the embryonic dataset was re-analysed using these filters.

6.3.5.1 Differential Expression between C59X Mutants and Wildtypes.

The number of genes differentially expressed between the C59X mutants and wildtypes at a p value ≤0.05 when probesets that target chromosome 2 are excluded was 1184 (Table 6.7) approximately 200 fewer than when using the original filters (Chapter 5, Table 5.7). As before, no genes remained significant after multiple test correction (FDR threshold 0.05).

Unadjusted p value	Significant Genes	% of Genes	Expected by Chance
p≤0.05	1184	9	677
p≤0.01	199	1	135
p≤0.001	12	0	14
p≤0.0001	1	0	1
FDR 0.05 threshold	0		
Bonferroni Correction for 15830 tests	0		

Table 6.7. Differentially Expressed Genes between Embryonic C59X Mutants and Wildtypes.

The AltAnalyze software predicts 1246 ($p \leq 0.05$) differentially expressed genes when chromosome 2 is removed of which 260 genes are also among those nominally significant in the Partek analysis. Of 1019 genes significantly differentially expressed in easyExon, 263 of them are among those nominally significant in the Partek analysis. When considering the results of all 3 analyses, 157 genes are predicted to be differentially expressed in all 3. The fold change of the expression difference was not considered with this data. This is because, using what I consider to be a more robust analysis in Partek GS with the more stringent intensity threshold, the addition of a fold change filter in addition to a significance filter seemed excessively stringent (as observed in Chapter 5.3.8.3.1). The degree by which significant differential expression results from the embryonic data replicated in the adult dataset was evaluated using Partek GS (Table 6.8). This analysis took the data generated using the more stringent approach. Significant genes were defined at a series of p value thresholds beginning at $p \leq 0.05$ in the embryonic data with a progressively more stringent threshold applied ($p \leq 0.01$, $p \leq 0.001$ and $p \leq 0.0001$). There was no obvious concordant significantly differentially expressed genes in the adult and embryonic datasets.

Differential Expression	O.R	$\chi 2$
p≤0.05	1.11	p = 0.216
p≤0.01	1.22	p = 0.273
p≤0.001	2.98	p = 0.172
p≤0.0001	0	p = 0.937

Table 6.8 Replication of Differential Expression Results in C59X Embryonic and Adult Following the Removal of Low Expressing and Chromosome 2 Targeting Probesets. When more stringent filters were applied to the data no overlap was observed between embryonic and adult studies. O.R. is the odds ratio that a finding is significant in the adult data conditional on it being significant in the embryonic data. An O.R. <1 means findings that are significant in the embryonic data are not more likely to be significant in the adult data compared with those that are not significant in the embryonic data.

6.3.5.2 Differential Splicing Between Embryonic C59X Mutants and Wildtypes.

The number of differentially spliced genes between embryonic C59X mutants and wildtypes following the removal of low intensity and chromosome 2 probesets decreased by ~400 genes at an unadjusted p value of ≤0.05 (Table 6.9). Although fewer than before, there were still an appreciable number of genes (Chapter 5, table 5.8) significant following multiple test correction.

Unadjusted p value	Significant Genes	% of Genes	Expected by Chance
p≤0.05	1440	11	677
p≤0.01	544	4	135
p≤0.001	140	1	14
p≤0.0001	58	0	1
FDR 0.05 threshold	88		
Bonferroni Correction for 15830 tests	16		

Table 6.9. Number of Differentially Spliced Genes between Embryonic C59X Mutants and Wildtypes.

In AltAnalyze 412 genes are predicted to be differentially spliced (p≤0.05 FIRMA) of which 130 overlap with genes predicted by Partek GS. The easyExon analysis using the MiDAS algorithm (p≤0.05) predicts 316 significant differential splice events of which 114 overlap with those significant in Partek GS. 55 significant differentially spliced genes were common to all 3 analyses.

Revisiting the overlap of differential splicing results between C59X mutants and wildtypes in embryonic and adult data using the more robust filtering criteria, there were no more replications between embryonic and adult data than expected by chance at any p value stringency (Table 6.10). This suggests the highly significant overlap before was likely driven by false positive results due to a combination of low intensity probesets and polymorphisms linked to the C59X mutation on chromosome 2. The apparent lack of overlap in the more stringent analyses can be interpreted in four ways. First, the results of the adult experiment could be confounded by the *Snca* deletion, and a high proportion of changes seen between C59X mutants and wildtypes could be attributable to the *Snca* deletion. Second, changes observed in the adult mutants could be different from those seen in embryonic mutants due to the developmental regulation

of the expression or splice changes. Third, the results from both adult and embryonic studies were chance findings (although the number of FDR and Bonferroni significant genes would argue against this (Chapter 3, Table 3.10). Fourth, there could be an as yet unaccounted for technical confounder in the embryonic data.

Differential Splicing	**O.R**	**$\chi 2$**
p≤0.05	0.97	p = 0.472
p≤0.01	0.87	p = 0.360
p≤0.001	1.38	p = 0.279
p≤0.0001	2.29	p = 0.109

Table 6.10 Replication of Differential Splicing events observed in embryonic and adult studies following the removal of low expressing and chromosome 2 targeting probesets. When more stringent filters were applied to the data no overlap was observed between embryonic and adult studies. O.R. is the odds ratio that a finding is significant in the adult data conditional on it being significant in the embryonic data. An O.R. <1 means findings that are significant in the embryonic data are not more likely to be significant in the adult data compared with those that are not significant in the embryonic data.

6.3.6 Expression and Splicing in Zfp804a Following increased Intensity Filtering.

6.3.6.1 Differential Expression of Zfp804a.

Using the original less stringent intensity filters in the adult data no significant differential expression was observed in either the female or the male analyses using Partek GS (Table 6.11), although a trend was observed for upregulation in the mutants. In AltAnalyse there is the option to make intensity filters less stringent. The default is to exclude probesets with a non log expression below 70. I reduced this to exclude only probesets with a nonlog expression below 1. A similar non significant result was observed in AltAnalyze when using this less stringent intensity cut-off. In easyExon (without the default fold change filter), significant upregulation of Zfp804a was observed in the C59X female and male mutants ($p = 0.008$ and $p= 0.01$ respectively). I repeated the analysis in Partek GS using the extended metaprobeset file (as no core probesets target Zfp804a) with the mean intensity <6 exclusion criterion, and in AltAnalyze excluding probesets with a non log (absolute) expression below 70. Significant differential expression was observed in both AltAnalyze (female $p< 0.02$; male $p<0.02$) and Partek GS (females $p < 0.01$ and males $p <0.0007$) concordant with the easyExon results. In the embryonic data, no differential expression is observed between mutants and wildtype irrespective of intensity filter.

In the RNAseq data Zfp804a was significantly upregulated in mutants (fold change = 1.56), $p = 2.26x10^{-7}$) and this remained significant following Benjamini and Hochberg (1995) FDR correction for multiple testing. Thus, significant upregulation of Zfp804a was consistently observed in Partek GS, easyExon, AltAnlayze and the RNAseq data for adult C59X mutants, but no differential expression was observed in the embryonic C59X mutants.

	Differential Expression of Zfp804a P value		
Differential Expression	**Adult Female C59X**	**Adult Male C59X**	**Embryonic C59X**
Partek GS (Log2 Intensity >3)	0.18	0.75	0.84
AltAnalyze (No nonLog expression Filter)	0.17	0.83	Not Significant
easyExon	**0.008**	**0.01**	Not Significant

With More Stringent Intensity Filters	Differential Expression of Zfp804a P value		
Differential Expression	**Adult Female C59X**	**Adult Male C59X**	**Embryonic C59X**
Partek GS (Log2 Intensity >6)	**0.01**	**0.007**	0.58
AltAnalyze (Non log expression >70)	**0.02**	**0.02**	Not Significant

Table 6.11 Expression of Zfp804a in Embryonic and Adult C59X mutants. When using the most stringent intensity cut offs significant upregulation of Zfp804a is observed in female and male adult C59X mutants in all 3 software tools. Differential expression is not observed in embryonic C59X mutants.

6.3.6.2 Differential Splicing of Zfp804a.

In the adult data with the default intensity filter (Max log2 intensity <3), significant differential splicing was observed at Zfp804a using Partek GS in both female and male datasets (Females $p = 5.16x10^{-9}$; Males $p = 3.07x10^{-8}$) (Table 6.12). Using all three packages (Partek GS, AltAnalyze and easyExon) significant differential splicing at probeset 5023975 was observed in the analysis of female mice, and the same probeset showed the same effect using Partek GS in male mice. Significant differential splicing was also observed in males in AltAnalyze and easyExon, but at different probesets (4929502 and 4710659 in AltAnalyze and 4929502 in easyExon). There were an inadequate number of alignments in the RNAseq data for differential splicing to be calculated.

When more stringent filters are applied in both Partek GS and AltAnalyze the significant differential splicing in female adult mice is nominal (p=0.05) and in embryonic C59X mice it disappears. Significant differential splicing in Male C59X mutants is observed in all 3 software programmes and appears to be robust to more stringent intensity filters. In the adult males, probeset 5424882 ($p<0.007$) was significantly differentially spliced between C59X mutants and wildtypes in AltAnalyze and Partek GS while probeset 4929502 was significant in AltAnalyze and easyExon.

Zfp804a appeared to be upregulated in the adult C59X mutants (male and female), but differential splicing robust to software and threshold changes was observed only in the adult male analyses (Fig. 6.5). As the spliced probeset varied depending on the software programme, the splice differences may reflect technical artefacts. Overall, no significant differential expression and splicing changes in *Zfp804a* were observed in the embryonic study using the most conservative criteria (Fig. 6.6). As upregualtion of *Zfp804a* in C59X mutants relative to wildtypes in observed only in adult mice, the difference may be the result of an unknown technical confounder.

	Differential Expression of Zfp804a Pvalue		
Differential Splicing	**Adult Female C59X**	**Adult Male C59X**	**Embryonic C59X**
Partek GS (Log2 Intensity >3)	**$5.16x10^{-9}$**	**$3.07x10^{-8}$**	**0.002**
AltAnalyze (No nonLog expression Filter)	**0.009**	**0.04**	Not Significant
easyExon	**0.04**	**0.04**	Not Significant

With More Stringent Intensity Filters	Differential Expression of Zfp804a Pvalue		
Differential Splicing	**Adult Female C59X**	**Adult Male C59X**	**Embryonic C59X**
Partek GS (Log2 Intensity >6)	0.05	**1.28x10-8**	0.07
AltAnalyze (Non log expression >70)	Not Significant	**0.02**	Not Significant

Table 6.11 Splicing of *Zfp804a* in Embryonic and Adult C59X Mutants. When more stringent filters are applied in both Partek GS and AltAnalyze the significant differential splicing in female and embryonic C59X mutants disappears. Significant differential splicing in Male C59X mutants is observed in all 3 software programmes and appears to be robust to more stringent intensity filters.

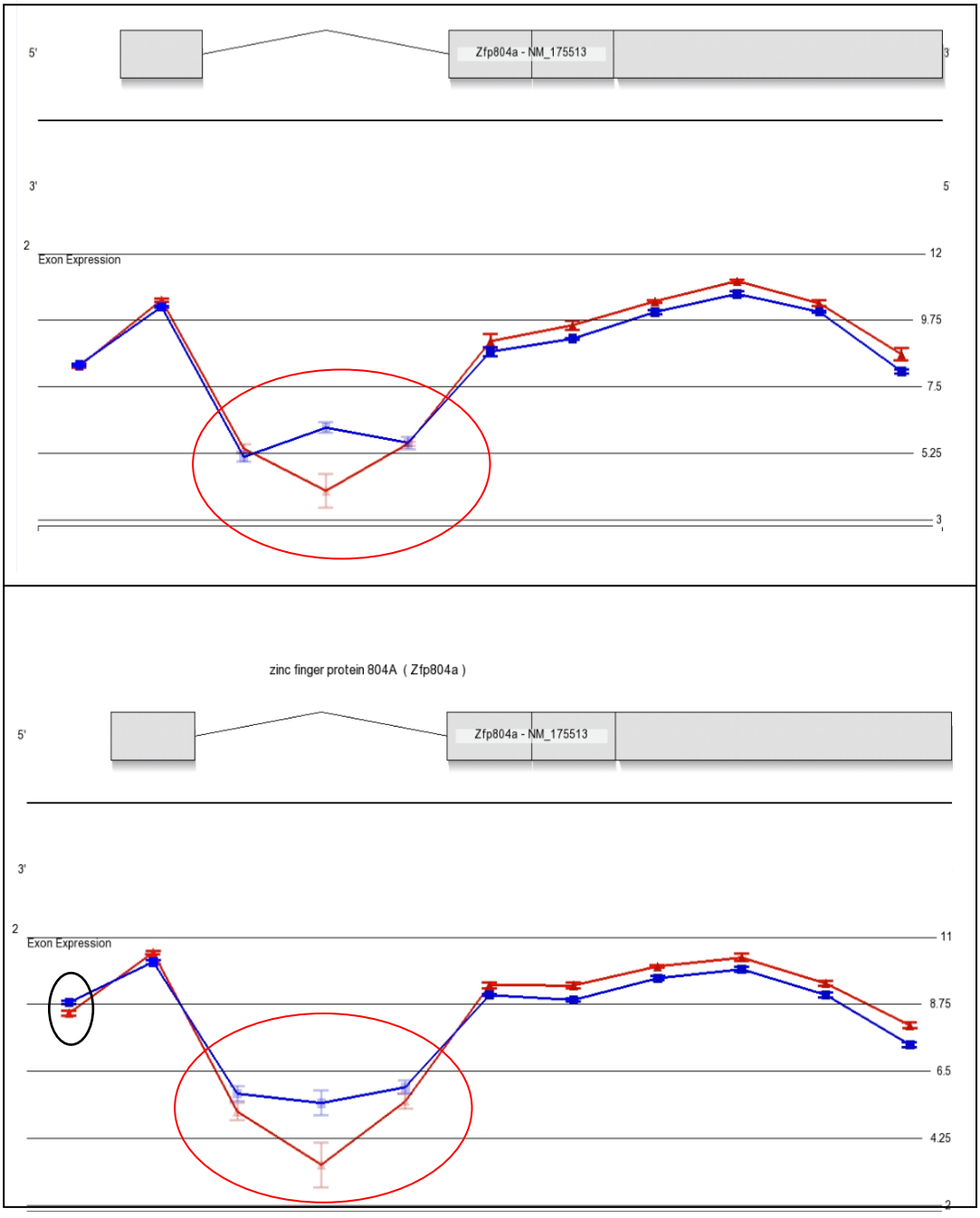


Figure 6.5 Differential Expression and Splicing of *Zfp804a* in Partek GS in female adult mice (top) and male adult mice (bottom). When probesets with expression values below a mean log2 intensity of 6 are excluded from the analysis (drawn transparent and circled in red) significant upregulation of Zfp804a in female and male C59X mutants is observed. The significant differential splicing from the central of the 3 circled probesets is removed. Differential splicing is observed for the 5' probeset (circled black) in male C59X mutants relative to increased mutant expression in all other probesets. **Mutants** Red **Wildtype** Blue.

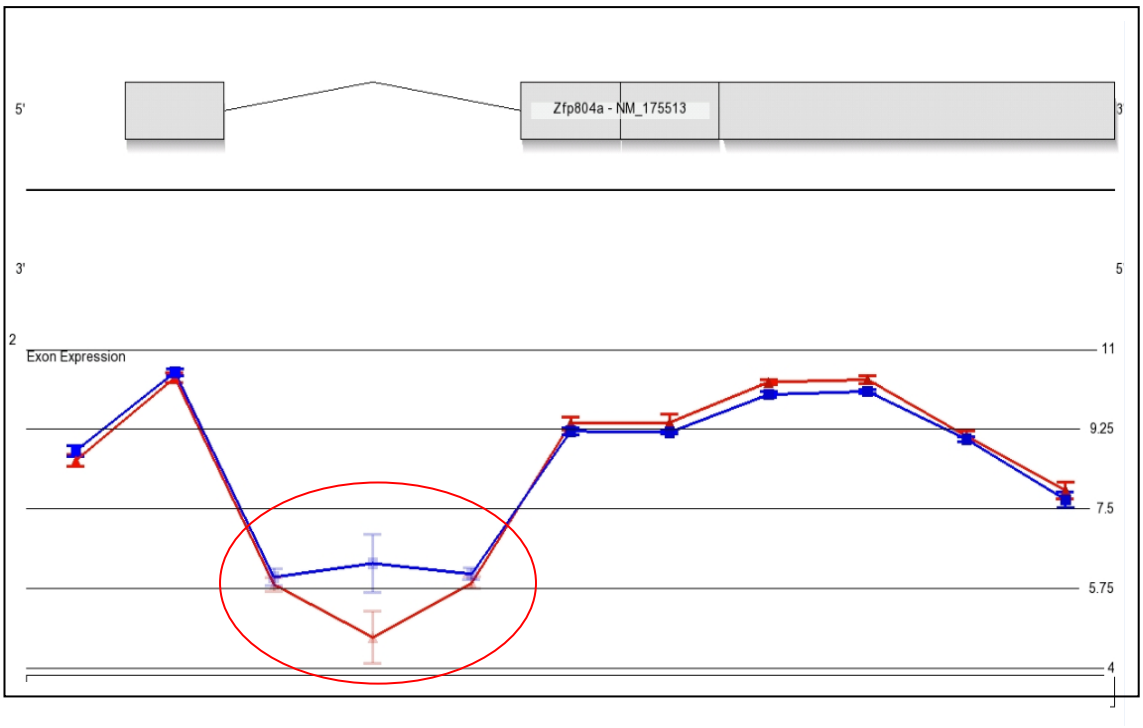


Figure 6.6 Differential Expression and Splicing of *Zfp804a* in Partek GS. Using a log2 intensity threshold <6 the 3 circled probesets are excluded and so no significant differential expression or splicing was observed between embryonic C59X mutants and wildtypes. Expression for each of the probesets targeting Zfp804a are displayed in a 5' to 3' direction with expression plotted on the y axis using a log2 scale. The blue line represents Wildtypes and the red line represents C59X Mutants.

The probeset predicted to be differentially spliced in Partek GS using lower stringencies targeted sequence containing a dbSNP (rs28042740) and was downregulated in C59X mutants. Based on earlier findings of linked variants causing reduced hybridisation efficiency, in the mutants I hypothesised that the differential splicing in Partek GS could have be generated due to this sequence variant. Only perfect match alleles to the probeset were present in the RNAseq data suggesting reduced hybridisation efficiency at the probeset did not explain the differential splicing. This rules out the effects of a C59X mutants specific SNP generating a false positive. The possibility of the probeset cross-hybridising was ruled out by ensuring that the probeset sequence specifically targeted Zfp804a (Chapter 3.3.4.11). This suggests the differential splicing at this probeset could be an artefact of the low expression of the probeset. When using more stringent filters this probeset is excluded.

To determine if any strain specific alleles lay within the *Zfp804a* all exons of the gene were inspected in the RNAseq data. In addition to the C59X mutation, the mutant line differs from wildtype at two points. Both are known dbSNPs, both in exon 4, and are targeted by probeset 5147435. One is non-synonymous (H782R) the other synonymous (V340V).

6.4 Discussion.

The removal of probes with known SNPs is an important part of the analysis workflow when using human samples (Kwan et al., 2007 & 2008) and options to do so are provided within the Partek GS software. This option is omitted from the mouse array workflow and the issue of probes targeting mouse sequences with SNPs is largely absent from the literature and Affymetrix support. This is most likely due to the presumption that inbred mouse strains will be genetically identical, but as I have shown, this is an important consideration when mouse lines are derived by backcrossing when genetic linkage can generate correlations between experimental groups and genotype. Gene level analyses will take into account intensities at a larger number of probes than exon level data and therefore false positives due to polymorphisms are much more likely to be prevalent in splicing data. Polymorphisms found at the centre of a probe sequence can generate a 2 fold decrease in expression measure relative to a near 0 fold change when the polymorphism is found at either end of the probe sequence (Benovoy et al., 2008) and thus are more likely to cause erroneous results. Several methods have been developed to overcome this problem when using human data. Most methods deal with the issue by masking probes that target sequence containing SNPs then carrying out expression analysis (Duan et al., 2008). Only one method has been developed which considers mouse array data and uses a post analysis statistical method to remove false positives caused by differential hybridisation efficiency as a result of polymorphisms in probe sequences (Alberts et al., 2007).

All probesets targeting chromosome 2 were excluded from the present analysis. Whilst this is not an approach that has been reported in the literature, the excess of significant results on the same chromosome as the mutation of interest in probesets spanning dbSNPs was suggestive of a technical artefact and, in my opinion, justified a conservative approach to data analysis. Without the removal of these false positives on chromosome 2 the results would impact on multi-locus pathway analysis, and conclusions about biological functions would have been made based on these inaccurate results.

A number of C59X mutant specific dbSNP were identified, including a nonsynonymous SNP found in exon 4 of *Zfp804a*. Nonsynonymous genetically linked mutations could have an impact on the relevant protein's function in the C59X mutants specifically

independent of the effects of the C59X mutation, and therefore influence the expression and splicing results of the present study. Unfortunately, the presence of such variants could not be adjusted for in the present study, leaving the caveat that any expression or splicing changes observed may not be attributable to *Zfp804a*.

Probesets that cause many of the technical artefacts are difficult to filter out using the currently available filtering methods and as exon arrays are relatively new there is no general consensus yet as to which probesets are problematic. As more results are generated this will allow this information to become available which should aid appropriate analyses (Whistler et al., 2010). Due to the number of different pre- and post-analytical filtering and the different software and algorithms available to analyse exon arrays, it is difficult to determine the best approach. A more systematic approach would increase the identification of robust splicing events and lessen the requirement for time consuming visual inspection of the data (Whistler et al., 2010).

I discovered that a linear relationship between the size of a gene and significant alternative splicing existed and this is in accordance with an mRNA sequencing study (Pan et al., 2008). Increased alternative splicing in larger genes may be attributable to a real biological increase of splicing in genes with more exons or as a result of the increased multiple testing in larger genes. Due to the weak relationship observed it is most likely not attributable to multiple testing.

Covarying for *Snca* expression was not possible in the adult data therefore the overlap analysis offered an alternative way of determining how confident to be in the data. When using the most conservative filtering (excluding probesets targeting chromosome 2 and with a mean log2 intensity <6) no overlap was observed between the embryonic and adult data that was more than would be expected by chance. This makes it difficult to interpret the results. The different gene expression and splicing profiles could reflect developmental regulation by *Zfp804a*, but could also represent changes that are a direct or indirect effect of the deletion in *Snca.* The findings in both experiments could be attributable to chance and this would account for the differences, although the overlap of multiple test corrected data and the control of type I error analysis (Chapter 3, pp.88.) would discount this. Finally it can not be disregarded that an as yet unaccounted for technical confounder in the embryonic data is responsible for the different findings between embryonic and adult C59X mutants. As it was difficult to determine which of these hypotheses was correct, the confidence in the data was not enough to extrapolate

meaningful conclusions about *Zfp804a* gene's function nor the potential ways in which mutations in this gene influence the pathophysiology of schizophrenia. The lack of any overlap does not negate the embryonic data which I regarded as the most reliable due to the consistent deletion of the *Snca* locus in all mice. These were therefore used in further downstream analyses described in the next section. Based upon the work of the present chapter, the downstream analyses are informed by a better understanding of the technical limitations of the data and analytic methods.

Chapter 7. Relevance of altered Zfp804a function for Schizophrenia.

7.1 Introduction.

The use of microarray experiments to study gene expression is now common place, but the output of these experiments at the level of single transcripts is generally inadequate for making functional inferences. The data need to be placed in the context of biological processes to infer knowledge about the changes in function related to the experimental conditions. One method commonly used to acquire functional relevance to microarray results is pathway analysis. Whilst the benefits of pathway analysis have been exemplified in some studies (Mootha et al., 2003) the potential for bias and ambiguity must be considered.

Ascertaining biological insights from a microarray study is a time consuming and demanding task if each gene were to be considered individually. Investigating the data at the level of biological pathways enables information to be gained about the entire dataset simultaneously without highly specific *a priori* functional hypotheses, which is one of the main advantages of pathway analysis (Drăghici et al., 2003).

Small sample sizes can limit the power of a gene expression study. Difficulty in distinguishing between true signal and noise is also often amplified by high variability between samples and the large number of genes that are tested (Mootha et al., 2003). Pathway analysis can offer a way of gaining more insight into the data from small samples because the data are viewed at the level of biological pathways rather than individual genes, the latter often being known to act as sets of coregulated gene sets. Power can then be increased by considering changes in multiple sets of functionally related genes. In essence, a large expression change observed in a single gene may provide less information than multiple smaller changes in a group of genes all belonging to the same biological pathway (Subramanian et al. 2005).

The basic hypothesis is that if changed function in a pathway results from the presence of the mutation at Zfp804a, genes found within that pathway would be enriched for those with significant differential expression or splicing. One major caveat of pathway analysis is the fact that functionally related genes are often co-regulated which means

observations cannot be strictly considered as providing fully independent evidence for a given set of genes. If a particular pathway shows significant changes in the expression or splicing of multiple members, this could simply be attributable to a chance fluctuation affecting an entire group of genes due to their correlated expression (Mootha et al., 2003). This is important to consider in pathway analysis where often independence of genes is assumed, and where doing so can increase the false positive rate (Emmert-streib & Glazko, 2011).

Often the approach taken to pathway analysis is to restrict pathways to those thought relevant to what is already understood about the condition. This user definition of how the gene lists and pathways are chosen can bias the results toward biological processes already known to be relevant to the pathophysiology of the disorder being studied. In other words, significant findings have high plausibility simply because they are defined in advance to have such plausibility. In addition the statistics can prioritise gene sets with more genes ascribed to them than gene sets that include genes demonstrating the greatest expression changes, which may be counter intuitive (Damian & Gorfine, 2004).

The numerous pathway analysis tools available can hinder the interpretation of expression results. The use of a variety of algorithms, databases and thresholds makes it difficult for a standardised approach to be followed and no consensus or gold standard currently exists which further impairs the comparison of results across studies.

One important aspect of pathway analysis is the use of an appropriate comparator reference set of genes (hereafter called a background list) (Huang et al., 2009a) against which the changes in specific target pathways can be assessed. The significance of overrepresentation in the target versus background sets can be determined using a number of tests, for example Fisher's exact test, the Chi squared test or the hypergeometric distribution (Huang et al., 2009a). As with pathways analysis in general there is no gold standard for background or the statistical approach. Generally the larger (i.e., the more comprehensive) the list the more accurate the estimate of enrichment (Huang et al., 2009a). Often several pathway analyses tools are used to gauge the robustness of results (Huang et al 2009b).

In order to understand the consequences of the Zfp804a mutation, I used complete lists of significant results for the pathway analysis. Enrichment analysis (EA) was performed using Metacore (GeneGO) (Inc., St Joseph, MI) and DAVID (Database for Annotation, Visualisation and Integration Discovery (Dennis et al., 2003; Huang et al., 2009a; Huang et al., 2009b). There are many different pathway tools available each using different annotation databases, but one common annotation database widely used is the Gene Ontology (GO) database (The Gene Ontology Consortium, 2000). The GO database consists of a hierarchy of 5 levels with level 1 representing the most general terms and 5 the most specific. Since level 1 incorporates all level terms, and genes may be members of multiple categories, the pathways are overlapping in membership and are not independent. This is generally not compensated for in available pathway tools (Jantzen et al., 2011). The occurrence of similar terms in functionally related pathways can obviously result in multiple similar pathways being simultaneously identified as significant, which might be misinterpreted as increased evidence for the relevance of the broad biological function to the question being studied (Jantzen et al., 2011). Both Metacore GeneGO and DAVID use the GO ontology as well as other annotation databases to allow a more thorough evaluation of functional enrichment in the gene list. The current level of annotation does however bias results in favour of systems for which more information is known.

As well as assessing the consequences of the C59X mutation through pathway analysis of the array data, the expression results were also considered within the context of human genetic data. It was my hypothesis that genes identified as showing altered expression or splicing would contain downstream mediators of the effects of ZNF804A on disease risk. The polygenic model of schizophrenia suggests the involvement of thousands of genetic variants (ISC, 2009) and therefore I postulated that a number of the downstream targets of ZNF804A might contain genetic variants that influence disease risk. Genes identified in the exon array as showing altered expression and splicing were therefore tested for genetic association using large genome-wide schizophrenia case-control association resources available from the Psychiatric GWAS Consortium (https://pgc.unc.edu/). To identify genes influenced by ZNF804A that might mediate disease risk, I used two gene-wide analyses, one which had been previously implemented and validated by the host department (Moskvina et al., 2009), the other based on the Simes' tests, the applicability of which for extracting valid gene-wide

genetic association values allowing for multiple testing has been demonstrated (Li et al., 2011).

Aim

The overall aim of this section is to assess the gene expression data in the context of biological pathways, and to determine if any of the genes showing altered expression or splicing are relevant to disease risk by assessing association with disease in a large case-control genetic dataset.

7.2 Methods

7.2.1 Pathway Analysis.

Pathway analysis was carried out using Metacore GeneGO and DAVID seeking biological pathways enriched for genes differentially expressed and spliced in embryonic C59X mutants compared to wildtype mouse brain.

7.2.1.1 Gene and Background lists

When referring to the 'gene list' I refer to all genes significantly differentially expressed or spliced at the relevant statistical threshold. In order to evaluate enrichment a background list must be provided for a comparator as discussed above. Default background lists are provided in both DAVID and Metacore GeneGO, but these lists represent either the complete mouse transcriptome or all genes included on specific array platforms. Both are inappropriate as a comparator group as these sets include genes that are not included in the analysis (e.g. genes not expressed in brain). The use of default sets would therefore inflate the significance of almost all pathways relevant to brain expression as only brain expressed genes have any chance of showing significant effects in the present study. There is no consensus as to the most appropriate background list to use, but in principle, the most appropriate set would include only those genes that were evaluated for significant changes in expression or splicing.

Both Metacore and DAVID offer the option to upload a customised background list. To construct the most relevant background list, I compiled a list consisting of all genes which had been entered into the expression or splicing analyses. This meant I included

only genes that met with the most stringent criteria (excluding probesets on chromosome 2 and with a mean Log2 Intensity <6). This same background list was used for both expression and splicing pathway analyses.

Of 13,535 genes included in the background list, 28 had no gene symbol or other useful identifier and were removed. Different transcripts of the same gene can be significantly differentially spliced meaning the same gene identifier will be represented in the list more than once, biasing the results related to this gene and its associated pathways. Duplicate genes symbols were removed to allow only one representation, of which there were 163 leaving 13,344 unique gene symbols. Using MGI Biomart (http://biomart.informatics.jax.org/biomart/martview/06197243657fc4ec9d186dc2cb2df738) gene symbols were converted to Entrez IDs, of which 12,943 unique identifiers were found and these comprised the background list for all analyses.

1184 genes were significantly differentially expressed at $p<0.05$ of which 1 had no gene symbol and the remaining 1183 were unique. 199 unique genes were significantly differentially expressed at $p<0.01$. These arbitrary thresholds were used to obtain gene lists of adequate size for pathway analysis. The gene symbols were converted to 1141 and 191 unique Entrez IDs respectively. The alternatively spliced gene lists were made up of either the genes significant at $p<0.01$ (544) or the genes significant at $p<0.001$ (140) of which all had gene symbols and no duplicate gene symbols were present. 530 and 135 unique Entrez IDs were identified in MGI Biomart for each list respectively. Pathway analysis was conducted in both Metacore and DAVID with the species selected as *Mus musculus*.

7.2.1.2 Metacore GeneGO

Metacore (GeneGo, Inc., St Joseph, MI, http://www.genego.com/metacore.php) is based on a collection of protein-protein, protein-DNA, protein-RNA, protein-compound and compound-compound interactions. The GeneGO database has been manually curated, with the content based on literature published from 2002-present.

Gene lists of Entrez IDs were uploaded into Metacore and the number of IDs recognised by Metacore was determined. Options to filter data based on fold change and significance are available but the lists entered had already been filtered for significance, and I did not apply fold change thresholds. P value thresholds for gene selection for

pathway analyses are generally not as conservative as those applied for inference from single locus analyses since the idea is to obtain information from multiple changes that are not themselves necessarily highly significant. Enrichment Analysis was carried out using the GO ontologies of 'Biological processes', 'Molecular Functions' and 'Localisations' in addition to the proprietary GeneGO datasets of 'Pathway Maps' and 'Cellular Process Networks'. The GeneGo Pathway Maps are made up of about 650 signalling and metabolic pathways. The GeneGo Process Network ontology is made up of 110 cellular and molecular processes each comprised of a set network of protein interactions which contain information about empirically validated interactions between the products of the genes. Annotations were curated by Metacore using an oracle database consisting of information from full text articles. Only empirically validated data are included in the database and the database is updated daily. I chose to include the Metacore GeneGO ontologies in addition to the GO ontologies as the GeneGO ontologies are more frequently updated and are, according to Metacore, more comprehensive. I also included GO ontologies. Following upload of the gene and background lists the Entrez gene IDs were mapped onto gene IDs from the GO ontologies or Metacore's own GeneGO database.

In Metacore all statistics are calculated according to the gene ID and whether or not the genes were associated with a particular pathway relative to the background list. A number of the GO processes have no actual gene content and for this reason, these "empty terms" are excluded. After each analysis was performed the results were presented in a histogram ordered according to the negative log p value for enrichment of the pathway for transcripts showing significant changes in expression or splicing that meet the significance thresholds described above. Enrichment analysis statistics were calculated using the hypergeometric p value. This evaluates the significance of the number of genes surpassing the chosen threshold in the test gene set given the distribution of results in the background comparator set (Tavazoie et al., 1999). Multiple testing was controlled using the FDR set at a threshold of 0.05. In Metacore the FDR is determined using the q-value (Storey, 2002), where a q value of 0.05 corresponds to a FDR of 0.05.

7.2.1.3 DAVID

DAVID (v6.7) (Database for Annotation, Visualisation and Integration Discovery) was also used for pathway analysis (Dennis et al., 2003; Huang et al., 2009a; Huang et al., 2009b). The gene list and background list were uploaded into DAVID and the Entrez IDs were then mapped to DAVID gene IDs. Each DAVID gene ID is unique to account for redundancy in the input gene list, although all lists I compiled had already had duplicates removed. I then used the functional annotation chart option to carry out enrichment or overrepresentation analysis. Statistics in DAVID are based on an adaptation of the Fisher's exact test called the EASE score, which is a one tail Fisher's exact probability value (Fig. 7.1). The EASE score is more conservative than the Fisher's exact test. In DAVID for a pathway to be considered significant, that pathway must contain at least 2 genes from the test set. This is because a finding derived from a single gene in a pathway is neither likely to be robust nor can in implicate convergence in a pathway as convergence implies multiple lines of evidence. As the intention is to discover overrepresented pathways the EASE score in DAVID is always a one tailed p value (Huang et al., 2009b).

		Background		
		Hit	No Hit	Total
Gene List	Hit	18 (19-1)	332-18	**332**
	No Hit	214-18	14431	**14959-332**
	Total	**214**	**14959-214**	**14959**

		Background	
		Hit	No Hit
Gene List	Hit	18	314
	No Hit	196	14431

Figure 7.1. The Ease Score Method. In the above example a gene list of 332 'significant' genes is compared to a background list of 14959 genes. 19 genes from the significant gene list hit the pathway being tested and 214 genes from the background list are also found in this pathway. To determine if this pathway is overrepresented in the gene list 2x2 contingency tables are compiled. In DAVID a modified version of the Fisher's Exact Test is used to determine if the gene list is overrepresented in a particular pathway. The modification is that the positive count (i.e., number of genes in the pathway that are found in the gene list) is penalised by subtracting 1. In the above example 19 genes from the pathway are found in the 'significant' gene list which is entered as 18 (19-1). To determine the significance of this a one tail Fishers Exact Test is then used (Huang et al., 2009b). Example taken from http://david.abcc.ncifcrf.gov/

To prevent redundancy among pathways, only one type of database is usually included in pathway tools but this can limit the depth of information obtained. In contrast, DAVID allows multiple databases to be examined and explored simultaneously. In order to avoid the issue of redundancy, a specific DAVID ID is used representing all possible identifiers of one gene. This enables that gene to be cross referenced across all

the different databases irrespective of what identifier is used by each database (Huang et al., 2009b).

To avoid repetition of GO ontologies from different levels of the hierarchy, and to reduce multiple testing, I used the GO FAT database (Dennis et al., 2003). The GO FAT database was developed by DAVID to filter out broad terms with numerous child terms, thereby including specific terms with less repetition at lower levels in the hierarchy (Huang et al., 2009b).

7.2.3 Genetic Analysis using the PGC database

7.2.3.1 Investigating PGC Schizophrenia top hits for Differential Expression and Splicing in C59X mutants.

The mouse orthologues at loci showing genome wide significance in the Psychiatric GWAS consortium (PGC) (PGC, 2011a) schizophrenia GWAS were specifically investigated for differential expression and splicing in the embryonic exon array data. MIR137 was not targeted by any core probesets and so could not be investigated. PCGEM1 is found on chromosome 2 and so was also omitted from the analysis. The MHC region was also excluded in this analysis given the imprecision of mapping the association signal due to long range LD in this region (PGC, 2011a).

7.2.3.2 Investigating the C59X Differentially Expressed and Spliced Genes for Association with Psychosis.

In order to investigate a relationship between the expression and splicing data from my mouse experiments, and the human genetic data the significant genes were investigated in the Psychiatric GWAS Consortium (PGC) schizophrenia and bipolar disorder datasets (PGC, 2011a; PGC, 2011b). The schizophrenia PGC carried out meta-analyses of GWAS data from 17 studies consisting of ~9000 cases and ~12,000 controls of European Ancestry. The most significant 81 SNPs identified were followed up in an independent sample of ~8,000 cases and ~21,000 controls (PGC, 2011a). The PGC Bipolar Disorder Working Group carried out combined GWAS on 7481 patients with bipolar disorder and 9250 controls. In this study the top 34 SNPs were tested in an

independent sample of 4,496 bipolar disorder and 42,422 controls. Both schizophrenia and bipolar disorder PGC studies used data collected from subjects of white European ancestry. These datasets are the largest GWAS datasets available. To investigate differentially spliced and expressed genes in these GWAS datasets, I used gene-wide estimates of significance (discussed below).

7.2.3.2.1 Approximation Method using Brown's p value

Across a gene, genotypes at many SNPs may be correlated with each other due to linkage disequilibrium. If this non independence is not allowed for when combining the SNP p values, highly inaccurate gene-wide p values can result. Traditionally, this is dealt with though permutation testing requiring the availability of individual genotypes but Moskvina and colleagues (Moskvina et al., 2011) reported a method that allows gene-wide p values to be derived from summary association statistics in the absence of individual genotype data. The derived P values, based upon theoretical approximation of the Fisher's Statistic (Brown, 1975), I refer to as Browns P values. All Browns P values were derived from the PGC data by Dr Moskvina.

7.2.3.2.2 Simes' P value

An alternative approach to establishing gene-wide significance value is based on the Simes' method, the validity of which has been demonstrated for GWAS data (Li et al., 2011). Using this approach the p values of each SNP are ranked from most significant to least. The p value ranked 1 is then multiplied by the number of markers divided by its rank (n/1). This process is repeated for the SNP ranked 2 (and so on) until all SNPs had been adjusted. The smallest p value is the Simes' corrected p value. The Simes' corrected p values for the PGC SZ and BP datasets were based on the data published by the PGC (2011). I personally calculated the Simes' values for the genes of highest interest, but otherwise accessed a database available in the host department.

7.3 Results

7.3.1 Pathway Analysis based on embryonic expression data.

Pathway analysis was carried out on the embryonic expression data. Chromosome 2 probesets were excluded as were probesets with mean $\log_2$ intensity <6. Both Metacore and DAVID were used for the analysis.

7.3.1.1 Pathways Enriched for Differentially Expressed Genes.

Gene sets comprising those differentially expressed at two thresholds, $p<0.05$ and $p<0.01$ were chosen for the analysis. When converted to unique Entrez IDs this resulted in lists of 1411 and 191 genes respectively. Both lists along with the background list of 12,943 Entrez IDs were uploaded into Metacore GeneGO and DAVID to enable enrichment of the gene lists in pathways to be quantified relative to the background list.

1137 of the 1141 Entrez IDs at $p<0.05$ mapped to DAVID IDs. Of the smaller list (n= 191) based upon the more stringent threshold for differential expression ($p<0.01$), 190 of them mapped to DAVID IDs. The 'functional annotation chart' option was chosen to test pathways for enrichment for genes in the differential expression dataset. The GO ontologies 'biological processes' (BP), 'molecular functions' (MF) and 'cellular component' (CC) were investigated as were KEGG and PANTHER databases. For each GO ontology, the FAT database was chosen. In Metacore, all 1141 Entrez identifiers were recognised. For each gene list the analysis was run in DAVID and then in Metacore first using the GO ontologies (as in DAVID), then using Metacore's own proprietaty GeneGO databases which have been curated differently to the GO ontologies.

$P \leq 0.05$

In DAVID 21 pathways were significantly enriched for genes with altered expression between C59X mutants and wildtype following a Bonferroni correction for the pathways tested. The top 10 most significant pathways are presented in Table 7.1. 7 of the 10 pathways related to the mitochondria or energy metabolism.

Differentially Expressed Genes between C59X mutants and Wildtypes (p≤0.05)

DAVID GO, KEGG, PANTHER	pValue	Bonferroni
GO:0005739~mitochondrion	5.42E-14	2.32E-11
GO:0044429~mitochondrial part	7.49E-13	3.21E-10
GO:0005743~mitochondrial inner membrane	1.35E-10	5.80E-08
GO:0031966~mitochondrial membrane	1.95E-10	8.36E-08
GO:0006091~generation of precursor metabolites and energy	2.24E-10	5.43E-07
GO:0006412~translation	3.62E-10	8.78E-07
GO:0005740~mitochondrial envelope	6.27E-10	2.69E-07
GO:0019866~organelle inner membrane	1.16E-09	4.98E-07
GO:0006413~translational initiation	1.35E-07	3.27E-04
mmu05016:Huntington's disease	4.33E-07	7.41E-05

Table 7.1 DAVID annotation categories significantly enriched for genes which show differential expression in the embryonic expression data. The pathways displayed above are the top 10 most significant pathways enriched for genes with significant (p≤0.05) differential expression between C59X mutants and wildtype. Enrichment was tested for GO biological processes, molecular functions and cellular components as well as the KEGG and PANTHER databases. **P value** was generated using the EASE score, a modified Fishers Exact Test. Bonferroni represents the p value following multiple test correction for the number of pathways in each database.

For Metacore the results are divided so that first the GO ontologies are considered, as these correspond to the same ontologies investigated in DAVID, then the results using the GeneGO databases (Metacore proprietary databases) are presented.

Non proprietary GO ontologies. The GO ontologies of Biological processes, molecular functions and localisations (cellular components) were each investigated separately in Metacore and KEGG and PANTHER databases were not included. The top 10 most significantly enriched pathways for genes differentially expressed between C59X mutants and wildtype for the three GO ontologies are displayed in table 7.2. Of the 9 GO ontologies in the top 10 DAVID results ('Huntington's disease' is a KEGG pathway) 8 are replicated in the top 10 from Metacore. All pathways relating to translation and the mitochondria replicated. The Metacore algorithm identified the GO biological processes of 'translation' ($p = 1.32 \times 10^{-8}$) and 'translational initiation' ($p = 1.67 \times 10^{-8}$) as enriched for differentially expressed genes. Both were significant at FDR threshold of 0.05.

Differentially Expressed Genes between C59X mutants and Wildtypes (p≤0.05)

Metacore GO Biological Processes	pValue	FDR
cellular respiration	3.42E-11	Significant
respiratory electron transport chain	1.02E-10	Significant
oxidation-reduction process	7.06E-09	Significant
translation*	1.32E-08	Significant
small molecule metabolic process	1.45E-08	Significant
translational initiation*	1.67E-08	Significant
electron transport chain	2.07E-08	Significant
negative regulation of protein ubiquitination	1.97E-07	Significant
energy derivation by oxidation of organic compounds	2.45E-07	Significant
mitochondrial ATP synthesis coupled electron transport	2.62E-07	Significant

Metacore GO Molecular functions	pValue	FDR
structural constituent of ribosome	1.39E-07	Significant
translation initiation factor activity	3.06E-07	Significant
oxidoreductase activity	3.27E-07	Significant
oxidoreductase activity, acting on NADH or NADPH, quinone or similar compound as acceptor	1.23E-05	Significant
catalytic activity	3.23E-05	Significant
gamma-catenin binding	6.99E-05	Significant
translation factor activity, nucleic acid binding	7.74E-05	Significant
NADH dehydrogenase activity	7.92E-05	Significant
NADH dehydrogenase (ubiquinone) activity	7.92E-05	Significant
NADH dehydrogenase (quinone) activity	7.92E-05	Significant

Metacore GO Localizations	pValue	FDR
cytoplasmic part	6.61E-15	Significant
mitochondrion*	7.15E-14	Significant
mitochondrial part*	2.15E-13	Significant
mitochondrial inner membrane*	9.51E-11	Significant
Cytoplasm	2.30E-10	Significant
mitochondrial envelope*	3.81E-09	Significant
organelle inner membrane*	5.29E-09	Significant
mitochondrial membrane*	2.97E-08	Significant
respiratory chain	3.22E-08	Significant
macromolecular complex	4.85E-08	Significant

Table 7.2 Metacore Analysis of Non proprietary GO ontologies. The results for each analysis (Biological processes, molecular function and localisation) are generated and presented separately. * Represents pathways that were one of the 10 most significant pathways in DAVID. These include translation and the mitochondrion.

Proprietry GeneGO database: The Metcore GeneGO process networks of 'translation initiation' as well as 'translation in mitochondria' were significant following multiple test correction (FDR 0.05 threshold) (Table 7.3). Although the GeneGo databases use

different classification methods the pathways that were the most significant relate to translation and the mitochondria which were prominent in the GO ontology results.

Differentially Expressed Genes between C59X mutants and Wildtypes (p≤0.05)

Metacore GeneGO Pathway Maps	pValue	FDR
Oxidative phosphorylation	1.88E-10	Significant
Ubiquinone metabolism	1.52E-04	Significant
Immune response_Signaling pathway mediated by IL-6 and IL-1	7.40E-04	Not Significant
Tricarbonic acid cycle	1.41E-03	Not Significant
Development_Thyroliberin signalling	1.44E-03	Not Significant
Immune response_IL-5 signalling	2.81E-03	Not Significant
Development_Prolactin receptor signalling	3.07E-03	Not Significant
Aminoacyl-tRNA biosynthesis in cytoplasm	3.17E-03	Not Significant
Aminoacyl-tRNA biosynthesis in cytoplasm/ Rodent version	3.17E-03	Not Significant
GTP-XTP metabolism	3.85E-03	Not Significant

Metacore GeneGO Process Networks	pValue	FDR
Translation_Translation initiation	2.05E-07	Significant
Translation_Translation in mitochondria	2.76E-05	Significant
Translation_Regulation of initiation	8.68E-05	Significant
Protein folding_Folding in normal condition	2.50E-04	Significant
Immune response_IL-5 signalling	2.19E-03	Not Significant
Translation_Elongation-Termination	4.28E-03	Not Significant
Translation_Elongation-Termination_test	4.28E-03	Not Significant
Protein folding_Protein folding nucleus	5.83E-03	Not Significant
Protein folding_Response to unfolded proteins	6.18E-03	Not Significant
Proteolysis_Ubiquitin-proteasomal proteolysis	0.01	Not Significant

Table 7.3. Metacore Analysis of proprietary GeneGO ontologies annotation categories significantly enriched for genes which show differential expression in the embryonic expression data. The p values displayed are unadjusted but all 4 pathways remained significant following correction for multiple testing using the FDR (q value) with a threshold set at 0.05.

p≤0.01

When analysing the smaller list (n= 191) based upon the more stringent threshold for differential expression (p<0.01) four pathways were significant following Bonferroni correction using the DAVID functional annotation chart analysis (Table 7.4). Each of the 4 pathways involved translation and included ‘translation’ and ‘translational initiation’ both observed when using the less stringent list (Table 7.1). The ‘mitochondrion’ pathway was also significant when using the more stringent p value threshold (p≤0.01). The finding that all 4 Bonferroni significant pathways were related

to translation, 2 of which were identified at the less stringent threshold, suggests Zfp804a may play a role in regulating genes which control translation.

Differentially Expressed Genes between C59X mutants and Wildtypes (p≤0.01)

DAVID GO, KEGG, PANTHER	pValue	Bonferroni
GO:0006413~translational initiation †	2.71E-06	0.002
GO:0006412~translation †	1.35E-05	0.01
GO:0005852~eukaryotic translation initiation factor 3 complex	1.49E-05	0.003
GO:0003743~translation initiation factor activity	4.99E-05	0.01
GO:0008135~translation factor activity, nucleic acid binding	6.70E-04	0.18
GO:0030027~lamellipodium	1.88E-03	0.31
GO:0031252~cell leading edge	3.37E-03	0.48
GO:0006644~phospholipid metabolic process	3.55E-03	0.96
GO:0019637~organophosphate metabolic process	4.95E-03	0.99
GO:0005739~mitochondrion †	8.79E-03	0.82

Table 7.4 Pathways overrepresented for genes with significant differential expression at p<0.01 in the analysis of embryonic brains using DAVID (v6.7). In the GO ontologies, KEGG and Panther databases, 4 pathways were significantly enriched after Bonferroni correction. † Pathway is in the top 10 when using the p value threshold p≤0.05 in corresponding DAVID analysis

Metacore Non proprietary GO ontologies Taking the same list of 191 genes which show differential expression in the embryonic expression data at p≤0.01 and analysing them in Metacore using the GO onotology pathways (Table 7.5). 8 of the top 10 DAVID pathways were also significantly enriched using the same GO ontologies in Metacore. These pathways related to translation and the mitochondria.

Differentially Expressed Genes between C59X mutants and Wildtypes (p≤0.01)

Metacore GO Biological Processes	pValue	FDR
translation* †	1.36E-05	Significant
translational initiation* †	3.29E-05	Significant
cellular protein metabolic process	4.13E-05	Significant
protein metabolic process	5.35E-05	Significant
sulfur amino acid metabolic process	1.30E-04	Not Significant
cellular metabolic process	5.09E-04	Not Significant
metabolic process	5.80E-04	Not Significant
cellular carbohydrate metabolic process	6.92E-04	Not Significant
phospholipid metabolic process*	1.30E-03	Not Significant
organophosphate metabolic process*	1.49E-03	Not Significant

Metacore GO Molecular functions	pValue	FDR
translation initiation factor activity* †	3.10E-06	Significant
translation factor activity, nucleic acid binding* †	1.39E-04	Significant
catalytic activity †	2.60E-04	Significant
intramolecular oxidoreductase activity, interconverting aldoses and ketoses	5.00E-04	Not Significant
isomerase activity	1.00E-03	Not Significant
alditol:NADP+ 1-oxidoreductase activity	1.31E-03	Not Significant
S-methyl-5-thioribose-1-phosphate isomerise activity	1.31E-03	Not Significant
tRNA (guanine) methyltransferase activity	1.31E-03	Not Significant
monosaccharide binding	1.47E-03	Not Significant
oxidoreductase activity †	2.24E-03	Not Significant

Metacore GO Localizations	pValue	FDR
eukaryotic translation initiation factor 3 complex*	3.46E-07	Signficiant
cytoplasmic part †	6.51E-06	Significant
cytoplasm †	2.43E-05	Significant
mitochondrial matrix	9.68E-04	Not Significant
eukaryotic translation initiation factor 2B complex	1.38E-03	Not Significant
mitochondrion* †	1.96E-03	Not Significant
intracellular part	2.58E-03	Not Significant
Intracellular	2.76E-03	Not Significant
mitochondrial part †	3.19E-03	Not Significant
Arp2/3 protein complex	3.38E-03	Not Significant

Table 7.5 Pathways overrepresented for genes with significant differential expression at p<0.01 in the analysis of embryonic brains using Metacore Non proprietary GO databases. * Represents pathways that were one of the top 10 pathways in the DAVID analysis of the same gene list (191 genes) † Represents pathway that is in the top 10 when using p value threshold p≤0.05 in corresponding Metacore analysis. **FDR** p value following Benjamini & Hochberg (1995) False Discovery Rate Correction, set at a 0.05 threshold.

Metacore Proprietry GeneGO database: The GeneGO databases are curated by Metacore and use a different method to assign genes into relevant pathways (7.2.1.2). Therefore the lists of genes in each pathway will not be exactly the same. Despite this, pathways relating to 'translation initiation' and 'regulation of translation initiation' were the most significantly enriched pathways for genes with differential expression in the embryonic data set in both the GeneGO datasets. None of the pathways were significant following multiple test correction using a FDR of threshold of 0.05. The pathway 'translation in mitochondria' is significant and the 'mitochondrion' pathway was a significant GO cellular component when using the GO ontologies.

Differentially Expressed Genes between C59X mutants and Wildtypes (p≤0.01)

Metacore GeneGO Pathway Maps	pValue	FDR
Translation _Regulation of translation initiation	1.48E-03	Not Significant
Immune response_MIF-JAB1 signalling	3.35E-03	Not Significant
Estrone metabolism / Human version	8.68E-03	Not Significant
Estrone metabolism	8.68E-03	Not Significant
Androstenedione and testosterone biosynthesis and metabolism p.1	0.01	Not Significant
Androstenedione and testosterone biosynthesis and metabolism p.1/ Rodent version	0.01	Not Significant
Sphingolipid metabolism / Human version	0.02	Not Significant
Sphingolipid metabolism	0.02	Not Significant
Apoptosis and survival_NGF activation of NF-kB	0.03	Not Significant
Oxidative phosphorylation †	0.03	Not Significant

Metacore GeneGO Process Networks	pValue	FDR
Translation_Translation initiation †	6.12E-04	Not Significant
Apoptosis_Endoplasmic reticulum stress pathway	1.74E-03	Not Significant
Translation_Regulation of initiation †	2.55E-03	Not Significant
Translation_Translation in mitochondria †	9.25E-03	Not Significant
Inflammation_MIF signalling	0.02	Not Significant
Inflammation_Kallikrein-kinin system	0.05	Not Significant
Immune response_IL-5 signalling †	0.07	Not Significant
Blood coagulation	0.07	Not Significant
Inflammation_Inflammasome	0.09	Not Significant
Neurophysiological process_Long-term potentiation	0.12	Not Significant

Table 7.6 Pathways overrepresented for genes with significant differential expression at p<0.01 in the analysis of embryonic brains using Metacore's propriety GeneGO databases. Pathways relating to 'translation initiation' were significant using this gene list as well as when using a larger gene list with less stringent p values (p≤0.05). † Pathway is in the top 10 when using p value threshold p≤0.05 in corresponding Metacore GeneGO analysis. **FDR** significance following Benjamini & Hochberg (1995) False Discovery Rate Correction, set at 0.05 threshold.

Using the same GO ontologies but different pathway analysis tools and algorithms produced consistent results. There was also concordance with Metacore's own curated databases. Genes differentially expressed in C59X mutants appeared to be consistently overrepresented in pathways relating to translation and the mitochondrion. Translation initiation is known to negatively regulate gene expression in response to stress (Harding et al., 2000). Differential expression in a number of components of the translation initiation pathway in C59X mutants could affect the rate of translation under certain cell conditions such as stress.

7.3.1.2 Pathways Enriched for Genes Differentially Spliced between C59X Mutants and wildtypes.

Selection for genes differentially spliced at $p<0.01$ and $p<0.001$ resulted in lists of 530 and 135 unique Entrez IDs respectively which were compared with the background set of genes (12,943 unique Entrez IDs). All 530 Entrez IDs mapped to DAVID IDs and all identifiers were recognised in Metacore. The analyses were conducted on the same databases as described in 7.3.1.1.

$p \leq 0.01$

Entering the list of 530 genes into DAVID resulted in no pathways that were significant following Bonferroni correction (Table 7.7). The most significant pathways was 'axon guidance' from the KEGG database. The GO biological process 'translation' was also significant for differentially spliced genes as observed for the genes showing differential expression in the embryonic data.

Pathways enriched for Genes Differentially Spliced between C59X mutants and Wildtypes (p≤0.01)

DAVID GO, KEGG, PANTHER	pValue	Bonferroni
mmu04360:Axon guidance	5.73E-04	0.07
GO:0031252~cell leading edge	6.10E-04	0.20
GO:0006412~translation	1.13E-03	0.86
GO:0006418~tRNA aminoacylation for protein translation	1.82E-03	0.96
GO:0043038~amino acid activation	1.82E-03	0.96
GO:0043039~tRNA aminoacylation	1.82E-03	0.96
GO:0005768~endosome	1.85E-03	0.49
P00034:Integrin signalling pathway	2.24E-03	0.17
GO:0016876~ligase activity, forming aminoacyl-tRNA and related compounds	2.43E-03	0.72
GO:0016875~ligase activity, forming carbon-oxygen bonds	2.43E-03	0.72

Table 7.7 Pathways overrepresented for genes with significant differential splicing at p≤0.01 in the analysis of embryonic brains using DAVID (v6.7). The most significant pathways was axon guidance from the KEGG database.

Non proprietary GO ontologies. Considering the same GO ontologies in Metacore the biological process 'axon guidance' was the most significant pathway but did not survive the FDR correction for all the pathways in the GO category (Table 7.8). Although the 'axon guidance' pathway that was top of the DAVID list was from the KEGG database the fact that the same pathway from two different databases is the most significant in DAVID and Metacore suggests that genes differentially spliced between C59X mutants and wildtypes are enriched in pathways related to axon guidance.

Pathways enriched for Genes Differentially Spliced between C59X mutants and Wildtypes (p≤0.01)

Metacore GO Biological Processes	pValue	FDR
axon guidance*	3.26E-05	Not Significant
cell morphogenesis involved in differentiation	3.46E-05	Not Significant
cell-substrate adhesion	4.36E-05	Not Significant
chemotaxis	6.64E-05	Not Significant
Taxis	6.95E-05	Not Significant
cellular component organization or biogenesis at cellular level	1.07E-04	Not Significant
cell morphogenesis involved in neuron differentiation	1.39E-04	Not Significant
cellular component organization at cellular level	1.69E-04	Not Significant
semaphorin-plexin signalling pathway	1.88E-04	Not Significant
axonogenesis	2.01E-04	Not Significant

Metacore GO Molecular functions	pValue	FDR
protein binding	1.66E-06	Significant
binding	2.26E-05	Significant
semaphorin receptor activity	3.40E-04	Not Significant
transferase activity, transferring acyl groups other than amino-acyl groups	4.69E-04	Not Significant
choline transmembrane transporter activity	7.11E-04	Not Significant
transferase activity, transferring acyl groups	9.56E-04	Not Significant
actin binding	1.36E-03	Not Significant
semaphorin receptor binding	1.38E-03	Not Significant
monovalent cation:hydrogen antiporter activity	1.38E-03	Not Significant
ligase activity, forming aminoacyl-tRNA and related compounds*	1.76E-03	Not Significant

Metacore GO Localizations	pValue	FDR
organelle part	1.62E-06	Significant
cell leading edge*	1.68E-06	Significant
lamellipodium	3.81E-06	Significant
intracellular organelle part	4.05E-06	Significant
stress fiber	1.88E-04	Significant
actin filament bundle	3.01E-04	Significant
cell body	3.93E-04	Significant
cell projection	4.14E-04	Significant
endosome*	5.15E-04	Significant
actomyosin	5.34E-04	Significant

Table 7.8 Pathways overrepresented for genes with significant differential splicing at p≤0.01 in the analysis of embryonic brains using Metacore. The most significant biological process was 'axon guidance', which was also the most significant pathway in the DAVID analysis. * Pathway was one of the top 10 pathways in the corresponding DAVID analysis.

Metacore Proprietry GeneGO database: Using Metacore's proprietary GeneGO databases 2 pathways were significant following FDR correction using a 0.05 threshold; the proprietary pathway map 'cell adhesion ECM remodelling' ($p = 4.24x10^{-5}$) and the proprietary process network 'development neurogenesis axonal guidance ($p = 1.33x10^{-5}$). Both pathways are significantly enriched following an FDR correction at a threshold of 0.05 (Table 7.9).

Pathways enriched for Genes Differentially Spliced between C59X mutants and Wildtypes (p≤0.01)

Metacore GeneGO Pathway Maps	pValue	FDR
Cell adhesion_ECM remodelling	4.243E-05	Significant
Cell adhesion_Chemokines and adhesion	1.342E-03	Not Significant
Cell adhesion_Endothelial cell contacts by non-junctional mechanisms	5.113E-03	Not Significant
Cytoskeleton remodeling_Integrin outside-in signaling	5.859E-03	Not Significant
Development_Role of HDAC and calcium/calmodulin-dependent kinase (CaMK) in control of skeletal myogenesis	7.558E-03	Not Significant
Cytoskeleton remodeling_Cytoskeleton remodeling	8.810E-03	Not Significant
Transport_RAB5A regulation pathway	1.623E-02	Not Significant
Development_Role of CDK5 in neuronal development	1.623E-02	Not Significant
Nitrogen metabolism	1.743E-02	Not Significant
Development_VEGF-family signalling	1.911E-02	Not Significant

Metacore GeneGO Process Networks	pValue	FDR
Development_Neurogenesis_Axonal guidance	1.328E-05	Significant
Cytoskeleton_Actin filaments	1.825E-03	Not Significant
Inflammation_Complement system	3.864E-03	Not Significant
Cell adhesion_Integrin-mediated cell-matrix adhesion	6.286E-03	Not Significant
Cell adhesion_Attractive and repulsive receptors	9.440E-03	Not Significant
Signal transduction_Androgen receptor nuclear signalling	1.412E-02	Not Significant
DNA damage_Core	1.898E-02	Not Significant
Cell adhesion_Synaptic contact	1.988E-02	Not Significant
DNA damage_Checkpoint	2.413E-02	Not Significant
Cell cycle_G2-M	2.928E-02	Not Significant

Table 7.9 Pathways overrepresented for genes with significant differential splicing at p≤0.01 in the analysis of embryonic brains using Metacore's Proprietary GeneGO databases. Following a FDR correction for all pathways in the database both The 'Cell adhesion ECM remodelling' and 'Development Neurogenesis Axonal guidance' were significant.

p≤0.001

The most significant pathway in DAVID when using the more stringent p≤0.001 threshold was the KEGG pathway 'axon guidance' ($p = 8.09x10^{-5}$) and it remained significant after a Bonferroni correction for all pathways tested (p = 0.006) (Table 7.10). This pathway was also the most significant pathway when using the p≤0.01 threshold, but did not survive the Bonferroni correction. 8 genes differentially spliced at p<0.001 are found in this pathway (*Epha2, Sema6b, Plxnb2, Sema6a, Rac3, Robo1, Plxna1 and Robo3*). The 'cell adhesion' pathway was also significant.

Pathways enriched for Genes Differentially Spliced between C59X mutants and Wildtypes (p≤0.001)

DAVID GO, KEGG, PANTHER	pValue	Bonferroni
mmu04360:Axon guidance †	8.09E-05	5.57E-03
GO:0044420~extracellular matrix part	7.66E-04	0.14
GO:0005604~basement membrane	2.55E-03	0.40
GO:0007155~cell adhesion	4.01E-03	0.97
GO:0022610~biological adhesion	4.06E-03	0.97
GO:0015629~actin cytoskeleton	5.45E-03	0.66
GO:0030036~actin cytoskeleton organization	9.06E-03	1.00
GO:0031252~cell leading edge †	1.16E-02	0.90
GO:0030029~actin filament-based process	1.17E-02	1.00
GO:0030030~cell projection organization	1.18E-02	1.00

Table 7.10 Pathways overrepresented for genes with significant differential splicing at p≤0.001 in the analysis of embryonic brains using DAVID (v6.7). The most significant pathway was axon guidance as observed at p≤0.01. This pathway was the only one to remain significant following a Bonferroni correction. † Pathway is in the top 10 pathways in corresponding DAVID analysis at p value threshold p≤0.01.

Non proprietary GO ontologies: 6 of the top 10 pathways from DAVID overlap in Metacore GO analysis of same 3 databases at p≤0.001 threshold (Table 7.11). 'Cell adhesion' ($p = 9.27x10^{-5}$) and 'biological adhesion' ($p = 9.86x10^{-5}$) were significantly enriched biological processes, but did not remain significant following FDR correction. Both these pathways were nominally significant in DAVID (p = 0.01 for both) but neither had survived Bonferroni correction for all pathways. Interestingly, these same 2 pathways ('cell adhesion' and 'biological adhesion') were also identified as enriched for genes showing differential expression following knockdown of ZNF804A in a neural cell line (Hill et al., 2012). Of the genes within these pathways one gene, *Lama4* was found to be differentially spliced between C59X mutants and wildtypes (p=0.00048) and also differentially expressed following knockdown of *ZNF804A* (significant using

two ZNF804A siRNA conditions p=0.0139 p=0.0155, Hill et al., 2012). Mutations in this gene cause a mild muscular dystrophy (Patton et al., 2001) and bleeding disorder (Thyboll et al., 2002).

Pathways enriched for Genes Differentially Spliced between C59X mutants and Wildtypes (p≤0.001)

Metacore GO Biological Processes	pValue	FDR
cell adhesion*	9.27E-05	Not Significant
biological adhesion*	9.86E-05	Not Significant
regulation of neuron migration	1.13E-04	Not Significant
cellular process	2.47E-04	Not Significant
positive regulation of histone H3-K9 methylation	3.35E-04	Not Significant
semaphorin-plexin signalling pathway †	3.94E-04	Not Significant
cellular component organization or biogenesis at cellular level †	5.02E-04	Not Significant
negative regulation of translational initiation in response to stress	6.65E-04	Not Significant
regulation of translational initiation in response to stress	6.65E-04	Not Significant
cellular component organization or biogenesis	7.28E-04	Not Significant
Metacore GO Molecular functions	PValue	FDR
transferase activity, transferring acyl groups other than amino-acyl groups †	1.30E-05	Significant
transferase activity, transferring acyl groups †	1.00E-04	Significant
binding †	2.92E-04	Significant
tyrosine-tRNA ligase activity	3.48E-04	Significant
interleukin-8 receptor binding	3.48E-04	Significant
sphingosine N-acyltransferase activity	3.48E-04	Significant
N-acyltransferase activity	5.09E-04	Significant
protein binding †	5.89E-04	Significant
aminoacyl-tRNA ligase activity	1.08E-03	Significant
ligase activity, forming aminoacyl-tRNA and related compounds †	1.08E-03	Significant
Metacore GO Localizations	**pValue**	**FDR**
lamellipodium †	1.40E-04	Significant
cell leading edge* †	1.98E-04	Significant
actin cytoskeleton*	6.56E-04	Not Significant
eukaryotic translation initiation factor 2B complex	7.75E-04	Not Significant
filamentous actin	8.98E-04	Not Significant
actin filament	1.73E-03	Not Significant
stress fibre †	2.37E-03	Not Significant
basement membrane*	2.50E-03	Not Significant
extracellular matrix part*	2.55E-03	Not Significant
organelle part †	2.69E-03	Not Significant

Table 7.11. Pathways overrepresented for genes with significant differential splicing at p≤0.001 in the analysis of embryonic brains using Metacore. 'Cell adhesion' and 'biological adhesion' were the most significant biological processes. * Pathway was one of the top 10 pathways in the corresponding DAVID analysis. † Pathway is in the top 10 pathways in the Metacore analysis when using the less stringent p value threshold p≤0.01.

***Proprietry GeneGO database*:** When genes meeting the more stringent p value threshold ($p \leq 0.001$) for differential splicing were examined using the proprietary database, the same 'development neurogenesis axonal guidance' pathway was significantly enriched ($p = 1.57 \times 10^{-4}$) as seen with the less stringent p value threshold ($p \leq 0.01$) as was the 'cell adhesion ECM remodelling' pathway, ($p = 0.017$) although only the former remained significant following multiple test correction (Table 7.12). *Lama4* is also found in the Metacore proprietary GeneGO database in the 'cell adhesion ECM remodelling' pathway.

Pathways enriched for Genes Differentially Spliced between C59X mutants and Wildtypes (p≤0.001)

Metacore GeneGO Pathway Maps	pValue	FDR
Development_Slit-Robo signaling	2.93E-03	Not Significant
Immune response_Classical complement pathway	3.38E-03	Not Significant
Immune response_Lectin induced complement pathway	3.88E-03	Not Significant
DNA damage_DNA-damage-induced responses	5.71E-03	Not Significant
DNA damage_Role of NFBD1 in DNA damage response	1.09E-02	Not Significant
Nitrogen metabolism †	1.30E-02	Not Significant
Nitrogen metabolism/ Rodent version	1.52E-02	Not Significant
Cell adhesion_ECM remodelling †	1.68E-02	Not Significant
Cytoskeleton remodeling_Fibronectin-binding integrins in cell motility	2.00E-02	Not Significant
Cell adhesion_Alpha-4 integrins in cell migration and adhesion	2.84E-02	Not Significant

Metacore GeneGO Process Networks	pValue	FDR
Development_Neurogenesis_Axonal guidance †	1.57E-04	Significant
Cytoskeleton_Actin filaments †	1.94E-03	Not Significant
Cell adhesion_Attractive and repulsive receptors †	2.98E-03	Not Significant
Inflammation_Complement system †	1.18E-02	Not Significant
Cytoskeleton_Regulation of cytoskeleton rearrangement	2.53E-02	Not Significant
DNA damage_BER-NER repair	3.35E-02	Not Significant
DNA damage_Checkpoint †	4.06E-02	Not Significant
DNA damage_Core †	4.13E-02	Not Significant
Protein folding_Protein folding nucleus	7.83E-02	Not Significant
Transcription_Transcription by RNA polymerase II	8.14E-02	Not Significant

Table 7.12. Pathways overrepresented for genes with significant differential splicing at p≤0.001 in the analysis of embryonic brains using Metacore's proprietary GeneGO databases. The only pathway that was significant following the FDR correction was the process network 'development neurogenesis axonal guidance'.

The KEGG Pathway of 'axon guidance' significantly enriched in the DAVID analysis for gene sets showing differential splicing at both thresholds, comprises genes involved in axon guidance during brain development and is therefore an attractive candidate pathway for schizophrenia. As this is a KEGG pathway it is not tested in Metacore, however the Metacore proprietary pathway of 'development neurogenesis axonal guidance' was FDR significant at both thresholds. There were 6 differentially spliced genes that were common to the 'axon guidance' pathway (at the p≤0.001 P value threshold) in KEGG and the proprietary Metacore GeneGo process 'Development Neurogenesis Axonal Guidance'; *Robo1, Robo3, Epha2*, *Plxna1, Plxnb2* and *Sema6a*. Multiple isoforms of both *Robo1* and *Robo3* have been demonstrated, the different isoforms of *Robo3* having been shown to have different functions and roles in

embryogenesis (Camurri et al., 2005). Both are also cell adhesion molecules. Different splice variants can have a large impact on the function of this general class of molecules (Walsh & Doherty, 1991). In addition to roles in axon guidance, the Plexin *Plxnb2* is known to have important roles during the development of the neocortex including the generation, differentiation and migration of cortical cells (Hirschberg et al., 2010).

Using the more stringent threshold ($P \leq 0.001$) for selecting differentially spliced genes, the GO molecular function pathway 'interleukin-8 receptor binding' was one of the significant pathways in Metacore ($p = 3.48 \times 10^{-4}$) that survives an FDR correction (0.05 threshold). Increased levels of IL-8 have been observed in the serum of mothers of patients with schizophrenia spectrum disorder (Brown et al., 2004) and have been thought to be associated with aberrant brain development (Gilmore & Jarskog, 1997).

7.3.1.3. The effect of Probeset Number on Pathway Analysis.

In Chapter 6.3.4 I showed a significant correlation between the number of probes and the differential splice p value. Although the correlation was weak ($r = 0.061$) the effect was significant ($P<0.01$). Whilst that analysis could not distinguish between the possibility that such genes tend to be more significant due to a true biological effect (increased occurrence of altered splicing in genes with more exons) or because of the effects of multiple testing on genes with more probesets, both phenomena could bias the results of pathway analyses. That is if the main driver of significance with respect to differential splicing is simply number of probesets, pathways identified as being enriched for differentially spliced genes could simply reflect pathways which happen to be comprised of genes containing large numbers of probesets regardless of the function of those genes.

To determine if the significant pathways were enriched for genes with large numbers of probesets, I re-ran the pathway analyses using a gene list selected on the basis of probeset number rather than differential splicing p value. Genes were first ranked by number of probesets (range from 2-119 for the whole dataset). To make the comparison between differentially spliced genes and those selected by probeset number, I ensured the number of genes in the latter list was comparable to that selected by splicing p value (at a threshold $P \leq 0.01$). Thus, I selected the 530 genes with the most probesets which

also had Entrez IDs. In these 530 genes the probeset number ranged from 29-119. As before, gene symbols were converted to unique Entrez IDs and, as before, this gene set was compared using both Metacore and DAVID against the same background list as previously used (n = 12,943). All 530 IDs were recognised in Metacore and all 530 mapped to DAVID IDs. The same ontologies were investigated as described in 7.3.1.1.

In DAVID, 98 pathways were significant following Bonferroni correction. In both DAVID (Table 7.13) and Metacore the GO biological processes of 'biological adhesion' and 'cell adhesion' were significantly enriched for the geneset selected by probeset number (Metacore; cell adhesion $p = 1.21x10^{-16}$, biological adhesion $p = 1.59x10^{-16}$, both highly significant following multiple test correction (FDR 0.05 in Metacore).

Pathways enriched for the 530 genes with the most probesets

DAVID GO, KEGG, PANTHER	pValue	Bonferroni
GO:0005581~collagen	4.06E-19	1.62E-16
GO:0044420~extracellular matrix part	3.99E-18	1.59E-15
GO:0005201~extracellular matrix structural constituent	5.60E-17	5.50E-14
GO:0005524~ATP binding	6.20E-15	3.08E-12
GO:0022610~biological adhesion	9.76E-15	1.98E-11
GO:0007155~cell adhesion	9.76E-15	1.98E-11
GO:0032559~adenyl ribonucleotide binding	1.31E-14	6.48E-12
GO:0008092~cytoskeletal protein binding	7.95E-14	3.93E-11
GO:0001882~nucleoside binding	8.96E-14	4.43E-11
GO:0003774~motor activity	1.90E-13	9.42E-11

Table 7.13 Top 10 Significant Pathways enriched for Genes with the Highest Number of Probesets in DAVID (v6.7). The 10 most significant pathways enriched for 530 genes with the largest number of probesets included the biological processes of 'cell adhesion' and 'biological adhesion' and the cellular component term 'extracellular matrix part'.

Considering the Metacore proprietary GeneGO databases the process network 'Development Neurogenesis Axonal guidance' was the most significant pathway following FDR correction (0.05 threshold) (Table 7.14). This pathway was also the most significant pathway in the differential splice data.

Pathways enriched for the 530 genes with the most probesets

Metacore GeneGO Pathway Maps	pValue	FDR
cAMP/ Ca(2+)-dependent Insulin secretion	7.49E-05	Significant
Neurophysiological process_Receptor-mediated axon growth repulsion	1.28E-04	Significant
wtCFTR and delta508 traffic / Clathrin coated vesicles formation (norm and CF)	3.88E-04	Significant
Immune response_Classical complement pathway	6.19E-04	Significant
Immune response_Alternative complement pathway	6.19E-04	Significant
Immune response_Lectin induced complement pathway	8.14E-04	Significant
Transcription_Ligand-Dependent Transcription of Retinoid-Target genes	8.29E-04	Significant
Neurophysiological process_ACM regulation of nerve impulse	2.07E-03	Not Significant
Neurophysiological process_Netrin-1 in regulation of axon guidance	2.07E-03	Not Significant
Development_Osteopontin signaling in osteoclasts	5.33E-03	Not Significant

Metacore GeneGO Process Networks	pValue	FDR
Development_Neurogenesis_Axonal guidance	1.38E-08	Significant
Cell adhesion_Cell-matrix interactions	2.00E-08	Significant
Cytoskeleton_Actin filaments	3.73E-08	Significant
Cell adhesion_Attractive and repulsive receptors	4.26E-08	Significant
Cytoskeleton_Regulation of cytoskeleton rearrangement	1.17E-03	Significant
Cell adhesion_Cadherins	3.57E-03	Not Significant
Development_Neuromuscular junction	5.27E-03	Not Significant
Development_Cartilage development	5.65E-03	Not Significant
Cell adhesion_Integrin-mediated cell-matrix adhesion	1.81E-02	Not Significant
Signal transduction_Androgen receptor signaling cross-talk	1.99E-02	Not Significant

Table 7.14. Top 10 Significant Pathways enriched for Genes with the Highest Number of Probesets in Metacore Proprietary GeneGO databases. The most significant process network was 'development neurogenesis axonal guidance'**.**

Evidently, the same pathways that were enriched for differentially spliced genes were also enriched for genes with large probeset number. This does not imply that the true experimental pathway results were being driven by a statistical or biological artefact arising from probeset number, but given the significant (though weak) correlation between splicing p value and probeset number, it is consistent with the hypothesis that they might be. It was therefore important to adjust all pathway analyses by probeset number. The issue of gene size has been addressed previously for pathway analysis of GWAS data and RNAseq data (Jia et al., 2011; Young et al., 2010), but review of the literature revealed no publications which have endeavoured to make such corrections for exon array splicing data.

To adjust for the effect of probeset number on pathway analysis I carried out a linear regression between the number of probesets in a gene and the –log p value. In such an analysis, the standardised residual essentially reflects the contribution to the p value that is independent of the effect of probeset number. Genes were then ranked on the basis of the largest to smallest residuals, in effect identifying ranking the genes based upon their splicing p value independent of probeset number. To allow a direct comparison to the previous analysis of the experimental groups, the top 530 ranked genes were taken. Of these, 500 genes were also present in the unadjusted 530 most significantly differentially spliced genelist. This almost complete overlap in the top gene sets for adjusted and unadjusted analyses implying most of the variance in splicing p values was not related to probeset number, a result not entirely surprising given the correlation between p value and probeset number was weak.

Pathway analysis in both Metacore and DAVID was then carried out with these genes using all criteria as described previously (7.3.1.1). When correcting for gene size the 'axon guidance' pathway in the KEGG database remained the most significant pathway and the GO biological processes 'axon guidance' was significant as was the GeneGO process network ' development neurogenesis axonal guidance. Only the GeneGO pathway remains significant following multiple test correction using the FDR (0.05 threshold). 'Cell adhesion' remains significant in DAVID and Metacore GO analysis and 'cell adhesion ECM remodelling' was still a significant Metacore proprietary GeneGO pathway.

The overlap of results from the analyses in Metacore was determined using contingency tables (described in Chapter 3.2.11) to compare the outputs based upon the top sets derived from the adjusted and unadjusted analyses in GeneGO Pathway Maps and GO Molecular Functions. A significant overlap was observed for both ($p < 0.01$). Genes that showed significant differential splicing were still enriched in pathways relating to axon guidance in particular during development and cell adhesion.

7.3.2 Genetic Analysis

7.3.2.1 Investigating PGC Schizophrenia Top Hits for Significant Differential Expression and Splicing in the Embryonic Dataset.

The Psychiatric GWAS consortium (PGC) identified genome-wide significant association for 10 loci in schizophrenia. The mouse orthologues of the genes or miRNA nearest to each of the 10 SNPs were specifically investigated to determine if any of them were significantly differentially expressed or spliced in the C59X mutants. Due to the LD structure within the MHC (6p21.3-p22.1) this region was excluded. Both *miR137* and *PCGEM1* were also excluded as they did not qualify for the analysis as *miR137* was targeted by no core probesets and *PCGEM1* was on chromosome 2. None of the top PGC genes showed differential splicing between those with and without the C59X mutation (Table 7.5). *Ccdc68* showed nominally significant differential expression (p=0.02). The function of CCDC68 is not known. A 1.2 Mb deletion including this gene and TCF4 has been identified in a patient with Pitt-Hopkins Syndrome, although the resultant phenotype is likely attributable to the deletion of TCF4 (Zweier et al., 2007).

Gene Symbol	RefSeq	Gene Assignment	Differential Expression Pvalue	Differential Splice Pvalue
Csmd1	NM_053171	CUB and Sushi multiple domains 1	0.60	0.98
Mmp16	NM_019724	matrix metallopeptidase 16	0.09	0.66
Cnnm2	NM_033569	cyclin M2	0.11	0.62
Nt5c2	NR_028353	5'-nucleotidase, cytosolic II	0.54	0.56
Stt3a	NM_008408	STT3, subunit of the oligosaccharyltransferase complex, homolog	0.30	0.53
Ccdc68	NM_201362	Coiled-coil domain containing 68	**0.02**	0.63
Tcf4	NM_013685	transcription factor 4	0.76	0.49

Table 7.5 PGC Schizophrenia Top Genome-Wide Association Results. No altered splicing was observed in the top PGC schizophrenia genes in C59X mutants. Only Ccdc68 showed nominally significant altered expression in mice with the C59X mutation. Both *Mir137* and *PCGEM*1 were excluded as no core probesets target *miR137* and *PCGEM1* is located on chromosome 2 which had been excluded from the exon array analyses.

7.3.2.2 Relevance of Genes Differentially Expressed in C59X Mutants to Disease Risk.

My hypothesis was that genes identified as showing differential expression or splicing in mutants compared with wildtype would include genes that are downstream mediators of the effects of ZNF804A on disease risk, and that the gene sets might be enriched for genes associated with psychosis (schizophrenia and bipolar disorder). In the PGC datasets (https://pgc.unc.edu/), the differentially expressed gene sets (Table 7.6) were not clearly significantly enriched for genes showing nominally significant evidence for association to either disorder, though the geneset selected at a differential expression threshold p≤0.05 was nominally significantly enriched for genes showing evidence for association to bipolar disorder using Brown's method.

Differential Expression	**Brown SZ**		**Simes' SZ**	
	OR	**P VALUE**	**OR**	**P VALUE**
p≤0.05	1.20	0.10	1.32	0.06
P≤0.01	0.77	0.31	0.91	0.42
P≤0.001	4.01	0.12	2.05	0.30
P≤0.0001	4.66	0.93	2.74	0.89

Differential Expression	**Brown BP**		**Simes' BP**	
	OR	**P VALUE**	**OR**	**P VALUE**
p≤0.05	1.38	**0.01**	1.20	0.06
P≤0.01	1.05	0.50	1.37	0.13
P≤0.001	0.88	0.60	1.31	0.56
P≤0.0001	No Genes Significant		3.49	0.91

Table 7.6 Genetic Relevance of Genes Differentially Expressed in C59X Mutants. OR is the odds ratio that a gene will be nominally significant in the respective PGC dataset conditional on it being significantly differential expressed at the indicated P value threshold in the embryonic expression data. **Brown** Brown method used to calculate gene wide significance. **Simes'** Simes' method used to calculate gene wide significance. **SZ** schizophrenia **BP** bipolar disorder.

7.3.2.3 Relevance of Genes Differentially Spliced in C59X Mutants to Disease Risk.

Genes that were significantly differentially spliced (Table 7.7) were also not enriched for genes showing evidence (P≤0.05) for association with schizophrenia. Three of the tests in the bipolar dataset were significant but there was inconsistency between the Simes' and Brown generated gene-wide p values. Overall, there is some weak evidence that genes identified as being differentially expressed and spliced are enriched for genes showing evidence for association to bipolar disorder but the findings are far from conclusive.

Differential Splicing	**Brown SZ**		**Simes' SZ**	
	OR	**P VALUE**	**OR**	**P VALUE**
p≤0.05	0.94	0.34	1.12	0.13
P≤0.01	1.10	0.33	1.10	0.27
P≤0.001	1.04	0.51	1.11	0.40
P≤0.0001	0.57	0.33	1.10	0.48

Differential Splicing	**Brown BP**		**Simes' BP**	
	OR	**P VALUE**	**OR**	**P VALUE**
p≤0.05	1.22	0.06	1.28	**0.01**
P≤0.01	1.50	**0.01**	1.21	0.13
P≤0.001	1.59	0.10	1.53	0.09
P≤0.0001	2.44	**0.04**	1.95	0.07

Table 7.7 Relevance of Genes Differentially Spliced in C59X Mutant to Disease Risk. OR is the odds ratio that a gene will be nominally significant in the respective PGC dataset conditional on it being significantly differential spliced at the indicated P value threshold in the embryonic expression data. **Brown** Brown method used to calculate gene wide significance. **Simes'** Simes' method used to calculate gene wide significance. **SZ** schizophrenia **BP** bipolar disorder.

7.4 Discussion.

Pathways enriched for genes showing evidence for differential expression and splicing were investigated using DAVID and Metacore which are both widely used pathway analysis tools. Despite the different algorithms a consensus between the two sets of results was apparent. Genes with altered expression are enriched in pathways related to translation, most robustly 'translation initiation'. Translation rate in a cell is known to alter in response to certain conditions where cells need to respond rapidly such as stress (Sonenberg & Hinnebusch, 2009).

Genes with altered splicing were enriched in the pathway 'axon guidance' from the KEGG database as well as the GeneGO process network 'Development Neurogenesis Axonal Guidance' and both were significant following Bonferroni and FDR correction respectively. These pathways are involved in developing the neuronal network. The gene *Robo3* found in this pathway was differentially spliced in C59X mutants. The alternative isoforms of this gene are known to have distinct roles in embryogenesis (Camurri et al., 2005). The *disrupted in schizophrenia 1* (*DISC1*) gene, which as its name suggests, has been implicated in schizophrenia and other psychiatric disorders (Millar et al., 2000) is thought to have a role in axon guidance (Chen et al., 2011), as is semaphorin *Sema3a*, expression of which has been reported to be increased in schizophrenia in the cerebellum (Eastwood et al., 2003). The pathway 'SLIT/ROBO axon guidance' was significantly enriched in genes differentially expressed between neurons derived from human inducible pluripotent stem cells (hiPSC) from schizophrenia fibroblast reprogramming compared to control fibroblast reprogramming (Brennand et al., 2011).

One of the prominent functional pathways enriched in genes which show differential splicing was 'cell adhesion'. Initially regarded with extreme caution as it contains genes with large numbers of probesets, further investigation revealed that when controlling for the number of probesets, cell adhesion categories were still amongst the most significantly enriched pathways. This suggests the 'cell adhesion' pathway is enriched for genes which are directly or indirectly regulated by Zfp804a.

The KEGG pathway 'cell adhesion molecule' has been observed to be enriched for genes with significant association signals in schizophrenia and bipolar disorder GWAS data (O'Dulshaine et al., 2011). The 'cell adhesion molecule' pathway genes *NRXN1* and *CNTNAP2* were associated in both association datasets (O'Dulshaine et al., 2011).

Genes involved in cell adhesion have also been reported as associated with autism spectrum disorder (ASD) in a GWAS study (Wang et al., 2009). Neither *NRXN1* or *CNTNAP2* showed evidence for differential splicing between C59X mutants and wildtypes but both are found in the GO biological process of 'cell adhesion' (which differs to the KEGG pathway of 'cell adhesion molecule'), which was significantly enriched for differentially spliced genes following correction for the number of probesets, though this did not survive the stringent Bonferroni correction (nominal p = 0.03, Bonferroni p value = 1). This is the same pathway significantly enriched for genes differentially expressed between ZNF804A knockdown and wildtype (Hill et al., 2012) which also did not survive Bonferroni correction for all pathways in the relevant GO database.

An RNA sequencing study looking at expression differences in differentiating and mature neurons found increased expression in differentiating neurons in several genes (*NRXN1*, *NRXN3*, *NLGN1*, *CTNNA2*, *NCAM1*, *CHL1*, *ELAVL4* and *PCDH9*) with functions in cell adhesion (Lin et al., 2011). The neurexins and neuroligins form part of this pathway and are necessary for effective neurotransmission and have been associated with schizophrenia (Kirov et al., 2009b; Walsh et al., 2008). None of these 9 genes were significantly differentially spliced in my study, but the catenins *Ctnnb1* (p = 0.006) and *Ctnnd2* ($p = 8.57x10^{-5}$) (also part of the cell adhesion pathway) were. The finding that another catenin (*CTNND1*), also in the cell adhesion pathway, was differentially expressed following *ZNF804A* knockdown in a neural cell line (Hill et al., 2012) derived from human foetal brain may suggest catenin function is directly or indirectly regulated by ZNF804A.

The 'Cell adhesion extracellular matrix (ECM) remodelling pathway' was significantly enriched for genes with altered splicing. This pathway is involved in embryonic development and includes the genes *Lama4* and *Syndecan-2* which were differentially spliced. *Lama4* is also part of the 'cell adhesion' and 'biological adhesion' pathways which were significantly enriched for genes differentially expressed following the knockdown of ZNF804A (Hill et al., 2012). *Lama4* itself was found to be differentially expressed in response to ZNF804A knockdown (Hill et al., 2012). The significance of cell adhesion pathways in studies of schizophrenia risk genes (O'Dulshaine et al.,

2011), ZNF804A knockdown (Hill et al., 2012), differentiating neurons (Lin et al., 2011) and autism (Wang et al., 2009) suggests that aberrant cell adhesion processes in the brain during development could underlie neurodevelopmental disorders (O'Dulshaine et al., 2011).

Pathways enriched for genes with the largest number of probesets were similar to those obtained with differentially spliced genes. This highlights the potential for artefacts influencing pathway analyses. This is a particularly critical issue in differential splice data as the way in which differential splicing is calculated may be influenced by the number of probesets in the gene, which in the majority of instances is related to gene size. Since brain expressed genes are generally larger, and such genes are more likely to be significantly differentially spliced, pathways relevant to brain function are more likely to be significantly enriched by chance. However, importantly, the main findings reported above survived adjustment for probeset number.

As mentioned in the methods (7.2.1) both Metacore and DAVID are described as singular enrichment analysis as they follow a linear procedure. The output from both is a long list of pathways ordered by enrichment p value. The investigator chooses which pathways to investigate further and therefore sometimes the most relevant biological pathways for the data can be overlooked (Huang et al., 2008).

The mouse orthologues of genome wide significant genes from the PGC SZ GWAS were specifically investigated for differential expression and splicing; only Ccdc68 had nominally significant differential expression. Little is known about the function of this gene so it is difficult to postulate how aberrant expression of this gene may mediate the effects of ZNF804A on disease risk. Overall, there was no substantial evidence that the human orthologues of genes with altered expression and splicing in the present experiments were mediating the effects of ZNF804A on disease risk, although weak evidence for a link between differential splicing and bipolar disorder risk was observed. As GWAS datasets enlarge, it will be possible to test this more powerfully.

In conclusion genes showing significant differences between Zfp804a mutant and wildtype mouse brain in their expression and splicing were enriched in pathways associated with translation, axon guidance and cell adhesion. The latter two are known to be important processes during development and fit with a neurodevelopmental hypothesis of schizophrenia. However, the human orthologues of genes differentially

expressed and spliced were not clearly enriched for associations to either schizophrenia or bipolar disorder, thereby providing no evidence that would suggest the human orthologues of these genes are responsible for mediating the effects of ZNF804A on disease risk.

Chapter 8. General Discussion.

8.1 Research Findings.

ZNF804A is a strongly supported schizophrenia susceptibility gene, and at the time this thesis started, there were few other genes implicated at convincing levels of evidence from which insights into schizophrenia pathogenesis could be derived. The function of ZNF804A protein is currently unknown, but a hypothesis suggests it regulates gene expression and splicing. I have investigated this hypothesis by determining the consequences of disrupted ZNF804A in the brains of mice who carry a nonsense mutation in the mouse orthologue Zfp804a (a summary of the results is displayed in Table 8.1).

Levels of Zfp804a mRNA transcript were essentially unchanged between mutants carrying the nonsense and wildtypes indicating the nonsense mutation (C59X) did not activate the nonsense mediated decay (NMD) surveillance mechanism. In the absence of a suitable antibody to empirically determine if Zfp804a protein was expressed, it was postulated that the Zfp804a protein was disrupted.

Affymetrix exon array analysis of RNA extracted from whole brain of embryonic and of adult mice revealed no genes that showed significant expression differences between mice with and without the nonsense mutation. My data do not then support the hypothesis that Zfp804a is involved in transcription regulation, a finding that contrasts with that of others (Hill et al., 2012; Gigenti et al., 2012). This may reflect the absence of a significant expression difference between C59X mutants and wildtypes in Zfp804a, compared to the knockdown and overexpression of ZNF804A observed in the published studies. The discrepancy could also reflect the conflicting results sometimes observed between *in vitro* and *in vivo* gene expression studies (Tatenhorst et al., 2005; Lund et al., 2006; Tsai et al., 2007; Lisle et al., 2008). Expression differences could arise due to the complexities of regulation of certain pathways *in vivo* (Lisle et al., 2008). There are however examples where there is agreement between *in vivo* and *in vitro* studies (Suryo Rahmanto et al., 2007; Ma et al., 2007), but caution should be taken in extrapolating *in vitro* results to complex brain pathways (Tatenhorst et al., 2005).

Results Chapter	Chapter Outline	Differential Expression Candidates	Differential Splicing Candidates
Chapter 3: Expression analysis in Adult ENU mutant mice	Brain mRNA from adult Zfp804a mutant and wildtype mice were analysed using an Exon array (Affymetrix) to determine genes differentially expressed and spliced between Zfp804a mutants and wildtype controls.	*Arc* *Dusp1* *Egr2* *Npas4* *Nr4a1* *Snca*	*2010106G01Rik* *Centb5/Acap3* *Colec12* *Fam171b/D4300389N05Rik* *Dffa* *Frzb* *Itga6* *Itgav* *Lrp4* *Rapgef4* *Rhobtb2/Prdm4* *Slc39a13* *Ssfa2*
Chapter 4: RNA Sequencing	Whole transcriptome RNA Sequencing of 4 male mice analysed in the previous array experiment (2 Zfp804a mutants and 2 wildtypes).	*Arc* *Dusp1* *Npas4* *Nr4a1*	*Itga6* *Dffa*
Chapter 5: Expression Analysis in Embryonic Mice	The analyses carried out in Chapter 3 were repeated using embryonic tissue comparing expression and splicing differences between Zfp804a mutants and wildtypes.	*Npas4* *Ogn*	*Rhobtb2/Prdm4*
Chapter 6: Technical Artefacts	The results of the exon array analyses on embryonic and adult data following the exclusion of chromosome 2 probesets (due to an unusual excess of results on this chromosome) and the application of more stringent intensity filters.	*Npas4* *Ogn*	*Prdm4*

Chapter 7: Relevance of altered Zfp804a function for Schizophrenia	Pathway analysis in Metacore and DAVID. Lists of genes significantly differentially expressed or spliced between Zfp804a mutants and wildtypes were tested for enrichment in biological pathways	**Genes enriched in:** Translation Translation initiation	**Genes enriched in:** Axon Guidance Development Neurogenesis Axonal Guidance (*Robo3*) Cell Adhesion (*Ctnnb1; Ctnnd2*) Cell Adhesion Extra Cellular Matrix Remodelling (*Lama 4; Syndecan-2*)

Table 8.1. Summary of Results. Following each experimental chapter a number of candidate genes were identified as having robust differential expression or splicing in mice with a nonsense mutation in Zfp804a relative to wildtype controls.

Differential splicing was evident between C59X mutants and wildtypes in both embryonic and adult mice with a large number of genes (more than chance) remaining significant following multiple test correction. However a caveat to this is that the false positive rate in a differential splicing study might be greater than that for differential expression analysis because of the additional complexities of the former (Bemmo et al. 2008 and Chapter 3 of this thesis).

I employed quantitative RNA sequencing (RNAseq) to provide a guide as to the validity of the array results. It was also viewed as a tool with which to discover novel transcript variants, which is not possible using the exon array as analysis is restricted to known transcripts. Although statistical evaluations were performed on the sequencing data, the studies were based on two animals in each experimental group so I regarded the findings as a rough guide to the validity of the array results rather than as a robust confirmatory test.

Through RNAseq I identified (and subsequently confirmed using RT-PCR) a novel Zfp804a transcript containing an alternative exon 5' to the Refseq exon 1. This was present in both mutants and wildtypes. The alternative exon 1 skips the constitutive exon 1 and is spliced to exon 2 of *Zfp804a*. Characterising this transcript was not possible within the timeframe of this PhD but doing so will be important to fully understand *Zfp804a* function.

From the same RNAseq data, I found a deletion that was present in the C57BL/6JHsdOla strain that had been used in the breeding programme. This mutation has been reported before (Specht & Schoepfer, 2001), but this was unknown to the team who had been breeding the mutant mice. This deletion, which spans *Snca*, confounded the analysis of the adult mice since the majority of wildtype mice were deletion carriers whereas only one of the mutant mice were. Effects in gene expression variation due to the *Snca* deletion have not been reported before (Specht & Schoeffer, 2001), but nevertheless a caveat of the adult expression results was that the *Snca* deletion correlating with C59X genotype may have confounded the results.

Although of low quality, the RNAseq data also allowed me to identify a number of strain specific cDNA sequence variants at sites corresponding to probesets, many of which were providing data indicative of splice differences between C59X mutants and

wildtype. These sequence variants were largely confined to genes on chromosome 2, the same chromosome to which *Zfp804a* maps, and most likely are indicative of genetic linkage to the C59X mutation. Since these variants appeared to be largely responsible for an excess of significant splicing results on chromosome 2, for subsequent analysis, I excluded genes mapping to that chromosome. This was a conservative approach, but only removed a modest proportion of the genome and as such was unlikely to impact upon the results observed in pathway analyses.

Embryonic tissue was used to assess splicing and expression differences between C59X mutants and wildtypes during development. Impaired development is predicted to influence predisposition to schizophrenia (Weinberger, 1986). The genes with differential expression and splicing between C59X mutants and wildtype did not significantly overlap between embryonic and adult mice datasets, but the *Snca* deletion in the adult mouse data unfortunately prevented me from determining if changes in expression and splicing as a result of the Zfp804a mutation differed developmentally (between embryonic and adult datasets).

Prior to carrying out pathway analysis I established that probe number correlated with the likelihood of a significant alternative splicing result. In response to this finding I corrected for probe number. One of the prominent functional pathways enriched in genes which show differential splicing between C59X mutants and wildtypes was cell adhesion. Cell adhesion pathways have been implicated previously in schizophrenia, bipolar disorder (O'Dulshaine et al., 2011), autism (Wang et al., 2009) and in genes differentially expressed following knockdown of ZNF804A (Hill et al., 2012). Significant pathways also included axon guidance and extracellular matrix remodelling both important during embryonic development. Processes which are prominent during development therefore may be aberrant in C59X mutants following the disruption of Zfp804a. Genes differentially expressed and spliced between mutants and wildtypes were not enriched for association with disease in a large case-control genetic dataset.

8.2 Limitations of the data

The biggest limitation of the study, with regards to accurately determining genes with altered expression and splicing as a result of disrupted Zfp804a, was the presence of strain specific sequence variants found in the C59X mutants. This included a

nonsynonymous SNP found in exon 4 of *Zfp804a*, in addition to the artefacts relating to hybridisation. Nonsynonymous genetically linked mutations may impact upon the relevant protein's function. This in turn could affect expression and splicing results in the C59X mutants. Adjusting for the potential effects of these variants was not possible in this study and so leaves the caveat that expression and splicing differences observed between C59X mutants and wildtypes may not be attributable specifically to *Zfp804a* disruption.

Sample sizes meant the power of the studies were limited. 5 biological replicates per experimental group are recommended (Affymetrix) for determining differential splicing on the exon array, but the limited availability of C59X homozygotes meant this couldn't be met in the experiments on adult mice. The RNAseq study consisted of just 2 samples per experimental group and therefore the power to determine statistical differences in expression and splicing was very low.

Discrepancies occurring when assessing the overlap of results produced by different software may arise due to the annotation methods used. Previous studies have found this and note the importance of manually checking the curation (Bemmo et al., 2008). This was noted in this study as one of the results from Partek GS is annotated as *Prdm4*, however when the sequence predicted to be differentially spliced was entered into BLAT (UCSC) the sequence was found within the *Rhobtb2* gene found on a different chromosome. The two genes have different Refseq IDs but have the same Affymetrix transcript cluster ID annotation.

The correlation structure of genes is a caveat of pathway analysis but there is no standardised way known to deal with it aside from determining the LD between every gene. To address this future studies could use gene set enrichment analysis (GSEA). GSEA uses permutations and therefore may control for the correlation observed in the expression of co-regulated genes.

Finally, the absence of an antibody is an important limitation since I have no direct data regarding the impact of the mutation on protein abundance.

8.3 Future Work.

To complement the work carried out in this thesis results could be considered at the individual gene level. Genes with the most significant differential expression and splicing between C59X mutants and wildtypes would be confirmed using a real time

PCR technique in independent tissue derived from C59X mutant and wildtype mice that have undergone additional backcrossing to the C57/BL6JHsdOla strain. Further to this the confirmed splice and expression differences observed in C59X mutants relative to wildtypes could be linked to the psychosis associated risk variant (rs1344706) using mRNA from human post-mortem samples.

An experiment could be carried out in which the C59X mutation is rescued. This could be done by promoting read through of the premature termination codon (Kayali et al., 2012). The absence of expression and splicing differences between rescued C59X mutants and wildtypes would imply the differences observed previously were a consequence of the C59X mutation specifically.

The chromatin-immunoprecipitation sequencing technique (ChIPseq) could be utilised which would allow the distinction to be made between downstream targets which are either directly or indirectly regulated by Zfp804a. Genes found to interact directly with Zfp804a are more likely to be aetiologically relevant and would be more beneficial in elucidating the mechanisms by which ZNF804A may be linked to schizophrenia and psychosis. This approach would ideally rely on the genesis of a sensitive and specific antibody.

A transgenic approach could be taken in which the Zfp804a gene is knocked out in the mouse. This would avoid the complications that have arisen in this thesis as a result of using ENU random mutagenesis to create a mutation in one strain and then backcrossing to another strain for congenicity. By removing the problem of strain specific sequence variation any observed expression and splice differences between mice with Zfp804a knocked out and wildtype could be more confidently attributed to Zfp804a.

8.4 General Conclusion.

The data generated for this thesis demonstrate that a PTC within exon 2 of Zfp804a predominately effects the splicing of genes and suggests a role for ZNF804A in the regulation of RNA processing. When addressing the hypothesis that ZNF804A may be a transcription factor this cannot be confirmed from the data generated here, but at the same time does not rule out the possibility that effects on gene expression only occur in a very small number of genes. Downstream targets of Zfp804a were enriched in

pathways involved in axon guidance during development and cell adhesion. This provides additional support for these pathways in the underlying pathophysiology of schizophrenia and the relevance of developmental pathways to the aetiology of the disorder.

Appendix.

Chapter 2 Appendices

Appendix 2.1 Zfp804a Exon1-3 Primers

The primers were used to amplify between exon 1 and exon 3 of Zfp804a under the conditions described in Chapter 2.4.2.

Left primer	ctctcagcaagaacgggaac
Right primer	cgagcaaattctctctgtttca
Product Size:	208

Appendix 2.2 Zfp804a Exon 2 Primers

Primers used to amplify within exon 2 of Zfp804a spanning the C59X mutation for use in the high resolution melt analysis (HRMA)

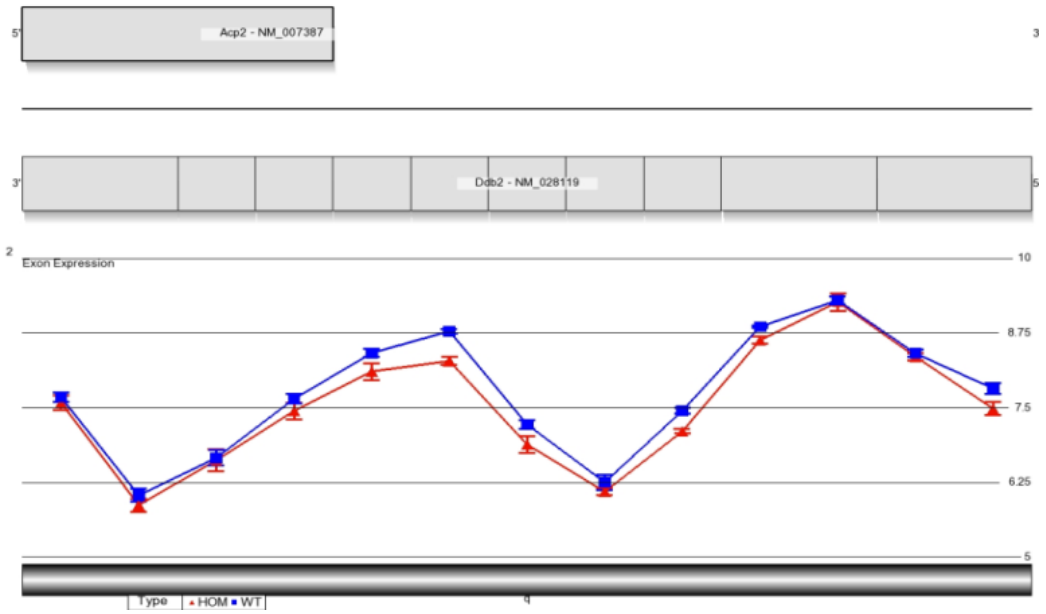
Left Primer	ccaaagctctggaggatctg
Right Primer	tgggcgtggtcatatgagtt
Product Size	109

Chapter 3 Appendices

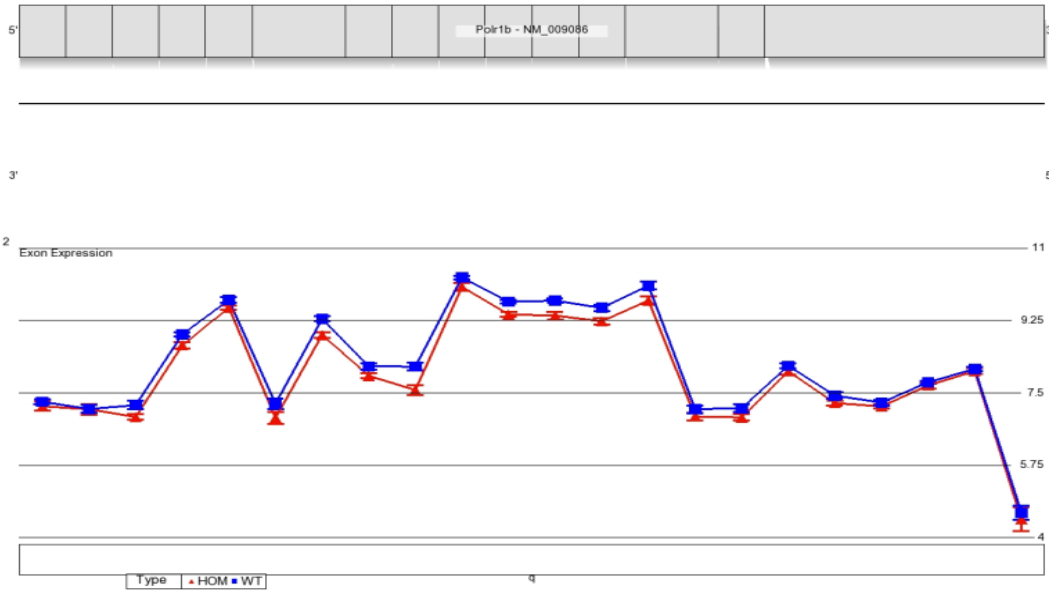
Appendix 3.1 Geneviews of Differentially Expressed Genes between Adult C59X Mutants and Wildtypes.

The Geneviews of 9 genes significantly differentially expressed between C59X mutant and wildtype following a FDR correction at a threshold of 0.05 in the combined adult sample. Of the 9 only 2 looked like differential gene expression (Mettl5 and Nfe212) with two others possibly being differential expression of known alternative transcripts. The remaining 5 showed differential expression in only a subset of the probesets within the transcripts, which could suggest differential expression of novel isoforms. Of the 3 Bonferroni significant transcripts only Mettl5 displays attributes of a differentially expressed gene, with all probesets showing differential expression in the geneview.

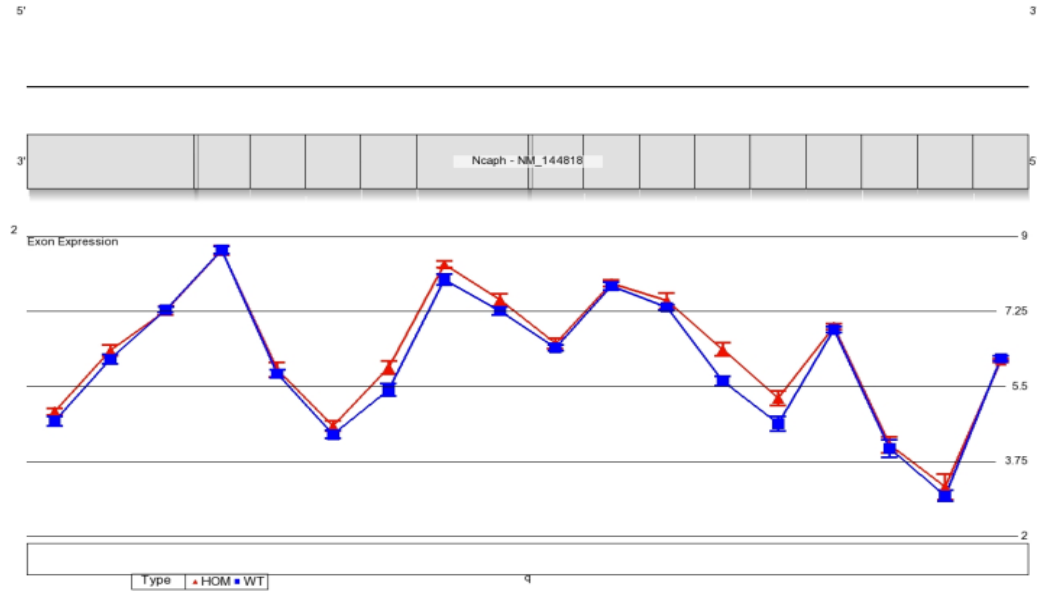
Ddb2


Acp2 - NM_007387
Ddb2 - NM_028119
Exon Expression
10
8.75
7.5
6.25
5
Type
HOM
WT

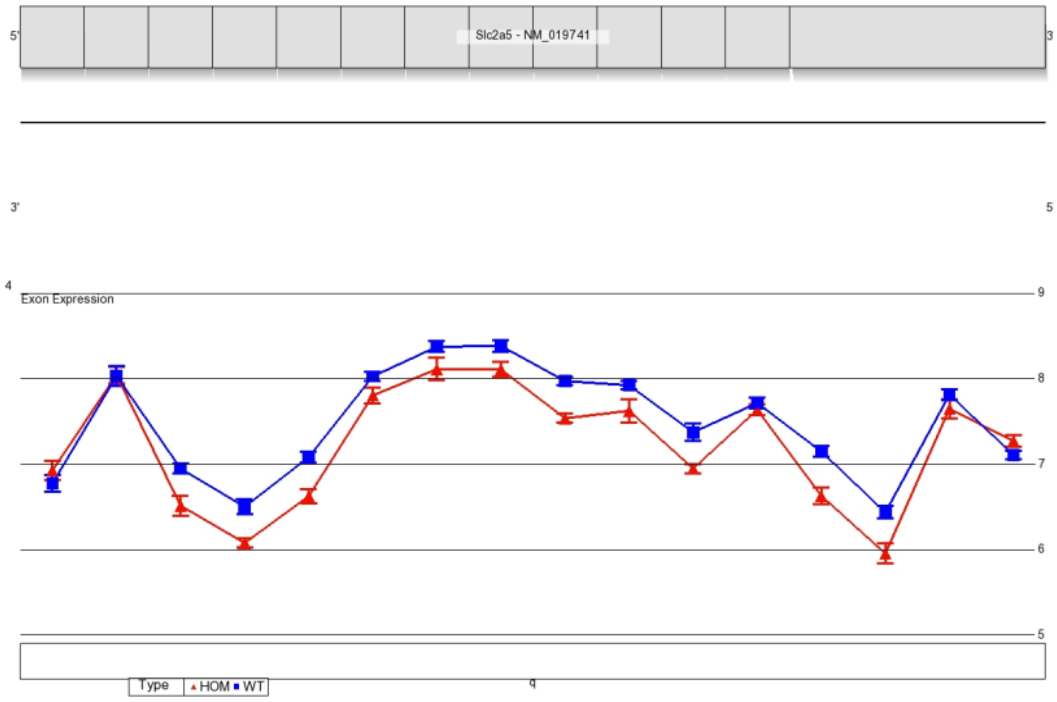
Polr1b


Polr1b - NM_009086
Exon Expression
11
9.25
7.5
5.75
4
Type
HOM
WT

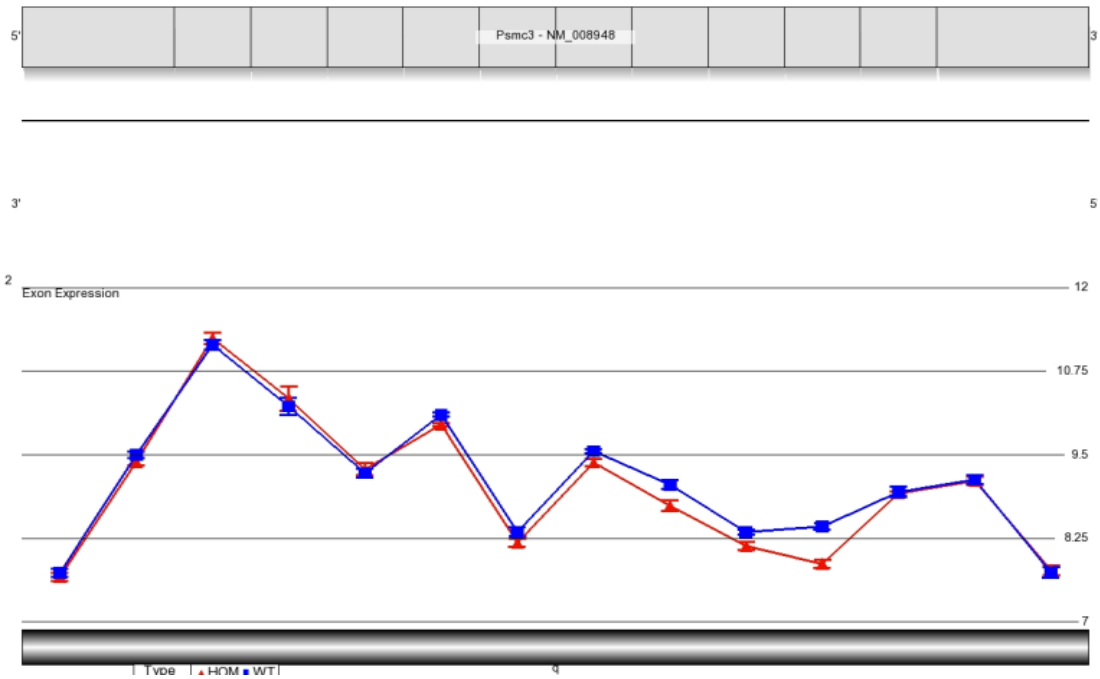
Ncaph



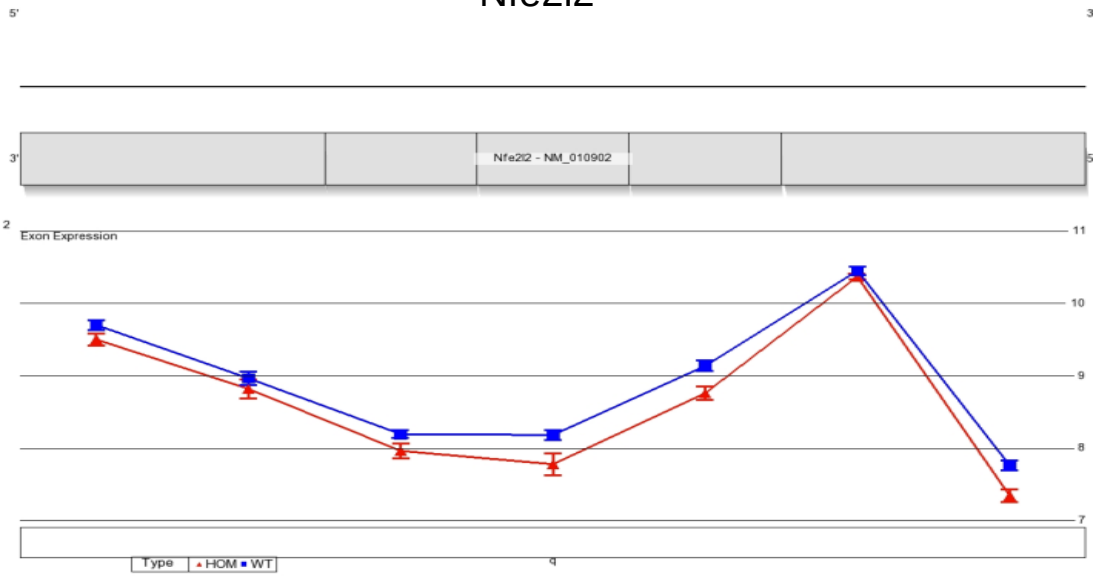
Slc2a5



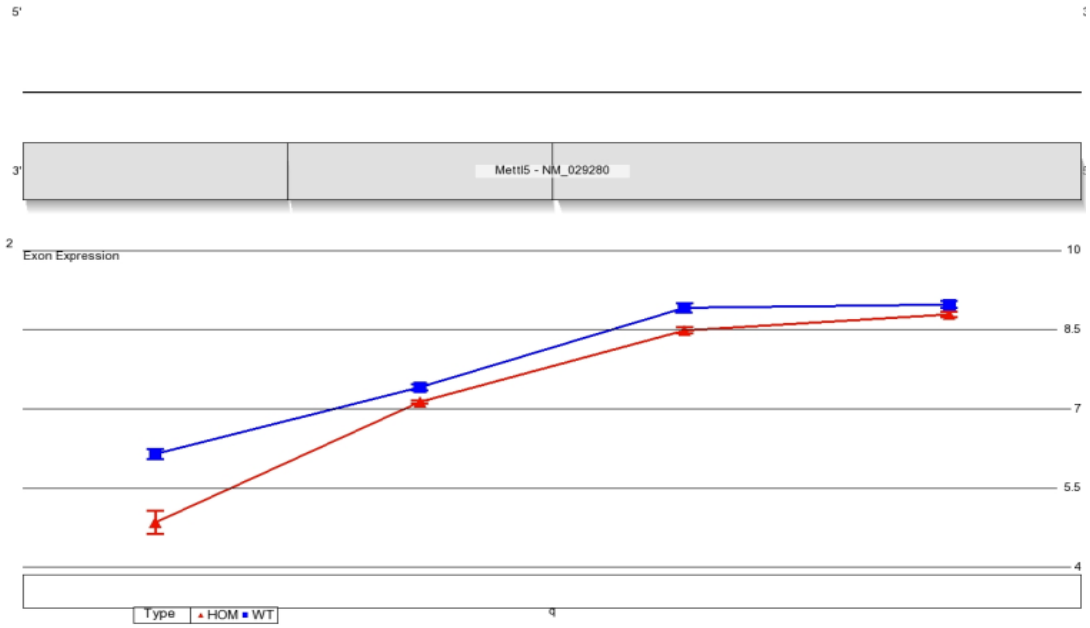
Psmc3



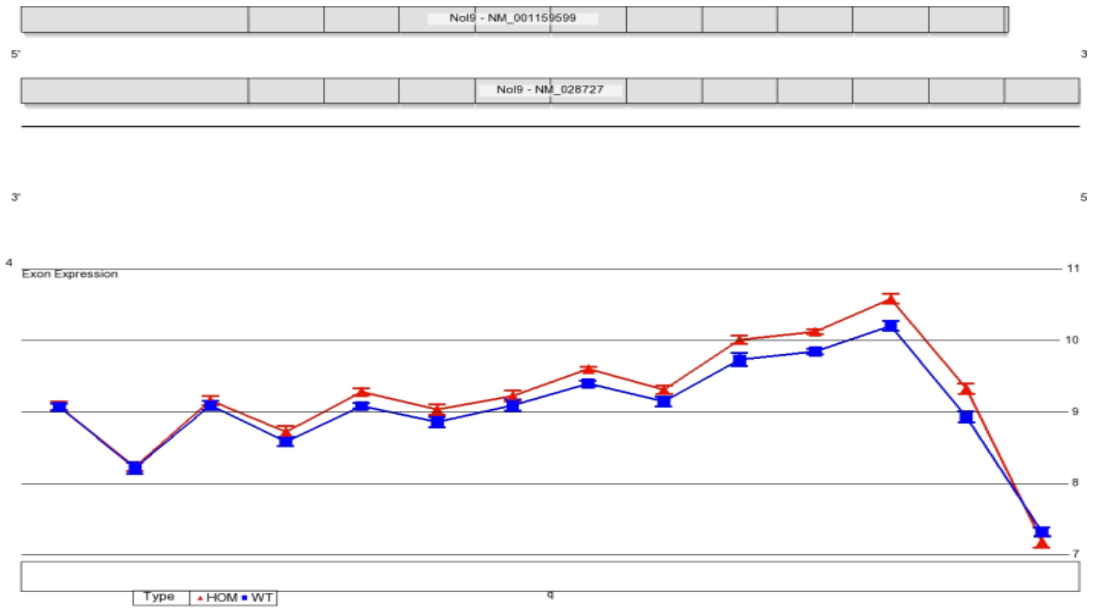
Nfe2l2



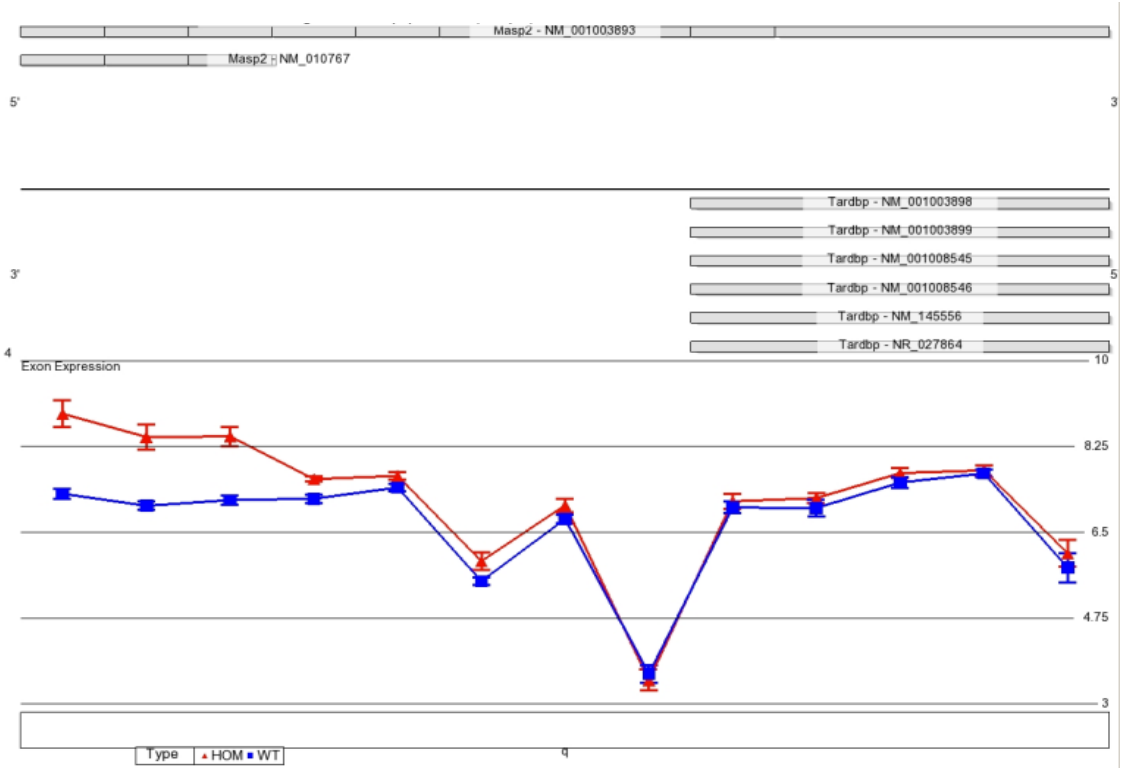
Mettl5


5'
3'
Mettl5 - NM_029280
Exon Expression
10
8.5
7
5.5
4
Type
HOM
WT

Nol9


Nol9 - NM_001159599
Nol9 - NM_028727
5'
3'
Exon Expression
11
10
9
8
7
Type
HOM
WT

Masp2


Masp2 - NM_001003893
Masp2 - NM_010767
5'
3'
Tardbp - NM_001003898
Tardbp - NM_001003899
Tardbp - NM_001008545
Tardbp - NM_001008546
Tardbp - NM_145556
Tardbp - NR_027864
Exon Expression
10
8.25
6.5
4.75
3
Type
HOM
WT

Chapter 4 Appendices

Appendix 4.1 TopHat

Scripts used to align sequence Files in TopHat (v.1.3.2) (Trapnell et al., 2009)

$ tophat –-output-dir s_1_thout -–solexa1.3-quals -–num-threads 10 –-library-type fr-unstranded –r 40 genome s_1_1_sequence.txt,s_1_3_sequence.txt

Where s_1_1 and s_1_3 represent the paired end reads for sample 1 a C59X mutant. The other C59X mutant was annotated as s_2_1; s_2_3 and the two wildtypes were s_3_1; s_3_3 and s_5_1; s_5_3.

The use of each option in the script is explained below and derived from the TopHat manual (http://tophat.cbcb.umd.edu/manual.html).

--output-dir This option defined the output directory in which the results files were placed e.g., s_1_thout

--solexa1.3-quals This option was used because the sequences were in fastq format produced using the Illumina GA pipeline version 1.5 and this option is recommended for versions 1.3 or later and states that the quality scores are in encoded using the phred scale (base 64).

--num-threads 10 10 threads were used to align reads.

--library-type fr-unstranded This option is the default and was used as it is suitable for sequencing produced using the Illumina Truseq protocol. This means TopHat treated the reads as strand specific. Reads at the left end of the transcript were mapped to the transcript strand and reads at the right end were mapped to the opposite strand.

-r 40 The –r option represents the mean inner distance between mate pairs. This parameter must be set for paired end runs. As adapters are ligated to the ends of sequences this option ensures sequencing begins at the cDNA not the adapter. The length of the adapter sequences was subtracted from the length of the library fragment size to determine the average distance between the ends of the paired end reads.

Appendix 4.2 SamTools

Script used to Index Binary Alignment Files (BAM) in order to view them in the Integrative Genomics Viewer. This was carried out using Samtools (http://samtools.sourceforge.net/)

$ samtools index accepted_hits.bam

Appendix 4.3 Cufflinks

Scripts used for RNAseq Transcript Assembly in Cufflinks (v.1.2.1) (Trapnell et al., 2010) (http://cufflinks.cbcb.umd.edu/)

$ cufflinks –p 8 –o s_1_clout –b genome.fa –u –N –g genes.gtf s_1_thout/accepted_hits.bam

-p reflects the number of threads used and can be either be 4 or 8. 8 was chosen as the faster option.

-b The NCBI 37.1build genome fasta file was provided to Cufflinks to allow the bias detection and correction algorithm to be run. This option was included to improve the accuracy of transcript abundance estimates.

-u This option was included so that cufflinks can accurately weight reads that map to multiple sites in the genome by running an estimation procedure.

-N represents an upper quartile normalisation. Rather than normalising to the total number of fragments mapping to an individual loci (default), cufflinks takes the upper quartile.

-g The NCBI build 37.1 reference annotation was supplied to Cufflinks to give additional information to increase the accuracy of transcript assembly. Novel genes/isoforms are still included in the output as well as reference transcripts.

Appendix 4.4 CuffMerge

Assemblies were merged by first using gedit (a text editor) to create a file called 'assemblies.txt' within which the assembly files for each sample were listed:

./s_1_clout/transcripts.gtf

./s_2_clout/transcripts.gtf

./s_3_clout/transcripts.gtf

./s_5_clout/transcripts.gtf

And then running CuffMerge (http://cufflinks.cbcb.umd.edu/)

$ cuffmerge -g genes.gtf -s genome.fa –p 8 assemblies.txt

The –s option was included to provide Cuffmerge with gDNA reference sequences (*Mus musculus* NCBI build 37.1) to help remove artefacts and define transcript fragments.

Appendix 4.5 CuffDiff

The following script in Cuffdiff (http://cufflinks.cbcb.umd.edu/) was used to compare the relative amounts of assembled transcripts represented in each sample and determine if the difference was statistically significant

$ cuffdiff –o diff_out -b genome.fa –p 8 –L mut,wt –u merged_asm/merged.gtf
s_1_thout/accepted_hits.bam,./s_2_thout/accepted_hits.bam
s_3_thout/accepted_hits.bam,./s_5_thout/accepted_hits.bam

Appendix 4.6 Primers Targeting an Alternative Zfp804a Isoform.

The primers were used to amplify between alternative exon 1 and exon 2 of Zfp804a under the conditions described in Chapter 2.4.2.

Left primer	CCACCACCTCAAAGGAGCTA
Right primer	GTTTGTGGGCGTGGTCATA
Product Size:	173bp

Chapter 5 Appendices

Appendix 5.1 Embryonic Sample Background. All brain tissue used on the embryonic array study was derived from E18.5 mice from a heterozygous intercross from parent mice with 98.5% C57BL/6J genome from either the 7th or 8th generation.

Parent ID		**Genotype**	**Generation**	**C57BL/6J Genome (%)**
Female	**E27AD2**	Het	F7	98.5
Male	E27AJ2	Het	F7	98.5
Parent ID		**Genotype**	**Generation**	**C57BL/6J Genome (%)**
Female	**E27AO4**	Het	F7	98.5
Male	E27P3	Het	F7	98.5
Parent ID		**Genotype**	**Generation**	**C57BL/6J Genome (%)**
Female	**B8E1**	Het	F8	98.5
Male	E27T2	Het	F7	98.5
Parent ID		**Genotype**	**Generation**	**C57BL/6J Genome (%)**
Female	**E27AM2**	Het	F7	98.5
Male	E27X0	Het	F7	98.5
Parent ID		**Genotype**	**Generation**	**C57BL/6J Genome (%)**
Female	**B8A2**	Het	F8	98.5
Male	E27VO	Het	F7	98.5

Appendix 5.2 Gender PCR Primers

Ssty (forward): CTGGAGCTCTACAGTGATGA

Ssty (reverse): CAGTTACCAATCAACACATCAC Product: 343bp

Om1a (forward): TTACGTCCATCGTGGACAGCAT

Om1a (reverse): TGGGCTGGGTGTTAGTCTTAT Product: 245bp

References

Abdueva D, et al. (2007) Experimental Comparison and Evaluation of the Affymetrix Exon and U133Plus2 GeneChip Arrays. PLoS ONE 2:e913.

Affymetrix (2005a) Exon probeset annotations and transcript cluster groupings. In. http://media.affymetrix.com/support/technical/whitepapers/exon_probeset_trans_clust_whitepaper.pdf

Affymetrix (2005b) Alternative Transcript Analysis Methods for Exon Arrays. Affymetrix GeneChip® Exon Array Whitepaper Collection: 1-13. http://media.affymetrix.com/support/technical/whitepapers/exon_alt_transcript_analysis_whitepaper.pdf

Affymetrix (2006) Identifying and validating alternative splicing events: An Introduction to managing data provided by GeneChip® Exon Arrays. Affymetrix Genechip© Exon Array Technical Note: 1-16. In. http://media.affymetrix.com/support/technical/technotes/id_altsplicingevents_technote.pdf

Affymetrix (2007) Quality Assessment of Exon and Gene Arrays. Affymetrix GeneChip® Gene and Exon Array Whitepaper Collection: 1-18. http://media.affymetrix.com/support/technical/whitepapers/exon_gene_arrays_qa_whitepaper.pdf

Affymetrix (2008) QC metrics for exon and gene design expression arrays. Quick Reference Card: 1-4. http://media.affymetrix.com/support/downloads/quick_reference_cards/qc_metrics_exon_gene_qrc.pdf

Alberts R, et al. (2007) Sequence polymorphisms cause many false cis eQTLs. PLoS One 2:e622.

Allardyce J, et al. (2001) Comparison of the incidence of schizophrenia in rural Dumfries and Galloway and urban Camberwell. Br J Psychiatry 179:335-339.

Anton ES, et al. (1999) Distinct functions of alpha3 and alpha(v) integrin receptors in neuronal migration and laminar organization. Neuron 22:277-289.

Arseneault L, et al. (2002) Cannabis use in adolescence and risk for adult psychosis: longitudinal prospective study. BMJ 325:1212-1213.

Bacchelli E, et al. (2003) Screening of nine candidate genes for autism on chromosome 2q reveals rare nonsynonymous variants in the cAMP-GEFII gene. Mol Psychiatry 8:916-924.

Bainbridge MN, et al. (2010) Whole exome capture in solution with 3 Gbp of data. Genome Biol 11:R62.

Balog Z, et al. (2011) ZNF804A may be associated with executive control of attention. Genes Brain Behav 10:223-227.

Bani-Yaghoub M, et al. (2007) A switch in numb isoforms is a critical step in cortical development. Dev Dyn 236:696-705.

Bark C, et al. (2004) Developmentally regulated switch in alternatively spliced SNAP-

25 isoforms alters facilitation of synaptic transmission. J Neurosci 24:8796-8805.

Bashyam MD (2009) Studies on nonsense mediated decay reveal novel therapeutic options for genetic diseases. Recent Pat DNA Gene Seq 3:7-15.

Becker J, et al. (2012) Evidence for the involvement of ZNF804A in cognitive processes of relevance to reading and spelling. Transl Psychiatry.

Bemmo A, et al. (2008) Gene expression and isoform variation analysis using Affymetrix Exon Arrays. BMC Genomics 9.

Benes FM, et al. (1994) Myelination of a key relay zone in the hippocampal formation occurs in the human brain during childhood, adolescence, and adulthood. Arch Gen Psychiatry 51:477-484.

Benjamini Y & Hochberg Y (1995) Controlling the false discovery rate: A practical and powerful approach to multiple testing. J R Statist Soc B 57:289-300.

Benovoy D, et al. (2008) Effect of polymorphisms within probe-target sequences on olignonucleotide microarray experiments. Nucleic Acids Res 36:4417-4423.

Blotta S, et al. (2009) Identification of novel antigens with induced immune response in monoclonal gammopathy of undetermined significance. Blood 114:3276-3284.

Bolstad B (2008) Preprocessing and Normalization for Affymetrix GeneChip Expression Microarrays. In: Methods in Microarray Normalization (Stafford P, ed), pp 41-59. USA: CRC Press.

Bolstad BM (2002) Comparing the effects of background, normalization and summarization on gene expression estimates. Unpublished Manuscript. http://bmbolstad.com/stuff/components.pdf

Bolstad BM, et al. (2003) A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. Bioinformatics 19:185-193.

Bonneh-Barkay D, & Wiley CA. (2009) Brain extracellular matrix in neurodegeneration. Brain Pathology 19:573-585.

Bora E, et al. (2009) Theory of mind impairment in schizophrenia: meta-analysis. Schizophrenia Research 109.

Borrell J, et al. (2002) Prenatal immune challenge disrupts sensorimotor gating in adult rats. Implications for the etiopathogenesis of schizophrenia. Neuropsychopharmacology 26:204-215.

Bosley TM, et al. (2007a) Clinical characterization of the HOXA1 syndrome BSAS variant. Neurology 69:1245–1253.

Bosley TM, et al. (2007b) Clinical characterization of the HOXA1 syndrome BSAS variant Neurology 69:1245–1253.

Boutz PL., et al. (2007) A post-transcriptional regulatory switch in polypyrimidine tract-binding proteins reprograms alternative splicing in developing neurons. Genes Dev 21:1636–1652.

Bowler AE & Torrey EF (1990) Influenza and schizophrenia. Archives of General Psychiatry, 47:876-877.

Brennand KJ, et al. (2011) Modelling schizophrenia using human induced pluripotent stem cells. Nature 473:221-225.

Brennand KJ, et al. (2011) Modelling schizophrenia using human induced pluripotent stem cells. Nature 473:221-225.

Brown AS, et al. (2005) Maternal exposure to toxoplasmosis and risk of schizophrenia in adult offspring. American Journal of Psychiatry 162:767-773.

Brown AS, et al. (2000) Maternal exposure to respiratory infections and adult schizophrenia spectrum disorders: A prospective birth cohort study. Schizophrenia Bulletin 26:287-295.

Brown AS, & Susser, ES. (2008) In utero infection and adult schizophrenia. Ment Retard Dev Disabil Res Rev 8:51-57.

Brown AS, et al. (2004) Elevated maternal interleukin-8 levels and risk of schizophrenia in adult offspring. Am J Psychiatry 161:889-895.

Brown MB (1975) A method for combining non-independent, one-sided tests of significance. Biometrics 31:987-992.

Buonocore F, et al. (2010) Effects of cis-regulatory variation differ across regions of the adult human brain. Hum Mol Genet 19:4490-4496.

Camurri L, et al. (2005a) Evidence for the existence of two Robo3 isoforms with divergent biochemical properties. Molecualr and Cellular Neuroscience 30:485-493.

Camurri L, et al. (2005b) Evidence for the existence of two Robo3 isoforms with divergent biochemical properties. Mol Cell Neurosci 30:485-493.

Cardno AG, & Gottesmann, II. (2000) Twin studies of schizophrenia: from bow-and-arrow concordances to star wars Mx and functional genomics. Am J Med Genet 97:12-17.

Chang TY, et al. (2008) easyExon--a Java-based GUI tool for processing and visualization of Affymetrix exon array data. BMC Bioinformatics 9.

Charlet BN, et al. (2002) Dynamic antagonism between ETR-3 and PTB regulates cell type-specific alternative splicing. Mol Cell 9:649–658.

Chen H, et al. (2010) Differential expression and alternative splicing of genes in lumbar spinal cord of an amyotrophic lateral sclerosis mouse model. Brain Res 1340:52-69.

Chen M, et al. (2012) Evidence of IQ-modulated association between ZNF804A gene polymorphism and cognitive function in schizophrenia patients. Neuropsychopharmacology 37:1572-1578.

Chen M, et al. (2012) Evidence of IQ-modulated association between ZNF804A gene polymorphism and cognitive function in schizophrenia patients. Neuropsychopharmacology 37:1572-1578.

Chen S-Y, et al. (2011) Disrupted-in-Schizophrenia 1–mediated axon guidance involves TRIO-RAC-PAK small GTPase pathway signaling. Proc Natl Acad Sci U S A 108: 5861–5866.

Chung HJ, et al. (2010) Mouse Homologue of the Schizophrenia Susceptibility Gene ZNF804A as a Target of Hoxc8. Journal of Biomedicine and Biotechnology 2010.

Cocchella A, et al. (2010) The refinement of the critical region for the 2q31.2q32.3 deletion syndrome. Am J Med Genet B Neuropsychiatr Genet 153B:1342-1346.

Coghill EL, et al. (2002) A gene-driven approach to the identification of ENU mutants in the mouse. Nat Genet 30:255-256.

Collarini EJ, et al. (1992) Down-regulation of the POU transcription factor SCIP is an early event in oligodendrocyte differentiation in vitro. Development 116:193-200.

Collarini EJ, et al. (1992) Down-regulation of the POU transcription factor SCIP is an early event in oligodendrocyte differentiation in vitro. Development 116:193-200.

Cooper, SJ (1992) Schizophrenia after prenatal exposure to 1957 A2 influenza epidemic. Br J Psychiatry 161:394-396.

Cousijn H, et al. (2012) Schizophrenia risk gene ZNF804A does not influence macroscopic brain structure: an MRI study in 892 volunteers. Mol Psychiatry 17:1155-1157.

Coyle J (2006) Glutamate and schizophrenia: beyond the dopamine hypothesis. Cell Mol Neurobiol 26:365-384.

Craddock N, & Owen, MJ. (2010) The Kraepelinian dichotomy - going, going... but still not gone. Br J Psychiatry 196:92-95.

Crow TJ (1994) Prenatal exposure to influenza as a cause of schizophrenia. There are inconsistancies and contradictions in the evidence. The British Journal of Psychiatry 164:588-592.

Crow TJ, et al. (1991) Schizophrenia and Influenza. The Lancet 338:116-117.

Cummings E, et al. (2010) Clinical symptomatology and the psychosis risk gene ZNF804A. Schizophr Res 122:273-275.

Dalman C, et al. (1999) Obstetric complications and the risk of schizophrenia: a longitudinal study of a national birth cohort. Arch Gen Psychiatry 56:234-240.

Damian D & Gorfine M (2004) Statistical concerns about the GSEA procedure. Nature Genetics 36:663.

De Arcangelis A, et al. (1999) Synergistic activities of alpha3 and alpha6 integrins are required during apical ectodermal ridge formation and organogenesis in the mouse. Development 126:3957-3968.

Della BC, et al. (2008) Dissecting an alternative splicing analysis workflow for Genechip Exon 1.0 ST Affymetrix arrays. BMC Genomics 9.571

doi:10.1186/1471-2164-9-571

Dennis G Jr, et al. (2003) DAVID: Database for Annotation, Visualization, and Integrated Discovery. Genome Biol 4:P3.

Dikeos DG, et al. (2006) Distribution of symptom dimensions across Kraepelinian divisions. Br J Psychiatry 189.

Doniger SW, et al. (2003) MAPPFinder: using Gene ontology and genMAPP to create a global gene-expression profile from microarray data. Genome Biology 4:Article R7.

Donohoe G, et al. (2010) The psychosis susceptibilty gene ZNF804A: Associations, functions and phenotypes. Schizophrenia Bulletin 36:904-909.

Donohoe G, et al. (2011) ZNF804A risk allele is associated with relatively intact gray matter volume in patients with schizophrenia. Neuroimage 54:2132-2137.

Dostie J, et al. (2000) Nuclear Eukaryotic Initiation Factor 4e (Eif4e) Colocalizes with Splicing Factors in Speckles. Journal of Cell Biology 148:239-246.

Draghici S, et al. (2003) Global functional profiling of gene expression. Genomics 81:98-104.

Duan S, et al. (2008) SNPinProbe_1.0: A database for filtering out probes in the Affymetrix GeneChip® Human Exon 1.0 ST array potentially affected by SNPs. Bioinformation 2:469–470.

Dwyer S, et al. (2010) No evidence that rare coding variants in ZNF804A confer risk of schizophrenia. Am J Med Genet B Neuropsychiatr Genet 153B:1411-1416.

Eastwood SL, et al. (2003) The axonal chemorepellant semaphorin 3A is increased in the cerebellum in schizophrenia and may contribute to its synaptic pathology. Mol Psychiatry 8:148-155.

Eisenhart C (1947) The assumptions underlying the analysis of variance. Biometrics 3:1-21.

Emig D, et al. (2010) AltAnalyze and DomainGraph: analyzing and visualizing exon expression data. Nucleic Acids Res 38.

Emmert-Streib F & Glazko GV (2011) Pathway Analysis of expression data: Deciphering functional building blocks of complex diseases. PLoS Computational Biology 7. e1002053.

Erickson R (1996) Mouse models of human genetic disease: which mouse is more like a man? Bioessays 18:993-998.

Eriksen JL, Janus, CG. (2007) Plaques, tangles, and memory loss in mouse models of neurodegeneration. Behav Genet 37:79-100.

Esslinger C, et al. (2009) Neural mechanisms of a genome-wide supported psychosis variant. Science 324605.

Esslinger C, et al. (2011) Cognitive state and connectivity effects of the genome-wide significant psychosis variant in ZNF804A. Neuroimage 54:2514-2523.

Ewing B, & Green, P. (1998) Base-calling of automated sequencer traces using phred. II. Error probabilities. Genome Res 8:186-194.

Ewing B, et al. (1998) Base-calling of automated sequencer traces using phred. I. Accuracy assessment. Genome Res 8:175-185.

Farabaugh PJ, et al. (1993) A novel programed frameshift expresses the POL3 gene of retrotransposon Ty3 of yeast: frameshifting without tRNA slippage. Cell 74:93-103.

Farina A, et al. (1963) Birth order of recovered and nonrecovered schizophrenics. Archives of General Psychiatry, 9:224-228.

Fatemi SH, et al. (2008) Maternal infection leads to abnomral gene regulation and brain atrophy in mouse offspring: implications for genesis of neurodevelopmetal diosrders. Schizophrenia Research 99:56-70.

Faustino NA, & Cooper TA (2003) Pre-mRNA splicing and human disease. Genes and Development 17:419-437.

Freedman R (2003) Schizophrenia. New England Journal of Medicine 349:1738-1749.

Frischmeyer PA, & Dietz, HC. (1999) Nonsense-mediated mRNA decay in health and disease. Hum Mol Genet 8:1893-1900.

Gajović S, et al. (2006) Unexpected rescue of alpha-synuclein and multimerin1 deletion in C57BL/6JOlaHsd mice by beta-adducin knockout. Transgenic Res 15:255-259.

Gamazon ER, et al. (2010) Comprehensive survey of SNPs in the Affymetrix exon array using the 1000 Genomes dataset. PLoS One 5:e9366.

Gamsjaeger R, et al. (2004) Sticky fingers: zinc-fingers as protein-recognition motifs. Langmuir 20:5885-5890.

Gardina P, & Turpaz Y (2008) Exon Array Analysis for the detection of Alternative Splicing. In: Methods in Microarray Normalization (Stafford P, ed), pp 205-231. USA: CRC Press.

Gilmore JH, & Jarskog, LF (1997) Exposure to infection and brain development: cytokines in the pathogenesis of schizophrenia. Schizophr Res 24:365-367.

Girgenti MJ, et al. (2012) ZNF804A regulates expression of the schizophrenia-associated genes PRSS16, COMT, PDE4B and DRD2. PLoS ONE 7:e32404.

Gottesman I (1991) Schizophrenia Genesis: The Origins of Madness. New York: W.H Freeman.

Gottesman II, & Shields J (1967) A polygenic theory of schizophrenia. Proc Natl Acad Sci 58:199–205.

Graham R, et al. (2005) Distinguishing different DNA heterozygotes by high-resolution melting. Clin Chem 51:1295-1298.

Green EK, et al. (2010) The bipolar disorder risk allele at CACNA1C also confers risk of recurrent major depression and of schizophrenia. Mol Psychiatry 15:1016-1022.

Griswold AJ, et al. (2012) Evaluation of copy number variations reveals novel candidate genes in autism spectrum disorder-associated pathways. Hum Mol Genet 21:3513-3523.

Hampton RY, (2002) ER-associated degradation in protein quality control and cellular regulation. Curr Opin Cell Biol 14:476-482.

Hansen KD, et al. (2010) Biases in Illumina transcriptome sequencing caused by random hexamer priming. Nucleic Acids Res 38:e131.

Harding HP, et al. (2000) Regulated translation initiation controls stress-induced gene expression in mammalian cells. Mol Cell 6:1099-1108.

Hargreaves A, et al. (2012) ZNF804A and social cognition in patients with schizophrenia and healthy controls. Mol Psychiatry 17:118-119.

Harismendy O, et al. (2009) Evaluation of next generation sequencing platforms for population targeted sequencing studies. Genome Biology 10.R32.

Harrison PJ (2000) Postmortem studies in schizophrenia. Dialogues Clin Neurosci 2:349-357.

Hashimoto R et al. (2010) The impact of a genome-wide supported psychosis variant in the ZNF804A gene on memory function in schizophrenia. Am J Med Genet B Neuropsychiatr Genet 153B:1459-1464.

Hennessy AV, et al. (1964) Asian influenza: occurrence and recurrence, a community and family study. Military Medicine, 129:38-50.

Herr AJ, et al. (2000) One protein from two open reading frames: mechanism of a 50 nt translational bypass. EMBO J 19:2671–2680.

Hill MJ, & Bray NJ (2011) Allelic differences in nucelar protein binding at a genome-wide signficant risk variant for schizophrenia in ZNF804A. Molecular Psychiatry 16:787-789.

Hill MJ, et al. (2012) Knockdown of the psychosis susceptibilty gene ZNF804A alters expression of genes involved in cell adhesion. Human Molecular Genetics 21:1018-1024.

Hirschberg A, et al. (2010) Gene Deletion Mutants Reveal a Role for Semaphorin Receptors of the Plexin-B Family in Mechanisms Underlying Corticogenesis. Molecular and Cellular Biology 30:764-780.

Hitotsumachi S, et al. (1985) Dose-repetition increases the mutagenic effectiveness of N-ethyl-N-nitrosourea in mouse spermatogonia. Proc Natl Acad Sci 82:6619–6621.

Hodges C, et al. (2009) Nucleosomal fluctuations govern the transcription dynamics of RNA polymerase II. Science 325:626-628.

Holm S (1979) A simple sequentially rejective multiple test procedure. Scandinavian Journal of Statistics 6:65–70.

Howes OD, et al. (2009) Elevated striatal dopamine function linked to prodromal signs of schizophrenia. Arch Gen Psychiatry 66:13-20.

Huang DW, et al. (2009a) Bioinformatics enrichment tools: paths towards the comprehensive functional analysis of large gene lists. Nucleic Acids Research 37:1-13.

Huang DW, et al. (2009b) Systematic and Integrative analysis if large gene lists using DAVID Bioinformatics resources. Nature Protocols 4:44-57.

Humphreys DT, et al. (2005) MicroRNAs control translation initiation by inhibiting eukaryotic initiation factor 4E/cap and poly(A) tail function. PNAS 102:16961-16966.

Ichikawa-Tomikawa N, et al. (2012) Laminin α1 is essential for mouse cerebellar development. Matrix Biology 31:17-28.

International Schizophrenia Consortium (ISC) (2008) Rare chromosomal deletions and duplications increase risk of schizophrenia. Nature 455:237-241.

International Schizophrenia Consortium (ISC) (2009a) Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. Nature 460:748-752.

Iozzo RV (1999) The biology of the small leucine-rich proteoglycans. Functional network of interactive proteins. J Biol Chem 274:18843–18846.

Irizarry RA, et al. (2003) Exploration, normalization, and summaries of high density oligonucleotide array probe level data. Biostatistics 4:249-264.

Irizarry RA, et al. (2003b) Summaries of Affymetrix GeneChip probe level data. Nucleic Acids Res 31:e15.

Jablensky A (2006) Subtyping schizophrenia: implications for genetic research. Mol Psychiatry 11:815-836.

Jantzen SG, et al. (2011a) GO trimming: Systematically reducing redundancy in large gene ontology datasets. BMC Research Notes 4.

Javitt DC, et al. (2008) Neurophysiological biomarkers for drug development in schizophrenia. Nature Rev Drug Discov 7:68-83.

Jia P, et al. (2012) A bias-reducing pathway enrichment analysis of genome-wide association data confirmed association of the MHC region with schizophrenia. J Med Genet 49:96-103.

Johnson JM, et al. (2003) Genome-wide survey of human alternative pre-mRNA splicing with exon junction microarrays. Science 302.

Johnson JM, et al. (2003) Genome-wide survey of human alternative pre-mRNA splicing with exon junction microarrays. Science 302:2141-2144.

Jung S, et al. (2012) Decreased expression of extracellular matrix proteins and trophic factors in the amygdala complex of depressed mice after chronic immobilization stress. BMC Neurosci 13.

Kalsotra A, et al. (2008) A postnatal switch of CELF and MBNL proteins reprograms alternative splicing in the developing heart. Proc Natl Acad Sci U S A 105:20333-20338.

Kaneko H, et al. (2011) Hyperuricemia cosegregating with osteogenesis imperfecta is

associated with a mutation in GPATCH8. Hum Genet 130:671-683.

Kayali R, et al. (2012) Read-through compound 13 restores dystrophin expression and improves muscle function in the mdx mouse model for Duchenne muscular dystrophy. Hum Mol Genet 21:4007-4020.

Kendell RE, & Kemp IW (1989) Maternal Influenza in the Etiology of Schizophrenia. Archives of General Psychiatry 46:878-882.

Kent, WJ (2002a) BLAT - the BLAST-like alignment tool. Genome Res 12:656-664.

Kent WJ, et al. (2002b) The human genome browser at UCSC. Genome Res 12:996-1006.

Kim AH, et al. (2012) Experimental validation of candidate schizophrenia gene ZNF804A as target for hsa-miR-137. Schizophr Res 141:60-64.

Kim S-Y, & Volsky DJ (2005) PAGE: Parametric analysis of gene set enrichment. BMC Bioinformatics 6.

Kirov G, et al. (2005) Finding schizophrenia genes. Clin Invest 115:1440-1448.

Kirov G, et al. (2009a) Support for the involvement of large copy number variants in the pathogenesis of schizophrenia. Hum Mol Genet 18:1497-1503.

Kirov G, et al. (2009b) Neurexin 1 (NRXN1) deletions in schizophrenia. Schizophr Bull 35:851-854.

Kirov G, et al. (2012) De novo CNV analysis implicates specific abnormalities of postsynaptic signalling complexes in the pathogenesis of schizophrenia. Mol Psychiatry 17:142-153.

Koike H, et al. (2006) Disc1 is mutated in the 129S6/SvEv strain and modulates working memory in mice. Proc Natl Acad Sci U S A 103:3693-3697.

Kraepelin E (1899) Psychiatry: A Textbook for Students and Physicians. New Delhi: Amerind Publishing Co.

Kukita A, et al. (1990) Osteoinductive factor inhibits formation of human osteoclast-like cells. Proc Natl Acad Sci U S A 87:3023–3026.

Kuswanto et al. (2012) Genome-wide supported psychosis risk variant in ZNF804A gene and impact on cortico-limbic WM integrity in schizophrenia. Am J Med Genet B Neuropsychiatr Genet 159B:255-262.

Kwan T, et al. (2007) Heritability of alternative splicing in the human genome. Genome Res 17:1210-1218.

Kwan T, et al. (2008) Genome-wide analysis of transcript isoform variation in humans. Nat Genet 40:225-231.

Langmead B, et al. (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol 10:R25.

Leimeister C, et al. (2002) Developmental expression and biochemical characterization of Emu family members. Dev Biol 249:204-218.

Lencz T, et al. (2007) Converging evidence for a pseudoautosomal cytokine receptor gene locus in schizophrenia. Molecular Psychiatry 12:572-580.

Lencz T, et al. (2010) A schizophrenia risk gene, ZNF804A, influences neuroanatomical and neurocognitive phenotypes. Neuropsychopharmacology 35:2284-2291.

Lett TA, et al. (2011) ANK3, CACNA1C and ZNF804A gene variants in bipolar disorders and psychosis subphenotype. World J Biol Psychiatry 12:392-397.

Li MX, et al. (2011) GATES: a rapid and powerful gene-based association test using extended Simes procedure. Am J Hum Genet 88:283-293.

Liew M, et al. (2004) Genotyping of single-nucleotide polymorphisms by high-resolution melting of small amplicons. Clin Chem 50:1156-1164.

Lin M, et al. (2011) RNA-Seq of human neurons derived from iPS cells reveals candidate long non-coding RNAs involved in neurogenesis and neuropsychiatric disorders. PLoS One 6:e23356.

Lipska BK, et al. (1995) Neonatal excitotoxic hippocampal damage in rats causes post-pubertal changes in pre-pulse inhibition of startle and it's disruption by apomorphine. Psychopharmacology 122:35-43.

Lisle JW, et al. (2008) Metastatic Osteosarcoma Gene Expression Differs In Vitro and In Vivo. Clin Orthop Relat Res 466:2071–2080.

Luco RF, & Misteli, T. (2011) More than a splicing code: integrating the role of RNA, chromatin and non-coding RNA in alternative splicing regulation. Curr Opin Genet Dev 21:366-372.

Lund S, et al. (2006) The dynamics of the LPS triggered inflammatory response of murine microglia under different culture and in vivo conditions. J Neuroimmunol 180:71–87.

Lykke-Andersen J. (2002) Identification of a human decapping complex associated with hUpf proteins in nonsense-mediated decay. Mol Cell Biol 22:8114-8121.

Ma S, et al. (2007) The significance of LMO2 expression in the progression of prostate cancer. J Pathol 211:278–285.

Mangalore R, & Knapp M (2007) Cost of schizophrenia in England. J Ment Health Policy Econ 10:23-41.

Markel P, et al. (1997) Theoretical and empirical issues for marker-assisted breeding of congenic mouse strains. Nat Genet 17:280-284.

McCaughan KK, et al. (1995) Translational termination efficiency in mammals is influenced by the base following the stop codon. Proc Natl Acad Sci U S A 92:5431–5435.

McGrath J, et al. (2004) A systematic review of the incidence of schizophrenia: the distribution of rates and the influence of sex, urbanicity, migrant status and methodology. BMC Med 28:2-13.

McGrath JJ, & Murray RM (2003) Risk factors for schizophrenia: from conception to

birth. In: Schizophrenia (Weinberger DR, Hirsch SR, eds), pp. 232-250. Oxford: Blackwells.

Mednick SA, et al. (1988) Adult Schizophrenia following prenatal exposure to an influenza epidemic. Archive of General Psychiatry 45:189-192.

Mefford HC, et al. (2008) Recurrent rearrangements of chromosome 1q21.1 and variable pediatric phenotypes. N Engl J Med 359:1685-1699.

Meyer U, et al. (2007) The neurodevelopmental impact of prenatal infections at different times of pregnancy: The earlier the worse? Neuroscientist 13:241-256.

Millar JK, et al. (2000) Disruption of two novel genes by a translocation co-segregating with schizophrenia. Hum Mol Genet 9:1415-1423.

Miner JH, et al. (2004) Compositional and structuralrequirements for laminin and basement membranes during mouse embryo implantation and gastrulation. Development 131:2247-2256.

Mootha VK et al. (2003) PGC-1alpha-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. Nature Genetics 34:267-273.

Mortazavi A, et al. (2008) Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nat Methods 5:621-628.

Morton AJ, et al. (2005) A combination drug therapy improves cognition and reverses gene expression changes in a mouse model of Huntington's disease. Eur J Neurosci 21:855-870.

Moskvina V, et al. (2009) Gene-wide analyses of genome-wide association data sets: evidence for multiple common risk alleles for schizophrenia and bipolar disorder and for overlap in genetic risk. Mol Psychiatry 14:252-260.

Moskvina V, et al. (2011) Evaluation of an Approximation Method for Assessment of overall significance of multiple-dependent tests in a genome-wide association study. Genetic Epidemiology 35:861-866.

Mouse Genome Sequencing Consortium, et al. (2002) Initial sequencing and comparative analysis of the mouse genome. Nature 420:520-562.

Murray CJL, & Lopez, AD (1996) The Global Burden of Disease: A Comprehensive Assessment of Mortality and Disability from Diseases, Injuries, and Risk Factors in 1990 and Projected to 2020. Cambridge, MA: Harvard University Press,.

Murray RM, & Lewis SW (1987) Is Schizohrenia a Neurodevelopmental Disorder? British Medical Journal 295:681-682.

Murray RM, et al. (1985) Genes and Environment in Schizophrenia. In: Genetics Aspects of Human Behaviour (Tsuboi TST, ed). Tokyo: Igaku Shoin.

Mössner R, et al. (2012) The schizophrenia risk gene ZNF804A influences the antipsychotic response of positive schizophrenia symptoms. Eur Arch Psychiatry Clin Neurosci 262:193-197.

Nagy E, &. Maquat, LE (1998) A rule for termination-codon position within intron-containing genes: when nonsense affects RNA abundance. Trends Biochem Sci 23:198-199.

Neu-Yilik G, et al. (2011) Mechanism of escape from nonsense-mediated mRNA decay of human beta globin transcripts with nonsense mutations in the first exon. RNA 17:843-854.

Nielsen JA, et al. (2004) Myelin transcription factor 1 (Myt1) modulates the proliferation and differentiation of oligodendrocyte lineage cells. Mol Cell Neurosci 25:111-123.

Nolan PM et al. (2000) A systematic, genome-wide, phenotype-driven mutagenesis programme for gene function studies in the mouse. Nat Genet 25:440-443.

Nurtdinov RN, et al. (2003) Low conservation of alternative splicing patterns in the human and mouse genomes. Hum Mol Genet 12:1313-1320.

O'Callaghan E, et al. (1991a) Schizophrenia after prenatal exposure to 1957 A2 influenza epidemic. The Lancet 337:1248-1250.

O'Callaghan E, et al. (1991b) Season of Birth in Schizophrenia. Evidence for confinement of an excess of winter births to patients without a family history of mental disorder. The British Journal of Psychiatry 158:764-769.

O'Donovan MC, et al. (2008) Identification of novel schizophrenia loci by genome-wide association and follow-up. Nature Genetics 40:1053-1055.

O'Donovan MC, et al. (2009) Genetics of psychosis; insights from views across the genome. Hum Genet 126:3-12.

O'Dushlaine C, et al. (2011) Molecular pathways involved in neuronal cell adhesion and membrane scaffolding contribute to schizophrenia and bipolar disorder susceptibility. Mol Psychiatry 16:286-292.

Okada T, et al. (2012) Expression analysis of a novel mRNA variant of the schizophrenia risk gene ZNF804A. Schizophr Res 141:277-278.

Okoniewski MJ, & Miller CJ (2008) Comprehensive Analysis of Affymetrix Exon Arrays Using BioConductor. PLoS Comput Biol 4:e6.

Okoniewski MJ, et al. (2007) High correspondence between Affymetrix exon and standard expression arrays. Biotechniques 42:181-185.

Oksman M, et al. (2006) Brain reward in the absence of alpha-synuclein. Neuroreport 17:1191-1194.

Owen MJ, et al. (2005) Schizophrenia: genes at last? Trends Genet 21:518-525.

Owen MJ, et al. (2011) Neurodevelopmental hypothesis of schizophrenia. Br J Psychiatry 198:173-175.

Palais RA, et al. (2005) Quantitative heteroduplex analysis for single nucleotide polymorphism genotyping. Anal Biochem 346:167-175.

Pan Q, et al. (2008) Deep surveying of alternaitve splicing complexity in the human transcriptome by high-throughput sequencing. Nature Genetics 40:1413-1415.

Patton BL, et al. (2001) Properly formed but improperly localized synaptic specializations in the absence of laminin alpha4. Nat Neurosci 4:597-604.

Paulus FM, et al. (2011) Partial support for ZNF804A genotype-dependent alterations in prefrontal connectivity. Hum Brain Mapp.

Pedrosa E, et al. (2011) Development of patient-specific neurons in schizophrenia using induced pluripotent stem cells. J Neurogenet 25:88-103.

PGC (2011a) Genome-wide association study identifies five new schizophrenia loci. Nat Genet 43:969-976.

PGC (2011b) Large-scale genome-wide association analysis of bipolar disorder identifies a new susceptibility locus near ODZ4. Nat Genet 43:977-983.

Pieler T, & Bellefroid E (1994) Perspectives on zinc finger protein function and evolution--an update. Mol Biol Rep 20:1-8.

Pinkstaff JK, et al. (1999) Integrin subunit gene expression is regionally differentiated in adult brain. J Neurosci 19:1541-1556.

Piontkewitz Y, et al. (2012) Effects of Risperidone treatment in adolescence on hippocampal neurogenesis, parvalbumin expression and vascularization following prenatal immune activation in rats. Brain, Behaviour and Immunity 26:353-363.

Purdom E, et al. (2008) FIRMA: a method for detection of alternative splicing from exon array data. Bioinformatics 24:1707-1714.

Quwailid MM et al. (2004) A gene-driven ENU-based approach to generating an allelic series in any gene. Mamm Genome 15:585-591.

Raghavachari et al. (2012) A systematic comparison and evaluation of high density exon arrays and RNA-seq technology used to unravel the peripheral blood transcriptome of sickle cell disease. BMC Med Genomics 5.

Rantakallio P, et al. (1997) Association between central nervous system infections during childhood and adult onset schizophrenia and other psychoses: a 28-year follow-up. Int J Epidemiol 26:837-843.

Rasetti R, et al. (2011) Altered cortical network dynamics: a potential intermediate phenotype for schizophrenia and association with ZNF804A. Arch Gen Psychiatry 68:1207-1217.

Rebbapragada I, & Lykke-Andersen J. (2009) Execution of nonsense-mediated mRNA decay: what defines a substrate? Curr Opin Cell Biol 21:394-402.

Revil T, et al. (2010) Alternative splicing is frequent during early embryonic development in mouse. BMC Genomics 11.

Riley B, et al. (2010) Replication of association between schizophrenia and ZNF804A in the Irish Case-Control Study of Schizophrenia sample. Mol Psychiatry 15:29-37.

Ririe KM, et al. (1997) Product differentiation by analysis of DNA melting curves during the polymerase chain reaction. Anal Biochem 245:154-160.

Roberts A, et al. (2011) Improving RNA-Seq expression estimates by correcting for fragment bias. Genome Biol 12:R22.

Roberts G (1991) Schizophrenia: a neuropathological perspective. British Journal of Psychiatry, 158:8-17.

Robinson JT, et al. (2011) Integrative genomics viewer. Nat Biotechnol 29:24-26.

Robinson MD, & Speed, TP. (2009) Differential splicing using whole-transcript microarrays. BMC Bioinformatics 10.

Rose EJ, & Donohoe G (2012) Brain vs Behavior: An Effect Size Comparison of Neuroimaging and Cognitive Studies of Genetic Risk for Schizophrenia. Schizophr Bull.

Ross CA, et al. (2006) Neurobiology of schizophrenia.
. Neuron 52:139-153.

Ross DE, & Margolis (2005) Neurogenetics: Insights into degenerative diseases and approached to schizophrenia. Clin Neurosci Res 5:3-14.

Ruoslahti E (1996) Brain extracellular matrix. Glycobiology 6:489–492.

Saha S, et al. (2005) A systematic review of the prevalence of schizophrenia. PLoS Medicine 2:e141.

Samuelsson AM, et al. (2006) Prenatal exposure to interleukin-6 results in inflammatory neurodegeneration in hippocampus with NMDA/GABA(A) dysregulation and impaired spatial learning. Am J Physiol Regul Integr Comp Physiol 290:R1345-1356.

Schroeder A, et al. (2006) The RIN: an RNA integrity number for assigning integrity values to RNA measurements. BMC Mol Biol 7.

Sham PC, et al. (1992) Schizophrenia following pre-natal exposure to influenza epidemics between 1939 and 1960. The British Journal of Psychiatry 160:461-466.

Shi J, et al. (2009) Common variants on chromosome 6p22.1 are associated with schizophrenia. Nature 460:753-757.

Simes RJ (1986) An improved Bonferroni Procedure for multiple tests of significance. Biometrika 73:751-754.

Snyder SH (2006) Dopamine Receptor Excess and mouse madness. Neuron 49:484-485.

Sommer T, & Wolf, DH (1997) Endoplasmic reticulum degradation: reverse protein flow of no return. FASEB J 11:1227-1233.

Sonenberg N, & Hinnebusch AG (2009) Regulation of Translation Initiation in Eukaryotes: Mechanisms and Biological Targets. Cell 135:731-745.

Sorek R, et al. (2004) How prevalent is functional alternative splicing in the human genome? Trends Genet 20:68-71.

Specht CG, & Schoepfer R (2001) Deletion of the alpha-synuclein locus in a

subpopulation of C57BL/6J inbred mice. BMC Neurosci 2.

Specht CG, & Schoepfer, R. (2004) Deletion of multimerin-1 in alpha-synuclein-deficient mice. Genomics 83:1176-1178.

Sprooten E, et al. (2012) An investigation of a genomewide supported psychosis variant in ZNF804A and white matter integrity in the human brain. Magn Reson Imaging 30:1373-1380.

Stalder L, & Mühlemann O. (2008) The meaning of nonsense. Trends Cell Biol 18:315-321.

Steen RG, et al. (2006) Brain volume in first-episode schizophrenia: systematic review and meta-analysis of magnetic resonance imaging studies. Br J Psychiatry 188:510-518.

Stefanis NC, et al. (2012) Variation in Psychosis Gene ZNF804A Is Associated With a Refined Schizotypy Phenotype but Not Neurocognitive Performance in a Large Young Male Population. Schizophr Bull.

Stefansson H, et al. (2009) Common Variants coferring risk of schizophrenia. Nature 460:744-748.

Stefansson H, et al. (2008) Large recurrent microdeletions associated with schizophrenia. Nature 455:232-236.

Stefansson H et al. (2009) Common variants conferring risk of schizophrenia. Nature 460:744-747.

Steinberg S, et al. (2010) Expanding the range of ZNF804A variants conferring risk of psychosis. Molecular Psychiatry:16:59-66.

Steinberg S, et al. (2011) Common variants at VRK2 and TCF4 conferringrisk of schizophrenia. Human Molecular Genetics 20:4076-4081.

Stevens JC, et al. (2007) Quiet mutations in inbred strains of mice. Trends Mol Med 13:512-519.

Storey JD (2002) A direct approach to fasle discovery rates. J R Statist Soc B 64:479-498.

Subramanian A, et al. (2005) Gene-set enrichment analysis: A knowledge based approach for intepreting genome-wide expression profiles. PNAS 102:15545-15550.

Sullivan PF, et al. (2008) Genomewide association for schizophrenia in the CATIE study: results of stage 1. Mol Psychiatry 13:570-584.

Sullivan PF (2010) The Psychiatric GWAS Consortium: Big science comes to psychiatry. Neuron 68:182-186.

Suryo Rahmanto Y, et al. (2007) Identification of distinct changes in gene expression after modulation of melanoma tumor antigen p97(melanotransferrin) in multiple models in vitro and in vivo. Carcinogenesis 28:2172–2183.

Susser E, et al. (1996) Schizophrenia after prenatal famine. Further evidence. Arch Gen Psychiatry 53:25-31.

Sørensen HJ, et al. (2009) Association Between Prenatal Exposure to Bacterial Infection and Risk of Schizophrenia. Schizophr Bull 35:631–637.

Talkowski ME et al. (2012) Sequencing chromosomal abnormalities reveals neurodevelopmental loci that confer risk across diagnostic boundaries. Cell 149:525-537.

Tamhane AC, & Dunlop DD (2000) Statistics and Data Analysis from Elementary to Intermediate: Prentice Hall.

Tatenhorst L, et al. (2005) Genes associated with fast glioma cell migration in vitro and in vivo. Brain Pathol 15:46–54.

Tavazoie S, et al. (1999) Systematic determination of genetic network architecture. Nature Genetics 22:281-285.

Thorvaldsdóttir H, et al. (2012) Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. Brief Bioinform.

Thurin K, et al. (2012) Effects of ZNF804A on neurophysiologic measures of cognitive control. Mol Psychiatry.

Thyboll J, et al. (2002) Deletion of the laminin alpha4 chain leads to impaired microvessel maturation. Mol Cell Biol 22:1194-1202.

Tian Y, et al. (2011) Exon expression and alternatively spliced genes in Tourette Syndrome. Am J Med Genet B Neuropsychiatr Genet 156B:72-78.

Tischfield MA, et al. (2005) Homozygous HOXA1 mutations disrupt human brainstem, inner ear, cardiovascular and cognitive development. Nature Genetics 37:1035–1037.

Torkamani A, et al. (2008) Pathway analysis of seven common diseases assessed by genome-wide association. Genomics 92:265-272.

Trapnell C, et al. (2009) TopHat: discovering splice junctions with RNA-Seq. Bioinformatics 25:1105-1111.

Trapnell C, et al. (2010) Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nat Biotechnol 28:511-515.

Trapnell C, et al. (2012) Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. Nat Protoc 7:562-578.

Tsai MH, et al. (2007) Gene expression profiling of breast, prostate, and glioma cells following single versus fractionated doses of radiation. Cancer Res 67:3845–3852.

Van Den Bossche MJ, et al. (2012) Less cognitive and neurological deficits in schizophrenia patients carrying risk variant in ZNF804A. Neuropsychobiology 66:158-166.

Van Gelder RN, et al. (1990) Amplified RNA synthesized from limited quantities of heterogeneous cDNA. Proc Natl Acad Sci U S A 87:1663-1667.

Venables JP. (2006) Unbalanced alternative splicing and its significance in cancer.

Bioessays 28:378-386.

Visscher, PM (1999) Speed congenics: accelerated genome recovery using genetic markers. Genetical Research 74:81-85.

Voineskos AN, et al. (2011) The ZNF804A gene: characterization of a novel neural risk mechanism for the major psychoses. Neuropsychopharmacology 36:1871-1878.

Walker EF. (1994) Developmentally moderated expressions of the neuropathology underlying schizophrenia. Schizophr Bull 20:453-480.

Walsh FS, & Doherty P (1991a) Structure and function of the gene for neural cell adhesion molecule. Seminars in Neuroscience 3:271–284.

Walsh FS, & Doherty P (1991b) Glycosylphosphatidylinositol anchored recognition molecules that function in axonal fasciculation, growth and guidance in the nervous system. Cell Biol Int Rep 15:1151-1166.

Walsh T, et al. (2008) Rare structural variants disrupt multiple genes in neurodevelopmental pathways in schizophrenia. Science 320:539-543.

Walter H, et al. (2010) Effects of a genome-wide supported psychosis risk variant on neural activation during a theory-of-mind task. Mol Psychiatry 16:462-470.

Walters JTR et al. (2010) Psychosis susceptibility gene ZNF804A and cognitive performance in schizophrenia. Arch Gen Psychiatry 67:692-700.

Wang ET, et al. (2008) Alternative isoform regulation in human tissue transcriptomes. Nature 456:470-476.

Wang K et al. (2009) Common genetic variants on 5p14.1 associate with autism spectrum disorders. Nature 459:528-533..

Wassink TH, et al. (2001) Chromosomal abnormalities in a clinic sample of individuals with autistic disorder. Psychiatric Genetics 11:57–63.

Wassink TH, et al. 2012) Influence of ZNF804a on brain structure volumes and symptom severity in individuals with schizophrenia. Arch Gen Psychiatry 69:885-892.

Weckx S, et al. (2005) novoSNP, a novel computational tool for sequence variation discovery. Genome Res 15:436–442.

Wei Q, et al. (2012) Association of the ZNF804A gene polymorphism rs1344706 with white matter density. Prog Neuropsychopharmacol Biol Psychiatry 36:122-127.

Wei Q, et al. (2012b) No association of ZNF804A rs1344706 with white matter integrity in schizophrenia A tract-based spatial statistics study. Neurosci Lett.

Weinberger D. R. (1986) The pathogenesis of schizophrenia: a neurodevelopmental theory. In: The Neurology of Schizophrenia. ed Nasrallah, H & Weinberger, DR. pp 387–405: Elsevier.

Weiss RB, et al. (1987) Slippery runs, shifty stops, backward steps, and forward hops: -2, -1, +1, +2, +5, and +6 ribosomal frameshifting. Cold Spring Harb Symp Quant Biol 52:687-693.

Whistler T, et al. (2010) The comparison of different pre- and post-analysis filters for determination of exon-level alternative splicing events using Affymetrix arrays. Journal of Biomolecular Techniques 21:44-53.

Williams H, et al. (2010) Fine mapping of ZNF804A and genome-wide significant evidence for its involvement in schizophrenia and bipolar disorder. Molecular Psychiatry:1-13.

Williams HJ, et al. (2011) Most genome-wide significant susceptibility loci for schizophrenia and bipolar disorder reported to date cross-traditional diagnostic boundaries. Hum Mol Genet 20:387-391.

Wong AH, & Van Tol HH (2003) Schizophrenia: from phenomenology to neurobiology. Neurosci Biobehav Rev 27:269–306.

Xiao B, et al. (2011) To the editor: association of ZNF804A polymorphisms with schizophrenia and antipsychotic drug efficacy in a Chinese Han population. Psychiatry Res 190:379-381.

Xing Y, et al. (2008) MADS: a new and improved method for analysis of differential alternative splicing by exon-tiling microarrays. RNA 14:1470-1479.

Yeo GW, et al. (2005) Identification and analysis of alternative splicing events conserved in human and mouse. PNAS 102:2850-2855.

Young MD, et al. (2010) Gene ontology analysis for RNA-seq: accounting for selection bias. Genome Biol 11:R14.

Zhang W, (2008) Evaluation of genetic variation contributing to differences in gene expression between populations. Am J Hum Genet 82:631-640.

Zhang F, et al. (2011a) Evidence of sex-modulated association of ZNF804A with schizophrenia. Biological Psychiatry 69:914-917.

Zhang R, et al. (2011b) Is the conserved mammalian region of ZNF804A locus associated with schizophrenia? A population-based genetics analysis. Schizophrenia Research 133:159-164.

Zhang J, et al (2012) Association analysis of ZNF804A (zinc finger protein 804A) rs1344706 with therapeutic response to atypical antipsychotics in first-episode Chinese patients with schizophrenia. Compr Psychiatry 53:1044-1048.

Zollino M, et al. (2011) Integrated analysis of clinical signs and literature data for the diagnosis and therapy of a previously undescribed 6p21.3 deletion syndrome. European Journal of Human Genetics 19:239–242.

Zweier C, et al. (2007) Haploinsufficiency of TCF4 Causes Syndromal Mental Retardation with Intermittent Hyperventilation (Pitt-Hopkins Syndrome). The American Journal of Human Genetics 80:994-1001.