Characterising the function of *ZNF804A*: a top genome-wide association study hit for schizophrenia

By

Ria M. Chapman

A thesis submitted to Cardiff University for the degree of Doctor of Philosophy

Supervisors: Prof. Derek Blake and Prof. Michael O'Donovan

April 2013

Summary

Schizophrenia is a debilitating psychiatric disorder with high heritability. Genome wide association studies (GWASs) have identified association between schizophrenia and an intronic single nucleotide polymorphism (SNP) in *zinc finger domain containing 804A* (*ZNF804A*). The biological functions of ZNF804A are largely unknown. Thus, the aim of this thesis was to determine the function of ZNF804A.

Yeast two-hybrid (Y2H) screens were used to determine the protein binding partners of ZNF804A in the brain. This identified ZNF804A interacted with transcription factors (zinc finger protein 40 (ZNF40); trans-acting transcription factor 1 (Sp1) and basic helix-loop-helix family, member e40 (Bhlhe40)) and regulators of pre-mRNA splicing, including RNA binding protein, fox-1 and -2 (Rbfox1 and RBFOX2) and neuro-oncological ventral antigen 2 (NOVA2). Therefore, we hypothesised that, via interactions with its protein binding partners, ZNF804A may have a role in regulating transcription and pre-mRNA splicing.

Exon arrays were used to determine the effects of *ZNF804A* knockdown on gene expression in SH-SY5Y neuroblastoma cells. Enrichment analysis on the differentially expressed or spliced genes indicated a significant effect of *ZNF804A* knockdown on genes involved in nervous system development, particularly synaptic contact and axon guidance. Among the most significantly differentially expressed genes were the known schizophrenia susceptibility genes *reelin* (*RELN*) and *neuropeptide Y* (*NPY*). Several alternative splicing events were confirmed empirically, including increased exclusion of exon 11a of *enabled homolog* (*ENAH*). Consistent with our hypothesis, splicing of exon 11a of *ENAH* is known to be regulated by RBFOX2.

In complementary experiments, exon arrays were used to identify differentially expressed genes in a stable cell line expressing *myc-ZNF804A*. Enrichment analysis on the differentially expressed genes indicated an over-representation of genes involved in the regulation of epithelial to mesenchymal transition and receptor-mediated axon growth repulsion. Among the genes with the largest fold changes in expression was a gene implicated in synapse development: *secreted protein acidic and rich in cysteine* (*SPARC*). Enrichment analysis of the alternatively spliced genes indicated a significant effect of *myc-ZNF804A* over-expression on genes involved in cell-matrix interactions.

These data suggest that ZNF804A regulates the expression of genes implicated in processes underlying schizophrenia pathology, and provide the first evidence that ZNF804A may be involved in the regulation of alternative splicing.

Acknowledgments

I would like to thank Prof. Derek Blake and Prof. Mick O'Donovan for allowing me the opportunity to work on the ZNF804A project. I would like to express my gratitude to Derek for his guidance and constructive critique throughout my PhD studies. I am very grateful for his patience and support.

I am hugely indebted to the Blake lab (Dr. Caroline Tinsley, Dr. Adrian Waite, Marc Forrest and Michelle Doyle) for their technical knowledge and insightful discussions of my research. I would particularly like to thank Dr. Caroline Tinsley for welcoming me as a friend and for her contributions to the ZNF804A project.

To the girls on the second floor; thank you for being fantastic friends and officemates.

I would like to say a big thank you to my Mum and Dad, for their constant support and for giving me the strength to stay focused and motivated. To Amelia and Alice, thank you for all of your love, laughter and encouragement. And of course, my utmost thanks to Greg, for his love and unwavering faith in me.

List of abbreviations

°C	degrees celsius
3-AT	3-aminotriazole
aCGH	array-comparative genomic hybridisation
A	adenine
ADHD	attention hyperactivity deficit disorder
amp	ampicillin
ANOVA	analysis of variance
APS	ammonium persulphate
ASD	autism spectrum disorder
ATCC	American type culture collection
BCA	bicinchoninic acid
BCAs	balanced chromosomal abnormalities
BLAST	Basic Local Alignment Search Tool
BLAT	BLAST-like alignment tool
bp	base pairs
BSA	bovine serum albumin
C	cysteine
C2H2	$cysteine_2$-$histidine_2$
CBS	central biotechnology services
CDCV	common disease common variant
cDNA	complementary DNA
CDRV	common disease rare variant
ChIP	chromatin immunoprecipitation
CNV	copy number variant
DLPFC	dorsolateral prefrontal cortex
DMEM	Dulbecco's modified eagle media
DMP	dimethyl pimelimidate
DMSO	dimethyl sulfoxide
DoC	deoxycholate
DSHB	developmental studies hybridoma bank
DSM	diagnostic and statistical manual of mental disorders
DTT	dithiothreitol
ECACC	european collection of cell cultures
EDTA	ethylenediaminetetraacetic acid
EGTA	ethylene glycol tetraacetic acid
ESTs	expressed sequence tags

EtBr	ethidium bromide
FCS	foetal calf serum
FDR	false discovery rate
fwd	forward
FRT	Flp recognition target
G	guanine
GABA	γ-aminobutyric acid
GFP	green fluorescent protein
GO	gene ontology
GWAS	genome wide association study
h	hour
HGNC	HUGO gene nomenclature committee
HEK	human embryonic kidney
hnRNP	heterogeneous nuclear ribonucleoprotein
ICD	International classification of diseases
IP	immunoprecipitation
IPA	isopropanol
IPTG	isopropyl β-$_D$-1-thiogalactopyranoside
kan	kanamycin
kb	kilobases
kDa	kilodalton
LB	Luria-Bertani
LD	linkage disequilibrium
LDS	lithium dodecyl sulphate
Log_2	logarithm to the base 2
M	molar
MADS	microarray analysis of differential splicing
MHC	major-histocompatibility complex
MIDAS	microarray detection of alternative splicing
min	minute
MRI	magnetic resonance imaging
mRNA	messenger RNA
NCBI	National Centre for Biotechnology Information
NGS	Next generation sequencing
NICE	National Institute for Health and Clinical Excellence
NLS	nuclear localisation signal

NMDA	*N*-methyl-$_D$-aspartic acid
OD	optical density
PAGE	polyacrylamide gel electrophoresis
PBS	phosphate buffered saline
PCR	polymerase chain reaction
PFA	paraformaldehyde
PGC	psychiatric GWAS consortium
PGS	Partek Genomics Suite
QA	quality assessment
Q-PCR	quantitative PCR
rev	reverse
RNAi	RNA interference
rpm	revolutions per minute
RRM	RNA recognition motifs
RT-PCR	reverse-transcription PCR
sd	standard deviation
SDS	sodium dodecyl sulphate
sec	second
siRNA	small interfering RNA
SMART	simple modular architecture research tool
SNP	single nucleotide polymorphism
snRNPs	small nuclear ribonucleoproteins
SUMO	small ubiquitin-related modifier
T	thymine
TAE	tris-acetate-EDTA
TE	tris-EDTA
UCSC	University of California Santa Cruz
UV	ultra-violet
V	volts
v/v	volume/volume
w/v	weight/volume
WHO	World Health Organisation
Y2H	yeast two-hybrid
YC	yeast complete
YPAD	yeast-peptone-adenine-glucose
ZnF	zinc finger

Contents

Chapter 1: Introduction

1.1. Introduction

Schizophrenia is a major psychiatric disorder that alters an individual's perception, thoughts, behaviour and mood. Schizophrenia affects as many as 1% of the world's population, making it the most common form of psychotic disorder.

The first symptoms of schizophrenia present in early adulthood and persist throughout the life-time. On average, males experience the onset of disease 1.07 years earlier than females (Eranti et al., 2012). Schizophrenia is associated with reduced fertility and may result in a reduced life expectancy, due to an increased risk of suicide, health problems, and risky behaviour (Bundy et al., 2011; Chang et al., 2011). The difficulties experienced by people with schizophrenia are compounded by social adversity and isolation, poverty, homelessness and the stigma associated with mental illness (NICE, 2010; Overton and Medina, 2008). As such, schizophrenia is considered among the top ten leading causes of disability worldwide (WHO, 2001). Additionally, a diagnosis of schizophrenia places strain on the family and friends of the patient (Baronet, 1999) whilst, in economic terms, schizophrenia places large demands on the healthcare system. For example, in 2004/2005 the societal cost of schizophrenia in England alone was £6.7 billion (Mangalore and Knapp, 2007).

The incidence rate of schizophrenia in England from 1950-2009 was 15 people per 100 000 per year (Kirkbride et al., 2012). The lifetime prevalence of schizophrenia is estimated to be 0.4% (Saha et al., 2005).

Although medication remains paramount in the treatment of schizophrenia, the drugs which are currently available are ineffective for as many as one third of patients and have undesirable side effects, which can often lead to poor compliance (Conley and Buchanan, 1997; Kahn et al., 2008; Lewis et al., 2006; Lieberman et al., 2005). The lack of effective treatments is generally attributed to our poor understanding of the biological aetiology of schizophrenia. Therefore, it is hoped that an improved understanding of the underlying molecular pathophysiology of schizophrenia may enable more effective treatments to be developed (Insel and Scolnick, 2006).

1.2. What is schizophrenia?

The modern concept of schizophrenia was first recognised by Emil Kraepelin (1855-1926). Kraepelin defined two courses of mental illness: a progressive deterioration of symptoms, with increasing impairment; and a fluctuating pattern, with frequent relapses and full recovery between episodes (Kraepelin, 1896). Kraepelin termed these mental illnesses *dementia praecox* (later known as schizophrenia) and manic-depressive insanity (later known as bipolar disorder). Kraepelin defined the primary symptoms of *dementia praecox* as delusions, hallucinations, formal thought disorder, and negativism. Although Kraepelin was pessimistic about the prognosis of *dementia praecox*, other psychiatrists believed the long-term outcome of the disease was more positive. Indeed, the term 'schizophrenia' was coined by Eugen Bleuler in 1911 to describe a group of related mental disorders in a way which did not suggest a deteriorating course (Bleuler, 1950). In contrast to Kraepelin, Bleuler believed that the essence of schizophrenia was negative symptoms, rather than delusions and hallucinations. Later, Kurt Schneider proposed that the primary deficits in schizophrenia were positive symptoms (Schneider, 1959). The modern-day international diagnostic criteria for schizophrenia (including the World Health Organisation's (WHO) International

Classification of Disease's (ICD) and the American Psychiatric Association Diagnostic and Statistical Manual (DSM)-IV)) incorporate Kraepelin chronicity, Bleulerian negative symptoms and Schneiderian positive symptoms (Table 1.1) (Tandon et al., 2009).

Despite the use of common diagnostic criteria, many psychiatrists continue to argue that the clinical definition of schizophrenia is arbitrary because there are no known biomarkers for schizophrenia and no disorder-specific brain abnormality (Carroll and Owen, 2009). The findings of genetic studies support this argument and suggest that, rather than being a single disease entity, schizophrenia may represent a group of disorders with similar pathology. For example, copy number variations (CNVs) in *neurexin 1* (*NRXN1*) have been associated with both schizophrenia and other neurodevelopmental disorders, such as autism (Kim et al., 2008; Kirov et al., 2009b). The issue of defining schizophrenia is important, not only because a diagnosis of schizophrenia has a huge impact on a patient's life; but also to facilitate research into its aetiology, pathogenesis and treatment.

1.3. Clinical symptoms of schizophrenia

The clinical presentation of schizophrenia falls into three collections of symptoms: positive, negative and cognitive. Positive symptoms of schizophrenia are those that are in excess of normal functions, such as delusions and hallucinations (Lindermayer and Khan, 2006). The most common hallucinations in schizophrenia are auditory, with as many as 75% of patients perceiving sounds without the presence of an auditory source (Shinn et al., 2012). The negative symptoms of schizophrenia involve an absence or reduction in affective and conative functions. These include, abulia (loss of motivation), alogia (poverty of speech), anhedonia (inability to experience pleasure), avolition (lack of initiative), apathy (lack of interest) and reduced social drive (Bleuler, 1950). These symptoms occur frequently and

DSM-IV diagnostic criteria for schizophrenia
A. characteristic symptoms: at least two of the following each present for a significant amount of time during a one month period. i) delusions ii) hallucinations iii) disorganised speech iv) grossly disorganised or catatonic behaviour v) negative symptoms
B. social or occupational dysfunction
C. duration: continuous signs of disturbance persist for at least 6 months
D. exclusion of mood disorders
E. exclusion of known organic causes (e.g.: substance abuse or a known brain disorder)

Table 1.1. DSM-IV diagnostic criteria for schizophrenia

(Adapted from the American Psychiatric Association, 2000.)

persistently in schizophrenia and are considered to be associated with poor long-term prognosis (Kirkpatrick et al., 2001; Strauss et al., 2010). The current medications have only modest effects on the negative symptoms of schizophrenia. Therefore, these symptoms remain a debilitating component of schizophrenia pathology (Erhart et al., 2006; Leucht et al., 1999). The cognitive impairments associated with schizophrenia include general intelligence (IQ), verbal memory, attention and executive function (Fioravanti et al., 2005; Keefe et al., 2006; Reichenberg, 2010). However, data suggests that deficits in IQ may not be universally characteristic in schizophrenia (Weickert et al., 2000). For example, some patients experience intellectual decline from childhood (David et al., 1997; Reichenberg et al., 2005), while others have an IQ within the normal range directly before the onset of the illness and do not show impairments in intellectual performance after the first episode (Reichenberg et al., 2005; Weickert et al., 2000). The cognitive profile of schizophrenia is highly heterogeneous; for example, as many as 27% of patients may have no neuropsychological deficits (Palmer et al., 1997). Despite this, the importance of cognitive symptoms, and their broad presence at the onset of disease, has led to the suggestion that they should be incorporated into the major diagnostic criteria for schizophrenia (Keefe, 2007).

1.4. The neurobiology of schizophrenia

The underlying molecular basis of schizophrenia remains elusive, yet there are several neuroimaging, neuropathological and neurochemical findings which have been reproducibly associated with the disease.

There are two reproducible neuroanatomical findings that distinguish schizophrenia patients from controls: enlarged ventricles and a reduction in whole brain volume at both the outset of illness and in the longer-term (Shepherd et al., 2012). Although the clinical relevance of these

changes is unclear, their presence at the outset of disease provides evidence for the neurodevelopmental hypothesis of schizophrenia (see section 1.5) (MacDonald and Schulz, 2009).

Post-mortem studies of brains from patients with schizophrenia have identified several changes associated with the disease. These include an increased density of pyramidal cells in the prefrontal cortex, without changes in total neuron number and without neuronal loss (Selemon and Goldman-Rakic, 1999). The absence of neuronal loss precludes neurodegeneration and supports the neurodevelopment hypothesis of schizophrenia (see section 1.5) (Harrison, 1997). The increase in neuronal density may be due to decreased soma size or decreased interneuronal space, known as neuropil (Figure 1.1) (Akbarian et al., 1995; Harrison, 1999; Selemon and Goldman-Rakic, 1999). Additionally, dendrites were found to be shorter and less branched in schizophrenia patients (Black et al., 2004; Guidotti, 2000). Furthermore, there was a reduction in the density of glutamate dendritic spines in the forebrains of schizophrenia patients (Glantz and Lewis, 2000; Kalus et al., 2000; Lewis et al., 2003).

Neuropathological studies have also identified changes in the interneurons of the cortex. Specifically, several studies have shown a decrease in the mRNA and protein abundance of glutamic acid decarboxylase (GAD) 67, one of the key enzymes for synthesis of the inhibitory neurotransmitter γ-aminobutyric acid (GABA) (Fatemi, 2000; Guidotti, 2000; Impagnatiello, 1998). Impaired synthesis and uptake of GABA was found to be associated with schizophrenia (Akbarian et al., 1995; Volk et al., 2000). The reduction in GAD67 in schizophrenia was associated with a decrease in the mRNA and protein abundance of reelin, a large extracellular matrix protein that is encoded by the gene *reelin* (*RELN*) and is critical

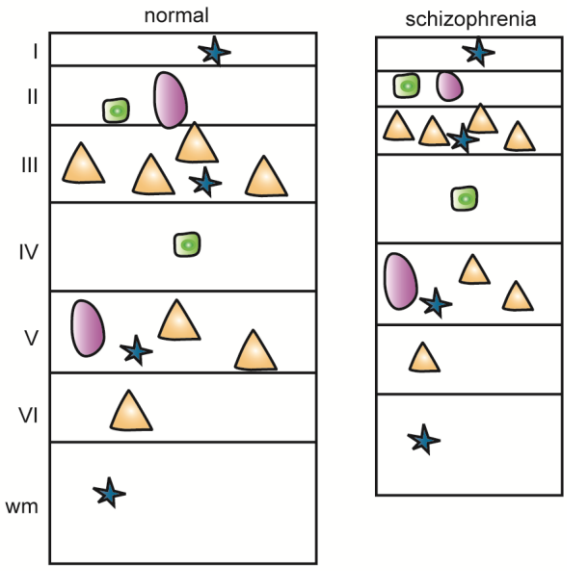


Figure 1.1. Schematic diagram of the cytoarchitectural features of schizophrenia

In schizophrenia, the number of neurons is unchanged but the somal size of the pyramidal neurons (yellow) is reduced, and the neurons are more densely packed. This reflects a reduction in the neuropil volume. The glial cells (blue) are unaffected. For clarity, potential changes in the distribution or organisation of the interneuron populations (purple and green) are omitted. I-VI represents the cortical layers; wm = white matter. (Adapted from Harrison, 1999.)

for the correct migration of neurons during development (Impagnatiello et al., 1998; Fatemi et al., 2000; Guidotti et al., 2000; Fatemi, 2005). Studies using heterozygous *reln* mice revealed that deficits in *reln* were associated with deficits in GABAergic systems and deficits in the molecular composition of synapses in the adult brain (Nullmeier et al., 2011; Ventruti et al., 2011). These data suggest that decreased *reln* expression may contribute to disease via impaired GABA neurotransmission and consequently, a reduction in neuronal connectivity (Costa et al., 2001; Costa et al., 2002).

The myelin hypothesis of schizophrenia was first presented by Hakak and colleagues (2001) following transcriptomic analysis of post-mortem brain tissue which identified differential expression of several genes involved in myelination in patients compared to controls. Subsequently, several studies have shown a reduction in oligodendrocyte number in various brain regions in schizophrenia (Byne et al., 2006; Vostrikov et al., 2007). It is postulated that these changes may contribute to aberrant neuronal connectivity in disease (Schmitt et al., 2011).

The neurochemical hypotheses of schizophrenia have been largely driven by observations that antipsychotic drugs modulate neurochemical pathways. For instance, the glutamate hypothesis developed after it was shown that psychomimetic drugs, such as phencyclidine and ketamine, elicited psychosis-like symptoms reminiscent of schizophrenia by blocking neurotransmission at *N*-methyl-$_{D}$-aspartic acid- (NMDA) type glutamate receptors and inducing a NMDA receptor hypofunction state (Javitt and Zukin, 1991; Olney et al., 1999). The resulting increase in glutamate release is thought to be linked to the behavioural deficits observed (Moghaddam et al., 1997). Data from post-mortem analyses of brain tissue obtained from patients with schizophrenia support a role for dysregulation of glutamatergic systems in

disease pathology. For example, Oni-Orisan (2008) reported a reduced abundance of the presynaptic protein vesicular glutamate transporter 1 (VGLUT1) in the anterior cingulate cortex (ACC) but not dorsolateral prefrontal cortex of patients with schizophrenia. However, the transcript expression of *VGLUT1* was unaltered in the ACC in schizophrenia (Oni-Orisan et al., 2008; Eastwood and Harrison, 2010). Additionally, alterations in glutamate receptor expression in schizophrenia have been found (Beneyto et al., 2007). Recent genetic studies of copy number variations in schizophrenia are consistent with a role for glutamate synapse disruption in disease (Walsh, 2008; Kirov, 2012); these data are discussed in more detail in section 1.6.3.2.

The dopamine hypothesis proposes that the symptoms of schizophrenia arise due to dopaminergic overactivity (Howes and Kapur, 2009). This hypothesis emerged after it was observed that anti-psychotic drugs were dopamine receptor D_2 antagonists (Carlsson et al., 1957). Molecular imaging studies confirm that all antipsychotics block striatal D2 receptors (reviewed by Frankle and Laruelle, 2002). There is considerable evidence to suggest that the neurochemical pathways underlying schizophrenia pathology may converge and interact to contribute to disease (Tost and Meyer-Lindenberg, 2011). For example, using imaging techniques Stone and colleagues (2010) showed the relationship between hippocampal glutamate and striatal dopamine was altered in patients at high risk of psychosis.

1.5. The neurodevelopmental hypothesis of schizophrenia

The neurodevelopmental hypothesis of schizophrenia suggests that the disorder arises as a result of a pathogenic process which occurs earlier in life than the onset of features of the illness (Weinberger, 1987). Considerable evidence from a wide range of studies supports this model of schizophrenia (Fatemi and Folsom, 2009). For example, Brown and colleagues

showed poor maternal nutrition and maternal infection in the prenatal period was associated with increased risk for schizophrenia in offspring (Brown and Patterson, 2011; Brown et al., 1996). Other studies revealed children who later went on to develop schizophrenia had subtle cognitive impairments; memory and language difficulties; and motor abnormalities (Chua and Murray, 1996; Erlenmeyer-Kimling et al., 2000; Walker et al., 1996). Additionally, the offspring of schizophrenics had a higher incidence of minor physical abnormalities, which may be indicative of a disruption in early neurodevelopment (Lawrie et al., 2001). It has been argued that an abnormality in late neurodevelopment, such as aberrant synaptic pruning in adolescence, may increase risk for schizophrenia (Feinberg, 1982; Lewis and Levitt, 2002). Consistent with this, Keshavan (1999) proposed a 'two-hit' neurodevelopmental hypothesis which posited that abnormalities at critical time points in both early and late neurodevelopment may act in concert to increase risk for schizophrenia.

Often schizophrenia is co-morbid with child-onset neurodevelopmental disorders such as intellectual disability, autism and attention deficit hyperactivity disorder (ADHD) (Stahlberg et al., 2004). This co-morbidity, and recent findings which point to genetic overlap between neurodevelopmental disorders, have led some psychiatrists to suggest that these disorders may represent a continuum of genetic and environmentally induced neurodevelopmental impairment (Owen et al., 2011).

1.6. Genetics of schizophrenia

1.6.1. Genetic epidemiology

The meta-analysis of modern (post-1980) family studies of schizophrenia indicated that first degree relatives were nearly 10 times more likely to be affected with schizophrenia than comparison subjects (Kendler and Diehl, 1993). This finding was consistent with a large

population-based study which showed that offspring whose parents both had schizophrenia had a risk of developing the disease of greater than 25% (Gottesman et al., 2010). Research using twin studies showed that the proportion of liability for schizophrenia that may be attributed to genetic differences between individuals was approximately 80% (Cardno and Gottesman, 2000; Sullivan et al., 2003). Furthermore, the results of adoption studies for schizophrenia demonstrated that the genetic contribution was relatively more important to the aetiology of schizophrenia than the environmental contribution (Kety, 1987). However, data suggests environmental factors specifically associated with a disadvantaged socioeconomic position, such as unemployment and a single-parent household, may increase risk for psychosis more in children with genetic susceptibility to disease (Wicks et al., 2010).

In summary, the results of genetic epidemiological studies strongly support the notion of a major genetic component to schizophrenia. Such data are important as they provide a strong case for the search for genes that underlie susceptibility to schizophrenia.

1.6.2. Genetic models of schizophrenia

The genetic architecture of complex disorders such as schizophrenia is determined by the number, frequency and penetrance of risk alleles that contribute to disease in the population. In this context, there are two primary models used to explain the genetic architecture of schizophrenia: 1) the common disease-rare variant (CDRV) hypothesis, which states that multiple rare variants contribute to risk for common disease; and 2) the common disease-common variant (CDCV) hypothesis, which states that variants of modest effect, with a relatively high frequency in the population, explain risk for common diseases (Lander, 1996). Data presented in section 1.6.3 suggest that risk for schizophrenia is influenced by the combined effect of both rare and common variants (Figure 1.2) (Mowry and Gratten, 2013).

However, the proportion of genetic susceptibility that can be explained by either rare or common variants remains a point of controversy (McClellan and King, 2010; Rodriguez-Murillo et al., 2012).

1.6.3. Candidate gene discovery in schizophrenia

1.6.3.1. Early genetic studies of schizophrenia

The early genetic studies of schizophrenia focused on linkage studies. Linkage analysis uses genetic markers to map co-segregation of chromosome regions with disease in related individuals. Linkage studies were based on the assumption that a relatively small number of genes of major effect could be identified. The early linkage studies for schizophrenia identified many regions of the genome which showed linkage with disease. For example, there was reproducible linkage between chromosome 8p and disease (Kendler et al., 1996; Pulver et al., 1995). Fine-mapping of chromosome 8p identified a haplotype in *neuregulin* (*NRG1*) that was associated with disease (Stefansson et al., 2002). *NRG1* was considered to be a likely schizophrenia candidate gene because its expression was highest in the brain and the protein had known key roles in neurodevelopmental processes (Corfas et al., 1995; Yarden and Sliwkowski, 2001). However, it remains unclear how genetic variation in *NRG1* mediates susceptibility for schizophrenia. Additionally, on chromosome 6p22-24, linkage for *dystrobrevin binding protein 1* (*DTNBP1*) in schizophrenia was reported in more than 10 independent samples (Williams et al., 2005). The disease-associated *DTNBP1* haplotypes appeared to associate with low *DTNBP1* expression (Bray et al., 2005). Reductions in the amount of DTNBP1 at the presynapse led to alterations in glutamate output, this may suggest that genetic alteration in *DTNBP1* may increase risk for schizophrenia via modulation of glutamate output (Jentsch et al., 2009; Numakawa et al., 2004; Talbot et al., 2004).

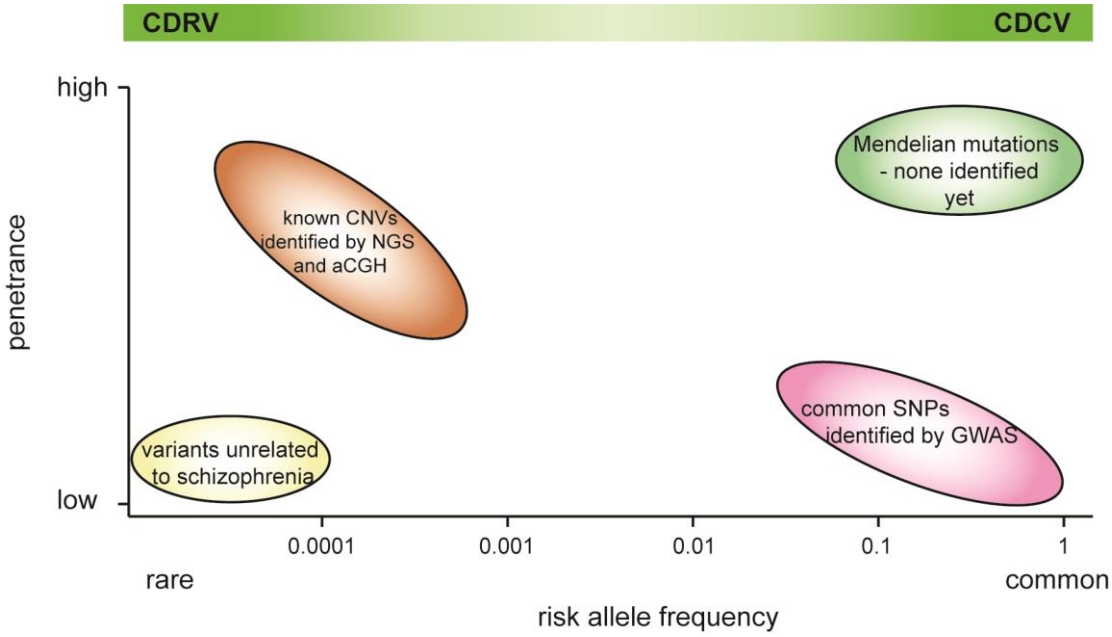


Figure 1.2. Schematic representation of the current genetic findings in schizophrenia, in relation to the proposed genetic models of the disease

Genome wide association studies (GWAS) have identified common single nucleotide polymorphisms (SNPs) with small effect size associated with schizophrenia. These data are consistent with the common disease common variant (CDCV) hypothesis. Rare disease-causing variants that have large effect, such as copy number variations (CNVs), have been identified by array-comparative genomic hybridisation (aCGH) and next generation sequencing (NGS). These data are consistent with the common disease rare variant (CDRV) hypothesis. (Adapted from Owen, 2009; Mowry and Gratten, 2013).

Notwithstanding these examples, many of the published linkage studies show inconsistent findings and a lack of replication (Owen, 2009). This was likely to arise because these studies were underpowered to detect rare alleles of large effect (section 1.6.3.2) or common alleles of small effect (section 1.6.3.3).

1.6.3.2. Studies of structural variation: common disease-rare variant

Rare variation can manifest as either rare single nucleotide variants (SNVs), copy number variations (CNVs) or large chromosomal abnormalities such as excisions, insertions and translocations of genetic material (Beckmann et al., 2007). The first chromosomal aberration associated with schizophrenia was a reciprocal translocation between 1q42 and 11q14.3, which co-segregated with schizophrenia, major depression and bipolar disorder in a large Scottish family (St Clair et al., 1990). The translocation disrupted *disrupted in schizophrenia 1* and *2* (*DISC1* and *DISC2*). The molecular characterisation of the DISC1 protein implied it may be involved in neural development; therefore, disruption of *DISC1* may contribute to schizophrenia by affecting neuronal development (Brandon and Sawa, 2011).

CNVs are submicroscopic duplications and deletions of the genome (Hastings et al., 2009). The consequences of CNVs may be loss of gene function, disruption of a regulatory element or the creation of a novel gene fusion. The first CNV found associated with schizophrenia was a microdeletion of *velo-cardio-facial syndrome* (*VCFS*) on chromosome 22q11.2 (Lindsay et al., 1995). Subsequently, several studies showed schizophrenia patients had an increased mutational burden of rare CNVs (Kirov et al., 2008; Magri, 2010; Walsh et al., 2008). The CNVs observed in schizophrenia arise both from rare *de novo* mutations and inherited CNVs (Kirov et al., 2009a; Kirov et al., 2012; Liao et al., 2012).

Locus	CNV type	References	Genes of particular interest within this region
1q21.1	deletion	(Glessner et al., 2010; Grozeva et al., 2012; Levinson et al., 2011; Stefansson et al., 2008; Vacic et al., 2011; Walsh et al., 2008)	*ERBB4*, *GJA8*
7q36.3	duplication	(Vacic et al., 2011)	*VIPR2*
2p16.3	deletion	(Walsh et al., 2008;(Kirov et al., 2009b; Magri et al., 2010; Vrijenhoek et al., 2008)	*NRXN1*
3q29	deletion	Grozeva, 2012; Magri, 2010; Mulle, 2010; Levinson, 2011	*DLG1*
3q29	duplication	(Vacic et al., 2011)	*CR597873*, *SDHALP2*
15q11.2	deletion	(Grozeva et al., 2012; Kirov et al., 2009b; Magri, 2010; Stefansson et al., 2008)	*CYFIP1*
15q11.2	duplication	(Liao et al., 2012)	*UBE3A*
15q13.3	deletion	(Grozeva et al., 2012; Levinson et al., 2012; Stefansson et al., 2008)	*CHRNA7*
15q13.3	duplication	(Ingason et al., 2011)	*UBE3A*, *ATP10A*
16p11.2	duplication	(Grozeva et al., 2012; McCarthy et al., 2009)	*DOC2A*
16p13.1	deletion	(Ingason et al., 2011)	*NTAN1, NDE1*
16p13.1	duplication	(Ingason et al., 2011; Magri, 2010)	*NDE1*
17p12	deletion	(Kirov et al., 2009b; Magri, 2010; Vacic et al., 2011)	*PMP22*
17q12	deletion	(Magri, 2010; Moreno-De-Luca et al., 2010; Vacic et al., 2011)	*CCL* genes
22q11.2	deletion	(Grozeva et al., 2012; Kirov et al., 2009a; Vacic et al., 2011)	*VCFS*, *BCR*

Table 1.2. Summary of the CNVs identified in schizophrenia

The studies of CNVs in schizophrenia have identified genomic loci which harbour multiple structural variants associated with disease; these loci are outlined in Table 1.2. Two large, genome-wide scans of rare CNVs reported a significant over-representation of copy number changes which affected single genes, namely a duplication of *vasoactive intestinal peptide receptor gene* (*VIPR2*) and a deletion of *neurexin 1* (*NRXN1*) (Kirov et al., 2009b; Vacic et al., 2011).

Enrichment analysis of genes containing CNVs showed the genes were enriched for involvement in signalling networks controlling neurodevelopment (Walsh, McClellan et al. 2008). For example, there were deletions of the *v-erb-a erythroblastic leukemia viral oncogene homolog 4 (avian)* (*ERBB4*) gene, which encodes a type 1 transmembrane tyrosine kinase receptor for neuregulin; and deletions of the *glutamate receptor, metabotropic 7* (*GRM7*) gene, which encodes a G protein-coupled metabotropic glutamate receptor (Walsh et al., 2008). Furthermore, *de novo* CNVs were significantly enriched for genes which function in synaptic plasticity such as *syntaxin 1A* (*brain*) (*STX1A*), *discs, large homolog 1* (*Drosophila*) (*DLG1*) and *rabphilin 3a homolog* (*mouse*) (*RPH3A*) (Glessner et al., 2010; Kirov et al., 2012). These findings support the glutamate hypothesis of schizophrenia (see section 1.4) and suggest that CNVs may lead to pathogenesis via influences on neural development and cognition (Kirov et al., 2012; Walsh et al., 2008).

Genetic studies have shown that some rare structural variants confer risk for other neuropsychiatric diseases in addition to schizophrenia. For example, CNVs in *NRXN1* were associated with schizophrenia and other neurodevelopmental disorders, such as autism (Kim et al., 2008; Kirov et al., 2009b). Additionally, the 22q11.2 deletion was implicated in schizophrenia, ADHD and autism spectrum disorder (ASD) (Sahoo et al., 2011).

Furthermore, CNVs at 16p13.1 were found in both schizophrenia patients and patients with autism and mental retardation (Behjati et al., 2008; Ingason et al., 2011). These data provide further evidence for an overlap in the underlying biology of schizophrenia and other mental illnesses.

Studies using small-exome sequencing have reported that schizophrenia cases had a higher rate of putatively functional SNVs than would be expected by chance (Girard et al., 2011; Xu et al., 2011). Specifically, using targeted exome sequencing of 14 parent-case trios Girard and colleagues found 15 *de novo* mutations in eight of the probands, of which four were nonsense mutations in the genes *low density lipoprotein receptor-related protein 1* (*LRP1*); *karyopherin alpha 1* (*KPNA1*); *zinc finger protein 480* (*ZNF480*); and *ALS2 C-terminal like* (*ALS2CL*) (Girard et al., 2011). Xu and colleagues sequenced the exomes of 53 affected and 22 unaffected trios and identified 34 *de novo* point mutations, of which 19 affected evolutionarily conserved residues and were predicted by PolyPhen to be damaging (Xu et al., 2011). Both studies showed that the ratio of nonsense to missense *de novo* changes was significantly higher than expected, suggesting that the mutations were likely to cause schizophrenia (Girard et al., 2011; Xu et al., 2011). Additionally, Girard and colleagues calculated that the *de novo* mutation rate in the probands was significantly higher than the normal *de novo* mutation rate (1000 Genomes Project Consortium, 2010; Girard et al., 2011). In a recent, larger exome sequencing study, Need and colleagues considered 166 affected individuals and selected more than 5000 SNVs. These were then evaluated in an independent cohort of 2617 cases and 1800 controls, which found that no single variant met the cut-off for genome-wide significant association with disease (Need et al., 2012). However, even larger exome sequencing studies are forthcoming and these should help pinpoint definitive associations (Sullivan et al., 2012).

1.6.3.3. Genome-wide association studies: common disease-common variant

Genetic association studies compare the allele or genotype frequencies between two groups of individuals, usually cases and unaffected controls. For schizophrenia, the cases are defined by the discrete DSM-IV and ICD-10 diagnostic categories. The increasing size of association studies means that large birth cohort control datasets are often used. The association of an allele with disease denotes either that the allele is directly associated with disease, and has potential functional relevance to disease pathology; or that the allele is indirectly associated with disease through its non-random association with another genomic variant (Corvin et al., 2010; Wang et al., 2005). The non-random association between alleles at two or more loci is known as linkage disequilibrium (LD) (Slatkin, 2008). The basis of LD is the phenomenon of recombination, which is the exchange of DNA segments between chromosomes during meiosis. Recombination is less likely to occur between genetic loci which are physically close together (<50kb) therefore, these loci have a higher probability of being inherited alongside one another. Most polymorphic SNPs tend to be in strong LD with one another, therefore a significant proportion of common variation can be determined using LD mapping of a subset of the SNPs that capture most of the allelic variation in a region. Thus, the phenomenon of LD can be used in association studies to reduce costs and increase the feasibility of a study.

The association signal may be better localised using high-density LD mapping of the region, known as fine-mapping (Corvin et al., 2010). Yet, to date, fine-mapping studies have not identified new common alleles with larger effect sizes than their tagging SNPs (Park et al., 2010). The power of an association study to detect a susceptibility variant is influenced by the

frequency of the disease causing allele, the effect size and the sample size (Hirschhorn and Daly, 2005; Zondervan and Cardon, 2004).

Traditionally, association studies searched for common variants associated with disease in known candidate genes. Instead of a candidate gene approach, it was proposed that susceptibility variants could be identified using indirect genotyping of thousands of common SNPs (Collins et al., 1997; Risch and Merikangas, 1996). Subsequently, a genome-wide haplotype map (HapMap) that catalogued the common genetic variation throughout the human genome was created (Gibbs et al., 2003). The most recent HapMap lists the genotypes of over three million SNPs in 270 individuals from four populations (West European, West African, Han Chinese and Japanese) (Frazer et al., 2007). The freely available HapMap (available at http://hapmap.ncbi.nlm.nih.gov/) enabled the generation of SNP assays for genome-wide association studies (GWASs). The term GWAS refers to association using a dense array of genetic markers that capture a large proportion of genetic variation. GWAS data are analysed using logistic regression, with the dependent variable case-control status and a SNP genotype as an independent variable. The output of a logistic regression is the identity of the reference allele and an odds ratio with its standard error along with a statistic and a P value that tests whether the odds ratio differs from one (Corvin et al., 2010). In GWAS, statistical tests are done for every SNP (approximately 1×10^{-6} logistic regression models). To correct for multiple testing, a SNP is considered associated with disease if its P value surpasses the 'genome-wide significance' threshold of $P < 7.2 \times 10^{-8}$. This is equivalent to a Bonferroni correction of the traditional 0.05 Type I error level for 1×10^{-6} statistical tests (Dudbridge and Gusnanto, 2008). In traditional candidate gene studies, the genes were selected based on knowledge of pathophysiology or gene location; as such, these studies were 'hypothesis-driven'. By contrast, GWASs are hypothesis-free as they evaluate variation

throughout the genome. This is an important consideration for a disorder such as schizophrenia, where pathogenesis is highly complex and largely unknown (Corvin et al., 2010).

To date, there have been 15 published GWASs of schizophrenia (summarised in Table 1.3) and four meta-analyses of GWAS data. The early GWASs for schizophrenia used DNA pooling (Kirov et al., 2009c; Mah et al., 2006; Shifman et al., 2008). DNA pooling entails estimating the allele frequency from aggregated cases and aggregated controls rather than individual DNA samples (Sham et al., 2002). These studies identified common variants associated with schizophrenia in genes including *RELN* (Shifman et al., 2008) and *retinol binding protein 1* (*RBP1*) (Kirov et al., 2009c). *RELN* was only identified as associated with schizophrenia in women, suggesting that it may be a female-specific risk factor for the disorder (Shifman et al., 2008). These results were particularly promising as both RELN and RBP1 had been previously implicated in the aetiology of schizophrenia (see section 1.4) (Fatemi, 2005; Kalkman, 2006). However, the practice of pooling DNA for GWAS was criticised as it reduced the ability to accurately predict allele frequencies of the samples that comprised the pools, compared to if each subject was genotyped individually (Corvin et al., 2010; Schosser et al., 2010). Although individual genotyping was initially less economically viable, advances in technology led to decreasing costs and it became feasible to perform a GWAS using individual genotyping.

Several of the early GWASs using individual genotyping failed to identify associations that reached genome-wide significance, potentially due to small sample sizes (Lencz et al., 2007; Need et al., 2009) (Table 1.3). Consequently, the study performed by O'Donovan and colleagues, which was the first to identify a polymorphism associated with disease with

Reference	Population	Sample size cases/controls Discovery	Replication	Genes identified with genome-wide significance
Mah, 2006	European	320/325	n/a	none
Lencz, 2007	European	178/144	200/300	*CSF2RA*
Sullivan, 2007	European	738/733	n/a	none
O'Donovan, 2008	European	479/2937	6829/9897	*ZNF804A*
Shifman, 2008	Ashkenazi Jewish	2274/4401	n/a	*RELN* (only in women)
Shi, 2009	European/African-American	2681/2653 European; 1286/973 African-American	8008/19077	MHC
Need, 2009	European	871/863	1460/12995	none
Kirov, 2009c	European	574/605 (trios)	n/a	*RBP1*
Athanasiou, 2011	European	201/305	2663/13780	none
Djurovic, 2010	European	230/336	435/10258	none
Ikeda, 2011	Asian	575/564	n/a	none
Yamada, 2011	Asian	120 (trios)		none
Yue, 2011	Han Chinese	746/1599	4027/5603	*ZKSCAN4, NKAPL, PGBD1, TSPAN18*
Levinson, 2012	European	2461 individuals (631 pedigrees)	n/a	none
Bergen, 2012	European	1507/2093		none, but combining these samples with others that had been previously reported (2111/2535) gave GWA in *MHC* region.
Stefansson, 2009	European	2663/13 498		none, but considering top markers in combined samples gave MHC region, *NRGN*, *TCF4*
Purcell, 2009	European	3322/3587	8008/19 077	MHC region
Ripke, 2011	European	9394/12 462	8442/21 397	*MIR137*, *PCGEM1*, *TRIM26*, *CSMD1*, *MMP16*, *CNNM2*, *NT5C2*, *STT3A*, *CCDC68*, *TCF4*

Table 1.3. Summary of the results of GWASs of schizophrenia

genome-wide significance, is considered to be a landmark paper in schizophrenia genetics (O'Donovan et al., 2008). In their study O'Donovan and colleagues used a discovery sample of 479 cases and 2937 controls and a replication sample of 6829 cases and 9897 controls. They identified a variant (rs1344706, risk allele T) in intron two of *zinc finger binding protein 804A* (*ZNF804A*) that was significantly associated with schizophrenia ($P = 1.61 \times 10^{-7}$; Odds ratio (OR) 1.12) (O'Donovan, 2008). Initially, the significance of the association did not reach the genome-wide threshold; however, the significance exceeded this cut-off when the affected group was extended to include bipolar disorder cases ($P = 9.96 \times 10^{-9}$) (O'Donovan et al., 2008).

Subsequent GWASs have combined the case and control subjects from different consortia to increase sample sizes. In 2009, three GWASs that used combined samples were published (Purcell et al., 2009; Shi et al., 2009; Stefansson et al., 2009). Purcell and colleagues (2009) genotyped the International Schizophrenia Consortium (ISC) case-control sample (3322 cases and 3587 controls) and identified significant association with disease at the major histocompatibility complex (MHC) locus. When the genotyped SNPs from the Molecular Genetics of Schizophrenia (MGS) (Shi et al., 2009) and the SGENE (Stefansson et al., 2009) consortia were included in the analysis, to give a sample size of 8008 cases and 19 077 controls, the MHC region had a combined significance value which exceeded genome-wide significance.

Using only the SGENE genotyped SNPs (2663 cases and 13 498 controls), Stefansson and colleagues did not find any SNPs that exceeded the genome-wide significance for association (Stefansson et al., 2009). However, when the top markers were evaluated in a combined sample from the ISC (Purcell, 2009) and the MGS (Shi et al., 2009), several markers in the

MHC region, a marker located upstream of the *neurogranin* (*NRGN*) gene and a marker in intron four of *transcription factor 4* (*TCF4*) were significantly associated with disease (Stefansson et al., 2009).

Additionally, the Irish Schizophrenia Genomics Consortium found significant association signals in the MHC region, *TCF4* and *ZNF804A* using the combined samples from the Psychiatric GWAS Consortium (PGC) schizophrenia sample, the SGENE+ consortium sample (Stefansson et al., 2009) and the Welcome Trust Case Control Consortium 2 (WTCCC2) replication dataset (Strange, 2012 in press).

The largest published meta-analysis of GWASs to date was performed by Ripke and colleagues in 2011. Ripke and colleagues combined data from 17 studies to give 51 695 independent subjects. This analysis identified markers within five novel candidate genes as associated with schizophrenia, in addition to replicating the association of markers within the MHC region and *TCF4* (Ripke et al., 2011). The strongest association was at a SNP within an intron of *microRNA 137* (*MIR137*) ($P = 1.6 \times 10^{-11}$), a gene which has been implicated in neurodevelopment (Ripke et al., 2011). Consistent with previous findings, the statistical significance of the associations increased when the case group was extended to include patients with bipolar disorder (O'Donovan et al., 2008; Ripke et al., 2011; Williams et al., 2011).

To summarise, GWASs have identified a number of highly likely susceptibility genes for schizophrenia including *ZNF804A*, *TCF4*, the MHC region and *MIR137*. However, the percentage of the genetic basis of schizophrenia which is unaccounted for remains large. Recent data suggests that as the size of GWAS samples increases, so does the confidence

with which a GWAS can identify genes associated with disease. However, it is conceivable that it will never be possible to identify all of the common variants which contribute to schizophrenia susceptibility using the methods currently available (Bergen and Petryshen, 2012; Ripke et al., 2011).

1.6.3.4. Risk-conferring molecular pathways

The large number of alleles which confer risk for schizophrenia are expected to cluster in molecular pathways which are relevant to disease pathology (Mowry and Gratten, 2013). The current evidence suggests genes which function at the synapse may be relevant to disease pathology (Figure 1.3). For example, pathway analysis of SNP data from the ISC and Genetic Association Information Network studies showed the cell adhesion molecule pathway was significantly associated with both schizophrenia and bipolar disorder (O'Dushlaine et al., 2011). This pathway is involved in synaptic formation. Additionally, enrichment analyses of CNV data implicate synaptic genes in the pathogenesis of schizophrenia (Glessner et al., 2010; Kirov et al., 2012) (see section 1.6.3.2).

1.7. ZNF804A

ZNF804A (NM_194250; OMIM: 612282) is located on the cytogenetic band 2q32.1 of chromosome two. *ZNF804A* encodes a mRNA transcript of 4.7 kb comprising of four coding exons extending from 595 bp to 4224 bp (Figure 1.4). Recently, a novel transcriptional variant of *ZNF804A* was identified in human lymphoblastoid cell lines and post mortem brain samples (Okada et al., 2012). This variant contained an additional exon in intron two and was predicted to generate an immature 88 amino acid protein (Okada et al., 2012). There are two known paralogs of *ZNF804A*: *G patch domain containing 8* (*GPATCH8*) and *zinc finger domain containing 804B* (*ZNF804B*). At the time of the GWAS discovery, *ZNF804A* was a

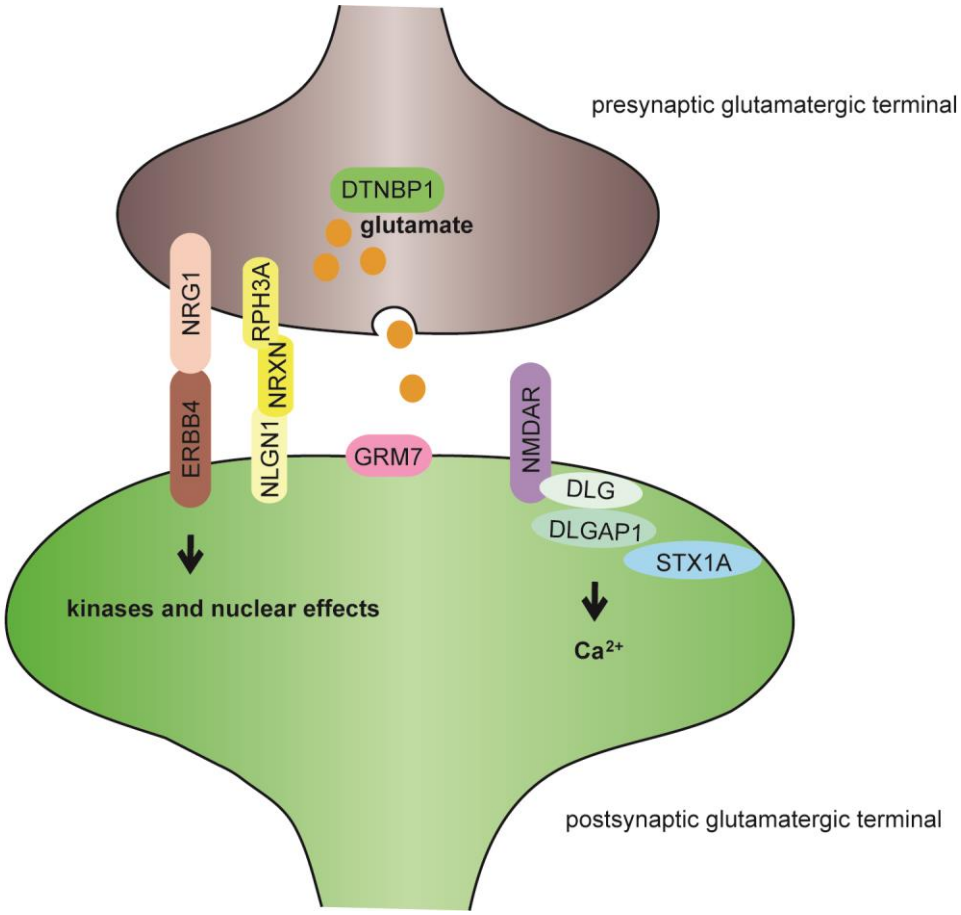


Figure 1.3. Genes implicated in schizophrenia converging at the synapse

The schematic shows the relationship between schizophrenia susceptibility genes which are involved in glutamate synapse function. Linkage studies implicate a role for *neuregulin 1* (*NRG1*) in schizophrenia while deletions of the NRG1 receptor, *v-erb-a erythroblastic leukemia viral oncogene homolog 4 (avian)* (*ERBB4*) have been found in schizophrenia. NRG1-ERBB4 signalling regulates synapse morphology and function. Linkage studies also implicate a role for *dystrobrevin binding protein 1* (*DTNBP1*) in disease. Data shows DTNBP1 regulates glutamate release. Genetic studies of schizophrenia have identified *de novo* CNVs in genes encoding components of the postsynaptic density, including *discs, large homolog 1* and *2* (*Drosophila*) (*DLG1* and *DLG2*), *discs, large (Drosophila) homolog-associated protein 1* (*DLGAP1*) and *syntaxin 1A* (*brain*) (*STX1A*). DLG1, DLG2 and DLGAP1 associate the *N*-methyl-$_D$-aspartate receptor (NMDAR) with downstream pathways which modulate calcium (Ca^{2+}) release. CNVs have also been identified in *glutamate receptor, metabotropic 7* (*GRM7*), *rabphilin 3a homolog* (*mouse*) (*RPH3A*) and neurexin 1 (*NRXN1*). NRXN1 is a synaptic cell adhesion molecule that, with neuroligin-1 (NLGN1), connects pre- and postsynaptic terminals. For references, see text. (Adapted from Harrison, 2005; Kirov, 2012.)

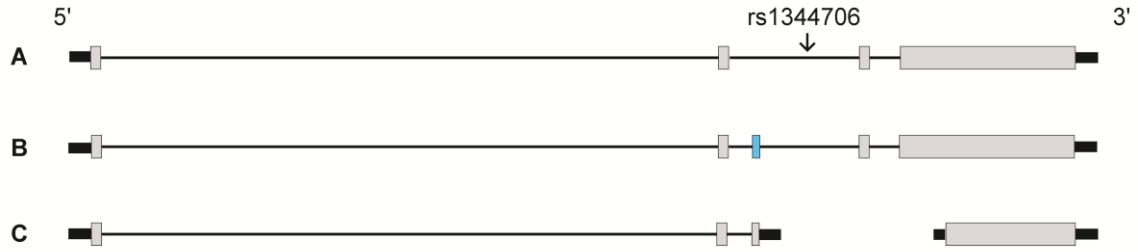


Figure 1.4. *ZNF804A* gene structure

This schematic depicts the gene structure of the known *ZNF804A* transcript variants (not to scale). The location of the disease-associated SNP (rs1344706) in intron two is also indicated. **(A)** The *ZNF804A* NCBI RefSeq (NM_194250) suggests that the mRNA transcript comprises of four coding exons. The four exons are represented by grey boxes, while the intronic sequences and the untranslated regions are represented by black lines and black rectangles respectively. This mRNA transcript encodes a predicted protein of 1209 amino acids. **(B)** Recently, Okada and colleagues (2012) identified a novel *ZNF804A* transcript variant which included an additional exon termed 2.2. Exon 2.2 is represented by a blue box. This mRNA transcript encodes a predicted immature protein of 88 amino acids. **(C)** Two alternative *ZNF804A* mRNAs are described in NCBI's AceView: bAug10-unspliced and cAug10. These mRNA transcripts encode predicted proteins of 602 amino acids and 54 amino acids respectively.

gene of unknown function. As such, subsequent studies have endeavoured to establish how *ZNF804A* may contribute to susceptibility for disease using genetics, functional molecular biology, brain imaging, neuropsychology testing and pharmacology. The findings of these studies are discussed here.

1.7.1. Genetic studies of *ZNF804A*

The genetic association between *ZNF804A* and schizophrenia has been independently replicated (Purcell et al., 2009; Riley et al., 2009; Steinberg et al., 2011; Strange, 2012). Consistent with these data, meta-analysis in 60 000 subjects supported an association between rs1344706 and schizophrenia ($P = 2.5 \times 10^{-11}$), and a combined schizophrenia and bipolar disorder affected group ($P = 4.1 \times 910^{-13}$) (Williams et al., 2011). Using fine mapping, Williams and colleagues showed that rs1344706 was most likely to be the disease causing allele (Williams et al., 2011).

It is possible that rare variation in *ZNF804A* may also increase susceptibility for schizophrenia. While Dwyer and colleagues (2010) showed that no rare, non-synonymous SNVs in *ZNF804A* were associated with schizophrenia, Steinberg and colleagues (2011) identified two CNVs spanning at least part of *ZNF804A* in psychosis patients. Specifically, Steinberg and colleagues identified a deletion of *ZNF804A* in an individual with schizophrenia and an individual with anxiety disorder, and a duplication of *ZNF804A* in an individual with bipolar disorder. These data were contrary to evidence from other CNV studies that did not find CNVs in *ZNF804A* associated with schizophrenia (Kirov et al., 2009a; Walsh et al., 2008; Xu et al., 2008). Interestingly, the current evidence suggests that rare variation in *ZNF804A* may also contribute to risk for other neuropsychiatric diseases. Notably, inspection of autism CNV datasets showed that complete or partial duplications of

ZNF804A were associated with autism (Griswold et al., 2012; Steinberg et al., 2011). Furthermore, Talkowski and colleagues (2012) identified a pathogenic chromosomal rearrangement of *ZNF804A* in a neurodevelopment disorder case that was inherited from a similarly affected parent. Taken together, these data highlight the importance of *ZNF804A* in healthy neurodevelopment, and are consistent with the hypothesis that *ZNF804A*'s contribution to disease extends beyond schizophrenia (O'Donovan et al., 2008).

1.7.2. Functional studies of ZNF804A

The *ZNF804A* transcript was predicted to generate a protein of 1209 amino acids. Bioinformatics sequence analysis predicted ZNF804A has a C2H2 type zinc finger (ZnF) domain at the N-terminus of the protein (O'Donovan et al., 2008). The C2H2 type of ZnF domains can bind to DNA, RNA and proteins, and are prevalent in transcription factors (Matthews and Sunde, 2002). Therefore, it was speculated that ZNF804A may be a regulator of gene expression (O'Donovan et al., 2008). Further bioinformatics analysis of the ZNF804A amino acid sequence revealed two putative nuclear localisation signals (D. Blake, personal communication). The nuclear localisation of transiently expressed and endogenous ZNF804A has been confirmed experimentally (Girgenti et al., 2012).

There have been very few published functional studies of ZNF804A. Notably, using microarrays, Hill and colleagues (2012) showed that, after knockdown of *ZNF804A* to 60% of wild type levels in human neural progenitor cells, there was differential expression of genes enriched for the gene ontology (GO) process cell adhesion. The authors proposed these findings may suggest a role for ZNF804A in processes such as neural migration, neurite outgrowth and synapse formation (Hill et al., 2012a). Girgenti and colleagues (2012) investigated the effects of over-expressing *ZNF804A* on a subset of genes which had been

previously implicated in schizophrenia. The results showed that transient over-expression of *myc-ZNF804A* altered the expression of *dopamine receptor D2* (*DRD2*), *phosphodiesterase 4B, cAMP-specific* (*PDE4B*), *catechol-O-methyltransferase* (*COMT*) and *protease, serine, 16 (thymus)* (*PRSS16*) between three- and five-fold (Girgenti et al., 2012). Subsequent chromatin immunoprecipitation (ChIP) showed ZNF804A bound directly to the promoter/enhancer regions of *PRSS16* and *COMT* and up-regulated their transcription (Girgenti et al., 2012). Recently, Umeda-Yano and colleagues (2013) showed that over-expression of ZNF804A in human embryonic kidney (HEK) cells led to differential expression of genes associated with TGFβ signalling. In summary, the current literature is consistent with initial speculation that ZNF804A may be a regulator of gene expression.

1.7.3. The effects of the SNP rs1344706 on *ZNF804A*

The disease-associated SNP rs1344706 is located in intron two of *ZNF804A* (O'Donovan et al., 2008). The literature suggests that intronic regions of DNA are important for the regulation of gene expression; for example, *trans*-acting regulatory proteins may bind to intronic regions (Maston et al., 2006). Using electromobility shift assays, Hill and colleagues demonstrated that the risk allele of rs1344706 resulted in reduced binding of unidentified nuclear protein in neural cells (Hill and Bray, 2011). This result may suggest that the SNP rs1344706 may disrupt binding of *trans*-acting regulatory proteins and consequently influence *ZNF804A* expression (Hill and Bray, 2011). Subsequently, Hill and Bray (2012) investigated the effects of SNP rs1344706 on the *cis* regulation of *ZNF804A*. The results showed that the disease-associated allele was associated with a relative decrease in *ZNF804A* expression during the second trimester of foetal brain development with no effect on the regulation of *ZNF804A* in the three adult brain regions examined (Hill and Bray, 2012).

Consequently, the authors proposed that the disease-associated SNP rs1344706 may contribute to disease risk in early brain development.

1.7.4. The effects of the SNP rs1344706 on brain structure and function

Initial reports suggested that the disease-associated SNP rs1344706 was associated with differences in brain structure in healthy controls (Lencz et al., 2010; Voineskos et al., 2011), and that rs1344706 affected brain volumes in schizophrenia patients, but not controls (Donohoe et al., 2011). However, these studies used small sample sizes (<110) and therefore, they were superseded by a recent study which used 892 healthy controls to examine the effects of the risk allele on brain structure (Cousijn et al., 2012). These data showed there was no effect of the disease-associated SNP on total brain volume, grey or white matter volume, or the volume of specific brain structures (Cousijn et al., 2012). While these data suggest that the genetic variation in *ZNF804A* is more likely to increase risk for disease through changes in brain function than structure, this conclusion would be more definitive if a case-control sample had been used.

Functional magnetic resonance imaging (fMRI) studies showed that the rs1344706 risk allele was associated with altered connectivity in the brains of healthy controls (Esslinger et al., 2011; Esslinger, 2009; Rasetti et al., 2011; Walter et al., 2011). To investigate the hypothesis that *ZNF804A* may contribute to disease through alterations in connectivity, researchers used tasks that probe the core neuropsychological deficits of schizophrenia, such as cognition. The results showed that the rs1344706 risk allele was associated with less impaired cognitive performance in patients but not healthy controls (Chen et al., 2012; Esslinger et al., 2009; Walters et al., 2010). This suggests that the risk allele may be associated with a subtype of patients for which deficits in cognition are a less significant feature of the disease (Walters et

al., 2010). Walters (2010) argued that, because deficits in cognition are less severe in patients with bipolar disorder than those with schizophrenia, these findings may be indicative of *ZNF804A*'s overlapping role in schizophrenia and bipolar disorder. Further neuropsychological studies showed association between SNP rs1344706 and impairments in the theory of mind network, suggesting that *ZNF804A* may contribute to disease through impairment of social cognition (Walter et al., 2011). However, this result was not replicated in a subsequent study (Hargreaves et al., 2012).

1.7.5. The effects of the SNP rs1344706 on neuropharmacology

The influence of the risk allele of *ZNF804A* on neuropharmacology was assessed by genotyping schizophrenia patients for the SNP rs1344706, treating the patients with atypical anti-psychotics and measuring the response of their positive and negative symptoms (Mossner et al., 2012). The data showed that the risk allele was associated with less improvement in the positive symptoms than the protective allele. This finding was consistent with the association between the risk allele and psychotic symptoms (O'Donovan et al., 2008; Williams et al., 2011). These data suggest either that *ZNF804A* may directly perturb the action of anti-psychotics, or that the rs1344706 risk allele may denote a distinct subtype of schizophrenia that is more resistant to treatment with atypical anti-psychotics (Mossner et al., 2012; Xiao et al., 2011).

1.8. The aims of this thesis

At the outset of this study, the biological function of ZNF804A had not been experimentally demonstrated. Therefore, nothing was known about the cellular processes ZNF804A might be involved in, and how this may contribute to susceptibility for schizophrenia. As such, the association of *ZNF804A* with disease represented a unique opportunity to understand more

about the biological aetiology of schizophrenia. The primary aim of this study was to gain a greater understanding of the biological function of ZNF804A. Firstly, the putative protein binding partners of ZNF804A were identified using yeast two-hybrid (Y2H) screening (Chapter Three). These data suggest a potential role for ZNF804A in transcriptional regulation and pre-mRNA splicing. Consequently, in Chapters Four and Five, human exon arrays were used to investigate the effects of knocking down and over-expressing *ZNF804A* on genome-wide gene expression and pre-mRNA splicing. The novel findings presented in this thesis are summarised and discussed in Chapter Six.

Chapter 2: Materials and Methods

2.1. Reagents and kits

All chemicals used for experiments were of analytical grade and were purchased from Fisher, Sigma or Invitrogen unless otherwise stated. All molecular biology reagents were purchased from New England Biolabs or Promega. All kits were used as per manufacturer's instructions.

2.2. Maintenance of cell lines and media

2.2.1. Bacterial cell culture

Propagation of DNA expression plasmids was performed using *Escherichia coli* (*E. coli*) XL1-Blue cells (Stratagene), genotype: *recA1 endA1 gyrA96 thi-1 hsdR17 supE44 relA1 lac[F' proAB lac1^{q}ZΔM15* Tn10(Tet^{r})*]*^{c}. Bacterial cells were grown in Luria-Bertani (LB) media (liquid culture) and on LB-agar media (plate culture) (Tryptone (pancreatic digest of casein) 10 g/l, yeast extract 5 g/l, NaCl 5 g/l and agar 15 g/l). Media was autoclaved and cooled to below 50°C prior to addition of the appropriate antibiotic. For bacterial cell culture, the final concentrations of antibiotics used were: ampicillin 100 µg/ml (LB-amp), kanamycin 30 µg/ml (LB-kan), zeocin 50 µg/ml (LB-zeo) and tetracycline 100 µg/ml (LB-tet). In liquid culture, *E. coli* strains were grown for 16-20h at 37°C in LB with shaking at 200 rpm for all recombinant DNA methods. In solid culture, *E. coli* strains were grown on LB plates for 16-20h and stored at 4°C for up to one month. Before picking colonies, the LB plates were removed from storage one hour before use, inverted and allowed to dry. For long term storage of *E. coli* strains, fresh cultures in mid-log phase (optical density measured at 600 nm (OD_{600}) = 0.6-0.8) were frozen in 15% (v/v) glycerol and stored at -80°C.

2.2.2. Yeast cell culture

The L40 strain of *Saccharomyces cerevisiae* (*S. cerevisiae*) (Invitrogen) was used for all yeast work. The known genotype of L40 is: *MAT*a *his3-Δ200 trp1-901 leu2-3,112 ade2 lys2-801am LYS2*::(lexAop)4-*HIS3 URA3*::(lexAop)8-lacZ *GAL4*. L40 were grown in yeast-peptone-adenine-glucose (YPAD) media. To prepare YPAD for liquid culture, 1% (w/v) yeast extract, 2% (w/v) Bacto-Peptone and 0.01% (w/v) adenine media was autoclaved. To prepare YPAD for plate culture, Bacto-agar (Invitrogen) was added to YPAD media to a final concentration of 2% (w/v). Media was autoclaved and cooled to below 50°C prior to addition of 20% (w/v) glucose to a final concentration of 2% (v/v). For maintenance of bait strains, zeocin was added to YPAD at a final concentration of 300 µg/ml following autoclaving (YPADZ300).

For Y2H screening, yeast complete (YC) media (minimal defined media for yeast) was prepared: 0.67% (w/v) yeast nitrogen base (Qbiogene), 0.0652% (w/v) complete supplemental mixture (CSM) minus listed amino acids (Qbiogene) as indicated; 2% (v/v) glucose was added after autoclaving. YC-agar plates were made by adding 2% (w/v) Bacto-agar (Invitrogen) to YC media prior to autoclaving. For selection of the bait plasmid, zeocin was added to autoclaved YC media at a final concentration of 300 µg/ml (YCZ300). L40 were grown at 30°C with shaking at 200rpm or incubated for 2-4 days at 30°C on solid growth media. Plates were stored at 4°C for up to one month. For long term storage of strains, fresh cultures in stationary phase were supplemented with 15% (v/v) glycerol and stored at -80°C.

2.2.3. Mammalian cell culture

COS-7 (African green monkey kidney), HEK293T (human embryonic kidney 293T) and SH-SY5Y (neuroblastoma) cell lines were purchased from the European Collection of Cell

Cultures (ECACC) and the American Type Culture Collection (ATCC). The 9E10 (anti-myc) hybridoma cell line was obtained from the Developmental Studies Hybridoma Bank (DSHB) at the University of Iowa. The Flp-In TREx -293 cell line was purchased from Invitrogen. The Flp-In TREx -293 cell line is a derivative of HEK293 fibroblasts (obtained from ATCC) which has been genetically modified to contain a Flp recognition target (FRT) site.

Cells were cultured in complete media: Dulbecco's modified eagle media (DMEM, Invitrogen) supplemented with 10% (v/v) foetal calf serum (FCS; PAA laboratories), 1% (v/v) penicillin/streptomycin at 37°C, 5% CO_2 in a humidified incubator. For the Flp-In TREx -293 cells, the cell culture media was supplemented with zeocin (100 µg/ml) and blasticidin (15 µg/ml) according to the manufacturer's instructions. Cells were cultured in T175 cm^2 flasks and passaged twice weekly by first washing with 10 ml Ca^{2+} and Mg^{2+} free Hank's balanced salt solution (Sigma Aldrich) and then trypsinising with 5 ml of 1 X trypsin-ethylenediaminetetraacetic acid (TE). When required, cells were counted using a haemocytometer and seeded at the appropriate density. For long term storage, the cells were resuspended in freezing medium: 90% complete medium and 10% dimethyl sulfoxide (DMSO), and frozen in automated controlled-rate freezing apparatus before being transferred to liquid nitrogen or kept at -80°C for long term storage.

2.2.4. Growth and maintenance of hybridomas

Hybridomas were grown in DMEM supplemented with 10% (v/v) FCS and 1% (v/v) penicillin/streptomycin. Frozen cells were established in a 75 cm^2 tissue culture flask with 20% (v/v) FCS, and then transferred to a 175 cm^2 flask containing 10% (v/v) FCS after recovery. Cells were passaged as described above. When the cells were confluent, the

supernatant was removed and centrifuged for 5 min at 4000 rpm. The supernatant was stored at -20°C until the antibody was purified.

2.3. Molecular biology

2.3.1. Polymerase chain reaction

For polymerase chain reaction (PCR), oligonucleotide primers were designed complementary to 18-24 base pairs of the target sequence with an estimated melting temperature of 60°C. PCR primers are listed in Appendix 1. When required, restriction endonuclease sites were added to the 5' end of the primer sequence. To increase the efficiency of restriction enzyme cleavage, three base pairs were included at the 5' end of the restriction enzyme recognition sequence. PCR was routinely performed using Red Taq (Sigma). When proof-reading was required, enzymes such as PfuUltra or Easy A (Stratagene) were used. PCRs were performed in a PTC-220 DNA Engine DyadTM Peltier Thermal Cycler (MJ Research Inc.) as outlined in Table 2.1. PCR products were separated by gel electrophoresis using 0.8-2% agarose gels as described in section 2.3.3.

Name	Volume (µl)	Final concentration	Program
10X PCR buffer	5	1 x	1. 95°C for 5 min
Forward primer (10 µM)	1	0.2 µM	2. 95°C for 30 sec
Reverse primer (10 µM)	1	0.2 µM	3. 56-60°C for 30 sec
Deoxynucleotide mix (10 mM)	1	200 µM	4. 72°C for 1 min/kb
cDNA (10 ng/µl)	5	50 ng	5. Go to step 2; 20-40
DNA polymerase (5 U/µl)	0.5	0.05 U/µl	6. 72°C for 7 min
Sterile distilled water	up to 50		7. 4°C for ∞

Table 2.1 Typical PCR conditions

2.3.2. Screening colonies by PCR

To confirm the presence of an insert by PCR, bacterial or yeast colonies were individually inoculated into 50 µl of the appropriate media using sterile needles (Fisher). An aliquot of 5

µl of this mixture was used as a template for PCR using oligonucleotide primers designed complementary to the insert and Red Taq DNA polymerase. The PCR programme used 30 cycles. The resulting PCR products were electrophoresed on an agarose gel to confirm the presence of a band of the expected size. For colonies which contained the insert, the remaining 45 µl was used to inoculate 5 ml LB containing the appropriate antibiotic and this was grown overnight (37°C, 200 rpm). Plasmid DNA was prepared from this culture as described in section 2.4.1.

2.3.3. Agarose gel electrophoresis

Agarose gels were typically made as 1% (w/v) agarose (Invitrogen) in 1 X TAE buffer (40 mM tris-acetic acid, 10 mM EDTA pH 8.0). However, 0.5-2% (w/v) agarose gels were used depending on the size of the expected fragments. The agarose was dissolved in TAE buffer by heating in a microwave and the solution was cooled to 50°C before addition of ethidium bromide (EtBr) solution to a final concentration of 0.5 µg/ml. Where appropriate, DNA samples were prepared for electrophoresis using 6 X gel loading buffer (30% (v/v) glycerol, 20 mM EDTA pH8.0, 0.25% (w/v) bromophenol blue, 0.25% (w/v) xylene cyanol in sterile distilled water) before loading into wells. A 1 kb Plus DNA Ladder (Invitrogen) was used as a size standard. The gel was submerged in 1 X TAE and run at 100V for 30 min or until an appropriate resolution was achieved. EtBr stained nucleic acid was visualised using a UV transilluminator system (Bio-Rad Gel Doc 2000).

2.3.4. Preparation of RNA

Total RNA was prepared from mammalian cells using the RNeasy RNA extraction kit (Qiagen). Cells were lysed directly using RLT buffer and homogenised using a syringe and needle (23 gauge). For routine use, the quality of RNA was determined by measuring

absorbance at 260 and 280 nm. For exon array, the quality of the RNA was determined by Central Biotechnology Services (CBS) at Cardiff University, using the Agilent 2100 bioanalyser. For PCR and quantitative-PCR (Q-PCR), the samples received DNase treatment to remove residual DNA: 2.5-5 µg of RNA was treated with DNase using the Ambicon Turbo DNase free kit according to manufacturer's instructions. The RNA was stored at -80°C.

2.3.5. First strand cDNA synthesis

To generate 1st strand cDNA, 2.5-5 µg RNA was reverse transcribed using Protoscript M-MuLV First strand cDNA Synthesis Kit (NEB) according to the manufacturer's instructions. Oligo-dT priming [$d(T)_{23}VN$] was used as it ensures that all cDNA copies terminate at the 3' end of the mRNA and produces the longest contiguous cDNA. A control reaction without reverse transcriptase (–RT) was prepared to examine the levels of residual genomic DNA (gDNA). Typically, the –RT sample was prepared using pooled RNA. A non-template RT control (5 µl of sterile distilled water instead of RNA in the RT reaction) was performed to examine the potential contamination of the reagents in the kit. The cDNA products were diluted with sterile distilled water (1:5 when the cDNA was destined for PCR and 1:20 when the cDNA was destined for Q-PCR) and were stored at -20°C.

2.3.6. Quantitative-PCR

Gene expression was quantitated by Q-PCR using either Sensimix no rox SYBR-Green (Bioline) or Solaris primer/probes (Thermo Scientific) and a Qiagen Rotor-gene 3000 real time PCR machine (Qiagen) following the conditions outlined in Table 2.2. When SYBR-Green was used, the melt curve analysis was performed from 55°C to 95°C. The threshold cycle (C_t) values, the fractional cycle number at which fluorescence passes the threshold, were calculated by the Qiagen Rotor-Gene 3000 software. The Q-PCR mastermix was

Sensimix		
Name	**Volume (µl)**	**Program**
SYBR-Green PCR master mix	12.5	1. 95°C for 10 min
Forward primer (100 µM)	0.75	2. 95°C for 15 sec
Reverse primer (100 µM)	0.75	3. 55-60°C for 15 sec
Sterile distilled water	6	4. 72°C for 15 sec
cDNA (1.5 ng/µl)	5	5. Go to step 2; 40 times
Solaris		
Name	**Volume (µl)**	**Program**
Solaris master mix (2X)	12.5	1. 95°C for 15 min
Solaris primer/probe set (20X)	1.25	2. 95°C for 15 sec
Sterile distilled water	10.25	3. 60°C for 1 min
cDNA (1.5 ng/µl)	1	4. Go to step 2; 40 times

Table 2.2 Typical Q-PCR conditions

prepared using a PCR setup pipetting robot (CAS-1200, Corbett Life Science, Qiagen). The samples were run in triplicate. A –RT control, a non-template RT control (5 µl of sterile distilled water instead of RNA in the RT reaction) and a non-template Q-PCR control (5 µl of sterile distilled water instead of cDNA) were included in each Q-PCR run.

The details of the Q-PCR primer sequences used in this thesis are given in Appendices 1.6 and 1.7. Both the transcript- and exon-level target validations used intron-flanking primers wherever possible to eliminate gDNA amplification. For accurate and reproducible Q-PCR measurements, it was important that the primer sets used had high amplification efficiency; a PCR assay has 100% efficiency if the PCR amplicon doubles in quantity during the geometric phase of the PCR amplification. To assess the efficiency of the primer sets, Q-PCR assays were performed in duplicate using, wherever possible, a four log span of input cDNA. Where the transcript abundance was too low to use a four log span, the dilutions were adjusted accordingly. The amplification efficiency was calculated from a standard curve of the C_t value against log input of cDNA where efficiency = $(10^{-1/slope} -1) \times 100$. Only primers with amplification efficiency between 90% and 110% were used in experiments.

For data analysis, a relative quantification method known as the comparative C_t method was applied because the aim was to compare transcript-level and exon-level abundance among different groups. When using the comparative C_t method, it is important to apply a normalisation strategy. Housekeeping genes are commonly used as an internal control in order to normalise the Q-PCR data for the amount of RNA that is added to a reaction. In this thesis, beta actin (*ACTB*) was selected as the reference gene (also known as the 'control'), because the exon array analyses showed that the mRNA level of *ACTB* was not altered after *ZNF804A* knockdown or over-expression (data not shown). For transcript-level targets, the

target mRNA was quantified relative to *ACTB*. For the exon-level targets, the transcript variant of interest was quantified relative to an mRNA common to all transcript variants. The comparative C_t method was as follows: for each condition the three technical replicate C_t values were averaged using the mean. This average C_t value was subtracted from the average C_t value of the control to produce the ΔC_t value. The next step was to calculate the $\Delta\Delta C_t$ value by subtracting the ΔCt values of the experimental group from the ΔC_t values of the un-treated/control group. The negative $\Delta\Delta C_t$ value was then used as the exponent of two (based on the assumption that the reaction doubles the amount of product per cycle – see below). Finally, the $2^{-\Delta\Delta Ct}$ values were multiplied by 100 to give the abundance as a percentage relative to the control sample and represented as a graph. The error bars represent the limits of the standard deviation (sd) of the triplicate C_t values prior to normalisation. The upper limit was calculated as $(2^{-(\Delta\Delta Ct - sd)} - 2^{-\Delta\Delta Ct})$ and lower limit was calculated as $(2^{-\Delta\Delta Ct} - (2^{-(\Delta\Delta Ct + sd)})$.

It was important that the PCR amplification efficiencies of both the target and the control primer set were relatively equivalent. To assess this, Q-PCR assays were performed and the ΔC_t values were calculated and plotted against log input cDNA. An R^2 less than 0.1 signified that the two primer sets had relatively equivalent amplification efficiencies. The primer amplification efficiencies are given in Appendices 1.6 and 1.7.

2.3.7. Mammalian cell transfection and drug treatment

Cells were seeded in 6-well plates at a density of 1×10^5 (COS-7) or 5×10^5 (HEK293T) cells per well. Transfections were carried out using Fugene-6 (Roche) as per manufacturer's instructions. For proteasome inhibition, 24h after transfection the cells were treated with lactacystin (10 μM) (Sigma) for 17h.

2.3.8. Gene silencing using small interfering RNA

SH-SY5Y cells were transfected with small interfering RNA (siRNA) duplexes using Lipofectamine RNAi MAX (Life Technologies) according to manufacturer's instructions. Briefly, SH-SY5Y cells were seeded in 12-well plates (5 x 10^4 cells per well) in DMEM supplemented with 5% (v/v) FCS. Three wells were used per condition. The following day, cells were transfected with *ZNF804A*-specific siRNA (Dharmacon) or *GAPDH*-specific siRNA at a final concentration of 50 nM (Table 2.3). The growth media was changed 24h post-transfection and the following day the transfection protocol was repeated. Forty-eight hours later, the cells from the three wells were pooled. RNA was extracted as described in section 2.3.4 and frozen at -80°C.

Duplex	Sense strand 5' -
siZNFA	GGAAAAUACCAUAGCAAAAUU
siZNFB	CCAGGAAAGAUGAAAGAAAUU
siGAPDH	GUCAACGGAUUUGGUCGUAUU

Table 2.3 The siRNA duplexes used in this study

2.4. Cloning

2.4.1. Purification of plasmid DNA

DNA was isolated from small scale cultures (5 ml) or large scale cultures (>100 mls) of *E. coli*. Plasmid DNA was purified using the Qiaprep Spin Miniprep or plasmid Maxiprep kit (Qiagen) according to the manufacturer's instructions. DNA was quantified using a spectrophotometer (BioPhotometer, Eppendorf) at 260/280 nm (A260).

2.4.2. Restriction digest of DNA

All restriction enzymes were purchased from New England Biolabs. For preparation of vectors for cloning, 1 µg of purified vector was cut with the appropriate restriction enzymes

(approximately 20 units (U), where 1 U of restriction enzyme will completely digest 1 µg of substrate DNA in a 50 µl in 1h) in a total volume of 50 µl for 2h at 37°C followed by addition of 10 U calf intestinal alkaline phosphatase (CIP) for 1h at 37°C. The restriction digest products were separated by agarose gel electrophoresis as described in section 2.3.3 and the fragment was excised and purified using the Qiagen Gel extraction kit. For digestion of purified PCR products, the entire purified product was mixed with the appropriate 10X buffer and restriction enzyme (approximately 20 U) for 1-2h at 37°C. The digest was purified using the Qiagen PCR purification kit. For digestion of purified DNA, for example to confirm the presence of an insert, 1 µl of DNA was digested with the appropriate restriction enzyme (approximately 10 U) and 10X buffer in a total volume of 20 µl for 1h at 37°C. The restriction digest products were separated by agarose gel electrophoresis.

2.4.3. Ligation of DNA fragments into vectors

Ligation of digested DNA into vectors was performed using T4 DNA ligase (Promega). A typical reaction used a 3:1 molar ratio of insert to vector. In a 10 µl reaction 1 µl T4 DNA ligase was used. Reactions were centrifuged briefly before incubation overnight at 4°C. Control reactions containing insert only or vector only were prepared to confirm complete digestion of the insert and dephosphorylation of the vector. Typically, 5 µl of the ligation reaction was used to transform *E. coli* (section 2.4.6).

2.4.4. Preparation of chemically competent *E. coli* XL1-Blue cells

Competent cells were prepared from *E. coli* XL1-Blue. An *E. coli* XL1-Blue glycerol stock was streaked onto LB-tet plates and grown overnight at 37°C. A single colony was inoculated into 10 ml LB-tet media and grown overnight (37°C, 200 rpm). 5 ml of the overnight culture was used to inoculate 500 ml LB-tet, which was grown at 37°C with agitation (200 rpm) until

the cells reached mid-log phase (OD_{600} approximately 0.5). The culture was cooled on ice for 2h and the bacteria were pelleted by centrifugation at 4000 rpm for 20 min at 4°C. The pellet was gently resuspended in 250 ml fresh, ice-cold, filter sterilised salt buffer (100 mM calcium chloride ($CaCl_2$), 70 mM manganese chloride ($MnCl_2$), 40 mM sodium acetate (NaOAc) pH 5.5) and incubated on ice for 45 min. After incubation, the bacteria were pelleted by centrifugation at 4000 rpm for 10 min at 4°C. The pellet was resuspended in a total of 50 ml salt buffer and 11.5 ml ice-cold filter sterilised 80% (v/v) glycerol was added drop-wise with gentle agitation, to give a final concentration of 15% (v/v). Single aliquots of 500 µl were stored at -80°C.

2.4.5. Preparation of electrocompetent *E. coli* XL1-Blue cells

An *E. coli* XL1-Blue glycerol stock was streaked onto LB-tet plates and grown overnight at 37°C. A single colony was picked, inoculated into 5 ml LB-tet broth and grown overnight (37°C, 200 rpm). The 5 ml of overnight culture was used to inoculate 500 ml LB-tet and this was grown at 37°C with agitation (200 rpm) until the cells reached mid-log phase (OD_{600} approximately 0.5). The culture was cooled on ice for 30 min, transferred to ten pre-cooled 50 ml Falcon tubes and centrifuged (2000g, 4°C, 15 min). The pellets were resuspended in a total of 500 ml cold sterile distilled water and re-centrifuged. This was repeated in a total of 250 ml cold sterile distilled water. The pellets were washed and pooled in a total of 10 ml sterile, cold, 10% glycerol, re-centrifuged and resuspended in a total of 1 ml sterile, cold 10% glycerol. Single aliquots were stored at -80°C.

2.4.6. Transformation of *E. coli* XL1-Blue using heat-shock method

Cells were removed from the -80°C freezer and thawed on ice. 50 µl of cells were added to pre-chilled 2059 polypropylene tubes. 5 µl of ligation reaction (section 2.4.3) or 10 ng diluted

plasmid was added to the cells and incubated on ice for 30 min. Cells were heat-shocked at 42°C for 1 min using a water bath and chilled on ice for 2 min. 1 ml LB media was added and the cells were incubated at 37°C for 1h to enable the expression of antibiotic resistance genes. Cells were transferred to a microcentrifuge tubes and centrifuged in a desktop microfuge (13 000 rpm, 1 min). The pellet was resuspended in 100 µl fresh LB, plated onto LB plates containing the appropriate selective antibiotic and incubated overnight at 37°C.

2.4.7. Transformation of *E. coli* XL1-Blue using electroporation

Cells were removed from the -80°C freezer and thawed on ice. 5 µl of ligation reaction (section 2.4.3) or Y2H 'prey' plasmids extracted from yeast (section 2.5.4), were added to cells and the sample was incubated on ice for 1 min. The sample was added to a pre-chilled 2mm electroporation cuvette and electroporated with a 5 ms pulse of 2.5 kV, 600 Ω, 10 µF using the MicroPulser electroporation apparatus (Bio-Rad). Cells were recovered in 1 ml LB and incubated at 37°C for 1h to enable the expression of antibiotic resistance genes. Cells were transferred to a microcentrifuge tube and centrifuged in a desktop microfuge (13 000 rpm, 1 min). The pellet was resuspended in 100 µl fresh LB, plated onto LB plates containing the appropriate selective antibiotic and incubated overnight at 37°C.

2.4.8. DNA sequencing

Purified templates were sequenced by SourceBioscience Life Sciences Oxford DNA sequencing services. All sequencing was performed using ABI BigDye Terminators V3.1 (Applied Biosystems). The sequencing chromatogram traces were checked using Chromas Lite, version 2.01 (Technelysium Pty Ltd).

2.5. Yeast two-hybrid screening

The version of the yeast two-hybrid (Y2H) system used in this study is based on the Hybrid Hunter interaction trap (Invitrogen).

2.5.1. Generating the Y2H bait strain

The bait plasmids (N-term-ZNF804A-pHybLex/ZeoA and C-term-ZNF804A-pHybLex/ZeoA) and the bait strains used in this study were obtained from C.L. Tinsley. The bait strains were generated by transforming the bait plasmids into L40 using the lithium acetate (LiAc) method as follows: to generate a starter culture 10 ml of YPAD was inoculated with a colony of L40 and shaken overnight at 30°C. The starter culture was diluted to an OD_{600} of 0.4 in 50 ml of YPAD and grown for an additional 2h. The cells were centrifuged at 2500 rpm and the pellet was resuspended in 40 ml sterile distilled water. The cells were centrifuged at 2500 rpm, resuspended in 2 ml of 1X LiAc/0.5X TE and incubated at rt for 10 min. For each transformation, 1 µg plasmid DNA and 100 µg denatured salmon sperm DNA was mixed with 100 µl of the yeast suspension. Subsequently, 700 µl of 1X LiAc/40% polyethylene glycol (PEG) -3350/1X TE was added and the sample was mixed well. The solution was incubated at 30°C for 30 min and then 88 µl DMSO was added. To heat shock, the sample was incubated at 42°C for 7 min. To allow expression of antibiotic resistance by the plasmid, the sample was then incubated at 30°C for 1h. The mixture was centrifuged in a table top microcentrifuge (13 000 rpm, 10s) and the supernatant was removed. The cell pellet was resuspended in sterile distilled water and re-centrifuged. The transformation mix was resuspended in 100 µl sterile distilled water and plated on YPADZ300 plates. The plates were incubated at 30°C for 3-4 days.

2.5.2. Testing for self-activation of the bait plasmid

To confirm that the bait plasmid did not self-activate, transformants were selected from the YPADZ300 plates and transferred to 50 µl sterile distilled water in microcentrifuge tubes. These colonies were then streaked onto the YPADZ300 plates containing 3-aminotriazole (3-AT) at the following concentrations: 0; 5; 10; 15 and 20 mM. Bait strains that grew on plates containing 20 mM 3-AT were deemed unsuitable for use in the Y2H experiments. In subsequent Y2H screens for baits that could be used, YPADZ300 plates were supplemented with the lowest 3-AT concentration at which no growth was observed after five days.

2.5.3. Small scale transformation of yeast using the lithium acetate method

The bait strain was co-transformed with the prey mouse brain cDNA library (Invitrogen) in the pPC86 vector or the human foetal brain cDNA library (Invitrogen) in the pDEST™22 vector according to the transformation of yeast (TRAFO) protocol (Agatep et al., 1998; Gietz and Woods, 2002). A 5 X TRAFO scale screen was used. A starter culture was generated by inoculating a yeast bait strain colony into two aliquots of 5 ml YPAD media and incubating at 30°C with 200 rpm overnight. The OD_{600} was measured using a spectrophotometer and the cell titre was determined using the formula OD_{600} of 1 = 1 $x10^7$ cells. A volume of cells that yielded 1.25 x 10^8 cells was used. The culture was centrifuged at 3000 x g for 5 min and resuspended in 25 ml YPADZ300. This was incubated at 30°C with shaking at 200 rpm until the cells had reached mid-log phase (OD_{600} of 2). The cells were harvested at 3000 x g for 5 min and the pellet washed with 12.5 ml sterile distilled water and centrifuged again. The pellet was resuspended in 1.5 ml sterile 100 mM LiAc and incubated in a water bath at 30°C for 15 min. The cells were harvested as above and the components of the transformation mix added to the pellet in the following order: 34% (v/v) PEG3000, 100 mM LiAc, 150 µg boiled salmon sperm DNA, 20 µg prey plasmid in sterile distilled water. The pellet was resuspended

by vigorous vortexing, incubated for 30 min at 30°C and the tube mixed gently several times after 10 and 20 min to mix the contents thoroughly. The suspension was then heat shocked at 42°C for 50 min and mixed by inversion for 15 sec at 5 min intervals. The cells were harvested as previously and the cell pellet was gently resuspended in 2 ml of sterile distilled water. 400 µl aliquots were plated onto five 150 mm plates containing MM-Trp-Lys-His-Ura, Z300 (MM-TrpZ300) and 3-AT. These were incubated at 30°C for 4-7 days. Colonies that grew on the histidine deficient medium were re-streaked onto fresh selective plates. The transformation efficiency was determined by plating dilutions of the transformed yeast onto MM-TrpZ300 plates and was used to estimate the number of clones screened (Table 2.4). In error, the transformation efficiency of the N-terminal screen using the human foetal brain cDNA library was not measured.

cDNA library	Bait	Number of clones screened
Human foetal brain	C	8×10^5
Mouse brain	N	3×10^6
Mouse brain	C	1.5×10^6

Table 2.4 The transformation efficiency

2.5.4. Extraction of the prey yeast DNA

The yeast containing prey plasmids were isolated by growing in MM-TrpZ300 media. Plasmid DNA from interacting clones was purified using the RPM Yeast Plasmid Isolation Kit (Q-Biogene). The resulting DNA was amplified by PCR using pDEST™22 forward and reverse primers (see Appendix 1.5 for primer sequences) and the resulting products analysed by agarose gel electrophoresis as described in section 2.3.3. When a clear DNA band was obtained, the PCR product was PCR purified and sent for direct sequencing using the pPC86 vector forward sequencing primer (section 2.4.8). When a clear DNA band was not observed by agarose gel electrophoresis, 5 µl of the yeast DNA was transformed into electrocompetent

E. coli XL1-Blue using electroporation (section 2.4.7). Three colonies were picked for each clone, grown overnight and the DNA isolated (section 2.4.1). The resulting DNA was digested with the restriction enzyme BSRG1 to determine the size of the insert. The 5' end of the cDNA was sequenced using pPC86 vector forward sequencing primer and the identity of the insert was determined using the National Centre Biotechnology Information (NCBI) basic local alignment search tool (BLAST).

2.6. Protein analysis

2.6.1. Sample preparation

For the preparation of protein samples from transfected cell lines, 5 x 10^5 HEK293T cells, myc-ZNF804A Flp-In TREx or GFP-TCF4 Flp-In TREx cells were seeded and after 24h the cells were transfected or induced as appropriate. The following day, protein lysates were prepared. Cells were washed twice with phosphate buffered saline (PBS) and lysed in 250 µl 2X sample buffer (0.125 M Tris pH 6.8, 4% sodium dodecyl sulphate (SDS), 20% (v/v) glycerol, 5% (v/v) β-mercaptoethanol, 0.001 (w/v) bromophenol blue). Cells were scraped, placed in a microcentrifuge tube and sonicated with a Vibra-Cell ultrasonic processor (Sonics) using 10 sec pulses with a 15 sec rest at 50W for two cycles. Samples were stored at -20°C and 15 µl was used for analysis by SDS- polyacrylamide gel electrophoresis (SDS-PAGE). Samples were boiled for 5 min at 95°C and cooled on ice before use.

2.6.2. SDS-PAGE

Protein samples were separated on 10% (v/v) acrylamide gels by SDS-PAGE, under denaturing and reducing conditions using the Mini Protean III Gel System. The resolving gel contained 380 mM Tris-hydrogen chloride (Tris-HCl), pH 8.8, 10% (v/v) acrylamide (30% (w/v) acrylamide; 0.8% (w/v) bis-acrylamide (37.5:1)), 0.1% (w/v) SDS, 0.1% (w/v)

ammonium persulphate (APS) and 0.08% N,N,N,N-Tetramethylethylenediamine (TEMED). The stacking gel consisted of 125 mM Tris-HCl pH 6.8, 5% (v/v) acrylamide (30% (w/v) acrylamide: 0.8% (w/v) bis-acrylamide (37.5:1)), 0.1% (w/v) SDS, 0.1% (w/v) APS and 0.1% TEMED. The resolving gel was layered with isopropanol (IPA) during the polymerisation process to ensure that the resolving gel was packed tightly and had an even interface with the stacking gel. Following polymerisation, the IPA was discarded and the top of the gel was gently washed with sterile distilled water. The stacking gel was then layered onto the resolving gel and a lane-forming comb was inserted. When the stacking gel had polymerised, the gels were immobilised in the clamp system, submerged in SDS-PAGE running buffer (25 mM Tris-base, 192 mM glycine, 1% (w/v) SDS) and samples were loaded. A pre-stained, broad range molecular weight marker (7-175 kDa, NEB) was loaded as a reference on each gel to allow confirmation of band size after migration. Gels were electrophoresed at 150V for 70 min.

2.6.3. Western blotting

Following electrophoresis, proteins from the SDS-PAGE gels were transferred to nitrocellulose membrane using the Mini Trans-blot electrophoretic transfer cell. Gels were submerged in transfer buffer (25 mM Tris-base, 192 mM glycine, 1% (w/v) SDS, 20% (v/v) methanol) and proteins were transferred for 1h at 70V for a single gel or at 80V for two gels. The membrane was blocked either for 1h at room temperature or overnight at 4°C in blocking solution (5% (w/v) dried skimmed milk powder in Tris buffered saline-Tween (TBST) (150 mM NaCl, 50 mM Tris-HCl pH 7.5, 1% (v/v) Tween-20)). After blocking, the membrane was incubated with the primary antibody diluted to the appropriate concentration in blocking solution at room temperature for 1h or at 4°C overnight. The primary antibodies used in this thesis are listed in Table 2.5. The membrane was washed twice in TBST for 5 min with

shaking and then washed once in blocking solution for 5 min with shaking. The membrane was incubated with the appropriate secondary antibody diluted in 10 ml blocking solution at room temperature for 1h or at 4°C overnight, in both instances the membrane was protected from the light. The secondary antibodies used in this thesis are listed in Table 2.5. The membrane was washed three times with TBST for 5 min with shaking, before visualisation using the Odyssey Infrared Imaging System (LI-COR Biosciences). The images were assembled using Adobe Photoshop CS2 and Illustrator, and were cropped and adjusted for brightness and contrast but otherwise not manipulated.

2.6.4. Purification of monoclonal antibody from hybridomas

To purify the antibody, a 1 ml protein-G column was packed in an Econo-Pac® disposable chromatography column, equilibrated with PBS and a frit was inserted. The supernatant (200 ml) was poured over the column repeatedly for 1h. The purified antibody was eluted with ImmunoPure® IgG Elution buffer in 1 ml fractions and neutralised with 50 µl/ml Tris pH 9.0. The column was washed with PBS and stored at 4°C in PBS with 0.025% (v/v) sodium azide.

2.7. Proteomics

2.7.1. Crosslinking of the anti-myc antibody to protein A agarose

Anti-myc antibody-conjugated protein A beads were prepared by coupling the anti-myc antibody (9E10) to protein A agarose (Invitrogen). Protein A agarose was prepared in a 15 ml falcon tube by washing 1 ml of 50% protein A agarose slurry with PBS by centrifugation at 2000 rpm for 5 min. 2 mg of affinity-purified antibody was diluted 1:1 with PBS to a final volume of 4 ml and incubated with the protein A agarose for 2h at room temperature with rotation. The beads were centrifuged at 2000 rpm for 5 min and were washed twice with 5 ml

Antibody	Specificity	Dilution for western blot	Dilution for cell culture immunofluorescence	Source
Primary antibodies				
Rabbit polyclonal				
3077	human ZNF804A	1:50	–	in-house, described Appendix 3.1
001	mouse Zfp804a	1:50	–	in-house, described Appendix 3.2
002	mouse Zfp804a	1:50	–	in-house, described Appendix 3.2
D-14	human ZNF804A	1:400	–	Santa Cruz
S-16	rat and mouse ZNF804A	1:400	–	Santa Cruz
P-13	rat and mouse ZNF804A	1:400	–	Santa Cruz
Mouse monoclonal				
9E10	c-myc	1:400	1:200	DSHB
12G10	α-tubulin	1:10000	–	DSHB
FK2	mono- and poly- ubiquitin conjugates	1:1000	–	Biomol
GFP	green fluorescent protein	1:1000	–	Covance
SC35	SC-35 (nuclear speckle marker)		1:2000	Abcam
Secondary antibodies				
Alexa Fluor 488	mouse/rabbit IgG	–	1:1000	Invitrogen
Alexa Fluor 568	mouse/rabbit IgG	–	1:1000	Invitrogen
Alexa Fluor 680	rabbit IgG	1:10000	–	Invitrogen
IRDye 800	mouse IgG	1:10000	–	Rockland Immunocytochemicals

Table 2.5 The antibodies used in this thesis

0.1 M borate buffer (0.2 M di-sodium tetraborate, 0.2 M boric acid, pH 9.0). The antibody was cross-linked to the protein A beads by adding 20 mM dimethyl pimelimidate (DMP) in 9.5 ml borate buffer and incubating with rotation at room temperature for 30 min. The beads were centrifuged at 2000 rpm for 5 min and the reaction was stopped by washing the beads in 5 ml 0.1 M ethanolamine, pH 8.0. Unreacted DMP was quenched by incubating the beads with 5 ml 0.1 M ethanolamine at room temperature for 1h. The beads were then centrifuged at 2000 rpm for 5 min, washed in PBS and centrifuged again. To remove any uncoupled IgG the beads were washed with 5 ml of ImmunoPure® IgG Elution Buffer (Pierce). The beads were then washed twice with PBS and resuspended as a 50% slurry in PBS with 0.05% sodium azide. The beads were stored at 4°C.

2.7.2. Immunoprecipitation

For immunoprecipitation, cells were seeded in 10 cm dishes at 2 x 10^6 cells per dish. The following day, the cells were either transfected with 1-6 µg DNA or induced to express the gene of interest with 1 µg/ml tetracycline. 24h later, the cells were washed in 5 ml PBS and lysed in RIPA buffer (150 mM NaCl, 50 mM Tris, pH 8.0, 1% (v/v) Triton X-100, 0.5% (v/v) sodium deoxycholate (DoC), 1 mM ethylene glycol tetraacetic acid (EGTA) in sterile distilled water) and homogenised using a polytron. The samples were incubated on ice for 30 min and were then centrifuged at 25 000 rpm for 35 min at 4°C in a SW41Ti rotor (Beckman) to remove insoluble protein. To limit the non-specific binding of the sample with the protein A agarose beads, the supernatant was pre-cleared with 50 µl packed protein A agarose beads (equilibrated in RIPA buffer) at 4°C for 2h with rotation. The protein A agarose beads were pelleted by centrifugation at 1000 rpm for 5 min at 4°C. The supernatant was then divided into two samples and incubated with either 50 µl packed 9E10-conjugated protein A beads (IP sample) or 50 µl packed protein A agarose beads (control sample) overnight at 4°C with

rotation. The beads were pelleted by centrifugation at 1000 rpm for 5 min at 4°C and the supernatant was discarded. The beads were extensively washed with RIPA buffer and the protein was eluted from the beads using 100 µl 1 X NuPAGE lithium dodecyl sulphate (LDS) (Invitrogen) sample buffer containing 30 mM dithiothretol (DTT) at room temperature for 5 min. The samples were boiled at 95°C to denature the immune complexes and cooled on ice before centrifugation at 13 000 rpm for 1 min to pellet the beads. The presence of myc-tagged protein in the supernatant was analysed by western blot.

2.8. Cell biology

2.8.1. Immunocytochemistry

COS-7 cells were seeded on sterile glass coverslips (22 mm x 22 mm) in a 6-well plate format at 1 x 10^5 cells per well. 24h post-seeding the cells were transfected with a DNA construct expressing the gene of interest. The following day, the cells were washed twice with 2 ml PBS and fixed with either 1 ml 4% (w/v) paraformaldehyde (PFA) in PBS at 4°C for 15 min or with 1 ml ice cold methanol at -20°C for 10 min. The PFA was removed and the cells permeabilised with 2 ml PBS and 0.1% (v/v) Triton X-100 at 4°C for 15 min. Following permeabilisation, the cells were washed three times with PBS for 5 min with shaking. The cells were blocked with 1 ml 10% (v/v) FCS for 20 min at room temperature and incubated with the primary antibody in 1 ml PBS at room temperature with shaking for 1h. The cells underwent two 5 min washes with PBS and were incubated with the secondary antibody in 1 ml PBS at room temperature with shaking for 1h. The antibodies used in this study are presented in Table 2.5. Following a further two PBS washes, the nuclei were stained with Hoechst (1 µg/ml) at room temperature for 15 min. Following a further two PBS washes, the coverslips were mounted onto glass slides with Aqua Poly/Mount (Polyscience, Inc.). Cells were visualised using a Leica DMRA2 epifluorescent microscope or with a Zeiss

LSM 510META confocal microscope. The images were captured using the confocal microscope with a 63x oil immersion objective lens and a pinhole setting of 1 airy unit. The images were assembled using ImageJ v1.421, Adobe Photoshop CS2 and Illustrator, in which images were cropped and adjusted for brightness and contrast but otherwise not manipulated.

2.8.2. The tetracycline-inducible, stable Flp-In TREx expression system

The tetracycline-controlled gene expression system was first described by Gossen and Bujard (1992) to facilitate the controlled and reversible expression of genes to a defined level. Subsequently, Yao and colleagues developed the tetracycline-inducible system in which expression of the target gene is under control of a pCMV promoter which is repressed by two tandem tetracycline operator sequences ($TetO_2$) (Yao et al., 1998). This means that the promoter is only active when tetracycline is present in the cell culture media. Thereafter, the tetracycline-inducible, stable Flp-In TREx expression system was developed by Invitrogen. This system incorporates a yeast DNA recombination system which uses the Flp recombinase enzyme and site-specific recombination to facilitate integration of a desired gene at a FRT site at a single point in the genome (Figure 2.1). In this study, the Flp-In TREx -293 cell line was used as the 'host cell line'. The Flp-In TREx -293 cell line has been genetically modified to contain the FRT site necessary for site-specific recombination and insertion of the gene of interest.

The target gene was cloned into a pcDNA5/FRT/TO expression vector in frame with the pCMV/$TetO_2$ promoter according to the manufacturer's instructions (Figure 2.1A). The primer sequences used are given in Appendices 1.1 and 1.2. The pcDNA5/FRT/TO expression vector and the pOG44 vector, which encodes the Flp recombinase enzyme, were co-transfected into the Flp-In TREx -293 cell line at a ratio of 9:1, according to the

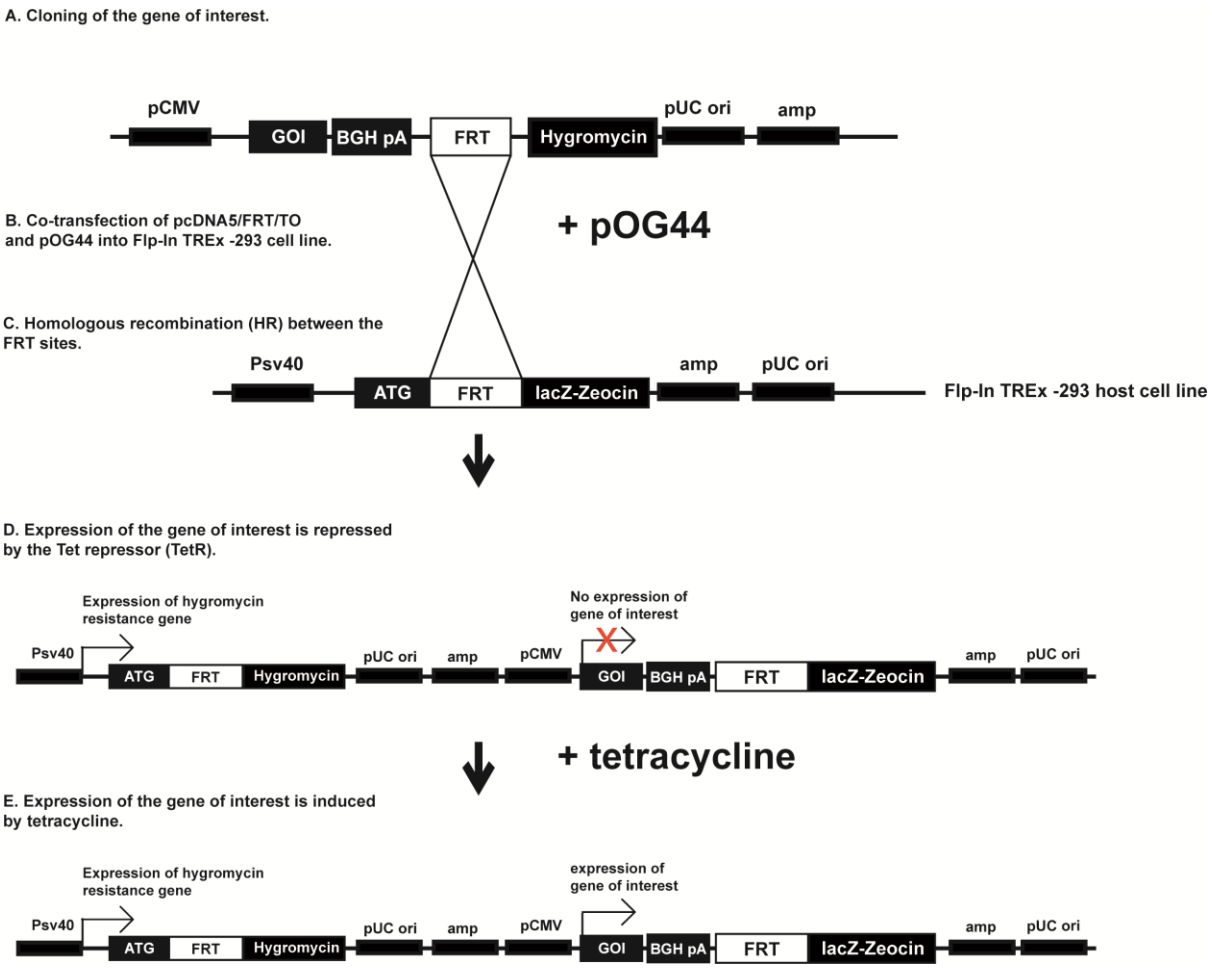


Figure 2.1 Schematic overview of the Flp-In TREx system

Flp-In TREx expression cell line generation. **(A)** The gene of interest was cloned into the pcDNA5/FRT/TO expression vector in-frame with a pCMV/$TetO_2$ promoter. **(B)** The pcDNA5/FRT/TO expression vector was co-transfected into the Flp-In TREx -293 host cell line with the pOG44 vector which expressed the Flp recombinase enzyme. **(C)** The Flp recombinase enzyme mediated a homologous recombination event between the Flp recognition target (FRT) sites on the pcDNA5/FRT/TO vector and in the Flp-In TREx -293 cell line. The result was incorporation of the gene of interest into the Flp-In TREx -293 cell genome and gene expression under the control of a tetracycline inducible promoter. **(D)** 150 µg/ml hygromycin B was used to select for cells which had undergone successful recombination. **(E)** Treatment with 1 µg/ml tetracycline induced gene expression.

manufacturer's instructions (Figure 2.1B). To generate a control cell line which expressed the hygromycin B and blasticidin resistance genes, an empty pcDNA5/FRT expression vector and the pOG44 vector were transfected into the Flp-In TREx -293 cell line. Following co-transfection of the appropriate expression construct and the pOG44 plasmid, the successful stable transformants were selected in complete media containing hygromycin B (150 µg/ml) and blasticidin (15 µg/ml). After 10 days, single hygromycin B-resistant, blasticidin-resistant foci were isolated using cloning rings and expanded to generate individual clonal cell lines. The individual clonal cell lines were maintained in media containing hygromycin B and blasticidin as recommended by the manufacturer. Where appropriate, target gene expression was induced using 1 µg/ml of tetracycline. For long term storage the cells (> 3 x 10^6 per cyrovial) were frozen in freezing medium: 90% complete medium and 10% DMSO in an automated controlled-rate freezing apparatus and stored at -80°C.

2.9. Exon array

2.9.1. Preparation of RNA for exon array

To prepare RNA for use in the exon array experiments assessing the effect of knocking down *ZNF804A*, RNAi was performed on SH-SY5Y cells as described in section 2.3.7 and the RNA was extracted as described in section 2.3.4. A mock transfection sample using only the RNAi transfection reagents was also prepared. To prepare RNA for use in the exon array experiments assessing the effects of *ZNF804A* over-expression, myc-ZNF804A and control Flp-In TREX expression cells were seeded in 6 cm dishes at 1 x 10^6 cells per dish. The following day, the cells were treated with 1 µg/ml of tetracycline and 24h post induction, the RNA was extracted as described in section 2.3.4 and frozen at -80°C.

The exon arrays were performed by CBS, Cardiff University. Prior to performing the exon arrays, CBS assessed the quality and quantity of the RNA using the Agilent 2100 bioanalyser. First strand cDNA was prepared using the Ambicon WT Expression Kit (Applied Biosystems), fragmented and terminally labelled using the GeneChip WT Terminal Labelling and Hybridization Kit (Affymetrix) following methods suggested by the manufacturer. The cDNA samples were hybridised to the Affymetrix GeneChip human exon 1.0 ST Array plate at 45°C for 17h, as described in the Affymetrix User's Manual. Following hybridisation, the chips were washed and stained using the GeneChip fluidics Station 450/250 according to the GeneChip Expression Wash, Stain and Scan User's Manual. The chips were scanned using the Affymetrix GeneChip Scanner 3000.

2.9.2. Importing the exon array data into the Partek Genomics Suite

All exon array data was analysed using tools in the Partek Genomics Suite (PGS) software (v6.5; Partek Inc.). The .CEL files were imported into the PGS according to the manufacturer's guidelines outlined in the 'Importing exon data into Partek Genomics Suite' tutorial (available from www.partek.com). The .CEL files were labelled and grouped according to whether they represented a treated or control sample. Table 2.6 outlines the library files used for import, these library files are available from the Affymetrix website (www.affymetrix.com).

Table 2.6 The library files used to analyse the exon arrays

Information contained within the file	Version
Core metaprobe set list	HuEx-1_0-st-v2.r2.dt1.hg18.core.mps
Probe set annotation	HuEx-1_0-st-v2.na32.hg19.probeset.csv
Transcript annotation	HuEx-1_0-st-v2.na32.hg19.transcript.csv
Defines which probes are in each probe set	HuEx-1_0-st-v2.r2.pgf
Layout of array	HuEx-1_0-st-v2.r2.cdf
Quality control content	HuEx-1_0-st-v2-r2.qcc

Briefly, the .CEL files were imported using the interrogating and control probes. Probe set reliability is ranked from more to less reliable as 'core', 'extended' or 'full'. For the analysis presented in this thesis the 'core' metaprobe set list was used (Affymetrix, 2005). This represented approximately 22 000 RefSeq genes and full-length GenBank mRNAs. The robust multiarray averaging (RMA) algorithm was used to summarise the probe-level data to a single value for each probe set (Irizarry et al., 2003a). The RMA algorithm consists of quantile normalisation followed by fitting of the data to a model of expression and probe affinities using an algorithm called median polish (Okoniewski, 2008). For each probe this algorithm corresponded to: perfect match $(PM)_{ij} = e_i + a_j + E_{ij}$, where e is the chip effect and a represents the probe affinity for the j^{th} probe on the i^{th} array (Okoniewski, 2008). Adjustments were made for probe GC content on pre-background-subtracted values. The exon-level probe sets were summarised to the gene-level using the probe set mean. The output information was set to include an additional calculation of the median and mean of the raw data.

2.9.3. Quality assessment

Quality assessment (QA) of the exon array data was performed using the QA metrics in the PGS according to Affymetrix GeneChip QA of exon and gene arrays whitepaper (v1.1) (2007).

To determine how the samples clustered, agglomerative hierarchical clustering was performed in the PGS on the gene-summarised expression values of the most variably expressed genes. Hierarchical clustering computes a dendrogram that assembles all of the elements (the samples) into a tree. In agglomerative clustering, each object forms a separate group and each step of the computation joins the groups which are closest to one another. The

intensity plots presented were standardised by applying a mean set to 0.0 and a standard deviation of 1.0.

2.9.4. Identifying differentially expressed genes

Exon arrays can provide estimates of gene expression by virtually assembling probe sets which correspond to exons into transcript clusters that correspond to genes (Kapur et al., 2007). The exon-level signal intensities were summarised to the gene-level by calculating the mean $\log_2$ signal intensity of all of the probe sets across the gene. The differentially expressed genes were detected by a one-way ANOVA model using method of moments (Eisenhart, 1947) with treatment group type as the ANOVA factor as follows:

$$Y = \mu + Z + \varepsilon$$

Where Y represents the expression of a gene, μ is the mean expression of the gene, Z is group type to group type effect and ε is the error term. The output information was P value, fold change and mean ratio.

To overcome the problem of multiple testing, a list of statistically significantly differentially expressed genes was generated using a maximum false discovery rate (FDR) of either 0.05 or 0.01. The FDR is defined as the proportion of false positives among the declared differentially expressed genes (Benjamini and Hochberg, 1995). For example, if 100 genes were declared differentially expressed with a maximum FDR 0.05, a maximum number of 5 false positive results would be expected. The FDR method is often used in microarray studies to correct for multiple testing and control the type I error rate (Pawitan et al., 2005). In this thesis, the raw ANOVA P values, rather than FDR adjusted P values (the q values), are presented.

2.9.5. Empirical validation of differentially expressed genes

The differentially expressed genes were selected for empirical validation by Q-PCR based on: 1) the magnitude of the fold change, where a large fold change was preferred; 2) the number of known transcript variants of the gene, where a single transcript variant was preferred; and 3) the DNA sequence, where sequences with 40-60% GC content which were amenable to Q-PCR were preferred. The Q-PCR primers were designed, optimised and validated for use with the $\Delta\Delta C_t$ method as described in section 2.3.6. The sequences of the Q-PCR primers used in this study are given in Appendices 1.6 and 1.7. Please note, only one biological replicate used in the exon array was analysed using Q-PCR.

2.9.6. Identifying alternative splicing events

When analysing alternative splicing data, an essential step is to normalise the exon-level signal to the corresponding gene-level signal, this ensure that any changes in transcript abundance between samples are not misrepresented as changes in splicing. To identify alternatively spliced transcripts, the alternative splicing one-way ANOVA model in the PGS was performed with treatment group type as the ANOVA factor. To avoid the lack of detection of a probe set being misconstrued as alternative splicing, only probe sets with a $\log_2$ signal intensity of greater than three were included in the analysis. The model was as follows:

$$Y = \mu + Z + E + Z^*E + \varepsilon$$

Where Y represents the expression of a gene, μ is the mean expression of the gene, Z is group type to group type effect, E is the exon-exon effect (alternative splicing independent to group type), Z*E represents an exon expressing differently in different tissues (alternative splicing dependent to tissue type) and ε is the error term.

To overcome the statistical issues associated with multiple testing, the gene-level alternative splicing one-way ANOVA P values were corrected using the conservative Bonferroni method (Whistler, 2010). The Bonferroni corrected alternative splicing P values are presented. Transcripts were considered alternatively spliced when the corresponding Bonferroni corrected P value surpassed a set threshold. The threshold P values were chosen to allow a manageable list of statistically significant alternatively spliced transcripts to be generated. It is difficult to interpret splicing patterns in transcripts with few markers therefore, all genes represented by less than five probe sets across the transcript cluster were removed (Affymetrix, 2006; Whistler, 2010). To ensure the focus was on known genes, any transcript clusters not represented by a HUGO gene nomenclature committee (HGNC) gene symbol were also removed (Whistler, 2010). Additionally, to reduce the type I error rate, any gene with high differential expression across the transcript (defined as a fold change greater than five) was also removed (Affymetrix, 2006; Whistler, 2010).

The literature suggests that probe-level estimation, rather than gene-level estimation, improves the detection of differential splicing in Affymetrix GeneChip exon array studies (Laajala et al., 2009). Therefore, to implement these findings in the PGS, I took advantage of the probe set-level results which were generated alongside the alternative splicing one-way ANOVA in the PGS; these methods have been previously used in the literature (Gillett et al., 2009). The probe set-level alternative splicing one-way ANOVA P values were filtered to select for probe sets showing significant differential expression, determined at a FDR 0.05, and no differential expression at the gene-level (transcript $P < 0.05$). The genes were sorted by the fold change of the differentially expressed probe set, and the probe sets with the largest fold changes were considered in the downstream analysis.

Following identification of statistically significant alternative splicing events, the data was subjected to detailed manual analysis to determine if the statistical analysis had identified a potential true alternative splicing event (Affymetrix, 2006; Whistler et al., 2010). This was achieved using the PGS 'geneviews'. A PGS geneview graphically plots the transcript's probe set $\log_2$ signal intensities and the corresponding fold change in these probe set $\log_2$ signal intensities between the groups alongside transcript variant information. The transcript variant information is available from the University of California Santa Cruz (UCSC) genome browser database. Throughout this thesis, the GRCh37/hg19 assembly was used.

2.9.7. Empirical validation of alternative splicing events

The alternative splicing events were selected for empirical validation if: 1) the event corresponded to a known alternative splicing event; 2) the splicing event could not be explained by changes in gene expression; 3) the difference in $\log_2$ signal intensity was present across all of the probe sets in the same exon; and 4) the alternative splicing event was not at the extreme 3' or 5' end of the sequence (Whistler, 2010). To confirm the alternative splicing event, RT-PCR was performed using primers designed to target constitutive exons flanking the exon of interest (Figure 2.2). The PCR products were resolved by agarose gel electrophoresis (section 2.3.3). The expected result was two bands, which represent exon inclusion (larger band) and exon exclusion (smaller band). Any change in the relative intensities of these two bands between sample groups was likely to be indicative of alternative splicing (Affymetrix, 2006). Please note, only one biological replicate used for the exon array was evaluated using RT-PCR. If two products were detected by RT-PCR, Q-PCR was used to quantify the amount of the spliced exon. Please note, only one biological replicate sample used in the exon array was evaluated using Q-PCR. Q-PCR primers were designed, optimised and validated for use with the $\Delta\Delta C_t$ method as described in section 2.3.6.

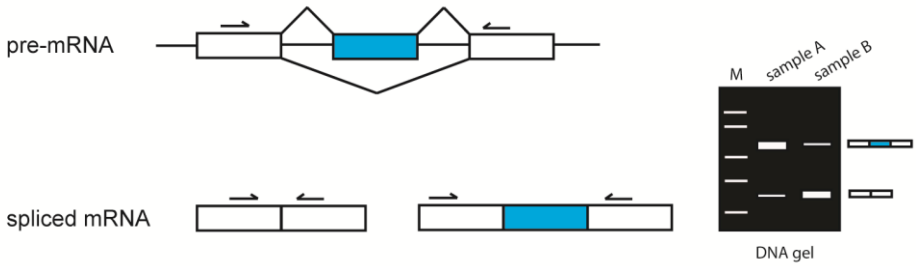


Figure 2.2 Using RT-PCR to assess alternative splicing

To confirm alternative splicing events, RT-PCR primers (arrows) were designed complementary to the consituitive exons (white boxes) flanking the spliced exon (blue box). The PCR products were separated by agarose gel electrophoresis. The bands of DNA correspond to transcript variants with exon inclusion (higher molecular weight DNA band) and exon exclusion (lower molecular weight DNA band). Changes in the relative abundance of the DNA in each band between samples denotes alternative splicing. In the hypothetical cassette exon shown, sample A shows greater inclusion of the spliced exon while sample B shows greater exclusion of the spliced exon. (Adapted from Affymetrix, 2006.) M = size marker.

The sequences of the Q-PCR primers used to validate the alternative splicing events are described in Appendix 1.7. For the exon-level targets, the transcript variant of interest was quantified relative to a mRNA common to all transcript variants. This was to ensure that quantification of alternative splicing was accurate and was not influenced by possible differences in transcript abundance between samples. An alternative splicing event was considered validated if there was a statistically significant difference in the abundance of the spliced exon present between the treated samples and the controls.

2.10. Enrichment analysis

The enrichment analysis was carried out using GeneGo MetaCore™ (Thomson Reuters). The gene lists destined for analysis were uploaded into GeneGo MetaCore™ according to the manufacturer's instructions. Where appropriate, the P value and fold change for each gene were included in the upload and the species was specified. The analysis was performed using the enrichment analysis workflow in GeneGo MetaCore™. Briefly, the enrichment analysis method maps the selected data onto GeneGo MetaCore™ ontologies in terms of their respective sets of genes or network objects. The significance is evaluated based on the size of the intersection between the selected data and the ontology using a hypergeometric distribution. The resulting P value represents the probability for a particular mapping of an experiment to a particular ontology to arise by chance, considering the numbers of genes in an experiment versus the number of genes in the process within the 'full set' of all genes in the ontologies. These mappings are sorted by their P values and are represented graphically in the form of a histogram. Where appropriate, the Affymetrix GeneChip human exon array 1.0 ST v2 array was set as the background comparison list. Unless otherwise stated, both up- and down-regulated genes were used in the enrichment analysis. All reported pathways and biological processes are listed according to their enrichment score provided by the software

package as –log(P value). All reported pathways and biological processes were statistically significant, determined at a FDR of 0.05, according to GeneGo MetaCore™. The uncorrected P values are shown. The gene ontology (GO) processes represent cellular processes as defined by GO. GeneGo process networks represent content which is manually created by GeneGo MetaCore™ on the basis of GO processes and GeneGo pathway maps.

2.11. Bioinformatics

Gene sequences were obtained from the RefSeq presented by the National Centre for Biotechnology Information (NCBI; http://www.ncbi.nlm.nih.gov/BLAST/). DNA sequence data was analysed using BLAST presented by the NCBI and the BLAST-like alignment tool (BLAT) (Kent, 2002) presented by UCSC genome browser (http://genome.ucsc.edu/) (Fujita et al., 2011; Kent et al., 2002). All UCSC analysis used the February 2009 human reference sequence GRCh37/hg19 assembly. The putative alternative splicing events were analysed using UCSC genome browser using the additional tracks: Affy Exon Array (Affymetrix), spliced expressed sequence tags (ESTs) and Alt Events. General analysis of DNA and protein sequences were performed with ClustalW (Larkin et al., 2007), SMART (Simple Modular Architecture Research Tool; http://smart.emblheidelberg.del) (Schultz et al., 1998), Kalign (http://www.ebi.ac.uk/Tools/msa/kalign/) and Boxshade 3.21 (http://www.ch.embnet.org). DNA to protein translation and calculation of predicted molecular weights was carried out using the ExPASy (Expert Protein Analysis System) proteomics server from the Swiss Institute of Bioinformatics (Gasteiger et al., 2003).

2.12. Statistical methods

The statistical analyses related to the exon array data were performed in the PGS v6.5 and are described in section 2.10. All other statistical analyses were performed using SPSS 16.0 for

Windows (LEAD Technologies Inc.). Data was analysed by independent Student's T test or one-way ANOVA. Post hoc comparisons were performed using Tukey. For these statistical comparisons, P values of ≤ 0.05 were regarded as significant.

Chapter 3: Characterising ZNF804A and identification of its protein binding partners

3.1. Introduction

The genome-wide significant association between the SNP rs1344706 in intron two of *ZNF804A* and schizophrenia was a landmark finding in psychiatric genetics (O'Donovan et al., 2008). Yet, at the time of the GWAS discovery, *ZNF804A* represented a gene within the annotation gap; that is, a gene whose function was unknown and could not be accurately inferred using bioinformatics techniques. Aside from a report that suggested that ZNF804A may be a putative binding partner for ataxin-1 (ATXN1) (Lim et al., 2006); a protein implicated in processes such as transcription and pre-mRNA processing (de Chiara et al., 2009; Irwin et al., 2005; Lam et al., 2006; Lim et al., 2008; Orr, 2010; Tsai et al., 2004; Yue et al., 2001), there was no published information regarding ZNF804A thus, very little could be inferred about the protein's function or its possible contribution to disease pathogenesis.

Several strategies can be used to determine the function of a protein. Often, because most proteins function as part of a complex with other proteins, identifying a protein's binding partners can provide useful insight into its function. If two proteins interact, it is likely that their functions are related. Protein-protein interactions can be identified using biochemical techniques, such as immunoaffinity purification; or using genetic methods, such as the yeast two-hybrid (Y2H) system (Fields and Song, 1989; Phizicky and Fields, 1995). The Y2H system is often preferred over biochemical methods because it does not require high quantities of purified proteins or good quality antibodies and it enables the simultaneous isolation of interacting proteins with their encoding genes (Van Criekinge and Beyaert,

1999). For example, Y2H studies implicate DISC1; a protein encoded by the *DISC1* gene, which has been associated with major mental illnesses, in processes such as neuronal development as a consequence of its interactions with proteins such as nuclear distribution element-like (NUDEL) (Camargo et al., 2007).

The subcellular localisation of a protein can also provide information about its biological function; this is because localisation determines the environment in which the protein operates and which proteins it is able to interact with. For example, if a protein localises to the nucleus, this may suggest that it is involved in processes specific to the nucleus, such as gene expression. To establish a protein's localisation in mammalian cells, the protein sequence is transiently expressed with an epitope tag that can be detected by immunocytochemistry and visualised using confocal microscopy. Alternatively, the epitope tag may encode a protein such as green fluorescent protein (GFP), which fluoresces when exposed to a specific wavelength of light.

The homologs of a protein of interest can also provide insight into its function. Homologs are genes which share common descent either via speciation (orthologs) or gene duplication (paralogs) (Koonin, 2005). Theories of population genetics suggest that paralogs may be preserved in the genome by so-called 'subfunctionalisation'; this implies that protein paralogs may share similar or co-operative functions (Force et al., 1999; Lynch and Force, 2000). *ZNF804A* has two known paralogs: *GPATCH8* and *ZNF804B*. At the outset of this study, the molecular function of these paralogs was unknown; investigating the function of these proteins may provide insight into the biological role of ZNF804A.

The aim of the experiments described in this Chapter was to gain initial insights into the possible functions of ZNF804A. To achieve this, Y2H studies were employed to identify the potential protein binding partners of ZNF804A. Additionally, expression vectors were used to identify the localisation of transiently expressed ZNF804A and GPATCH8 in cultured cells. Data described in this Chapter implicate a role for ZNF804A in the regulation of gene expression; particularly DNA transcription and pre-mRNA processing.

3.2. Yeast two-hybrid screening with ZNF804A bait strains

For Y2H screening, the ZNF804A cDNA sequence was divided into two non-overlapping N- (amino acid 1 to 630) and C- (amino acid 631 to 1209) terminal baits (Figure 3.1) and cloned in-frame into the bait plasmid pHybLex/Zeo (section 2.4). The bait plasmids were then transformed into *Saccharomyces cerevisiae* (*S. cerevisiae*) L40 to generate the bait strain (section 2.5.1). The bait plasmids and the bait strains were kindly provided by C.L. Tinsley. Small scale Y2H screens were performed with a human foetal brain cDNA library (section 2.5.3). The prey plasmid DNA was isolated from clones which putatively interacted with ZNF804A and the cDNA sequence was obtained by 5' sequencing using a vector primer (section 2.5.4). Protein sequence alignments were made using the standard protein BLAST program. Only amino acid sequences which were in-frame with the prey vector were considered positive results. Y2H screening using the N-terminal bait identified 13 positive clones, while Y2H screening using the C-terminal bait identified 39 positive clones (Table 3.1).

To rationalise the assortment of ZNF804A-interacting proteins isolated from the Y2H screens, the available literature for each protein was obtained from the NCBI PubMed and web of science databases using the protein's name and symbol as search terms, and manually

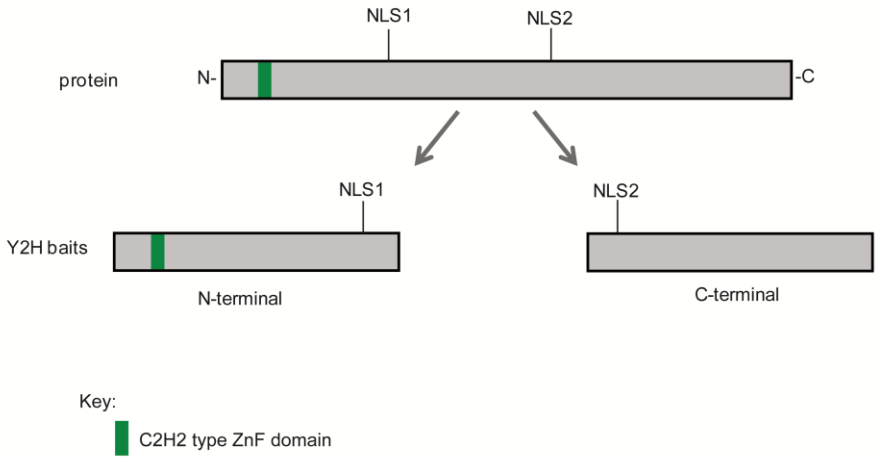


Figure 3.1 Schematic representation of the ZNF804A Y2H baits

The ZNF804A cDNA sequence was divided into two non-overlapping N- (amino acid 1 to 630) and C- (amino acid 631 to 1209) terminal baits. NLS = putative nuclear localisation signal (D. J. Blake, personal communication); ZnF = zinc finger; Y2H = yeast two-hybrid.

Bait	Clone	RefSeq	Protein name	Protein symbol	Amino acid alignment	% identity	E score
N	32a	NP_001989.2	Fibulin-2 isoform b precursor	FBLN2	731-1011	99	0.0E+00
N	28a	NP_002105.2	Zinc finger protein 40	ZNF40	2026-2222	100	7.0E-123
N	1a	NP_001138582.1	NEL-like 2 (chicken) isoform d	NELL2	373-649	99	0.0E+00
N	7a	NP_001138579.1	NEL-like 2 (chicken) isoform a	NELL2	336-596	100	0.0E+00
N	18a	NP_001138579.1	NEL-like 2 (chicken) isoform a	NELL2	404-681	100	0.0E+00
N	31	NP_001138579.1	NEL-like 2 (chicken) isoform a	NELL2	225-474	100	1.0E-176
N	6	NP_006702.1	RNA binding protein S1, serine-rich domain	RNPS1	140-245	100	2.0E-67
N	11	NP_006702.1	RNA binding protein S1, serine-rich domain	RNPS1	140-263	100	3.0E-67
N	17a	NP_006702.1	RNA binding protein S1, serine-rich domain	RNPS1	33-245	99	2.0E-81
N	24b	NP_006702.1	RNA binding protein S1, serine-rich domain	RNPS1	140-263	100	2.0E-67
N	33a	NP_006702.1	RNA binding protein S1, serine-rich domain	RNPS1	33-245	99	2.0E-81
N	34a	NP_006702.1	RNA binding protein S1, serine-rich domain	RNPS1	38-263	99	1.0E-78
C	17a	NP_001121.2	Amino-terminal enhancer of split	AES	1-121	100	2.0E-88
C	73b	NP_001395.1	Eukaryotic translation elongation factor 1 gamma	EEF1G	7-311	99	0.0E+00
C	11a	NP_002105.2	Zinc finger protein 40	ZNF40	2196-2494	99	0.0E+00
C	13a	NP_002105.2	Zinc finger protein 40	ZNF40	2197-2496	96	0.0E+00
C	14a	NP_002105.2	Zinc finger protein 40	ZNF40	2026-2309	99	0.0E+00
C	18a	NP_002105.2	Zinc finger protein 40	ZNF40	2172-2419	100	9.0E-164
C	24a	NP_002105.2	Zinc finger protein 40	ZNF40	2026-2118	88	5.0E-48
C	30a	NP_002105.2	Zinc finger protein 40	ZNF40	2196-2478	99	0.0E+00
C	52a	NP_002105.2	Zinc finger protein 40	ZNF40	2178-2477	100	0.0E+00
C	76a	NP_002105.2	Zinc finger protein 40	ZNF40	2172-2456	99	0.0E+00
C	56a	NP_002507.1	Neuro-oncological ventral antigen 2	NOVA2	3-283	100	0.0E+00
C	4a	NP_002779.1	Proteasome (proteasome, macropain) subunit, alpha type, 3	PSMA3	1-244	99	5.0E-179
C	10b	NP_002779.1	Proteasome (proteasome, macropain) subunit, alpha type, 3	PSMA3	1-255	100	0.0E+00
C	23a	NP_002779.1	Proteasome (proteasome, macropain) subunit, alpha type, 3	PSMA3	1-255	100	0.0E+00
C	28a	NP_002779.1	Proteasome (proteasome, macropain) subunit, alpha type, 3	PSMA3	1-255	99	0.0E+00
C	29a	NP_002779.1	Proteasome (proteasome, macropain) subunit, alpha type, 3	PSMA3	1-255	100	0.0E+00
C	31a	NP_002779.1	Proteasome (proteasome, macropain) subunit, alpha type, 3	PSMA3	1-255	100	2.0E-175

C	32a	NP_002779.1	Proteasome (proteasome, macropain) subunit, alpha type, 3	PSMA3	1-255	99	0.0E+00
C	46a	NP_002779.1	Proteasome (proteasome, macropain) subunit, alpha type, 3	PSMA3	1-255	99	0.0E+00
C	50a	NP_002779.1	Proteasome (proteasome, macropain) subunit, alpha type, 3	PSMA3	1-255	99	0.0E+00
C	54b	NP_002779.1	Proteasome (proteasome, macropain) subunit, alpha type, 3	PSMA3	1-240	99	4.0E-173
C	55a	NP_002779.1	Proteasome (proteasome, macropain) subunit, alpha type, 3	PSMA3	1-222	94	5.0E-149
C	90a	NP_056176.2	R3H domain containing 1	R3HDM1	238-344	99	5.0E-63
C	7a	NP_000970.1	Ribosomal protein L18	RPL18	2-188	100	2.0E-132
C	85a	NP_001019.1	Ribosomal protein S25	RPS25	8-125	100	1.0E-77
C	51a	NP_001076048.1	RNA binding protein, fox-1 homolog (C. elegans) 2	RBFOX2	61-214	97	2.0E-103
C	26a	NP_054878.5	SET domain containing 2	SETD2	2260-2539	99	1.0E-178
C	48a	NP_054878.5	SET domain containing 2	SETD2	2260-2541	99	1.0E-178
C	53b	NP_054878.5	SET domain containing 2	SETD2	2260-2517	100	9.0E-163
C	75a	NP_054878.5	SET domain containing 2	SETD2	2093-2186	100	2.0E-59
C	79b	NP_054878.5	SET domain containing 2	SETD2	2093-2369	95	0.0E+00
C	41a	NP_689699.2	Sterile alpha motif domain containing 11	SAMD11	413-653	100	1.0E-163
C	43a	NP_689699.2	Sterile alpha motif domain containing 11	SAMD11	304-572	100	0.0E+00
C	2a	NP_689699.2	Sterile alpha motif domain containing 11	SAMD11	413-681	100	0.0E+00
C	8a	NP_689699.2	Sterile alpha motif domain containing 11	SAMD11	413-621	98	3.0E-137
C	25b	NP_689699.2	Sterile alpha motif domain containing 11	SAMD11	n/a	n/a	n/a
C	33a	NP_689699.2	Sterile alpha motif domain containing 11	SAMD11	n/a	n/a	n/a
C	49a	NP_689699.2	Sterile alpha motif domain containing 11	SAMD11	n/a	n/a	n/a
C	60c	NP_055959.1	Zinc finger, CCHC domain containing 14	ZCCHC14	664-942	100	0.0E+00

Table 3.1 The putative ZNF804A-interactors identified in Y2H screens using a human foetal brain cDNA library. The prey plasmid DNA was isolated from clones which putatively interacted with ZNF804A and the cDNA sequence was obtained by 5' sequencing using a vector primer. Protein sequence alignments were made using the standard protein BLAST program. The amino acid alignment and the % identity and E score for each alignment is shown. A value of 'n/a' is given when the identity of the clone was confirmed using colony PCR. (N = N-terminal ZNF804A bait; C = C-terminal ZNF804A bait.)

reviewed. The proteins were sorted into categories based on their common known and/or predicted roles in biological processes (Table 3.2). This analysis showed that the Y2H screens identified multiple proteins implicated in the regulation of transcription or pre-mRNA processing. These data were consistent with the results of additional Y2H screening using the N- and C- terminus ZNF804A baits and a mouse brain cDNA library (Table 3.3) (mouse brain cDNA library screening performed by C.L. Tinsley). For instance, RNA binding protein, fox-1 homolog (*C. elegans*) 2 (RBFOX2) and RNA binding protein S1, serine-rich domain (RNPS1) were identified as putative ZNF804A-interactors in both of the screens. Consistent with their functions in transcription regulation or pre-mRNA processing, many of the putative ZNF804A-interactors contain protein domains with the capacity to bind nucleic acids; these include RNA recognition motifs (RRM), KH-1 domains, ZAS domains and ZnF domains (Table 3.4).

In order to identify potential enrichment of the putative ZNF804A-interactors for specific biological processes, enrichment analysis for gene ontology (GO) processes was performed using GeneGo MetaCore™ software (section 2.10). The enrichment analysis was performed against the default background list of genes in GeneGo MetaCore™ as the composition of the cDNA libraries was unknown. The results are presented as a ratio of the number of genes within the Y2H dataset annotated to a particular GO process compared to the total number of genes in that GO process, alongside a P value for the enrichment. Figure 3.2A shows genes belonging to the GO terms 'mRNA metabolic process' (5/675 $P=4.08 \times 10^{-5}$) and 'gene expression' (9/3834 $P=1.17 \times 10^{-4}$) were significantly enriched among the putative ZNF804A-interactors identified in the human foetal brain cDNA library Y2H screen. The interactors identified in the mouse brain cDNA library Y2H screens were also significantly enriched for the GO term 'mRNA metabolic process' (5/675 $P=2.69 \times 10^{-5}$), in addition to

Biological process	Protein name	Symbol	Accession	Number of clones	Bait	References
Transcription regulation	Zinc finger protein 40	ZNF40	NP_002105.2	9	N (1); C (8)	Otsuka et al., 1995
	Amino-terminal enhancer of split	AES	NP_001121.2	1	C	Tetsukati, 2000
	SET domain containing 2	SETD2	NP_054878.5	5	C	Edmunds, 2008
Pre-mRNA processing	RNA binding protein S1, serine-rich domain	RNPS1	NP_006702.1	6	N	Sakashita et al., 2004
	Neuro-oncological ventral antigen 2	NOVA2	NP_002507.1	1	C	Ule et al., 2005
	RNA binding protein, fox-1 homolog (C. elegans) 2	RBFOX2	NP_001076048.1	1	C	Yeo et al., 2009
Translation	Ribosomal protein S25	RPS25	NP_001019.1	1	C	Marion and Marion, 1988
	Ribosomal protein L18	RPL18	NP_000970.1	1	C	Tsurugi et al., 1978
	Eukaryotic translation elongation factor 1 gamma	EEF1G	NP_001395.1	1	C	Riis et al., 1990
Cell adhesion	Fibulin 2	FBLN2	NP_001989.2	1	N	Pfaff et al., 1995
	NEL-like 2 (chicken) isoform a	NELL2	NP_001138579.1	3	N	Aihara et al., 2003
Protein degradation	Proteasome (proteasome, macropain) subunit, alpha type, 3	PSMA3	NP_002779.1	11	C	Fedorova et al., 2011
Proteins of unknown function	R3H domain containing 1	R3HDM1	NP_056176.2	1	C	
	Sterile alpha motif domain containing 11	SAMD11	NP_689699.2	7	C	
	Zinc finger, CCHC domain containing 14	ZCCHC14	NP_055959.1	1	C	

Table 3.2 The identity of putative ZNF804A-interacting proteins identified in Y2H screens using a human foetal brain cDNA library

The proteins are grouped according to their common known and/or predicted roles in biological processes, according to the current literature. (N = N-terminal ZNF804A bait; C = C-terminal ZNF804A bait).

Biological process	Protein name	Symbol	Accession	Number of clones	Bait	References
Transcription regulation	Trans-acting transcription factor 1	Sp1	NP_038700.2	1	C	Li et al., 2004
	Basic helix-loop-helix family, member e40	Bhlhe40	NP_035628.1	1	C	Yamada and Miyamoto, 2005
	Glucocorticoid modulatory element binding protein 1	Gmeb1	NP_064669.2	1	N	Kaul et al., 2000
	Melanoma antigen, family D, 1	Maged1	NP_062765.1	8	C	Mouri et al., 2012
Pre-mRNA processing	RNA binding protein, fox-1 homolog (C. elegans) 1	Rbfox1	NP_899011.2	1	C	Yeo et al., 2009
	RNA binding protein, fox-1 homolog (C. elegans) 2	Rbfox2	NP_780596.1	2	C	Jin et al., 2003
	CUGBP, Elav-like family member 4	Celf4	NP_573458.2	1	C	Ladd et al., 2001
	Ribonucleic acid binding protein S1	Rnps1	NP_001073597.1	1	N	Sakashita et al., 2004
Protein degradation	Proteasome (proteasome, macropain) subunit, alpha type 3	Psma3	NP_035314.3	26	C	Fedorova et al., 2011
	Tetratricopeptide repeat domain 3	Ttc3	NP_033467.2	3	C	Suizu et al., 2009
	DnaJ (Hsp40) homolog, subfamily A, member 3	Dnaja3	NP_076135.3	3	C	Rowley et al., 1994
Proteins of unknown function	G patch domain-containing 8	Gpatch8	NP_001152964.1	1	C	

Table 3.3 The identity of putative ZNF804A-interacting proteins identified in Y2H screens using a mouse brain cDNA library, grouped according to biological process

The proteins are grouped according to their common known and/or predicted roles in biological processes, according to the current literature. (N = N-terminal ZNF804A bait; C = C-terminal ZNF804A bait.) (The mouse brain cDNA library screen was performed by C. L. Tinsley.)

Domain name	Domain function	ZNF804A-interactors containing this domain
RNA recognition motif (RRM)	binds RNA	RNPS1; RBFOX2; Rbfox1; Celf4
K homology RNA-binding domain, type 1(KH-1)	binds ssRNA or DNA	NOVA2
K homology RNA-binding domain, PCBP_like (PCBP_like_KH)	binds ssRNA or DNA	NOVA2
zinc finger (ZnF) domain	binds RNA, DNA and protein	Gpatch8; Sp1; ZCCHC14
ZAS domain	binds DNA	ZNF40
R3H_encore_like (R3H)	predicted to bind ssDNA or ssRNA	R3HDM1
G patch domain	predicted to bind RNA	Gpatch8
heterogeneous nuclear ribonucleoprotein R, Q family (hnRNP-R-Q)	binds RNA	Rbfox1
helix-loop-helix domain (HLH)	binds DNA	Bhlhe40
SAND domain	binds DNA	Gmeb1

Table 3.4 The putative ZNF804A-interactors contain a number of conserved protein domains that can bind to DNA or RNA

The details of the predicted protein domains were obtained from the RefSeq on the NCBI database.

'RNA splicing' (4/339 P=3.43 x 10^{-5}) and 'regulation of alternative nuclear mRNA splicing, via splicesome' (2/20 P=5.98 x 10-5) (Figure 3.2B). The combined list of ZNF804A-interactors was enriched for genes in the GO term 'gene expression' (14/3834 P = 1.35 x 10^{-5}) and 'negative regulation of transcription, DNA dependent' (7/1026 P = 1.02 x 10^{-4}) (Figure 3.2C). These enrichments survived 5% FDR correction, as determined by GeneGo MetaCore™. The uncorrected P values are shown.

Often, it is possible to use Y2H data to identify a putative minimal binding region between two interacting proteins. However, this requires the identification of multiple clones for the same interactor. The majority of putative ZNF804A-interactors that were identified corresponded to only a single clone and as such, identifying a minimal binding region was not possible. However, multiple clones corresponding to zinc finger protein 40 (ZNF40) and RNPS1 were identified. Analysing the sequence of these clones suggests that the minimal binding region of ZNF804A to ZNF40 was upstream of ZNF40's C-terminal C2H2 type ZnF domains (Figure 3.3A) and that the minimal binding region of ZNF804A to RNPS1 included the RRM domain (Figure 3.3B).

3.3. Validation of Y2H results

The Y2H system may identify false-positive results which can be 'technical' false positives, that cannot be reproduced under identical experimental conditions; or 'biological' false positives, which do not occur *in vivo* (Huang and Bader, 2009). Therefore, true interactors must be confirmed using an alternative method; traditionally co-immunoprecipitation is used. Co-immunoprecipitation involves transiently expressing epitope tagged-proteins in mammalian cell lines and precipitating the proteins using an antibody raised against the epitope tag.

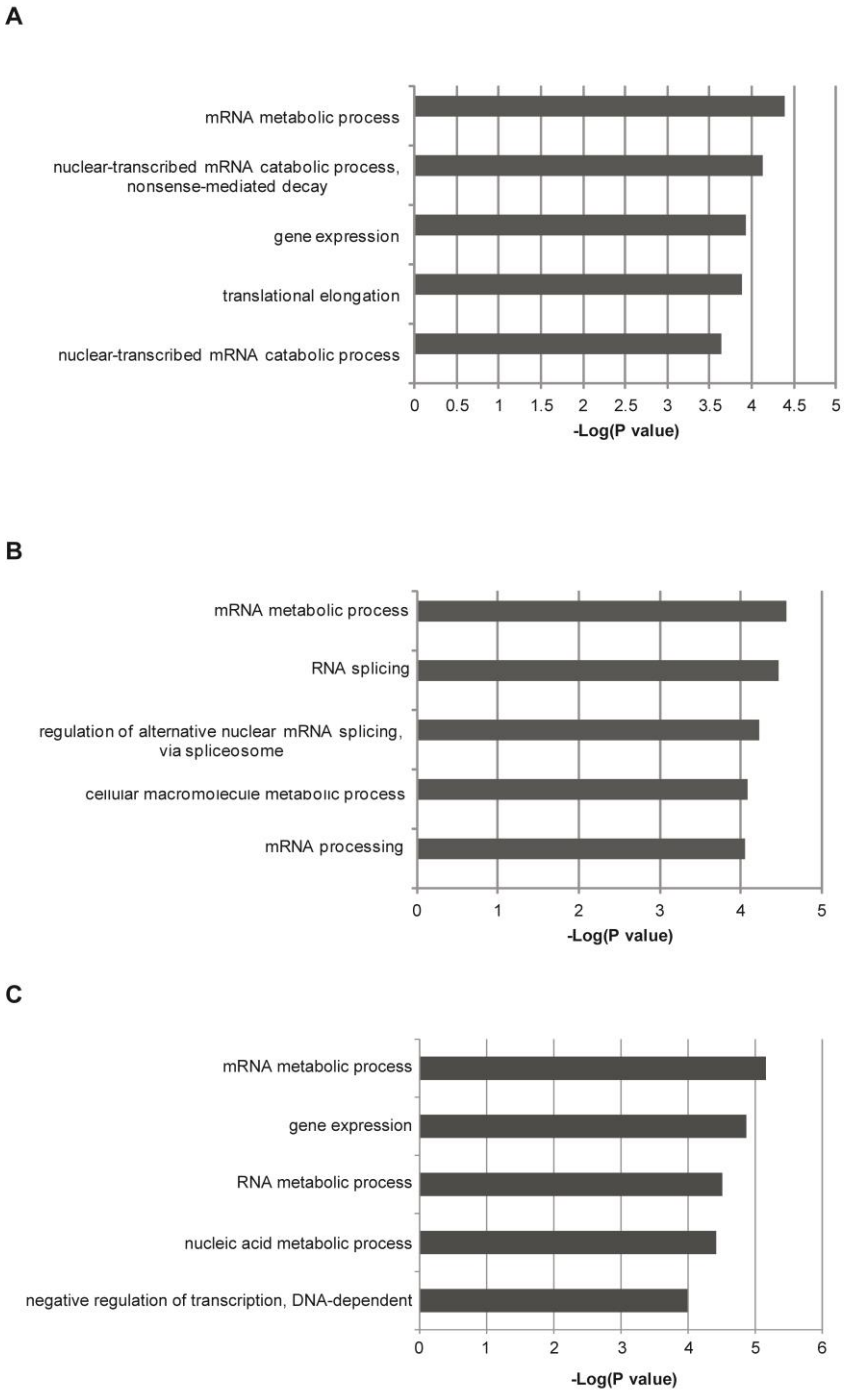


Figure 3.2 GO processes that were enriched for the putative ZNF804A-interactors

The list of putative ZNF804A-interactors identified in the Y2H screens using either **(A)** a human foetal brain cDNA library, **(B)** a mouse brain cDNA library or **(C)** the combined list of putative interactors were imported into GeneGo MetaCore™ and analysed using the enrichment analysis workflow for GO processes. The top five statistically significant GO processes are presented. These enrichments survived 5% FDR correction, as determined by GeneGo MetaCore™. The uncorrected P values are shown.

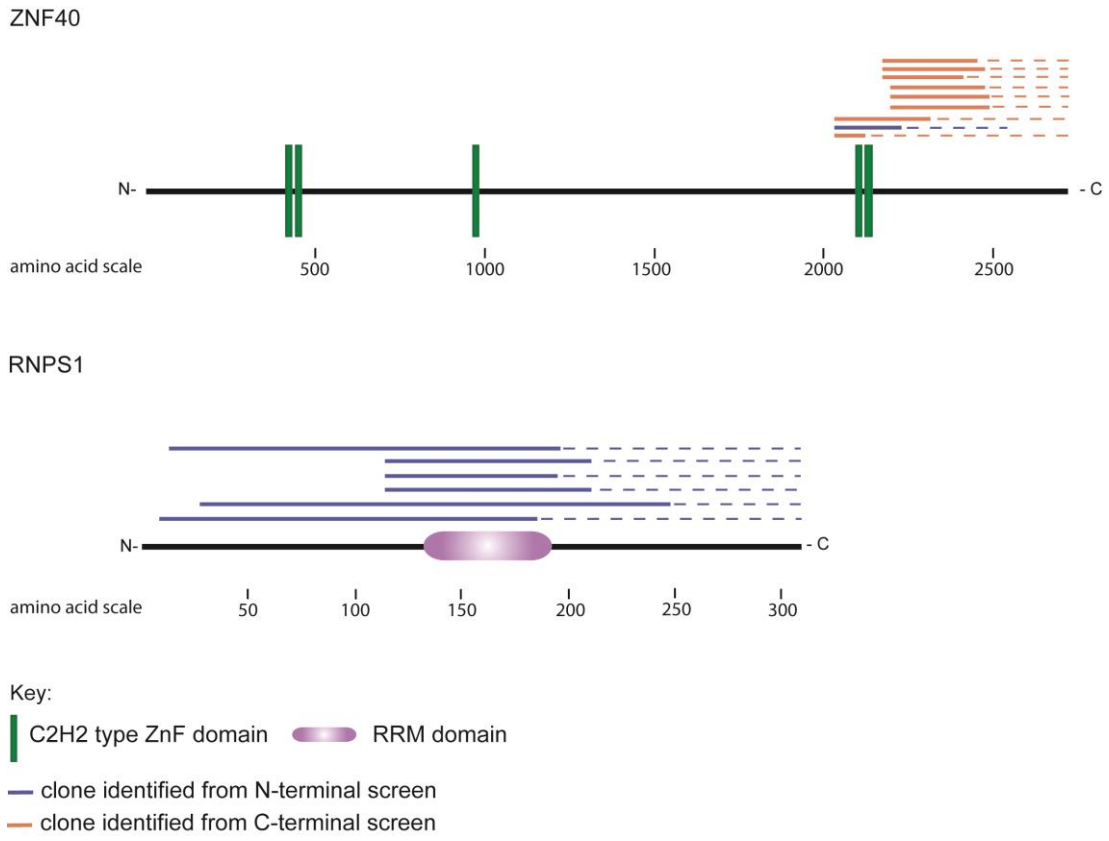


Figure 3.3 The distribution of the ZNF40 and RNPS1 clones

The distribution of the ZNF40 and RNPS1 clones isolated from a human foetal brain cDNA library screen using ZNF804A as the bait. The 5' cDNA sequence was obtained by sequencing of the cDNA clones identified in the Y2H screen (solid line). The remainder of the sequence was inferred from the size of insert obtained by restriction digest of the prey plasmid using the enzyme BSRG1 (dashed line). The locations of the sequences encoding predicted protein domains are indicated.

3.3.1. Detecting exogenous ZNF804A in cultured mammalian cells

At the outset of this study, attempts to synthesise tagged-ZNF804A in cultured mammalian cells using *ZNF804A* expression vectors had proved unsuccessful (C. L. Tinsley, personal communication; the vectors and cell lines used are summarised in Appendix 2). It was hypothesised that the inability to detect tagged-ZNF804A may indicate that over-expression of *ZNF804A* was cytotoxic or that ZNF804A was produced at extremely low levels. An alternative hypothesis was that few cells took up the vector cDNA. Previous studies have used conditional, stable cell lines to successfully overcome these difficulties (Reeves et al., 2002). Therefore, I generated a tetracycline-inducible, stable cell line expressing *myc-ZNF804A* (Chapter Five). Data presented in Chapter Five show the myc-ZNF804A Flp-In TREx cell line produced very low but detectable levels of myc-ZNF804A protein (section 5.2). Myc-ZNF804A was not detected after immunoprecipitation using anti-myc antibody (9E10) -conjugated beads, and was not detected by immunocytochemistry (section 5.2). The unsuccessful attempts to immunoprecipitate myc-ZNF804A from the myc-ZNF804A Flp-In TREx cell line meant that the cell line could not be used in co-immunoprecipitation experiments to validate the Y2H data.

While my studies were in progress, Girgenti and colleagues presented immunocytochemistry data which showed that transiently expressed myc-tagged ZNF804A localised to the nucleus in rat neural progenitor cells (Girgenti et al., 2012). Girgenti and colleagues also used subcellular fractionation and western blots probed with an anti-ZNF804A antibody (Santa Cruz) to evaluate the localisation of endogenous ZNF804A. These data showed endogenous ZNF804A was predominantly detected in the nuclear fraction and migrated to 136 kDa (Girgenti et al., 2012). To investigate whether our expression constructs were ineffective at producing ZNF804A, I obtained this vector cDNA (a kind gift from M. Girgenti). Please note

that, although this vector was designated pCAG-hZNF804A-myc by Girgenti and colleagues, DNA sequence analysis showed that the myc epitope tag was at the N-terminus of ZNF804A (data not shown).

To determine the utility of the pCAG-hZNF804A-myc expression vector, HEK293T cells were transfected with pCAG-hZNF804A-myc and the following day, protein lysates were prepared (section 2.6.1). ZNF804A protein levels were analysed by western blotting alongside a protein lysate prepared from cells transfected with our *myc-ZNF804A* expression vector (pCMV-myc-ZNF804A) (section 2.6.3). A lysate prepared from cells transfected with a *myc-TCF4* expression vector (pCMV-myc-TCF4) served as a positive control for the utility of the anti-myc antibody (9E10). The blots were incubated with the 9E10 antibody and our four custom anti-ZNF804A antibodies: 3078, 3077, 001 and 002 (C. L. Tinsley designed and characterised the custom anti-ZNF804A antibodies; these antibodies are described in Appendix 3). The western blot incubated with the 9E10 antibody showed myc-ZNF804A was not detected in protein lysates prepared from cells transfected with pCAG-hZNF804A-myc or pCMV-myc-ZNF804A (Figure 3.4). Myc-TCF4 was clearly detected in the protein lysate prepared from cells transfected with pCMV-myc-TCF4, confirming the utility of the 9E10 antibody. The blot probed with the 3078 antibody showed that a product (labelled A) which migrated to 175 kDa was detected all of the samples - including the negative controls - but appeared more abundant in the pCMV-myc-ZNF804A sample. However, the control blot incubated with the anti-α-tubulin antibody suggested that the pCMV-myc-ZNF804A sample had greater overall protein abundance. Furthermore, a product of this size was not detected on the blot probed with the 9E10 antibody. Therefore, it was concluded that this product was unlikely to be ZNF804A. The blot incubated with the 3077 antibody showed that a very faint product (labelled B) which migrated to 150 kDa was detected in the pCMV-myc-ZNF804A

sample. This product was not detected on the blot probed with the 9E10 antibody. Therefore, it was concluded that this product was unlikely to be ZNF804A.

In summary, these data suggest that pCAG-hZNF804A-myc did not produce detectable levels of myc-ZNF804A in HEK293T cells. Consistent with these findings, immunocytochemistry using the 9E10 antibody showed COS-7 cells transfected with pCAG-hZNF804A-myc did not produce detectable levels of myc-ZNF804A (section 2.8.1) (data not shown).

3.3.2. Degradation of myc-ZNF804A by the proteasome

Many intracellular proteins are degraded by the ubiquitin-proteasome system (UPS) following post-translational modification with ubiquitin or SUMO (small ubiquitin-related modifier) (Guo et al., 2012). Degradation by the UPS regulates a protein's half-life; this is a particularly important consideration for regulatory proteins, such as transcription factors (Desterro et al., 2000; Muratani and Tansey, 2003; Wang et al., 2008; Wang et al., 2011). The ubiquitin or SUMO modifiers facilitate interaction between the target protein, E3 ligases and the proteasome (Wang et al., 2008). Interestingly, the Y2H data presented in Table 3.3 shows that ZNF804A putatively interacted with an E3 ubiquitin ligase, Ttc3 (Suizu et al., 2009), and Mage-d1, which has been implicated in protein degradation (Mouri et al., 2012). Furthermore, additional Y2H data suggest that the ZNF804A paralog, GPATCH8, may interact with an E2 conjugating enzyme, Ubc9 (unpublished Y2H screening performed by C.L. Tinsley using a mouse brain cDNA library and a GPATCH8 bait), which transfers SUMO to the substrate (Desterro et al., 1997; Johnson and Blobel, 1997). Therefore, it was tempting to speculate that myc-ZNF804A protein was not detected after transient expression of *myc-ZNF804A* because the protein was post-translationally modified and targeted to the

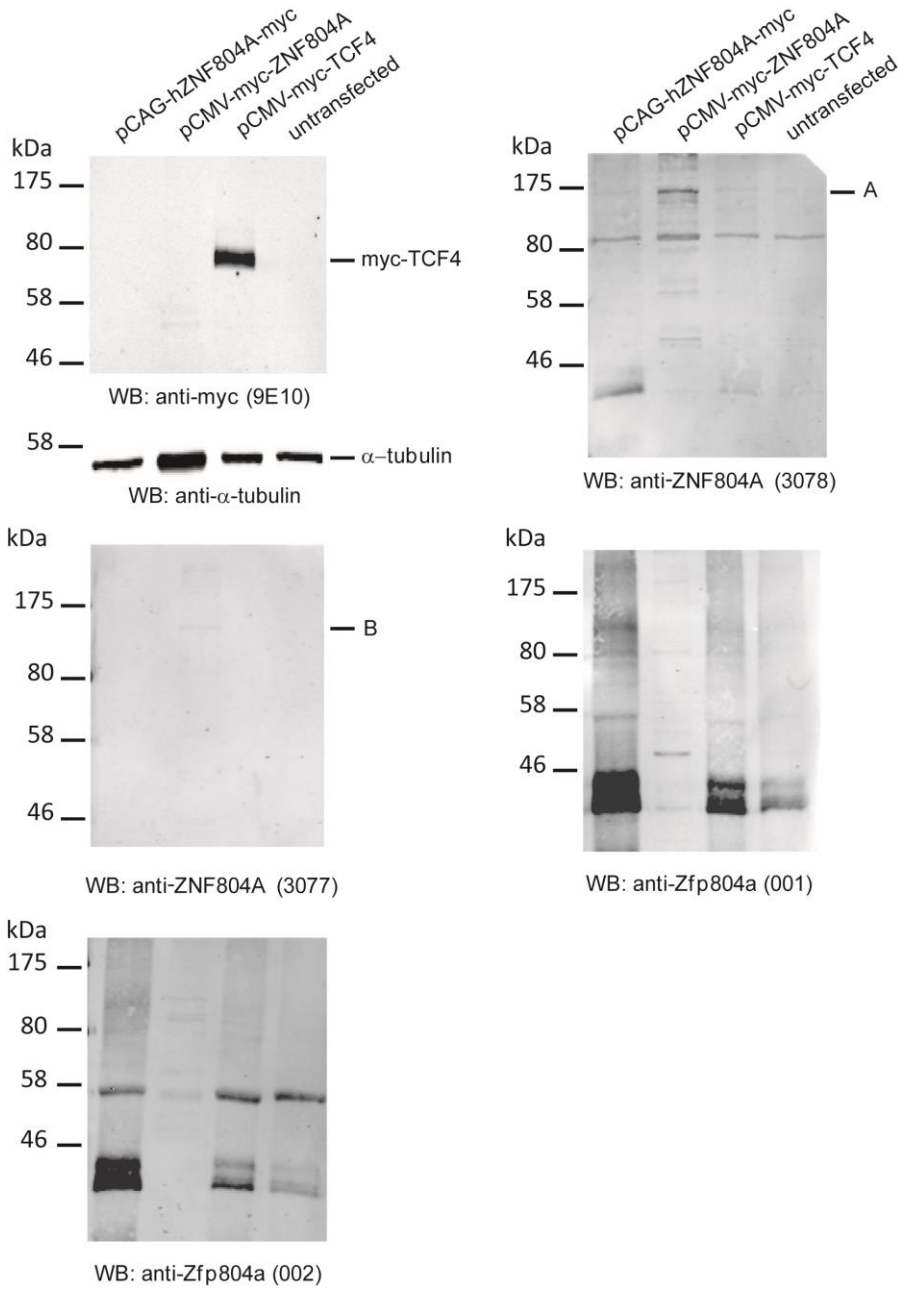


Figure 3.4 Western blot analysis of myc-ZNF804A

Protein lysates were prepared from HEK293T cells transfected with the plasmids indicated. Proteins were detected with either the anti-myc antibody (9E10) only or one of the four custom anti-ZNF804A antibodies raised against human ZNF804A (3077 and 3078) or the mouse homolog Zfp804a (002 and 001). The custom anti-ZNF804A antibodies are described in Appendix 3. A control protein lysate was prepared from cells transiently expressing myc-TCF4 to indicate the utility of the 9E10 antibody. On a separate but identical blot, the detection of endogenous α-tubulin was used as an internal loading control. The antibody dilutions used are given in Table 2.5. Single colour blots are shown. For imaging, the highest infrared fluorescence detection was used. The labels 'A' and 'B' indicate potential proteins of interest.

proteasome for degradation. To investigate this possibility, HEK293T cells were transfected with the pCMV-myc- ZNF804A expression vector and after 24h, the cells were treated with a proteasome inhibitor (lactacystin) for a further 18h (section 2.3.7). The hypothesis was that treating the cells with a proteasome inhibitor would prevent the putative degradation of ZNF804A. The abundance of myc-ZNF804A present in cells treated with a proteasome inhibitor and in cells left untreated was analysed by western blotting. The blots were probed with either the 9E10 antibody and a goat polyclonal anti-ZNF804A (D-14) antibody (Santa Cruz); or the 3077 antibody and an anti-ubiquitinylated proteins antibody (FK2). The blot probed with the FK2 antibody showed that the abundance of ubiquitinylated proteins increased after treatment with a proteasome inhibitor; this confirmed that the proteasome was successfully inhibited (Figure 3.5). The blots probed with the D-14, 9E10 and 3077 antibodies each showed a product that migrated to 175 kDa was detected in transfected cells only when the proteasome was inhibited (Figure 3.5). It is highly likely this product was myc-ZNF804A because it was detected by all three antibodies. These data are consistent with the hypothesis that myc-ZNF804A was targeted to the proteasome for degradation.

Having established that proteasome inhibition enabled myc-ZNF804A to be detected, a further two commercial anti-ZNF804A antibodies, P-13 and S-16 (Santa Cruz) were assessed. P-13, which was raised against mouse and rat Zfp804a, detected a very faint product that migrated to 175 kDa and was likely to be myc-ZNF804A (data not shown). S-16, which was also raised against mouse and rat Zfp804a, did not detect myc-ZNF804A (data not shown).

To evaluate whether the subcellular localisation of ZNF804A could be determined following proteasome inhibition, COS-7 cells expressing *myc-ZNF804A* were treated with lactacystin.

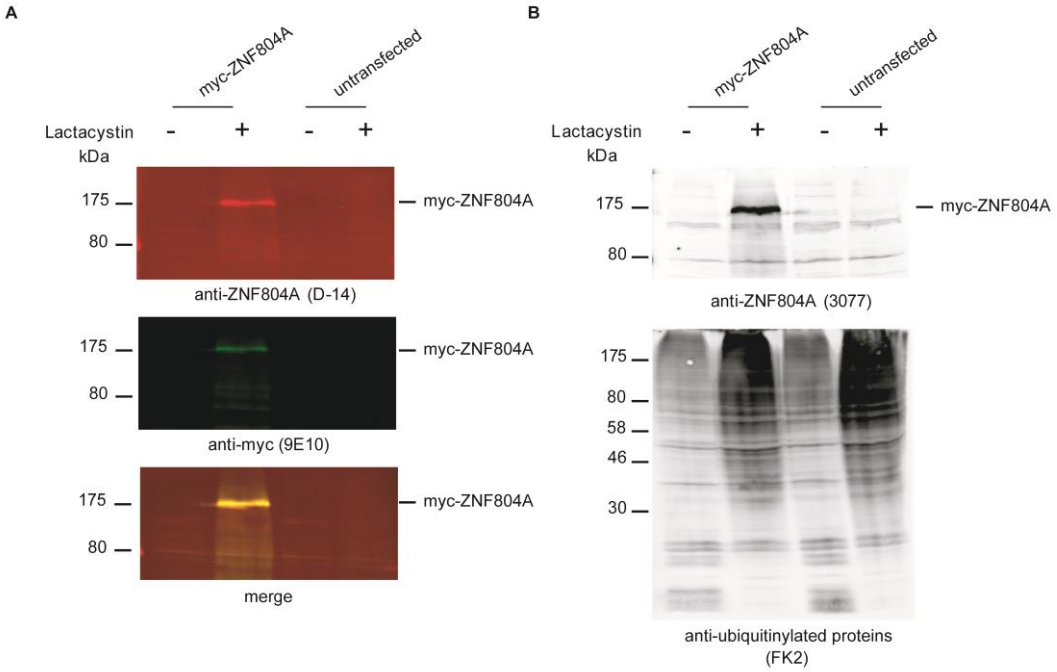


Figure 3.5 Western blot analysis of myc-ZNF804A after proteasome inhibition

Cultured HEK293T cells were transfected with a myc-ZNF804A expression vector. The following day, the cells were treated with a proteasome inhibitor (lactacystin; 10 µM). 18h after treatment, protein lysates were prepared and ZNF804A levels were analysed by western blotting. **(A)** Two-colour western blot probed with an anti-myc antibody (9E10) and an anti-ZNF804A (D-14) antibody (Santa Cruz). The blot was imaged using the same infrared fluorescence detection signal intensity (5) for the individual red and green channels. The merged image is shown to illustrate that the two antibodies detect the same product. **(B)** Two-colour western blot probed with a custom anti-ZNF804A (3077) antibody (described in Appendix 3.1) and the anti-ubiquitin (FK2) antibody which served as a control for successful proteasome treatment. Black and white images are shown for clarity. The antibody dilutions used are given in Table 2.5.

Myc-ZNF804A was detected using the 9E10 antibody while the nucleus was counter-stained using Hoechst. Consistent with the findings reported above, myc-ZNF804A was readily detected when the proteasome was inhibited whereas myc-ZNF804A was only detected in an extremely small number of cells when the proteasome was not inhibited. Myc-ZNF804A was localised homogenously throughout the nucleus and, in some cells, myc-ZNF804A was also evident in the cytoplasm (Figure 3.6).

3.4. Assessing the cellular localisation of transiently expressed GPATCH8

Paralogs may be used to provide insight into the biological role of a protein of unknown function because paralogs may be preserved in the genome by so-called 'subfunctionalisation' (Force et al., 1999; Lynch and Force, 2000). The molecular functions of the proteins encoded by the two *ZNF804A* paralogs: *GPATCH8* and *ZNF804B*, have not been elucidated. However, consistent with the possibility that ZNF804A and GPATCH8 may co-operate to share the same function as the protein encoded by their ancestral gene, data presented in Table 3.3 suggest these proteins putatively interact. Therefore, investigating the subcellular localisation of GPATCH8 may provide insight into the function of ZNF804A. Alignment of human GPATCH8 and ZNF804A amino acid sequences using ClustalW showed that they share 23% identity (Figure 3.7). Notably, the amino acid sequences showed homology in the region surrounding the C2H2 type ZnF domain. To determine the subcellular localisation of GPATCH8, COS-7 cells expressing GPATCH8-*enhanced yellow-fluorescent protein* (-EYFP) were processed for confocal microscopy. The nuclei were counter-stained using Hoechst and nuclear speckles were detected using the anti-SC35 antibody. The nuclear speckle marker was used because the Y2H data suggest ZNF804A may interact with factors involved in pre-mRNA processing (Tables 3.2 and 3.3), and nuclear speckles are enriched for pre-mRNA splicing factors (Perraud et al., 1979). Figure 3.8 shows

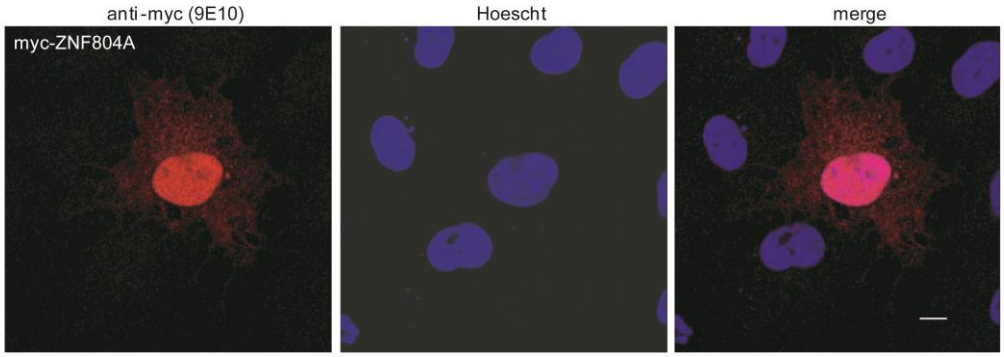
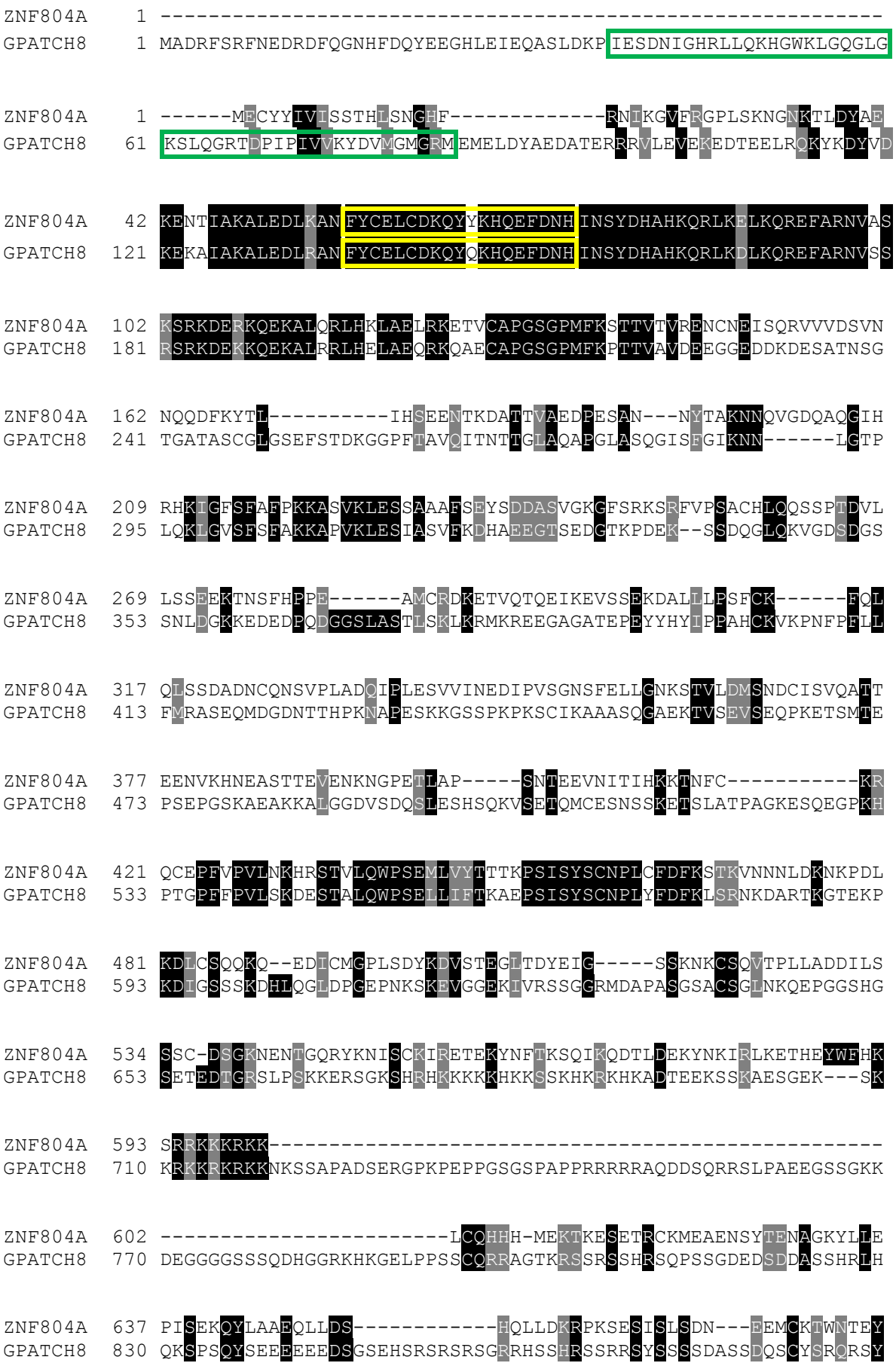


Figure 3.6 Localisation of myc-ZNF804A in COS-7 cells

Cultured COS-7 cells were transfected with a myc-ZNF804A expression vector. The following day, the cells were treated with proteasome inhibitor (lactacystin; 10 µM). 18h after treatment the cells were stained with an anti-myc antibody (9E10) and a marker for the nucleus (Hoechst; 1 µg/ml). Scale bar 10µm. The antibody dilutions used are given in Table 2.5.


ZNF804A 1 ----------
GPATCH8 1 MADRFSRFNEDRDFQGNHFDQYEEGHLEIEQASLDKPIESDNIGHRLLQKHGWKLGQGLG
ZNF804A 1 ------MECYYIVISSTHLSNGHF----------RNIKGVFRGPLSKNGNKTLDYAE
GPATCH8 61 KSLQGRTDPIPIVVKYDVMGMGRMEMELDYAEDATERRRVLEVEKEDTEELRQKYKDYVD
ZNF804A 42 KENTIAKALEDLKANFYCELCDKQYYKHQEFDNHINSYDHAHKQRLKELKQREFARNVAS
GPATCH8 121 KEKAIAKALEDLRANFYCELCDKQYQKHQEFDNHINSYDHAHKQRLKDLKQREFARNVSS
ZNF804A 102 KSRKDERKQEKALQRLHKLAELRKETVCAPGSGPMFKSTTVTVRENCNEISQRVVVDSVN
GPATCH8 181 RSRKDEKKQEKALRRLHELAEQRKQAECAPGSGPMFKPTTVAVDEEGGEDDKDESATNSG
ZNF804A 162 NQQDFKYTL----------IHSEENTKDATTVAEDPESAN---NYTAKNNQVGDQAQGIH
GPATCH8 241 TGATASCGLGSEFSTDKGGPFTAVQITNTTGLAQAPGLASQGISFGIKNN------LGTP
ZNF804A 209 RHKIGFSFAFPKKASVKLESSAAAFSEYSDDASVGKGFSRKSRFVPSACHLQQSSPTDVL
GPATCH8 295 LQKLGVSFSFAKKAPVKLESIASVFKDHAEEGTSEDGTKPDEK--SSDQGLQKVGDSDGS
ZNF804A 269 LSSEEKTNSFHPPE------AMCRDKETVQTQEIKEVSSEKDALLLPSFCK------FQL
GPATCH8 353 SNLDGKKEDEDPQDGGSLASTLSKLKRMKREEGAGATEPEYYHYIPPAHCKVKPNFPFLL
ZNF804A 317 QLSSDADNCQNSVPLADQIPLESVVINEDIPVSGNSFELLGNKSTVLDMSNDCISVQATT
GPATCH8 413 FMRASEQMDGDNTTHPKNAPESKKGSSPKPKSCIKAAASQGAEKTVSEVSEQPKETSMTE
ZNF804A 377 EENVKHNEASTTEVENKNGPETLAP-----SNTEEVNITIHKKTNFC----------KR
GPATCH8 473 PSEPGSKAEAKKALGGDVSDQSLESHSQKVSETQMCESNSSKETSLATPAGKESQEGPKH
ZNF804A 421 QCEPFVPVLNKHRSTVLQWPSEMLVYTTTKPSISYSCNPLCFDFKSTKVNNNLDKNKPDL
GPATCH8 533 PTGPFFPVLSKDESTALQWPSELLIFTKAEPSISYSCNPLYFDFKLSRNKDARTKGTEKP
ZNF804A 481 KDLCSQQKQ--EDICMGPLSDYKDVSTEGLTDYEIG-----SSKNKCSQVTPLLADDILS
GPATCH8 593 KDIGSSSKDHLQGLDPGEPNKSKEVGGEKIVRSSGGRMDAPASGSACSGLNKQEPGGSHG
ZNF804A 534 SSC-DSGKNENTGQRYKNISCKIRETEKYNFTKSQIKQDTLDEKYNKIRLKETHEYWFHK
GPATCH8 653 SETEDIGRSLPSKKERSGKSHRHKKKKKHKKSSKHKRKHKADTEEKSSKAESGEK---SK
ZNF804A 593 SRRKKKRKK----------
GPATCH8 710 KRKKRKRKKNKSSAPADSERGPKPEPPGSGSPAPPRRRRRAQDDSQRRSLPAEEGSSGKK
ZNF804A 602 ----------LCQHHH-MEKTKESETRCKMEAENSYTENAGKYLLE
GPATCH8 770 DEGGGGSSSQDHGGRKHKGELPPSSCQRRAGTKRSSRSSHRSQPSSGDEDSDDASSHRLH
ZNF804A 637 PISEKQYLAAEQLLDS----------HQLLDKRPKSESISLSDN---EEMCKIWNTEY
GPATCH8 830 QKSPSQYSEEEEEEDSGSEHSRSRSRSGRRHSSHRSSRRSYSSSSDASSDQSCYSRQRSY

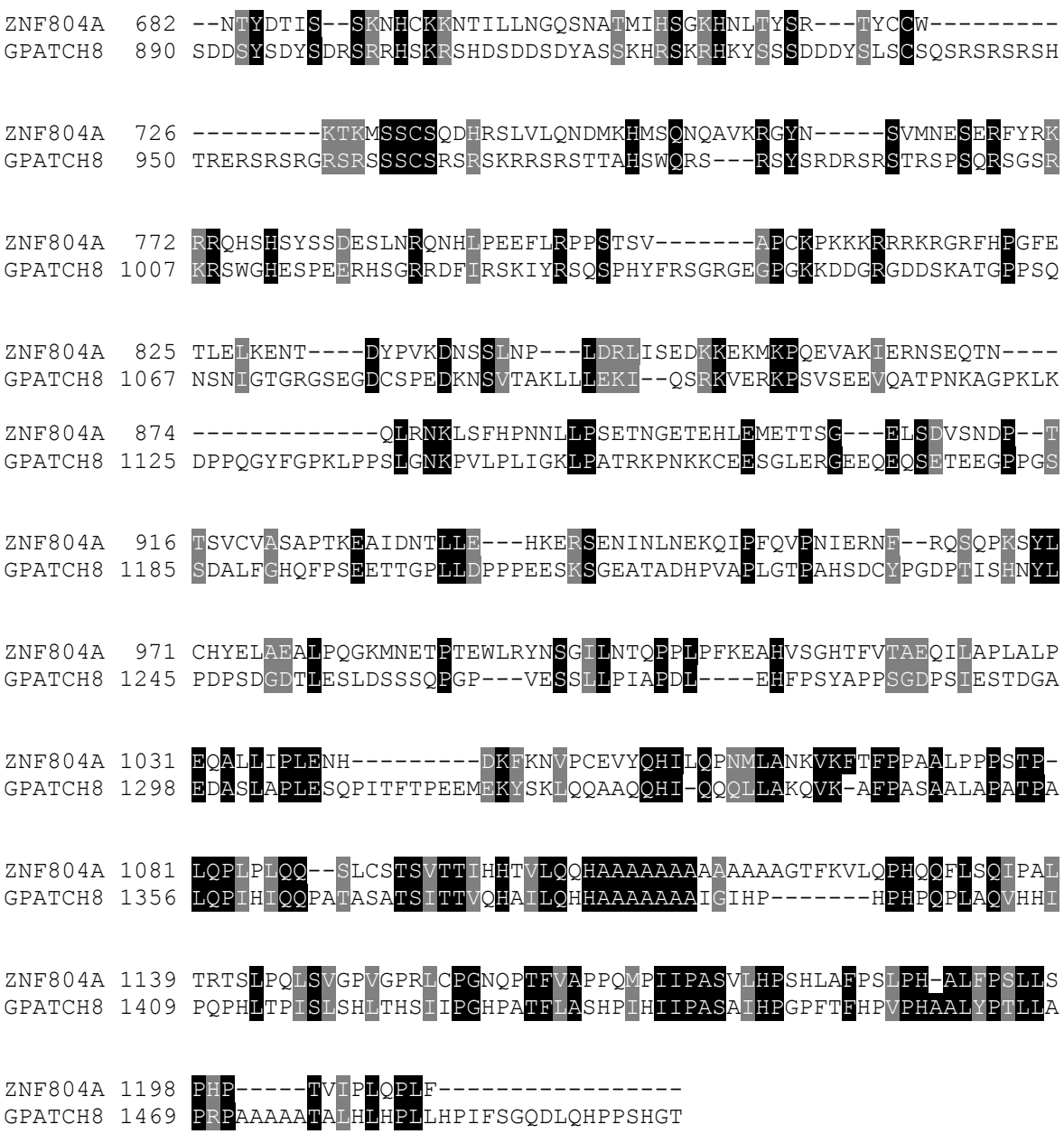


Figure 3.7 Alignment of human ZNF804A and GPATCH8 amino acid sequences

The human ZNF804A and GPATCH8 amino acid sequences were aligned using ClustalW, Kalign and Boxshade servers (section 2.11). The proteins shared 23% amino acid similarity. The regions of identity are coloured black whereas similar residues are highlighted in grey. The conserved C2H2 type ZnF domain is highlighted yellow on the alignment. The alignment indicated that the C2H2 type ZnF domain was highly conserved between these two proteins.

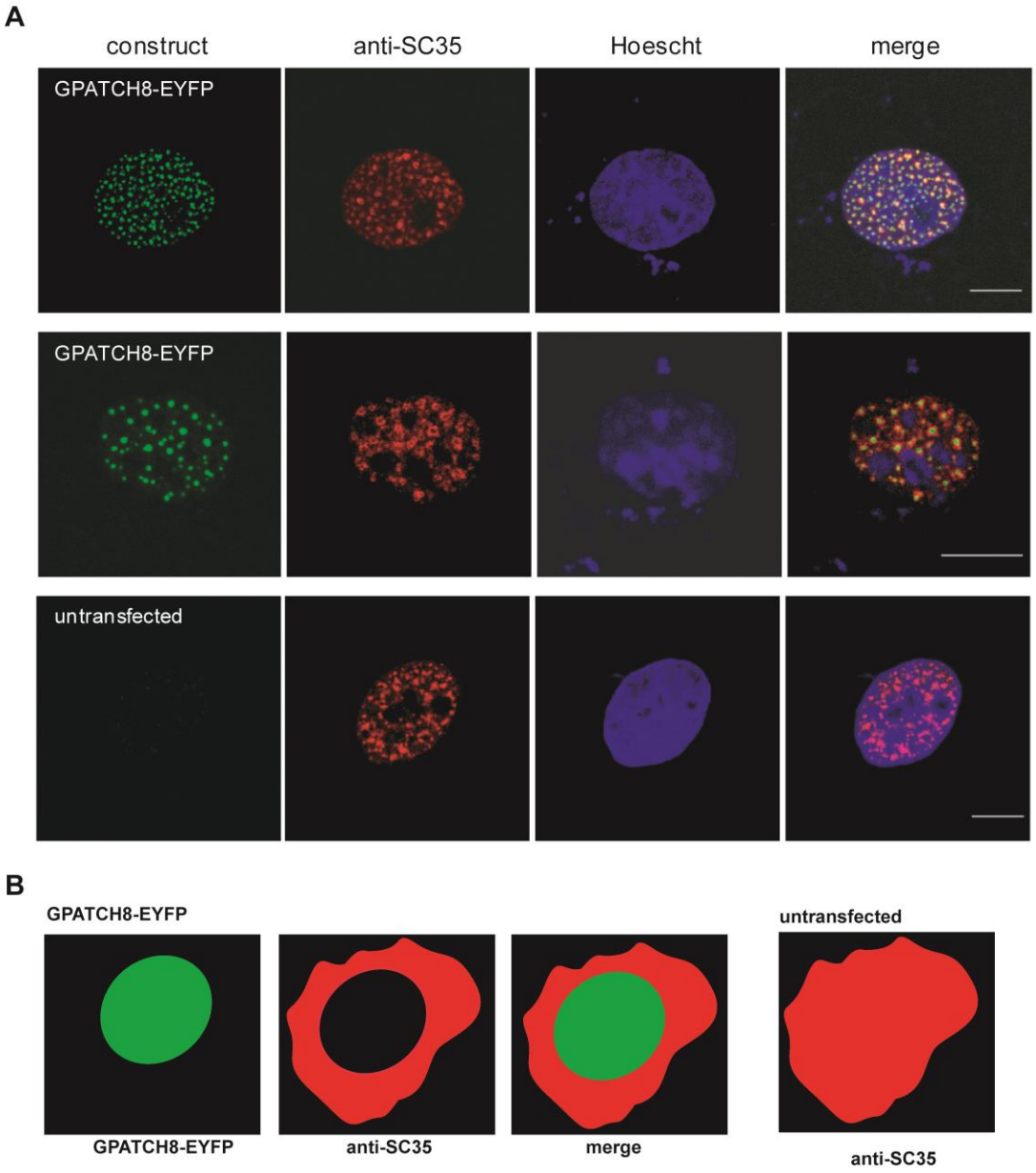


Figure 3.8 Localisation of GPATCH8-EYFP in COS-7 cells.

(A) Cultured COS-7 cells were transfected with a GPATCH8-EYFP expression vector. 24h post-transfection cells were stained markers for the nucleus (Hoechst; 10 µM) and nuclear speckles (anti-SC35). The cells were imaged by confocal microscopy using a 63x lens (top row) and a 100x lens (second row). Scale bar 10 µm. **(B)** Schematic representation of the change in nuclear speckle morphology in GPATCH8-EYFP transfected cells. This change in morphology was not observed in untransfected cells or cells transfected with an EYFP vector only (data not shown). The antibody dilutions used are given in Table 2.5.

GPATCH8-EYFP formed punctae within the nucleus. Some GPATCH8- EYFP punctae appeared to localise within atypical ring tori formed by nuclear speckles. The examination of untransfected cells (Figure 3.8) and cells transfected with EYFP vector only (data not shown) suggest that the nuclear speckles only formed ring tori when GPATCH8-EYFP was over-expressed.

3.5. Discussion

The principle aim of the work described in this Chapter was to provide the first insights into the biological function of ZNF804A by identifying its protein binding partners. The binding partners of ZNF804A were determined using Y2H screens with N- and C-terminal ZNF804A baits and human foetal and mouse brain cDNA libraries. Proteins belonging to the GO processes 'gene expression' and 'RNA splicing' were enriched among the putative ZNF804A-interactors. Thus, these data may suggest a functional role for ZNF804A as a regulator of gene expression at the level of transcription and/or pre-mRNA processing.

The suggestion that ZNF804A may be involved in transcriptional regulation is consistent with the current literature. For example, studies have reported ZNF804A interacted directly with the promoter/enhancer regions of *COMT* and *PRSS16*, which are known schizophrenia susceptibility genes, and up-regulated expression of these genes (Girgenti et al., 2012). Additionally, using a hypothesis-free, genome-wide microarray approach Hill and colleagues (2012) showed that knockdown of *ZNF804A* in human neural progenitor cells led to changes in expression of genes implicated in cell adhesion. Recently, studies have reported that ZNF804A over-expression in HEK293 cells altered the expression of genes involved in TGFβ signalling (Umeda-Yano et al., 2013).

The data presented in this Chapter expand on the current understanding of ZNF804A's potential role in transcriptional regulation by providing evidence that ZNF804A may interact with proteins which have known roles in pre-mRNA processing. Pre-mRNA processing (also known as 'splicing') is the process by which intronic sequences in the primary transcript (the pre-mRNA) are excised and the exonic sequences are joined together to produce a mature, translatable mRNA (Black, 2003). This process is catalysed by the spliceosome; a large structure in which numerous proteins co-operate to mediate the splicing reaction, and is directed by sequences at the intron/exon junctions called splice sites (Wahl et al., 2009). Alterations in splice site choice, via skipping of exons, retention of introns or the use of alternative splice sites, promoters or poly-A tails can change the resulting mRNA sequence; such changes are broadly termed 'alternative splicing' (Figure 3.9) (Black, 2003; Chen and Manley, 2009). Alternative splicing gives rise to changes in the mature mRNA sequence and consequently, can result in different protein products that may have different functions. As such, alternative splicing is considered to contribute greatly to proteomic diversity. The decision as to which exon is excised and which is included in the mature mRNA is complex and involves RNA sequence elements and many protein regulators. Several of the putative ZNF804A-interactors identified here are known pre-mRNA splicing regulators. For instance, ZNF804A interacted with RNPS1, a SR-protein which localises to nuclear speckles and regulates constitutive pre-mRNA splicing (Loyer et al., 1998; Mayeda, 1999; Sakashita et al., 2004). ZNF804A also interacted with CUGBP, Elav-like family member 4 (Celf4), a member of the Bruno/CELF (CUG-BP- and Etr-3-like factor) family of proteins which have multiple roles in post-transcriptional regulation, including roles in alternative splicing and mRNA translation (Barreau et al., 2006).

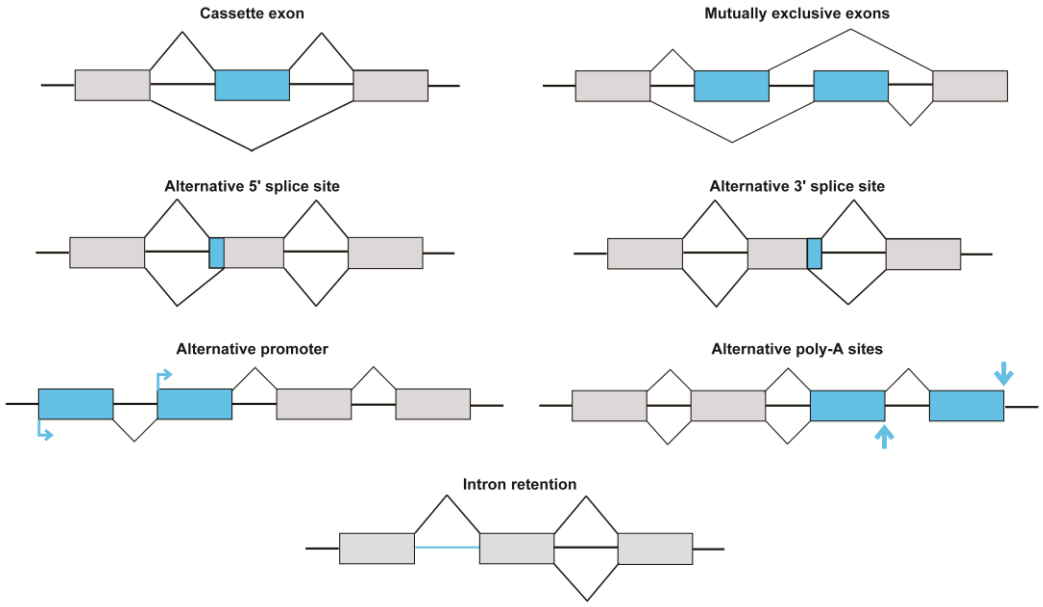


Figure 3.9 Patterns of alternative splicing

The primary transcripts from a gene can undergo many patterns of alternative splicing. The constitutive exons that are present in all mature mRNAs are represented as grey boxes. The alternatively spliced exons which may or may not be included in the mature mRNA are represented as blue boxes. The most common form of alternative splicing is a cassette exon, in which the whole exon may or may not be included in the mature mRNA. Paired cassette exons may be mutually exclusive, in which one exon or the other is included, but not both. The exons may be lengthened or shortened using different 5' or 3' splice sites. The transcriptional initiation at different promoters, or the start of polyadenylation (poly-A) at alternative sites, can lead to changes in the mature mRNA sequence. The exclusion of an intronic sequence may also be repressed, such that the intron is retained. A single gene may have many different alternative splicing patterns; this generates proteomic diversity. (Black, 2003)

In addition, ZNF804A interacted with tissue-specific alternative splicing factors (NOVA2, Rbfox1 and RBFOX2). In mouse brain, Nova2 is present exclusively in neurons and localises to the nucleus (Bemmo et al., 2008; Jensen et al., 2000; Yang et al., 1998). Using cross-linking and immunoprecipitation, Ule and colleagues showed that Nova2 binds RNAs that function at the neuronal synapse and are involved in neuronal inhibition (Ule, 2003). Subsequently, using splicing sensitive microarrays Ule and colleagues demonstrated that Nova2 regulated the splicing of targets which form an interaction module at the synapse (Ule et al., 2005). The Rbfox family of splicing factors comprises of three members, Rbfox1-3. The Y2H results indicate that ZNF804A interacted with both Rbfox1 and RBFOX2. Rbfox1 and Rbfox2 regulate splicing of many neuronal transcripts by binding to the consensus sequence (U)GCAUG in introns flanking alternatively spliced exons (Gehman et al., 2011; Yeo et al., 2009). The location of the binding sequence, either upstream or downstream of the alternative exon, determines whether the result of alternative exon usage will be increased or repressed exon inclusion (Jin et al., 2003; Underwood et al., 2005; Yeo et al., 2009). It is interesting to note that the only putative ZNF804A-interactor identified at the outset of this study, ATXN1, has also been implicated in pre-mRNA processing (Lim et al., 2006; Orr, 2010). Although no interaction between ATXN1 and ZNF804A was identified here, it is possible an interaction may be detected if more clones were screened.

Several proteins (including RNPS1 and RBFOX2) were identified in both the human foetal and mouse brain cDNA library Y2H screens. These data support the notion that these proteins are true ZNF804A-interactors. However, it would have been preferable to confirm the Y2H data in cultured mammalian cells using co-immunoprecipitation. This required ZNF804A over-expression in mammalian cells; yet at the outset of this study, transiently expressed ZNF804A could not be detected in mammalian cells. Data presented here support

these findings and show that transiently expressed ZNF804A was degraded by the proteasome (section 3.3.2). Proteins, particularly those involved in regulating gene expression, are targeted for degradation by post-translational modifications such as ubiquitination and SUMOylation (Desterro et al., 2000; Muratani and Tansey, 2003). Therefore, these data may be consistent with a role for ZNF804A in the regulation of gene expression. The finding that ZNF804A is only reliably detected after proteasome inhibition is inconsistent with the literature which suggests transiently expressed myc-ZNF804A has been readily detected in rat neural progenitor cells (Girgenti et al., 2012) and HEK293 cells (Umeda-Yano et al., 2013) without inhibiting the proteasome. To further understand ZNF804A's biological function, it will be useful to establish the nature of the post-translational modifications that target ZNF804A to the proteasome for degradation and under what cellular conditions these post-translational modifications occur. Having established that myc-ZNF804A migrated to 175 kDa (rather than 136 kDa as reported (Girgenti et al., 2013)), it is feasible that data presented in Figure 3.4 show that the 3078 antibody detected very low levels of myc-ZNF804A when the proteasome was not inhibited. Further experiments using proteasome inhibitors are needed to confirm this hypothesis.

Preliminary attempts have been made to immunoprecipitate exogenously synthesised myc-ZNF804A in the presence of a proteasome inhibitor using anti-myc antibody-conjugated beads. To date, immunoprecipitation of myc-ZNF804A has been unsuccessful; therefore these experiments require further optimisation. These experiments are important to enable validation of the Y2H data in mammalian cells.

To facilitate the understanding of ZNF804A's function in mammalian cells *in vitro,* the paralog of *ZNF804A, GPATCH8,* was used. GPATCH8-EYFP localised in punctae in the

nucleus within rings formed by nuclear speckles, suggesting GPATCH8 may have a role in pre-mRNA splicing (Figure 3.8). As ZNF804A and GPATCH8 are paralogs and putatively interact, it is tempting to interpret these data as further evidence that ZNF804A may function in the regulation of gene expression and particularly, pre-mRNA splicing. However, Figure 3.6 shows myc-ZNF804A localised homogenously throughout the nucleus and did not form punctae. It is possible that proteasome inhibition may have affected nuclear architecture and consequently altered the localisation of ZNF804A. Alternatively, accumulation of myc-ZNF804A following proteasome inhibition may have altered its nuclear localisation compared with endogenous levels. The potential inability of ZNF804A to localise to nuclear speckles does not preclude its involvement in pre-mRNA processing as little is understood about where in the nucleus splicing takes place (Han et al., 2011).

Notably, anti-SC35 staining showed that the nuclear speckles underwent a change in morphology from an orb-shape to a ring-torus shape when *GPATCH8-EYFP* was over-expressed. Using fluorescent microscopy, nuclear speckles are visible as punctae which vary in size and shape (Misteli, 2000; Spector and Lamond, 2011). However, electron microscopy shows that nuclear speckles are in fact composed of two morphologically distinct structures: a central region consisting of interchromatin granules, and a peripheral region containing nascent transcripts (Fakan, 1994). Interestingly, previous literature suggests that knockdown of a G-patch domain-containing protein called Son in HeLa cells led to a comparable change in nuclear speckle morphology (Sharma et al., 2010). Consequently, Sharma and colleagues proposed that Son has a role in maintaining the correct nuclear speckle structure. It is tempting to speculate that GPATCH8 may also have a role in maintaining nuclear speckle structure.

In summary, the results presented in this Chapter suggest ZNF804A putatively interacts with a range of proteins involved in regulating gene expression, including transcription factors and regulators of pre-mRNA processing. These data imply that, via interactions with its binding partners, ZNF804A may influence gene transcription and splicing of pre-mRNA transcripts. Consistent with a potential role for ZNF804A in pre-mRNA processing, the *ZNF804A* paralog and putative interactor, GPATCH8, localised in punctate within nuclear speckles enriched for pre-mRNA splicing factors (Spector and Lamond, 2011). Having established ZNF804A may have a role in the regulation of transcription and pre-mRNA splicing, human exon arrays were used to investigate the effects of knocking down and over-expressing *ZNF804A* on the cellular transcriptome (Chapters Four and Five). The working hypothesis was that manipulating *ZNF804A* mRNA levels would alter the expression and pre-mRNA splicing of transcripts regulated, directly or indirectly, by ZNF804A. Data presented in Chapters Four and Five are consistent with a role for ZNF804A in transcription and pre-mRNA processing. In particular, the data show that knockdown of *ZNF804A* altered splicing of a known target of RBFOX2 called *enabled homolog* (*ENAH*) (section 4.6.2).

Chapter 4: Investigating the effects of depleting *ZNF804A* on the cellular transcriptome

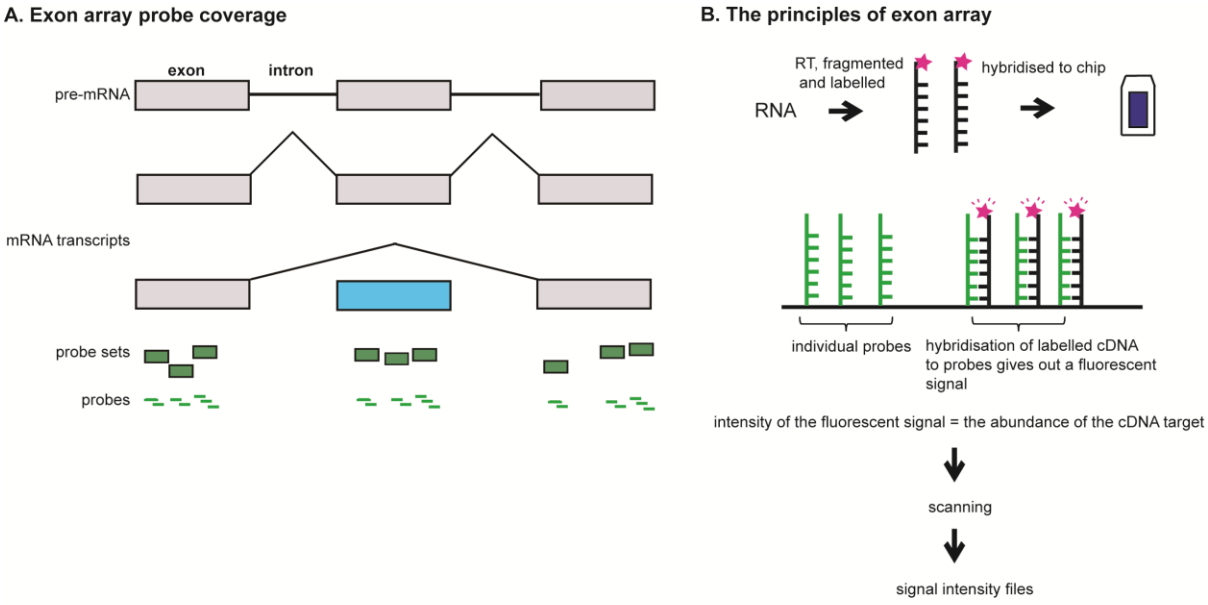
4.1. Introduction

In Chapter Three, Y2H screening was used to identify the putative protein binding partners of ZNF804A. Data presented in Chapter Three suggest ZNF804A putatively interacts with a range of proteins involved in regulating gene expression including transcription factors and regulators of pre-mRNA processing, such as RNPS1, members of the RBFOX family and NOVA2. Fluorescent microscopy showed the ZNF804A paralog and putative interactor, GPATCH8, localised in punctate within nuclear speckles enriched for pre-mRNA splicing factors (Spector and Lamond, 2011). These data imply that, via interaction with its protein binding partners, ZNF804A may influence transcription and pre-mRNA processing.

RNA interference (RNAi) is a mechanism by which double stranded RNA (dsRNA) can trigger the sequence-specific suppression of mRNA (Fire et al., 1998). The task of specific gene knockdown *in vitro* has been facilitated by the development of small interfering RNAs (siRNAs) (Elbashir et al., 2001). RNAi is a useful tool to investigate gene function, particularly when combined with systematic analysis of the downstream consequences of mRNA knockdown. For example, transcriptional targets of transcription factors can be identified by depleting the endogenous transcription factor and analysing any subsequent changes in gene expression using microarrays (Yu et al., 2010). The combination of siRNA and splicing-sensitive microarray technologies (known as exon arrays; Figure 4.1) has also proved useful in identifying the targets of predicted regulators of pre-mRNA splicing (Cheung et al., 2009; Hung et al., 2008a; Warzecha et al., 2009). The aim of the experiments

described in this Chapter was to investigate the role of ZNF804A in the regulation of transcription and pre-mRNA processing using a combination of siRNA, exon array technologies and pathway analysis. The workflow is shown in Figure 4.2 and is described as follows: siRNA was used to knockdown endogenous *ZNF804A* in the human neuroblastoma cell line SH-SY5Y (Ross and Biedler, 1985). The SH-SY5Y cell line was chosen because it endogenously expresses *ZNF804A* and it has many of the properties of mature neurons (Pahlman et al., 1990; Vaughan et al., 1995). Subsequently, the Affymetrix GeneChip human exon 1.0 ST array was used to profile changes in the cellular transcriptome and the data was analysed in the context of biological pathways.

The hypothesis was that knockdown of endogenous *ZNF804A* would lead to changes in the expression and pre-mRNA splicing of transcripts regulated, directly or indirectly, by ZNF804A. Yet, because very few studies have analysed the effects of depleting a splicing regulator on genome-wide pre-mRNA processing (Cheung et al., 2009; Hung et al., 2008a; Warzecha et al., 2009), it was difficult to predict how knockdown of *ZNF804A* would influence splicing if ZNF804A was a true regulator of pre-mRNA processing. Therefore, to establish a benchmark comparison, siRNA was used to deplete the mRNA of a gene (*glyceraldehyde 3-phosphate dehydrogenase* (*GAPDH*)) which has no known role in pre-mRNA processing. It was predicted that there would be relatively more changes in pre-mRNA splicing in *ZNF804A*-depleted cells than in *GAPDH*-depleted cells when compared to mock transfections.


A. Exon array probe coverage
exon
intron
pre-mRNA
mRNA transcripts
probe sets
probes
B. The principles of exon array
RT, fragmented and labelled
RNA
hybridised to chip
individual probes
hybridisation of labelled cDNA to probes gives out a fluorescent signal
intensity of the fluorescent signal = the abundance of the cDNA target
scanning
signal intensity files

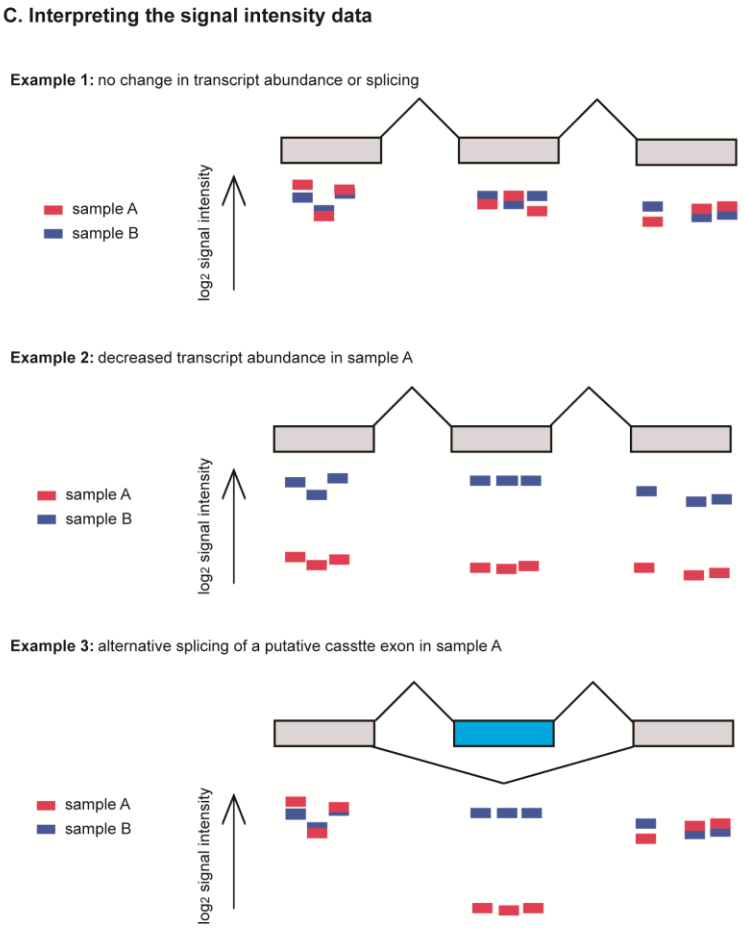

C. Interpreting the signal intensity data
Example 1: no change in transcript abundance or splicing
sample A
sample B
log2 signal intensity
Example 2: decreased transcript abundance in sample A
sample A
sample B
log2 signal intensity
Example 3: alternative splicing of a putative casstte exon in sample A
sample A
sample B
log2 signal intensity

Figure 4.1 Schematic overview of the exon array design, method and data interpretation

The Affymetrix GeneChip human exon 1.0 ST array is used to evaluate changes in gene expression and splicing. **(A)** Pre-mRNA is composed of exons and introns. The array is a comprehensive platform that contains approximately 5.4 million probes grouped into 1.4 million probe sets interrogating over 1 million exon clusters. **(B)** The RNA samples are prepared for exon array by reverse transcription (RT), fragmentation and labelling. The cDNA is then hybridised to the chip. **(C)** The signal intensity data from each individual probe is summarised to the probe set-level. The probe set level (exon-level) signal intensity data can be virtually assembled into transcript clusters (gene-level) using a meta-probe set list. The assembly of exon-level data into transcript clusters allows accurate quantification of changes in pre-mRNA splicing (Whistler, 2010) and gene expression (Kapur et al., 2007). In example 1 there is no difference in the $\log_2$ signal intensities for any of the probe sets across the gene, implying there is no difference in transcript abundance or splicing between the sample groups. In example 2 all of the probe sets in sample A have smaller $\log_2$ signal intensities, implying the transcript cDNA is less abundant in sample A. In example 3 the $\log_2$ signal intensities of the probe sets in the second exon (blue) suggest this exon is less abundant in sample A, but the $\log_2$ signal intensities of the probe sets in the first and third exons are unchanged between the two samples. Therefore, it is likely the second exon is alternatively spliced in sample A.

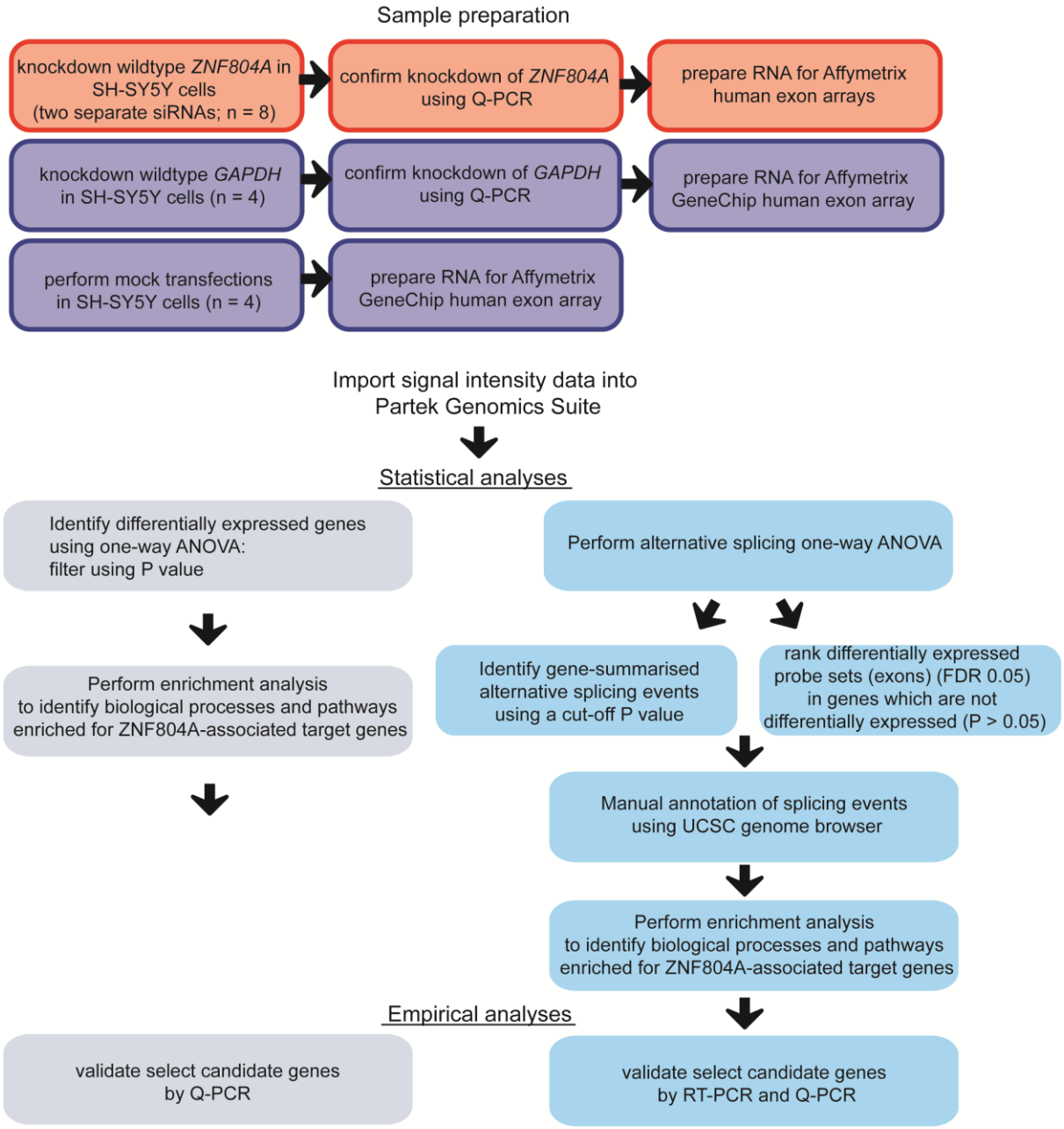


Figure 4.2 Flowchart illustrating the experimental design

Short interfering RNAs (siRNAs) were used to knockdown the mRNA of interest in SH-SY5Y cells. RNA was prepared and exon arrays were performed by Central Biotechnology Services (CBS), Cardiff University. The raw .CEL files which contained the signal intensity information were imported into the Partek Genomics Suite (PGS) using the core metaprobe set annotation file. The robust multiarray averaging (RMA) algorithm was used to summarise the probe-level data to a single value for each probe set. Two parallel analyses were performed to identify changes in gene expression and pre-mRNA splicing. The results were filtered to reduce false-positives, and a selection of changes was empirically confirmed.

4.2. The knockdown of wildtype *ZNF804A* mRNA using siRNA-mediated RNAi

The aim of these experiments was to examine the effects of *ZNF804A* knockdown. Therefore, to ensure that only specific on-target effects of *ZNF804A* knockdown were analysed, two siRNA duplexes were designed to target exons two and three of *ZNF804A* and the signal intensity files from these samples were pooled to give a '*ZNF804A*-knockdown' group for the analysis (Jackson, Bartz et al. 2003, Berns, Hijmans et al. 2004). Appendix 5.1 depicts the location of the sites of siRNA directed cleavage, in the context of *ZNF804A*'s structure. C.L.Tinsley designed and optimised the siRNA duplexes used to knockdown *ZNF804A* (designated siZNFA and siZNFB) and *GAPDH* (designated siGAP) and performed the knockdown experiments (section 2.3.8). Mock transfections were also performed to control for any changes in the transcriptome in response to exposure to the transfection reagent. To ensure that any changes in gene expression or pre-mRNA splicing reported were robust and reproducible; four independent biological replicates were performed. Reverse transcription (RT) reactions were performed to generate 1st strand cDNA and the level of transcript knockdown was assessed using quantitative-PCR (Q-PCR) analysis in collaboration with C.L. Tinsley (sections 2.9.1 and 2.3.6). The Q-PCR primer sequences used are listed in Appendix 1.4. Appendix 5.1 illustrates the location of the Q-PCR primers, in the context of the gene's structure and in relation to the sites of siRNA directed cleavage. Figure 4.3A shows *ZNF804A* was depleted to a mean average of 18.7% (sd = 8.4) and a mean average of 23.1% (sd = 8.8) of endogenous levels using siZNFA and siZNFB respectively. Figure 4.3B shows *GAPDH* was depleted to 14.3% (sd = 3.2) of endogenous levels using siGAP. These RNA samples were used for the subsequent exon array (n = 4 for each condition). The quality of the RNA was assessed on the Agilent 2100 bioanalyser (Central Biotechnology Services (CBS), Cardiff University). The RNA integrity number (RIN) for each sample was 10. This meant that the RNA was intact and suitable for gene expression analysis.

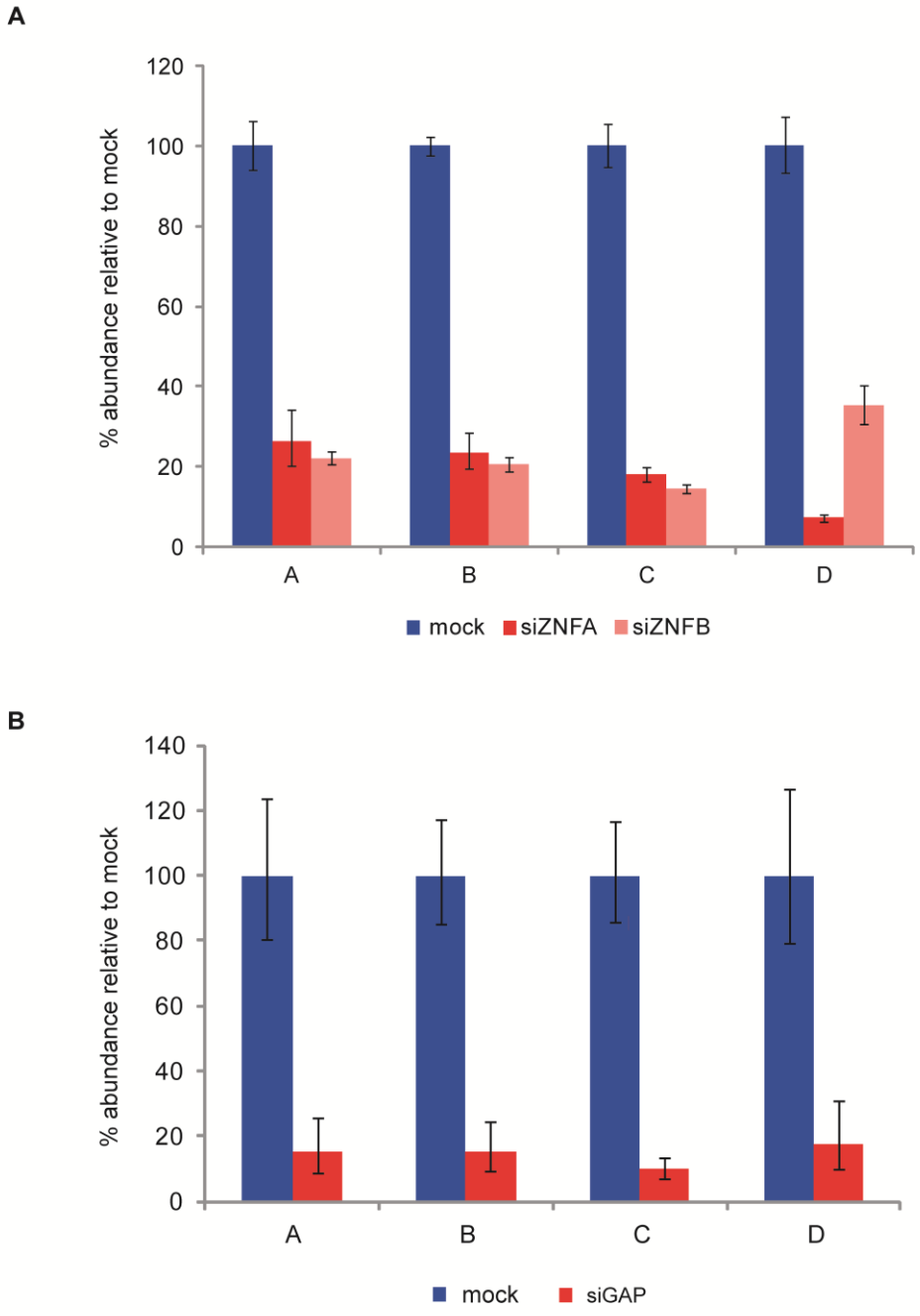


Figure 4.3 *ZNF804A* and *GAPDH* were depleted after siRNA treatment

In four independent experiments (A-D), SH-SY5Y cells were treated twice at 48h intervals with either a siRNA duplex designed to target *ZNF804A* (siZNFA or siZNFB) or a siRNA duplex designed to target *GAPDH* (siGAP). The following day, RNA was prepared. **(A)** Q-PCR to quantify *ZNF804A* knockdown was performed in triplicate. The raw C_t values were normalised to *GAPDH* levels. **(B)** Q-PCR to quantify *GAPDH* knockdown was performed in triplicate. The raw C_t values were normalised to *ACTB* levels. The error bars represent the standard deviation of the three raw C_t values. These RNA samples were used for the subsequent exon array experiments.

4.3. Processing of the exon array chips and quality assessment

The exon arrays were performed by CBS, Cardiff University (section 2.9.1). Briefly, the RNA samples were converted into labelled cDNA, hybridised to the Affymetrix GeneChip human exon array 1.0 ST plate and scanned. I received and analysed the resulting .CEL files which contained the raw signal intensity information for each probe set. To eliminate non-specific effects of siRNA treatment, the differentially expressed and alternatively spliced transcripts in *ZNF804A*-depleted cells were identified relative to the *GAPDH*-depleted cells (Jackson, Bartz et al. 2003, Berns, Hijmans et al. 2004). Therefore, the .CEL files were imported into the Partek Genomics Suite (PGS) and analysed in three groups: 1) pooled siZNF-treated samples (n = 8) versus siGAP-treated samples (n = 4); 2) pooled siZNF-treated samples (n = 8) versus mock samples (n = 4); and 3) siGAP-treated samples (n = 4) versus mock samples (n = 4). The .CEL files were imported using the core metaprobe set and robust multiarray averaging (RMA) algorithm (section 2.9.2). RMA is a model-based method that generates robust signal estimates by down-weighing probes that perform poorly (Irizarry et al., 2003b; Lockstone, 2011). Supplementary data provided in Appendices 4A and B show quality assessment metrics performed in the PGS confirmed RMA normalisation was successful. To identify related samples, agglomerative hierarchical clustering was performed (section 2.9.3), and the resulting cluster dendrograms are shown in Figures 4.4 and 4.5. The clustering was performed using the gene-summarised expression values from the top 100, 1000 and 5000 most differentially expressed genes. The clustering patterns correlated with *ZNF804A* knockdown. These data suggest that *ZNF804A*-depleted cells were systematically different to the control cell lines.

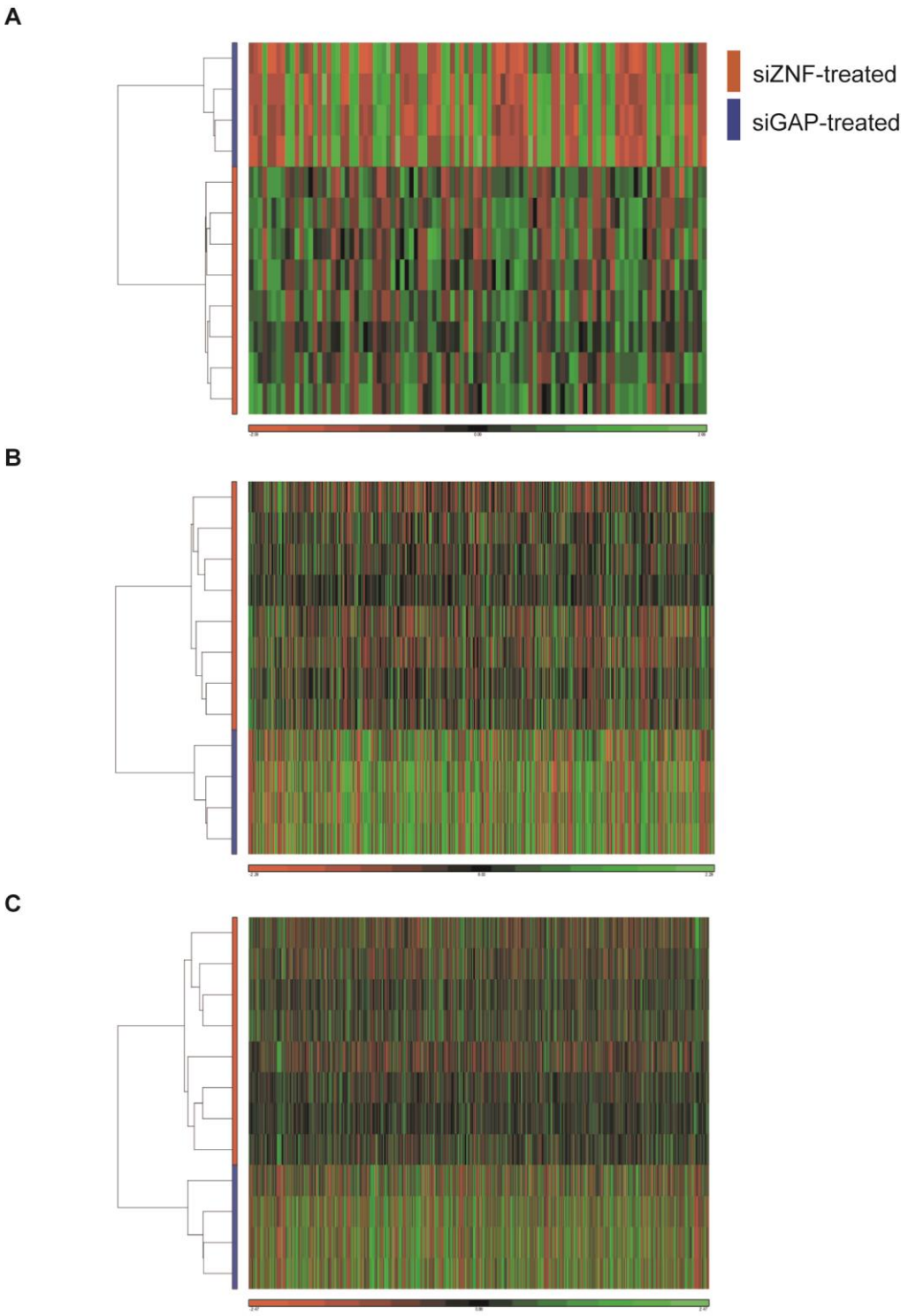


Figure 4.4 Hierarchical clustering of the most differentially expressed genes between siZNF-treated and siGAP-treated samples

Hierarchical clustering of the **(A)** 100, **(B)** 500 and **(C)** 1000 most differentially expressed genes between siZNF-treated samples (red) and siGAP-treated samples (blue). The heatmap colours are artificial: green = relative up-regulation and red = relative down-regulation black = no difference.

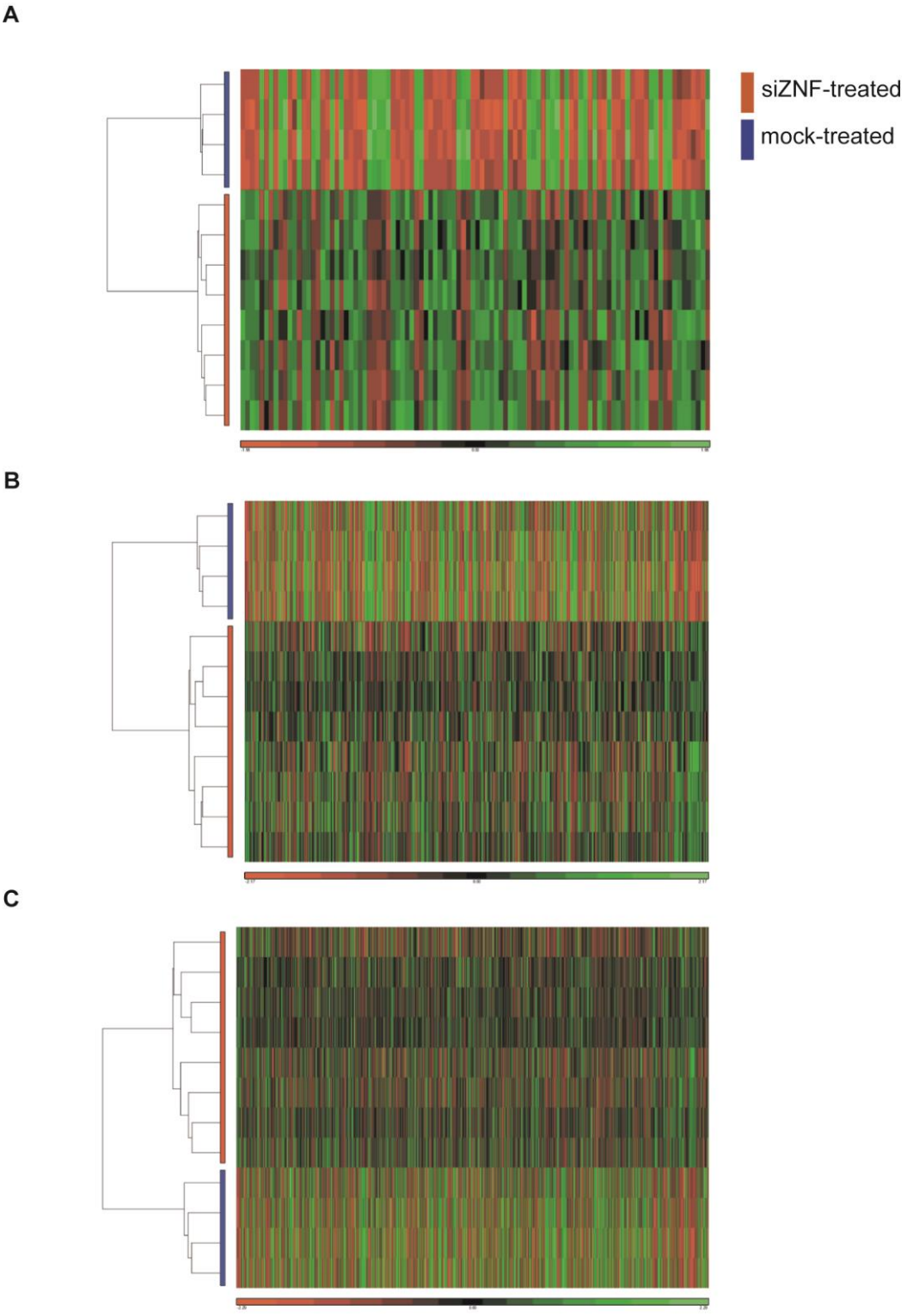


Figure 4.5 Hierarchical clustering of the most differentially expressed genes between the siZNF-treated and the mock samples

Hierarchical clustering of the **(A)** 100, **(B)** 500 and **(C)** 1000 most differentially expressed genes between siZNF-treated samples (red) and mock samples (blue). The heatmap colours are artificial: green = relative up-regulation and red = relative down-regulation black = no difference.

4.4. The PGS 'geneview' of *ZNF804A*

The PGS represents exon array data graphically in the form of a 'geneview'. A geneview contains three key pieces of information: 1) the mean log_2 signal intensity for each probe set assigned to the gene for each sample group 2) the fold change in log_2 signal intensity for each probe set between sample groups and 3) the known mRNA transcript variants. Figure 4.6A shows the geneview for *ZNF804A*. This geneview shows that only probe sets in exon four of *ZNF804A* had reduced log_2 signal intensity after *ZNF804A* knockdown. This implies only exon four of *ZNF804A* was depleted after siZNF treatment. These data contradict the Q-PCR analysis of *ZNF804A* knockdown which was performed using a primer/probe pair designed complementary to the 5' end of *ZNF804A* (Figure 4.3; Appendix 5.1). Figure 4.6B shows the *GAPDH* geneview was consistent with a reduction in log_2 signal intensity across every probe set complementary to *GAPDH*. Therefore, it is unlikely that the *ZNF804A* geneview reflects the RNAi method used to deplete *ZNF804A*. The most parsimonious explanation for the *ZNF804A* geneview is that non-specific mRNA cross-hybridised with the probe sets complementary to exons 1-3 of *ZNF804A* on the exon array and artificially increased the signal intensity readings for these probe sets. However, further experiments using RNA-sequencing or northern blotting are required to explore the possibility that there may be more than one transcript variant of *ZNF804A*.

4.5. Identifying genes with altered expression after *ZNF804A* knockdown

To identify genes with altered expression in *ZNF804A*-depleted cells, a one-way ANOVA was performed on the gene-summarised expression values, between the pooled siZNF-treated samples and the siGAP-treated samples (section 2.9.4). The expression of 579 genes differed significantly (FDR 0.01; Table 4.1). Of these genes, 263 had a fold change of greater than 1.5 or less than -1.5 and 66 had a fold change of greater than 2 or less than -2. The top 10

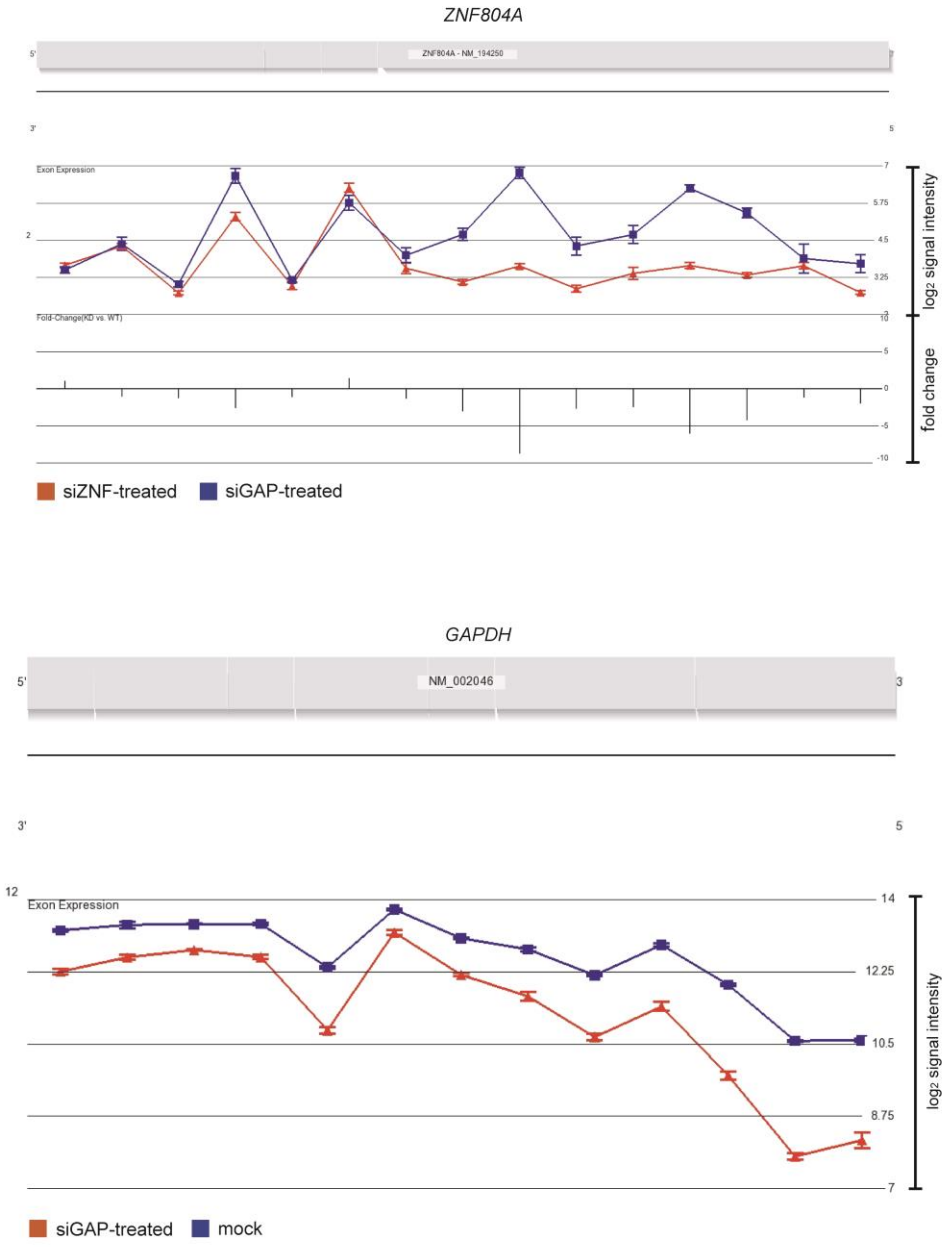


Figure 4.6 The geneviews of *ZNF804A* and *GAPDH*

(A) There was reduced *ZNF804A* mRNA abundance in *ZNF804A*-depleted cells ($P = 2.13 \times 10^{-8}$; FC = -2.01; one-way ANOVA). The geneview of *ZNF804A* was the same when either siGAP-treated or mock samples were used as the control. **(B)** There was reduced *GAPDH* mRNA abundance in *GAPDH*-depleted cells compared to the mock samples. FC = fold change.

Gene name	Gene symbol	P value	FC
alanyl (membrane) aminopeptidase	*ANPEP*	1.17E-09	-2.53
inositol polyphosphate-5-phosphatase	*INPP5A*	1.54E-09	2.04
retinal degeneration 3	*RD3*	6.93E-09	2.15
protocadherin beta 14	*PCDHB14*	1.65E-08	2.50
zinc finger protein 804A	*ZNF804A*	2.13E-08	-2.01
caspase 3	*CASP3*	2.27E-08	-2.81
MARVEL domain containing 1	*MARVELD1*	3.28E-08	2.17
F-box and leucine-rich repeat protein 2	*FBXL2*	3.31E-08	2.00
calcium channel, voltage-dependent, T type, alpha 1G subunit	*CACNA1G*	3.45E-08	-1.94
solute carrier family 36	*SLC36A1*	5.68E-08	1.84
notch 3	*NOTCH3*	8.22E-08	-2.02
protein tyrosine phosphatase domain containing 1	*PTPDC1*	8.64E-08	1.39
ATPase, H+ transporting, lysosomal 42kDa, V1 subunit C1	*ATP6V1C1*	1.01E-07	3.37
tribbles homolog 2 (Drosophila)	*TRIB2*	1.03E-07	1.97
placental growth factor	*PGF*	1.07E-07	-1.69
bone morphogenetic protein 2	*BMP2*	1.56E-07	1.62
solute carrier family 9 (sodium/hydrogen exchanger)	*SLC9A6*	2.30E-07	-2.00
protein S (alpha)	*PROS1*	2.44E-07	1.65
T-cell leukemia translocation altered gene	*TCTA*	2.48E-07	2.25
EPH receptor A6	*EPHA6*	2.50E-07	2.16
sestrin 3	*SESN3*	2.77E-07	-3.11
connector enhancer of kinase suppressor of Ras 2	*CNKSR2*	3.19E-07	-1.56
salt-inducible kinase 2	*SIK2*	3.36E-07	-1.88
vasoactive intestinal peptide	*VIP*	3.52E-07	3.15
LAG1 homolog, ceramide synthase 2	*LASS2*	3.99E-07	2.29
nidogen 2 (osteonidogen)	*NID2*	4.29E-07	-1.51
Kruppel-like factor 7 (ubiquitous)	*KLF7*	5.41E-07	1.53
sphingomyelin phosphodiesterase 3, neutral membrane (neutral sphingomyelinase II)	*SMPD3*	6.76E-07	-2.32
potassium voltage-gated channel, shaker-related subfamily, member 3	*KCNA3*	6.95E-07	1.61
chromosome 12 open reading frame 53	*C12orf53*	7.16E-07	1.63
chromosome 6 open reading frame 48	*C6orf48*	8.02E-07	1.53
tissue factor pathway inhibitor (lipoprotein-associated coagulation inhibitor)	*TFPI*	8.74E-07	-1.45
adrenergic, alpha-2C-, receptor	*ADRA2C*	8.85E-07	1.50
MYST histone acetyltransferase 1	*MYST1*	1.05E-06	-1.39
GDNF family receptor alpha 2	*GFRA2*	1.09E-06	2.13

Table 4.1 The most significantly differentially expressed genes after *ZNF804A* knockdown

The .CEL files were imported into the PGS using the core metaprobe set and RMA normalisation. The .CEL files were assigned as either siZNF- or siGAP-treatment. A one-way ANOVA was performed on the gene-summarised expression values, using treatment as the candidate variable in the ANOVA model. The expression of 579 genes differed significantly (FDR 0.01). The top 35 genes ranked by P value are shown. FC = fold change.

positive and negative fold changes are shown in Table 4.2.

Using siRNA to initiate the RNAi pathway can result in off-target effects (Jackson, Bartz et al. 2003, Berns, Hijmans et al. 2004). To negate these off-target effects, the siGAP-treated samples were used as the negative control condition. To ensure using the siGAP-treated samples as the negative control condition did not alter the conclusions drawn from the study, a one-way ANOVA was performed on the gene-summarised expression values between the siGAP-treated samples and the untreated, mock samples, and the list was compared to the 579 genes identified as differentially expressed between siZNF-treated samples and siGAP-treated samples. Using a 1% FDR, the expression of 47 genes differed significantly between the siGAP-treated samples and mock samples (gene list provided in Appendix 5.2). There was an overlap of only 10 genes between these 47 genes and the 579 identified as differentially expressed between siZNF-treated and siGAP-treated samples. This suggests that the overwhelming majority of the 579 differentially expressed genes were true *ZNF804A*-related events.

4.5.1. Enrichment analysis of the genes with altered expression after *ZNF804A* knockdown

Some of the genes that were differentially expressed in *ZNF804A*-depleted cells relative to the *GAPDH*-depleted cells have been consistently implicated in neuropsychiatric disease using a range of genetics and post-mortem studies, including *reelin* (*RELN*), *neuropeptide Y* (*NPY*) and components of glutamate signalling, such as *glutamate receptor 7* (*GRM7*) (Table 4.3).

While it was tempting to place emphasis on these individual genes, it was more informative

Gene name	Gene symbol	P value	FC
Up-regulated genes			
EGF-like repeats and discoidin I-like domains 3	*EDIL3*	2.49E-04	4.11
reelin	*RELN*	1.03E-05	3.91
LY6/PLAUR domain containing 1	*LYPD1*	2.28E-06	3.56
ATPase, H+ transporting, lysosomal 42kDa, V1 subunit C1	*ATP6V1C1*	1.01E-07	3.37
vasoactive intestinal peptide	*VIP*	3.52E-07	3.15
Rho GTPase activating protein 36	*ARHGAP36*	3.11E-05	3.11
histone cluster 1, H4h	*HIST1H4H*	3.59E-06	2.83
cyclin-dependent kinase inhibitor 1A (p21, Cip1)	*CDKN1A*	2.13E-06	2.69
ER degradation enhancer, mannosidase alpha-like 2	*EDEM2*	1.40E-06	2.59
achaete-scute complex homolog 1 (Drosophila)	*ASCL1*	2.25E-04	2.54
Down-regulated genes			
hepatocyte growth factor (hepapoietin A; scatter factor)	*HGF*	4.30E-05	-12.43
cellular retinoic acid binding protein 2	*CRABP2*	1.04E-04	-4.04
ADAM metallopeptidase with thrombospondin type 1 motif, 3	*ADAMTS3*	8.69E-06	-3.24
decorin	*DCN*	2.09E-05	-3.11
sestrin 3	*SESN3*	2.77E-07	-3.11
serpin peptidase inhibitor, clade F (alpha-2 antiplasmin, pigment epithelium derived factor), member 1	*SERPINF1*	2.38E-06	-2.91
fms-related tyrosine kinase 1 (vascular endothelial growth)	*FLT1*	6.79E-05	-2.84
caspase 3, apoptosis-related cysteine peptidase	*CASP3*	2.27E-08	-2.81
transmembrane protein with EGF-like and two follistatin-l	*TMEFF1*	2.89E-05	-2.58
coiled-coil domain containing 80	*CCDC80*	6.86E-05	-2.57

Table 4.2 The genes with the largest fold changes in expression after *ZNF804A* knockdown

The .CEL files were imported into the PGS using the core metaprobe set and RMA normalisation. The .CEL files were assigned as either siZNF- or siGAP-treatment. A one-way ANOVA was performed on the gene-summarised expression values, using treatment as the candidate variable in the ANOVA model. The expression of 579 genes differed significantly (FDR 0.01). The top 10 positive and negative fold changes in expression are shown. (FC = fold change.)

to use a systematic, statistical approach to identify particular biological processes which were enriched for *ZNF804A*-related target genes. Therefore, enrichment analysis was performed using GeneGo MetaCore™ software (section 2.10). GeneGo MetaCore™ was chosen as the analysis software because it utilises both GO processes and GeneGo's manually curated processes and pathways. To ensure that only genes that could be detected as differentially expressed in the exon array were included in the enrichment analysis, the background comparison was set as the Affymetrix GeneChip human exon array 1.0 ST-v2. Of the 579 differentially expressed genes, 544 were represented in GeneGo MetaCore™. Numerous GO process terms were identified as significantly enriched among the differentially expressed genes including 'neuron projection development' (72/774 P = 3.35 $x10^{-15}$), 'cell projection organisation' (82/986 P = 1.96 x 10^{-14}) and 'nervous system development' (142/2250 P = 2.66 x 10^{-14}) (Figure 4.7A). These enrichments survived 5% FDR correction for multiple testing. Genes belonging to GeneGo process networks 'synaptic contact as an element of cell adhesion' (21/167 P = 8.39 x 10^{-6}) and 'axonal guidance as an element of neurogenesis and development' (21/219 P = 4.5 x 10^{-4}) were significantly enriched among the differentially expressed genes (Figure 4.7B). These enrichments survived 5% FDR correction for multiple testing. The differentially expressed genes belonging to 'synaptic contact as an element of cell adhesion' included *neuroligin 2* (*NLGN2*), *neurofascin* (*NFASC*) and *regulating synaptic membrane exocytosis 1* (*RIMS1*). There was no apparent difference in the enrichment categories of the up-regulated and down-regulated genes (data not shown).

4.5.2. GWAS enrichment analysis

If the genes with altered expression after *ZNF804A* knockdown represent the causal effects of *ZNF804A*'s contribution to disease, it was hypothesised that these genes may show genetic associations with neuropsychiatric disease. Many secondary studies of GWAS data use gene-

Gene name	Gene symbol	Evidence for involvement in pathophysiology of schizophrenia	siZNF vs. siGAP		siZNF vs. mock	
			P value	FC	P value	FC
Neuroligin 2	*NLGN2*	Rare mutations in *NLGN2* are associated with disease (Sun et al., 2011).	1.51E-04	-1.29	3.86E-03	-1.25
Caspase 3	*CASP3*	Decreased expression in oligodendrocytes in schizophrenia patients (Kerns et al., 2010).	2.27E-08	-2.81	5.39E-07	-2.05
Reelin	*RELN*	Post-mortem brain studies show a decrease in *RELN* expression in schizophrenia (Guidotti et al., 2000; Impagnatiello et al., 1998); genetics analysis identified association between common variation in *RELN* and disease (Kahler et al., 2008; Shifman et al., 2008) and disruptions in RELN signalling cascade have been associated with disease (Costa et al., 2002).	1.03E-05	3.91	8.86E-04	2.35
Ephrin B2	*EFNB2*	Genetic analysis identified association with schizophrenia (Zhang et al., 2010). Also, EFNB2 is an essential component of RELN pathway and regulates neuronal migration (Senturk et al., 2011)	3.88E-06	-2.14	1.87E-04	-1.61
Neuropeptide Y	*NPY*	Post-mortem brain studies of schizophrenia and bipolar disorder patients show decrease in *NPY* expression in frontal cortex (Kuromitsu et al., 2001) and disorganisation of NPY containing neurons in DPFC (Ikeda et al., 2004). Additionally, impaired NPY signalling is implicated in pathophysiology of psychiatric disorders (Caceda et al., 2007; Eaton et al., 2007).	8.13E-06	2.05	3.36E-05	1.86
Glutamate receptor 7	*GRM7*	*GRM7* encodes a G-protein-coupled receptor for glutamate. Glutamate dysfunction has been implicated in pathophysiology of schizophrenia (Carlsson et al., 2001). CNVs in *GRM7* have been found in schizophrenia patients (Saus et al., 2010).	4.85E-06	1.26	6.72E-06	1.26
Neurotensin	*NTS*	NTS is implicated in dopamine and NMDA interactions (Li et al., 2010; Tanganelli et al., 2012).	4.31E-06	1.95	1.11E-07	2.80

Table 4.3 The literature implicates some of the differentially expressed genes in the neurobiology of schizophrenia. (FC = fold change.)

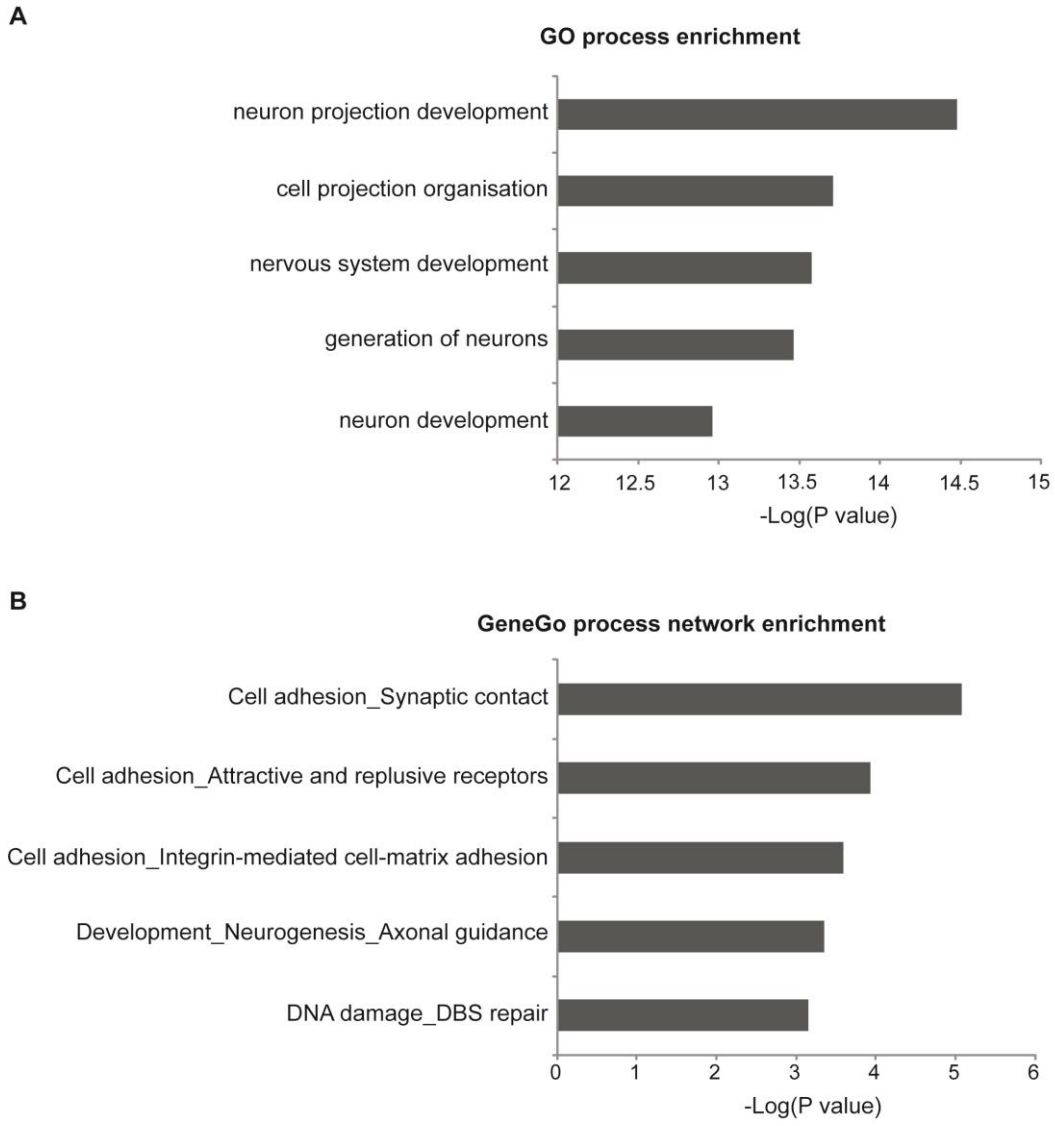


Figure 4.7 Biological processes that were significantly enriched for genes showing differential expression after *ZNF804A* knockdown

The list of genes showing differential expressed between siZNF-treated and siGAP-treated samples using a 1% FDR (section 4.5.1) was imported into GeneGo MetaCore™. Of the 579 differentially expressed genes, 544 were represented in GeneGo MetaCore™. The enrichment analysis was performed using Affymetrix GeneChip human exon array 1.0 ST v-2 as the background list. The bar charts show the top statistically significant **(A)** GO processes and **(B)** GeneGo process networks identified by the enrichment analysis. The enrichments presented were statistically significant when 5% FDR correction was applied, as determined by GeneGo MetaCore™. The uncorrected P values are shown.

based tests rather than information from individual SNP markers to assess whether a group of genes are genetically associated with disease. Gene-based association P values (also known as gene-wide P values) are calculated by combining the information from several SNP markers (Moskvina et al., 2011). Here, gene-wide P values from the schizophrenia and bipolar disorder psychiatric GWAS consortium (PGC) datasets (kindly provided by V. Moskvina) were used to test the hypothesis that there were more genes significantly associated with disease (gene-wide $P < 0.05$) in the differentially expressed gene list than would be expected by chance. The significance of the over-representation was tested using a normal distribution. The results show that more genes that have a gene-wide significant association with disease were differentially expressed in *ZNF804A*-depleted cells (FDR 0.05) than expected by chance ($P = 0.03$ for schizophrenia and $P = 0.04$ for bipolar disorder) (Table 4.4). Manual inspection of the genomic location of the differentially expressed genes which were genetically associated with disease suggested they were not in close proximity to one another (data not shown). Therefore, it is unlikely that this result reflects LD between the SNPs. This finding suggests that the genes with altered expression after *ZNF804A* knockdown are moderately enriched for genes genetically associated with disease. These data provide independent support for a causal link between the genes with altered expression after *ZNF804A* knockdown and neuropsychiatric disease. When a more stringent cut-off for differential expression was used (FDR 0.01), there was no significant difference from the normal distribution ($P = 0.23$ for schizophrenia and $P = 0.22$ for bipolar disorder; Table 4.4). However, there were more genes genetically associated with disease observed than expected, suggesting that these data did follow the same trend.

To establish whether the genes with altered expression after *ZNF804A* knockdown were more strongly associated with disease than genes which did not show altered expression, a Mann

Whitney U test was used. The null hypothesis of this test was that the gene-wide P values of the differentially expressed genes had the same distribution as the gene-wide P values of the non-differentially expressed genes. The results showed there was no significant difference in the distribution of the gene-wide P values for differentially expressed and non-differentially expressed genes for either the schizophrenia dataset (P = 0.868) or the bipolar disorder dataset (P = 0.808) (data not shown).

PGC dataset	exon array significance cut-off	number of overlapping gene symbols (PGC/exon array)	number of genes significantly associated with disease (P < 0.05) in the list of differentially expressed genes		sd	P value
			observed	expected		
schizophrenia	FDR 0.05	1367/1755	85	68.35	8.91	**0.03**
	FDR 0.01	446/578	26	22.3	5.09	0.23
bipolar disorder	FDR 0.05	1367/1755	85	69.45	8.98	**0.04**
	FDR 0.01	446/578	28	23.85	5.26	0.22

Table 4.4 Enrichment analysis of the genes significantly associated with neuropsychiatric disease (gene-wide P < 0.05) in the lists of genes which were differentially expressed after *ZNF804A* knockdown

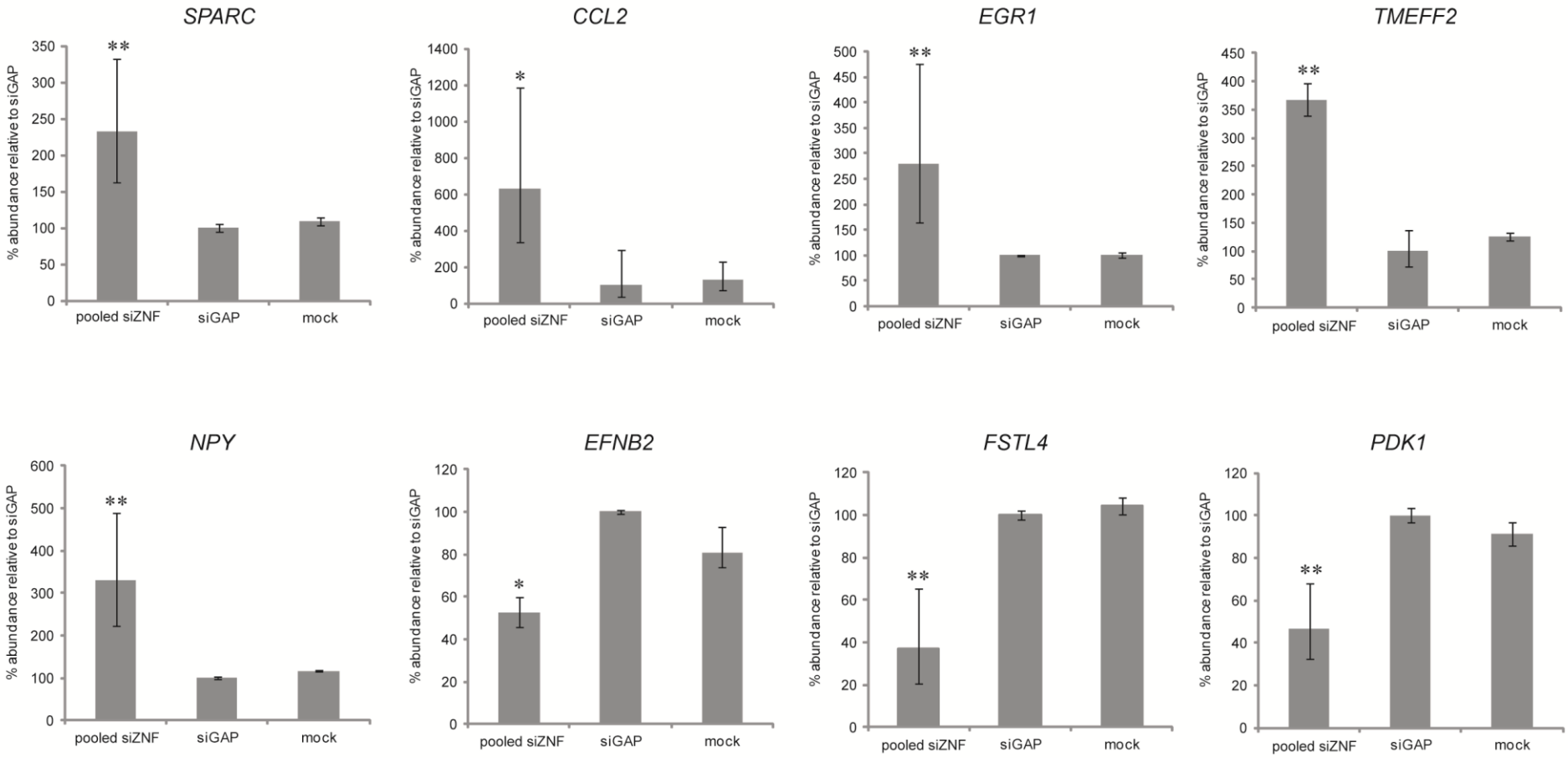
The PGC datasets were filtered using the gene symbols of the genes differentially expressed after *ZNF804A* knockdown relative to the siGAP-treated samples. The number of overlapping gene symbols is shown after *ZNF804A* was removed from the analysis. A filter was applied to retain only the genes with a gene-wide association P < 0.05. The expected number of retained genes was calculated using a significance level of 0.05. The significance of the over-representation was tested using a normal distribution. (PGC = Psychiatric GWAS consortium; sd = standard deviation.)

4.5.3. Q-PCR validation of gene expression changes after *ZNF804A* knockdown

Currently, there is no 'gold standard' for analysing gene expression measurements therefore, it is important to consider the reliability of the array results by validating a selection of the identified target genes experimentally (Allison et al., 2006; Chuaqui et al., 2002). Here, Q-PCR was used to validate the gene expression data from the exon array. Please note, only one biological replicate used in the exon array was analysed using Q-PCR. Q-PCR was chosen as the validation method due to its ability to detect small changes in fold change and its excellent target specificity (section 2.9.5).

For validation, eight genes were selected from the list of genes differentially expressed between siZNF-treated and siGAP-treated samples based on: 1) the magnitude of fold change, where a large fold change was preferred; 2) the number of known transcript variants, where a single transcript variant was preferred; and 3) the DNA sequence, where sequences with 40-60% GC content which were amenable to Q-PCR were preferred. The array data was considered verified if the Q-PCR results demonstrated a statistically significant difference in the abundance of mRNA between the pooled siZNF-treated samples and both of the control samples, with the fold change in the same direction as the array data. Of the eight genes selected for validation, all eight were verified as true results (Figure 4.8A). For example, *NPY*, *chemokine* (*C-C motif*) *ligand 2* (*CCL2*), *secreted protein acidic and rich in cysteine* (*SPARC*), *transmembrane protein with EGF-like and two follistatin-like domains 2* (*TMEFF2*) and *early growth response 1* (*EGR1*) showed up-regulation after *ZNF804A* knockdown while *pyruvate dehydrogenase lipoamide kinase isozyme 1* (*PDK1*), *follistatin-like 4* (*FSTL4*), *ephrin-B2* (*EFNB2*) showed down-regulation after *ZNF804A* knockdown. Comparison of the Q-PCR data with the exon array data showed that the magnitude of fold change calculated by Q-PCR was very similar to that calculated by exon array analysis

A


SPARC
CCL2
EGR1
TMEFF2
NPY
EFNB2
FSTL4
PDK1
% abundance relative to siGAP
pooled siZNF
siGAP
mock
*
**

B

Gene symbol	Exon array		Q-PCR	
	P value	FC	P value	FC
SPARC	1.25E-04	2.29	>0.001	2.32
FSTL4	8.32E-03	-2.06	6.00E-03	-2.70
NPY	8.14E-06	2.05	1.00E-03	3.31
PDK1	5.28E-03	-2.04	>0.001	-2.13
EGR1	4.49E-04	2.65	2.00E-03	3.68
TMEFF2	6.80E-04	2.37	>0.001	2.80
CCL2	4.96E-03	3.64	1.20E-02	6.29
EFNB2	3.87E-06	-2.14	3.00E-03	-1.90

Figure 4.8 Q-PCR assessment of gene expression changes after *ZNF804A* knockdown

To assess a selection of the gene expression changes using Q-PCR, RNA from one biological replicate used for the exon array was treated with DNase and cDNA was prepared. The cDNA from each condition was run by Q-PCR in triplicate for each primer set (primer sequences are given in Appendix 1.4). The mean average raw C_t value was normalised to *ACTB* and the data was analysed using the ΔC_t method. The data from the siZNFA- and siZNFB-treated sample was pooled to reflect the exon array analysis. **(A)** Bar graphs presenting the target mRNA abundance as a percentage of the target mRNA abundance in the siGAP-treated sample. The error bars represent the standard deviation of the raw C_t values (three technical replicates for each condition). **(B)** The statistical analysis was carried out on the ΔC_t values using one-way ANOVA and Tukey post-hoc. The fold change was calculated by dividing the siGAP-treated $\Delta\Delta C_t$ value by the pooled siZNF-treated $\Delta\Delta C_t$ value. When this number was less than one, indicating a negative fold change, the reciprocal fold change is listed. The summary statistics from the exon array are listed for comparison. * = $P < 0.05$ ** = $P < 0.001$ compared to both siGAP-treated and mock samples. (FC = fold change.)

(Figure 4.8B). These data imply that the exon array analysis was able to accurately determine changes in gene expression in *ZNF804A*-depleted cells.

4.6. Investigating changes in pre-mRNA splicing after *ZNF804A* knockdown

Since the advent of exon arrays, several algorithms have been proposed to identify alternative splicing events. These algorithms include the alternative splicing ANOVA in the PGS, the microarray detection of alternative splicing (MIDAS) (Affymetrix, 2006), the splicing index (SI) (Srinivasan et al., 2005) and the microarray analysis of differential splicing (MADS) (Xing et al., 2008). In this Chapter, two analysis approaches using the statistical tests available in the PGS were employed (section 2.9.6). In both approaches, the alternative splicing one-way ANOVA was performed on probe set $\log_2$ signal intensities using only probe set $\log_2$ signal intensities which were greater than three to ensure the analysis included only probe sets detected in at least one of the sample groups (Whistler, 2010). The output of the alternative splicing one-way ANOVA was displayed in the PGS both at the gene-level and at the probe set-level. This meant that each gene was attributed an alternative splicing P value and each individual probe set was attributed a P value for differential expression. Inherently, gene-level analyses of alternative splicing take into consideration any differences in gene-wide expression which may influence discovery of true changes in splicing. Therefore, most published papers present and validate only the results of gene-level analyses. However, the literature suggests that analysis of individual probe set $\log_2$ signal intensities at the exon-level (equivalent to the probe-set level) may also identify true alternative splicing events (Laajala et al., 2009; Warzecha et al., 2009). Therefore, both the gene-level (section 4.6.1) and the probe set-level (section 4.6.2) output derived from the alternative splicing one-way ANOVA were used to identify alternative splicing events.

4.6.1. Analysing alternative splicing after *ZNF804A* knockdown using the alternative splicing one-way ANOVA gene-level output

To reduce the number of false-positive results, the gene list was filtered to exclude any gene with less than five probe sets or gene expression changes greater than five-fold and a conservative Bonferroni correction for multiple testing was applied. The Bonferroni corrected P values are presented here. After Bonferroni correction, an arbitrary cut-off of $P < 1 \times 10^{-6}$ was chosen to generate a manageable list of statistically significant alternatively spliced genes (Whistler, 2010). This analysis revealed the splicing of 116 genes was altered after *ZNF804A* knockdown, relative to siGAP-treated samples (Table 4.5). The geneviews of the 116 alternatively spliced genes detailed in Table 4.5 were viewed alongside the gene information in the UCSC genome browser to determine if the statistically significant events resembled true changes in splicing (Langer et al., 2010; Whistler, 2010). This analysis is presented as supplementary data in Appendix 5.3. Many of the 116 statistically significant changes in splicing corresponded to complex changes in transcript usage which would be difficult to evaluate empirically. However, 17 of the events corresponded to putative alternative splicing of a cassette exon. Three of these cassette exons corresponded to known alternative splicing events. Thus, these three events were the focus of the confirmation analyses (see below).

Signal induced proliferation-associated 1 like 1 (SIPA1L1) was identified as statistically significantly alternatively spliced (Bonferroni corrected $P = 2.57 \times 10^{-11}$; alternative splicing one-way ANOVA). The geneview for *SIPA1L1* showed one probe set had a larger, negative fold change following *ZNF804A* knockdown than the other probe sets (Figure 4.9A). This probe set corresponded to exon 13a (chr14:72171438-72171500) of *SIPA1L1* which is known to be alternatively spliced (Figure 4.9B). Therefore, RT-PCR was used to test the hypothesis

Gene name	Gene symbol	Gene expression P value	Bonferroni corrected alternative splicing P value	FC
calcium channel, voltage-dependent, T type, alpha 1G subunit	*CACNA1G*	3.43E-08	0.00E+00	-1.97
laminin, beta 2 (laminin S)	*LAMB2*	2.57E-05	1.98E-35	2.42
fms-related tyrosine kinase 1	*FLT1*	6.79E-05	5.39E-34	-2.84
notch 3	*NOTCH3*	8.22E-08	1.59E-33	-2.02
G protein-coupled receptor 64	*GPR64*	2.75E-04	1.57E-30	3.19
Rho GTPase activating protein 36	*ARHGAP36*	3.36E-05	3.18E-30	3.29
WNK lysine deficient protein kinase 1	*WNK1*	6.48E-05	3.54E-30	-1.29
myosin X	*MYO10*	6.64E-06	5.39E-30	-1.90
slit homolog 1 (Drosophila)	*SLIT1*	8.62E-05	9.38E-26	-1.52
alanyl (membrane) aminopeptidase	*ANPEP*	1.08E-09	1.51E-25	-2.63
zinc finger protein 804A	*ZNF804A*	2.13E-08	1.98E-22	-2.01
ATP-binding cassette, sub-family C (CFTR/MRP), member 3	*ABCC3*	1.03E-03	4.00E-22	1.45
reelin	*RELN*	1.03E-05	9.25E-22	3.91
protein tyrosine phosphatase, receptor type, E	*PTPRE*	2.08E-03	3.22E-20	2.16
LY6/PLAUR domain containing 1	*LYPD1*	2.28E-06	3.06E-19	3.56
protein tyrosine phosphatase, receptor type, R	*PTPRR*	1.38E-01	3.32E-19	-1.17
transforming, acidic coiled-coil containing protein 2	*TACC2*	1.80E-03	1.17E-18	-1.21
thrombospondin, type I, domain containing 4	*THSD4*	1.94E-04	4.01E-17	2.05
sestrin 3	*SESN3*	2.77E-07	8.38E-17	-3.11
leucine rich repeat containing 7	*LRRC7*	3.32E-06	1.28E-16	1.73
glutamate decarboxylase 1 (brain, 67kDa)	*GAD1*	2.53E-01	4.18E-16	-1.07
integrin, alpha 5	*ITGA5*	1.18E-02	4.81E-15	-1.20
decorin	*DCN*	2.09E-05	6.16E-15	-3.11
ATP-binding cassette, sub-family C (CFTR/MRP), member 5	*ABCC5*	5.57E-01	1.79E-14	-1.06
family with sequence similarity 184, member A	*FAM184A*	8.33E-05	2.35E-14	-1.31
PALM2-AKAP2 readthrough	*PALM2-AKAP2*	1.38E-01	3.58E-14	-1.11
protocadherin gamma subfamily C, 5	*PCDHGC5*	2.51E-04	5.98E-14	1.23
polymerase (DNA directed), theta	*POLQ*	2.54E-04	1.35E-13	-1.42
elastin	*ELN*	1.16E-05	3.10E-13	1.45
laminin, alpha 1	*LAMA1*	5.49E-02	4.22E-13	-1.17
T-box 3	*TBX3*	3.83E-04	5.35E-13	2.30
ATPase, Ca++ transporting, ubiquitous	*ATP2A3*	4.37E-04	6.40E-13	-1.18
U-box domain containing 5	*UBOX5*	4.88E-04	8.14E-13	-1.25
kinesin family member 21A	*KIF21A*	5.66E-03	8.32E-13	-1.41
retinoic acid induced 14	*RAI14*	4.15E-04	9.43E-13	-2.51
B-cell CLL/lymphoma 11A (zinc finger protein)	*BCL11A*	1.53E-05	9.80E-13	-1.62
caspase 4, apoptosis-related cysteine peptidase	*CASP4*	2.91E-06	1.07E-12	1.80

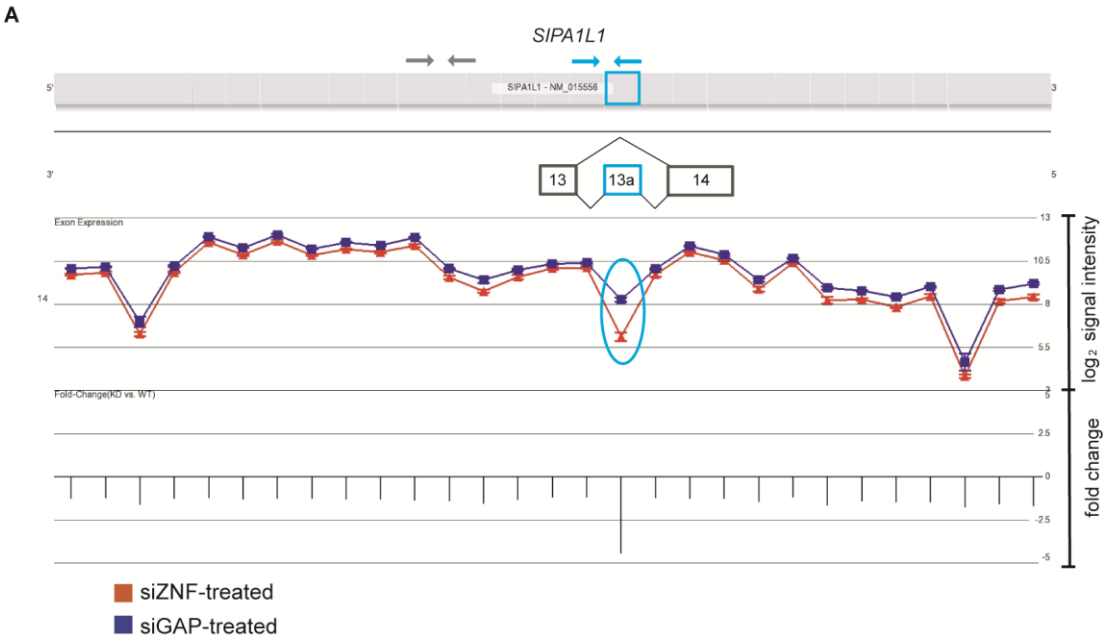
G protein-coupled receptor 162	*GPR162*	1.00E+00	1.16E-12	-1.00
vasoactive intestinal peptide	*VIP*	3.52E-07	1.26E-12	3.15
reticulon 4	*RTN4*	9.15E-01	1.50E-12	1.01
aspartate beta-hydroxylase	*ASPH*	1.84E-05	2.29E-12	-1.79
fibroblast growth factor receptor-like 1	*FGFRL1*	1.77E-05	3.92E-12	-1.62
matrilin 2	*MATN2*	8.59E-03	4.83E-12	2.27
immunoglobulin superfamily, member 1	*IGSF1*	3.65E-06	6.02E-12	2.47
topoisomerase (DNA) II alpha 170kDa	*TOP2A*	1.77E-04	8.31E-12	-1.99
Rho GTPase activating protein 28	*ARHGAP28*	2.55E-03	8.35E-12	-2.20
chimerin (chimaerin) 1	*CHN1*	5.03E-01	1.03E-11	-1.12
pre-B-cell leukemia homeobox interacting protein 1	*PBXIP1*	3.67E-03	2.35E-11	-1.63
signal-induced proliferation-associated 1 like 1	*SIPA1L1*	2.46E-04	2.57E-11	-1.43
complement factor H	*CFH*	9.35E-02	2.59E-11	1.20
lipid phosphate phosphatase-related protein type 3	*LPPR3*	5.19E-05	2.89E-11	-1.31
immunoglobulin superfamily, member 9	*IGSF9*	1.44E-05	3.36E-11	-1.32
signal-induced proliferation-associated 1	*SIPA1*	4.74E-05	4.59E-11	-1.39
RNA binding protein with multiple splicing	*RBPMS*	6.73E-04	1.03E-10	-1.77
collagen, type XXVII, alpha 1	*COL27A1*	4.09E-01	1.04E-10	1.06
active BCR-related gene	*ABR*	6.56E-02	1.19E-10	1.18
phosphodiesterase 11A	*PDE11A*	3.68E-05	1.77E-10	1.74
CD44 molecule (Indian blood group)	*CD44*	2.51E-03	2.04E-10	1.83
tissue factor pathway inhibitor 2	*TFPI2*	1.54E-04	2.87E-10	1.72
serpin peptidase inhibitor, clade F	*SERPINF1*	2.38E-06	3.56E-10	-2.91
MAX interactor 1	*MXI1*	1.87E-06	3.87E-10	-1.55
matrix metallopeptidase 15	*MMP15*	1.12E-04	6.60E-10	2.01
E1A binding protein p400	*EP400*	5.87E-02	7.31E-10	-1.12
phosphatidylinositol-3,4,5-trisphosphate-dependent Rac exchange factor 1	*PREX1*	2.16E-06	1.13E-09	-1.22
neuron navigator 2	*NAV2*	3.51E-02	1.21E-09	-1.30
tight junction protein 2 (zona occludens 2)	*TJP2*	2.81E-03	1.48E-09	-1.76
nuclear transcription factor Y, alpha	*NFYA*	9.45E-03	1.50E-09	-1.39
BR serine/threonine kinase 2	*BRSK2*	2.71E-05	1.64E-09	-1.13
ryanodine receptor 1 (skeletal)	*RYR1*	3.86E-02	2.03E-09	1.13
WD repeat and SOCS box-containing 1	*WSB1*	7.90E-01	2.43E-09	-1.02
C1q and tumor necrosis factor related protein 6	*C1QTNF6*	1.10E-04	3.58E-09	1.84
Sema domain, immunoglobulin domain (Ig), short basic domain, secreted, (semaphorin) 3F	*SEMA3F*	1.86E-03	3.72E-09	-1.19
metastasis suppressor 1	*MTSS1*	4.15E-05	4.49E-09	-1.32
cytochrome b5 reductase 2	*CYB5R2*	2.82E-03	5.13E-09	-1.22
transcription elongation factor A (SII), 2	*TCEA2*	3.66E-03	6.25E-09	-1.52

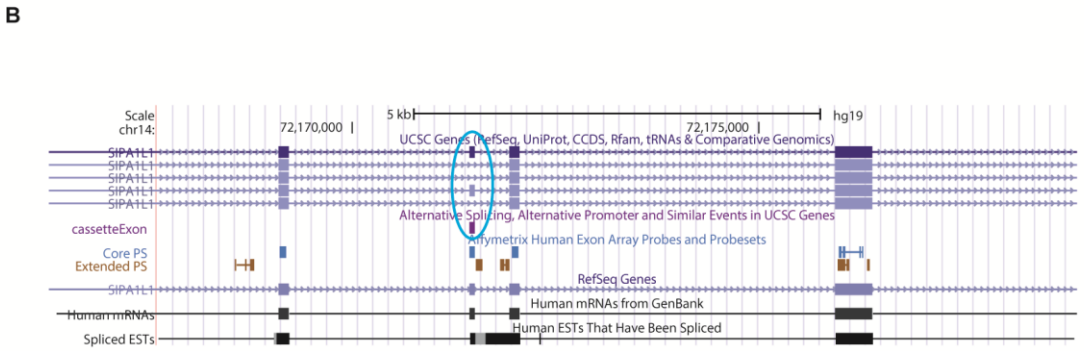
microtubule associated serine/threonine kinase 1	*MAST1*	4.29E-04	6.39E-09	-1.24
LIM homeobox 8	*LHX8*	1.90E-04	7.41E-09	-1.30
pyruvate kinase, muscle	*PKM2*	2.30E-03	7.76E-09	-1.23
collagen, type IV, alpha 2	*COL4A2*	1.90E-03	8.55E-09	-1.27
beta-1,4-N-acetyl-galactosaminyl transferase 4	*B4GALNT4*	1.03E-05	1.60E-08	-1.28
sphingomyelin phosphodiesterase 3, neutral membrane (neutral sphingomyelinase II)	*SMPD3*	6.77E-07	1.64E-08	-2.32
extra spindle pole bodies homolog 1 (S. cerevisiae)	*ESPL1*	5.57E-03	1.68E-08	-1.36
KIAA0649	*KIAA0649*	1.07E-05	2.45E-08	-1.27
transmembrane protein with EGF-like and two follistatin-like domains 2	*TMEFF2*	6.80E-06	3.08E-08	2.37
ankyrin repeat domain 12	*ANKRD12*	3.65E-04	3.13E-08	-1.55
casein kinase 2, alpha prime polypeptide	*CSNK2A2*	4.45E-06	3.25E-08	-1.39
glutamine-fructose-6-phosphate transaminase 2	*GFPT2*	3.18E-04	3.47E-08	-1.39
prostaglandin F receptor (FP)	*PTGFR*	6.88E-05	3.65E-08	1.72
zinc finger protein 395	*ZNF395*	5.56E-06	3.68E-08	-1.44
astrotactin 1	*ASTN1*	8.86E-04	4.56E-08	1.75
anoctamin 3	*ANO3*	1.69E-04	5.14E-08	1.56
platelet-derived growth factor receptor, alpha polypeptide	*PDGFRA*	5.34E-03	5.20E-08	-1.27
family with sequence similarity 190, member B	*FAM190B*	1.86E-05	5.60E-08	-1.47
sulfatase 2	*SULF2*	1.30E-04	7.43E-08	2.02
SRSF protein kinase 2	*SRPK2*	8.56E-04	7.61E-08	-1.53
chromosome 1 open reading frame 85	*C1orf85*	4.74E-03	9.78E-08	-1.40
cellular retinoic acid binding protein 2	*CRABP2*	1.04E-04	1.01E-07	-4.04
tripartite motif-containing 9	*TRIM9*	8.56E-05	1.14E-07	-1.25
plexin D1	*PLXND1*	5.28E-01	1.23E-07	-1.03
DENN/MADD domain containing 5B	*DENND5B*	3.63E-05	1.26E-07	-1.61
potassium large conductance calcium-activated channel, subfamily M, alpha member 1	*KCNMA1*	3.08E-04	1.43E-07	1.42
regulator of G-protein signalling 9	*RGS9*	1.17E-06	2.35E-07	1.80
cadherin 2, type 1, N-cadherin (neuronal)	*CDH2*	4.58E-04	2.66E-07	-1.44
EPH receptor B2	*EPHB2*	5.25E-06	2.78E-07	-1.66
protein kinase C and casein kinase substrate in neurons 3	*PACSIN3*	6.98E-04	2.89E-07	-1.82
potassium channel tetramerisation domain containing 12	*KCTD12*	4.44E-04	3.65E-07	1.74
DIX domain containing 1	*DIXDC1*	1.52E-04	3.81E-07	-1.62
small nuclear RNA activating complex, polypeptide 3, 50kDa	*SNAPC3*	2.02E-03	3.95E-07	-1.40
insulin-like growth factor 2 (somatomedin A)	*IGF2*	1.86E-03	5.23E-07	-1.83
spectrin repeat containing, nuclear envelope 1	*SYNE1*	7.72E-04	6.20E-07	1.26
coiled-coil domain containing 50	*CCDC50*	4.15E-06	6.56E-07	-1.73

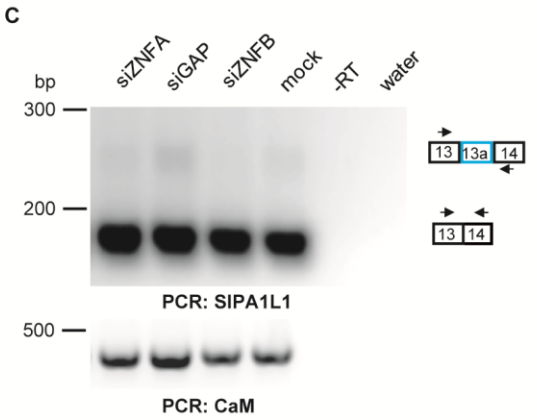
spondin 1, extracellular matrix protein	*SPON1*	2.71E-02	7.09E-07	1.30
WEE1 homolog (S. pombe)	*WEE1*	1.24E-04	7.10E-07	-1.37
chromosome 9 open reading frame 86	*C9orf86*	1.14E-01	8.32E-07	1.07
brain-derived neurotrophic factor	*BDNF*	1.97E-03	8.37E-07	1.37
chromosome 20 open reading frame 117	*C20orf117*	3.78E-01	9.51E-07	-1.04

Table 4.5 The genes showing alternative splicing after *ZNF804A* knockdown

The .CEL files were imported into the PGS using the core metaprobe set and RMA normalisation. The .CEL files were assigned as either siZNF- or siGAP-treatment. To identify alternatively spliced transcripts, an alternative splicing one-way ANOVA was performed on the probe set $\log_2$ signal intensities greater than three. Treatment was chosen as the candidate variable in the ANOVA model. The gene list was filtered to exclude genes containing fewer than five markers or a fold change greater than five. Alternative splicing ANOVA P values were corrected using the conservative Bonferroni method. After Bonferroni correction, an arbitrary cut-off of $P < 1 \times 10^{-6}$ was chosen to generate a manageable list of statistically significant alternatively spliced genes. Using this threshold, the splicing of 116 genes was altered after *ZNF804A* knockdown. (FC = fold change.)


A
SIPA1L1
SIPA1L1 - NM_015556
13
13a
14
Exon Expression
log2 signal intensity
Fold-Change(KD vs. WT)
fold change
siZNF-treated
siGAP-treated


B
Scale
chr14:
72,170,000
5 kb
72,175,000
hg19
UCSC Genes (RefSeq, UniProt, CCDS, Rfam, tRNAs & Comparative Genomics)
SIPA1L1
Alternative Splicing, Alternative Promoter and Similar Events in UCSC Genes
cassetteExon
Core PS
Extended PS
Affymetrix Human Exon Array Probes and Probesets
RefSeq Genes
Human mRNAs from GenBank
Human ESTs That Have Been Spliced
Spliced ESTs


C
bp
siZNFA
siGAP
siZNFB
mock
-RT
water
300
200
500
13
13a
14
PCR: SIPA1L1
PCR: CaM

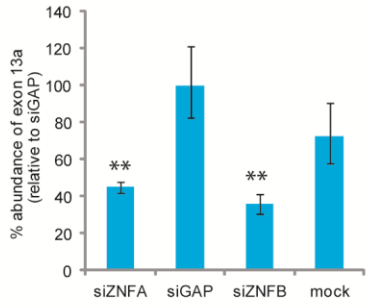

D
% abundance of exon 13a
(relative to siGAP)
140
120
100
80
60
40
20
0
**
**
siZNFA
siGAP
siZNFB
mock

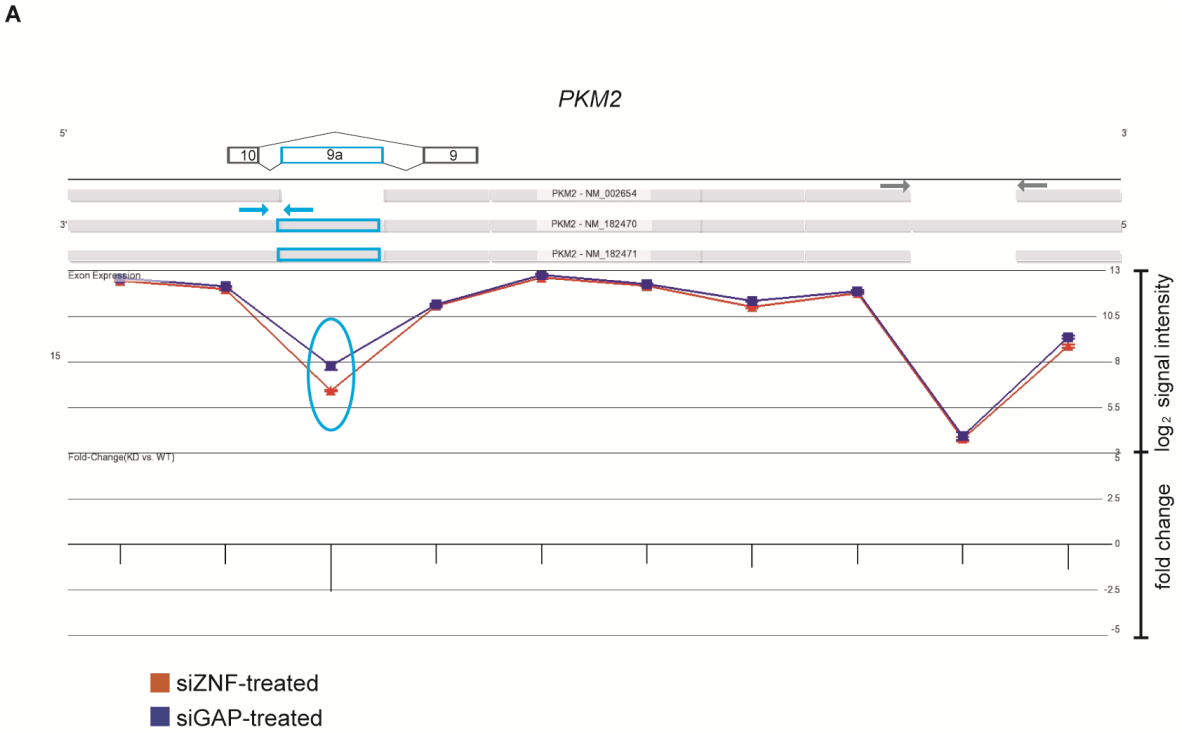
Figure 4.9 Exon 13a of *SIPA1L1* was alternatively spliced when *ZNF804A* was depleted

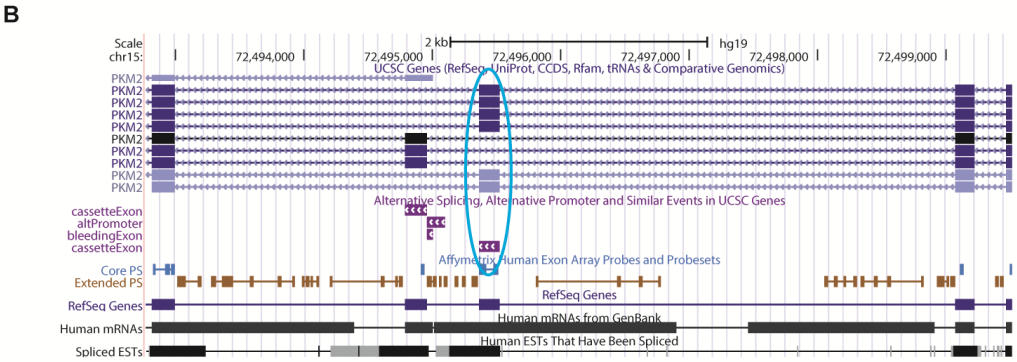
SIPA1L1 was identified as alternatively spliced (Bonferroni corrected $P = 2.57 \times 10^{-11}$; alternative splicing one-way ANOVA). The splicing event was confirmed empirically. **(A & B)** The geneview of *SIPA1L1* showed increased exclusion of the known cassette exon 13a when *ZNF804A* was depleted. **(C)** RT-PCR using primers complementary to the constitutive exons flanking exon 13a. *Calmodulin* (*CaM*) was used as a loading control. **(D)** Q-PCR using primers complementary to exon 13a (blue arrows) and a 'control' pair complementary to another region on the transcript (grey arrows). Please note, one biological replicate used in the exon array was analysed using Q-PCR. The expression of exon 13a was compared between the conditions using the $\Delta\Delta C_t$ method with the control primer set as the endogenous normaliser. The bar graph presented shows the percentage abundance of exon 13a in each condition relative to its abundance in the siGAP-treated sample. Q-PCR was performed in triplicate (three technical replicates) for each condition. The error bars represent the standard deviation of the three raw C_t values. The significance was assessed using one-way ANOVA, Tukey post-hoc. ** $P < 0.05$.

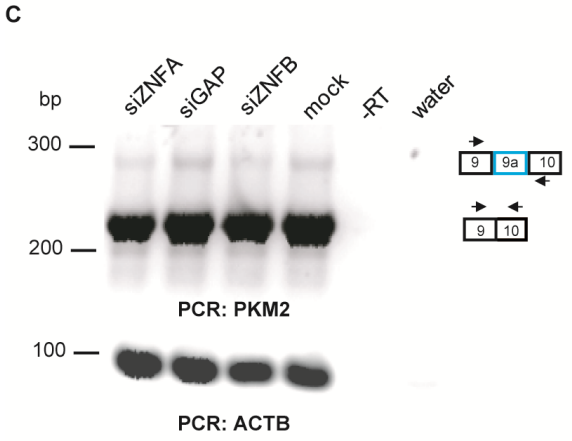
that there was increased exclusion of exon 13a of *SIPA1L1* in *ZNF804A*-depleted cells (section 2.9.7). RT-PCR showed that there was less of the longer transcript of *SIPA1L1* (the transcript including exon 13a) in siZNF-treated samples (Figure 4.9C). Q-PCR showed that there was statistically significantly less of exon 13a present after *ZNF804A* knockdown (Figure 4.9D). These data were consistent with the exon array analysis. These data suggest that knockdown of *ZNF804A* altered the splicing of exon 13a of *SIPA1L1*.

Figure 4.10A shows that there was increased exclusion of exon 9a of *pyruvate kinase, muscle* (*PKM2*) (chr15: 72495363-72495529) after *ZNF804A* knockdown (Bonferroni corrected P $=7.76 \times 10^{-9}$; alternative splicing one-way ANOVA). Exon 9a of *PKM2* is known to be alternatively spliced (Figure 4.10B). RT-PCR analysis showed less of the longer transcript of *PKM2* was present in the siZNF-treated samples compared to the controls (Figure 4.10C). While this result was consistent with the exon array data, subsequent Q-PCR analysis showed that there was no significant difference in the abundance of exon 9a between the siZNF-treated samples and the siGAP-treated sample (Figure 4.10D). Therefore, the alternative splicing of exon 9a of *PKM2* after *ZNF804A* knockdown was not validated by the empirical analysis.

The third gene containing an alternatively spliced cassette exon was *nuclear transcription factor Y, alpha* (*NFYA*) (Bonferroni corrected $P = 1.5 \times 10^{-9}$; alternative splicing one-way ANOVA). The geneview of *NFYA* showed exon 1a (chr6:41048550-41048636) was excluded significantly more after *ZNF804A* knockdown (Figures 4.11A and B). RT-PCR showed that there was a reduced amount of the longer transcript and a reciprocal increase in the amount of the shorter transcript following *ZNF804A* knockdown (Figure 4.11C). This was consistent


A
PKM2
10
9a
9
PKM2 - NM_002654
PKM2 - NM_182470
PKM2 - NM_182471
Exon Expression
Fold-Change(KD vs. WT)
log2 signal intensity
fold change
siZNF-treated
siGAP-treated


B
Scale
chr15:
2 kb
hg19
72,494,000
72,495,000
72,496,000
72,497,000
72,498,000
72,499,000
UCSC Genes (RefSeq, UniProt, CCDS, Rfam, tRNAs & Comparative Genomics)
PKM2
Alternative Splicing, Alternative Promoter and Similar Events in UCSC Genes
cassetteExon
altPromoter
bleedingExon
cassetteExon
Affymetrix Human Exon Array Probes and Probesets
Core PS
Extended PS
RefSeq Genes
Human mRNAs from GenBank
Human mRNAs
Human ESTs That Have Been Spliced
Spliced ESTs


C
bp
siZNFA
siGAP
siZNFB
mock
-RT
water
300
200
100
9
9a
10
PCR: PKM2
PCR: ACTB

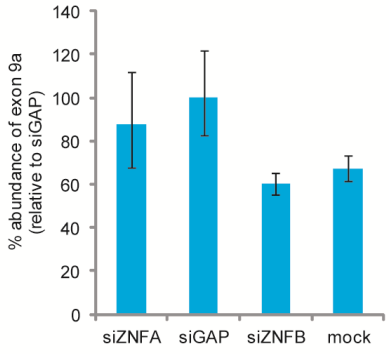
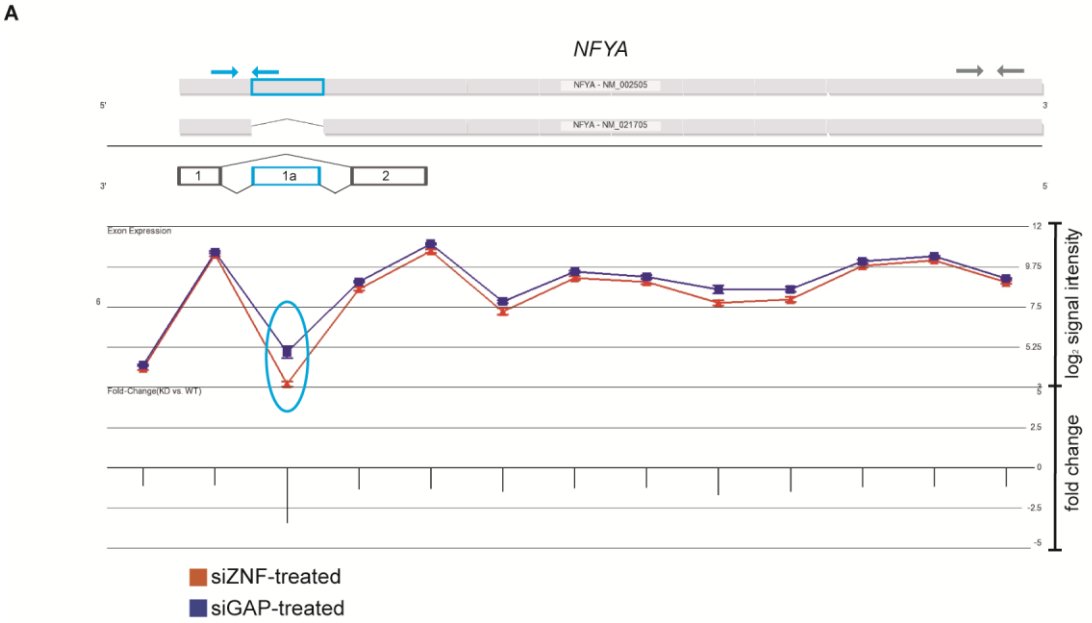
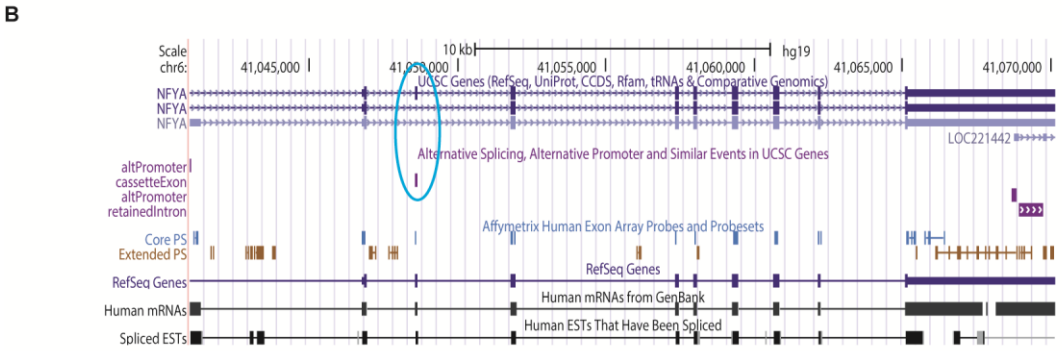
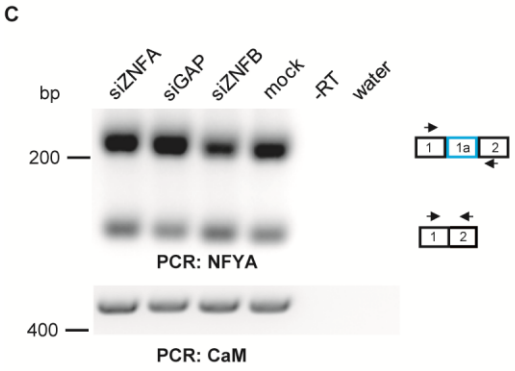

D
% abundance of exon 9a
(relative to siGAP)
140
120
100
80
60
40
20
0
siZNFA
siGAP
siZNFB
mock

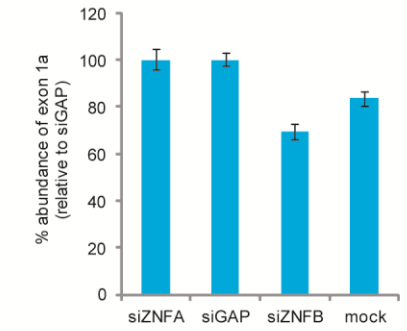
Figure 4.10 The empirical assessment of alternative splicing of exon 9a in *PKM2*

PKM2 was identified as alternatively spliced (Bonferroni corrected P = 7.76 x 10^{-9}; alternative splicing one-way ANOVA) however; this splicing event was not confirmed empirically. **(A & B)** The *PKM2* geneview suggested increased exclusion of a known cassette exon 9a when *ZNF804A* was depleted. **(C)** RT-PCR using primers complementary to the constitutive exons flanking exon 9a. *Beta actin* (*ACTB*) was used as a loading control. **(D)** Q-PCR using primers complementary to exon 9a (blue arrows) and a 'control' pair complementary to amplify another region on the transcript (grey arrows). Please note, one biological replicate used in the exon array was analysed using Q-PCR. The expression of exon 9a between the conditions was assessed using the $\Delta\Delta C_t$ method with the control primer set as the endogenous normaliser. The bar graph presented shows the percentage abundance of exon 9a in each condition relative to its abundance in the siGAP-treated sample. Q-PCR was performed in triplicate (three technical replicates) for each condition. The error bars represent the standard deviation of the three raw C_t values.


A
NFYA
NFYA - NM_002505
NFYA - NM_021705
1
1a
2
Exon Expression
Fold-Change(KD vs. WT)
log2 signal intensity
fold change
siZNF-treated
siGAP-treated


B
Scale
chr6:
10 kb
hg19
41,045,000
41,050,000
41,055,000
41,060,000
41,065,000
41,070,000
UCSC Genes (RefSeq, UniProt, CCDS, Rfam, tRNAs & Comparative Genomics)
NFYA
NFYA
NFYA
LOC221442
Alternative Splicing, Alternative Promoter and Similar Events in UCSC Genes
altPromoter
cassetteExon
altPromoter
retainedIntron
Affymetrix Human Exon Array Probes and Probesets
Core PS
Extended PS
RefSeq Genes
RefSeq Genes
Human mRNAs
Human mRNAs from GenBank
Spliced ESTs
Human ESTs That Have Been Spliced


C
bp
siZNFA
siGAP
siZNFB
mock
-RT
water
200
400
1
1a
2
PCR: NFYA
PCR: CaM


D
% abundance of exon 1a
(relative to siGAP)
120
100
80
60
40
20
0
siZNFA
siGAP
siZNFB
mock

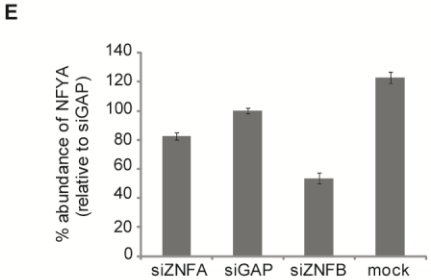
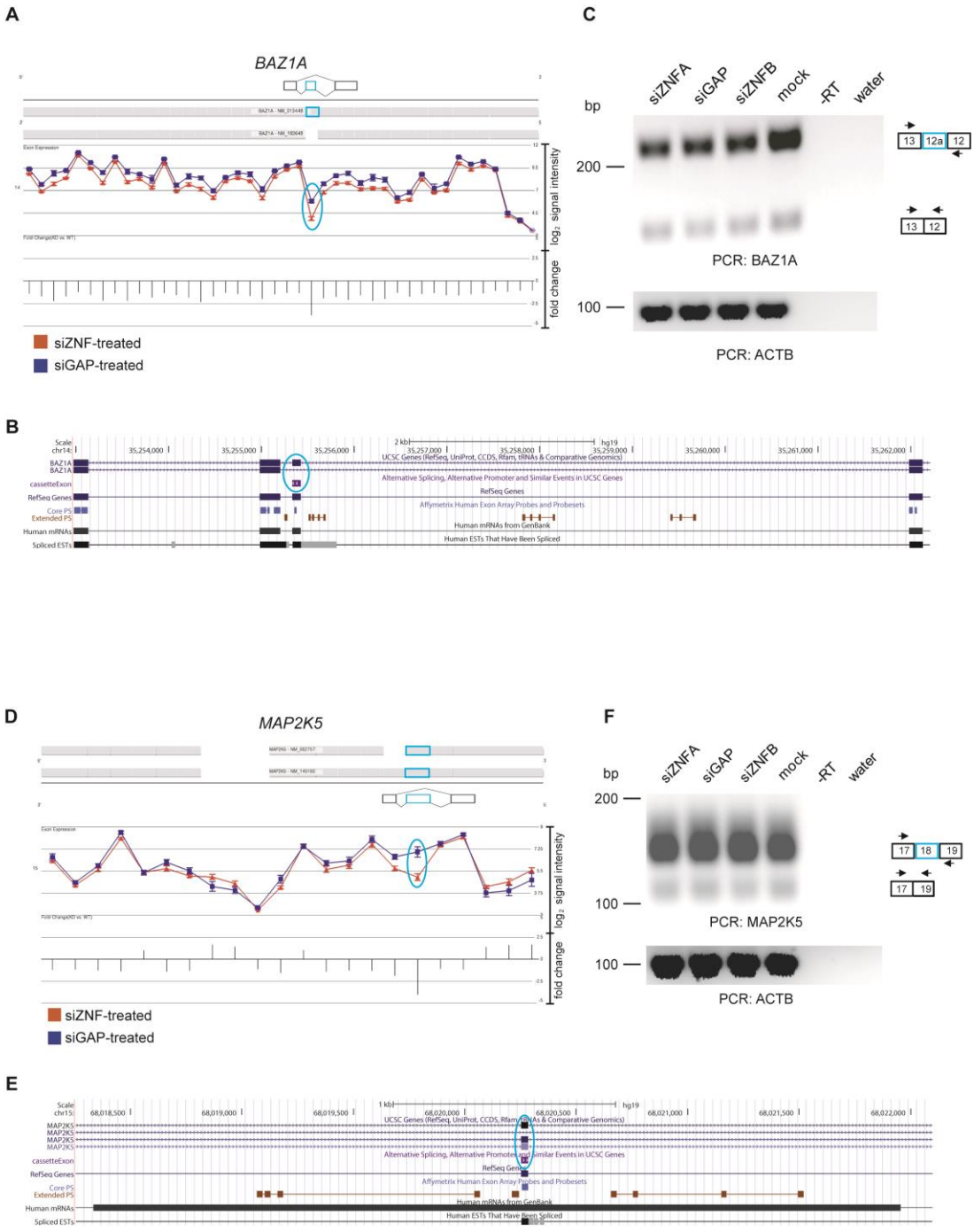

E
% abundance of NFYA
(relative to siGAP)
140
120
100
80
60
40
20
0
siZNFA
siGAP
siZNFB
mock

Figure 4.11 Validation of splicing of a cassette exon in *NFYA*

NFYA was identified as alternatively spliced (Bonferroni corrected $P = 1.5 \times 10^{-9}$; alternative splicing one-way ANOVA) however; this splicing event was not confirmed empirically. **(A & B)** In *ZNF804A*-depleted cells, there was increased exclusion of the known cassette exon 1a in *NFYA* **(C)** RT-PCR using primers complementary to the constitutive exons flanking exon 1a. *Calmodulin* (*CaM*) was used as a loading control. **(D)** Q-PCR using primers complementary to exon 1a (blue arrows) and a 'control' pair complementary to another region on the transcript (grey arrows). Please note, one biological replicate used in the exon array was analysed using Q-PCR. The expression of exon 1a was compared between the conditions using the control primer set as the endogenous normaliser in the $\Delta\Delta C_t$ method. The bar graph presented shows the percentage abundance of exon 1a in each condition relative to its abundance in the siGAP-treated sample. **(E)** Q-PCR using the 'control' pair of *NFYA* primers. Please note, one biological replicate used in the exon array was analysed using Q-PCR. The expression of *NFYA* was compared between the conditions using the $\Delta\Delta C_t$ method with *beta actin* (*ACTB*) set as the endogenous normaliser. The bar graph presented shows the percentage abundance of *NFYA* relative in each condition relative to its abundance in the siGAP-treated sample. Q-PCR was performed in triplicate (three technical replicates) for each condition. The error bars represent the standard deviation of the three raw C_t values.

with increased exclusion of exon 1a when *ZNF804A* was depleted. Q-PCR showed that there was less of exon 1a in the siZNFB-treated sample, relative to the siGAP-treated and mock samples (Figure 4.11D). However, there was no difference in the amount of exon 1a of *NFYA* in siZNFA-treated sample. Further inspection of the Q-PCR data showed that the expression of *NFYA* transcript was decreased after *ZNF804A* knockdown relative to the controls (Figure 4.11E). As the level of exon 1a was normalised to the amount of total *NFYA* transcript in each of the samples, this may explain the discrepancy between the exon array and Q-PCR data. In summary, the alternative splicing of exon 1a of *NFYA* was not confirmed.

The visual inspection of the PGS geneviews revealed very few of the 116 alternative splicing events corresponded to known cassette exons in the UCSC genome browser. Consequently, the number of events which were selected for empirical validation was restricted. To increase the number of putative alternative splicing events evaluated alongside the UCSC genome browser, a less stringent significance cut-off of $P < 0.05$ rather than $P < 1 \times 10^{-6}$ was used. This gave a list of 448 genes which were putatively alternatively spliced. As previously, the geneviews were assessed alongside the UCSC genome browser. Three genes containing cassette exons that were putatively alternatively spliced in *ZNF804A*-depleted cells were selected. These genes, and the corresponding alternatively spliced exons, were: *bromodomain adjacent to zinc finger domain, 1A* (*BAZ1A*; exon 12a; chr14:35255332-35255427); *mitogen-activated protein kinase kinase 5* (*MAP2K5*; exon 18; chr15:68040569-68040595); and *REST corepressor 2* (*RCOR2*; exon 11; chr11:63679777-63680006). An alternative 5' prime splice site (chr17:75137137-75137189) in *SEC14-like 1 (S. cerevisiae*) (*SEC14L1*) was also chosen for validation. RT-PCR primers were designed to amplify the transcript variants of each gene (primer sequences are given in Appendix 1.7). Data presented in Figure 4.12 show knockdown of *ZNF804A* led to putative increased exclusion of known alternatively spliced


A
BAZ1A
log2 signal intensity
fold change
siZNF-treated
siGAP-treated
C
siZNFA
siGAP
siZNFB
mock
-RT
water
bp
200
100
PCR: BAZ1A
PCR: ACTB
B
UCSC Genes (RefSeq, UniProt, CCDS, Rfam, tRNAs & Comparative Genomics)
Alternative Splicing, Alternative Promoter and Similar Events in UCSC Genes
RefSeq Genes
Affymetrix Human Exon Array Probes and Probesets
Human mRNAs from GenBank
Human ESTs That Have Been Spliced
D
MAP2K5
E
F
PCR: MAP2K5

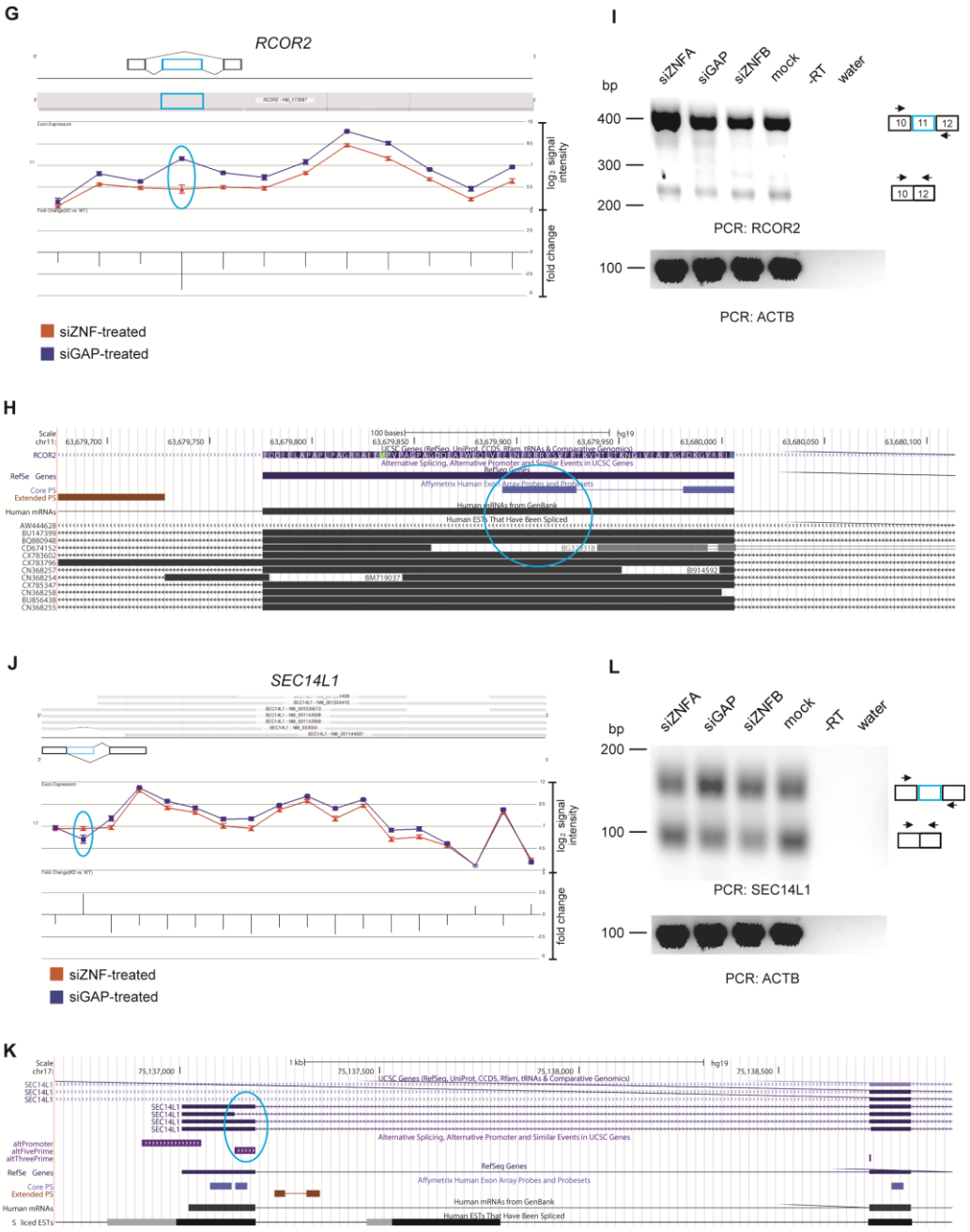


Figure 4.12 RT-PCR assessment of alternative splicing events

BAZ1A, *MAP2K5*, *RCOR2* and *SEC14L1* were identified as alternatively spliced after *ZNF804A* knockdown (alternative splicing one-way ANOVA; Bonferroni corrected $P < 0.05$). Manually annotation of the geneviews alongside the UCSC genome browser showed *BAZ1A* (**A**), *MAP2K5* (**D**) and *RCOR2* (**G**) had increased exclusion of a known cassette exon (**B**; **E**; **H** respectively). The *SEC14L1* geneview (**J**) showed alternative 5' splice site usage (**K**). RT-PCR primers were designed to amplify the constitutive exons flanking the spliced exon. None of the empirical validations was conclusive (**C**; **F**; **I**; **L**).

exons in *BAZ1A*, *MAP2K5* and *RCOR2*. However, RT-PCR showed that there was no obvious difference in the transcript variant abundance in the siZNF-treated samples relative to siGAP-treated and mock samples for each of these genes (Figure 4.12). The geneview for *SEC14L1* showed an alternative 5' splice site was used; however, RT-PCR showed no difference in the abundance of the two transcript variants. Therefore, alternative splicing of these four transcripts was not validated, however further Q-PCR analysis of exon levels is required to conclude confidently that these are not true splicing events.

4.6.2. Analysing alternative splicing after *ZNF804A* knockdown using the alternative splicing one-way ANOVA probe set-level output

To identify changes in splicing using the $\log_2$ signal intensities for each individual probe set, the probe set-level list derived from the alternative splicing one-way ANOVA was filtered to retain probe sets with differential expression (FDR 0.05) and to remove any probe sets present in differentially expressed genes ($P > 0.05$). The $\log_2$ signal intensity of 566 individual probe sets differed in genes without differential expression after *ZNF804A* knockdown (Table 4.6). The probe sets were ranked according to fold change in $\log_2$ signal intensity. The top 30 positive and top 30 negative ranked probe sets were manually annotated alongside the UCSC genome browser; four alternative splicing events were selected for empirical confirmation: *enabled homolog* (*Drosophila*) (*ENAH*), *GTPase activating protein (SH3 domain*) *binding protein 2* (*G3BP2*), *ATPase, class VI, type 11C* (*ATP11C*) and *syntaxin binding protein 1* (*STXBP1*).

The geneview of *ENAH* showed that there was increased exclusion of the known alternatively spliced exon 11a (chr1:225692693-225692755) after *ZNF804A* knockdown ($P = 8.2 \times 10^{-4}$; FC = -2.69; one-way ANOVA), although the *ENAH* transcript was not differentially

A

Gene name	Gene Symbol	Probe set ID	Probe set		Gene	
			P value	FC	P value	FC
protease, serine, 22	*PRSS22*	3677272	4.11E-04	1.90	1.27E-01	1.08
guanylate cyclase activator 1A (retina)	*GUCA1A*	2907084	7.54E-04	1.93	5.72E-02	1.27
centrosomal protein 41kDa	*TSGA14*	3072581	2.76E-04	1.94	6.25E-01	1.04
mechanistic target of rapamycin (serine/threonine kinase)	*MTOR*	2396609	5.16E-05	1.96	1.09E-01	1.22
DDB1 and CUL4 associated factor 11	*DCAF11*	3529551	1.08E-03	1.96	1.16E-01	1.08
kinesin family member 13B	*KIF13B*	3129597	1.46E-04	1.97	4.96E-01	1.04
prolyl 4-hydroxylase, alpha polypeptide II	*P4HA2*	2875234	1.40E-04	1.98	1.20E-01	1.24
phosphoserine phosphatise	*PSPH*	3051872	3.07E-05	2.02	3.54E-01	1.09
TIMP metallopeptidase inhibitor 3	*TIMP3*	3943508	9.70E-04	2.02	1.05E-01	1.15
solute carrier family 35, member F3	*SLC35F3*	2386041	1.10E-04	2.03	1.59E-01	1.13
mevalonate (diphospho) decarboxylase	*MVD*	3704307	5.14E-04	2.03	4.91E-01	1.05
leucyl/cystinyl aminopeptidase	*LNPEP*	2821467	1.14E-03	2.17	5.31E-02	1.19
dual specificity phosphatase 19	*DUSP19*	2518732	4.72E-04	2.17	5.07E-01	-1.04
homeodomain interacting protein kinase 3	*HIPK3*	3325929	4.91E-05	2.28	5.78E-02	1.43
transmembrane channel-like 4	*TMC4*	3870551	1.17E-03	2.36	1.02E-01	1.16
protein tyrosine phosphatase, receptor type, R	*PTPRR*	3461943	4.44E-06	2.37	1.38E-01	-1.17
Smad nuclear interacting protein 1	*SNIP1*	2407184	6.12E-04	2.39	1.83E-01	1.10
active BCR-related	*ABR*	3740042	5.29E-05	2.40	6.56E-02	1.18
KDEL (Lys-Asp-Glu-Leu) endoplasmic reticulum protein retention receptor 3	*KDELR3*	3945319	5.42E-05	2.54	7.18E-02	1.52
trinucleotide repeat containing 6B	*TNRC6B*	3946327	2.02E-06	2.56	3.89E-01	1.08
salt-inducible kinase 1	*SIK1*	3934129	8.50E-04	2.57	1.07E-01	1.16
lipocalin 8	*LCN8*	3230419	3.03E-04	2.60	7.91E-02	1.11
ankyrin repeat domain 10	*ANKRD10*	3525703	4.69E-05	2.63	5.44E-02	-1.18
protein phosphatase 4, regulatory subunit 4	*PPP4R4*	3549607	1.13E-03	2.78	9.61E-02	1.12
fucosyltransferase 8 (alpha (1,6) fucosyltransferase)	*FUT8*	3540553	2.03E-04	2.80	5.53E-01	1.04
sialic acid binding Ig-like lectin 1, sialoadhesin	*SIGLEC1*	3895653	6.48E-07	3.01	5.85E-02	1.13
ataxin 3	*ATXN3*	3576899	2.25E-04	3.57	7.90E-02	1.29
G protein-coupled receptor 162	*GPR162*	3402829	6.60E-07	3.77	1.00E+00	-1.00
ATP-binding cassette, sub-family C (CFTR/MRP), member 5	*ABCC5*	2708339	4.59E-05	4.51	5.57E-01	-1.06
complement factor H	*CFH*	2373397	9.77E-04	5.63	9.62E-02	1.18

B

Gene name	Gene Symbol	Probe set ID	Probe set		Gene	
			P value	FC	P value	FC
mitochondrial calcium uptake 2	*EFHA1*	3504794	1.08E-03	-3.80	8.68E-02	-1.64
mastermind-like domain containing 1	*MAMLD1*	3994752	1.72E-04	-2.97	4.12E-01	-1.07
tetratricopeptide repeat domain 39B	*TTC39B*	3199695	2.29E-04	-2.76	1.72E-01	-1.10
WNK lysine deficient protein kinase 1	*WNK1*	3400127	1.06E-05	-2.74	1.15E-01	-1.45
enabled homolog (Drosophila)	*ENAH*	2458376	8.23E-04	-2.69	5.48E-02	-1.35
coiled-coil-helix-coiled-coil-helix domain containing 5	*CHCHD5*	2500878	4.82E-05	-2.68	9.34E-02	-1.18
ATPase, class VI, type 11C	*ATP11C*	4024164	6.83E-04	-2.67	1.78E-01	-1.16
glutamate decarboxylase 1 (brain, 67kDa)	*GAD1*	2515021	1.83E-04	-2.66	2.50E-01	-1.07
coiled-coil domain containing 68	*CCDC68*	3808749	2.06E-04	-2.58	4.73E-01	-1.05
reticulon 4	*RTN4*	2553630	9.09E-04	-2.46	9.15E-01	1.01
GTPase activating protein (SH3 domain) binding protein 2	*G3BP2*	2773765	7.34E-04	-2.43	1.64E-01	-1.26
F-box protein 36	*FBXO36*	2531135	5.61E-04	-2.36	8.13E-02	-1.20
CKLF-like MARVEL transmembrane domain containing 4	*CMTM4*	3695161	4.83E-04	-2.28	2.83E-01	-1.15
frizzled family receptor 4	*FZD4*	3385518	6.05E-04	-2.26	3.12E-01	-1.08
plexin D1	*PLXND1*	2694864	6.57E-04	-2.23	5.35E-01	-1.03
Kv channel interacting protein 2	*KCNIP2*	3304096	7.82E-04	-2.21	9.47E-01	-1.00
collagen, type XXVII, alpha 1	*COL27A1*	3185987	7.89E-06	-2.21	4.27E-01	1.06
semaphorin 6B	*SEMA6B*	3846887	2.81E-06	-2.18	6.43E-01	1.03
LUC7-like (S. cerevisiae)	*LUC7L*	3674992	6.69E-04	-2.15	6.96E-02	-1.25
leucine rich repeat (in FLII) interacting protein 1	*LRRFIP1*	2534476	9.97E-04	-2.14	5.76E-01	1.07
kinesin family member 24	*KIF24*	3203937	5.26E-04	-2.08	2.97E-01	-1.05
gamma-aminobutyric acid (GABA) B receptor, 1	*GABBR1*	2947928	5.22E-04	-2.07	1.55E-01	-1.11
microtubule-associated protein 2	*MAP2*	2525534	2.44E-04	-2.06	3.57E-01	-1.07
caspase 8 associated protein 2	*CASP8AP2*	2916983	5.90E-04	-2.03	5.15E-02	-1.24
galactose mutarotase (aldose 1-epimerase)	*GALM*	2477959	5.87E-04	-2.02	9.04E-02	-1.30
latent transforming growth factor beta binding protein 3	*LTBP3*	3335309	6.89E-04	-2.01	2.10E-01	1.10
collagen, type XVI, alpha 1	*COL16A1*	2404672	5.61E-07	-1.99	7.47E-01	1.02
autism susceptibility candidate 2	*AUTS2*	3006703	5.69E-04	-1.97	1.38E-01	-1.23
syntaxin binding protein 1	*STXBP1*	3189965	9.70E-04	-1.96	9.09E-02	-1.06
supervillin	*SVIL*	3283067	8.38E-04	-1.92	9.81E-01	1.00

Table 4.6 Probe sets with changes in log_2 signal intensity after *ZNF804A* knockdown

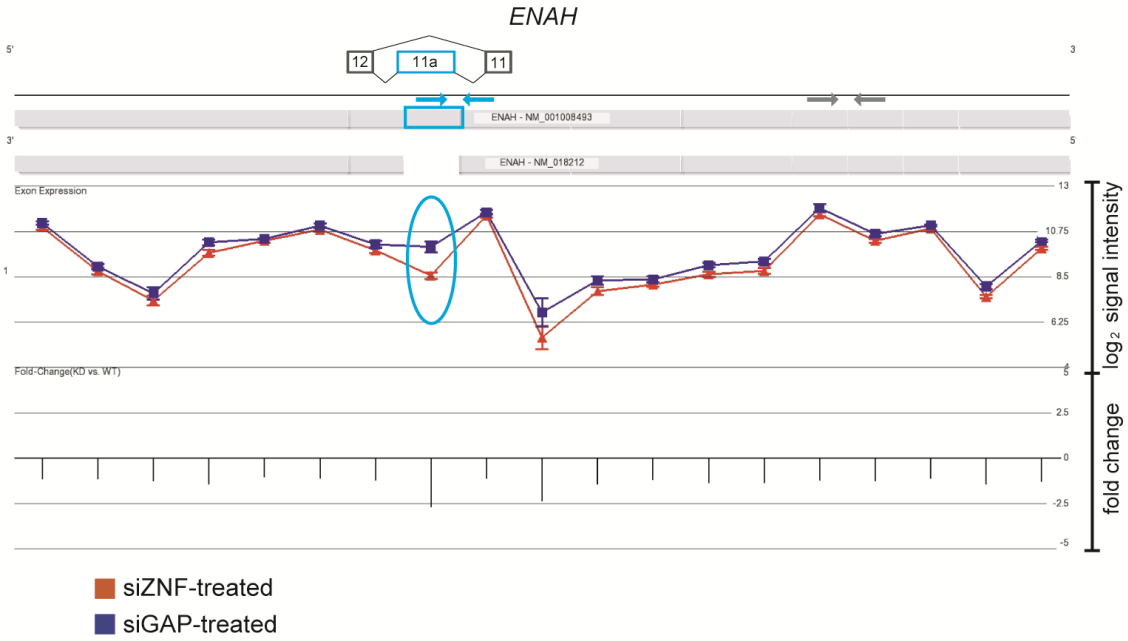
To identify changes in splicing, the probe set-level list derived from the alternative splicing one-way ANOVA was filtered to retain probe sets with altered log_2 signal intensity (FDR 0.05) and to remove any probe sets present in differentially expressed genes ($P > 0.05$). The top 30 **(A)** positive and **(B)** negative fold changes in probe set log_2 signal intensity are shown.

expressed (P = 0.055; one-way ANOVA) (Figures 4.13A and B). Consistent with this result, RT-PCR showed that there was less of the longer transcript of *ENAH* in siZNF-treated samples (Figure 4.13C). Q-PCR confirmed that there was less of the spliced exon present in *ZNF804A*-depleted cells relative to both of the siGAP-treatment and mock samples (one-way ANOVA and Tukey post-hoc P < 0.01; Figure 4.13D). Therefore, the empirical results for *ENAH* were consistent with the exon array analysis. These data suggest that *ZNF804A* may have a role in regulating the inclusion of exon 11a of *ENAH*.

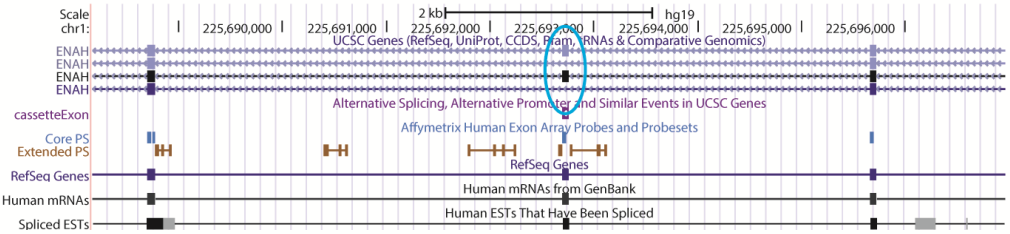
The geneview of *G3BP2* showed that there was alternative splicing of the known cassette exon 7a (chr4:76579167-76579265) when *ZNF804A* was depleted (P = 7.3 x 10^{-4}; FC = -2.43; one-way ANOVA), although the *G3BP2* transcript was not differentially expressed (P = 0.164; one-way ANOVA) (Figures 4.14A and B). RT-PCR and Q-PCR suggested that there was less of the longer transcript in the siZNFB-treated sample, but there was no change in the abundance of the longer transcript in the siZNFA-treated sample (Figures 4.14C and D). These data were inconsistent with the exon array results and therefore, the alternative splicing of exon 7a of *G3BP2* was not validated.

Confirming the two remaining selected alternative splicing events proved problematic. Firstly, the gene *ATP11C* showed increased exon exclusion of a known alternatively spliced exon, exon 29a (chrX: 138813810-138813914), when *ZNF804A* was depleted (P = 6.8 x 10^{-4}; FC = -2.67; one-way ANOVA) (Figure 4.15A). RT-PCR showed the abundance of the longer transcript of *ATP11C* decreased and that there was a reciprocal increase in the abundance of the shorter transcript in siZNF-treated samples (Figure 4.15B). This was consistent with the hypothesis that there was increased exclusion of exon 29a after *ZNF804A* knockdown. However, it proved impossible to design Q-PCR primers for exon 29a that satisfied the

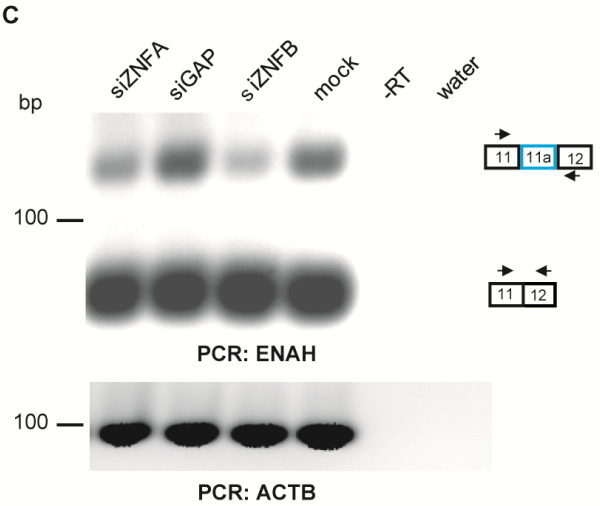
A


ENAH
12
11a
11
ENAH - NM_001008493
ENAH - NM_018212
Exon Expression
log2 signal intensity
Fold-Change(KD vs. WT)
fold change
siZNF-treated
siGAP-treated

B


Scale
chr1:
2 kb
hg19
225,690,000
225,691,000
225,692,000
225,693,000
225,694,000
225,695,000
225,696,000
UCSC Genes (RefSeq, UniProt, CCDS, Rfam, tRNAs & Comparative Genomics)
ENAH
ENAH
ENAH
ENAH
Alternative Splicing, Alternative Promoter, and Similar Events in UCSC Genes
cassetteExon
Affymetrix Human Exon Array Probes and Probesets
Core PS
Extended PS
RefSeq Genes
RefSeq Genes
Human mRNAs
Human mRNAs from GenBank
Spliced ESTs
Human ESTs That Have Been Spliced

C


siZNFA
siGAP
siZNFB
mock
-RT
water
bp
100
11
11a
12
11
12
PCR: ENAH
100
PCR: ACTB

D

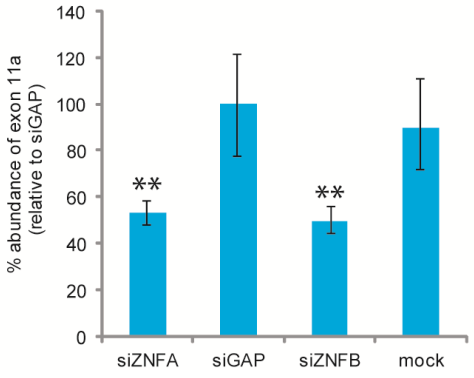

% abundance of exon 11a
(relative to siGAP)
140
120
100
80
60
40
20
0
**
**
siZNFA
siGAP
siZNFB
mock

Figure 4.13 Exon 11a of *ENAH* was alternatively spliced when *ZNF804A* was depleted

Analysis of the probe set-level list generated from the alternative splicing ANOVA identified a probe set in *ENAH* as differentially expressed ($P = 8.2 \times 10^{-4}$; FC = -2.69; one-way ANOVA) although the *ENAH* transcript was not differentially expressed (P = 0.055; one-way ANOVA). This splicing event was confirmed empirically. **(A & B)** The geneview for *ENAH* showed that after *ZNF804A* knockdown there was increased exclusion of the known cassette exon 11a. **(C)** RT-PCR using primers complementary to the constitutive exons flanking exon 11a. *Beta actin* (*ACTB*) was used as a loading control. **(D)** Q-PCR using primers complementary to exon 11a (blue arrows) and a 'control' pair complementary to another region on the transcript (grey arrows). Please note, one biological replicate used in the exon array was analysed using Q-PCR. The expression of exon 11a was compared between the conditions using the $\Delta\Delta C_t$ method with the control primer set as the endogenous normaliser. The bar graph presented shows the percentage abundance of exon 11a in each condition relative to the siGAP-treated sample. Q-PCR was performed in triplicate (three technical replicates) for each condition. The error bars represent the standard deviation of the three raw C_t values. The significance was assessed using a one-way ANOVA and Tukey post-hoc. ** $P < 0.05$.

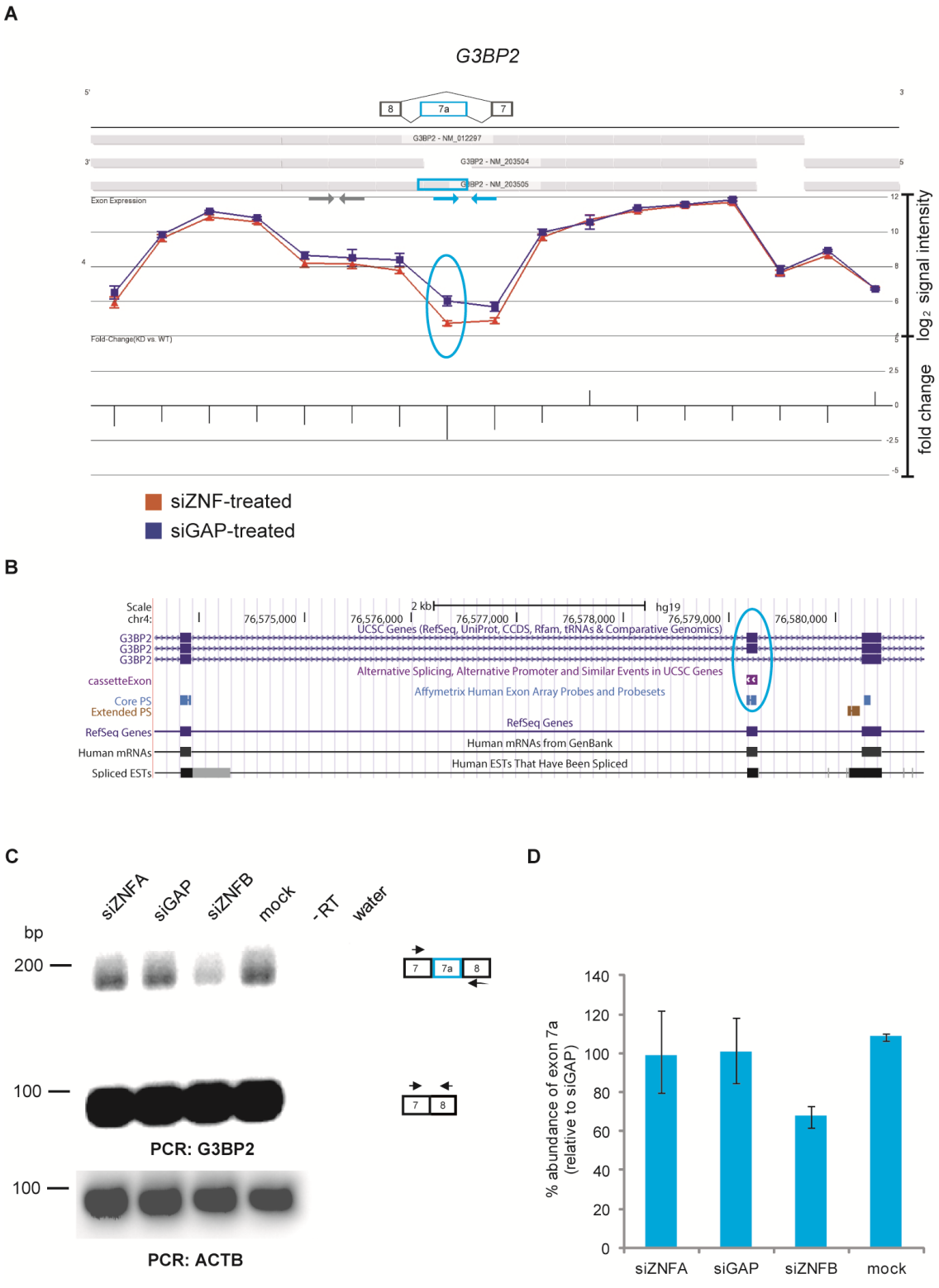

A
G3BP2
siZNF-treated
siGAP-treated
log2 signal intensity
fold change
B
C
siZNFA
siGAP
siZNFB
mock
-RT
water
bp
200
100
PCR: G3BP2
PCR: ACTB
D
% abundance of exon 7a (relative to siGAP)
siZNFA
siGAP
siZNFB
mock

Figure 4.14 The assessment of alternative splicing of exon 7a in *G3BP2*

Analysis of the probe set-level list generated from the alternative splicing ANOVA identified a probe set in *G3BP2* as differentially expressed ($P = 7.3 \times 10^{-4}$; FC = -2.43; one-way ANOVA) although the *G3BP2* transcript was not differentially expressed (P = 0.164). This alternative splicing event was not validated using empirical methods. **(A & B)** The PGS geneview showed this corresponded to increased exclusion of the known cassette exon 7a when *ZNF804A* was depleted. **(C)** RT-PCR using primers complementary to the constitutive exons flanking exon 7a. **(D)** Q-PCR using primers complementary to exon 7a (blue arrows) and a 'control' pair complementary to another region on the transcript (grey arrows). Please note, one biological replicate used in the exon array was analysed using Q-PCR. The expression of exon 7a was compared between the conditions using the $\Delta\Delta C_t$ method with the control primer set as the endogenous normaliser. The bar graph presented shows the percentage abundance of exon 7a in each condition relative to the siGAP-treated sample. Q-PCR was performed in triplicate (three technical replicates) for each condition. The error bars represent the standard deviation of the three raw C_t values.

desired amplification efficiencies required to use the comparative $\Delta\Delta C_t$ method. Therefore, the splicing of exon 29a was not examined by Q-PCR. However, the RT-PCR showed convincing evidence that *ZNF804A* knockdown altered the inclusion of exon 29a of *ATP11C*. Therefore, it is likely that ZNF804A has a role in regulating the splicing of this exon.

The second gene containing a splicing event which was problematic to confirm was *STXBP1*. The geneview analysis of *STXBP1* showed that there was alternative inclusion of exon 18a when *ZNF804A* was depleted ($P = 9.7 \times 10^{-4}$; FC = -1.96; one-way ANOVA), although the *STXBP1* transcript was not differentially expressed (P = 0.178; one-way ANOVA) (Figure 4.15C). RT-PCR demonstrated that the amount of the longer transcript was markedly reduced in the siZNF-treated samples (Figure 4.15D). However, RT-PCR amplified three bands, rather than the predicted two bands. This suggests that these RT-PCR primers may amplify a novel splice variant of *STXBP1*. The splicing of exon 18a of *STXBP1* was not validated by Q-PCR as the Q-PCR primers also amplified more than one product. However, as the RT-PCR showed convincing evidence of alternative splicing in *STXBP1*, it is highly likely that the splicing of exon 18a of STXBP1 was altered in *ZNF804A*-depleted cells.

In summary, knockdown of *ZNF804A* led to numerous statistically significant changes in splicing. Of the 11 splicing events selected for validation, two were unequivocally validated (*SIPA1L1* and *ENAH*), two were validated by RT-PCR only (*ATP11C* and *STXBP1*), three events were not validated (*PKM2*, *NFYA*, *G3BP2*) and four were inconclusive (*BAZ1A*, *MAP2K5*, *RCOR2* and *SEC14L1*).

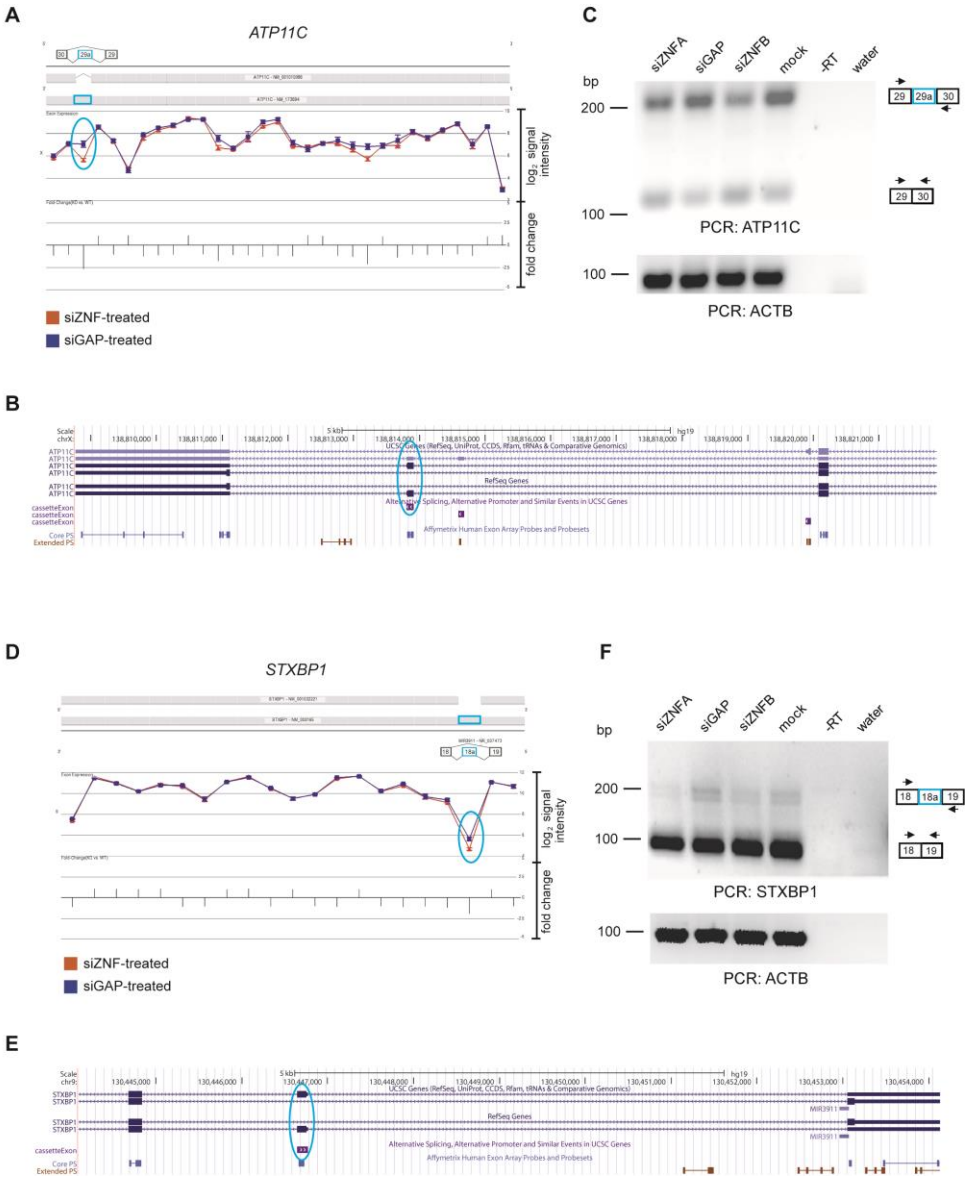


Figure 4.15 Validation of alternative splicing in *ATP11C* and *STXBP1*

Analysis of the probe set-level list generated from the alternative splicing ANOVA identified a probe set in *ATP11C* as differentially expressed ($P = 6.8 \times 10^{-4}$; FC = -2.67; one-way ANOVA) although the *ATP11C* transcript was not differentially expressed (P = 0.178; one-way ANOVA). **(A)** The *ATP11C* PGS geneview showed increased exclusion of a known cassette exon 29a **(B)** when *ZNF804A* was depleted. **(C)** RT-PCR using primers complementary to the constitutive exons flanking exon 29a confirmed the alternative splicing event. Analysis of the probe set-level list generated from the alternative splicing ANOVA identified a probe set in *STXBP1* as differentially expressed ($P = 9.7\ 10^{-4}$; FC = -1.96; one-way ANOVA) although the *STXBP1* transcript was not differentially expressed (P = 0.099; one-way ANOVA). **(D)** The *STXBP1* PGS geneview showed increased exclusion of known cassette exon 18a **(E)** when *ZNF804A* was depleted. **(D)** RT-PCR using primers complementary to the constitutive exons confirmed alternative splicing of exon 18a.

4.6.3. Investigating a change in transcription start site usage

In addition to changes in cassette exon splicing, the alternative splicing one-way ANOVA identified changes in transcript variant use after *ZNF804A* knockdown. Many of these events involved multiple transcript variants and were too complex to confirm empirically. However, the geneview of *PTPRR* demonstrated that there was a simple switch in transcript variant usage when *ZNF804A* was depleted (Figure 4.16A). This switch in transcript variant usage was confirmed using Q-PCR (Figure 4.16B). The isoforms of PTPRR arise from the differential use of distinct transcription start sites and promoters (Chirivi et al., 2004). Therefore, this result suggests ZNF804A may also have a role in determining transcription start site or promoter site choice.

4.6.4. Enrichment analysis of genes showing alternative splicing after *ZNF804A* knockdown

To identify particular biological processes which were enriched for *ZNF804A*-related alternative splicing events, enrichment analysis was performed using GeneGo MetaCore™ software. To ensure that only genes which could be detected as alternatively spliced in the exon array were included in the enrichment analysis, the background comparison was set as the Affymetrix GeneChip human exon array 1.0 ST-v2. Enrichment by GeneGo pathway map showed genes belonging to the term 'cell adhesion_extracellular matrix (ECM) remodelling' (4/52 $P = 2.93 \times 10^{-4}$) were significantly enriched among the 116 putatively alternatively spliced genes (section 4.6.1). This was the only pathway map enrichment which survived 5% FDR correction for multiple testing. The four genes belonging to this pathway map were *insulin-like growth factor 2 (somatomedin A)* (*IGF2*); *CD44 molecule (Indian blood group)* (*CD44*), *matrix metallopeptidase 15 (membrane-inserted)* (*MMP15*) and *collagen, type IV, alpha 1* (*COL4A1*). Consistent with these data, enrichment by GeneGo process network

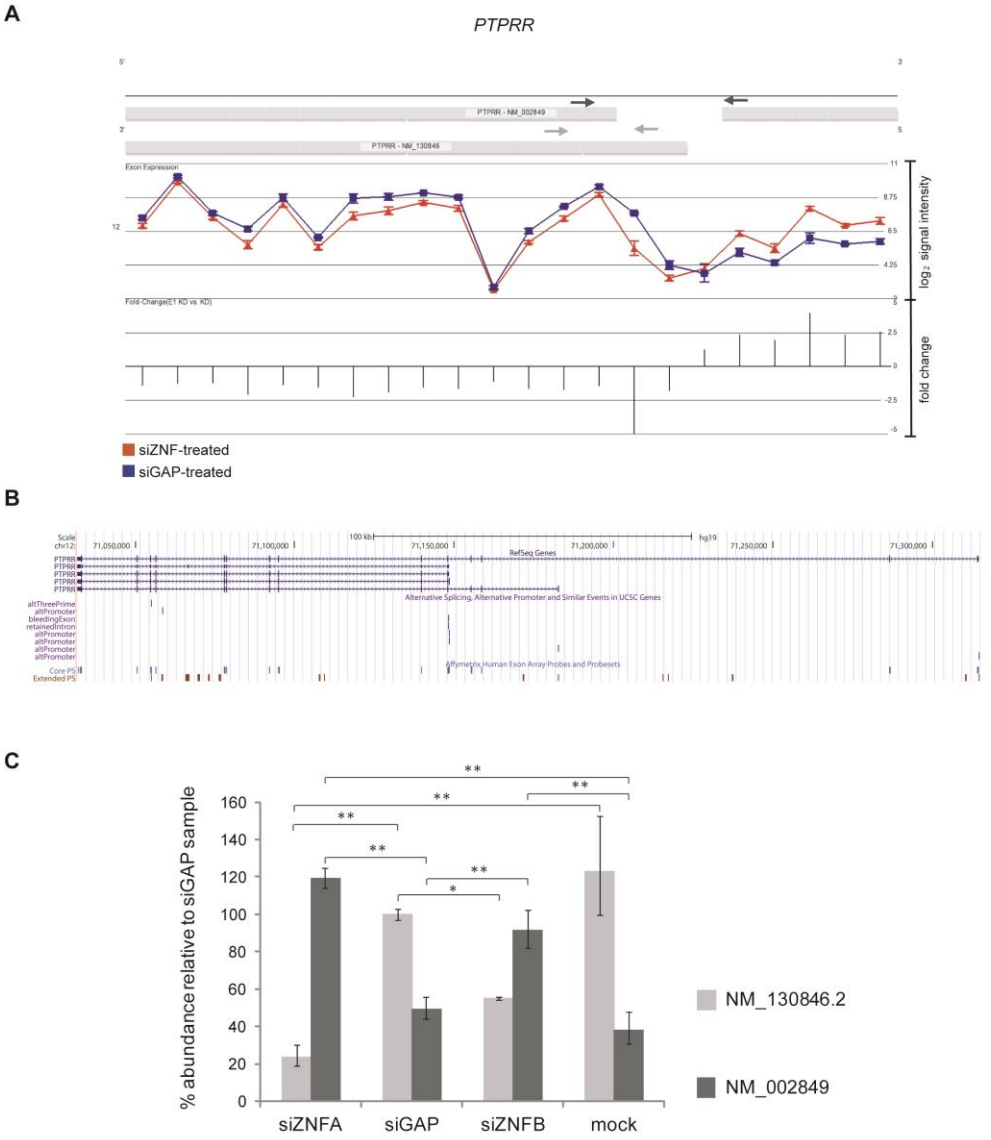


Figure 4.16 Empirical validation of alternative transcript usage in *PTPRR* after *ZNF804A* knockdown *PTPRR* was identified as alternatively spliced by alternative splicing one-way ANOVA ($P = 3.32 \times 10^{-19}$). **(A)** The *PTPRR* geneview showed a potential switch in transcript variant use in *ZNF804A*-depleted cells. **(B)** *PTPRR* has five RefSeq transcript variants. **(C)** Q-PCR using primers complementary to NM_130846.2 and NM_002849 (illustrated by arrows on the geneview). Please note, one biological replicate used in the exon array was analysed using Q-PCR. The expression of each *PTPRR* transcript was compared between the conditions using the $\Delta\Delta C_t$ method with *beta actin* (*ACTB*) as the endogenous normaliser. The bar graph presented shows the percentage abundance of each *PTPRR* transcript in each condition relative to the abundance of NM_130846.2 in the siGAP-treated sample. Q-PCR was performed in triplicate (three technical replicates) for each condition. The error bars represent the standard deviation of the three raw C_t values. The significance was assessed between the *ZNF804A*-depleted samples and both of the control samples using a one-way ANOVA and Tukey post-hoc test on ΔC_t ; ** = $P < 0.001$; * = $P < 0.05$.

identified 'cell adhesion_cell-matrix interactions' (13/211 P = 7.95 x 10^{-5}) as significantly enriched (Figure 4.17A). Genes belonging to the GeneGo process network term 'development_neurogenesis_axonal guidance' (12/230 P = 1.19 x 10^{-5}) were also enriched among the 116 putatively alternatively spliced genes. The 10 active genes annotated to this term were *reticulon 4* (*RTN4*); *plexin D1* (*PLXND1*); *EPH receptor B2* (*EPHB2*); *reelin* (*RELN*); *brain-derived neurotrophic factor* (*BDNF*); *sema domain, immunoglobulin domain (Ig), short basic domain, secreted, (semaphorin) 3F* (*SEMA3F*); *slit homolog 3 (Drosophila)* (*SLIT3*); *slit homolog 1 (Drosophila)* (*SLIT1*); *cadherin 2, type 1, N-cadherin (neuronal)* (*CDH2*) and *ryanodine receptor 1 (skeletal)* (*RYR1*). *EPHB2* was assigned to more than one node in the network. Enrichment by GO process showed the most significantly enriched terms related to neurodevelopment (Figure 4.17B). For example, 'neurogenesis' (52/1649 P = 9.03 x 10^{-24}) and 'nervous system development' (59/2400 P 8.36 x 10^{-22}).

The biological processes enriched among the alternatively spliced genes are consistent with those enriched among the differentially expressed genes (section 4.5.1). For instance, genes belonging to the GO process term 'nervous system development' and the GeneGo process network 'development_neurogenesis_axonal guidance' were significantly enriched among both datasets. It was hypothesised that genes which were alternatively spliced but were not differentially expressed may belong to different biological processes from those which were both alternatively spliced and differentially expressed. To test this hypothesis, the list of 116 alternatively spliced genes was filtered to exclude the 59 genes showing differential expression (FDR 0.01; corresponding to uncorrected P > 0.00026). Enrichment analysis identified genes belonging to GO process terms 'neurogenesis' (20/1649 P = 4.97 x 10^{-9}), 'nervous system development' (23/2400 P = 2.13 x 10^{-8}) and the GeneGo process network 'development_neurogenesis_axonal guidance' (6/230 P = 7.72 x 10^{-4}) as the top biological

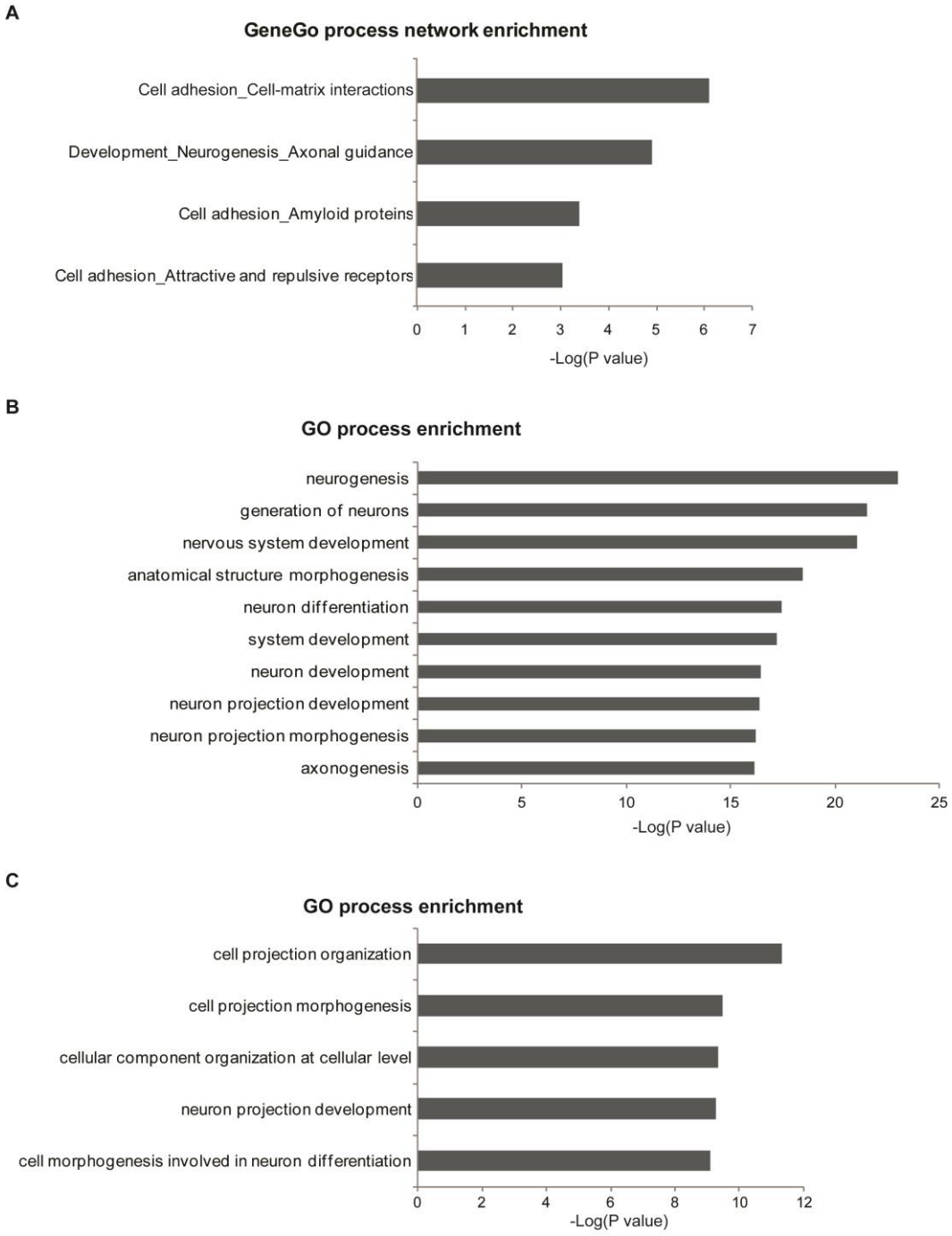


Figure 4.17 Biological processes significantly enriched for genes showing alternative splicing in *ZNF804A*-depleted cells

(A) The top statistically significant GeneGo process networks and **(B)** GO processes among the 116 alternatively spliced genes (section 4.6.1). **(C)** The top statistically significant GO processes among the 566 non-differentially expressed genes containing a differentially expressed probe set (section 4.6.2). Enrichment analysis was performed using GeneGo MetaCore™ with Affymetrix GeneChip human exon array 1.0 ST v-2 as the background list. These enrichments survived multiple test correction, as determined at a 5% FDR by GeneGo MetaCore™.

processes significantly enriched among the 57 alternatively spliced genes with no differential expression (data not shown). These enrichments survived 5% correction for multiple testing. These data imply that ZNF804A may influence both the expression and splicing of genes implicated in processes underlying nervous system development. Consistent with this interpretation, enrichment by GO process of the 566 non-differentially expressed genes containing a differentially expressed probe set (section 4.6.2) showed genes belonging to the term 'neuron projection development' (57/813 $P = 5.76 \times 10^{-10}$) were significantly over-represented among these genes (Figure 4.17C).

4.6.5. Comparing the relative number of alternative splicing events after *ZNF804A* or *GAPDH* knockdown

It was difficult to predict the number of alternative splicing events which would be expected following *ZNF804A* knockdown, if it was a true regulator of pre-mRNA processing. Therefore, to provide a benchmark with which to compare the effects of knocking down *ZNF804A* on splicing, the number of alternatively spliced genes after *ZNF804A* or *GAPDH* knockdown relative to the mock samples was determined using the same parameters as above. The splicing of 177 genes was altered after *ZNF804A* knockdown relative to the mock samples (in the interest of brevity, these data are not shown). The splicing of 13 genes was altered after *GAPDH* knockdown relative to the mock samples (presented as supplementary data in Appendix 5.4). This result suggests that knockdown of *ZNF804A* led to many more changes in splicing than would be expected if *ZNF804A* had no role in pre-mRNA processing.

4.7. Discussion

Data presented in Chapter Three implicate a role for ZNF804A in the regulation of transcription and pre-mRNA splicing. The aim of the experiments described in this Chapter was to investigate the role of ZNF804A in the regulation of transcription and pre-mRNA processing using a combination of siRNA, exon array technologies and pathway analysis. The hypothesis was that knockdown of endogenous *ZNF804A* would lead to changes in the expression and pre-mRNA splicing of transcripts regulated, directly or indirectly, by ZNF804A. The findings presented in this Chapter provide convincing evidence to support this hypothesis.

The knockdown of *ZNF804A* resulted in numerous statistically significant changes in gene expression. A selection of these gene expression changes were empirically validated using Q-PCR. Among the most differentially expressed genes were genes which have been implicated in schizophrenia pathology, including *RELN*, *NPY* and *NLGN2* (Table 4.3). Whilst it was tempting to place emphasis on these gene expression changes, based on *a priori* knowledge of schizophrenia genetics, it was more informative to use enrichment analysis to determine biological processes enriched for ZNF804A-related target genes. Enrichment analysis indicated that genes belonging to GO process 'nervous system development' and the GeneGo network process 'cell adhesion_synaptic contact' were significantly over-represented in the genes with altered expression in *ZNF804A*-depleted cells. These biological processes have been implicated in the neurobiology of schizophrenia (Bourgeron, 2009; Melom and Littleton, 2011; Zoghbi, 2003). Further, recent genetic data suggests risk genes for schizophrenia converge at the synapse (see section 1.6.3.4) (O'Dushlaine et al., 2011; Kirov et al., 2012; Glessner et al., 2010). Therefore, these data may reflect candidate mechanisms for ZNF804A's function in disease. Consistent with the notion that the genes with altered

expression after *ZNF804A* knockdown represent the causal effects of *ZNF804A*'s contribution to disease, the differentially expressed genes were moderately enriched for genes with genetic association with schizophrenia and bipolar disorder (section 4.5.2).

The knockdown of endogenous *ZNF804A* led to a number of statistically significant changes in pre-mRNA processing. Importantly, the knockdown of endogenous *ZNF804A* resulted in remarkably more statistically significant alternative splicing events than the knockdown of endogenous *GAPDH* (section 4.6.5). These data are consistent with a potential role for *ZNF804A* in pre-mRNA processing. The empirical confirmation of the exon array data was successful for five of the eleven events selected for validation. Although this number seems small, it is comparable with data obtained by previous groups when evaluating the targets of known pre-mRNA splicing factors using exon arrays. For example, knockdown of *heterogeneous nuclear ribonucleoprotein L* (*hnRNP L*), which encodes an abundant nuclear protein that is a known global regulator of alternative splicing, gave only 11 validated alternative splicing events across the genome (Hung et al., 2008).

Interestingly, knockdown of *ZNF804A* altered the splicing of exon 11a of *ENAH*, a known target of the putative ZNF804A-interactor, RBFOX2 (Chapter Three) (Yeo et al., 2009). It is tempting to speculate that ZNF804A and RBFOX2 may interact to co-regulate splicing of exon 11a of *ENAH* (Figure 4.18). ENAH is a member of the enabled (Ena)/vasodilator-stimulated phosphoprotein (VASP) family of proteins which are involved in a range of processes dependent on cytoskeleton remodelling and cell polarity, such as axon guidance and growth cone dynamics in migrating cells (Lebrand et al., 2004). Studies suggest that the two transcript variants, including and excluding exon 11a of *ENAH*, have different functional roles in the cell, possibly arising from additional serine and tyrosine phosphorylation sites

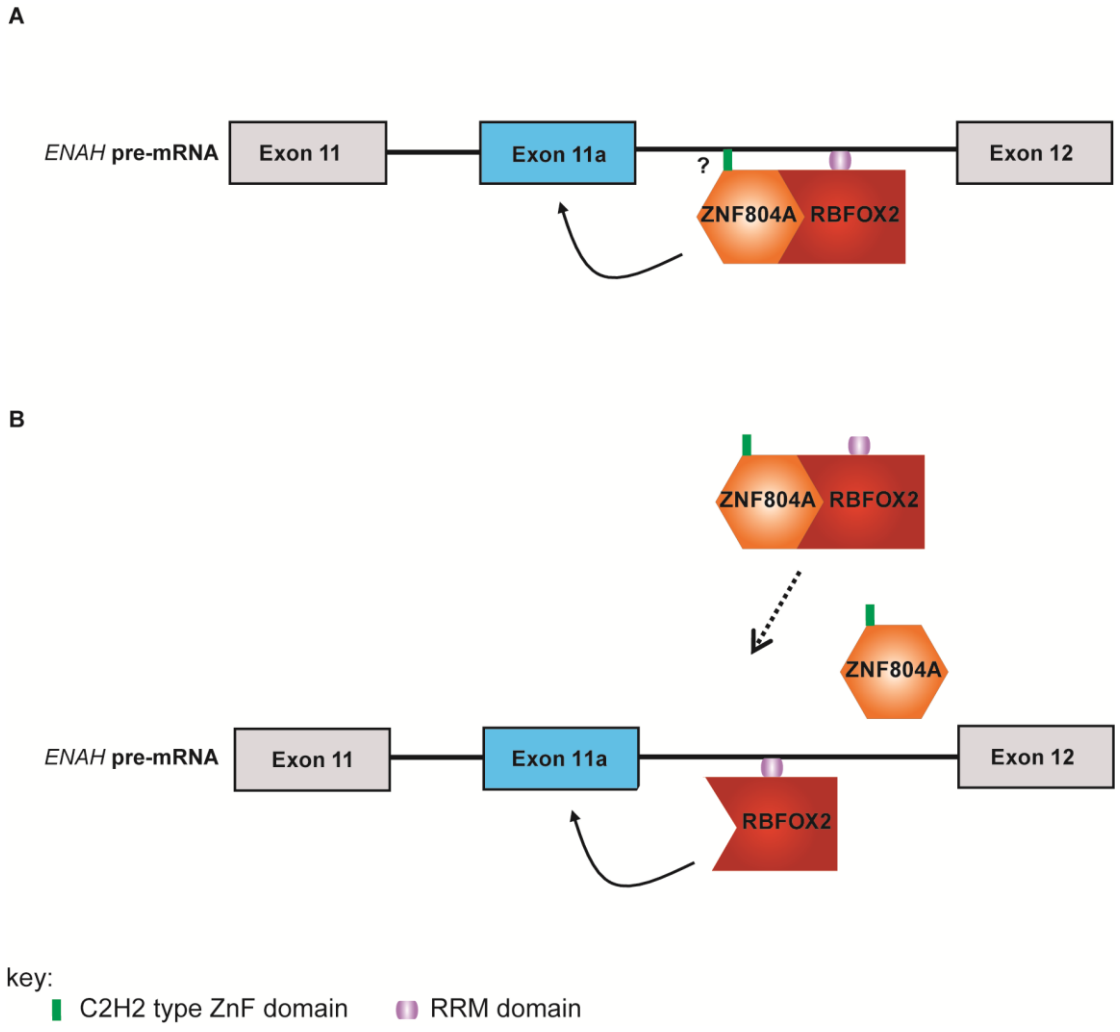


Figure 4.18 Models of ZNF804A/RBFOX2 mediated splicing of exon 11a of *ENAH*

(A) RBFOX2 and ZNF804A may interact with one another and both bind directly to RNA upstream of exon 11a of *ENAH* to promote its inclusion. **(B)** Alternatively, interaction between RBFOX2 and ZNF804A may be necessary for binding of RBFOX2 to RNA upstream of exon 11a to promote its inclusion. In each of these models, knockdown of either RBFOX2 or ZNF804A would result in exclusion of exon 11a of *ENAH*.

within the alternatively spliced exon (Barzik et al., 2005; Di Modugno et al., 2007). Therefore, aberrant splicing of exon 11a of *ENAH* may have downstream consequences on axonal guidance and nervous system development which may increase risk for neuropsychiatric disease. Support for this hypothesis comes from recent post-mortem transcriptome analyses of brain tissue which showed a statistically significant increase in the exclusion of exon 11a of *ENAH* in the Brodmann Area 10 and caudate of schizophrenia patients compared to controls (Cohen, 2012). Furthermore, this hypothesis is consistent with the data presented by Kahler and colleagues that showed that haplotypes of *ENAH* were associated with increased risk for schizophrenia (Kahler et al., 2008).

In addition to *ENAH*, splicing of exon 18a of *STXBP1* was altered in *ZNF804A*-depleted cells. STXBP1 (also known as MUNC18-1) is a member of the Sec1/munc18-like protein family. These proteins are highly conserved between species and are essential components of synaptic vesicle fusion protein complexes. There are two known isoforms of STXBP1 of which only the longer isoform is expressed in neural tissues, suggesting it has a role specific for synaptic transmission (Swanson et al., 1998; Tellam et al., 1995). The results presented here show that when *ZNF804A* was depleted, there was reduced expression of the neural-specific transcript variant. Interestingly, the splicing of a different family member, *STXBP2*, has been shown to be regulated by NOVA2 (Ule et al., 2005), a putative ZNF804A-interactor (Chapter Three). It is tempting to speculate that NOVA2 may also mediate splicing of *STXBP1*, potentially through interactions with ZNF804A. Aberrant splicing of *STXBP1* may contribute to susceptibility for schizophrenia via changes to neurotransmitter release and vesicle trafficking.

There is little functional information regarding the genes *SIPA1L1* and *ATP11C* and as such, it is not possible to infer the consequences of alternative splicing of these transcripts on the protein's function. However, deep sequencing of *SIPA1L1* in 138 schizophrenia patients and 285 controls identified a nonsense mutation in *SIPA1L1* in a schizophrenia case (Myers et al., 2011). Therefore, it is conceivable that alternative splicing in *SIPA1L1* may contribute to increased risk for schizophrenia.

The alternative splicing analysis identified a switch in transcript variant usage in *PTPRR* which was empirically confirmed. The isoforms of PTPRR arise from the differential use of distinct transcription start sites and promoters (Chirivi et al., 2004). PTPRR is an enzyme expressed predominantly in the brain (Augustine et al., 2000a). There are four known human protein isoforms of PTPRR (Augustine et al., 2000b), of which the mRNAs coding for PTPPBSα and PTPPBSγ were identified as differentially expressed in this Chapter. Studies in mice show PTPPBSα is a receptor-type protein tyrosine phosphatase (PTP) whereas PTPPBSγ is a cytosolic type PTP (van den Maagdenberg 1999). Each of these isoforms exhibits a different expression pattern in specific neural cell subtypes during development, suggesting that they have diverse roles in neurodevelopment (van den Maagdenberg 1999). Further, the predicted ligands of PTPPBSα suggest that this receptor is involved in neuronal development and synaptic plasticity (Chesini et al., 2011). This suggests that aberrant transcript variant usage of *PTPRR* may contribute to disease susceptibility via functional consequences on nervous system development. The alternative transcript usage of *PTPRR* after *ZNF804A* knockdown may suggest that ZNF804A has roles in the regulation of transcription start site and promoter site choice.

For *PKM2*, *NFYA* and *G3BP2*, the exon array data were not validated. It is important to consider the reasons behind this and what this infers about the methods used. Firstly, these three events may represent false-positive alternative splicing calls in the exon array analysis. For instance, the assessment of *NFYA* suggests that rather than alternative splicing of exon 1a, there was a change in transcript expression of the whole gene (Figure 4.12E). This may suggest that the alternative splicing ANOVA might not be sensitive enough to correct for small changes in transcript expression. Secondly, some or all of these three events may represent false-negative findings of the empirical analyses. The primary caveat of the empirical analyses is that only one of the independent biological replicates used in the exon array was evaluated by Q-PCR. It is possible that if all four of the biological replicates were evaluated by Q-PCR the increase in statistical power may support replication of the exon array data. Further Q-PCR experiments are required to investigate this possibility. In addition, there may be unknown transcript variants of these three genes which may impact the empirical assessment of the exon usage using RT-PCR and Q-PCR. Future experiments using northern blotting or RNA sequencing analysis may provide a more comprehensive assessment of the possible nature of alternative splicing of these genes after *ZNF804A* knockdown. Interestingly, both the RT-PCR and Q-PCR data appeared to suggest that the siRNA duplexes used to deplete *ZNF804A* may have had different effects on the splicing of exon 9a of PKM2 (Figure 4.10) and exon 7a of G3BP2 (Figure 4.14). However further inspection of the geneviews on the exon array showed that both siRNA duplexes influenced the splicing of the exons to the same extent (data not shown). Therefore, it is unlikely that the siRNA duplexes had different effects on splicing. It is important to note that the empirical analysis focused on known alternative splicing events because previous investigations showed that this filter reduces the number of false positives (Whistler, 2010). The bias to known pre-mRNA splicing events precludes the discovery of novel pre-mRNA splicing

events and as such, the estimates of empirically validated alternative splicing are very conservative.

The enrichment analysis indicated the alternatively spliced genes were enriched for genes involved in biological processes implicated in schizophrenia, such as nervous system development and axonal guidance (Amann-Zalcenstein et al., 2006; Ingason et al., 2007). These data may indicate alternative splicing of *ZNF804A*-related target genes could increase risk for disease. These data are consistent with enrichment by biological process of the genes differentially expressed after *ZNF804A* knockdown. Importantly, these data suggest ZNF804A may influence both the expression and splicing of genes implicated in processes underlying nervous system development.

Recent studies have reported knockdown of wildtype *ZNF804A* in human neural progenitor cells led to nominally significant differential expression of 151 genes; these genes were enriched for genes belonging to the GO term 'cell adhesion' (Hill et al., 2012a). Data presented in this Chapter also indicate a significant effect of *ZNF804A* knockdown on genes involved in cell adhesion, particularly synaptic contact. Only 20 genes were identified as differentially expressed in both studies (supplementary data presented in Appendix 5.5). The relatively small overlap between the datasets, and the finding that more genes showed altered expression in our study, may be explained by the difference in the level of knockdown achieved using the siRNA duplexes. Specifically, Hill and colleagues (2012) chose to knockdown wildtype *ZNF804A* to 60% to emulate unpublished observations of the cis-regulatory effects of the disease-associated SNP in *ZNF804A*, while siZNFA or siZNFB mediated knockdown of *ZNF804A* to an average of 18.7% (sd = 8.4) and 23.1% (sd = 8.8) of endogenous levels respectively.

In summary, the results presented in this Chapter are consistent with the hypothesis that *ZNF804A* plays a role in the regulation of gene expression and pre-mRNA processing (Chapter Three). Notably, *ZNF804A*-related targets were enriched for biological processes implicated in neuropsychiatric diseases and genes which are genetically associated with schizophrenia and bipolar disorder. The alternative splicing analysis showed knockdown of *ZNF804A* altered the splicing of exon 11a of *ENAH* which is regulated by RBFOX2 (Yeo et al., 2009), a putative interactor of ZNF804A (Chapter Three). This is consistent with the hypothesis that ZNF804A may interact with RNA-binding proteins to mediate splicing (Chapter Three). Further investigation is required to define the mechanisms of *ZNF804A*-related gene expression and pre-mRNA splicing and to develop an appreciation of how the disease-associated SNP in *ZNF804A* impacts on its function and increases risk for schizophrenia.

Chapter 5: Investigating the effects of over-expressing *myc-ZNF804A* on the cellular transcriptome

5.1. Introduction

The findings presented in Chapter Three suggest a role for ZNF804A in the regulation of transcription and pre-mRNA processing. Consistent with these findings, data presented in Chapter Four show that *ZNF804A*-depleted cells displayed numerous changes in gene expression and pre-mRNA splicing. Often, to complement knockdown experiments, over-expression experiments are used to study gene function in heterologous cells. For example, fibroblasts can be converted to induced pluripotent stem (iPS) cells through reprogramming by heterologous expression of four transcription factors (Oct4, Sox2, c-Myc and Klf4) (Takahashi and Yamanaka, 2006; Wernig et al., 2007). In addition, a recent paper by Wilbert and colleagues elucidated the role of LIN28 in mRNA regulation using splicing arrays in cells stably over-expressing *LIN28* and in *LIN28*-depleted cells (Wilbert et al., 2012). Therefore, the aim of this Chapter was to over-express *myc-ZNF804A* in mammalian cells and profile the ensuing genome-wide changes in gene expression and pre-mRNA splicing using Affymetrix GeneChip human exon 1.0 ST arrays (Figure 5.1). The hypothesis was that over-expression of *myc-ZNF804A* would lead to changes in the expression and pre-mRNA splicing of transcripts regulated, directly or indirectly, by ZNF804A.

5.2. Characterisation of the myc-ZNF804A Flp In-TREx cell line

At the outset of this project, our data suggested that transient transfection of *ZNF804A*-expression vectors into mammalian cells did not produce detectable levels of exogenous ZNF804A (C.L. Tinsley, personal communication, the vectors and cell lines used are summarised in Appendix 2). Data presented in Chapter Three support these findings and

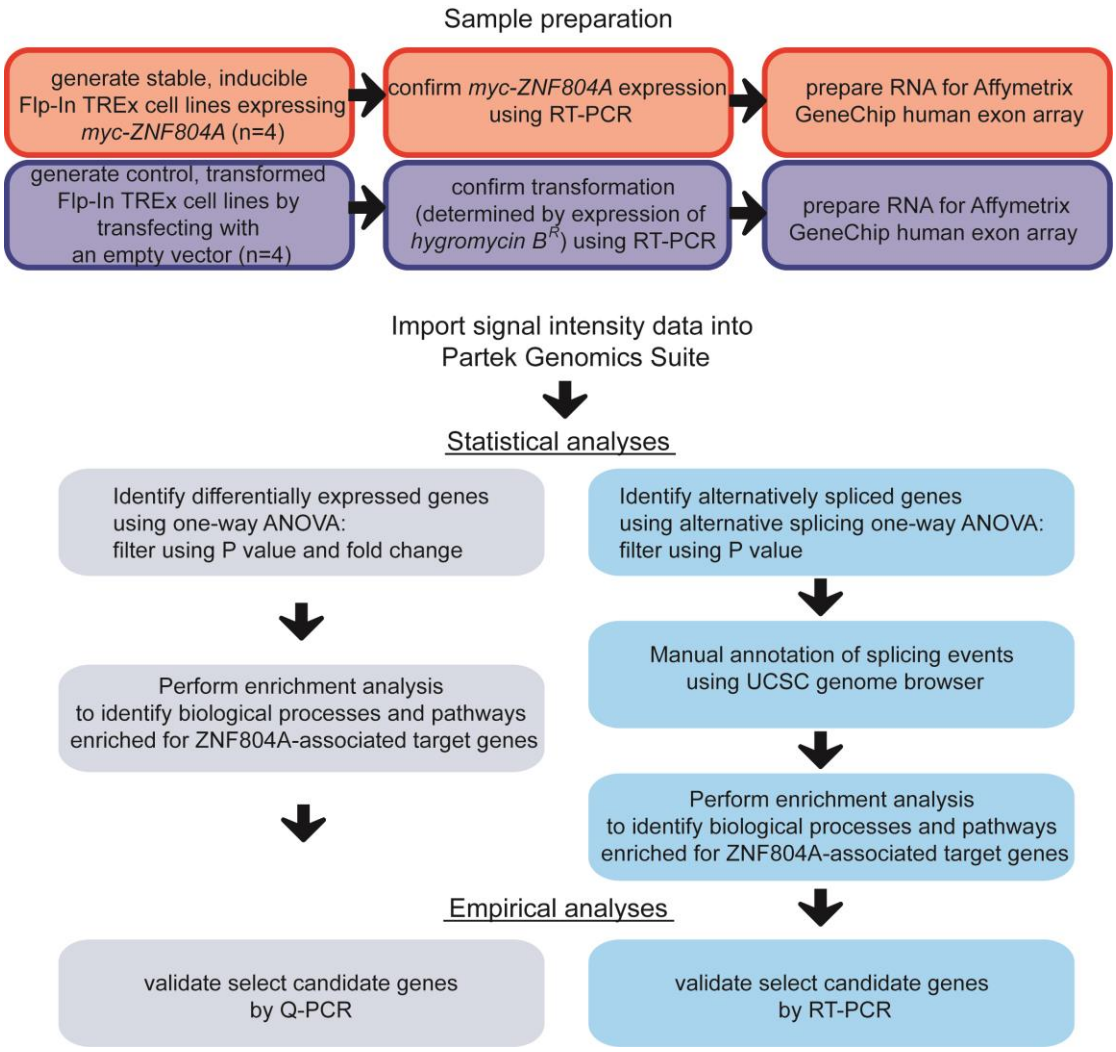


Figure 5.1 A flowchart illustrating the experimental design

The Flp-In TREx expression system was used to generate a stable, inducible cell line expressing *myc-ZNF804A*. Cultured myc-ZNF804A Flp-In TREx cells were induced to over-express *myc-ZNF804A* and RNA was prepared. Exon arrays were performed by Central Biotechnology Services (CBS), Cardiff University using Affymetrix GeneChip human exon array 1.0 ST. The raw .CEL files which contained the signal intensity information were imported into Partek Genomics Suite (PGS) using the core metaprobe set annotation file. The robust multiarray averaging (RMA) algorithm was used to summarise the probe-level data to a single value for each probe set. Two parallel analyses were performed to identify changes in either gene expression or pre-mRNA splicing. Differentially expressed gene lists were generated using a cut-off P value only or P value and fold change. A list of alternatively spliced genes was generated using a cut-off P value. Enrichment analysis was performed to identify biological processes enriched for ZNF804A-associated target genes. A selection of changes was empirically confirmed.

show that transiently expressed myc-ZNF804A was only reliably detected after proteasome inhibition (section 3.5.2). Prior to this discovery, it was hypothesised that the inability to detect exogenous ZNF804A may indicate that over-expression of *ZNF804A* was cytotoxic or that ZNF804A may be produced at extremely low levels. An alternative hypothesis is that the cDNA was only taken-up by a few cells. Previous studies have used tetracycline-inducible stable cell lines to successfully overcome such difficulties (Reeves et al., 2002). Therefore to express ZNF804A in a regulatable manner, a tetracycline-inducible, stable cell line expressing *myc-ZNF804A* was generated (section 2.8.2). *Myc-ZNF804A* was cloned into the pcDNA5/FRT/TO expression vector, co-transfected into Flp-In TREx -293 cells with the pOG44 vector and four successful transformants were selected. In control experiments, I also generated an inducible cell line expressing *GFP-TCF4*. TCF4 can be transiently expressed in mammalian cells and its cognate protein can be readily detected by western blotting and immunocytochemistry (Forrest et al., 2012).

RT-PCR analysis confirmed that the tetracycline-induced myc-ZNF804A and GFP-TCF4 Flp-In TREx cell lines expressed *myc-ZNF804A* or *GFP-TCF4* respectively (Figure 5.2A) (section 2.3.1). However, Figure 5.2A also shows that *myc-ZNF804A* and *GFP-TCF4* mRNA was detected in the uninduced sample. Q-PCR analysis showed the expression of *ZNF804A* was significantly increased in tetracycline-induced samples relative to the uninduced samples (Figure 5.2B) (section 2.3.6). Endogenous *ZNF804A* was not detected in wildtype HEK293T cells by Q-PCR (data not shown). Therefore, the *ZNF804A* detected in the uninduced samples was likely to be *myc-ZNF804A*. These data indicate that there was some *myc-ZNF804A* expression from the pCMV/$TetO_2$ promoter in the absence of tetracycline. This may occur through leaky repression or if there was small amounts of tetracycline present in the cell culture media, for example in the foetal calf serum used to supplement the media.

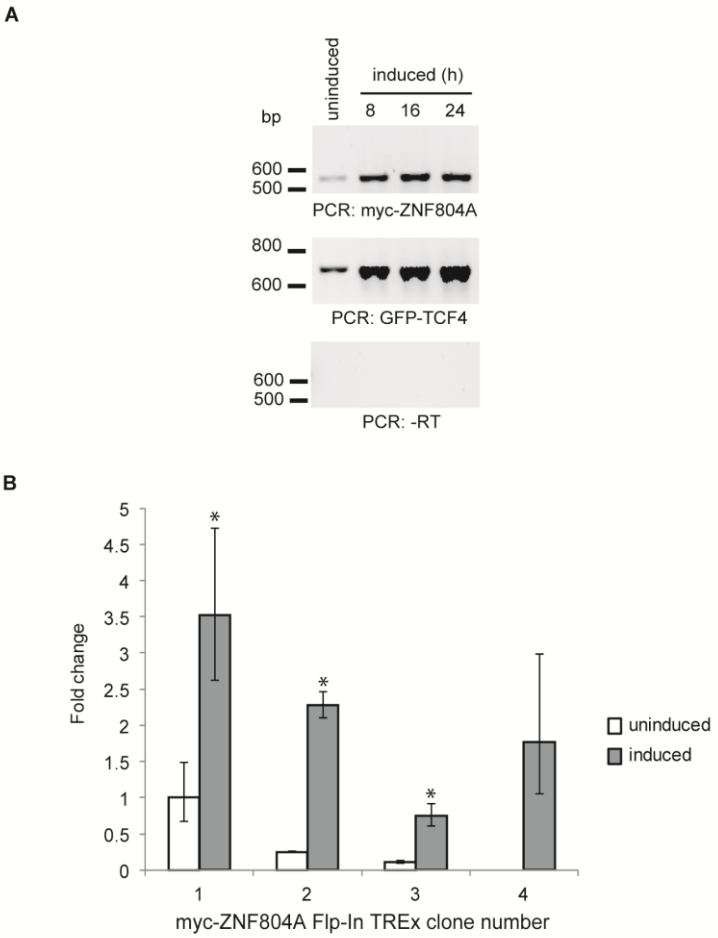


Figure 5.2 PCR analysis of *myc-ZNF804A* and *GFP-TCF4* expression

Cultured myc-ZNF804A or GFP-TCF4 Flp-In TREx cells were either treated with 1 µg/ml tetracycline to induce expression of the gene of interest (induced) or left untreated (uninduced). RNA was prepared at 8h intervals (0h uninduced; 8h post-induction; 16h post-induction and 24h post-induction). Reverse transcription (RT) reactions were performed to generate cDNA. As a control for genomic DNA contamination (gDNA), RT reactions were performed without reverse transcriptase (-RT) using myc-ZNF804A Flp-In TREx clone one as the template. **(A)** RT-PCR was performed using a forward primer complementary to the myc epitope tag and a reverse primer complementary to the gene of interest. Primer sequences are given in Appendices 1.1 and 1.5. (**B)** Q-PCR was performed in triplicate using a Solaris probe and primer pair designed complementary to *ZNF804A*. Primer and probe sequences are given in Appendix 1.4. The mean C_t value for each sample was normalised to *GAPDH* levels. The bar graph represents the fold change in *ZNF804A* mRNA abundance between the samples, relative to the abundance of *ZNF804A* mRNA in the uninduced sample of clone number one. The error bars represent the standard deviation of the three raw C_t values. (* denotes $P < 0.05$, paired t-test on ΔC_t values).

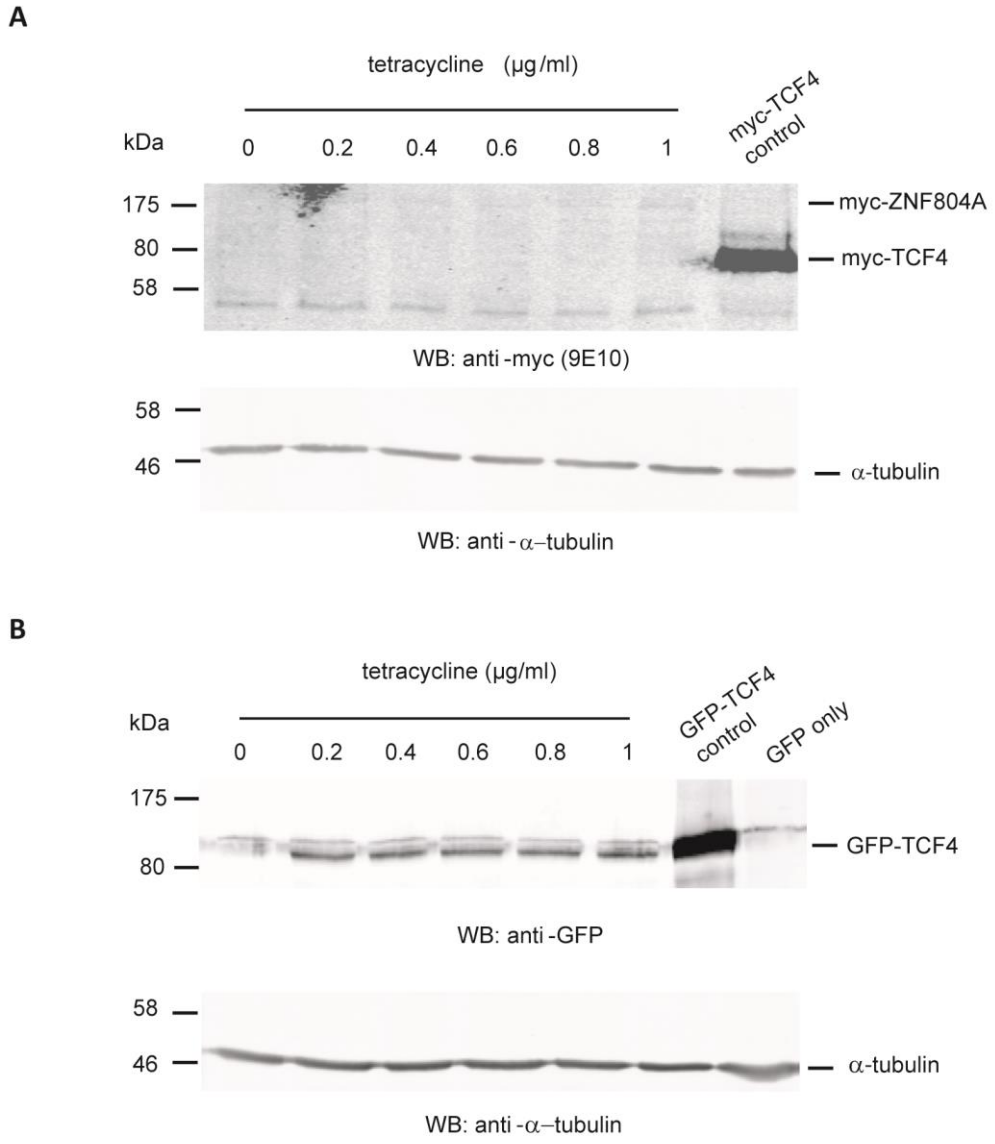


Figure 5.3 Western blot analyses of myc-ZNF804A and GFP-TCF4

Cultured myc-ZNF804A and GFP-TCF4 Flp-In TREx cells were induced with tetracycline at the concentrations shown. The following day, protein lysates were prepared. **(A)** The protein lysates were western blotted and the blot was incubated with an anti-myc antibody. A protein lysate prepared from HEK293T cells transiently expressing *myc-TCF4* served to confirm that the anti-myc antibody detected the myc epitope tag. **(B)** The protein lysates were western blotted and the blot was incubated with an anti-GFP antibody. Incubating a separate, but identical, blot with an anti-α-tubulin antibody confirmed the presence of similar amounts of protein in each sample. The antibody dilutions used are given in Table 2.5.

To evaluate whether myc-ZNF804A protein could be detected, protein lysates were prepared from tetracycline-induced myc-ZNF804A Flp-In TREx cells and western blotted (section 2.6). The blots were probed for the presence of myc-tagged proteins using an anti-myc antibody (9E10). Myc-ZNF804A migrated at 175 kDa and was only detected very weakly in induced myc-ZNF804A Flp-In TREx cells (Figure 5.3). These data are consistent with those presented in Chapter Three which show myc-ZNF804A was degraded by the proteasome and the abundance of myc-ZNF804A was extremely low. In control experiments, GFP-TCF4 was readily detected in induced GFP-TCF4 Flp-In TREx cells. This confirmed the utility of the Flp-In TREx expression system (Figure 5.3B).

To determine whether the subcellular location of myc-ZNF804A could be visualised, induced myc-ZNF804A Flp-In TREx cells were processed for immunocytochemistry using the 9E10 antibody while the nucleus was counter-stained using Hoechst (section 2.8.1). Consistent with the extremely low levels of myc-ZNF804A detected by western blot, myc-ZNF804A was not detected by immunocytochemistry (data not shown). The utility of the Flp-In TREx expression system for immunocytochemistry was confirmed using the tetracycline-induced GFP-TCF4 Flp-In TREx cell line (data not shown). Consistent with published findings, GFP-TCF4 localised homogenously throughout the nucleus (Forrest et al., 2012).

Having established that the myc-ZNF804A Flp-In TREx cells produced detectable levels of myc-ZNF804A protein, I wanted to determine if the cell line could be used to validate the Y2H data presented in Chapter Three using co-immunoprecipitation. To determine the utility of the myc-ZNF804A Flp-In TREx cells for immunoprecipitation, large-scale immunoprecipitation experiments were performed (section 2.7.2). The myc-ZNF804A Flp-In TREx cells were seeded in thirty 10 cm dishes and *myc-ZNF804A* expression was induced

with 1 µg/ml tetracycline. The following day, the cells were lysed in RIPA buffer and myc-ZNF804A was immunoprecipitated from the RIPA lysate using anti-myc antibody (9E10) - conjugated protein A beads. Transient transfection of a myc-TCF4 expression vector into four 10 cm dishes of HEK293T cells served as a control for immunoprecipitation of myc-tagged protein with 9E10-conjugated protein A beads. The immunoprecipitated proteins were western blotted and the blots were probed with the 9E10 antibody. Figure 5.4A shows that myc-ZNF804A was not detected; however myc-TCF4 was enriched after immunoprecipitation. The successful immunoprecipitation of myc-TCF4 indicates that the 9E10-conjugated protein A beads could immunoprecipitate myc-tagged proteins.

5.3. Preparation of samples for exon array

Having demonstrated that the *ZNF804A* transcript could be over-expressed in myc-ZNF804A Flp-In TREx cells, the cell line could be used to investigate potential changes in the cellular transcriptome in cells over-expressing *myc-ZNF804A* using exon array technology. Data presented in Figure 5.2A show uninduced myc-ZNF804A Flp-In TREx samples expressed *myc-ZNF804A*; therefore these samples were not appropriate negative controls. Instead, the preliminary exon array analysis was performed using wildtype HEK293T cells as the control sample because Q-PCR analysis showed that these cells did not express *ZNF804A*. Examination of these data suggested that the antibiotics used to maintain and/or induce the myc-ZNF804A Flp-In TREx cell line may influence pre-mRNA splicing. Therefore, a control, transformed Flp-In TREx cell line that expressed the hygromycin B resistance (H^R) and blasticidin resistance (B^R) genes was generated using an empty, circular pcDNA5/FRT expression vector. This cell line is referred to as the 'negative control'. In parallel, a further four clonal myc-ZNF804A Flp-In TREx cell lines were generated to control for any potential effects of the experimental conditions used to generate the stable cell lines on the cellular

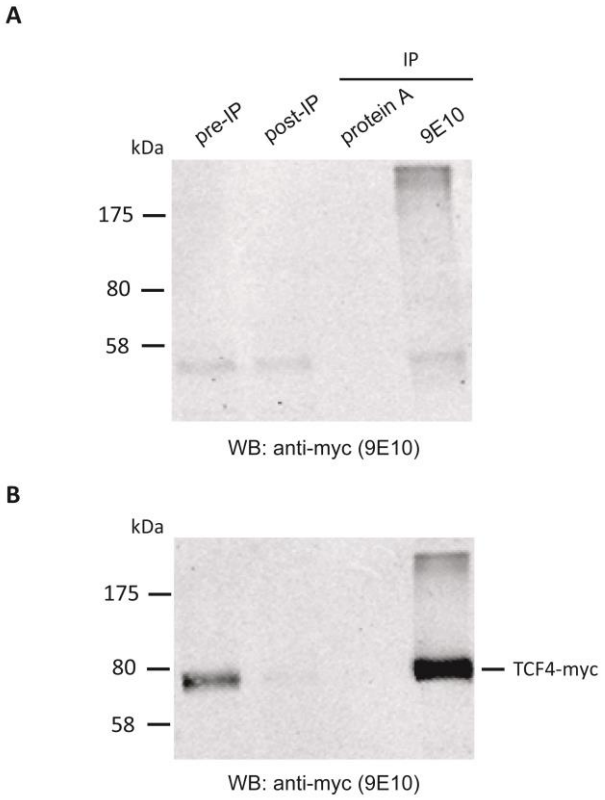


Figure 5.4 Immunoprecipitation of myc-tagged proteins

Expression of *myc-ZNF804A* was induced in myc-ZNF804A Flp-In TREx cells using 1 µg/ml of tetracycline. HEK293T cells transiently expressing *myc-TCF4* served as a positive control for the immunoprecipitation. The following day, RIPA extracts were prepared. Myc-ZNF804A and myc-TCF4 were immunoprecipitated from RIPA extracts using anti-myc antibody (9E10) -conjugated Protein A beads. A portion of the eluted immune complexes was western blotted alongside a 'Protein A only' control and pre- and post-IP samples. The blots were incubated with the 9E10 antibody. The antibody dilutions used are given in Table 2.5.

transcriptome.

To prepare the samples for exon array, the four clonal myc-ZNF804A Flp-In TREx cell lines and the four clonal negative control cell lines were seeded and 24h later, the cells were treated with 1 µg/ml tetracycline (section 2.9.1). The following day, RNA was extracted from the induced cells and 1st strand cDNA was prepared. In control reactions, the reverse transcriptase was excluded from the reaction mix to determine the levels of residual genomic DNA in the RNA samples. RT-PCR confirmed that *myc-ZNF804A* and H^R were expressed when the cells were treated with tetracycline (Figure 5.5). After 40 cycles of PCR, there was a very faint product in the no reverse transcriptase (-RT) control sample which may suggest that there was very low levels of residual genomic DNA in the RNA samples. The quality of the RNA was assessed on the Agilent 2100 bioanalyser (Central Biotechnology Services (CBS), Cardiff University). The RNA integrity number (RIN) for each sample was 10; this meant that the RNA was intact and suitable for gene expression analysis.

5.4. Processing of exon array chips and quality assessment

The RNA samples were converted into labelled cDNA for hybridisation to the Affymetrix GeneChip human exon array 1.0 ST and the arrays were processed by CBS. CBS performed an assessment of the array quality control parameters using the Affymetrix expression console; this assessment showed that all of the arrays met the required parameters for labelling and hybridisation. I received and analysed the resulting raw .CEL files which contained the signal intensity information for each probe set. The .CEL files were imported into the Partek Genomics Suite (PGS) using the core metaprobe set and RMA algorithm (section 2.9.2). The quality control metrics performed in the PGS showed that the RMA normalisation was successful (Appendix 4C) (section 2.9.3). To identify related samples,

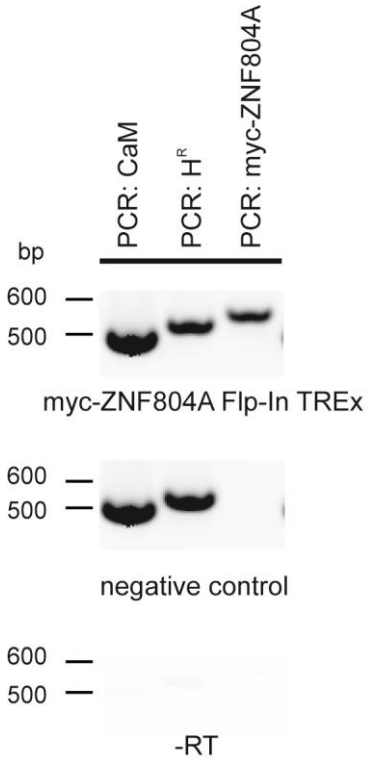


Figure 5.5 RT-PCR assessment of *myc-ZNF804A* and *hygromycin*^{R}

Cultured myc-ZNF804A Flp-In TREx and control cells were treated with 1 µg/ml tetracycline to induce gene expression. The following day, RNA was prepared. Reverse transcription (RT) reactions were performed to generate cDNA. RT reactions performed without reverse transcriptase (-RT) served as a control for residual levels of genomic DNA (gDNA). PCR was performed using primers complementary to the hygromycin resistance gene (*hygromycin*^{R}) or *myc-ZNF804A*. Primer sequences are given in Appendix 1.1 and 1.5. A representative example of the induced myc-ZNF804A and induced control cell line is shown.

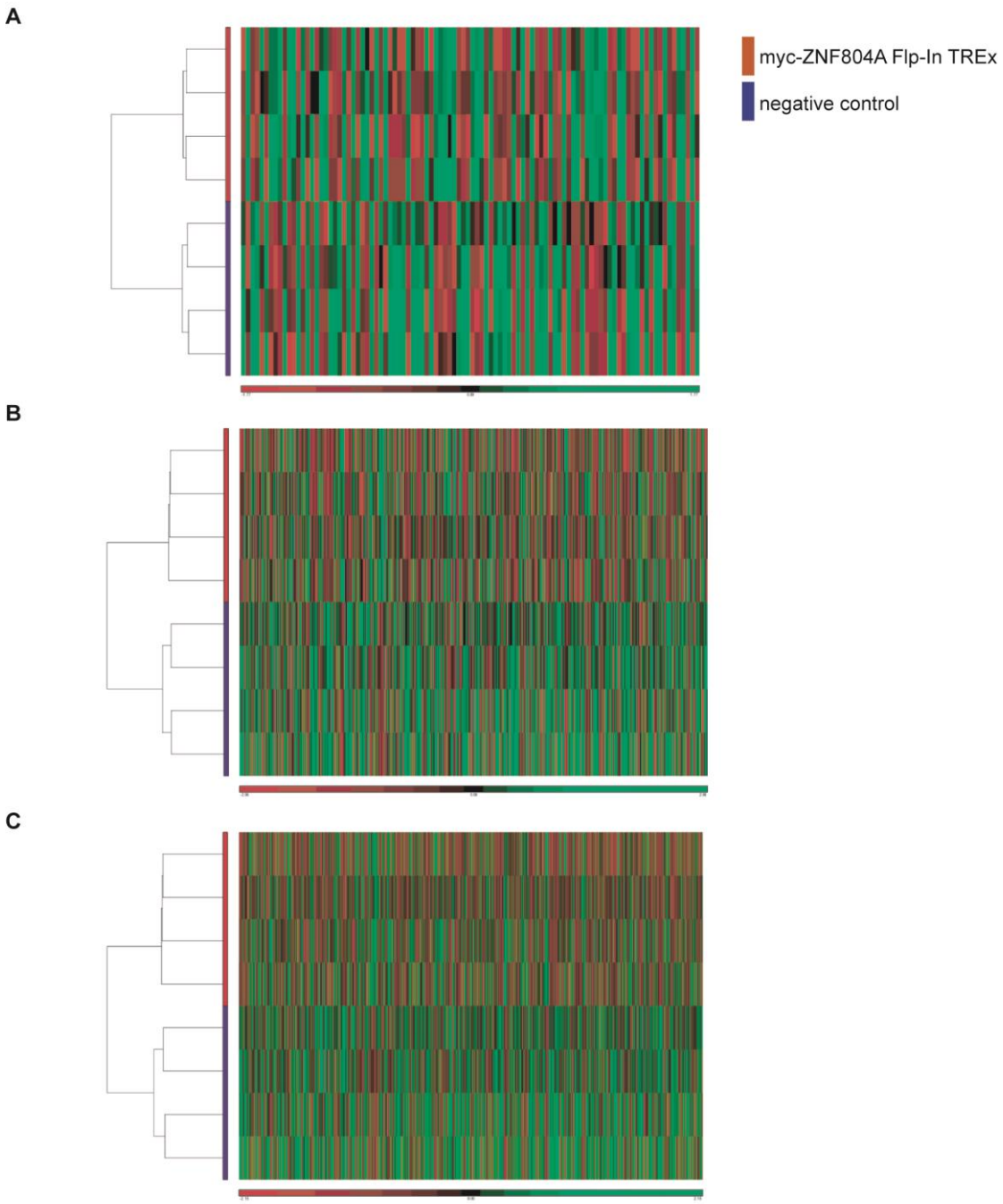


Figure 5.6 Hierarchical clustering of the most differentially expressed genes between ZNF804A-myc Flp-In TREx samples and control samples

Hierarchical clustering of the **(A)** 100, **(B)** 1000 and **(C)** 5000 most differentially expressed genes between the induced myc-ZNF804A Flp-In TREx samples (red) and negative control samples (blue). The heatmap colours are artificial: green = relative up-regulation and red = relative down-regulation black = no difference.

agglomerative hierarchical clustering was performed, and the resulting cluster dendrograms are shown in Figure 5.6. The clustering was performed using the gene-summarised expression values from the top 100, 1000 and 5000 most differentially expressed genes. The clustering patterns correlated with the over-expression of *myc-ZNF804A*. These data suggest that the cell lines over-expressing *myc-ZNF804A* were systematically different to the negative control cell lines.

5.5. Identifying genes with altered expression after *myc-ZNF804A* over-expression

To identify genes with altered expression in cells over-expressing *myc-ZNF804A*, a one-way ANOVA was performed on the gene-summarised expression values between the myc-ZNF804A Flp-In TREx and the negative control samples (section 2.9.4). The expression of 1257 genes differed with a nominally significant $P < 0.05$ (Table 5.1). Of these 1257 genes, 951 mapped to known HGNC gene symbols. After correction for multiple testing using a 5% FDR, only *ZNF804A* was statistically significantly differentially expressed (uncorrected $P = 2.58 \times 10^{-8}$; FC = 46.41) (Figure 5.7). However, the literature suggests judging whether a gene is differentially expressed by fold change rather than P value gives greater reproducibility between microarray platforms (Patterson et al., 2006; Guo et al., 2006). Moreover, from a biological perspective, differential gene expression may only be of interest if the fold change is biologically significant; this is likely to depend on the gene in question and the experimental context (McCarthy and Smyth, 2009). However, measurements of fold change do not control for variance between samples therefore, data outliers can have an effect of the fold change calculated. As such, many studies use a combination of a modest P value and a fold change cut-off to identify differentially expressed genes. For example, Patterson et al. (2006) required a gene to satisfy $P < 0.01$ or $P < 0.05$ and then used a fold change ranking of 1.5, 2 or 4. While Rahimov et al. (2012) required genes to have a fold change of greater

Gene name	Gene Symbol	P value	FC
zinc finger protein 804A	*ZNF804A*	2.58E-08	46.41
mitochondrial ribosomal protein L11	*MRPL11*	2.20E-04	1.19
chromosome 20 open reading frame 165	*C20orf165*	2.36E-04	-1.47
coiled-coil domain containing 142	*CCDC142*	2.69E-04	-1.21
tyrosylprotein sulfotransferase 2	*TPST2*	4.35E-04	1.16
zinc finger protein 208	*ZNF208*	4.80E-04	-1.68
lipase, endothelial	*LIPG*	5.68E-04	1.56
transcription factor 7-like 1 (T-cell specific, HMG-box)	*TCF7L1*	6.01E-04	-1.15
S-phase kinase-associated protein 1	*SKP1*	6.15E-04	-1.19
chromosome 5 open reading frame 45	*C5orf45*	8.67E-04	-1.22
SMAD family member 5 opposite strand	*SMAD5OS*	9.02E-04	-1.14
heterogeneous nuclear ribonucleoprotein K	*HNRNPK*	9.68E-04	1.17
family with sequence similarity 83, member D	*FAM83D*	9.71E-04	-1.07
thymine-DNA glycosylase	*TDG*	1.00E-03	1.09
ventral anterior homeobox 1	*VAX1*	1.13E-03	1.15
purinergic receptor P2X, ligand-gated ion channel, 5	*P2RX5*	1.14E-03	1.66
zinc finger protein 793	*ZNF793*	1.17E-03	-1.32
family with sequence similarity 40, member A	*FAM40A*	1.22E-03	-1.14
toll-like receptor adaptor molecule 1	*TICAM1*	1.29E-03	-1.16
Niemann-Pick disease, type C1	*NPC1*	1.52E-03	1.17
Kv channel interacting protein 4	*KCNIP4*	1.55E-03	-1.08
chromosome 3 open reading frame 36	*C3orf36*	1.58E-03	-1.20
PIH1 domain containing 2	*PIH1D2*	1.58E-03	1.26
aminopeptidase-like 1	*NPEPL1*	1.60E-03	-1.27
apolipoprotein L, 6	*APOL6*	1.60E-03	1.16
SWI/SNF related, matrix associated, actin dependent regulator of chromatin, subfamily c, member 1	*SMARCC1*	1.65E-03	1.07
N-acetylneuraminic acid synthase	*NANS*	1.68E-03	1.27
collagen, type XII, alpha 1	*COL12A1*	1.72E-03	-1.64
acyl-CoA binding domain containing 7	*ACBD7*	1.79E-03	1.56

Table 5.1 The top 30 differentially expressed genes after *myc-ZNF804A* over-expression

The .CEL files were imported into the PGS using the core metaprobe set and RMA normalisation. The .CEL files were assigned as either myc-ZNF804A Flp-In TREx or negative control. The genes with changes in expression associated with *myc-ZNF804A* over-expression were detected using a one-way ANOVA with type as the candidate variable in the ANOVA model. The gene list represents the top 30 nominally differentially expressed genes (uncorrected $P < 0.05$) with known HGNC gene symbols, sorted by P value.

Figure 5.7 The PGS 'geneview' of *ZNF804A* in cells over-expressing *myc-ZNF804A*

One-way ANOVA detected differential expression of *ZNF804A* between myc-ZNF804A Flp-In TREx samples and control samples. The individual probe set $\log_2$ signal intensities across the *ZNF804A* transcript were plotted in the PGS. The first three 5' probe sets were unaltered because these probe sets fell within the untranslated region of *ZNF804A* whereas the *myc-ZNF804A* sequence which was cloned into the pcDNA5/FRT/TO vector began at the start codon.

Gene name	Gene Symbol	P value	FC
zinc finger protein 804A	*ZNF804A*	2.58E-08	46.41
zinc finger protein 208	*ZNF208*	4.80E-04	-1.68
lipase, endothelial	*LIPG*	5.68E-04	1.56
purinergic receptor P2X, ligand-gated ion channel, 5	*P2RX5*	1.14E-03	1.66
collagen, type XII, alpha 1	*COL12A1*	1.72E-03	-1.64
acyl-CoA binding domain containing 7	*ACBD7*	1.79E-03	1.56
secreted protein, acidic, cysteine-rich (osteonectin)	*SPARC*	2.62E-03	5.28
pentraxin 3, long	*PTX3*	5.50E-03	1.61
tetratricopeptide repeat domain 30B	*TTC30B*	7.37E-03	1.52
Zic family member 1 (odd-paired homolog, Drosophila)	*ZIC1*	7.39E-03	-2.43
hypothetical LOC84856	*LOC84856*	7.53E-03	-1.83
inactivation escape 1 (non-protein coding)	*INE1*	9.64E-03	-1.58
zinc finger protein, X-linked	*ZFX*	1.06E-02	-1.52
PCI domain containing 2	*PCID2*	1.18E-02	-1.51
spondin 1, extracellular matrix protein	*SPON1*	1.82E-02	2.15
sprouty homolog 4 (Drosophila)	*SPRY4*	2.03E-02	1.87
NAD(P)H dehydrogenase, quinone 1	*NQO1*	2.16E-02	1.66
zinc finger protein 738	*ZNF738*	3.07E-02	-4.83
TLC domain containing 1	*TLCD1*	3.29E-02	1.62
hairy/enhancer-of-split related with YRPW motif 1	*HEY1*	3.69E-02	-1.77
v-myb myeloblastosis viral oncogene homolog (avian)	*MYB*	3.76E-02	1.60
ryanodine receptor 2 (cardiac)	*RYR2*	4.03E-02	-1.63
interferon regulatory factor 8	*IRF8*	4.22E-02	1.71
paired related homeobox 1	*PRRX1*	4.41E-02	1.67
chromosome 7 open reading frame 58	*C7orf58*	4.66E-02	1.58

Table 5.2 The genes that had a fold change in expression greater than +/- 1.5 when *myc-ZNF804A* was over-expressed

The .CEL files were imported into the PGS using the core metaprobe set and RMA normalisation. The .CEL files were assigned as either myc-ZNF804A Flp-In TREx or negative control. The genes with changes in expression associated with *myc-ZNF804A* over-expression were detected using a one-way ANOVA with type as the candidate variable in the ANOVA model. The gene list represents the nominally differentially expressed genes ($P < 0.05$) with known HGNC gene symbols and fold change of less than -1.5 or greater than 1.5.

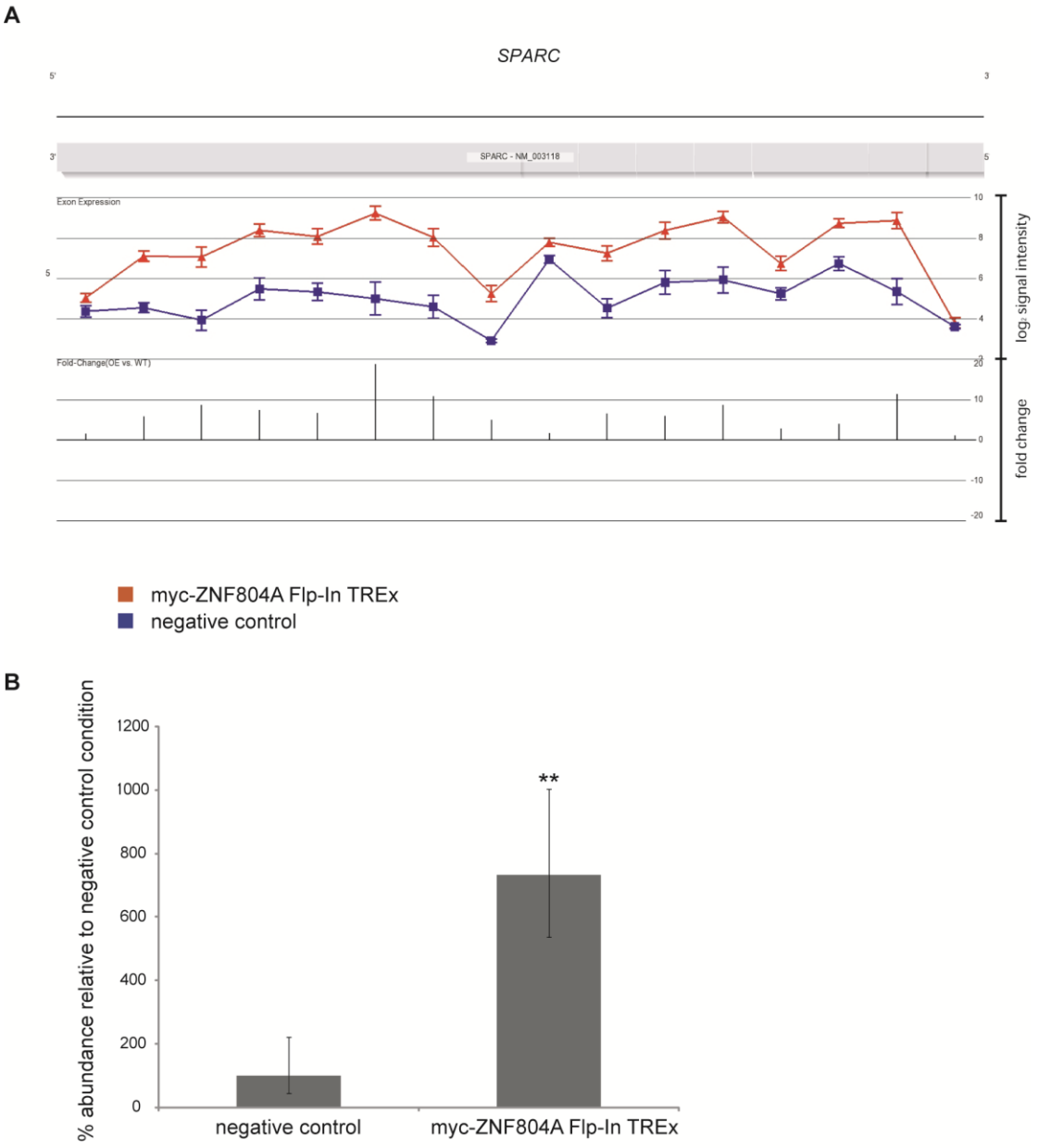


Figure 5.8 *SPARC* was up-regulated in cells over-expressing *myc-ZNF804A*

(A) One-way ANOVA detected differential expression of *SPARC* between myc-ZNF804A Flp-In TREx samples and negative control samples ($P = 2.62 \times 10^{-3}$; FC = 5.28). The individual probe set $\log_2$ signal intensities across the *SPARC* transcript were plotted in the PGS alongside the fold change in $\log_2$ signal intensity for each of the probe sets. **(B)** Q-PCR analysis of *SPARC* mRNA abundance (n = 4). Q-PCR was performed using primers designed complementary to *SPARC*. The primer sequences are given in Appendix 1.6. The Q-PCR reactions were performed in duplicate and the C_t values were normalised to *beta actin* (*ACTB*). The error bars represent the standard deviation of the raw C_t values. ** denotes $P < 0.001$ using independent samples T-test on the ΔC_t values.

than 1.2 and a nominal P < 0.01. Therefore, to identify expression changes which did not withstand multiple test correction but may still have biological relevance to the over-expression of *myc-ZNF804A*, a filter for fold change was applied. Only genes with nominal differential expression (P < 0.05) and known HGNC gene symbol were included in the analysis. All 951 genes had a fold change greater than 1 or less than -1 so a fold change greater than 1.5 or less than -1.5 was used. Twenty-five genes fulfilled these criteria (Table 5.2). Of these, four genes had a fold change greater than 2 or less than -2: *ZNF804A*; *SPARC*; *Zic family member 1 (odd-paired homolog, Drosophila)* (*ZIC1*) and *spondin 1, extracellular matrix protein* (*SPON1*). *SPARC* was up-regulated five-fold in cells over-expressing *myc-ZNF804A* (2.62×10^{-3}; FC = 5.28). Q-PCR analysis confirmed the up-regulation of *SPARC* when *myc-ZNF804A* was over-expressed (Figure 5.8). The up-regulation of *SPARC* is particularly notable because data presented in Chapter Four show *SPARC* was also up-regulated after *ZNF804A* knockdown (section 4.5.5; Figure 4.11).

5.5.1. Enrichment analysis of differentially expressed genes after *myc-ZNF804A* over-expression

To identify biological processes which were enriched for *ZNF804A*-related target genes, enrichment analysis was performed using GeneGo MetaCore™ software (section 2.10). Of the 1256 differentially expressed genes (excluding *ZNF804A*), 948 were annotated in GeneGo MetaCore™. Genes belonging to the GO processes 'regulation of RNA biosynthetic process' (250/3710 $P = 2.23 \times 10^{-13}$) and 'regulation of transcription, DNA-dependent' (249/3692 $P = 2.32 \times 10^{-13}$) were significantly enriched among the differentially expressed genes (Figure 5.9A). After 5% FDR correction for multiple testing, no GeneGo process networks were statistically significantly enriched among the differentially expressed genes. However, a number of GeneGo pathway maps were significantly enriched among the

differentially expressed genes (Figure 5.9B). GeneGo pathway maps reflect consensus knowledge on specific functional components in the intracellular cell signaling, regulatory processes, metabolic processes or disease-related processes. The most statistically significant GeneGo pathway map was 'development_regulation of epithelial-to-mesenchymal transition (EMT)' (10/64 $P = 2.29 \times 10^{-5}$) (Figure 5.10A). The differentially expressed genes belonging to the top 10 statistically significant GeneGo pathway maps are listed in Table 5.3. This shows that of the 10 genes belonging to 'development_regulation of epithelial-to-mesenchymal transition (EMT)', four were also annotated in 'development_TGF-beta-dependent induction of EMT via RhoA, PI3K and ILK' (8/46 $P = 6.95 \times 10^{-5}$; Figure 5.10B). Table 5.3 highlights that some of the differentially expressed genes belonged to multiple significantly enriched pathway maps. For example, *ras homolog gene family, member A* (*RHOA*), which encodes a small guanosine triphosphatase (GTPase) protein that is known to regulate the actin cytoskeleton, belonged to five pathway maps: 'cell adhesion_chemokines and adhesion'; 'development_TGF-beta-dependent induction of EMT via RhoA, PI3K and ILK'; 'development_Role of IL-8 in angiogenesis'; 'neurophysiological process_receptor-mediated axon growth repulsion' and 'cell adhesion_histamine H1 receptor signaling in the interruption of cell barrier integrity'.

5.6. Identifying changes in pre-mRNA splicing after *myc-ZNF804A* over-expression

To investigate the effect of over-expressing *myc-ZNF804A* on pre-mRNA splicing, an alternative splicing one-way ANOVA was performed (section 2.9.6). To ensure that only transcripts that were expressed in at least one sample group were analysed, only probe sets with a $\log_2$ signal intensity greater than three were included in the analysis. To reduce the number of false-positive results, the gene list was filtered to exclude any gene with less than five probe sets or gene expression changes greater than five-fold and a conservative

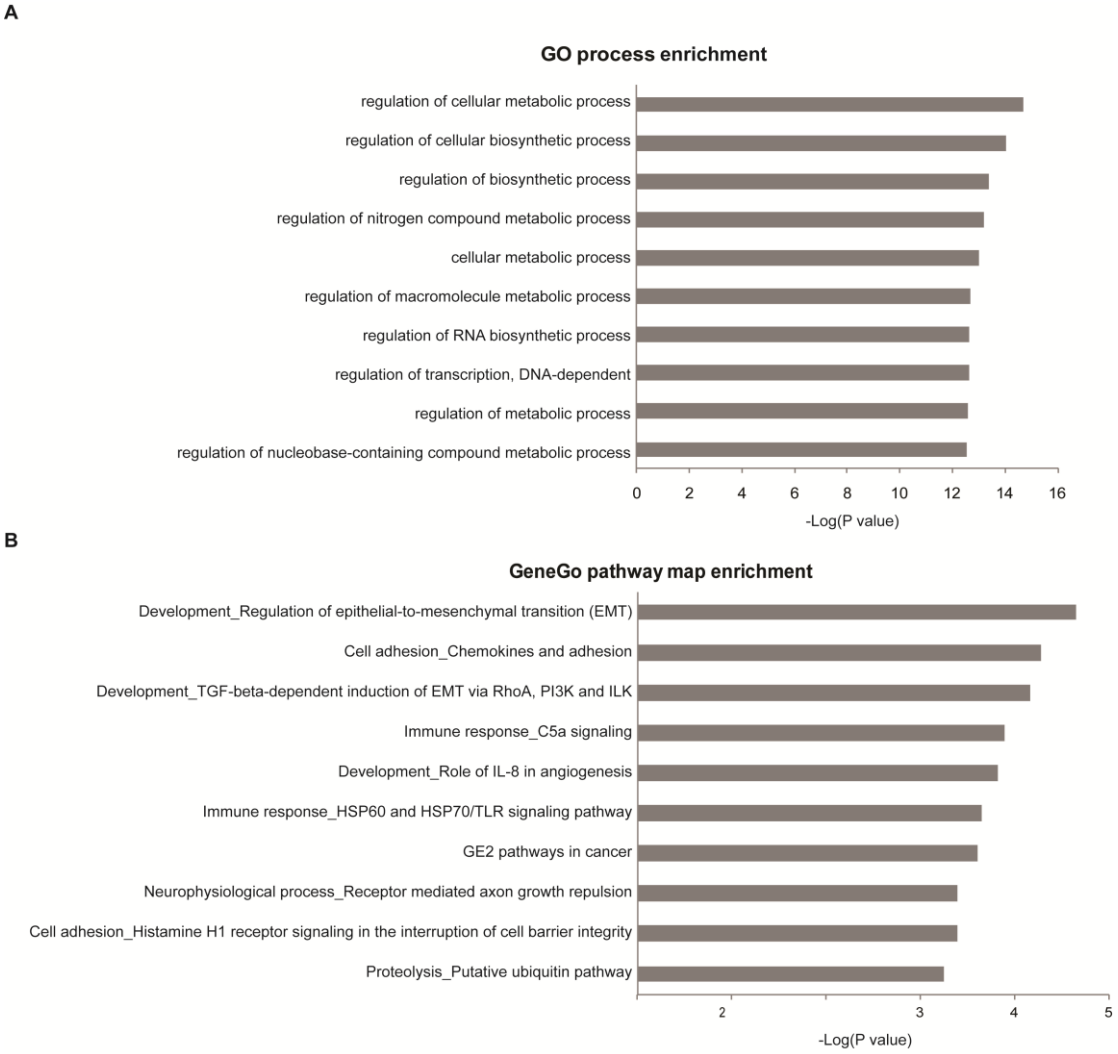
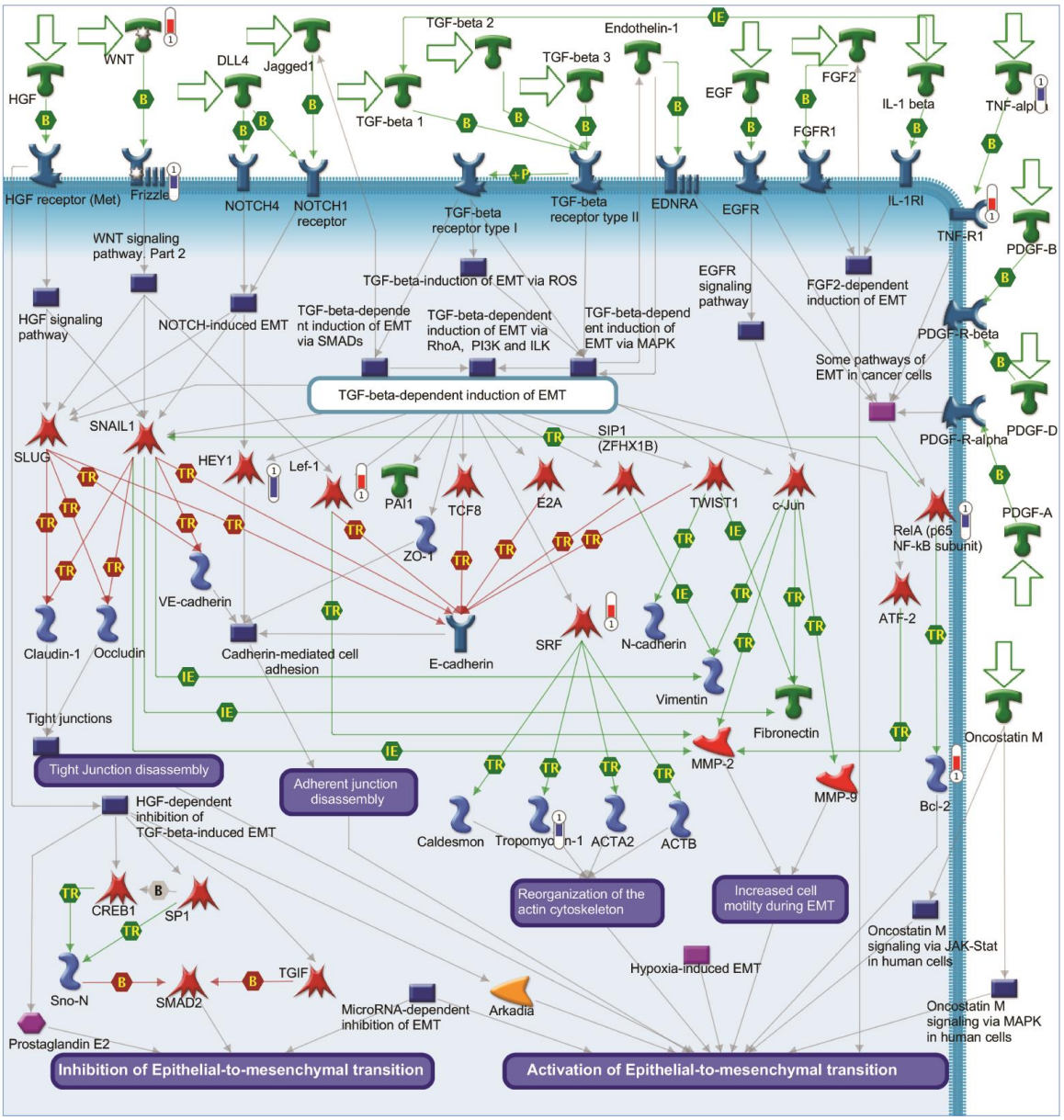


Figure 5.9 The top GO processes and GeneGo pathway maps that were significantly enriched for genes showing differential expression in cells over-expressing *myc-ZNF804A*

The list of 1256 genes (excluding *ZNF804A*) differentially expressed between myc-ZNF804A Flp-In TREx samples and negative control samples (one-way ANOVA, nominal $P < 0.05$) was imported into GeneGo MetaCore™. Enrichment analysis was performed as described in section 2.10. The bar charts show **(A)** the top 10 statistically significant GO processes and **(B)** the top 10 statistically significant GeneGo pathway maps identified by the enrichment analysis. GeneGo MetaCore™ determined these enrichments to be statistically significant at a 5% FDR; the uncorrected P values are shown.

A


WNT
DLL4
Jagged1
TGF-beta 2
TGF-beta 3
Endothelin-1
EGF
FGF2
IL-1 beta
TNF-alpha
HGF
TGF-beta 1
FGFR1
HGF receptor (Met)
Frizzled
NOTCH4
NOTCH1 receptor
TGF-beta receptor type I
TGF-beta receptor type II
EDNRA
EGFR
IL-1RI
TNFR1
PDGF-B
WNT signaling pathway, Part 2
TGF-beta-induction of EMT via ROS
EGFR signaling pathway
FGF2-dependent induction of EMT
HGF signaling pathway
NOTCH-induced EMT
TGF-beta-dependent induction of EMT via SMADs
TGF-beta-dependent induction of EMT via RhoA, PI3K and ILK
TGF-beta-dependent induction of EMT via MAPK
PDGF-R-beta
Some pathways of EMT in cancer cells
TGF-beta-dependent induction of EMT
PDGF-D
PDGF-R-alpha
SNAIL1
SLUG
SIP1 (ZFHX1B)
HEY1
Lef-1
PAI1
TCF8
E2A
TWIST1
c-Jun
RelA (p65 NF-kB subunit)
PDGF-A
ZO-1
VE-cadherin
ATF-2
Claudin-1
Occludin
Cadherin-mediated cell adhesion
E-cadherin
SRF
N-cadherin
Vimentin
Fibronectin
Oncostatin M
Tight junctions
MMP-2
Tight Junction disassembly
Adherent junction disassembly
MMP-9
Bcl-2
HGF-dependent inhibition of TGF-beta-induced EMT
Caldesmon
Tropomyosin-1
ACTA2
ACTB
CREB1
SP1
Reorganization of the actin cytoskeleton
Increased cell motilty during EMT
Oncostatin M signaling via JAK-Stat in human cells
Hypoxia-induced EMT
Sno-N
SMAD2
TGIF
MicroRNA-dependent inhibition of EMT
Arkadia
Oncostatin M signaling via MAPK in human cells
Prostaglandin E2
Inhibition of Epithelial-to-mesenchymal transition
Activation of Epithelial-to-mesenchymal transition

B

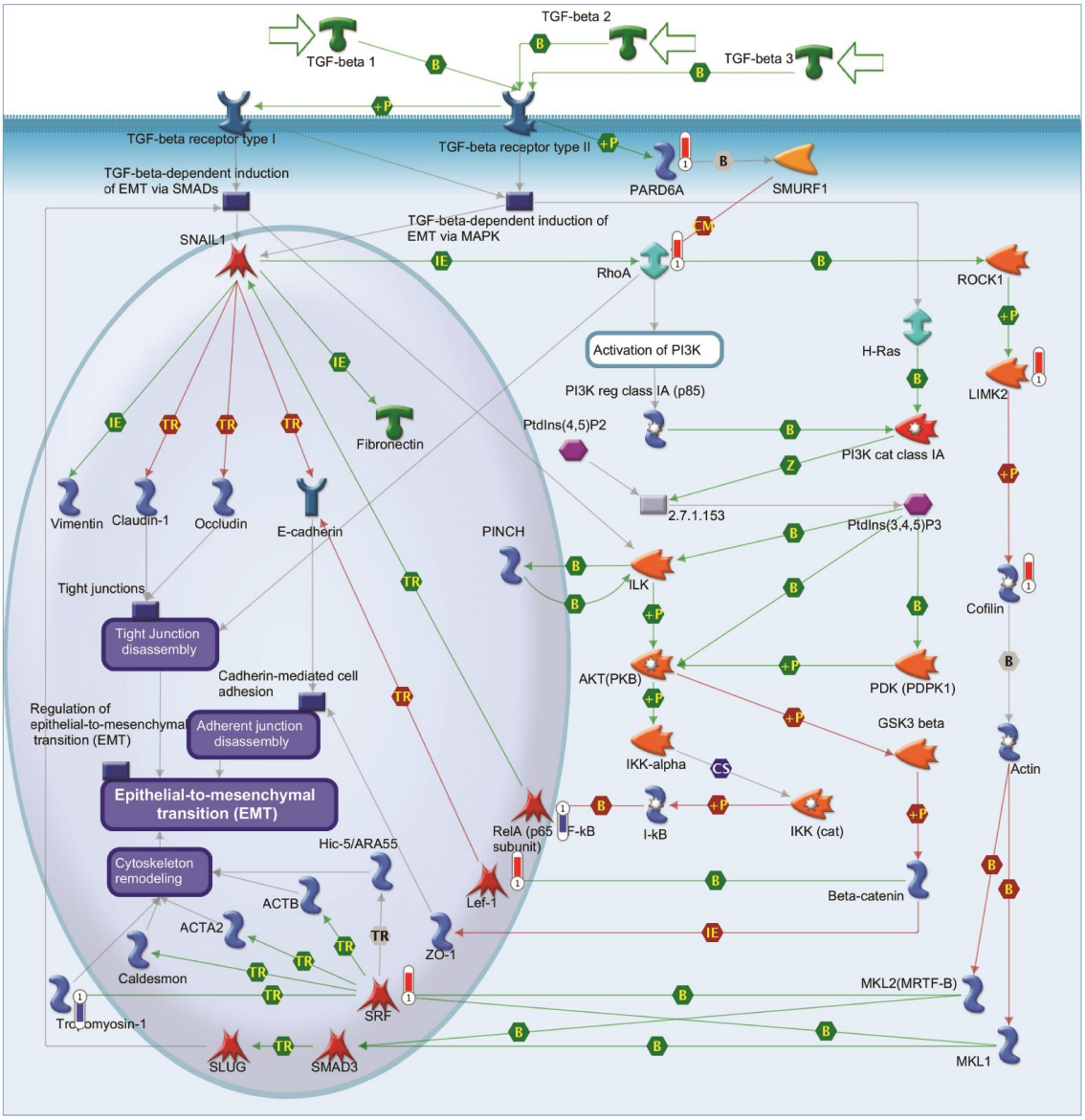


Figure 5.10 Visual representation of GeneGo pathway maps enriched among the differentially expressed genes. Enrichment analysis for GeneGo pathway maps was performed using GeneGo MetaCore™ (refer to Figure 5.9 for details). **(A)** Ten genes belonging to the pathway map 'development_regulation of epithelial-to-mesenchymal transition (EMT)' and **(B)** eight genes belonging to the pathway map 'development_TGF-beta-dependent induction of EMT via RhoA, PI3K and ILK' were differentially expressed in cells over-expressing *myc-ZNF804A*. The target genes are represented by red (up-regulation) or blue (down-regulation) bars. The GeneGo pathway map legend is provided in Appendix 6.

GeneGo pathway map	Active data
Development_Regulation of epithelial-to-mesenchymal transition (EMT)	*hairy/enhancer-of-split related with YRPW motif 1* (*HEY1*); *tropomyosin 1 (alpha)* (*TPM1*), *v-rel reticuloendotheliosis viral oncogene homolog A (avian)* (*RELA*); *tumor necrosis factor receptor superfamily, member 1A* (*TNFRSF1A*); *wingless-type MMTV integration site family, member 7B* (*WNT7B*); *lymphoid enhancer-binding factor 1* (*LEF1*); *serum response factor (c-fos serum response element-binding transcription factor)* (*SRF*); *frizzled-related protein* (*FRZB*); *B-cell CLL/lymphoma 2* (*BCL2*); *tumor necrosis factor* (*TNF*)
Cell adhesion_Chemokines and adhesion	*ras homolog gene family, member A* (*RHOA*); G-protein beta/gamma* (*guanine nucleotide binding protein (G protein), β polypeptide 1* and *2* (*GNB1* and *GNB2*); *guanine nucleotide binding protein (G protein), gamma 12* and *13* (*GNG13* and *GNG12*)); *vascular endothelial growth factor A* (*VEGF-A*); *NFKB activating protein* (*NKAP*); *LIM domain kinase 2* (*LIMK2*); *collagen, type IV, alpha 4* (*COLA4A*); *moesin* (*MSN*); *transcription factor 7-like 1* (*TCF7L1*); *laminin, alpha 4* (*LAMA4*); *syndecan 2* (*SDC2*); *paxillin* (*PXN*); *cofilin 1* (*CFL1*)
Development_TGF-beta-dependent induction of EMT via RhoA, PI3K and ILK.	*RHOA*; *TPM1*; *RELA*; *par-6 partitioning defective 6 homolog alpha (C. elegans)* (*PARD6A*); *LIMK2*; *LEF1*; *SRF*; *CFL1*
Immune response_C5a signaling	G-protein beta/gamma* (*GNB1*, *GNB2*, *GNG13*, *GNG12*); *mitogen-activated protein kinase 9* (*MAPK9*); *NKAP*; PKC* (*protein kinase D3* (*PRKD3*)); *phospholipase C, beta 1 (phosphoinositide-specific)* (*PLCB1*); *ribosomal protein S, kinase, 90kDa* (*RPS6KA1*); *BCL2*; *TNF*
Development_Role of IL-8 in angiogenesis	*RHOA*; *sterol regulatory element binding transcription factor 1* (*SREBF1*); *janus kinase 2* (*JAK2*); *VEGF-A*; *ubiquitin-conjugating enzyme E2N (UBC13 homolog, yeast)* (*UBE2N*); *NKAP*; *PXN*
Immune response_HSP60 and HSP70/ TLR signaling pathway	*CD80 molecule* (*CD80*); *MAPK9*; *UBE2N*; *NKAP*; *CD14 molecule* (*CD14*); *heat shock 60kDa protein 1 (chaperonin)* (*HSPD1*); *TNF*
PGE2 pathways in cancer	*prostaglandin E receptor 2 (subtype EP2), 53kDa* (*PTGER2*); G-protein beta/gamma* (*GNB1*, *GNB2*, *GNG13*, *GNG12*); *VEGF-A*; *TCF7L1*; *LEF1*; *solute carrier organic anion transporter family, member 2A1* (*SLCO2A1*); *TNF*
Neurophysiological process_Receptor-mediated axon growth repulsion	*RHOA*; Ephrin-A receptors* (*EPH receptor A6* (*EPHA6*) and *EPH receptor A2* (*EPHA2*)); *microtubule-associated protein tau* (*MAPT*); *LIMK2*; *Rho family GTPase 1* (*RND1*); *CFL1*
Cell adhesion_Histamine H1 receptor signaling in the interruption of cell barrier integrity	*RHOA*; G-protein beta/gamma* (*GNB1*, *GNB2*, *GNG13*, *GNG12*); *LIMK2*; *PLCB1*; *myosin, light chain 9, regulatory* (*MYL9*); *PXN*; *CFL1*
Proteolysis_Putative ubiquitin pathway	*cullin 1* (*CUL1*); *UBE2N*; *ubiquitin-like modifier activating enzyme 1* (*UBA1*); *S-phase kinase-associated protein 1* (*SKP1*); *parkinson protein 2, E3 ubiquitin protein ligase (parkin)* (*PARK2*)

Table 5.3 The active data in the top 10 statistically enriched GeneGo pathway maps

The differentially expressed genes that correspond to the active data in the top 10 statistically significant GeneGo pathway maps (Figure 5.9B). * denotes a protein family that is encoded by more than one gene; all differentially expressed genes which relate to the protein family are listed.

Bonferroni correction for multiple testing was applied. After Bonferroni correction, an arbitrary cut-off of $P < 0.05$ was chosen to generate a manageable list of statistically significant alternatively spliced genes (Whistler, 2010).This analysis identified 27 genes as alternatively spliced (Table 5.4). The putative alternative splicing events were manually annotated alongside the UCSC genome browser (data presented as supplementary information in Appendix 6.2). Four genes were identified as both statistically significantly differentially expressed ($P < 0.05$; FC +/- 1.5; Table 5.2) and alternatively spliced: *ZIC1*; *lipase, endothelial* (*LIPG*); *hypothetical LOC84856* (*LOC84856*) and *collagen, type XII, alpha 1* (*COL12A1*). Visual inspection of the PGS geneviews alongside the UCSC genome browser suggested that there may be potential novel alternative splicing events in *COL12A1* and *LIPG* and a potential alternative transcription start site in *ZIC1*. However, there was no identifiable alternative splicing in *LOC84856*. Three of the putative splicing events corresponded with known alternatively spliced exons: *O- linked N-acetylglucosamine (GlcNAc) transferase* (*OGT*), *nucleoporin like 1* (*NUPL1*) and *mucosa associated lymphoid tissue lymphoma translocation gene 1* (*MALT1*). However, preliminary RT-PCR analysis using primers designed complementary to the constitutive exons flanking the putatively spliced exons failed to amplify more than one transcript variant for these genes therefore, further empirical analysis is needed to determine if these are true alternative splicing events.

5.6.1. Enrichment analysis of genes showing alternative splicing after *myc-ZNF804A* over-expression

To identify biological processes for genes differentially spliced after *myc-ZNF804A* over-expression, enrichment analysis was performed using GeneGo MetaCore™ software. Of the 27 genes identified as alternatively spliced (section 5.6), all 27 were annotated in GeneGo MetaCore™. Genes belonging to the GeneGo process network ‘cell_adhesion_cell-matrix

Gene name	Gene Symbol	Bonferroni corrected alternative splicing P value
collagen, type XII, alpha 1	*COL12A1*	2.31E-09
Zic family member 1 (odd-paired homolog, Drosophila)	*ZIC1*	1.75E-06
O-linked N-acetylglucosamine (GlcNAc) transferase	*OGT*	2.53E-06
ryanodine receptor 2 (cardiac)	*RYR2*	3.30E-06
cyclin-dependent kinase 16	*CDK16*	6.66E-06
SET domain containing (lysine methyltransferase) 7	*SETD7*	8.31E-06
interleukin enhancer binding factor 3, 90kDa	*ILF3*	1.25E-05
protein phosphatase methylesterase 1	*PPME1*	1.88E-05
nucleoporin like 1	*NUPL1*	1.83E-04
LPS-responsive vesicle trafficking, beach and anchor containing	*LRBA*	1.35E-03
hypothetical LOC84856	*LOC84856*	1.99E-03
mucosa associated lymphoid tissue lymphoma translocation gene 1	*MALT1*	2.76E-03
NADH dehydrogenase (ubiquinone) Fe-S protein 2, 49kDa (NADH-coenzyme Q reductase)	*NDUFS2*	3.33E-03
zinc finger protein 536	*ZNF536*	3.94E-03
mediator complex subunit 1	*MED1*	4.30E-03
feline leukemia virus subgroup C cellular receptor 1	*FLVCR1*	5.87E-03
transcription factor 7-like 2 (T-cell specific, HMG-box)	*TCF7L2*	6.00E-03
collagen, type II, alpha 1	*COL2A1*	1.39E-02
kinesin family member 22	*KIF22*	1.66E-02
CD44 molecule (Indian blood group)	*CD44*	2.29E-02
erythrocyte membrane protein band 4.1-like 2	*EPB41L2*	2.40E-02
claudin domain containing 1	*CLDND1*	2.51E-02
lipase, endothelial	*LIPG*	2.63E-02
myelin expression factor 2	*MYEF2*	3.30E-02
GTPase, IMAP family member 7	*GIMAP7*	3.89E-02
sprouty homolog 1, antagonist of FGF signalling (Drosophila)	*SPRY1*	3.90E-02
leucine rich repeat containing 7	*LRRC7*	4.42E-02

Table 5.4 Genes identified as alternatively spliced in cells over-expressing *myc-ZNF804A*

The .CEL files were imported into the PGS using the core metaprobe set and RMA normalisation. The .CEL files were assigned as either myc-ZNF804A Flp-In TREx or negative control. The genes with changes in splicing associated with *ZNF804A* over-expression were detected using an alternative splicing one-way ANOVA. Alternative splicing ANOVA P values were corrected using the Bonferroni method. After Bonferroni correction, an arbitrary cut-off of $P < 0.05$ was chosen to generate a manageable list of statistically significant alternatively spliced genes. Using this cut-off, 27 genes were identified as alternatively spliced.

interactions' were significantly enriched among the alternatively spliced genes (6/207 P = 1.65 x 10^{-5}). This was the only GeneGo process network which remained significantly enriched after 5% FDR correction for multiple testing. The alternatively spliced genes belonging to this GeneGo process network were *CD44 molecule (Indian blood group)* (*CD44*), *collagen, type II, alpha 1* (*COL2A1*) and *COL12A1*. *CD44* mapped to four nodes within the network.

5.7. Discussion

The aim of the experiments described in this Chapter was to over-express *myc-ZNF804A* in mammalian cells and profile the ensuing genome-wide changes for differential gene expression and pre-mRNA splicing changes using Affymetrix GeneChip human exon 1.0 ST arrays. The hypothesis was that over-expression of *myc-ZNF804A* would lead to changes in the expression and pre-mRNA splicing of transcripts regulated, directly or indirectly, by ZNF804A. The data presented herein suggest that over-expression of *myc-ZNF804A* altered the expression of numerous genes but had relatively few effects on pre-mRNA processing.

The impetus to generate tetracycline-inducible cell lines expressing *myc-ZNF804A* came from unsuccessful attempts to transiently over-express myc-ZNF804A in mammalian cells (C. L. Tinsley, personal communication). It was hypothesised that the inability to detect exogenous ZNF804A may indicate that over-expression of *ZNF804A* was cytotoxic or that ZNF804A may be produced at extremely low levels. An alternative hypothesis is that the cDNA was only taken-up by a few cells. Previous studies have used conditional, stable expression to produce proteins that could not be produced by transient expression methods (Reeves et al., 2002). Conditional, stable expression lines are successful in such instances because they do not require high transfection efficiency, and potential negative effects of

gene expression on cell survival may be combated by growing the cells to confluency before inducing gene expression. Therefore, it was hypothesised that tetracycline-inducible, stable expression may generate detectable levels of myc-ZNF804A. Data presented here show the myc-ZNF804A Flp-In TREx cell line produced very low but detectable levels of myc-ZNF804A protein. In control experiments, GFP-TCF4 was produced to relatively high levels. This suggests that the low levels of myc-ZNF804A were specific to this protein and were not a consequence of using the Flp-In TREx system. Data presented in Chapter Three show myc-ZNF804A was only reliably detected after proteasome inhibition (section 3.3.2). Further studies using proteasome inhibitors will help determine if myc-ZNF804A produced by the Flp-In TREx cell line is degraded by the proteasome.

Immunoprecipitation experiments to isolate myc-ZNF804A from the protein lysates using 9E10-conjugated protein A beads were unsuccessful. Although control experiments show the 9E10-conjugated protein A beads could immunoprecipitate transiently expressed myc-tagged proteins, it remains to be determined whether myc-tagged protein could be immunoprecipitated from induced Flp-In TREx cells. One hypothesis to explain the inability to enrich myc-ZNF804A using immunoprecipitation is that the myc-ZNF804A protein adopted a conformation which rendered the myc-tag invisible to the 9E10-conjugated protein A beads. Further experiments using protein A beads conjugated with the anti-ZNF804A antibodies described in Chapter Three (D-14 and 3077) will help to evaluate this hypothesis.

Numerous genes were nominally differentially expressed in cells over-expressing *myc-ZNF804A*. However, only the differential expression of *ZNF804A* survived a multiple test correction of 5% FDR. The FDR method is often used in microarray studies to control for false-positives however, it is important to strike a balance between sensitivity and controlling

for type I errors (Pawitan et al., 2005). Rather than using a larger FDR cut-off and accepting a higher number of type I errors, fold change was used to highlight biologically meaningful changes in gene expression (McCarthy, 2009). This analysis showed that most of the differentially expressed genes had relatively small changes in expression. The gene showing the largest change in expression (aside from *ZNF804A*) was *SPARC*, which was up-regulated five-fold in cells over-expressing *myc-ZNF804A*. Importantly, this finding was confirmed by Q-PCR analysis. The Q-PCR data suggest the exon array method under-estimated the fold change in *SPARC* mRNA abundance by as much as two-fold. The standard deviation of the raw Ct values obtained from the Q-PCR analysis highlighted that there was considerable variability in *SPARC* mRNA abundance between the samples over-expressing *myc-ZNF804A*. This variability between samples may explain why differential expression of *SPARC* corresponded to a relatively less significant P value in the ANOVA on the exon array data, despite the large fold change in expression. The empirical validation of the up-regulation of *SPARC* confirms that the FDR method was too conservative, negating the identification of differentially expressed genes that could be empirically validated. This result is significant as it justifies using a nominal P value to identify the differentially expressed genes.

SPARC encodes a 32 kDa matricellular protein which is secreted by glial cells in the developing nervous system and is widely distributed with notably high abundance in synaptogenic areas (Mendis and Brown, 1994; Mendis et al., 1995; Vincent et al., 2008). SPARC has been implicated many cellular processes, including processes implicated in schizophrenia pathology, such as neuronal differentiation and synapse formation (Albrecht et al., 2012; Bhoopathi et al., 2011; Kucukdereli et al., 2011). Therefore, it is tempting to speculate that changes in *SPARC* expression, as a consequence of genetic variation in *ZNF804A*, may increase susceptibility for schizophrenia. Moreover, data presented in

Chapter Four show *SPARC* was also up-regulated in *ZNF804A*-depleted cells. It is tempting to interpret these data as further evidence that ZNF804A has a role in regulating the expression of *SPARC*.

Enrichment analysis by GO process showed genes belonging to general GO processes such as 'regulation of cellular metabolic process' and 'regulation of biosynthetic process' were enriched among the differentially expressed genes. There was notable redundancy among the most significantly enriched GO processes, for example 'regulation of cellular metabolic process', 'cellular metabolic process' and 'metabolic process' were all in the top 10 significantly enriched pathways. Consistent with ZNF804A's putative role in the regulation of gene expression (Chapters Three and Four), there was enrichment for genes belonging to the GO process 'regulation of transcription (DNA dependent)'.

Enrichment by GeneGo pathway map highlighted more specific biological processes which were enriched for ZNF804A-related target genes. The most significant GeneGo pathway map among the nominally differentially expressed genes was 'development_regulation of epithelial-to-mesenchymal transition (EMT)'. EMT describes the reversible process whereby epithelial cells lose cell-cell adhesions and apical-basolateral polarity, and gain mesenchymal cell characteristics such as motility and invasiveness. EMT is essential for neural tube formation in development. EMT is initiated by extracellular signals from the extracellular matrix and a number of secreted ligands (reviewed by Thiery and Sleeman, 2006). The primary and best-characterised of these inducers are members of the transforming growth factor beta (TGFβ) family (Polyak and Weinberg, 2009). TGFβ can induce EMT through multiple signalling mechanisms or via influencing the activities of other EMT-inducing signal transduction pathways, such as those involving Notch or Wnt (Bakin et al., 2000;

Massague, 2008; Miettinen et al., 1994; Timmerman et al., 2004). Data presented here show genes belonging to the GeneGo pathway map 'development_TGF-beta-dependent induction of EMT via RhoA, PI3K and ILK' were enriched among the differentially expressed genes. Importantly, these data are consistent with recent findings presented by Umeda-Yano and colleagues (2013) which showed that genes regulated by ZNF804A were associated with TGFβ signalling. The literature suggests TGFβ is the dominant extracellular signalling molecule required for axon specification in the developing brain (Yi et al., 2010). Other regulators of EMT which were differentially expressed also have additional roles in neurodevelopment. For example *HEY1*, a basic-helix-loop-helix-Orange (bHLH-O) transcriptional repressor was down-regulated in cells over-expressing *myc-ZNF804A*. HEY1 has been implicated in both the induction of EMT (Zavadil et al., 2004) and the maintenance of neural precursor cells (Sakamoto et al., 2003). These data imply over-expression of *myc-ZNF804A* alters the expression of critical EMT regulators, some of which have additional roles in nervous system development.

The enrichment analysis also identified genes belonging to the GeneGo pathway map 'neurophysiological process_receptor mediated axon growth repulsion' as enriched among the differentially expressed genes. In the developing nervous system, axons navigate to their targets by sensing attractive and repulsive molecules via receptors expressed on their growth cones. The importance of axon connectivity in nervous system development is highlighted by disorders that result from aberrant axon guidance (Engle, 2010). Genetic studies have identified association between genes implicated in aberrant axon guidance and schizophrenia, for example *Abelson helper integration site 1* (*AHI1*) (Amann-Zalcenstein et al., 2006; Ingason et al., 2007). Therefore, potential influences on axon guidance may be considered a candidate mechanism for ZNF804A's involvement in schizophrenia pathology.

Interestingly, genes belonging to the GeneGo pathway map 'proteolysis_putative ubiquitin pathway' were significantly enriched among the differentially expressed genes. This is consistent with data presented in Chapter Three that show ZNF804A was only reliably detected after the proteasome was inhibited. As discussed above, further experiments using proteasome inhibitors are required to determine if myc-ZNF804A produced by the myc-ZNF804A Flp-In TREx cell line was degraded by the proteasome.

Only 27 genes were identified as alternatively spliced in cells over-expressing *myc-ZNF804A*. This was a relatively low number of alternative splicing events; for instance in Chapter Four, 448 genes were identified as alternatively spliced (one-way ANOVA; Bonferroni corrected $P < 0.05$). Enrichment by GeneGo process network identified genes belonging 'cell_adhesion_cell-matrix interactions' as significantly enriched among the alternatively spliced genes. This could suggest a role for ZNF804A in regulating cytoskeletal structure which might be important in neuronal migration. Importantly, data presented in Chapter Four and in the literature (Hill et al., 2012a) also implicate a role for ZNF804A in cell adhesion.

In the context of the *ZNF804A*-knockdown data presented in Chapter Four, there were relatively fewer statistically significant changes in gene expression and pre-mRNA splicing in cells over-expressing *myc-ZNF804A*. A possible explanation for this finding is that unknown downstream mediators of ZNF804A's effects may not be expressed in the host cell line (HEK293) used to generate the myc-ZNF804A Flp-In TREx cell line. An alternative hypothesis is that the cells regulated the over-expression of myc-ZNF804A via the ubiquitin-proteasome pathway. While fewer statistically significant changes in the transcriptome were identified, the enrichment analyses presented herein are consistent with data presented in

Chapter Four. For example, both the over-expression and knockdown experiments implicate a role for ZNF804A in regulating the expression and splicing of genes involved in cell adhesion and axonal guidance.

Recent studies have reported that transient over-expression of *myc-ZNF804A* in rat neural progenitor cells altered the expression of four genes (*DRD2, PDE4B, COMT* and *PRSS16*) between three- and five-fold (Girgenti et al., 2012). Girgenti and colleagues evaluated the expression of these specific genes because they had been previously implicated in schizophrenia pathology. Based on the prediction that ZNF804A was a DNA transcription factor (O'Donovan et al., 2008), Girgenti and colleagues used ChIP to investigate the ability of myc-ZNF804A to bind to the promoter regions of these genes. These data showed that myc-ZNF804A bound directly to the promoter regions of *PRSS16* and *COMT* and up-regulated their expression (Girgenti et al., 2012). Using our global, hypothesis-free approach, these genes were not identified as differentially expressed in cells over-expressing *myc-ZNF804A*. This inconsistency may reflect experimental differences. Specifically, Girgenti and colleagues readily produced myc-ZNF804A in rat neural progenitor cells whereas myc-ZNF804A was only detected to very low levels in induced myc-ZNF804A Flp-In TREx cells derived from a HEK293 cell line.

In summary, the data presented in this Chapter show that over-expression of *myc-ZNF804A* using a tetracycline-inducible Flp-In TREx cell line resulted in nominal differential expression of numerous genes but had relatively few effects on pre-mRNA splicing. Enrichment analysis indicated a significant effect of *myc-ZNF804A* over-expression on genes involved in cell adhesion, axon guidance and regulating EMT during development; in particular TGFβ-induced EMT. Previous studies have indicated ZNF804A may regulate the

expression of genes involved in TGFβ signalling (Umeda-Yano et al., 2013). Therefore, further study to understand the link between ZNF804A, TGFβ signalling and the regulation of EMT may provide important insights into the function of ZNF804A and its contribution to schizophrenia pathology.

Chapter 6: General discussion

The aim of this thesis was to characterise the function of ZNF804A. This protein is encoded by the *ZNF804A* gene, which has been robustly and reproducibly implicated as a susceptibility gene for schizophrenia (O'Donovan et al., 2008; Purcell et al., 2009; Riley et al., 2009; Steinberg et al., 2011; Strange, 2012). At the outset of this study, the function of ZNF804A was unknown. The findings presented in this thesis provide convincing evidence that ZNF804A is involved in transcriptional regulation and pre-mRNA processing. Remarkably, ZNF804A-associated 'target' genes are implicated in biological processes relevant to schizophrenia pathophysiology, such as nervous system development.

Often, the first step to understanding the function of an uncharacterised protein is to establish which proteins it interacts with. Such data allows the uncharacterised protein to be placed in a functional context derived from known information about its interaction partners. In Chapter Three, Y2H screening was used to identify the putative protein binding partners of ZNF804A using human foetal brain and mouse brain cDNA libraries. The results presented show ZNF804A putatively interacted with proteins implicated in processes such as transcriptional regulation and pre-mRNA processing. For example, ZNF804A interacted with ZNF40, a known transcription factor, and RNA-binding proteins that have important roles in constitutive and alternative splicing such as RNPS1, NOVA2, Celf4, Rbfox1 and RBFOX2. The identification of these binding partners implicates a functional role for ZNF804A in transcriptional regulation and pre-mRNA processing. These data are consistent with recent, unpublished Y2H data presented by Rasko and colleagues at the 20th World Congress of Psychiatric Genetics that showed putative ZNF804A-interactors were enriched for the GO term 'RNA splicing' (Rasko et al., 2012). Furthermore, the observation that GPATCH8, a

paralog of ZNF804A, localised to nuclear speckles also implicates a role for ZNF804A in pre-mRNA processing (Chapter Three).

A key limitation of using genetic techniques such as the Y2H system to identify protein-protein interactions is that the results must be confirmed in a mammalian *in vitro* system to exclude the possibility that they are false-positive interactions. Previous attempts within our laboratory to transiently express ZNF804A had been unsuccessful (C.L. Tinsley, personal communication). Data presented in Chapter Three support these findings and show myc-ZNF804A was only reliably detected after proteasome inhibition (section 3.3.2). Data presented in Chapter Five show that the tetracycline-induced myc-ZNF804A Flp-In TREx cells produced very low but detectable levels of myc-ZNF804A (section 5.2). These data are inconsistent with data reported in the literature which show *myc-ZNF804A* expression vectors produced readily detectable levels of myc-ZNF804A in rat neural progenitor cells (Girgenti et al., 2012) and HEK293 cells (Umeda-Yano et al., 2013). To further understand ZNF804A's biological function, it will be useful to establish the nature of the post-translational modifications that target ZNF804A to the proteasome for degradation and under what cellular conditions these post-translational modifications occur.

Having established that ZNF804A putatively interacts with transcription factors and proteins involved in RNA splicing, it was hypothesised that ZNF804A, via interaction with its protein binding partners, may have a role in transcriptional regulation and pre-mRNA processing. The putative role of ZNF804A in gene expression was investigated using knockdown (Chapter Four) and over-expression (Chapter Five) experiments combined with Affymetrix GeneChip human exon 1.0 ST arrays. The hypothesis was that manipulation of *ZNF804A*

expression would lead to changes in the expression and pre-mRNA splicing of transcripts regulated, directly or indirectly, by ZNF804A.

Data presented in Chapter Four show that there were numerous changes in gene expression after *ZNF804A* knockdown. Among the most differentially expressed genes were genes which have been implicated in schizophrenia pathology, including *RELN*, *NPY* and *NLGN2* (Table 4.3). Interestingly, genes belonging to the GO term 'nervous system development' and the GeneGo process network 'cell adhesion_synaptic contact' were enriched among those showing differential expression (section 4.5.1). These processes have been implicated in schizophrenia pathophysiology (Bourgeron, 2009; Fatemi and Folsom, 2009; Melom and Littleton, 2011; Zoghbi, 2003). These data are consistent with genetic data that implicates a role for ZNF804A in other neurodevelopmental disorders (see section 1.7.1) (Griswold et al., 2012; Steinberg et al., 2011; Talkowski et al., 2012). Furthermore, the implication that ZNF804A has a role in regulating the expression of genes involved in neurodevelopment, and particularly synaptic contact, is consistent with recent genetic data suggests risk genes for schizophrenia converge at the synapse (see section 1.6.3.4) (O'Dushlaine et al., 2011; Kirov et al., 2012; Glessner et al., 2010). Importantly, using the schizophrenia and bipolar disorder PGC GWAS datasets, it was shown that the genes with altered expression in *ZNF804A*-depleted cells (FDR 0.05) were enriched for genes that are genetically associated with disease ($P < 0.05$) (section 4.5.2). This finding provides independent support for the causal involvement of ZNF804A-associated 'target' genes in disease.

Data presented in Chapter Five show that over-expression of *myc-ZNF804A* led to numerous nominally significant changes in gene expression. Enrichment analysis indicated a significant effect of *myc-ZNF804A* over-expression on the expression of genes involved in regulating

EMT during development; in particular TGFβ-dependent EMT (section 5.5.1). This is consistent with recent data presented by Umeda-Yano and colleagues (2013) which indicated that genes regulated by ZNF804A were associated with TGFβ signalling. Notably, the expression of *SPARC* was up-regulated in both the *ZNF804A* knockdown and over-expression experiments. *SPARC* encodes the protein SPARC which is associated with biological processes implicated in schizophrenia, such as neuronal differentiation and synapse development (Albrecht et al., 2012; Bhoopathi et al., 2011; Kucukdereli et al., 2011). Therefore, SPARC may be an interesting discovery target for future studies of ZNF804A.

The data presented in Chapter Four show that knockdown of wildtype *ZNF804A* altered the pre-mRNA splicing of a number of transcripts. Enrichment analysis of alternatively spliced genes indicated a significant effect of *ZNF804A* knockdown on genes involved in nervous system development, particularly synaptic contact and axonal guidance. As discussed above, these biological processes have been implicated in schizophrenia and therefore, aberrant splicing of these genes may represent candidate mechanisms for ZNF804A's role in disease. While most of the empirical validations presented in this thesis represent cassette exons, there were many complex changes in splicing pattern after *ZNF804A* knockdown, including alternative promoter usage in *PTPRR* which was empirically confirmed (section 4.6.3). Interestingly, knockdown of *ZNF804A* led to alternative splicing of exon 11a of *ENAH*, an exon whose splicing is known to be regulated by RBFOX2 (section 4.6.2) (Yeo et al., 2009). These data are consistent with data presented in Chapter Three which suggest that ZNF804A may have a role in pre-mRNA processing via interaction with its binding partners. Recent genome-wide transcriptome profiling of post-mortem brains showed a statistically significant increase in the exclusion of exon 11a of *ENAH* in the Brodmann Area 10 and caudate

associated with schizophrenia (Cohen, 2012); this finding provides independent evidence for the causal involvement of alternative splicing of *ENAH* in schizophrenia.

Data presented in Chapter Five suggest over-expression of *myc-ZNF804A* led to fewer changes in pre-mRNA splicing than in *ZNF804A*-knockdown cells. One potential explanation for this is that unknown downstream mediators of ZNF804A's effects were not expressed in the host cell line (HEK293) used to generate the myc-ZNF804A Flp-In TREx cell line. Consistent with the literature (Hill et al., 2012a) and data presented in Chapter Four, enrichment analysis of the alternatively spliced genes indicated a significant effect of *myc-ZNF804A* over-expression on genes involved in cell-matrix interactions.

The primary limitation of exon array-based methods is a heavy reliance on prior knowledge of the genome. To further evaluate the effects of *ZNF804A* on pre-mRNA processing, it may be informative to use sequence-based methods such as RNA-sequencing. Although RNA-sequencing technology is still under active development, it does have several advantages to the array-based methods; in particular, RNA-sequencing can differentiate more readily between different transcript variants and has lower background noise (Ozsolak and Milos, 2011).

It is important to highlight that the methods used here cannot indicate whether the changes in gene expression and pre-mRNA processing after *ZNF804A* knockdown or over-expression are due to direct actions of ZNF804A or due to secondary consequences. ZNF804A contains a C2H2 type ZnF domain at its N-terminus (O'Donovan, 2008). The C2H2 type ZnF domain can bind to DNA, RNA and proteins and is prevalent in known transcription factors (Matthews and Sunde, 2002). Interestingly, although ZnF domain-containing proteins may

contain between one and 40 C2H2 type ZnFs, very few proteins which use a single C2H2 type ZnF domain to bind to DNA have been identified, suggesting that at least two ZnF domains in tandem are required for DNA recognition (Iuchi, 2001). By contrast, evidence shows that a single C2H2 type ZnF domain is sufficient for RNA binding (Friesen and Darby, 2001). Therefore, it is plausible that ZNF804A may interact directly with RNA via its C2H2 type ZnF domain. Alternatively, ZNF804A may interact indirectly with RNA via interactions with its putative binding partners that can bind RNA (Chapter Three). Further experiments based on chromatin immunoprecipitation or RNA immunoprecipitation will be critical in determining the direct targets of ZNF804A. While the degradation of ZNF804A by the proteasome poses a challenge for techniques based on immunoprecipitation, it is notable that Girgenti and colleagues (2012) successfully used chromatin immunoprecipitation to show ZNF804A bound the promoter/enhancer regions of two known schizophrenia susceptibility genes in rat neural progenitor cells. If RNA immunoprecipitation experiments show ZNF804A interacts directly with pre-mRNA, it will be interesting to establish the potential mechanisms of ZNF804A-mediated splicing using minigene constructs. Minigene constructs contain a genomic segment from the gene of interest that includes the alternatively spliced region and flanking genomic regions (Cooper, 2005). Minigene reporter assays can be used to assess the cis-regulatory elements and trans-acting factors that are required for alternative splicing events.

In the context of the literature, the implication that ZNF804A may be involved in both transcriptional regulation and pre-mRNA processing seems unusual for a C2H2 type ZnF domain-containing protein. While C2H2 type ZnF domain-containing proteins are often implicated as regulators of DNA transcription, and can function at more than one level of gene expression, there are very few examples of C2H2 type ZnF domain-containing proteins

that influence pre-mRNA processing (Burdach et al., 2012). A notable example with potential similarities to ZNF804A is the C2H2 zinc-finger domain-containing protein synapse defective-9 (syd-9) which is specifically required for synaptic function in *Caenorhabditis elegans* (*C.elegans*) (Vijayaratnam et al., 2003; Wang et al., 2006). Like GPATCH8, syd-9 localises to nuclear speckles (Wang et al., 2006). Data presented by Wang and colleagues implicate a role for syd-9 in post-transcriptional modifications. Specifically, *C. elegans* null mutants for *unc-75* (the *C. elegans* homolog of the CELF/Bruno family of proteins) had less severe deficits that combined *unc-75* and *syd-9* null mutants; suggesting that syd-9 may function with unc-75 to regulate transcription and splicing (Wang et al., 2006).

In light of the data presented in this thesis, it is tempting to speculate that genetic variation in *ZNF804A* may increase risk for schizophrenia by altering ZNF804A's interactions with its binding partners and/or target genes in a manner which may have downstream consequences on gene expression and/or splicing of transcripts. Recent data presented by Hill and colleagues showed that the disease-associated SNP rs1344706 significantly reduced expression of RNA transcribed from the risk allele in human foetal brain (Hill et al., 2012b). The data presented in Chapter Four show genes belonging to the GO term 'nervous system development' were enriched among differentially expressed genes in *ZNF804A*-depleted cells. This suggests that it is feasible that a reduction in *ZNF804A* expression during foetal development could have downstream consequences on gene expression and pre-mRNA processing which could increase risk for disease.

The notion that dysregulated gene expression and splicing could contribute to schizophrenia has been documented in the literature; post-mortem brain studies of schizophrenia patients show alternative splicing of several schizophrenia candidate genes including *regulator of G-*

protein signalling 4 (*RGS4*), *neuregulin 1* (*NRG1*) and *glutamate receptor, metabotropic 3* (*GRM3*) (Ding and Hegde, 2009; Sartorius et al., 2008; Tan et al., 2007). However, very few published studies have assessed both gene-level and exon-level changes in schizophrenia (Wu et al., 2012; Cohen, 2012). Interestingly, aberrant splicing is also implicated in other neuropsychiatric diseases, such as autism. For example, using post-mortem genome-wide transcriptome profiling, Voineagu and colleagues showed a number of RBFOX1-dependent splicing events were dysregulated in autism (Voineagu et al., 2011).

In summary, the data presented in this thesis suggest that ZNF804A contributes to the transcriptional regulation and pre-mRNA splicing of genes implicated in processes which underlie schizophrenia, such as nervous system development. This work provides the basis for further investigation into how the disease-related SNP in *ZNF804A* may contribute to disease susceptibility via dysregulation of ZNF804A-associated transcriptional regulation and pre-mRNA splicing.

Affymetrix (2005). Whitepaper: Exon probe set annotations and transcript cluster groupings.

Affymetrix (2006). Technical note: Identifying and validating alternative splicing events.

Affymetrix (2007). Whitepaper: Quality assessment of exon and gene arrays.

Agatep, R., Kirkpatrick, R.D., R.A, Woods, R.A., and Gietz, R.D. (1998). Transformation of *Saccharomyces cerevisiae* by the lithium acetate/single-stranded carrier DNA/polyethylene glycol (LiAc/ss-DNA/PEG) protocol. (Technical tips online).

Aihara, K., Kuroda, S., Kanayama, N., Matsuyama, S., Tanizawa, K., and Horie, M. (2003). A neuron-specific EGF family protein, NELL2, promotes survival of neurons through mitogen-activated protein kinases. Brain Res Mol Brain Res *116*, 86-93.

Akbarian, S., and Huang, H.S. (2006). Molecular and cellular mechanisms of altered GAD1/GAD67 expression in schizophrenia and related disorders. Brain Res Rev *52*, 293-304.

Akbarian, S., Kim, J.J., Potkin, S.G., Hagman, J.O., Tafazzoli, A., Bunney, W.E., Jr., and Jones, E.G. (1995). Gene expression for glutamic acid decarboxylase is reduced without loss of neurons in prefrontal cortex of schizophrenics. Arch Gen Psychiatry *52*, 258-266.

Albrecht, D., López-Murcia, F.J., Pérez-González, A.P., Lichtner, G., Solsona, C., and Llobet, A. (2012). SPARC prevents maturation of cholinergic presynaptic terminals. Mol Cell Neurosci *49*, 364-374.

Allison, D.B., Cui, X., Page, G.P., and Sabripour, M. (2006). Microarray data analysis: from disarray to consolidation and consensus. Nat Rev Genet *7*, 55-65.

Amann-Zalcenstein, D., Avidan, N., Kanyas, K., Ebstein, R.P., Kohn, Y., Hamdan, A., Ben-Asher, E., Karni, O., Mujaheed, M., Segman, R.H., *et al.* (2006). AHI1, a pivotal neurodevelopmental gene, and C6orf217 are associated with susceptibility to schizophrenia. European Journal of Human Genetics *14*, 1111-1119.

American Psychiatric Association. (2000). Diagnostic and statistical manual of mental disorders. (4th ed., text rev.) Washington, DC

Aruga, J. (2004). The role of Zic genes in neural development. Mol Cell Neurosci *26*, 205-221

Athanasiou, M.C., Dettling, M., Cascorbi, I., Mosyagin, I., Salisbury, B.A., Pierz, K.A., Zou, W., Whalen, H., Malhotra, A.K., Lencz, T., *et al.* (2011). Candidate gene analysis identifies a polymorphism in HLA-DQB1 associated with clozapine-induced agranulocytosis. J Clin Psychiatry *72*, 458-463.

Augustine, K.A., Rossi, R.M., Silbiger, S.M., Bucay, N., Duryea, D., Marshall, W.S., and Medlock, E.S. (2000a). Evidence that the protein tyrosine phosphatase (PC12,Br7,Sl) gamma (-) isoform modulates chondrogenic patterning and growth. Int J Dev Biol *44*, 361-371.

Augustine, K.A., Silbiger, S.M., Bucay, N., Ulias, L., Boynton, A., Trebasky, L.D., and Medlock, E.S. (2000b). Protein tyrosine phosphatase (PC12, Br7,S1) family: expression characterization in the adult human and mouse. Anat Rec *258*, 221-234.

Bakin, A.V., Tomlinson, A.K., Bhowmick, N.A., Moses, H.L., and Arteaga, C.L. (2000). Phosphatidylinositol 3-kinase function is required for transforming growth factor beta-mediated epithelial to mesenchymal transition and cell migration. J Biol Chem *275*, 36803-36810.

Baronet, A.M. (1999). Factors associated with caregiver burden in mental illness: a critical review of the research literature. Clin Psychol Rev *19*, 819-841.

Barreau, C., Paillard, L., Mereau, A., and Osborne, H.B. (2006). Mammalian CELF/Bruno-like RNA-binding proteins: molecular characteristics and biological functions. Biochimie *88*, 515-525.

Barzik, M., Kotova, T.I., Higgs, H.N., Hazelwood, L., Hanein, D., Gertler, F.B., and Schafer, D.A. (2005). Ena/VASP proteins enhance actin polymerization in the presence of barbed end capping proteins. J Biol Chem *280*, 28653-28662.

Beckmann, J.S., Estivill, X., and Antonarakis, S.E. (2007). Copy number variants and genetic traits: closer to the resolution of phenotypic to genotypic variability. Nat Rev Genet *8*, 639-646.

Behjati, F., Shafaghati, Y., Firouzabadi, S.G., Kahrizi, K., Bagherizadeh, I., Najmbadi, H., Bint, S., and Ogilvie, C. (2008). M-banding characterization of a 16p11.2p13.1 tandem duplication in a child with autism, neurodevelopmental delay and dysmorphism. Eur J Med Genet *51*, 608-614.

Bemmo, A., Benovoy, D., Kwan, T., Gaffney, D.J., Jensen, R.V., and Majewski, J. (2008). Gene expression and isoform variation analysis using Affymetrix Exon Arrays. BMC Genomics *9*, 529.

Beneyto, M., Kristiansen, L.V., Oni-Orisan, A., McCullumsmith, R.E., and Meador-Woodruff, J.H. (2007). Abnormal glutamate receptor expression in the medial temporal lobe in schizophrenia and mood disorders. Neuropsychopharmacol *32*, 1888-1902.

Benjamini, Y., and Hochberg, Y. (1995). Controlling the false discovery rate - a practical and powerful approach to multiple testing. J R Stat Soc Series B-Methodological *57*, 289-300.

Bergen, S.E., O'Dushlaine, C.T., Ripke, S., Lee, P.H., Ruderfer, D.M., Akterin, S., Moran, J.L., Chambert, K.D., Handsaker, R.E., Backlund, L., *et al.* (2012). Genome-wide association study in a Swedish population yields support for greater CNV and MHC involvement in schizophrenia compared with bipolar disorder. Mol Psychiatr *17*, 880-886.

Bergen, S.E., and Petryshen, T.L. (2012). Genome-wide association studies of schizophrenia: does bigger lead to better results? Curr Opin Psychiatr *25*, 76-82.

Bhoopathi, P., Chetty, C., Dontula, R., Gujrati, M., Dinh, D.H., Rao, J.S., and Lakka, S.S. (2011). SPARC stimulates neuronal differentiation of medulloblastoma cells via the Notch1/STAT3 pathway. Cancer Res *71*, 4908-4919.

Black, D.L. (2003). Mechanisms of alternative pre-messenger RNA splicing. Annu Rev Biochem *72*, 291-336.

Black, J.E., Kodish, I.M., Grossman, A.W., Klintsova, A.Y., Orlovskaya, D., Vostrikov, V., Uranova, N., and Greenough, W.T. (2004). Pathology of layer v pyramidal neurons in the prefrontal cortex of patients with schizophrenia. Am J Psych *161*, 742-744.

Bleuler, E. (1950). Dementia Praecox or the Group of Schizophrenias (New York: International Universities Press).

Bourgeron, T. (2009). A synaptic trek to autism. Curr Opin Neurobiol *19*, 231-234.

Bray, N.J., Preece, A., Williams, N.M., Moskvina, V., Buckland, P.R., Owen, M.J., and O'Donovan, M.C. (2005). Haplotypes at the dystrobrevin binding protein 1 (DTNBP1) gene locus mediate risk for schizophrenia through reduced DTNBP1 expression. Hum Mol Genet *14*, 1947-1954.

Brown, A.S., and Patterson, P.H. (2011). Maternal Infection and Schizophrenia: Implications for Prevention. Schizophr Bull *37*, 284-290.

Brown, A.S., Susser, E.S., Butler, P.D., Richardson Andrews, R., Kaufmann, C.A., and Gorman, J.M. (1996). Neurobiological plausibility of prenatal nutritional deprivation as a risk factor for schizophrenia. J Nerv Ment Dis *184*, 71-85.

Bundy, H., Stahl, D., and MacCabe, J.H. (2011). A systematic review and meta-analysis of the fertility of patients with schizophrenia and their unaffected relatives. Acta Psychiatr Scand *123*, 98-106.

Burdach, J., O'Connell, M.R., Mackay, J.P., and Crossley, M. (2012). Two-timing zinc finger transcription factors liaising with RNA. Trends Biochem Sci *37*, 199-205.

Byne, W., Kidkardnee, S., Tatusov, A., Yiannoulos, G., Buchsbaum, M.S., and Haroutunian, V. (2006). Schizophrenia-associated reduction of neuronal and oligodendrocyte numbers in the anterior principal thalamic nucleus. Schizophr Res *85*, 245-253.

Caceda, R., Kinkead, B., and Nemeroff, C.B. (2007). Involvement of neuropeptide systems in schizophrenia: human studies. Int Rev Neurobiol *78*, 327-376.

Camargo, L.M., Collura, V., Rain, J.C., Mizuguchi, K., Hermjakob, H., Kerrien, S., Bonnert, T.P., Whiting, P.J., and Brandon, N.J. (2007). Disrupted in Schizophrenia 1 Interactome: evidence for the close connectivity of risk genes and a potential synaptic basis for schizophrenia. Mol Psychiatr *12*, 74-86.

Cardno, A.G., and Gottesman, 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.

Carlsson, A., Lindqvist, M., and Magnusson, T. (1957). 3,4-Dihydroxyphenylalanine and 5-hydroxytryptophan as reserpine antagonists. Nature *180*, 1200.

Carlsson, A., Waters, N., Holm-Waters, S., Tedroff, J., Nilsson, M., and Carlsson, M.L. (2001). Interactions between monoamines, glutamate, and GABA in schizophrenia: new evidence. Annu Rev Pharmacol Toxicol *41*, 237-260.

Carroll, L.S., and Owen, M.J. (2009). Genetic overlap between autism, schizophrenia and bipolar disorder. Genome Med *1*, 102.

Chang, C.K., Hayes, R.D., Perera, G., Broadbent, M.T., Fernandes, A.C., Lee, W.E., Hotopf, M., and Stewart, R. (2011). Life expectancy at birth for people with serious mental illness and other major disorders from a secondary mental health care case register in London. PLoS One *6*, e19590.

Chen, M., and Manley, J.L. (2009). Mechanisms of alternative splicing regulation: insights from molecular and genomics approaches. Nat Rev Mol Cell Biol *10*, 741-754.

Chen, M., Xu, Z., Zhai, J., Bao, X., Zhang, Q., Gu, H., Shen, Q., Cheng, L., Chen, X., Wang, K., *et al.* (2012). Evidence of IQ-modulated association between ZNF804A gene polymorphism and cognitive function in schizophrenia patients. Neuropsychopharmacol *37*, 1572-1578.

Chesini, I.M., Debyser, G., Croes, H., Ten Dam, G.B., Devreese, B., Stoker, A.W., and Hendriks, W.J. (2011). PTPBR7 binding proteins in myelinating neurons of the mouse brain. Int J Biol Sci *7*, 978-991.

Cheung, H.C., Hai, T., Zhu, W., Baggerly, K.a., Tsavachidis, S., Krahe, R., and Cote, G.J. (2009). Splicing factors PTBP1 and PTBP2 promote proliferation and migration of glioma cell lines. Brain *132*, 2277-2288.

Chirivi, R.G., Dilaver, G., van de Vorstenbosch, R., Wanschers, B., Schepens, J., Croes, H., Fransen, J., and Hendriks, W. (2004). Characterization of multiple transcripts and isoforms derived from the mouse protein tyrosine phosphatase gene Ptprr. Genes Cells *9*, 919-933.

Chua, S.E., and Murray, R.M. (1996). The neurodevelopmental theory of schizophrenia: evidence concerning structure and neuropsychology. Ann Med *28*, 547-555.

Chuaqui, R.F., Bonner, R.F., Best, C.J., Gillespie, J.W., Flaig, M.J., Hewitt, S.M., Phillips, J.L., Krizman, D.B., Tangrea, M.A., Ahram, M., *et al.* (2002). Post-analysis follow-up and validation of microarray experiments. Nat Genet *32 Suppl*, 509-514.

Cohen, O.S. (2012). Transcriptomic analysis of *postmortem* brain identifies dysregulated splicing events in novel candidate genes for schizophrenia. Schizophr Res.

Collins, F.S., Guyer, M.S., and Charkravarti, A. (1997). Variations on a theme: cataloging human DNA sequence variation. Science *278*, 1580-1581.

Conley, R.R., and Buchanan, R.W. (1997). Evaluation of treatment-resistant schizophrenia. Schizophrenia Bull *23*, 663-674.

Cooper, T.A. (2005). Use of minigene systems to dissect alternative splicing elements. Methods *37*, 331-340.

Corfas, G., Rosen, K.M., Aratake, H., Krauss, R., and Fischbach, G.D. (1995). Differential expression of ARIA isoforms in the rat brain. Neuron *14*, 103-115.

Corvin, A., Craddock, N., and Sullivan, P.F. (2010). Genome-wide association studies: a primer. Psychol Med *40*, 1063-1077.

Costa, E., J. Davis, D. R. Grayson, A. Guidotti, G. D. Pappas and C. Pesold (2001). Dendritic spine hypoplasticity and downregulation of reelin and GABAergic tone in schizophrenia vulnerability. Neurobiol Dis *8*, 723-742.

Costa, E., Chen, Y., Davis, J., Dong, E., Noh, J.S., Tremolizzo, L., Veldic, M., Grayson, D.R., and Guidotti, A. (2002). REELIN and schizophrenia: a disease at the interface of the genome and the epigenome. Mol Interv *2*, 47-57.

Cousijn, H., Rijpkema, M., Harteveld, A., Harrison, P.J., Fernandez, G., Franke, B., and Arias-Vasquez, A. (2012). Schizophrenia risk gene ZNF804A does not influence macroscopic brain structure: an MRI study in 892 volunteers. Mol Psychiatr.

David, A.S., Malmberg, A., Brandt, L., Allebeck, P., and Lewis, G. (1997). IQ and risk for schizophrenia: a population-based cohort study. Psychol Med *27*, 1311-1323.

Desterro, J.M., Rodriguez, M.S., and Hay, R.T. (2000). Regulation of transcription factors by protein degradation. Cell Mol Life Sci *57*, 1207-1219.

Desterro, J.M., Thomson, J., and Hay, R.T. (1997). Ubch9 conjugates SUMO but not ubiquitin. FEBS Lett *417*, 297-300.

Di Modugno, F., DeMonte, L., Balsamo, M., Bronzi, G., Nicotra, M.R., Alessio, M., Jager, E., Condeelis, J.S., Santoni, A., Natali, P.G., *et al.* (2007). Molecular cloning of hMena (ENAH) and its splice variant hMena+11a: epidermal growth factor increases their expression and stimulates hMena+11a phosphorylation in breast cancer cell lines. Cancer Res *67*, 2657-2665.

Ding, L., and Hegde, A.N. (2009). Expression of RGS4 Splice Variants in Dorsolateral Prefrontal Cortex of Schizophrenic and Bipolar Disorder Patients. Biol Psychiatry *65*, 541-545.

Djurovic, S., Gustafsson, O., Mattingsdal, M., Athanasiu, L., Bjella, T., Tesli, M., Agartz, I., Lorentzen, S., Melle, I., Morken, G., *et al.* (2010). A genome-wide association study of bipolar disorder in Norwegian individuals, followed by replication in Icelandic sample. J Affect Disord *126*, 312-316.

Donohoe, G., Rose, E., Frodl, T., Morris, D., Spoletini, I., Adriano, F., Bernardini, S., Caltagirone, C., Bossu, P., Gill, M., *et al.* (2011). ZNF804A risk allele is associated with relatively intact gray matter volume in patients with schizophrenia. Neuroimage *54*, 2132-2137.

Dudbridge, F., and Gusnanto, A. (2008). Estimation of significance thresholds for genomewide association scans. Genet Epidemiol *32*, 227-234.

Dwyer, S., Williams, H., Holmans, P., Moskvina, V., Craddock, N., Owen, M.J., and O'Donovan, M.C. (2010). No evidence that rare coding variants in ZNF804A confer risk of schizophrenia. Am J Med Genet B *153B*, 1411-1416.

Eastwood, S.L., and Harrison, P.J. (2010). Markers of Glutamate Synaptic Transmission and Plasticity Are Increased in the Anterior Cingulate Cortex in Bipolar Disorder. Biol Psychiatry *67*, 1010-1016.

Eaton, K., Sallee, F.R., and Sah, R. (2007). Relevance of neuropeptide Y (NPY) in psychiatry. Curr Top Med Chem *7*, 1645-1659.

Edmunds, J.W.M., L. C.; Clayton, A. (2008). Dynamic histone H3 methylation during gene induction: HYPB/Setd2 mediates all H3K36 trimethylation. Embo J *27*, 406 - 420.

Eisenhart, C. (1947). The assumptions underlying the analysis of variance. Biometrics *3*, 1-21.

Engle, E.C. (2010). Human Genetic Disorders of Axon Guidance. Cold Spring Harbor Perspectives in Biology *2*.

Elbashir, S.M., Harborth, J., Lendeckel, W., Yalcin, A., Weber, K., and Tuschl, T. (2001). Duplexes of 21-nucleotide RNAs mediate RNA interference in cultured mammalian cells. Nature *411*, 494-498.

Eranti, S.V., Maccabe, J.H., Bundy, H., and Murray, R.M. (2012). Gender difference in age at onset of schizophrenia: a meta-analysis. Psychol Med, 1-13.

Erhart, S.M., Marder, S.R., and Carpenter, W.T. (2006). Treatment of schizophrenia negative symptoms: future prospects. Schizophr Bull *32*, 234-237.

Erlenmeyer-Kimling, L., Rock, D., Roberts, S.A., Janal, M., Kestenbaum, C., Cornblatt, B., Adamo, U.H., and Gottesman, II (2000). Attention, memory, and motor skills as childhood predictors of schizophrenia-related psychoses: the New York High-Risk Project. Am J Psych *157*, 1416-1422.

Esslinger, C., Kirsch, P., Haddad, L., Mier, D., Sauer, C., Erk, S., Schnell, K., Arnold, C., Witt, S.H., Rietschel, M., *et al.* (2011). Cognitive state and connectivity effects of the genome-wide significant psychosis variant in ZNF804A. Neuroimage *54*, 2514-2523.

Esslinger, C.W., H.; Kirsch, P.; Erk, S.; Schnell, K.; Arnold, C.; Haddad, L.; Mier, D.; von Boberfeld, C. O.; Raab, K.; Witt, S. H.; Rietschel, M.; Cichon, S.; Meyer-Lindenburg, A.; (2009). Neural mechanisms of a genome-wide supported psychosis variant. Science *324*, 605.

Fakan, S. (1994). Perichromatin fibrils are in situ forms of nascent transcripts. Trends Cell Biol *4*, 86-90.

Fatemi, S. H., J. A. Earle and T. McMenomy (2000). Reduction in Reelin immunoreactivity in hippocampus of subjects with schizophrenia, bipolar disorder and major depression. Mol. Psychiatr *5*, 654-663.

Fatemi, S.H. (2005). Reelin glycoprotein: structure, biology and roles in health and disease. Mol Psychiatr *10*, 251-257.

Fatemi, S.H., and Folsom, T.D. (2009). The Neurodevelopmental Hypothesis of Schizophrenia, Revisited. Schizophrenia Bull *35*, 528-548.

Fedorova, O.A., Moiseeva, T.N., Nikiforov, A.A., Tsimokha, A.S., Livinskaya, V.A., Hodson, M., Bottrill, A., Evteeva, I.N., Ermolayeva, J.B., Kuznetzova, I.M., *et al.* (2011). Proteomic analysis of the 20S proteasome (PSMA3)-interacting proteins reveals a functional link between the proteasome and mRNA metabolism. Biochem Biophys Res Commun *416*, 258-265.

Feinberg, I. (1982). Schizophrenia caused by a fault in programmed synaptic elimination during adolescence. J Psych Res *17*, 319-334.

Fioravanti, M., Carlone, O., Vitale, B., Cinti, M.E., and Clare, L. (2005). A meta-analysis of cognitive deficits in adults with a diagnosis of schizophrenia. Neuropsychol Rev *15*, 73-95.

Fire, A., Xu, S., Montgomery, M.K., Kostas, S.A., Driver, S.E., and Mello, C.C. (1998). Potent and specific genetic interference by double-stranded RNA in Caenorhabditis elegans. Nature *391*, 806-811.

Forrest, M., Chapman, R.M., Doyle, A.M., Tinsley, C.L., Waite, A., and Blake, D.J. (2012). Functional analysis of TCF4 missense mutations that cause Pitt-Hopkins syndrome. Hum Mutat.

Frankle, W.G., and Laruelle, M. (2002). Neuroreceptor imaging in psychiatric disorders. Ann Nucl Med *16*, 437-446.

Frazer, K.A., Ballinger, D.G., Cox, D.R., Hinds, D.A., Stuve, L.L., Gibbs, R.A., Belmont, J.W., Boudreau, A., Hardenbol, P., Leal, S.M., *et al.* (2007). A second generation human haplotype map of over 3.1 million SNPs. Nature *449*, 851-861.

Friesen, W.J., and Darby, M.K. (2001). Specific RNA binding by a single C2H2 zinc finger. J Biol Chem *276*, 1968-1973.

Fujita, P.A., Rhead, B., Zweig, A.S., Hinrichs, A.S., Karolchik, D., Cline, M.S., Goldman, M., Barber, G.P., Clawson, H., Coelho, A., *et al.* (2011). The UCSC Genome Browser database: update 2011. Nucleic Acids Res *39*, D876-882.

Gasteiger, E., Gattiker, A., Hoogland, C., Ivanyi, I., Appel, R.D., and Bairoch, A. (2003). ExPASy: The proteomics server for in-depth protein knowledge and analysis. Nucleic Acids Res *31*, 3784-3788.

Gehman, L.T., Stoilov, P., Maguire, J., Damianov, A., Lin, C.H., Shiue, L., Ares, M., Jr., Mody, I., and Black, D.L. (2011). The splicing regulator Rbfox1 (A2BP1) controls neuronal excitation in the mammalian brain. Nat Genet *43*, 706-711.

Gibbs, R.A., Belmont, J.W., Hardenbol, P., Willis, T.D., Yu, F.L., Yang, H.M., Ch'ang, L.Y., Huang, W., Liu, B., Shen, Y., *et al.* (2003). The International HapMap Project. Nature *426*, 789-796.

Gietz, R.D., and Woods, R.A. (2002). Transformation of yeast by lithium acetate/single-stranded carrier DNA/polyethylene glycol method. Methods Enzymol *350*, 87-96.

Gillett, A., Maratou, K., Fewings, C., Harris, R.A., Jagodic, M., Aitman, T., and Olsson, T. (2009). Alternative splicing and transcriptome profiling of experimental autoimmune encephalomyelitis using genome-wide exon arrays. PLoS One *4*, e7773.

Girard, S.L., Gauthier, J., Noreau, A., Xiong, L., Zhou, S., Jouan, L., Dionne-Laporte, A., Spiegelman, D., Henrion, E., Diallo, O., *et al.* (2011). Increased exonic de novo mutation rate in individuals with schizophrenia. Nature Genet *43*, 860-863.

Girgenti, M.J., Loturco, J.J., and Maher, B.J. (2012). ZNF804a regulates expression of the Schizophrenia associated genes PRSS16, CDMT, PDE4B, and DRD2. PLoS One *7*, 1 - 5.

Glantz, L.A., and Lewis, D.A. (2000). Decreased dendritic spine density on prefrontal cortical pyramidal neurons in schizophrenia. Arch Gen Psychiatry *57*, 65-73.

Glessner, J.T., Reilly, M.P., Kim, C.E., Takahashi, N., Albano, A., Hou, C., Bradfield, J.P., Zhang, H., Sleiman, P.M., Flory, J.H., *et al.* (2010). Strong synaptic transmission impact by copy number variations in schizophrenia. Proc Natl Acad Sci U S A *107*, 10584-10589.

Gossen, M., and Bujard, H. (1992). Tight control of gene expression in mammalian cells by tetracycline-responsive promoters. Proc Natl Acad Sci U S A *89*, 5547-5551.

Gottesman, II, Laursen, T.M., Bertelsen, A., and Mortensen, P.B. (2010). Severe mental disorders in offspring with 2 psychiatrically ill parents. Arch Gen Psychiatry *67*, 252-257.

Griswold, A.J., Ma, D., Cukier, H.N., Nations, L.D., Schmidt, M.A., Chung, R.H., Jaworski, J.M., Salyakina, D., Konidari, I., Whitehead, P.L., *et al.* (2012). Evaluation of copy number variations reveals novel candidate genes in autism spectrum disorder-associated pathways. Hum Mol Genet *21*, 3513-3523.

Grozeva, D., Conrad, D.F., Barnes, C.P., Hurles, M., Owen, M.J., O'Donovan, M.C., Craddock, N., and Kirov, G. (2012). Independent estimation of the frequency of rare CNVs in the UK population confirms their role in schizophrenia. Schizophr Res *135*, 1-7.

Guidotti, A., Auta, J., Davis, J.M., Di-Giorgi-Gerevini, V., Dwivedi, Y., Grayson, D.R., Impagnatiello, F., Pandey, G., Pesold, C., Sharma, R., *et al.* (2000). Decrease in reelin and glutamic acid decarboxylase67 (GAD67) expression in schizophrenia and bipolar disorder: a postmortem brain study. Arch Gen Psychiatry *57*, 1061-1069.

Guo, L., Lobenhofer, E.K., Wang, C., Shippy, R., Harris, S.C., Zhang, L., Mei, N., Chen, T., Herman, D., Goodsaid, F.M., *et al.* (2006). Rat toxicogenomic study reveals analytical consistency across microarray platforms. Nature Biotechnol *24*, 1162-1169.

Guo, Z., Kanjanapangka, J., Liu, N., Liu, S., Liu, C., Wu, Z., Wang, Y., Loh, T., Kowolik, C., Jamsen, J., *et al.* (2012). Sequential Posttranslational Modifications Program FEN1 Degradation during Cell-Cycle Progression. Mol Cell *47*, 444-456.

Hakak, Y., Walker, J.R., Li, C., Wong, W.H., Davis, K.L., Buxbaum, J.D., Haroutunian, V., and Fienberg, A.A. (2001). Genome-wide expression analysis reveals dysregulation of myelination-related genes in chronic schizophrenia. Proc Natl Acad Sci U S A *98*, 4746-4751.

Han, J., Xiong, J., Wang, D., and Fu, X.D. (2011). Pre-mRNA splicing: where and when in the nucleus. Trends Cell Biol *21*, 336-343.

Hargreaves, A., Morris, D.W., Rose, E., Fahey, C., Moore, S., Cummings, E., Tropea, D., Gill, M., Corvin, A., and Donohoe, G. (2012). ZNF804A and social cognition in patients with schizophrenia and healthy controls. Mol Psychiatr *17*, 118-119.

Harrison, P.J. (1997). Schizophrenia: a disorder of neurodevelopment? Curr Opin Neurobiol *7*, 285-289.

Harrison, P.J. (1999). The neuropathology of schizophrenia. A critical review of the data and their interpretation. Brain *122 (Pt 4)*, 593-624.

Harrison, P.J., and Weinberger, D.R. (2005). Schizophrenia genes, gene expression, and neuropathology: on the matter of their convergence. Mol Psychiatr *10*, 40-68; image 45.

Hastings, P.J., Lupski, J.R., Rosenberg, S.M., and Ira, G. (2009). Mechanisms of change in gene copy number. Nat Rev Genet *10*, 551-564.

Hayashi-Takagi, A., and Sawa, A. (2010). Disturbed synaptic connectivity in schizophrenia: convergence of genetic risk factors during neurodevelopment. Brain Res Bull *83*, 140-146.

Hill, M.J., and Bray, N.J. (2011). Allelic differences in nuclear protein binding at a genome-wide significant risk variant for schizophrenia in ZNF804A. Mol Psychiatr *16*, 787-789.

Hill, M.J., Jeffries, A.R., Dobson, R.J., Price, J., and Bray, N.J. (2012a). Knockdown of the psychosis susceptibility gene ZNF804A alters expression of genes involved in cell adhesion. Hum Mol Genet *21*, 1018-1024.

Hill, M.J., and Bray, N.J. (2012b). Evidence That Schizophrenia Risk Variation in the ZNF804A Gene Exerts Its Effects During Fetal Brain Development. Am J Psychiatry *169*, 1301-1308.

Hirschhorn, J.N., and Daly, M.J. (2005). Genome-wide association studies for common diseases and complex traits. Nat Rev Genet *6*, 95-108.

Howes, O.D., and Kapur, S. (2009). The dopamine hypothesis of schizophrenia: version III--the final common pathway. Schizophr Bull *35*, 549-562.

Huang, H., and Bader, J.S. (2009). Precision and recall estimates for two-hybrid screens. Bioinformatics *25*, 372-378.

Hung, L.H., Heiner, M., Hui, J., Schreiner, S., Benes, V., and Bindereif, A. (2008a). Diverse roles of hnRNP L in mammalian mRNA processing: a combined microarray and RNAi analysis. RNA *14*, 284-296.

Ikeda, K., Iritani, S., Ueno, H., and Niizato, K. (2004). Distribution of neuropeptide Y interneurons in the dorsal prefrontal cortex of schizophrenia. Prog Neuropsychopharmacol Biol Psychiatry *28*, 379-383.

Ikeda, M., Aleksic, B., Kinoshita, Y., Okochi, T., Kawashima, K., Kushima, I., Ito, Y., Nakamura, Y., Kishi, T., Okumura, T., *et al.* (2011). Genome-wide association study of schizophrenia in a Japanese population. Biol Psychiatry *69*, 472-478.

Impagnatiello, F., Guidotti, A.R., Pesold, C., Dwivedi, Y., Caruncho, H., Pisu, M.G., Uzunov, D.P., Smalheiser, N.R., Davis, J.M., Pandey, G.N., *et al.* (1998). A decrease of reelin expression as a putative vulnerability factor in schizophrenia. Proc Natl Acad Sci U S A *95*, 15718-15723.

Ingason, A., Sigmundsson, T., Steinberg, S., Sigurdsson, E., Haraldsson, M., Magnusdottir, B.B., Frigge, M.L., Kong, A., Gulcher, J., Thorsteinsdottir, U., *et al.* (2007). Support for involvement of the AHI1 locus in schizophrenia. European Journal of Human Genetics *15*, 988-991.

Ingason, A., Rujescu, D., Cichon, S., Sigurdsson, E., Sigmundsson, T., Pietilainen, O.P., Buizer-Voskamp, J.E., Strengman, E., Francks, C., Muglia, P., *et al.* (2011). Copy number variations of chromosome 16p13.1 region associated with schizophrenia. Mol Psychiatr *16*, 17-25.

Insel, T.R., and Scolnick, E.M. (2006). Cure therapeutics and strategic prevention: raising the bar for mental health research. Mol Psychiatr *11*, 11-17.

Irizarry, R.A., Bolstad, B.M., Collin, F., Cope, L.M., Hobbs, B., and Speed, T.P. (2003a). Summaries of Affymetrix GeneChip probe level data. Nucleic Acids Res *31*, e15.

Irizarry, R.A., Hobbs, B., Collin, F., Beazer-Barclay, Y.D., Antonellis, K.J., Scherf, U., and Speed, T.P. (2003b). Exploration, normalization, and summaries of high density oligonucleotide array probe level data. Biostatistics *4*, 249-264.

Iuchi, S. (2001). Three classes of C2H2 zinc finger proteins. Cell Mol Life Sci *58*, 625-635.

Javitt, D.C., and Zukin, S.R. (1991). Recent advances in the phencyclidine model of schizophrenia. Am J Psychiatry *148*, 1301-1308.

Jensen, K.B., Dredge, B.K., Stefani, G., Zhong, R., Buckanovich, R.J., Okano, H.J., Yang, Y.Y., and Darnell, R.B. (2000). Nova-1 regulates neuron-specific alternative splicing and is essential for neuronal viability. Neuron *25*, 359-371.

Jentsch, J.D., Trantham-Davidson, H., Jairl, C., Tinsley, M., Cannon, T.D., and Lavin, A. (2009). Dysbindin Modulates Prefrontal Cortical Glutamatergic Circuits and Working Memory Function in Mice. Neuropsychopharmacol *34*, 2601-2608.

Jin, Y., Suzuki, H., Maegawa, S., Endo, H., Sugano, S., Hashimoto, K., Yasuda, K., and Inoue, K. (2003). A vertebrate RNA-binding protein Fox-1 regulates tissue-specific splicing via the pentanucleotide GCAUG. Embo J *22*, 905-912.

Johnson, E.S., and Blobel, G. (1997). Ubc9p is the conjugating enzyme for the ubiquitin-like protein Smt3p. J Biol Chem *272*, 26799-26802.

Kahler, A.K., Djurovic, S., Kulle, B., Jonsson, E.G., Agartz, I., Hall, H., Opjordsmoen, S., Jakobsen, K.D., Hansen, T., Melle, I., *et al.* (2008). Association analysis of schizophrenia on 18 genes involved in neuronal migration: MDGA1 as a new susceptibility gene. Am J Med Genet B Neuropsychiatr Genet *147B*, 1089-1100.

Kahn, R.S., Fleischhacker, W.W., Boter, H., Davidson, M., Vergouwe, Y., Keet, I.P.M., Gheorghe, M.D., Rybakowski, J.K., Galderisi, S., Libiger, J., *et al.* (2008). Effectiveness of antipsychotic drugs in first-episode schizophrenia and schizophreniform disorder: an open randomised clinical trial. Lancet *371*, 1085-1097.

Kalkman, H.O. (2006). The role of the phosphatidylinositide 3-kinase-protein kinase B pathway in schizophrenia. Pharmacol Ther *110*, 117-134.

Kalus, P., Muller, T.J., Zuschratter, W., and Senitz, D. (2000). The dendritic architecture of prefrontal pyramidal neurons in schizophrenic patients. Neuroreport *11*, 3621-3625.

Kapur, K., Xing, Y., Ouyang, Z., and Wong, W.H. (2007). Exon arrays provide accurate assessments of gene expression. Genome Biol *8*, R82.

Kaul, S., Blackford, J.A., Jr., Chen, J., Ogryzko, V.V., and Simons, S.S., Jr. (2000). Properties of the glucocorticoid modulatory element binding proteins GMEB-1 and -2: potential new modifiers of glucocorticoid receptor transactivation and members of the family of KDWK proteins. Mol Endocrinol *14*, 1010-1027.

Keefe, R.S., Bilder, R.M., Harvey, P.D., Davis, S.M., Palmer, B.W., Gold, J.M., Meltzer, H.Y., Green, M.F., Miller, D.D., Canive, J.M., *et al.* (2006). Baseline neurocognitive deficits in the CATIE schizophrenia trial. Neuropsychopharmacol *31*, 2033-2046.

Kendler, K.S., and Diehl, S.R. (1993). The genetics of schizophrenia: a current, genetic-epidemiologic perspective. Schizophr Bull *19*, 261-285.

Kendler, K.S., MacLean, C.J., O'Neill, F.A., Burke, J., Murphy, B., Duke, F., Shinkwin, R., Easter, S.M., Webb, B.T., Zhang, J., *et al.* (1996). Evidence for a schizophrenia vulnerability locus on chromosome 8p in the Irish Study of High-Density Schizophrenia Families. Am J Psychiatry *153*, 1534-1540.

Kent, W.J. (2002). BLAT--the BLAST-like alignment tool. Genome Res *12*, 656-664.

Kent, W.J., Sugnet, C.W., Furey, T.S., Roskin, K.M., Pringle, T.H., Zahler, A.M., and Haussler, D. (2002). The human genome browser at UCSC. Genome Res *12*, 996-1006.

Kerns, D., Vong, G.S., Barley, K., Dracheva, S., Katsel, P., Casaccia, P., Haroutunian, V., and Byne, W. (2010). Gene expression abnormalities and oligodendrocyte deficits in the internal capsule in schizophrenia. Schizophr Res *120*, 150-158.

Keshavan, M.S., and Hogarty, G.E. (1999). Brain maturational processes and delayed onset in schizophrenia. Dev Psychopathol *11*, 525-543.

Kety, S.S. (1987). The significance of genetic factors in the etiology of schizophrenia: results from the national study of adoptees in Denmark. J Psychiatr Res *21*, 423-429.

Kim, H.G., Kishikawa, S., Higgins, A.W., Seong, I.S., Donovan, D.J., Shen, Y., Lally, E., Weiss, L.A., Najm, J., Kutsche, K., *et al.* (2008). Disruption of neurexin 1 associated with autism spectrum disorder. Am J Hum Genet *82*, 199-207.

Kirkbride, J.B., Errazuriz, A., Croudace, T.J., Morgan, C., Jackson, D., Boydell, J., Murray, R.M., and Jones, P.B. (2012). Incidence of schizophrenia and other psychoses in England, 1950-2009: a systematic review and meta-analyses. PLoS One *7*, e31660.

Kirkpatrick, B., Buchanan, R.W., Ross, D.E., and Carpenter, W.T. (2001). A separate disease within the syndrome of schizophrenia. Arch Gen Psychiatry *58*, 165-171.

Kirov, G., Grozeva, D., Norton, N., Ivanov, D., Mantripragada, K.K., Holmans, P., Craddock, N., Owen, M.J., and O'Donovan, M.C. (2009a). Support for the involvement of large copy number variants in the pathogenesis of schizophrenia. Hum Mol Genet *18*, 1497-1503.

Kirov, G., Gumus, D., Chen, W., Norton, N., Georgieva, L., Sari, M., O'Donovan, M.C., Erdogan, F., Owen, M.J., Ropers, H.H., *et al.* (2008). Comparative genome hybridization suggests a role for NRXN1 and APBA2 in schizophrenia. Hum Mol Genet *17*, 458-465.

Kirov, G., Pocklington, A.J., Holmans, P., Ivanov, D., Ikeda, M., Ruderfer, D., Moran, J., Chambert, K., Toncheva, D., Georgieva, L., *et al.* (2012). De novo CNV analysis implicates specific abnormalities of postsynaptic signalling complexes in the pathogenesis of schizophrenia. Mol Psychiatr *17*, 142-153.

Kirov, G., Rujescu, D., Ingason, A., Collier, D.A., O'Donovan, M.C., and Owen, M.J. (2009b). Neurexin 1 (NRXN1) deletions in schizophrenia. Schizophr Bull *35*, 851-854.

Kirov, G., Zaharieva, I., Georgieva, L., Moskvina, V., Nikolov, I., Cichon, S., Hillmer, A., Toncheva, D., Owen, M.J., and O'Donovan, M.C. (2009c). A genome-wide association study in 574 schizophrenia trios using DNA pooling. Mol Psychiatr *14*, 796-803.

Kraepelin, E. (1896). Dementia praecox and paraphrenia (New York: Robert E. Kreiger).

Kucukdereli, H., Allen, N.J., Lee, A.T., Feng, A., Ozlu, M.I., Conatser, L.M., Chakraborty, C., Workman, G., Weaver, M., Sage, E.H., *et al.* (2011). Control of excitatory CNS synaptogenesis by astrocyte-secreted proteins Hevin and SPARC. Proc Natl Acad Sci USA *108*, E440-449.

Kuromitsu, J., Yokoi, A., Kawai, T., Nagasu, T., Aizawa, T., Haga, S., and Ikeda, K. (2001). Reduced neuropeptide Y mRNA levels in the frontal cortex of people with schizophrenia and bipolar disorder. Brain Res Gene Expr Patterns *1*, 17-21.

Laajala, E., Aittokallio, T., Lahesmaa, R., and Elo, L.L. (2009). Probe-level estimation improves the detection of differential splicing in Affymetrix exon array studies. Genome Biol *10*, R77.

Ladd, A.N., Charlet-B, N., and Cooper, T.A. (2001). The CELF family of RNA binding proteins is implicated in cell-specific and developmentally regulated alternative splicing. Mol Cell Biol *21*, 1285-1296.

Lander, E.S. (1996). The new genomics: global views of biology. Science *274*, 536-539.

Langer, W., Sohler, F., Leder, G., Beckmann, G., Seidel, H., Grone, J., Hummel, M., and Sommer, A. (2010). Exon array analysis using re-defined probe sets results in reliable identification of alternatively spliced genes in non-small cell lung cancer. BMC Genomics *11*, 676.

Larkin, M.A., Blackshields, G., Brown, N.P., Chenna, R., McGettigan, P.A., McWilliam, H., Valentin, F., Wallace, I.M., Wilm, A., Lopez, R., *et al.* (2007). Clustal W and Clustal X version 2.0. Bioinformatics *23*, 2947-2948.

Lawrie, S.M., Byrne, M., Miller, P., Hodges, A., Clafferty, R.A., Cunningham Owens, D.G., and Johnstone, E.C. (2001). Neurodevelopmental indices and the development of psychotic symptoms in subjects at high risk of schizophrenia. Br J Psychiatry *178*, 524-530.

Lebrand, C., Dent, E.W., Strasser, G.A., Lanier, L.M., Krause, M., Svitkina, T.M., Borisy, G.G., and Gertler, F.B. (2004). Critical role of Ena/VASP proteins for filopodia formation in neurons and in function downstream of netrin-1. Neuron *42*, 37-49.

Lencz, T., Morgan, T.V., Athanasiou, M., Dain, B., Reed, C.R., Kane, J.M., Kucherlapati, R., and Malhotra, A.K. (2007). Converging evidence for a pseudoautosomal cytokine receptor gene locus in schizophrenia. Mol Psychiatr *12*, 572-580.

Lencz, T., Szeszko, P.R., DeRosse, P., Burdick, K.E., Bromet, E.J., Bilder, R.M., and Malhotra, A.K. (2010). A schizophrenia risk gene, ZNF804A, influences neuroanatomical and neurocognitive phenotypes. Neuropsychopharmacol *35*, 2284-2291.

Leucht, S., Pitschel-Walz, G., Abraham, D., and Kissling, W. (1999). Efficacy and extrapyramidal side-effects of the new antipsychotics olanzapine, quetiapine, risperidone, and sertindole compared to conventional antipsychotics and placebo. A meta-analysis of randomized controlled trials. Schizophr Res *35*, 51-68.

Levinson, D.F., Duan, J., Oh, S., Wang, K., Sanders, A.R., Shi, J., Zhang, N., Mowry, B.J., Olincy, A., Amin, F., *et al.* (2011). Copy number variants in schizophrenia: confirmation of five previous findings and new evidence for 3q29 microdeletions and VIPR2 duplications. Am J Psychiatry *168*, 302-316.

Levinson, D.F., Shi, J., Wang, K., Oh, S., Riley, B., Pulver, A.E., Wildenauer, D.B., Laurent, C., Mowry, B.J., Gejman, P.V., *et al.* (2012). Genome-Wide Association Study of Multiplex Schizophrenia Pedigrees. Am J Psychiatry *169*, 963-973.

Lewis, D.A., Glantz, L.A., Pierri, J.N., and Sweet, R.A. (2003). Altered cortical glutamate neurotransmission in schizophrenia: evidence from morphological studies of pyramidal neurons. Ann N Y Acad Sci *1003*, 102-112.

Lewis, D.A., and Levitt, P. (2002). Schizophrenia as a disorder of neurodevelopment. Annu Rev Neurosci *25*, 409-432.

Lewis, S.W., Barnes, T.R.E., Davies, L., Murray, R.M., Dunn, G., Hayhurst, K.P., Markwick, A., Lloyd, H., and Jones, P.B. (2006). Randomized controlled trial of effect of prescription of clozapine versus other second-generation antipsychotic drugs in resistant schizophrenia. Schizophrenia Bull *32*, 715-723.

Li, L., He, S., Sun, J.M., and Davie, J.R. (2004). Gene regulation by Sp1 and Sp3. Biochem Cell Biol *82*, 460-471.

Li, Z., Boules, M., Williams, K., Gordillo, A., Li, S., and Richelson, E. (2010). Similarities in the behavior and molecular deficits in the frontal cortex between the neurotensin receptor subtype 1 knockout mice and chronic phencyclidine-treated mice: relevance to schizophrenia. Neurobiol Dis *40*, 467-477.

Liao, H.M., Chao, Y.L., Huang, A.L., Cheng, M.C., Chen, Y.J., Lee, K.F., Fang, J.S., Hsu, C.H., and Chen, C.H. (2012). Identification and characterization of three inherited genomic copy number variations associated with familial schizophrenia. Schizophr Res *139*, 229-236.

Lieberman, J.A., Stroup, T.S., McEvoy, J.P., Swartz, M.S., Rosenheck, R.A., Perkins, D.O., Keefe, R.S.E., Davis, S.M., Davis, C.E., Lebowitz, B.D., *et al.* (2005). Effectiveness of antipsychotic drugs in patients with chronic schizophrenia. New Engl J Med *353*, 1209-1223.

Lindermayer, J.P., and Khan, A. (2006). The American Psychiatric Publishing Textbook of Schizophrenia. In, J.A. Lieberman, ed. (American Psychiatric Publishing, Inc.).

Lindsay, E.A., Morris, M.A., Gos, A., Nestadt, G., Wolyniec, P.S., Lasseter, V.K., Shprintzen, R., Antonarakis, S.E., Baldini, A., and Pulver, A.E. (1995). Schizophrenia and chromosomal deletions within 22q11.2. Am J Hum Genet *56*, 1502-1503.

Lockstone, H.E. (2011). Exon array data analysis using Affymetrix power tools and R statistical software. Brief Bioinform *12*, 634-644.

Loyer, P., Trembley, J.H., Lahti, J.M., and Kidd, V.J. (1998). The RNP protein, RNPS1, associates with specific isoforms of the p34cdc2-related PITSLRE protein kinase in vivo. J Cell Sci *111 (Pt 11)*, 1495-1506.

MacDonald, A.W., and Schulz, S.C. (2009). What we know: findings that every theory of schizophrenia should explain. Schizophr Bull *35*, 493-508.

Magri, C., Sacchetti, E., Traversa, M., Valsecchi, P., Gardella, R., Bonvicini, C., Minelli, A., Gennarelli, M., and Barlati, S. (2010). New copy number variations in Schizophrenia. PLoS One *5*, 1 - 6.

Mah, S., Nelson, M.R., Delisi, L.E., Reneland, R.H., Markward, N., James, M.R., Nyholt, D.R., Hayward, N., Handoko, H., Mowry, B., *et al.* (2006). Identification of the semaphorin receptor PLXNA2 as a candidate for susceptibility to schizophrenia. Mol Psychiatr *11*, 471-478.

Mangalore, R., and Knapp, M. (2007). Cost of schizophrenia in England. J Ment Health Policy Econ *10*, 23-41.

Marion, M.J., and Marion, C. (1988). Ribosomal protein-s2, protein-s6, protein-s10, protein-s14, protein-s15 and protein-s25 are localized on the surface of mammalian 40-s subunits and stabilize their conformation - a study with immobilized trypsin. Febs Letters *232*, 281-285.

Massague, J. (2008). TGF beta in cancer. Cell *134*, 215-230.

Maston, G.A., Evans, S.K., and Green, M.R. (2006). Transcriptional regulatory elements in the human genome. Annu Rev Genomics Hum Genet *7*, 29-59.

Matthews, J.M., and Sunde, M. (2002). Zinc fingers - folds for many occasions. Life *54*, 351 - 355.

Mayeda, A.B., J.; Kobayashi, R.; Zhang, M. Q.; Gardiner, E. M.; Krainer, A. R. (1999). Purification and charasterization of human RNPS1: a general activator of pre-mRNA splicing. Embo J *18*, 4560- 4570.

McCarthy, D.J., and Smyth, G.K. (2009). Testing significance relative to a fold-change threshold is a TREAT. Bioinformatics *25*, 765-771.

McCarthy, S.E., Makarov, V., Kirov, G., Addington, A.M., McClellan, J., Yoon, S., Perkins, D.O., Dickel, D.E., Kusenda, M., Krastoshevsky, O., *et al.* (2009). Microduplications of 16p11.2 are associated with schizophrenia. Nat Genet *41*, 1223-1227.

McClellan, J., and King, M.C. (2010). Genomic analysis of mental illness: a changing landscape. Jama *303*, 2523-2524.

McGlashan, T.H., and Hoffman, R.E. (2000). Schizophrenia as a disorder of developmentally reduced synaptic connectivity. Arch Gen Psychiatry *57*, 637-648.

Melom, J.E., and Littleton, J.T. (2011). Synapse development in health and disease. Curr Opin Genet Dev *21*, 256-261.

Mendis, D.B., and Brown, I.R. (1994). Expression of the gene encoding the extracellular matrix glycoprotein SPARC in the developing and adult mouse brain. Brain Res Mol Brain Res *24*, 11-19.

Mendis, D.B., Malaval, L., and Brown, I.R. (1995). SPARC, an extracellular matrix glycoprotein containing the follistatin module, is expressed by astrocytes in synaptic enriched regions of the adult brain. Brain Res *676*, 69-79.

Miettinen, P.J., Ebner, R., Lopez, A.R., and Derynck, R. (1994).TGF-beta induced transdifferentiation of mammary epithelial-cells to mesenchymal cells - involvement of type-i receptors. J Cell Biol *127*, 2021-2036.

Misteli, T. (2000). Cell biology of transcription and pre-mRNA splicing: nuclear architecture meets nuclear function. J Cell Sci *113 (Pt 11)*, 1841-1849.

Moghaddam, B., Adams, B., Verma, A., and Daly, D. (1997). Activation of glutamatergic neurotransmission by ketamine: a novel step in the pathway from NMDA receptor blockade to dopaminergic and cognitive disruptions associated with the prefrontal cortex. J Neurosci *17*, 2921-2927.

Moreno-De-Luca, D., Mulle, J.G., Kaminsky, E.B., Sanders, S.J., Myers, S.M., Adam, M.P., Pakula, A.T., Eisenhauer, N.J., Uhas, K., Weik, L., *et al.* (2010). Deletion 17q12 Is a Recurrent Copy Number Variant that Confers High Risk of Autism and Schizophrenia. Am J Hum Genet *87*, 618-630.

Moskvina, V., O'Dushlaine, C., Purcell, S., Craddock, N., Holmans, P., and O'Donovan, M.C. (2011). Evaluation of an approximation method for assessment of overall significance of multiple-dependent tests in a genomewide association study. Genet Epidemiol *35*, 861-866.

Mossner, R., Schuhmacher, A., Wagner, M., Lennertz, L., Steinbrecher, A., Quednow, B.B., Rujescu, D., Rietschel, M., and Maier, W. (2012). The schizophrenia risk gene ZNF804A influences the antipsychotic response of positive schizophrenia symptoms. Eur Arch Psychiatry Clin Neurosci *262*, 193-197.

Mouri, A., Sasaki, A., Watanabe, K., Sogawa, C., Kitayama, S., Mamiya, T., Miyamoto, Y., Yamada, K., Noda, Y., and Nabeshima, T. (2012). MAGE-D1 regulates expression of depression-like behavior through serotonin transporter ubiquitylation. J Neurosci *32*, 4562-4580.

Mowry, B.J., and Gratten, J. (2013). The emerging spectrum of allelic variation in schizophrenia: current evidence and strategies for the identification and functional characterization of common and rare variants. Mol Psychiatr *18*, 38-52.

Myers, R.A., Casals, F., Gauthier, J., Hamdan, F.F., Keebler, J., Boyko, A.R., Bustamante, C.D., Piton, A.M., Spiegelman, D., Henrion, E., *et al.* (2011). A population genetic approach to mapping neurological disorder genes using deep resequencing. PLoS Genet *7*, e1001318.

Muratani, M., and Tansey, W.R. (2003). How the ubiquitin-proteasome system controls transcription. Nature Reviews Mol Cell Biol *4*, 192-201.

Need, A.C., Ge, D., Weale, M.E., Maia, J., Feng, S., Heinzen, E.L., Shianna, K.V., Yoon, W., Kasperaviciute, D., Gennarelli, M., *et al.* (2009). A genome-wide investigation of SNPs and CNVs in schizophrenia. PLoS Genet *5*, e1000373.

Need, A.C., McEvoy, J.P., Gennarelli, M., Heinzen, E.L., Ge, D., Maia, J.M., Shianna, K.V., He, M., Cirulli, E.T., Gumbs, C.E., *et al.* (2012). Exome sequencing followed by large-scale genotyping suggests a limited role for moderately rare risk factors of strong effect in schizophrenia. Am J Hum Genet *91*, 303-312.

NICE (2010). Schizophrenia: Core Interventions in the Treatment and Management of Schizophrenia in Adults in Primary and Secondary Care, Vol 82 (London: The British Psychological Society).

Nullmeier, S., Panther, P., Dobrowolny, H., Frotscher, M., Zhao, S., Schwegler, H., and Wolf, R. (2011). Region-specific alteration of GABAergic markers in the brain of heterozygous reeler mice. The European journal of neuroscience *33*, 689-698.

Numakawa, T., Yagasaki, Y., Ishimoto, T., Okada, T., Suzuki, T., Iwata, N., Ozaki, N., Taguchi, T., Tatsumi, M., Kamijima, K., *et al.* (2004). Evidence of novel neuronal functions of dysbindin, a susceptibility gene for schizophrenia. Hum Mol Genet *13*, 2699-2708.

O'Donovan, M.C., Craddock, N., Norton, N., Williams, H., Peirce, T., Moskvina, V., Nikolov, I., Hamshere, M., Carroll, L., Georgieva, L., *et al.* (2008). Identification of loci associated with schizophrenia by genome-wide association and follow-up. Nature Genet *40*, 1053-1055.

O'Dushlaine, C., Kenny, E., Heron, E., Donohoe, G., Gill, M., Morris, D., and Corvin, A. (2011). Molecular pathways involved in neuronal cell adhesion and membrane scaffolding contribute to schizophrenia and bipolar disorder susceptibility. Mol psychiatr *16*, 286-292.

Okada, T., R. Hashimoto, H. Yamamori, S. Umeda-Yano, Y. Yasuda, K. Ohi, M. Fukumoto, K. Ikemoto, Y. Kunii, H. Tomita, A. Ito and M. Takeda (2012). Expression analysis of a novel mRNA variant of the schizophrenia risk gene ZNF804A. Schizophr Res *141*, 277-278.

Okoniewski, M.J.M., C. J. (2008). Comprehensive analysis of affymetrix exon arrays using BioConductor. PLoS Comput Biology *4*, 1 - 6.

Olney, J.W., Newcomer, J.W., and Farber, N.B. (1999). NMDA receptor hypofunction model of schizophrenia. J Psychiatr Res *33*, 523-533.

Oni-Orisan, A., Kristiansen, L.V., Haroutunian, V., Ruff, J.H.M.-W., and McCullumsmith, R.E. (2008). Altered vesicular glutamate transporter expression in the anterior Cingulate cortex in schizophrenia. Biol Psychiatry *63*, 766-775.

Orr, H.T. (2010). Nuclear ataxias. Cold Spring Harb Perspect Biol *2*, a000786.

Otsuka, M., Fujita, M., Aoki, T., Ishii, S., Sugiura, Y., Yamamoto, T., and Inoue, J. (1995). Novel zinc chelators with dual activity in the inhibition of the kappa B site-binding proteins HIV-EP1 and NF-kappa B. J Med Chem *38*, 3264-3270.

Overton, S.L., and Medina, S.L. (2008). The stigma of mental illness. J Couns Dev *86*, 143-151.

Owen, M.J., O'Donovan, M.C., Thapar, A., and Craddock, N. (2011). Neurodevelopmental hypothesis of schizophrenia. Br J Psychiatry *198*, 173-175.

Owen, M.J., Williams, H.J., and O'Donovan, M.C. (2009). Schizophrenia genetics: advancing on two fronts. Curr Opin Genet Dev *19*, 266-270.

Ozsolak, F., and Milos, P.M. (2011). RNA sequencing: advances, challenges and opportunities. Nat Rev Genet *12*, 87-98.

Pahlman, S., Mamaeva, S., Meyerson, G., Mattsson, M.E., Bjelfman, C., Ortoft, E., and Hammerling, U. (1990). Human neuroblastoma cells in culture: a model for neuronal cell differentiation and function. Acta Physiol Scand Suppl *592*, 25-37.

Palmer, B.W., Heaton, R.K., Paulsen, J.S., Kuck, J., Braff, D., Harris, M.J., Zisook, S., and Jeste, D.V. (1997). Is it possible to be schizophrenic yet neuropsychologically normal? Neuropsychol *11*, 437-446.

Park, J.H., Wacholder, S., Gail, M.H., Peters, U., Jacobs, K.B., Chanock, S.J., and Chatterjee, N. (2010). Estimation of effect size distribution from genome-wide association studies and implications for future discoveries. Nat Genet *42*, 570-575.

Patterson, T.A., Lobenhofer, E.K., Fulmer-Smentek, S.B., Collins, P.J., Chu, T.-M., Bao, W., Fang, H., Kawasaki, E.S., Hager, J., Tikhonova, I.R., *et al.* (2006). Performance comparison of one-color and two-color platforms within the MicroArray Quality Control (MAQC) project. Nature Biotechnol *24*, 1140-1150.

Pawitan, Y., Michiels, S., Koscielny, S., Gusnanto, A., and Ploner, A. (2005). False discovery rate, sensitivity and sample size for microarray studies. Bioinformatics *21*, 3017-3024.

Perraud, M., Gioud, M., and Monier, J.C. (1979). [Intranuclear structures of monkey kidney cells recognised by immunofluorescence and immuno-electron microscopy using anti-ribonucleoprotein antibodies (author's transl)]. Ann Immunol *130C*, 635-647.

Pfaff, M., Sasaki, T., Tangemann, K., Chu, M.L., and Timpl, R. (1995). Integrin-binding and cell-adhesion studies of fibulins reveal a particular affinity for alpha IIb beta 3. Exp Cell Res *219*, 87-92.

Polyak, K., and Weinberg, R.A. (2009). Transitions between epithelial and mesenchymal states: acquisition of malignant and stem cell traits. Nat Rev Cancer *9*, 265-273.

Pulver, A.E., Lasseter, V.K., Kasch, L., Wolyniec, P., Nestadt, G., Blouin, J.L., Kimberland, M., Babb, R., Vourlis, S., and Chen, H. (1995). Schizophrenia: a genome scan targets chromosomes 3p and 8p as potential sites of susceptibility genes. Am J Med Genet *60*, 252-260.

Purcell, S.M., Wray, N.R., Stone, J.L., Visscher, P.M., O'Donovan, M.C., Sullivan, P.F., and Sklar, P. (2009). Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. Nature *460*, 748-752.

Rahimov, F., King, O.D., Leung, D.G., Bibat, G.M., Emerson, C.P., Jr., Kunkel, L.M., and Wagner, K.R. (2012). Transcriptional profiling in facioscapulohumeral muscular dystrophy to identify candidate biomarkers. Proc Natl Acad Sci USA *109*, 16234-16239.

Rasetti, R., Sambataro, F., Chen, Q., Callicott, J.H., Mattay, V.S., and Weinberger, D.R. (2011). Altered cortical network dynamics: a potential intermediate phenotype for schizophrenia and association with ZNF804A. Arch Gen Psychiatry *68*, 1207-1217.

Rasko, T., Haenisch, B., Szvetnik, A., Schaefer, M.H., Izsvak, Z., Cichon, S., Nothen, M.M., Andrade-Navarro, M.A., and Wanker, E.E. (2012). A first protein-protein interaction network for mood disorders and schizophrenia. Poster presented at: 20th World Congress of Psychiatric Genetics (Hamburg).

Reeves, P.J., Kim, J.M., and Khorana, H.G. (2002). Structure and function in rhodopsin: a tetracycline-inducible system in stable mammalian cell lines for high-level expression of opsin mutants. Proc Natl Acad Sci USA *99*, 13413-13418.

Reichenberg, A. (2010). The assessment of neuropsychological functioning in schizophrenia. Dialogues Clinical Neurosci *12*, 383-392.

Reichenberg, A., Weiser, M., Rapp, M.A., Rabinowitz, J., Caspi, A., Schmeidler, J., Knobler, H.Y., Lubin, G., Nahon, D., Harvey, P.D., *et al.* (2005). Elaboration on premorbid intellectual performance in schizophrenia: premorbid intellectual decline and risk for schizophrenia. Arch Gen Psychiatry *62*, 1297-1304.

Riley, B., Thiselton, D., Maher, B.S., Bigdeli, T., Wormley, B., McMichael, G.O., Fanous, A.H., Vladimirov, V., O'Neill, F.A., Walsh, D., *et al.* (2009). Replication of association between schizophrenia and ZNF804A in the Irish Case–Control Study of Schizophrenia sample. Mol Psychiatr *15*, 29-37.

Ripke, S., Sanders, A.R., Kendler, K.S., Levinson, D.F., Sklar, P., Holmans, P.A., Lin, D.Y., Duan, J., Ophoff, R.A., Andreassen, O.A., *et al.* (2011). Genome-wide association study identifies five new schizophrenia loci. Nat Genet *43*, 969-976.

Risch, N., and Merikangas, K. (1996). The future of genetic studies of complex human diseases. Science *273*, 1516-1517.

Riis, B., Rattan, S.I.S., Clark, B.F.C., and Merrick, W.C. (1990). Eukaryotic protein elongation-factors. Trends in Biochemical Sciences *15*, 420-424.

Rodriguez-Murillo, L., Gogos, J.A., and Karayiorgou, M. (2012). The genetic architecture of schizophrenia: new mutations and emerging paradigms. Annu Rev Med *63*, 63-80.

Ross, R.A., and Biedler, J.L. (1985). Presence and regulation of tyrosinase activity in human neuroblastoma cell variants in vitro. Cancer Res *45*, 1628-1632.

Rowley, N., Prip-Buus, C., Westermann, B., Brown, C., Schwarz, E., Barrell, B., and Neupert, W. (1994). Mdj1p, a novel chaperone of the DnaJ family, is involved in mitochondrial biogenesis and protein folding. Cell *77*, 249-259.

Saha, S., Chant, D., Welham, J., and McGrath, J. (2005). A systematic review of the prevalence of schizophrenia. PLoS Med *2*, e141.

Sahoo, T., Theisen, A., Rosenfeld, J.A., Lamb, A.N., Ravnan, J.B., Schultz, R.A., Torchia, B.S., Neill, N., Casci, I., Bejjani, B.A., *et al.* (2011). Copy number variants of schizophrenia susceptibility loci are associated with a spectrum of speech and developmental delays and behavior problems. Genet Med *13*, 868-880.

Sakamoto, M., Hirata, H., Ohtsuka, T., Bessho, Y., and Kageyama, R. (2003). The basic helix-loop-helix genes Hesr1/Hey1 and Hesr2/Hey2 regulate maintenance of neural precursor cells in the brain. J Biol Chem *278*, 44808-44815.

Sakashita, E., Tatsumi, S., Werner, D., Endo, H., and Mayeda, A. (2004). Human RNPS1 and Its Associated Factors: a Versatile Alternative Pre-mRNA Splicing Regulator In Vivo. Mol Cell Biol *24*, 1174-1187.

Sartorius, L.J., Weinberger, D.R., Hyde, T.M., Harrison, P.J., Kleinman, J.E., and Lipska, B.K. (2008). Expression of a GRM3 splice variant is increased in the dorsolateral prefrontal cortex of individuals carrying a schizophrenia risk SNP. Neuropsychopharmacol *33*, 2626-2634.

Saus, E., Brunet, A., Armengol, L., Alonso, P., Crespo, J.M., Fernandez-Aranda, F., Guitart, M., Martin-Santos, R., Menchon, J.M., Navines, R., *et al.* (2010). Comprehensive copy number

variant (CNV) analysis of neuronal pathways genes in psychiatric disorders identifies rare variants within patients. J Psychiatr Res *44*, 971-978.

Schneider, K. (1959). Clinical psychopathology (New York: Grune and Stratton).

Schosser, A., Pirlo, K., Gaysina, D., Cohen-Woods, S., Schalkwyk, L.C., Elkin, A., Korszun, A., Gunasinghe, C., Gray, J., Jones, L., *et al.* (2010). Utility of the pooling approach as applied to whole genome association scans with high-density Affymetrix microarrays. BMC Res Notes *3*, 274.

Schmitt, A., Hasan, A., Gruber, O., and Falkai, P. (2011). Schizophrenia as a disorder of disconnectivity. Eur Arch Psy Clin N *261*, 150-154.

Schultz, J., Milpetz, F., Bork, P., and Ponting, C.P. (1998). SMART, a simple modular architecture research tool: identification of signaling domains. Proc Natl Acad Sci USA *95*, 5857-5864.

Selemon, L.D., and Goldman-Rakic, P.S. (1999). The reduced neuropil hypothesis: a circuit based model of schizophrenia. Biol Psychiatry *45*, 17-25.

Senturk, A., Pfennig, S., Weiss, A., Burk, K., and Acker-Palmer, A. (2011). Ephrin Bs are essential components of the Reelin pathway to regulate neuronal migration. Nature *472*, 356-360.

Sham, P., Bader, J.S., Craig, I., O'Donovan, M., and Owen, M. (2002). DNA Pooling: a tool for large-scale association studies. Nat Rev Genet *3*, 862-871.

Sharma, A., Takata, H., Shibahara, K., Bubulya, A., and Bubulya, P.A. (2010). Son is essential for nuclear speckle organization and cell cycle progression. Mol Biol Cell *21*, 650-663.

Shepherd, A.M., Laurens, K.R., Matheson, S.L., Carr, V.J., and Green, M.J. (2012). Systematic meta-review and quality assessment of the structural brain alterations in schizophrenia. Neurosci Biobehav Rev *36*, 1342-1356.

Shi, J., Levinson, D.F., Duan, J., Sanders, A.R., Zheng, Y., Pe'er, I., Dudbridge, F., Holmans, P.A., Whittemore, A.S., Mowry, B.J., *et al.* (2009). Common variants on chromosome 6p22.1 are associated with schizophrenia. Nature *460*, 753-757.

Shifman, S., Johannesson, M., Bronstein, M., Chen, S.X., Collier, D.A., Craddock, N.J., Kendler, K.S., Li, T., O'Donovan, M., O'Neill, F.A., *et al.* (2008). Genome-wide association identifies a common variant in the reelin gene that increases the risk of schizophrenia only in women. PLoS Genet *4*, e28.

Shinn, A.K., Pfaff, D., Young, S., Lewandowski, K.E., Cohen, B.M., and Ongur, D. (2012). Auditory hallucinations in a cross-diagnostic sample of psychotic disorder patients: a descriptive, cross-sectional study. Compr Psychiatry *53*, 718-726.

Slatkin, M. (2008). Linkage disequilibrium--understanding the evolutionary past and mapping the medical future. Nat Rev Genet *9*, 477-485.

Spector, D.L., and Lamond, A.I. (2011). Nuclear speckles. Cold Spring Harb Perspect Biol *3*.

Srinivasan, K., Shiue, L., Hayes, J.D., Centers, R., Fitzwater, S., Loewen, R., Edmondson, L.R., Bryant, J., Smith, M., Rommelfanger, C., *et al.* (2005). Detection and measurement of alternative splicing using splicing-sensitive microarrays. Methods *37*, 345-359.

St Clair, D., Blackwood, D., Muir, W., Carothers, A., Walker, M., Spowart, G., Gosden, C., and Evans, H.J. (1990). Association within a family of a balanced autosomal translocation with major mental illness. Lancet *336*, 13-16.

Stahlberg, O., Soderstrom, H., Rastam, M., and Gillberg, C. (2004). Bipolar disorder, schizophrenia, and other psychotic disorders in adults with childhood onset AD/HD and/or autism spectrum disorders. J Neural Transm *111*, 891-902.

Stefansson, H., Ophoff, R.A., Steinberg, S., Andreassen, O.A., Cichon, S., Rujescu, D., Werge, T., Pietilainen, O.P., Mors, O., Mortensen, P.B., *et al.* (2009). Common variants conferring risk of schizophrenia. Nature *460*, 744-747.

Stefansson, H., Rujescu, D., Cichon, S., Pietilainen, O.P., Ingason, A., Steinberg, S., Fossdal, R., Sigurdsson, E., Sigmundsson, T., Buizer-Voskamp, J.E., *et al.* (2008). Large recurrent microdeletions associated with schizophrenia. Nature *455*, 232-236.

Stefansson, H., Sigurdsson, E., Steinthorsdottir, V., Bjornsdottir, S., Sigmundsson, T., Ghosh, S., Brynjolfsson, J., Gunnarsdottir, S., Ivarsson, O., Chou, T.T., *et al.* (2002). Neuregulin 1 and susceptibility to schizophrenia. Am J Hum Genet *71*, 877-892.

Steinberg, S., Mors, O., Borglum, A.D., Gustafsson, O., Werge, T., Mortensen, P.B., Andreassen, O.A., Sigurdsson, E., Thorgeirsson, T.E., Bottcher, Y., *et al.* (2011). Expanding the range of ZNF804A variants conferring risk of psychosis. Mol Psychiatr *16*, 59-66.

Stone, J.M., Howes, O.D., Egerton, A., Kambeitz, J., Allen, P., Lythgoe, D.J., O'Gorman, R.L., McLean, M.A., Barker, G.J., and McGuire, P. (2010). Altered Relationship Between Hippocampal Glutamate Levels and Striatal Dopamine Function in Subjects at Ultra High Risk of Psychosis. Biol Psychiatry *68*, 599-602.

Strange, A. (2012). Genome-wide association study implicates HLA-C*01:02 as a risk factor at the major histocompatibility complex locus in schizophrenia. Biol Psychiatry.

Strauss, G.P., Harrow, M., Grossman, L.S., and Rosen, C. (2010). Periods of recovery in deficit syndrome schizophrenia: a 20-year multi-follow-up longitudinal study. Schizophr Bull *36*, 788-799.

Suizu, F., Hiramuki, Y., Okumura, F., Matsuda, M., Okumura, A.J., Hirata, N., Narita, M., Kohno, T., Yokota, J., Bohgaki, M., *et al.* (2009). The E3 ligase TTC3 facilitates ubiquitination and degradation of phosphorylated Akt. Dev Cell *17*, 800-810.

Sullivan, P.F., Daly, M.J., and O'Donovan, M. (2012). Genetic architectures of psychiatric disorders: the emerging picture and its implications. Nature Rev Genet *13*, 537-551.

Sullivan, P.F., Kendler, K.S., and Neale, M.C. (2003). Schizophrenia as a complex trait: evidence from a meta-analysis of twin studies. Arch Gen Psychiatry *60*, 1187-1192.

Sullivan, P.F., Lin, D., Tzeng, J.Y., van den Oord, E., Perkins, D., Stroup, T.S., Wagner, M., Lee, S., Wright, F.A., Zou, F., *et al.* (2008). Genomewide association for schizophrenia in the CATIE study: results of stage 1. Mol Psychiatr *13*, 570-584.

Sun, C., Cheng, M.C., Qin, R., Liao, D.L., Chen, T.T., Koong, F.J., Chen, G., and Chen, C.H. (2011). Identification and functional characterization of rare mutations of the neuroligin-2 gene (NLGN2) associated with schizophrenia. Hum Mol Genet *20*, 3042-3051.

Swanson, D.A., Steel, J.M., and Valle, D. (1998). Identification and characterization of the human ortholog of rat STXBP1, a protein implicated in vesicle trafficking and neurotransmitter release. Genomics *48*, 373-376.

Takahashi, K., and Yamanaka, S. (2006). Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors. Cell *126*, 663-676.

Talbot, K., Eidem, W.L., Tinsley, C.L., Benson, M.A., Thompson, E.W., Smith, R.J., Hahn, C.G., Siegel, S.J., Trojanowski, J.Q., Gur, R.E., *et al.* (2004). Dysbindin-1 is reduced in intrinsic, glutamatergic terminals of the hippocampal formation in schizophrenia. The J Clin Invest *113*, 1353-1363.

Talkowski, M.E., Rosenfeld, J.A., Blumenthal, I., Pillalamarri, V., Chiang, C., Heilbut, A., Ernst, C., Hanscom, C., Rossin, E., Lindgren, A.M., *et al.* (2012). Sequencing chromosomal abnormalities reveals neurodevelopmental loci that confer risk across diagnostic boundaries. Cell *149*, 525-537.

Tan, W., Wang, Y.H., Gold, B., Chen, J.S., Dean, M., Harrison, P.J., Weinberger, D.R., and Law, A.J. (2007). Molecular cloning of a brain-specific, developmentally regulated neuregulin 1 (NRG1) isoform and identification of a functional promoter variant associated with schizophrenia. J Biol Chem *282*, 24343-24351.

Tandon, R., Nasrallah, H.A., and Keshavan, M.S. (2009). Schizophrenia, "just the facts" 4. Clinical features and conceptualization. Schizophr Res *110*, 1-23.

Tanganelli, S., Antonelli, T., Tomasini, M.C., Beggiato, S., Fuxe, K., and Ferraro, L. (2012). Relevance of dopamine D(2)/neurotensin NTS1 and NMDA/neurotensin NTS1 receptor interaction in psychiatric and neurodegenerative disorders. Curr Med Chem *19*, 304-316.

Tellam, J.T., McIntosh, S., and James, D.E. (1995). Molecular identification of two novel Munc-18 isoforms expressed in non-neuronal tissues. J Biol Chem *270*, 5857-5863.

Tetsukati, T.U., H.; Imai, H; Ono'l, T.; Sonta, S.; Takahashi, N.; Asamitsu, K.; Okamoto, T. (2000). Inhibition of nuclear factor-kB-mediated transcription by association with amino-terminal enhancer of split, a groucho-related protein lacking WD40 repeats. J Biol Chem *275*, 4383 - 4390.Ule, A.J., K. B.; Ruggiu, M.; Mele, A.; Ule, A.; Darnell, R. B. (2003). CLIP identifies nove-regulated RNA networks in the brain. Science *302*, 1212 - 1215.

Thiery, J.P., and Sleeman, J.P. (2006). Complex networks orchestrate epithelial-mesenchymal transitions. Nat Rev Mol Cell Bio *7*, 131-142.

Timmerman, L.A., Grego-Bessa, J., Raya, A., Bertran, E., Perez-Pomares, J.M., Diez, J., Aranda, S., Palomo, S., McCormick, F., Izpisua-Belmonte, J.C., *et al.* (2004). Notch promotes epithelial-mesenchymal transition during cardiac development and oncogenic transformation. Genes Dev *18*, 99-115.

Tost, H., and Meyer-Lindenberg, A. (2011). Dopamine-Glutamate Interactions: A Neural Convergence Mechanism of Common Schizophrenia Risk Variants. Biol Psychiatry *69*, 912-913.

Tsurugi, K., Collatz, E., Todokoro, K., Ulbrich, N., Lightfoot, H.N., and Wool, I.G. (1978). Isolation of eukaryotic ribosomal-proteins - purification and characterization of 60-s ribosomal-subunit protein LA, LB, LF, P1, P2, L13', L14, L18', L20, and L38. J Biol Chem *253*, 946-955.

Ule, J., Ule, A., Spencer, J., Williams, A., Hu, J.S., Cline, M., Wang, H., Clark, T., Fraser, C., Ruggiu, M., *et al.* (2005). Nova regulates brain-specific splicing to shape the synapse. Nat Genet *37*, 844-852.

Umeda-Yano, S., Hashimoto, R., Yamamori, H., Okada, T., Yasuda, Y., Ohi, K., Fukumoto, M., Ito, A., and Takeda, M. (2013). The regulation of gene expression involved in TGF-β signaling by ZNF804A, a risk gene for schizophrenia. Schizophr Res *146*, 273-278

Underwood, J.G., Boutz, P.L., Dougherty, J.D., Stoilov, P., and Black, D.L. (2005). Homologues of the Caenorhabditis elegans Fox-1 protein are neuronal splicing regulators in mammals. Mol Cell Biol *25*, 10005-10016.

Vacic, V., McCarthy, S., Malhotra, D., Murray, F., Chou, H.H., Peoples, A., Makarov, V., Yoon, S., Bhandari, A., Corominas, R., *et al.* (2011). Duplications of the neuropeptide receptor gene VIPR2 confer significant risk for schizophrenia. Nature *471*, 499-503.

Vaughan, P.F., Peers, C., and Walker, J.H. (1995). The use of the human neuroblastoma SH-SY5Y to study the effect of second messengers on noradrenaline release. Gen Pharmacol *26*, 1191-1201.

Ventruti, A., Kazdoba, T.M., Niu, S., and D'Arcangelo, G. (2011). Reelin deficiency causes specific defects in the molecular composition of the synapses in the adult brain. Neurosci *189*, 32-42.

Vincent, A.J., Lau, P.W., and Roskams, A.J. (2008). SPARC is expressed by macroglia and microglia in the developing and mature nervous system. Dev Dyn *237*, 1449-1462.

Voineagu, I., Wang, X., Johnston, P., Lowe, J.K., Tian, Y., Horvath, S., Mill, J., Cantor, R.M., Blencowe, B.J., and Geschwind, D.H. (2011). Transcriptomic analysis of autistic brain reveals convergent molecular pathology. Nature *474*, 380-384.

Voineskos, A.N., Lerch, J.P., Felsky, D., Tiwari, A., Rajji, T.K., Miranda, D., Lobaugh, N.J., Pollock, B.G., Mulsant, B.H., and Kennedy, J.L. (2011). The ZNF804A gene: characterization of a novel neural risk mechanism for the major psychoses. Neuropsychopharmacol *36*, 1871-1878.

Volk, D.W., Austin, M.C., Pierri, J.N., Sampson, A.R., and Lewis, D.A. (2000). Decreased glutamic acid decarboxylase67 messenger RNA expression in a subset of prefrontal cortical gamma-aminobutyric acid neurons in subjects with schizophrenia. Arch Gen Psychiatry *57*, 237-245.

Vostrikov, V.M., Uranova, N.A., and Orlovskaya, D.D. (2007). Deficit of perineuronal oligodendrocytes in the prefrontal cortex in schizophrenia and mood disorders. Schizophr Res *94*, 273-280.

Vrijenhoek, T., Buizer-Voskamp, J.E., van der Stelt, I., Strengman, E., Sabatti, C., Geurts van Kessel, A., Brunner, H.G., Ophoff, R.A., and Veltman, J.A. (2008). Recurrent CNVs disrupt three candidate genes in schizophrenia patients. Am J Hum Genet *83*, 504-510.

Wahl, M.C., Will, C.L., and Lührmann, R. (2009). The spliceosome: design principles of a dynamic RNP machine. Cell *136*, 701-718.

Walker, E.F., Lewine, R.R., and Neumann, C. (1996). Childhood behavioral characteristics and adult brain morphology in schizophrenia. Schizophr Res *22*, 93-101.

Walsh, T., McClellan, J.M., McCarthy, S.E., Addington, A.M., Pierce, S.B., Cooper, G.M., Nord, A.S., Kusenda, M., Malhotra, D., Bhandari, A., *et al.* (2008). Rare structural variants disrupt multiple genes in neurodevelopmental pathways in schizophrenia. Science *320*, 539-543.

Walter, H., Schnell, K., Erk, S., Arnold, C., Kirsch, P., Esslinger, C., Mier, D., Schmitgen, M.M., Rietschel, M., Witt, S.H., *et al.* (2011). Effects of a genome-wide supported psychosis risk variant on neural activation during a theory-of-mind task. Molecular psychiatr *16*, 462-470.

Walters, J.T., Corvin, A., Owen, M.J., Williams, H., Dragovic, M., Quinn, E.M., Judge, R., Smith, D.J., Norton, N., Giegling, I., *et al.* (2010). Psychosis susceptibility gene ZNF804A and cognitive performance in schizophrenia. Arch Gen Psychiatry *67*, 692-700.

Wang, W.Y., Barratt, B.J., Clayton, D.G., and Todd, J.A. (2005). Genome-wide association studies: theoretical and practical concerns. Nat Rev Genet *6*, 109-118.

Wang, Y., Gracheva, E.O., Richmond, J., Kawano, T., Couto, J.M., Calarco, J.A., Vijayaratnam, V., Jin, Y., and Zhen, M. (2006). The C2H2 zinc-finger protein SYD-9 is a putative posttranscriptional regulator for synaptic transmission. Proc Natl Acad Sci USA *103*, 10450-10455.

Wang, Y.T., Chuang, J.Y., Shen, M.R., Yang, W.B., Chang, W.C., and Hung, J.J. (2008). Sumoylation of specificity protein 1 augments its degradation by changing the localization and increasing the specificity protein 1 proteolytic process. J Mol Biol *380*, 869-885.

Wang, Y.T., Yang, W.B., Chang, W.C., and Hung, J.J. (2011). Interplay of posttranslational modifications in Sp1 mediates Sp1 stability during cell cycle progression. J Mol Biol *414*, 1-14.

Warzecha, C.C., Shen, S., and Xing, Y. (2009). The epithelial splicing factors ESRP1 and ESRP2 positively and negatively regulate diverse types of alternative splicing events. RNA Biol *6*, 546-546.

Weickert, T.W., Goldberg, T.E., Gold, J.M., Bigelow, L.B., Egan, M.F., and Weinberger, D.R. (2000). Cognitive impairments in patients with schizophrenia displaying preserved and compromised intellect. Arch Gen Psychiat *57*, 907-913.

Weinberger, D.R. (1987). Implications of normal brain development for the pathogenesis of schizophrenia. Arch Gen Psychiat *44*, 660-669.

Wernig, M., Meissner, A., Foreman, R., Brambrink, T., Ku, M., Hochedlinger, K., Bernstein, B.E., and Jaenisch, R. (2007). In vitro reprogramming of fibroblasts into a pluripotent ES-cell-like state. Nature *448*, 318-U312.

Whistler, T.C., C. F.; Lin, J. M.; Lonergran, W.; Reeves, W.C. (2010). The comparison of different pre- and post- analysis filter for determination of exon -leve alternative splicing events using affymetrix arrays. J Biomol Tech *21*, 44 - 53.

Wicks, S., Hjern, A., and Dalman, C. (2010). Social risk or genetic liability for psychosis? A study of children born in Sweden and reared by adoptive parents. Am J Psychiatry *167*, 1240-1246.

Wilbert, M.L., Huelga, S.C., Kapeli, K., Stark, T.J., Liang, T.Y., Chen, S.X., Yan, B.Y., Nathanson, J.L., Hutt, K.R., Lovci, M.T., *et al.* (2012). LIN28 Binds Messenger RNAs at GGAGA Motifs and Regulates Splicing Factor Abundance. Mol Cell.

Williams, H.J., Norton, N., Dwyer, S., Moskvina, V., Nikolov, I., Carroll, L., Georgieva, L., Williams, N.M., Morris, D.W., Quinn, E.M., *et al.* (2011). Fine mapping of ZNF804A and genome-wide significant evidence for its involvement in schizophrenia and bipolar disorder. Molecular psychiatr *16*, 429-441.

Williams, N.M., O'Donovan, M.C., and Owen, M.J. (2005). Is the dysbindin gene (DTNBP1) a susceptibility gene for schizophrenia? Schizophrenia Bull *31*, 800-805.

Wu, J.Q., Wang, X., Beveridge, N.J., Tooney, P.A., Scott, R.J., Carr, V.J., and Cairns, M.J. (2012). Transcriptome Sequencing Revealed Significant Alteration of Cortical Promoter Usage and Splicing in Schizophrenia. Plos One *7*.

Xiao, B., Li, W.Q., Zhang, H.X., Lv, L.X., Song, X.Q., Yang, Y.F., Li, W., Yang, G., Jiang, C.D., Zhao, J.Y., *et al.* (2011). Association of ZNF804A polymorphisms with schizophrenia and antipsychotic drug efficacy in a Chinese Han population. Psychiatry Res *190*, 379-381.

Xing, Y., Stoilov, P., Kapur, K., Han, A., Jiang, H., Shen, S., Black, D.L., and Wong, W.H. (2008). MADS: a new and improved method for analysis of differential alternative splicing by exon-tiling microarrays. RNA *14*, 1470-1479.

Xu, B., Roos, J.L., Dexheimer, P., Boone, B., Plummer, B., Levy, S., Gogos, J.A., and Karayiorgou, M. (2011). Exome sequencing supports a de novo mutational paradigm for schizophrenia. Nat Genet *43*, 864-868.

Xu, B., Roos, J.L., Levy, S., van Rensburg, E.J., Gogos, J.A., and Karayiorgou, M. (2008). Strong association of de novo copy number mutations with sporadic schizophrenia. Nat Genet *40*, 880-885.

Yamada, K., and Miyamoto, K. (2005). Basic helix-loop-helix transcription factors, BHLHB2 and BHLHB3; Their gene expressions are regulated by multiple extracellular stimuli. Front Biosci *10*, 3151-3171.

Yamada, K., Iwayama, Y., Hattori, E., Iwamoto, K., Toyota, T., Ohnishi, T., Ohba, H., Maekawa, M., Kato, T., and Yoshikawa, T. (2011). Genome-wide association study of schizophrenia in Japanese population. PLoS One *6*, e20468.

Yang, Y.Y., Yin, G.L., and Darnell, R.B. (1998). The neuronal RNA-binding protein Nova-2 is implicated as the autoantigen targeted in POMA patients with dementia. Proc Natl Acad Sci USA *95*, 13254-13259.

Yao, F., Svensjö, T., Winkler, T., Lu, M., Eriksson, C., and Eriksson, E. (1998). Tetracycline repressor, tetR, rather than the tetR-mammalian cell transcription factor fusion derivatives, regulates inducible gene expression in mammalian cells. Hum Gene Ther *9*, 1939-1950.

Yarden, Y., and Sliwkowski, M.X. (2001). Untangling the ErbB signalling network. Nature Rev Mol Cell Biol *2*, 127-137.

Yeo, G.W., Coufal, N.G., Liang, T.Y., Peng, G.E., Fu, X.D., and Gage, F.H. (2009). An RNA code for the FOX2 splicing regulator revealed by mapping RNA-protein interactions in stem cells. Nat Struct Mol Biol *16*, 130-137.

Yi, J.J., Barnes, A.P., Hand, R., Polleux, F., and Ehlers, M.D. (2010). TGF-beta Signaling Specifies Axons during Brain Development. Cell *142*, 144-157.

Yu, W., Lin, Z., Hegarty, J.P., John, G., Chen, X., Faber, P.W., Kelly, A.A., Wang, Y., Poritz, L.S., Schreiber, S., *et al.* (2010). Genes regulated by Nkx2-3 in siRNA-mediated knockdown B cells: Implication of endothelin-1 in inflammatory bowel disease. Mol Genet and Metab *100*, 88-95.

Yue, W.H., Wang, H.F., Sun, L.D., Tang, F.L., Liu, Z.H., Zhang, H.X., Li, W.Q., Zhang, Y.L., Zhang, Y., Ma, C.C., *et al.* (2011). Genome-wide association study identifies a susceptibility locus for schizophrenia in Han Chinese at 11p11.2. Nat Genet *43*, 1228-1231.

Zavadil, J., Cermak, L., Soto-Nieves, N., and Bottinger, E.P. (2004). Integration of TGF-beta/Smad and Jagged1/Notch signalling in epithelial-to-mesenchymal transition. Embo Journal *23*, 1155-1165.

Zhang, R., Zhong, N.N., Liu, X.G., Yan, H., Qiu, C., Han, Y., Wang, W., Hou, W.K., Liu, Y., Gao, C.G., *et al.* (2010). Is the EFNB2 locus associated with schizophrenia? Single nucleotide polymorphisms and haplotypes analysis. Psychiatry Res *180*, 5-9.

Zoghbi, H.Y. (2003). Postnatal neurodevelopmental disorders: meeting at the synapse? Science *302*, 826-830.

Zondervan, K.T., and Cardon, L.R. (2004). The complex interplay among factors that influence allelic association. Nat Rev Genet *5*, 89-100.

Appendix 1: The sequences of the PCR primers

Appendix 1.1: ZNF804A PCR primers

Primer name	Sequence	Purpose
ZNF804A primers		
HZNF804A-Cterm Y2HFWD (EcoRI)	CTGGAATTCTGGGGAAATATCTATTGGAACCAAT	Cloning C-term into pHybLex/Zeo
HZNF804A-Y2H R2 (Sal)	CTGGTCGACCTAGAAGAGAGGTTGCAAAGG	Cloning C-term into pHybLex/Zeo
ZNF804A-Y2H F1	CTGGAATTCATGGAGTGTTACTACATTGTCAT	Cloning N-term into pHybLex/Zeo
ZNF804A-Y2H R1	CTGGTCGACCTAGCATTTTCAGTGTAACTATTCTC	Cloning N-term into pHybLex/Zeo
ZNF EcoRI F1	CTGGAATTCGGATGGAGTGTTACTACATTGTCAT	Cloning full-length ZNF804A into YFPN1
ZNF-YFPN1_REV	CTGCGTCGACTGGAAGAGAGGTTGCAAAAGGA	Cloning full-length ZNF804A into YFPN2
ZNF_NT-YFPN1_REV	CTGCGTCGACTGAGCATTTTCAGTGTAACTATTCTC	Cloning N-terminal of ZNF804A into YFPN1
ZNF CT_EcoYFP FWD	CTGGAATTCTGGGGAAATATCTATTGGAACCAAT	Cloning C-terminal of ZNF804A into YFPN1
ZNF-Neo_EcoFWD	CTGGAATTCATGGAGTGTTACTACATTGTCAT	Cloning of full-length ZNF804A into pCIneo
ZNF-myc(neo)_MLU FWD	CTGACGCGTATGGCATCAATGCAGAAGCTG	Cloning of full-length ZNF804A into pCIneo
ZNF_myc_kozak fwd	CTGACGGGTATGGCATCAATGCAGAAGCTG	Cloning into pcDNA5/FRT/TO
ZNF_500_REV	TTGAAATCTTGCTGGTTATTAACT	RT-PCR
hZNF804A749fwd	GCATCTAAATCCAGGAAAGATC	sequencing
hZNF804A1497fwd	GTGGTAACAGTTTTGAGTTGTTACG	sequencing
hZNF804A2254fwd	GTTATGTCAGCATCATCATATGG	sequencing
hZNF804A2970fwd	CTTCCTTAAATCCTCTGGATAGG	sequencing
hZNF804A1202fwd	AGATTTGTCCCCAGTGC	sequencing

Appendix 1.2: TCF4 PCR primers

Primer name	Sequence	Purpose
hTCF4_GFPC1_Sal_Rev	CTGGTCGACTTACATCTGTCCCATGTGATTC	Cloning into pcDNA5/FRT/TO
hTCF4 kozak EGFP C1	CCGGTCGCCACCATGGTGA	Cloning into pcDNA5/FRT/TO
hTCF4_1467_Fwd	TCTGCGACTTCCCCTGACC	sequencing
TCF4_523_REV	ATGGCACTACTGTGAAGAGG	RT-PCR

Appendix 1.3: GPATCH8 PCR primers

Primer name	Sequence	Purpose
GPATCH8 primers		
hGPATCH8_YFP_age_rev	CTGCACCGGTCCCGTGCCATGGCTGGGGG	Cloning into pEYFP
hGPATCH8_start35xho_YFPN1 fwd	CTGCTCGAGATGGCGGACCGCTTCTCC	Cloning into pEYFP
hGPATCH8_sfi_myc_fwd	CTGATGGCCATGGAGGCCATGGCGGACCGCTTCTCC	Cloning into pCMV-myc
hGPATCH8_stop rev	CTGCGCGGCCGCTCACGTGCCATGGCTGGG	Cloning into pCMV-myc

Appendix 1.4: Solaris Q-PCR primers

Primer name	Sequence	Purpose
ACTB fwd	TGGAGAAAATCTGGCACCAC	Q-PCR
ACTB rev	GGTCTCAAACATGATCTGG	Q-PCR
ACTB probe	ACCGCGAGAAGATGACC	Q-PCR
GAPDH fwd	GCCTCAAGATCATCAGCAATG	Q-PCR
GAPDH rev	CTTCCACGATACCAAAGTTGTC	Q-PCR
GAPDH probe	GCCAAGGTCATCCATGA	Q-PCR
hZNF804A fwd	TCTCAGCAACGGACACTTTC	Q-PCR
hZNF804A rev	CTCAGCATAGTCCAGAGTT	Q-PCR
hZNF804A probe	CAAGAACGGGAACAAAAC	Q-PCR

Appendix 1.5: Vector and miscellaneous PCR primers

Primer name	Sequence	Purpose
Vector primers		
MYC FWD	ATGGCATCAATGCAGAAGCTG	RT-PCR
EGFP FWD	ACGGCATCAAGGTGAACTTCA	RT-PCR
M13rev	CAGGAAACAGCTATGAC	sequencing
pEYFP-N1 fwd	GCTGGTTTAGTGAACCGTCA	sequencing
pEYFP-N1 rev	TCGAAGTCCAGCTCGACCAG	sequencing
pCMV-myc fwd	ATGGCATCAATGCAGAAGCTG	sequencing
pCMV-myc rev	CACTGCATTCTAGTTGTGGTTTGTCCA	sequencing
pPC86 fwd	TATAACGCGTTTGGAATCACT	sequencing
pDEST22 fwd	TATAACGCGTTTGGAATCACT	RT-PCR
pDEST22 rev	GTCTCCAATCAAGGTTGTCGGCT	RT-PCR
Misc.		
Hygromycin fwd	GCTCATCGAGAGCCTGCG	RT-PCR
Hygromycin rev	GCTCATCGAGAGCCTGCG	RT-PCR

Appendix 1.6: PCR primers used to validate genes which showed differential expression after *ZNF804A* knockdown

Primer name	Forward primer sequence	Reverse primer sequence	Purpose	amplification efficiency	R^2	Validation for ΔCt
ACTB	ACGGCCAGGTCATCACCATTG	GGAGTTGAAGGTAGTTTCGTGGATG	Q-PCR	96	0.9975	–
NPY	GCGCTGCGACACTACATCAA	GGGCTGGATCGTTTTCCATA	Q-PCR	94	0.9926	0.0273
CCL2	GAAGAATCACCAGCAGCAAGTGT	GCTTGTCCAGGTGGTCCATG	Q-PCR	91	0.9991	0.0973
SPARC	TACATCGGGCCTTGCAAATAC	GGGTGACCAGGACGTTCTTG	Q-PCR	92	0.9995	0.0113
TMEFF2	CACAAGGAAATGCCCCAGAA	GATTAACCTCGTGGACGCTCTT	Q-PCR	99	0.9794	0.1074
EGR1	CTTCGCCTGCGACATCTGT	TTTGTCTGCTTTCTTGTCCTTCTG	Q-PCR	100	0.9985	0.0933
PDK1	AGCCATCATTGCACGTGTCTT	CCTTGACCATGCCACTGTACTC	Q-PCR	95	1.0000	0.1246
FSTL4	GTATGCGCTGCTACCAAGATTG	GCAGCTTGGCACAGAAATGAT	Q-PCR	90	0.9954	0.0764
EFNB2	CTGCTGCTGCCTCTGAAACA	CCTACTGGCCTCTTCGATCTCA	Q-PCR	96	0.999	0.0231

Appendix 1.7 PCR primers used to validate putative alternative splicing events after *ZNF804A* knockdown

Exon array validation – alternative splicing						
Primer name	**Forward primer sequence**	**Reverse primer sequence**	**Purpose**	**amplification efficiency**	**R^2**	**Validation for ΔC_t**
NFYA	CTGGAATTCTTCGACAGAGCAGATTGTTGT	CTGGCGGCCGCAGCTATGATGGGTTGGCCAGTTGA	RT-PCR			
	CTGGAATTCTTCGACAGAGCAGATTGTTGT	CTGGCGGCCGCAGCTCCACCTGGGCCTGGGCCTCAGTCT	Q-PCR exon	108	0.9969	0.0575
	TGGAGGCCTCGACGGTTAC	GCCGCAGCACGAAGTTAAAC	Q-PCR standard	105	0.9999	–
SIPA1L1	CTGGAATTCCAGATATTGGTGGCAGCGG	CTGGCGGCCGCAGCTATCCTGCCGCATTAATCTC	RT-PCR			
	CCTAGCTGGCAAAGAAGTGA	TTCATGCTGCTCGAGGACAA	Q-PCR exon	107	0.9979	0.7672
	CCGTCCCAGGTGCAGAGT	GGCATCTTCCCATCTCCTTTT	Q-PCR standard	93	0.9999	–
ENAH	GCAGCAAGTCACCTGTTATCT	CTGGACTCCATTGGCACTG	RT-PCR			
	TGTTATCTCCAGACGGGATTC	TCAGCCTGTCATAGTCAAGTCCTT	Q-PCR exon	93	0.9963	0.0154
	CAGAGTGGTGGGCAGGAAGA	CCCTTTAGGAATGGCACAGTTT	Q-PCR standard	92	0.9998	–
G3BP2	CCTCCTCCGGCAGAACCT	ACGAGGTGGCTGAGATTGAA	RT-PCR			
	CAAAGGCTTTCTCCTGGGCT	CGACTCTTGGCTGTGAGAC	Q-PCR exon	96	0.9937	0.0022
	GCTCCGGAATATTTACACAGGTTT	GCATCTACTCCACCATGAACATAGG	Q-PCR standard	95	0.9991	–
PKM2	CTGGAATTCGCACAGCACAGGGAAGATG	CTGGCGGCCGCAGCTGGATGGAGCCGACTGCATC	RT-PCR			
	AGTGAT GTGGCCAATG CAGT	TGGAACATGGCTGCCTCA	Q-PCR exon	104	0.9628	0.0059
	CCATGAATGTTGGCAAGGCC	TCACGGCACAGGAACAA	Q-PCR standard	100	0.9989	–
PTPRR 130846.2	TTCTCAAGCTCTCATTTAACGT	CTGTAAAGAATCATCAAACAC	Q-PCR	95	0.9974	0.1239
PTPRR 002849	TGGAATTACAGAAGTCTCTCC	GTCTTGTCTTAAGGAAAGCT	Q-PCR	93	0.9939	0.2325
ATP11C	CTGAGATTCTTCTGATAGTATTA	TGTCAGTATGCTTGTAGGAC	RT-PCR			
STXBP1	ACGAGGTGACCCAGGCCA	TCTTCAGTGTGTCCAGCAGT	RT-PCR			
BAZ1A	AAGCTAAGATGTGCAGTCTGAG	ATCTTCCAGGCAATAGCTGAA	RT-PCR			
RCOR2	TGGACCACAGATGAGCAGC	GGCTGTCGGAGCAGAGTG	RT-PCR			
SEC14L1	ACCGGCGGTCGAGCTACG	CTGAATATCCTCAAGGAGCC	RT-PCR			
MAP2K5	GGAGCAGTATGGAATTCATTC	CAACAATGCACTGCAGAAGC	RT-PCR			

Appendix 2: Summary of the *ZNF804A* expression constructs

Construct name	Description	Primers used to clone the insert (see Appendix 1.1)
FL-ZNF804A-YFPN1	Full-length ZNF804A cloned into pEYFP-N1	ZNF EcoRI F1 + ZNF-YFPN1_REV
C-term-ZNF804A-YFPN1	C-terminal ZNF804A cloned into pEYFP-N1	ZNF CT_EcoRIYFP FWD + ZNF_ YFPN1_REV
N-term-ZNF804A-YFPN1	N-terminal ZNF804A cloned into pEYFP-N2	ZNF EcoRI FWD + ZNF_NT-YFPN1_REV
FL-ZNF804A-myc	Full-length ZNF804A cloned into pCMV-myc	ZNF EcoRI F1 + ZNF-YFPN1_REV
FL-ZNF804A-myc-neo	Full-length ZNF804A-myc cloned into pCIneo	ZNF-myc(neo)_MLU FWD + ZNF-Sal REV(Y2HR2)

Appendix 2.1: The *ZNF804A* expression constructs

Prior to the start of my studies, a number of different *ZNF804A* expression constructs were used to try to transiently over-express ZNF804A. Each of these *ZNF804A* expression constructs was tested in COS-7, HEK293T, SH-SY5Y and HeLa cells. In each instance, no tagged-ZNF804A expression was detected by western blot analysis or immunocytochemistry using an antibody designed to target the tag.

Appendix 3: Description of the custom anti-ZNF804A antibodies

The custom antibodies used in this study were designed and generated by C.L.Tinsley. To generate the human ZNF804A antibodies, an N-terminus immunogen (highlighted in yellow in Appendix 3.1) with a thioredoxin-tag was used. The resulting antibodies were designated 3077 and 3078. To generate the mouse Zfp804a antibodies, an N-terminus immunogen (highlighted in yellow in Appendix 3.2) with a thioredoxin-tag was used. The resulting antibodies were designated 001 and 002.

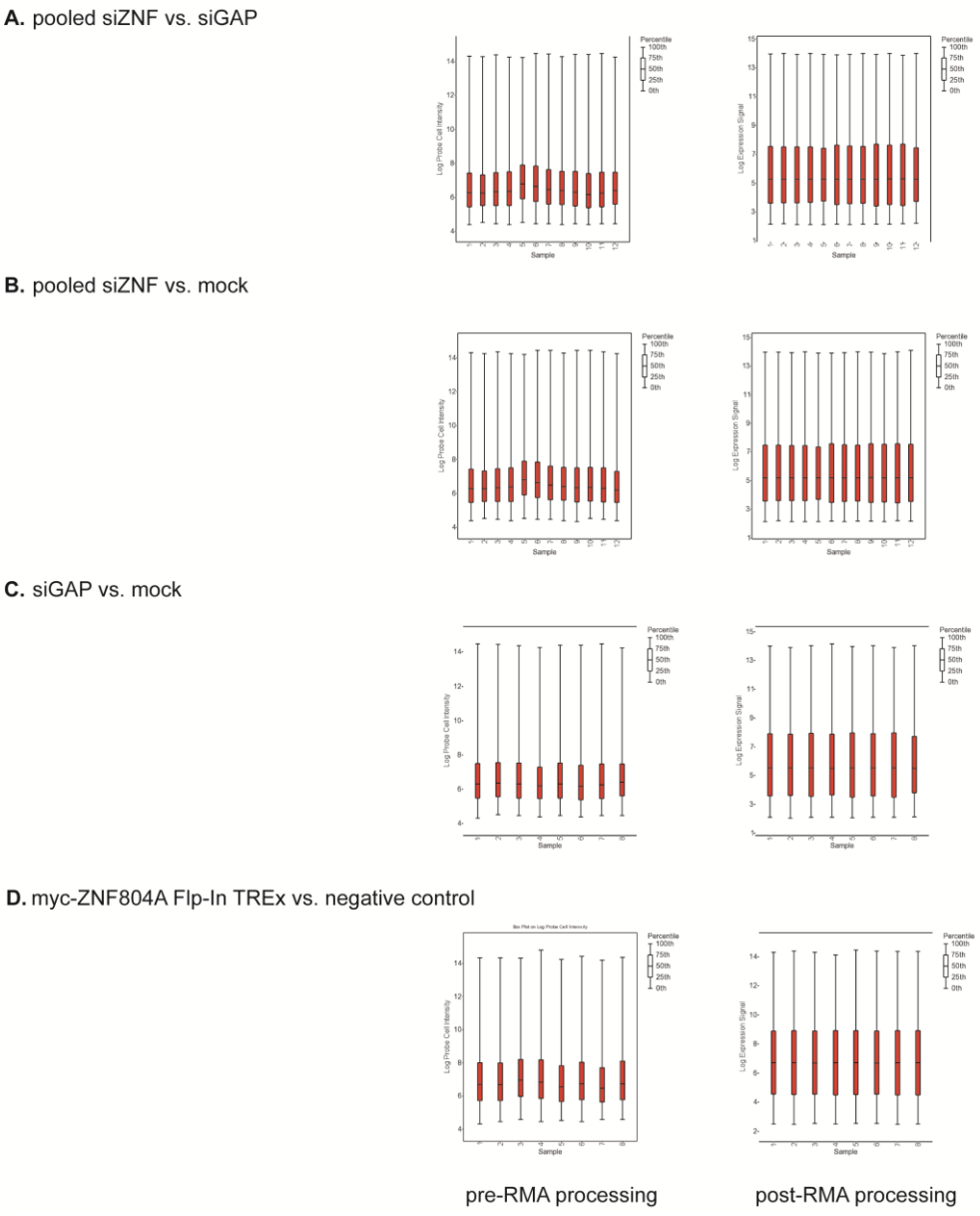
Appendix 3.1 The human ZNF804A amino acid sequence

MECYYIVISSTHLSNGHFRNIKGVFRGPLSKNGNKTLDYAEKENTIAKALEDLKANFYCELCDKQYYKHQ
EFDNHINSYDHAHKQRLKELKQREFARNVASKSRKDERKQEKALQRLHKLAELRKETVCAPGSGPMFKST
TVTVRENCNEISQRVVVDSVNNQQDFKYTLIHSEENTKDATTVAEDPESANNYTAKNNQVGDQAQGIHRH
KIGFSFAFPKKASVKLESSAAAFSEYSDDASVGKGFSRKSRFVPSACHLQQSSPTDVLLSSEEKTNSFHP
PEAMCRDKETVQTQEIKEVSSEKDALLLPSFCKFQLQLSSDADNCQNSVPLADQIPLESVVINEDIPVSG
NSFELLGNKSTVLDMSNDCISVQATTEENVKHNEASTTEVENKNGPETLAPSNTEEVNITIHKKTNFCKR
QCEPFVPVLNKHRSTVLQWPSEMLVYTTTKPSISYSCNPLCFDFKSTKVNNNLDKNKPDLKDLCSQQKQE
DICMGPLSDYKDVSTEGLTDYEIGSSKNKCSQVTPLLADDILSSSCDSGKNENTGQRYKNISCKIRETEK
YNFTKSQIKQDTLDEKYNKIRLKETHEYWFHKSRRKKKRKKLCQHHHMEKTKESETRCKMEAENSYTENA
GKYLLEPISEKQYLAAEQLLDSHQLLDKRPKSESISLSDNEEMCKTWNTEYNTYDTISSKNHCKKNTILL
NGQSNATMIHSGKHNLTYSRTYCCWKTKMSSCSQDHRSLVLQNDMKHMSQNQAVKRGYNSVMNESERFYR
KRRQHSHSYSSDESLNRQNHLPEEFLRPPSTSVAPCKPKKKRRRKRGRFHPGFETLELKENTDYPVKDNS
SLNPLDRLISEDKKEKMKPQEVAKIERNSEQTNQLRNKLSFHPNNLLPSETNGETEHLEMETTSGELSDV
SNDPTTSVCVASAPTKEAIDNTLLEHKERSENINLNEKQIPFQVPNIERNFRQSQPKSYLCHYELAEALP
QGKMNETPTEWLRYNSGILNTQPPLPFKEAHVSGHTFVTAEQILAPLALPEQALLIPLENHDKFKNVPCE
VYQHILQPNMLANKVKFTFPPAALPPPSTPLQPLPLQQSLCSTSVTTIHHTVLQQHAAAAAAAAAAAAAG
TFKVLQPHQQFLSQIPALTRTSLPQLSVGPVGPRLCPGNQPTFVAPPQMPIIPASVLHPSHLAFPSLPHA
LFPSLLSPHPTVIPLQPLF

Appendix 3.2 The mouse Zfp804a amino acid sequence

MECYYIVISSTHLSNGHFRNIKGVFRGPLSKNGNKTLDYAEKENTIAKALEDLKANFYCELCDKQYYKHQ
EFDNHINSYDHAHKQRLKELKQREFARNVASKSRKDERKQEKALQRLHKLAELRKETVCAPGSGPMFKST
TVTVRENLNDVSQRESVDPINNQQDLIPSEEKERDGTTALAETETASNCTANNCQIGDQSQAVHRHRIGF
SFAFPKKATVKLESCSAAAFSEYSDESSMEKEFSRKTRFVPSTSHLQLPPPTCELLSSEEKGNSPPPEAM
CTDKATAQTEERKITSNENNTLLTSSFCQLQHYIPTCSEADTCQNLAPFEDQLPMEAVIVNEDGPVSKSN
PNIIEKNPTVPNDRTSAQMTTEENITINDVTKMETRNKIDHEPLTPSSTIEESIRLQKRPDLCKRQCDPF
VPVLNKLGSTVLQWPSEMLAYTTTEPSISYSCNPLCFDFKSTKLNNNQDKNKLTLNDLFSEQKEDCCKGA
HADCKDVPIARVTNDETGHSKNDYPQVTTLSPVHVLSNGCDLGQNENVGQRYKHISCTNRQTNKYKFTRC
QIKRDMLNEKYDKMRLKETREHWFHKSRRKKKRRKLCRYHPGKSSKEPEGSGKTERSRQRTDEARKNPRE
PVLEKHPRSPERASDLHQLPDERPKAASTHLGEKETMNTTVNTESNDAAPGSQNCGGKNATVVNEQAKPL
VIHSVKQNLTYVRTYCCWKARTSRYQEDDGSLASQSNVKGPTQNQPVKRGYSSLTNDSERIHRKRRQHSC
SYSSDESLNQQHHLGEYLKPLSTSLISCQPKKKRRRKRSRLHIGDGTTKMKGNSNYPMKCSSLSQPDELA
KDCIKEDINPQENVSIEKNSEQTEQTEIKGMLHPYNPLLSEPSGEGEHSVTETTPCDSSQTSNDLATPVN
VTRDPSNSTTDNTLLEHNQRSQTTNSNEKQTPFKVTNPERNFRHSQAKSYICRYELAETIPQGKTNEAST
EWLCYNSGILNTQPPLQFKEAHVSGHAFVTTEQILAPLPLPEQALLIPLENHDKLKHLPCEVYQHIIQPN
MLTNKVKFTFPPPPLPPPSTPVQPLPLQRPFCSTSVTTIHHTVLQHHAAAAAAAAAAAAAAGTFKVLQPHQ
QFLPQVPALARTSIPQISVGTVGPRLCPGGQPALVASPQMPIIPASVLHPSPLAFPPLPHSFFPSLLSPH
PTVIPLQPLF

Appendix 4: Quality assessment of the RMA processing step



Appendix 4.1: Quality assessment of the RMA processing step in the exon array analyses. On import to the PGS the data was processed using the robust multiarray averaging (RMA) algorithm. The charts show the median distributions of the $\log_2$ expression signals across the (**A** - **C**) knockdown and (**D**) over-expression exon array data pre- and post RMA processing.

Appendix 5: Supplementary data for Chapter Four

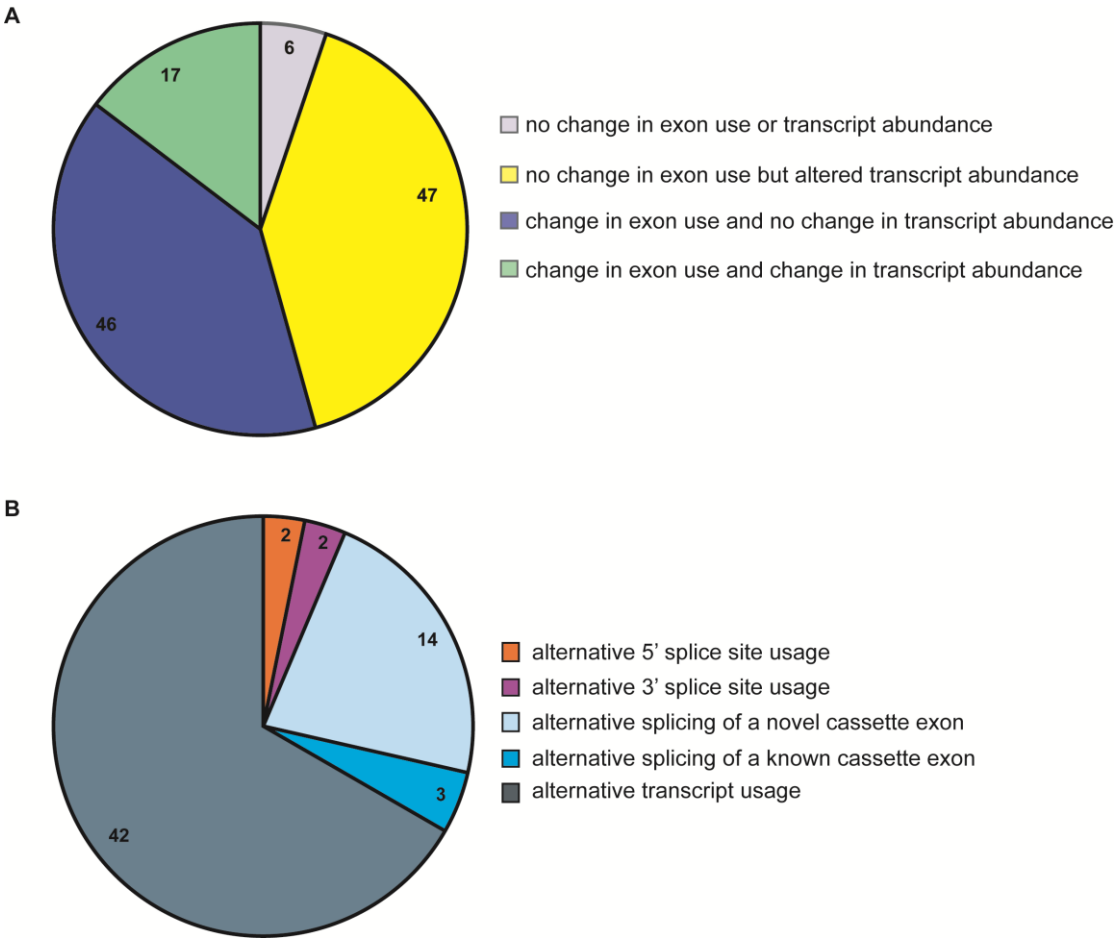
Appendix 5.1: Location of siRNA target sites and Q-PCR primer sequences in relation to gene structure

Two siRNA duplexes were designed to target exons two and three of *ZNF804A*. These siRNAs were designated siZNFA and siZNFB respectively. The location of the sites of siRNA directed cleavage are illustrated by the red arrows (not to scale). To confirm *ZNF804A* knockdown, Solaris Q-PCR primer/probe sets were designed complementary to *ZNF804A*. The location of the Q-PCR primers is shown by the black arrows (not to scale). In this schematic, exons are represented by grey boxes, introns are represented by black lines and untranslated regions are represented by black rectangles. (Fwd = forward; rev = reverse.)

Gene name	Gene symbol	P value	FC
ATP-binding cassette, sub-family C (CFTR/MRP), member 4	*ABCC4*	5.70E-08	-1.82
colony stimulating factor 1 (macrophage)	*CSF1*	2.81E-07	1.40
kinesin light chain 2	*KLC2*	3.73E-07	1.25
KIAA1539	*KIAA1539*	3.95E-07	1.59
tumor necrosis factor receptor superfamily, member 12A	*TNFRSF12A*	4.91E-07	2.37
RAB15, member RAS onocogene family	*RAB15*	1.43E-06	1.13
methyl CpG binding protein 2 (Rett syndrome)	***MECP2***	**1.53E-06**	**-1.78**
limb region 1 homolog (mouse)-like	*LMBR1L*	1.60E-06	1.90
inositol polyphosphate-5-phosphatase	***INPP5A***	**1.77E-06**	**-1.94**
ciliary neurotrophic factor receptor	***CNTFR***	**2.29E-06**	**1.56**
family with sequence similarity 198, member B	*FAM198B*	2.43E-06	1.79
TIMP metallopeptidase inhibitor 3	*TIMP3*	2.70E-06	-1.29
NEL-like 2 (chicken)	*NELL2*	2.85E-06	1.42
carnitine palmitoyltransferase 1C	*CPT1C*	3.33E-06	1.17
mediator complex subunit 25	*MED25*	3.77E-06	1.27
FYVE, RhoGEF and PH domain containing 1	***FGD1***	**4.43E-06**	**-1.38**
discs, large homolog 4 (Drosophila)	*DLG4*	4.64E-06	1.51
synaptotagmin IX	*SYT9*	4.85E-06	1.39
calcium channel, voltage-dependent, T type, alpha 1G subunit	*CACNA1G*	7.10E-06	1.55
motile sperm domain containing 2	*MOSPD2*	7.38E-06	-1.82
transmembrane protein 135	*TMEM135*	8.23E-06	-1.49
phosphate cytidylyltransferase 1, choline, beta	*PCYT1B*	8.40E-06	-1.76
placental growth factor	***PGF***	**1.04E-05**	**1.78**
5'-nucleotidase, ecto (CD73)	*NT5E*	1.05E-05	2.27
damage-specific DNA binding protein 2, 48kDa	*DDB2*	1.22E-05	1.50
CD68 molecule	*CD68*	1.24E-05	2.27
LIM homeobox 8	***LHX8***	**1.28E-05**	**1.39**
tissue factor pathway inhibitor (lipoprotein-associated coagulation inhibitor)	***TFPI***	**1.31E-05**	**1.84**
phosphatidic acid phosphatase type 2B	*PPAP2B*	1.41E-05	-1.18
clusterin	*CLU*	1.42E-05	1.50
insulin-like growth factor binding protein 4	*IGFBP4*	1.43E-05	1.27
annexin A1	*ANXA1*	1.48E-05	4.82
solute carrier family 7 (amino acid transporter light chain, L system), member 5	***SLC7A5***	**1.68E-05**	**-1.51**
transmembrane protein 132A	*TMEM132A*	1.79E-05	-1.47
WD repeat domain 18	*WDR18*	1.83E-05	1.10
roundabout, axon guidance receptor, homolog 1 (Drosophila)	***ROBO1***	**1.84E-05**	**-2.10**
glucuronic acid epimerase	*GLCE*	1.87E-05	1.68
enolase 3 (beta, muscle)	*ENO3*	1.88E-05	1.84
GrpE-like 2, mitochondrial (E. coli)	*GRPEL2*	1.90E-05	-1.38
toll-like receptor 4	*TLR4*	1.92E-05	1.97
pre-B-cell leukemia homeobox interacting protein 1	*PBXIP1*	1.96E-05	1.89
extracellular matrix protein 1	*ECM1*	1.98E-05	1.84
cathepsin A	*CTSA*	2.05E-05	1.87
secretogranin II	*SCG2*	2.05E-05	1.84
vesicle amine transport protein 1 homolog (T. californica)-like	*VAT1L*	2.06E-05	2.05
sestrin 3	***SESN3***	**2.11E-05**	**1.97**
v-raf murine sarcoma 3611 viral oncogene homolog	*ARAF*	2.12E-05	1.11

Appendix 5.2: The list of genes which showed differential expression after *GAPDH* knockdown relative to mock samples

The .CEL files were imported into the PGS using the core metaprobe set and RMA normalisation. The .CEL files were assigned as either siGAP-treatment or mock. A one-way ANOVA was performed on the gene-summarised expression values, using treatment as the candidate variable in the ANOVA model. The expression of 47 genes differed significantly (FDR 0.01). The ten genes in bold were also identified in the siZNF-treated versus siGAP-treated analysis with the same direction of fold change with respect to *GAPDH*.



Appendix 5.3: Manual annotation of the alternative splicing events in *ZNF804A*-depleted cells. The geneviews of the 116 genes that showed alternative splicing after *ZNF804A* knockdown were examined manually alongside the UCSC genome browser. **(A)** The statistically significant calls were attributed to the categories described. **(B)** The 63 events which corresponded with changes in exon usage (with and without changes in transcript abundance) were further categorised to establish the nature of the change.

Gene name	Gene Symbol	Gene expression P value	Bonferroni corrected alternative splicing P value	FC
fms-related tyrosine kinase 1	*FLT1*	2.99E-05	2.27E-24	2.42
annexin A1	*ANXA1*	1.48E-05	3.68E-14	4.82
G-protein coupled receptor 162	*GPR162*	4.32E-03	1.69E-13	1.23
La ribonucleoprotein domain family, member 6	*LARP6*	1.50E-01	6.33E-13	-1.08
integrin, alpha 5 (fibronectin receptor, alpha polypeptide)	*ITGA5*	9.18E-05	1.03E-11	1.46
glycoprotein (transmembrane) nmb	*GPNMB*	4.47E-05	2.22E-11	2.82
microtubule-associated protein 4	*MAP4*	2.15E-01	2.86E-11	1.08
toll-like receptor 4	*TLR4*	1.92E-05	1.51E-10	1.97
UDP-GlcNac:betaGal beta-1,3-N-acetylglucosaminyltransferase 4	*B3GNT4*	5.12E-06	4.40E-09	1.34
CD68 molecule	*CD68*	1.24E-05	1.06E-08	2.27
utrophin	*UTRN*	2.41E-01	6.42E-08	1.12
glutamate receptor, ionotrophic, AMPA 3	*GRIA3*	3.11E-04	2.04E-07	2.05
protocadherin 10	*PCDH10*	1.37E-04	2.78E-07	1.86

Appendix 5.4: The list of genes which showed alternative splicing after *GAPDH* knockdown, relative to mock samples

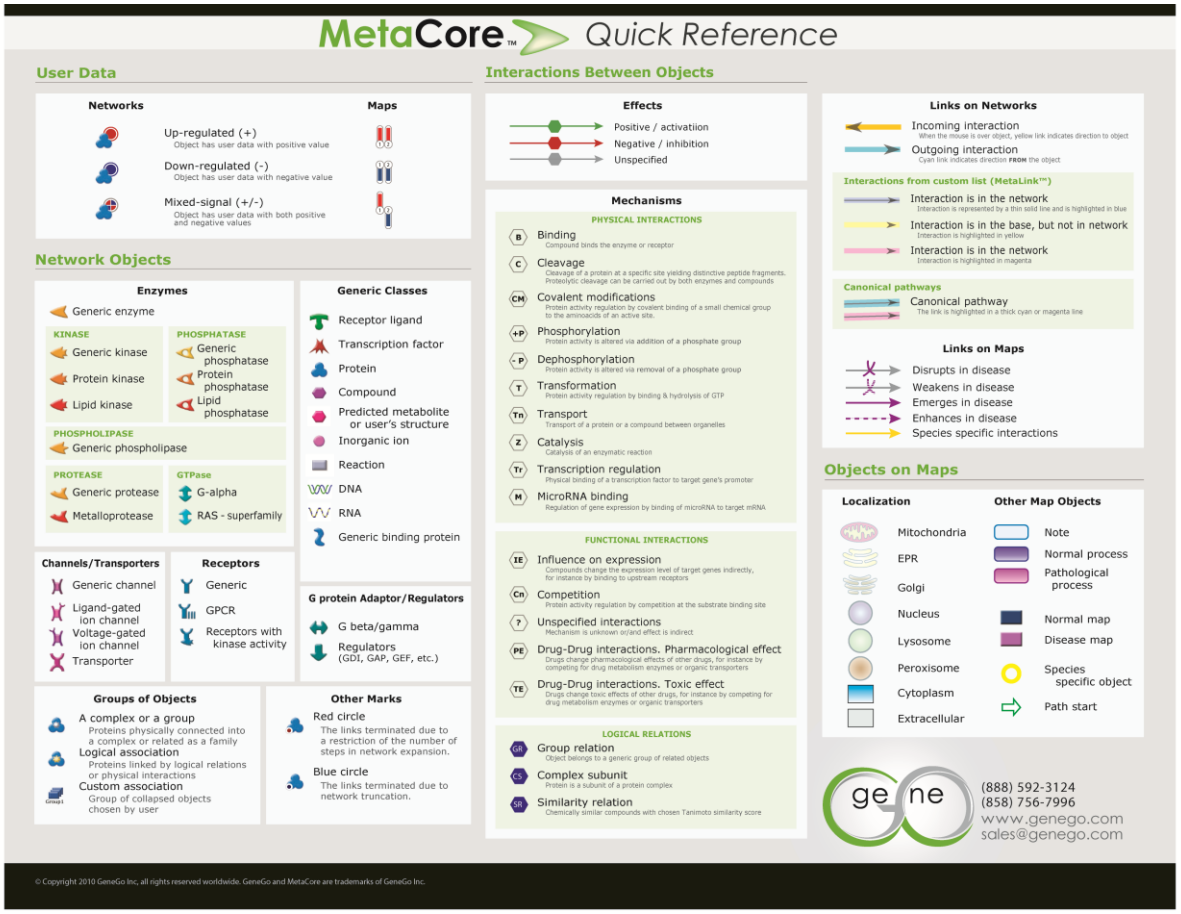
To identify alternatively spliced transcripts, an alternative splicing one-way ANOVA was performed on the probe set $\log_2$ signal intensities greater than three. Treatment was chosen as the candidate variable in the ANOVA model. The gene list was filtered to exclude genes containing fewer than five markers or a fold change greater than five. Alternative splicing ANOVA P values were corrected using the Bonferroni method. To enable a comparison of the number of statistically significant alternative splicing events in *GAPDH*-depleted and *ZNF804A*-depleted cells, a significance threshold of $P < 1 \times 10\text{-}6$ was used. The splicing of 13 genes was altered after *GAPDH* knockdown, relative to mock samples. (FC = fold change.)

Gene name	Gene symbol	RefSeq	pooled siZNFA+B		HSS150612		HSS150613	
			P value	Ratio	P value	Ratio	P value	Ratio
anthrax toxin receptor 1	*ANTXR1*	NM_032208	2.92E-02	0.82	4.23E-02	0.9	3.82E-02	0.9
ankyrin repeat and SOCS box containing 8	*ASB8*	NM_024095	4.22E-02	1.13	8.00E-04	1.22	2.37E-02	1.11
CD151 molecule (Raph blood group)	*CD151*	NM_004357	2.11E-04	1.40	3.15E-02	1.11	2.03E-02	1.06
CD83 molecule	*CD83*	NM_004233	6.38E-04	0.77	1.90E-03	0.75	3.14E-02	0.73
carboxylesterase 2	*CES2*	NR_036684	5.82E-02	1.26	2.44E-02	1.09	4.08E-02	1.11
coiled-coil-helix-coiled-coil-helix domain containing 7	*CHCHD7*	NM_001011667	6.83E-03	0.77	1.70E-02	0.87	1.45E-02	0.86
eukaryotic translation initiation factor 4A2	*EIF4A2*	NM_001967	3.51E-02	1.17	1.70E-03	1.15	5.00E-04	1.26
ELAV-like 1	*ELAVL1*	NM_001419	4.88E-02	0.83	3.94E-02	0.85	3.08E-02	0.88
eva-1 homolog A (C. elegans)	*FAM176A*	NM_001135032	1.86E-02	1.45	2.99E-02	1.28	2.36E-02	1.23
family with sequence similarity 46, member A	*FAM46A*	NM_017633	4.22E-02	0.77	1.03E-02	0.82	8.40E-03	0.82
FH2 domain containing 1	*FHDC1*	NM_033393	3.38E-03	0.58	7.20E-03	0.77	2.90E-03	0.77
monooxygenase, DBH-like 1	*MOXD1*	NM_015529	3.68E-04	0.56	4.45E-02	0.87	3.11E-02	0.86
mannose receptor, C type 2	*MRC2*	NM_006039	1.39E-03	0.64	1.91E-02	0.81	3.95E-02	0.84
organic solute carrier partner 1	*OSCP1*	NM_145047	2.07E-03	0.76	4.92E-02	0.93	4.77E-02	0.84
poly (ADP-ribose) polymerase 2	*PARP2*	NM_005484	4.27E-02	1.11	3.72E-02	1.15	3.14E-02	1.1
protease, serine, 35	*PRSS35*	NM_001170423	9.60E-05	0.84	1.77E-02	0.77	2.73E-02	0.86
quiescin Q6 sulfhydryl oxidase 2	*QSOX2*	NM_181701	3.55E-04	0.59	1.48E-02	0.78	2.46E-02	0.83
RAS, dexamethasone-induced 1	*RASD1*	NM_016084	1.35E-02	0.60	4.35E-02	0.81	1.90E-03	0.75
sperm associated antigen 16	*SPAG16*	NM_024532	2.27E-02	0.65	6.30E-03	0.82	1.80E-03	0.73
ubiquitin protein ligase E3C	*UBE3C*	NM_014671	1.13E-02	0.70	4.82E-02	0.91	2.30E-02	0.9

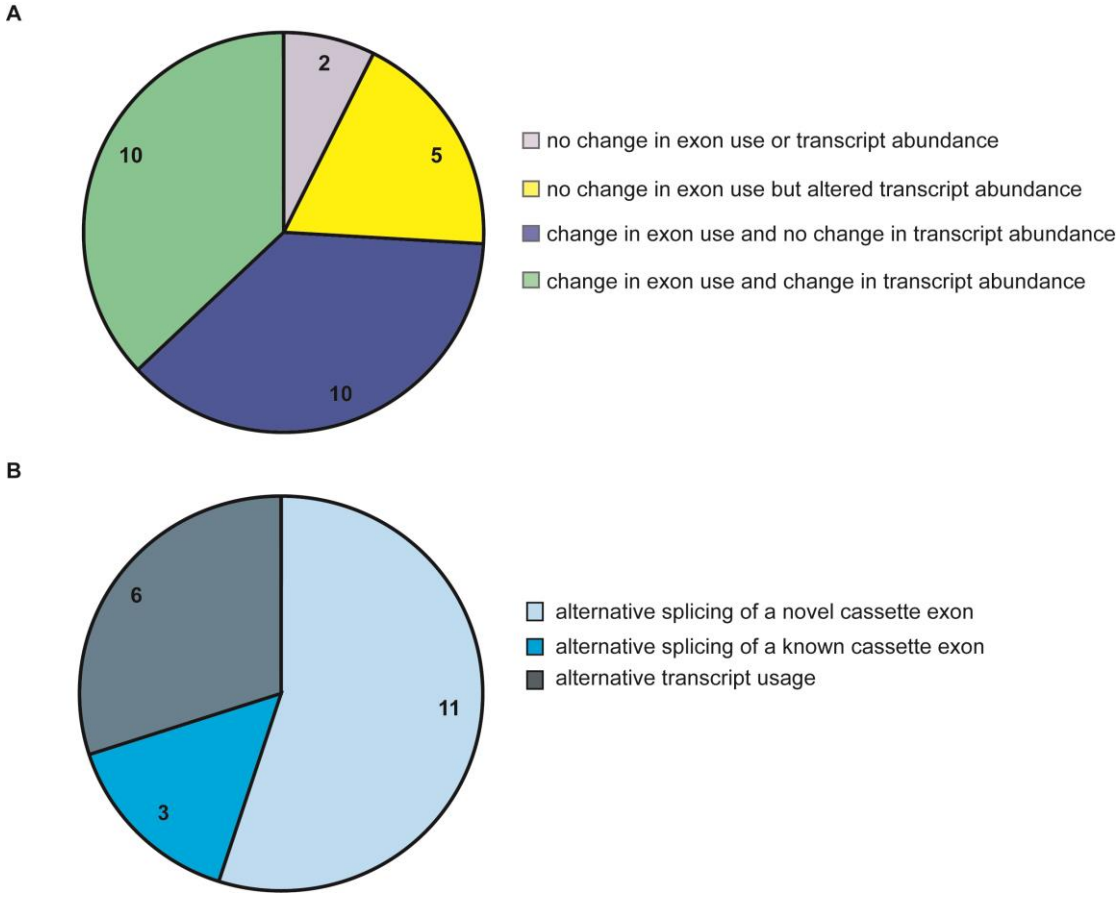
Appendix 5.5 The list of genes identified as nominally differentially expressed both in this study and by Hill and colleagues (2012)

Data presented in Chapter Four show there were numerous changes in gene expression in *ZNF804A*-depleted cells. To determine if these data were consistent with data published by Hill and colleagues (2012), the overlap between the two datasets was considered. The table below outlines the 20 genes which were nominally differentially expressed in the same direction in both of the datasets. HSS150612 and HSS150613 are siRNA duplexes used by Hill and colleagues to deplete *ZNF804A*.

Appendix 6: Supplementary data for Chapter Five



Appendix 6.1 GeneGo MetaCore™ pathway map legend



Appendix 6.2: Manual annotation of the alternative splicing events in cells over-expressing *myc-ZNF804A*. The geneviews of the 27 genes that showed alternative splicing after *myc-ZNF804A* over-expression were examined alongside the UCSC genome browser. **(A)** The statistically significant calls were attributed to the categories described. **(B)** The 20 events which corresponded with changes in exon usage (with and without changes in transcript abundance) were further categorised to establish the nature of the change.