Analysis of Dysbindin Interacting Genes in the Pathogenesis of Schizophrenia

Amy Gerrish

Department of Psychological Medicine

Cardiff University

ProQuest LLC
789 East Eisenhower Parkway
P.O. Box 1346
Ann Arbor, MI 48106-1346

Table of Contents

Acknowledgments

The completion of this thesis was only possible due to the help and support of a large number of people. First and foremost I would like to thank my supervisors Nigel and Lesley for giving me the opportunity to do a PhD within this department and for their advice and encouragement throughout. I must also thank Mick and Mike for their invaluable suggestions during many psychosis meetings.

Special thanks must go to Hywel and Nadine whose help and advice during my year as a technician provided me with a great foundation for this PhD. I know I wouldn't have got this far without their support. A particular mention must also go to Nick who one night in the Social suggested I apply for a PhD. I can only assume he had had one too many.

Thanks to everyone within the Psych Med department and specifically members of the middle office including the current residents Sarah, Didi, Denise, Jade, Evie and Irina who are never short of a sympathetic ear or indeed a bar of chocolate. I would also like to thank Becky and Dobril for their company through "the year of hell".

Huge thanks must also go to everyone of the second floor, in particular Caz, Lyn, Matt and Ade. Parts of this thesis just wouldn't have happened without their gallant attempts to teach functional biology to a geneticist. A special mention must also go to Valentina for her invaluable help with all the statistics.

To my friends and family, thanks for putting up with me the last few years and pretending to understand what I am on about. Special mention must go to Maz and Katy, who were always on hand with a bottle of wine or two when the going got tough, and to my parents for making me believe I could achieve anything I set my mind to, then keeping quiet when I chose science. Last but definitely not least I would like to thank Andrew for his love and support, not to mention cooking, over the last four years and whose words "I don't think it is meant to be easy" got me through even the most difficult days.

Summary

Dysbindin (dystrobrevin-binding protein 1; DTNBP1) has been implicated as a schizophrenia candidate gene. However while numerous positive associations have been reported, no non-synonymous alleles have been found which account for the association. A number of recent studies suggest that altered dysbindin expression may be the mechanism by which DTNBP1 variants confer susceptibility to schizophrenia. Therefore one objective of this study was to identify putative DTNBP1 cis-acting variants and perform association analyses of these variants with schizophrenia and allelic expression differences observed at the DTNBP1 locus.

While four variants were associated with schizophrenia, logistic regression suggested that the signal observed at these polymorphisms was not independent of the most associated SNP rs4715984. Comparison with the results of a DTNBP1 allelic expression assay revealed that seven SNPs were associated with differential expression. *Post hoc* analysis revealed that the majority of the expression differences were accounted for by variation at two loci (rs2619538 and rs13198512), one of which (rs13198512) was subsequently shown to directly affect transcription *in vitro* using a luciferase reporter gene assay. As rs4715984 was not correlated with allelic expression differences it implies that a reduction in dysbindin expression through cis-acting variation may not be the primary aetiological factor in schizophrenia pathogenesis. This was supported by further analysis of a schizophrenia risk haplotype previously reported to be associated with differential expression as the refined haplotype was no longer correlated.

A second objective of this thesis was to investigate the hypothesis that DTNBP1 could cause susceptibility to schizophrenia through its role within the BLOC-1 complex. Association analysis was performed on BLOC-1 genes which displayed evidence of being under the influence of cis-acting regulation. MUTED, a BLOC-1 gene previously reported as associated to schizophrenia was also investigated. However association results provided no compelling support for the hypothesis that DTNBP1 contributes to susceptibility to schizophrenia through the BLOC-1 complex.

Associated Publications

Gerrish, A., et al., *An examination of MUTED as a schizophrenia susceptibility gene.* Schizophrenia Research, 2009. **107**(1): p. 110-1.

Chapter 1: Introduction

1.1 Schizophrenia

Schizophrenia is a severe psychiatric disorder with a lifetime risk of ~1% [1]. The disease is characterised by psychotic symptoms, apathy, altered emotional reactivity and disorganised behaviour [2]. Although subtle cognitive and behavioural symptoms are present from early childhood, onset of the more distinguishing features is commonly in the late teens and early twenties. These symptoms often result in impaired functioning in work, school, parenting, independent living and interpersonal relationships.

Although outcomes are variable, the typical course of schizophrenia is one of relapses followed by partial remission. Even with treatment, only 20%-40% of patients diagnosed with schizophrenia or schizophrenia spectrum disorder (SSD) have been found to show a substantial clinical improvement after 5-6 years of treatment [3].

Due to the pervasiveness of associated deficits and often life-long course, schizophrenia is among the top ten leading causes of disease related disability in the world [4]. Sufferers are often the most vulnerable, isolated and disadvantaged individuals in society [5]. In addition to the core symptoms which result in impaired functioning, schizophrenic patients are also at an increased risk of alcohol and drug problems, violent victimisation, post traumatic stress disorder, housing instability and homelessness, smoking-related illness' and depression. The net result of exposure to these risks is an increased mortality due to suicide (estimated at 5%), accidents, or illness' such as respiratory and cardiovascular diseases [6, 7].

In addition to the suffering caused to patients and their families, schizophrenia also has an enormous economic impact. In 1996 the cost of schizophrenia to the UK economy

was estimated at £2.6 billion [8] and the combined economic and social costs of schizophrenia have been estimated to account for 2.3% of all burdens within developed countries [9].

At present schizophrenia cannot clearly be defined at a biological level and the disorder is inadequately treated by current therapies. Understanding schizophrenia from a genetic viewpoint will hopefully lead to a full understanding of the biological pathways involved in schizophrenia aetiology and as a result, successful treatment. The purpose of the following study was therefore to identify and understand genetic variants that influence schizophrenia susceptibility.

1.1.1 History of Schizophrenia

The condition now referred to as schizophrenia was first characterised in 1893 by the German psychiatrist Emil Kraepelin, under the name dementia praecox [10]. This disorder was established by grouping together several previously described syndromes which included hebephrenia, catatonia and paranoid dementia. The distinguishing features of dementia praecox were two fold. Firstly, Kraepelin noted that, unlike other dementia such as Alzheimer's disease which occurred later in life, dementia praecox usually began in late teens or early adulthood. Secondly, Kraepelin described dementia praecox as a deteriorating psychotic disorder whose primary disturbance was not one of mood but of cognition [11].

In addition to dementia praecox, Kraepelin also characterised the disorder manic depressive illness. By reclassifying the majority of the previously described insanities as either dementia praecox or manic depressive illness, Kraepelin introduced considerable order to the previously confused field of psychiatric disease classification. However there were a number of objections to Kraepelin's categorisations including his belief that dementia praecox was a form of dementia with a deteriorating course and no chance of recovery. To this end, dementia praecox was renamed by a Swiss psychiatrist Eugen Bleuler at the beginning of the 20th century. Bleuler realised that the illness was

not a type of dementia as some patients did improve rather than continue to deteriorate as Kraepelin had suggested. In light of this, and other observations, Bleuler coined the term schizophrenia, from the Greek schizen 'to split' and phren 'mind', to emphasise the fact that the disorder produces a severe fragmentation of thinking and personality. Although it is now known that split personality disorder is actually a separate and relatively rare disease (dissociative disorder), the term schizophrenia has remained.

1.1.2 Symptoms of Schizophrenia

Schizophrenia is characterised by a diverse set of signs and symptoms which arise from almost all domains of brain function including language, emotion, reasoning, motor activity and perception [12]. Although highly varied, the majority of symptoms displayed by schizophrenics fall into three broad categories; negative symptoms – the absence of certain functions or aspects of the mind that should be present in a normal individual, positive symptoms – the presence of certain features which are not present in a normal individual and cognitive impairment [7, 13]. Negative symptoms include social withdrawal, disorganisation and reduced will. Positive symptoms comprise of disorders of perception (hallucinations), inferential thinking (delusions) plus involuntary movements or actions (catatonia) [2, 13]. Symptoms of cognitive impairment include deficits in attention, intelligence, memory and executive function [14].

1.1.3 Classification of the Phenotype

At present there is no objective diagnostic test or easily measurable biomarker available to provide a secure diagnosis of schizophrenia [15]. As a result diagnosis is based solely upon the symptoms presented and reported and consequently on detailed clinical observation.

Cross-national studies at the beginning of the 20th century suggested that, due to the complexity of schizophrenia, there were international differences in the breadth and

style of schizophrenia diagnosis [16, 17]. In response to this the World Heath Organisation and the American Psychiatric Association produced criterion-based systems for the diagnosis of schizophrenia. Currently under the tenth and fourth revisions respectively, the International Classification of Disease (ICD) and the American Psychiatric Association's Diagnostic and Statistical Manual (DSM) objectively define the symptoms and characteristic impairments of schizophrenia in a relatively similar way [2]. The major differences between each approach consist of the DSM-IV requirements for social or occupational dysfunction (not included in ICD-10) and a six month duration of illness (vs. one month duration in ICD-10). This results in a somewhat narrower definition of the disorder in DSM-IV. Nevertheless consistency between the two systems is high [18] and the introduction of both has been shown to substantially improve the reliability of diagnosis [2].

1.1.4 Epidemiology of Schizophrenia

The annual incidence of schizophrenia has recently been estimated at between 8 and 43 per 100,000 individuals [19]. This level of incidence is relatively similar across a wide range of cultures and countries including developed and developing nations [15, 20]. The lifetime relative risk of developing schizophrenia ranges from 0.3-2% with an average of ~0.7% [21].

Previously it has generally been believed that the risk of developing schizophrenia over a lifetime is similar among males and females [22]. However gender differences in the clinical expression and outcome of schizophrenia have long been recognised [23] with women tending to have a later age of onset than men [24] and a more benign course of illness including fewer hospital admissions and increased social functioning [25]. Two recent meta-analyses have supported this observation and revealed a higher lifetime risk in males (relative risk male: female~1.4) [19, 26].

1.1.4.1 Environmental Risk Factors

A proportion of the risk of developing schizophrenia is attributable to environmental factors [3, 7, 27]. Substantial evidence has been reported, although mainly in western countries, which shows individuals born or brought up in an urban environment are more likely to develop schizophrenia than those in rural areas [28]. Migration has also been associated with an increased risk of schizophrenia. Other environmental risk factors include pre and perinatal events such as prenatal infections, famine during pregnancy, obstetric and perinatal complications and winter/spring births. More general environmental risk factors include lower social class, social stress, low IQ score and cannabis abuse [3, 29].

1.1.4.2 Genetic Risk Factors

Although several environmental risk factors have been implicated in schizophrenia, no one factor appears either necessary or sufficient to cause the disorder. In contrast, there is now overwhelming evidence that the risk of developing schizophrenia is largely attributable to genetic risk factors. This evidence comes in number of forms including family, twin and adoption studies.

By analysing over 40 family studies, Gottesman and Shields were able to show that schizophrenia runs in families [1]. In addition, it was observed that the risk of developing schizophrenia is increased in the relatives of affected individuals. The lifetime risk of developing schizophrenia in the general population is ~1%. This contrasts with the average risk to siblings (9%) and offspring (13%) of affected probands. Furthermore an identical (monozygotic) twin who shares 100% of their genes with a proband has a risk of almost 50% (Figure 1.1).

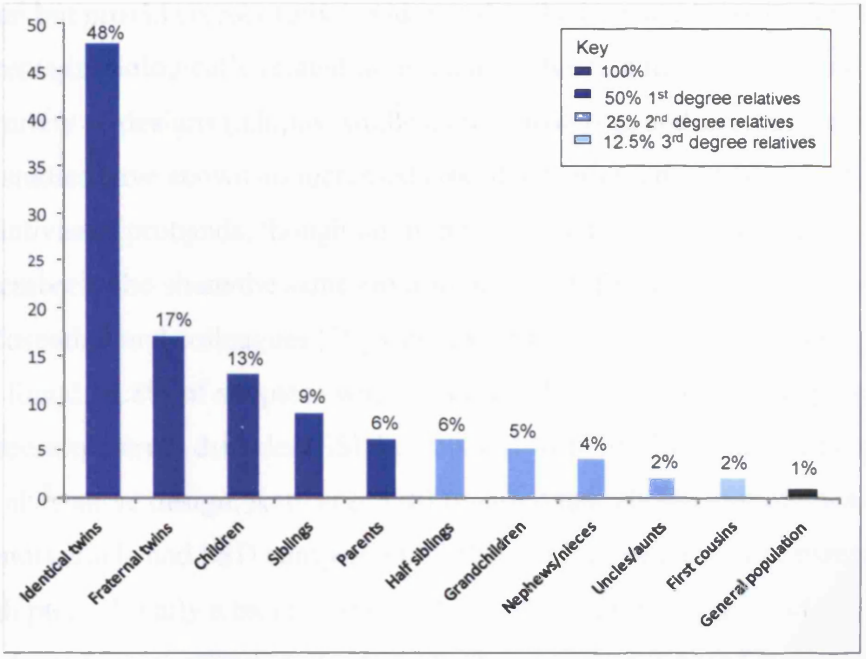


Figure 1.1. Risk of developing schizophrenia, for relatives of schizophrenia probands compared to the general population. The percentages indicated in the key refer to the proportion of genes shared. Data based on reviews by Gottesman [1].

Although family studies suggest that genetic factors may play a role in the aetiology of schizophrenia, they do not provide any evidence about the relative contribution of these factors. Twin studies can be used to determine whether familial clustering is due to genetic factors or the result of a shared environment. This class of family study compares the concordance rates of a disease between members of monozygotic (MZ) twin pairs who are genetically 100% identical and dizygotic (DZ) twins who share only 50% of their genes. A greater concordance rate between MZ than DZ twins reflects a genetic influence. Conversely, if a disease is caused by environmental factors then the rate of concordance between MZ and DZ twins would be the same [30]. Cardno and Gottesman [31] combined data from five relatively recent twin studies [32-36] and found that the average concordance rate between MZ twins was 50% compared to 1% in DZ twins using DSM-IIIR criteria.

One criticism of twin studies is that they assume that both types of twins share environmental influences to the same extent. Adoption studies do not rely on this

assumption but provid corroborative evidence that the increased risk of developing schizophrenia in biologically related individuals is due to a major genetic influence. Using a variety of designs (adoptee studies, cross-fostering and adoptee families) adoption studies have shown an increased risk of schizophrenia in biological first degree relatives of probands, though not in non-biologically related adopted or adoptive family members who share the same environment as probands [37-40]. These studies include Rosenthal and colleagues [39] who, examining subjects from Danish adoption registers, found 18.8% of adoptees whose biological parents were schizophrenic had schizophrenia spectrum disorder (SSD) compared with 10.1% of matched controls. Using an alternative design, Kety *et al* [38] reported that 20.3% of biological parents of affected individuals had SSD compared to 5.8% of adoptive parents or parents of control adoptees. Finally a more recent study showed a prevalence of schizophrenia of 9.4% in adopted away offspring of schizophrenic parents compared to a lifetime prevalence of 1.2% in control adoptees [41].

1.1.5 Neurobiology of Schizophrenia

The clinical features of schizophrenia described in section 1.1.2 are highly suggestive of a neurobiological basis of the disorder. However while a large amount of literature has been published investigating the neurobiology of schizophrenia, the precise nature of schizophrenia pathogenesis is still unclear [42]. Nevertheless over the past two decades a number of advances have been made in this field, the majority of which have been achieved due to developments in neuroimaging, electrophysiological and neuropathological technologies [42].

One notable and consistent finding in schizophrenia is the absence of cortical gliosis [43]. Indeed a number of studies have observed a reduction in glial density in schizophrenic patients [43]. Gliosis, which consists of an increase in glial cells within damaged areas of the CNS, is seen in various neurodegenerative disorders such as Alzheimer's disease [44]. The lack of gliosis in schizophrenia indicates that the cognitive impairments observed are not caused by a neurodegenerative process. This

conclusion is supported by studies which report little evidence of large scale neuronal loss in schizophrenic brains.

In the absence of degenerative changes, the attention of researchers has focused on alterations in the cytoarchitecture of neurons within the schizophrenia brain. Analysis of synapse and dendrite morphology has produced converging lines of evidence which suggest that disrupted cortical synaptic circuitry is a central deficit in schizophrenia [43, 45, 46]. MRI studies reveal subtle reductions in grey matter volume in the dorsolateral prefrontal cortex (DPFC) and hippocampus in subjects with schizophrenia [43, 45, 47]. While grey matter is composed primarily of neuropil and cell bodies, post-mortem studies suggests it is a decrease in neuropil that causes the reduction in grey matter as they show a decreased dendrite density along with normal or increased neurone density [43].

In addition to grey matter reductions a number of studies have reported a decrease in white matter in schizophrenic brains [42, 45, 46]. ^{1}H magnetic resonance spectroscopy (MRS) studies support this finding as they have identified lower N-acetylaspartate (NAA) levels in temporal and frontal cortex of patients with schizophrenia [48]. NAA is a putative measure of neuronal viability which includes the formation and maintenance of myelin. Furthermore a reduction in pre-synaptic markers such as synaptophysin indicates schizophrenics have a reduced number of axon terminals compared to healthy controls. This is supported indirectly by ^{31}P MRS which shows a decrease in phosphomonoester levels within the prefrontal and temporal cortices [43]. This decrease indicates a reduction in the synthesis of membrane phospholipids and consequently a decrease in the number of synapses.

If schizophrenia is a disorder of disrupted synaptic circuitry the underlying mechanism of susceptibility could be a malfunction in neurotransmitter chemistry. To date there have been a number of neurochemical theories which are based primarily on compounds that ameliorate or induce schizophrenia-like symptoms. These drugs have been shown to act on a number of neuronal receptors and affect neurotransmitter

release. The most prominent of these theories are the dopamine and glutamate hypotheses; although other theories involving neurotransmitters such as GABA, serotonin and acetylcholine have also been postulated.

The dopamine hypothesis of schizophrenia remains the most studied neurochemical theory relating to the disorder [49]. The classical dopamine hypothesis suggests that dopamine neurotransmission is hyperactive in schizophrenics [50] and is based on the indirect evidence that some anti-psychotic drugs block central dopamine receptors while dopamine agonists such as amphetamine produce a psychomimetic effect [51]. Even though this theory has been studied in depth, direct evidence for the dopamine hypothesis is still sparse. Furthermore it has been difficult to separate any reported observations, such as an increase in D2 receptors in schizophrenic patients, from the effect of prior antipsychotic drug treatment. Another limitation of the classical dopamine hypothesis is that its explanatory power is stronger for positive symptoms than other aspects of the disease such as cognitive impairment which has a greater lifetime prevalence than most positive symptoms and often better predicts a patient's outcome [42].

The abnormal cortical connectivity and functioning suspected to occur in schizophrenia automatically implicates another neurotransmitter, glutamate, due to the fact that majority of excitatory synapses in the brain are glutamatergic. Experimental evidence supports this association and suggests that schizophrenia may be related to deficient glutamate mediated excitatory neurotransmission via N-methyl-D-aspartic acid (NMDA) receptors. NMDA antagonists such as phencyclidine (PCP) and ketamine have been shown to trigger a number of schizophrenia-like symptoms including positive and negative symptoms along with cognitive deficits [51]. Furthermore post mortem studies of schizophrenia patients have reported, albeit not consistently, reduced expression of glutamate receptors, in particular NMDA receptors, in a variety of brain regions including the prefrontal cortex (PFC) and hippocampus [51].

Another neurotransmitter implicated in schizophrenia is gamma amino butyric acid (GABA). There is now substantial evidence that abnormalities in cortical and limbic GABA neuronal populations are a feature of the neurobiology of schizophrenia [49, 52, 53]. In addition, mRNA levels of glutamate decarboxylase, the major determinant of GABA synthesis have consistently been shown by post mortem studies to be reduced in schizophrenia patients, thus suggesting reduced GABA levels [54].

Although the majority of evidence is indirect, serotonergic and acetylcholinergic neurotransmission have also been suggested to be disrupted in schizophrenics. Serotonin antagonists such as clozapine and risperidone have been shown to have therapeutic benefits for schizophrenia cases [42, 51]. Additionally it has been suggested that the high prevalence of cigarette smoking in schizophrenics could be a compensatory response to a deficit in nicotinic cholinergic receptors [42, 51].

In summary, rather than having a distinctive characteristic, the neuropathology of schizophrenia appears to consist of a number of subtle changes that indicate an alteration in neuronal circuitry. The increased advancement of imaging and neuropathological technologies will hopefully provide more direct and compelling evidence of the specific abnormalities in schizophrenia which will in turn provide a clearer picture of its neuropathology. However one question that will still need resolving is whether a given abnormality represents a primary pathogenic event (cause), a secondary deleterious event (consequence) or a homeostatic response intended to restore normal brain function (compensation) [45]. As susceptibility genes are thought to represent a primary pathological event [46], the identification of schizophrenia susceptibility genes is likely to be crucial to achieve a complete understanding of schizophrenia pathology.

1.1.6 Modes of Inheritance

1.1.6.1. Polygenic Model

Evidence from family, twin and adoption studies has conclusively shown that genetic factors play a significant role in the aetiology of schizophrenia and the heritability of schizophrenia has been estimated to be >80% [55, 56]. However the almost exponential decline in relative risk observed as genetic distance from the proband increases (Figure 1.1) is incompatible with the hypothesis that schizophrenia is a single gene disorder or a collection of single gene disorders, even when incomplete penetrance is taken into account [55]. This observation is supported by Risch [57] who demonstrated that the relative risk data observed for schizophrenia is not compatible with one locus with a relative risk (λs) value of greater than 3.

Taken together, these findings suggest that the mode of inheritance for schizophrenia, like that of other common disorders such as Alzheimer's disease, bipolar disorder and diabetes, is non-Mendelian, complex and involves more than one locus [58]. Models of genetic susceptibility which consist of more than one susceptibility locus include oligogenic, a few genes of moderate effect and polygenic, many genes of weak effect.

The frequency of discordance between MZ twins (~50%) suggests that inheritance of a particular genotype does not confer a certainty of developing schizophrenia but rather a susceptibility to the disorder. This inherited susceptibility can be explained by a polygenic model, also known as a multifactorial liability threshold (MLT) model [59, 60]. In this situation the risk or liability of developing schizophrenia follows a continuous (normal) distribution in the general population (Figure 1.2a). Each susceptibility gene exerts a small effect and combines with other genetic and environmental risk factors in a predominantly additive manner. When the number of risk factors an individual possesses exceeds a critical threshold, the individual develops schizophrenia (or is diagnosed with the disorder). Furthermore an increase in risk factors past this threshold results in an increase in severity of the disease. The threshold

for schizophrenia is determined by the lifetime morbid risk of developing the disorder; therefore ~1% of the population will achieve the threshold for schizophrenia diagnosis.

As affected individuals possess a higher proportion of susceptibility genes than unaffected individuals, relatives of a proband have a higher probability of inheriting susceptibility alleles than the general population and therefore their mean risk increases, as shown in Figure 1.2.b. As the critical threshold level remains unchanged this shift means a greater proportion of siblings (~10%) exceed the threshold for schizophrenia diagnosis. This explains why the incidence of schizophrenia is higher in relatives of schizophrenics than in the general population and that the risk increases as the degree of genetic relatedness also increases as shown in Figure 1.1.

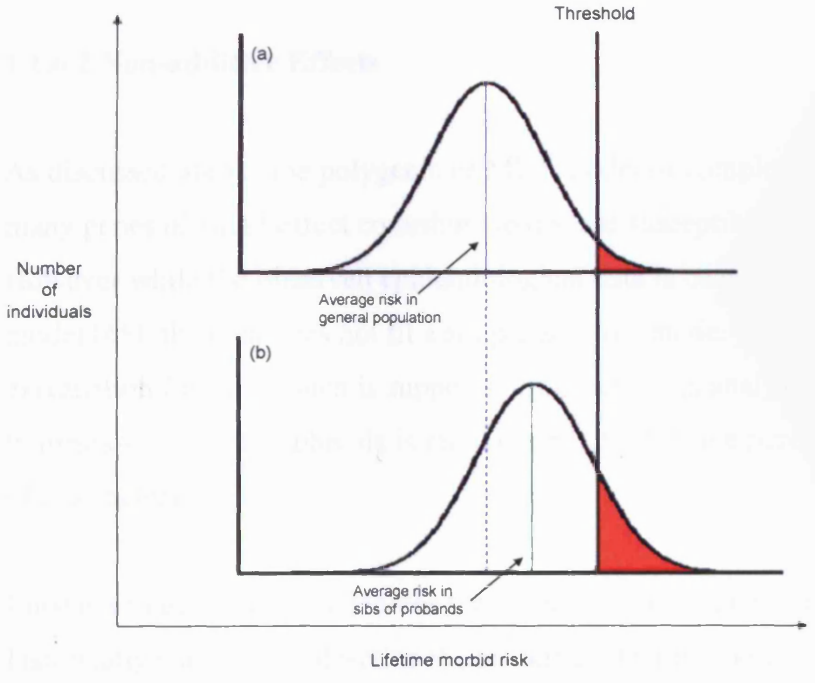


Figure 1.2. The multifactorial-liability threshold model for schizophrenia. The disorder has a continuous distribution in the population due to multiple genetic and environmental risk factors contributing to the risk of developing the disease. The lifetime morbidity risk of schizophrenia in the general population is given in a) where 1% of individuals (shown in red) have a combination of risk factors which produce a schizophrenia diagnosis. Unaffected individuals (shown in white) are under this threshold. The lifetime morbidity risk of siblings of schizophrenics is shown in b). As these individuals have a higher probability of inheriting susceptibility genes, their average liability is increased and as a result a greater proportion exceeds the threshold for schizophrenia diagnosis. Adapted from [61, 62].

In addition to an increased risk to those related to schizophrenic individuals, the MLT model can also explain why the risk for schizophrenia in an individual, increases with the number of affected relatives [56]. For example an individual with two affected siblings is more likely to develop schizophrenia than an individual with one affected and one unaffected sibling. This is because the parents of these offspring are likely to be closer to the critical threshold than parents of one proband and therefore are more liable to pass on risk genes to additional offspring. In addition, more severely affected schizophrenic individuals are more likely to have schizophrenic offspring due to the high number of risk factors they possess. Although a polygenic inheritance model implies multiple genes of weak affect, it should also be noted that even under this model, alleles of large effect are expected to occur by chance. Therefore the model can also account for families with a high density of illness whose susceptibility may be caused by alleles of large effect [63, 64].

1.1.6.2 Non-additive Effects

As discussed above, the polygenic or MLT model of complex diseases assumes that many genes of small effect contribute to disease susceptibility in an additive manner. However while the observed epidemiological data is best accounted for by the MLT model [65], the data does not fit a simple additive model particularly well [56]. One explanation for this, which is supported by recent twin analysis [33], is that the mode of transmission of schizophrenia is more complicated than a purely additive combination of risk factors [56].

Epistasis in general terms denotes the interaction between genes or gene products and historically was used to describe the masking effect of one genotype on another at separate loci. However epistasis in terms of genetic studies refers to two or more genes, each with their own susceptibility risk, acting in a multiplicative manner, whereby the total liability from n genes is greater than the sum of their individual susceptibilities [66].

Accounting for possible epistatic effects is particularly challenging in complex genetics. Nonetheless, a number of methods have been developed to test for epistasis within association data. These methods use logistic regression to fit multiple genotypes to affection status [67]. A successful example of this approach identified epistatic interactions between OLIG2 and CNP plus OLIG2 and ERBB4 [68]. These were independent of the main effects shown by each of the loci. Biological plausibility for these interactions was subsequently indicated by a correlation of the expression of these genes [68].

In addition to the gene-gene interactions involved in epistasis, gene-environmental interactions could also affect schizophrenia risk. A number of studies have been published which have investigated interactions between environmental risk factors and genetic susceptibility variants [69-73]. This includes the potential interaction between the COMT Val^{158}Met polymorphism, cannabis use and the development of psychosis [69]. However, at this time there are no gene-environment interactions which are unambiguously associated with schizophrenia.

In general although some advances have been made in hypothesising the mode of inheritance of schizophrenia; the number of susceptibility genes involved, the disease risk conferred by each locus, the extent of genetic heterogeneity and the degree of interaction between loci all remain unknown [63].

1.2 Identifying Complex Disease Genes

As the risk of developing schizophrenia is largely attributable to genetic risk factors a logical next step is to attempt to identify genes which confer susceptibility to schizophrenia. This in turn will hopefully lead to an understanding of the biological pathways involved in schizophrenia aetiology and as a result, successful treatment.

The two main methods employed to locate disease susceptibility genes are linkage and association analysis. Both methods depend on the existence of polymorphic genetic markers but do not require any prior knowledge of the pathogenesis of the disease [56]. Linkage analysis can detect susceptibility genes over large distances; however it is typically limited to detecting those of larger effect. In contrast, association analysis can detect genes of more moderate/minor effect although until recently with the advent of genome-wide association studies, only relatively small regions could be analysed at one time [56, 61, 62].

1.2.1 Linkage Analysis

The concept of linkage analysis, which has been particularly successful at detecting genes that cause Mendelian traits, is based on the phenomenon known as genetic linkage. Before cell division takes place during gametogenesis, homologous chromosomes pair up within a diploid cell. During meiosis these homologous pairs exchange genetic material in a process known as recombination which occurs at crossover points or chiasma. Therefore recombination produces new hybrid chromosomes where some loci originated from the maternal chromosome and some from the paternal. The process of recombination generates increased diversity among the human species and restructures genes and their alleles [56, 61, 62].

If two loci are on different chromosomes the probability that their alleles will be inherited together is 0.5. This is the phenomenon Mendel described as independent assortment. As a crossover point within a chromosome appears to occur in a random

fashion this is also true for loci far apart on the same chromosome. However the nearer two genes are to one another the less likely it is that a crossover point will occur between them and the more likely that they will be transmitted together. Therefore the probability that they will be inherited together is greater than 0.5. This represents a departure from the law of independent assortment and is known as genetic linkage.

Linkage analysis aims to establish if a genetic marker co-segregates with a phenotype within many independent families or over many generations of an extended pedigree. Although the marker itself may not be causing the disease/phenotype, the phenomenon of genetic linkage indicates that a susceptibility locus causing the phenotype resides in the same chromosomal region as the segregating marker.

The traditional approach to calculating the statistical evidence for linkage is the LOD score [74]. A LOD score is a logarithm of the odds ratio of the likelihood that the observed co-segregation of marker and illness is due to linkage, against the likelihood that the observed co-segregation occurs by chance. Morton suggested that a cumulative LOD score exceeding 3 should be regarded as good evidence for linkage, while a cumulative LOD score below –2 should be regarded as strong evidence against linkage [74]. A LOD score of 3 represents an odds ratio of 1000:1 in favour of linkage and corresponds to a p value of 0.05. This means that the observed data is 10^3–fold more likely to arise under the specified hypothesis of linkage than under the null hypothesis of no linkage [75].

1.2.1.1 Parametric and Non-parametric Analysis

Classic linkage analysis, known as parametric analysis, follows the cosegregation of markers and disease over a number of generations in large multiplex families. It involves the specification of a genetic model and has proved to be a very powerful method for detecting loci segregating in a Mendelian fashion [61]. However for complex diseases such as schizophrenia, where the mode of transmission is unknown, parametric linkage analysis is more challenging. As a result non-parametric analysis, which does not require a mode of inheritance to be defined, is often used. The most popular design for non-parametric analysis is the affected sib pairs (ASP) method. Unlike parametric analysis, which compares affected and unaffected family members, ASP examines allele sharing in affected siblings only. This is particularly suited for the analysis of complex diseases where unaffected siblings may be non-penetrant carriers of the susceptibility allele. Allele sharing can be defined in two ways, identity-by-state (IBS) or identity-by descent (IBD). Two alleles are characterised as IBS if they both have the same DNA sequence at the polymorphic site. If these two alleles are also both descended from a recent common ancestor then they are said to be IBD. IBD is regarded as a superior measure of allele sharing as it is more informative and less dependant on the exact marker allele frequency being known [61].

1.2.1.2 Limitations of Linkage Analysis

Linkage analysis can be a powerful and robust method of identifying genes that cause Mendelian disorders. However for most common diseases linkage analysis has achieved only limited success [76, 77]. This lack of success has been attributed to a number of factors. Firstly linkage analysis has been found to be less powerful at identifying common genetic variants that have a modest effect size [78]. In addition the standard set of microsatellites used in linkage analysis, which are spaced ~10cM apart, are unlikely to extract complete inheritance information [77]. While these confounders could potentially be resolved by using a denser set of markers, larger sample sizes and larger pedigrees, Hirschhorn and Daly [77] note that the Pro12Ala variant in the

PPARG gene which affects the risk of type 2 diabetes would only be detected using linkage analysis of over 1 million affected sib pairs.

1.2.2. Association Studies

An association study compares marker allele or genotype frequencies in a group of affected cases against a group of unaffected controls. Unlike linkage, association studies have been used to identify genes of minor effect in a number of complex genetic disorders. This includes the glucokinase and glycogen synthase genes in non-insulin dependant diabetes (NIDDM) and insulin dependant diabetes (IDDM) respectively [79, 80].

The term allelic association is used to describe a statistically significant difference in marker allele frequency between affected and unaffected individuals. Genotypic association refers to a significant difference in genotype frequency. An association between a marker and a certain disease can be explained by one of two genetic mechanisms: direct association, whereby the associated marker is the causal variant and indirect association due to linkage disequilibrium.

Linkage disequilibrium (LD) refers to the co-occurrence or correlation between two loci on the same chromosome. As discussed in section 1.2.1, a high correlation between two markers is likely to occur between loci that are close together (<50kb) as they are less likely to be separated by recombination during meiosis. The degree of LD between two loci is usually calculated by one of two measures, D' or r^2 [81, 82]. Both measures are based on the pairwise-disequilibrium coefficient D which is a measure of the co-variance between two loci. The value of D between two alleles (i.e. A and B) is calculated using the frequency of the two alleles (p_A and q_B) and the haplotype frequency (α_{AB}):

$$D_{AB} = \alpha_{AB} - p_A q_B$$

A limitation of using D as the degree of LD between two markers is that its possible value is constrained by the frequencies of each marker allele. Therefore in order to compare values of D between different pairs of markers with different allele frequencies, D is normalised to D' [82]. Unlike D, D' lies on a scale of 0-1 and is calculated using the theoretical maximal value of D (D_{max}) [81, 82].

$$D' = D/D_{max}$$

A D' value of 0 denotes no correlation between two loci while a value of 1 indicates complete LD. Complete LD between two loci indicates that one allele of a marker always occurs with an allele of the other marker. However a limitation of D' is that where a difference in allele frequency between two markers exists, a D' of 1 can occur even though the two markers are not in perfect correlation. For example, Allele A of locus 1 may always occur with allele B of loci 2. However allele B of locus 2 may occur with both allele A and a of the first polymorphism.

An alternative measure of LD which accounts for the above scenario is r^2. r^2 is measured using D plus the product of the allele frequencies at the two loci:

$$r^2 = \frac{D^2}{f(A_1)f(A_2)f(B_1)f(B_2)}$$

As with D', r^2 values lie between 0 and 1 however in this instance an r^2 value of 1 indicates a perfect correlation between the genotypes of two markers where the alleles of both markers always co-occur and therefore have the same frequency [83]. Because the genotype of one marker can be determined by the genotype of the second marker, two markers with an $r^2=1$ are termed proxies.

1.2.2.1 Association Study Design

In designing an association study for a particular gene or genomic region a number of so called "tag markers" are identified which account for all the known genetic variation at a particular locus under analysis, or at least to a specified minor allele frequency (MAF). If an association signal is detected then the polymorphisms tagged by the significant marker must be considered as potential risk variants along with the significant marker itself. Tagging a gene or genomic region poses a number of advantages. Firstly, by using tag markers a locus can be assayed without genotyping every polymorphism in the region, thereby dramatically reducing costs. In addition, no prior knowledge of the functional variants is needed as a functional role can be hypothesised, or investigated further, after putative risk allele(s) have been identified. Furthermore, all common, known genetic variation in a region will be assayed thereby reducing false negative rates. However there are limitations to a tagging approach. Rare variants are unlikely to be covered by this strategy. In addition, further analysis such as sequencing and extra genotyping, may be needed to find the causal variant(s).

1.2.2.1.1 Tag Marker Selection

In order to identify tag markers it is first necessary to genotype all polymorphisms at a locus within either a subset of your association sample or a representative population. From this the LD structure of the association sample can be estimated and tag markers selected. The HapMap database (www.hapmap.org) is a primary research tool which often allows the researcher to avoid this initial genotyping step. This publicly available database was created by the International Haplotype Map Consortium [84]. It contains details of over 5.5 million single nucleotide polymorphisms (SNPs) (HapMap phase II) which have been genotyped in 270 individuals from four populations (West European n=90, West African n=90, Han Chinese n=45, Japanese n=45).

Tag SNP selection is based on a number of factors. Firstly a researcher needs to decide how much of the genetic variation in a region to cover (i.e. to what MAF) and how

thoroughly (the degree of LD between the genotyped marker and the tagged SNP). Due to the large sample sizes needed to detect an association with rare events, a MAF $\geq$0.05 is often selected.

The two most common methods of tagging a locus involve the use of either haplotype tag SNPs (htSNPs) or pairwise tag SNPs (tag SNPs). Identifying htSNPs involves taking the haplotype diversity of a population in account rather than just a single marker-marker correlation. While this strategy can result in a reduction in the number of SNPs to be genotyped, and therefore a reduction in costs, interpreting the results can be problematic. Consequently a standard method of tag SNP selection has become pairwise analysis which involves measuring marker-marker (pairwise) LD via r^2 and subsequently selecting a set of independent markers. The ideal pairwise tag SNP selection strategy would be to genotype all markers where alleles were not in an r^2 of 1 with any other genotyped marker. However this threshold is often relaxed, i.e. r^2>0.8, to reduce the number of SNPs genotyped.

1.2.2.2 Family Based Association Studies

In addition to being the causal variant, or being in LD with the causal variant, a marker allele may also appear associated with disease if cases and controls are not ethnically comparable. In such a situation, known as population stratification, differences in allele frequency will emerge at all loci that differentiate between these two populations whether the alleles are causally related or not. Population stratification can also mask a true causal effect and produce a false negative result. It is therefore vital to ensure cases and controls are well matched for ethnicity. One strategy often employed is to use unaffected family members as controls, such as parents or siblings. The most popular family based association study design uses parent-proband trios, where both parents and their affected offspring are genotyped [61, 85]. If there is a distortion in the number of times an allele is transmitted to the affected offspring from a heterozygote parent, that is greater then expected by chance, then the allele is said to be associated. There

are numerous statistical methods to test for such an association, although the transmission/disequilibrium test (TDT) is the most widely used [85, 86].

Although family-based studies account for population stratification, a major drawback of this kind of sample is the difficulty in sample ascertainment. This can be particularly true of late onset disorders. Another limitation of the trio design is that at least two out of each trio needs to be genotyped for each variant. In addition, using unaffected siblings as controls results in a loss of power as they are over-matched to the cases [61].

1.2.2.3 Case Control Association Studies

A case control study design involves assaying the frequency of alleles or genotypes in a sample of affected individuals (cases) and comparing this with the frequency in a set of unaffected individuals (controls) from the same population. The main advantage of this design is the relative ease of sample collection compared to familial approaches which allows the construction of large sample sizes. In addition, as the analysis of each case is not reliant on an exact control (such as a parent or sibling) a large genotyped sample is often available for analysis which has greater power to detect real effects (see section 1.2.2.4). As discussed, population stratification is the main drawback of a case control study design. To protect against this, it is desirable that the sample is a homogenous population of similar ancestry [87]. Furthermore, a process of matching cases to controls for possible confounding factors such as age and sex is advantageous.

1.2.2.3.1 Statistical Evidence for Association in Case Control Studies

Statistical evidence for association in case-control studies comes from an estimation of the odds ratio or chi square statistic which are determined using a contingency table of either allele or genotype counts in cases and controls. Usually, the statistic is converted into a probability or p value. Typically a p value of 0.05 or less is used to indicate that the null hypothesis of no association can be rejected [61, 88] as such a value will occur by chance on only 5% of occasions, assuming all SNPs are independent. However as

many association studies test >20 SNPs such a p value can be expected to occur by chance for at least one variant. There are several methods that attempt to adjust the probability estimate for multiple comparisons, the most prominent of which are Bonferroni correction, experiment or gene wide adjustments and permutation methods [77]. Although statistical attempts to correct for multiple testing are justified in order to reduce the amount of false positives, it is highly likely these methods are over conservative in the presence of weak but true genetic effects [89]. Therefore the ideal method for assessing whether a reported association is a true effect is for it to be replicated in an independent sample [5].

1.2.2.4 Power of Association Studies

The major factors affecting the power of an association study to detect a susceptibility variant is the effect size, as well as the frequency, of the disease risk allele(s) [77, 90]. The more common a risk allele and/or the larger the effect size, the greater the power a study of fixed sample size has to find an association with that variant. The odds ratio (OR) of a variant is a measure of its effect size. It is defined as the odds of exposure to a susceptibility variant in cases compared to controls. For example, if a variant has an odds ratio of three, the odds of an individual with one copy of the risk allele being a case is three times higher than that of a control and the odds of an individual homozygote for the risk allele being a case compared to control are 9:1. Figure 1.3 illustrates the effect of both the OR and MAF of a disease susceptibility variant on the sample size required to provide 80% power to detect such variants at a $p<0.05$ level.

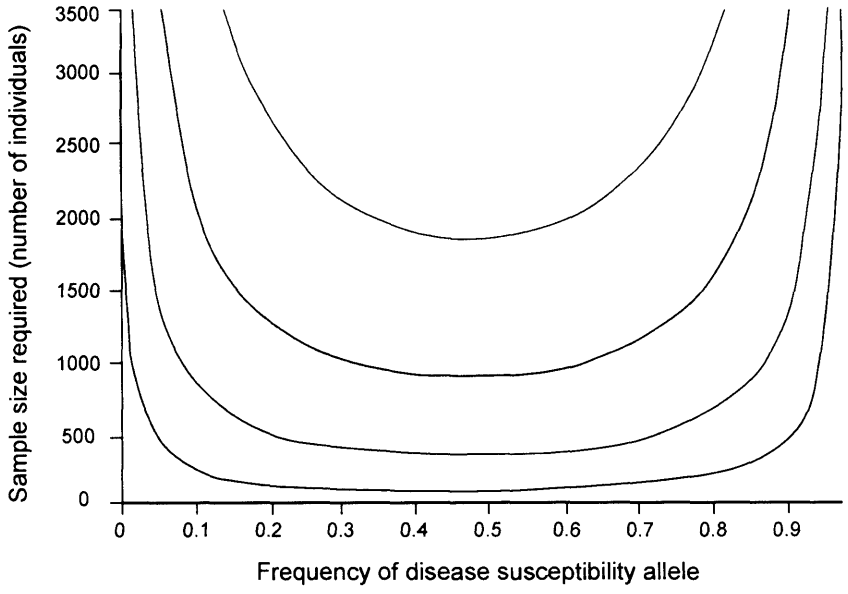


Figure 1.3. Effects of allele frequency and OR on sample size requirements. The total number of individuals (equal number of cases and controls) that are required in an association study to detect disease variants with allelic odds ratios of 1.2 (red), 1.3 (blue), 1.5 (green) and 2 (black) at various MAFs are shown. Numbers given are for a statistical power of 80% at a significance level of $p<0.05$. Adapted from [91] using Power and Sample Size Calculations software [92].

As can be seen from Figure 1.3 a sample of just under 500 cases and 500 controls (or 500 families with 2 parents and an affected offspring) is needed to have an 80% chance of detecting a risk allele with an OR of 1.5 and a MAF of 0.1. However to have the same level of power a sample of 4000 cases and 4000 controls would be required to detect a variant with the same effect size but a MAF of 0.01. In addition, although a sample of 500 cases and 500 controls have an 80% chance of detecting a risk allele with a MAF of 0.1 and an OR of 1.5, the sample size would need to be more than doubled to have the same chance of detecting a variant with an OR of 1.3.

It must also be noted that these figures assume a multiplicative model for risk and that either the disease variant is assayed itself or that there is perfect linkage disequilibrium between the test marker and disease variant. If this is not the case then even larger sample sizes may be required to achieve the same power.

As the allelic architecture of schizophrenia is completely unknown, the power of an association study to detect a risk variant can only be speculated. Two polarised views

have dominated much of the literature on the allelic frequency and risk affect of the variants that cause common diseases [91]. The common disease common variant (CDCV) hypothesis of schizophrenia and other complex disorders assumes that genetic variants that alter disease risk occur at relatively high frequency (MAF>0.05). Under this model, disease susceptibility is suggested to result from the joint action of several common variants and unrelated affected individuals share a significant proportion of disease alleles [93]. However, due to their common nature these variants are also likely to be of weak effect (see section 1.1.6.1). The extreme alternative to the CDCV hypothesis is the multiple rare variant hypothesis in which disease susceptibility is due to distinct genetic variants in different individuals and disease susceptibility alleles have low population frequencies (MAF<0.01) [94-98]. It is highly likely that both common alleles with a weak effect (OR<1.5) and rare variants (MAF<0.05) explain the instances of schizophrenia and therefore both need to be detected. As discussed above, this will require large sample sizes.

The advent of genome-wide association studies (GWAS) has also highlighted the need of increased samples sizes. GWAS studies have the capacity of assay over 500,000 polymorphisms at a time. Analysis of this number of variants necessitates that only a p value $<5x10^{-8}$ can be confidently reported as an association [99] and production of this level of significance requires large samples. It is therefore now becoming commonplace for research centres to form collaborations in order to produce sample sizes large enough to detect an effect. A case in point is the Wellcome Trust Case Control Consortium (WTCCC) which examined ~2,000 individuals for each of seven major complex diseases (bipolar disorder, coronary artery disease, Crohn's disease, hypertension, rheumatoid arthritis, and type 1 and type 2 diabetes) plus a shared set of ~3,000 controls [58]. However GWAS studies are aimed at detecting common variants only and therefore additional analysis, involving sequencing a large number of individuals, will be needed to investigate rare variants and schizophrenia.

1.3 DTNBP1 and Schizophrenia

The combination of linkage and association analyses has implicated a number of genes in the susceptibility of schizophrenia. These include neuregulin (NRG1) [100], D-amino oxidase (DAO) [101], D-amino oxidase activator (DAOA) [101], disrupted in schizophrenia 1 (DISC1) [102], catechol-O-methyltransferase (COMT) [103] and regulatory of G-protein signalling 4 (RSG4) [104]. However one of the most convincing susceptibility genes to date is DTNBP1.

The DTNBP1 gene encodes for the dystrobrevin binding protein 1 (commonly known as dysbindin) and is located on 6p24-22. Chromosome 6p24-22 is a well established schizophrenia linkage region [105-109] and DTNBP1 was first identified as a schizophrenia susceptibility gene through association analysis of this linkage region [109]. In this initial study Straub and colleagues typed 17 SNPs, plus an additional 3 simple sequence-length polymorphic markers (SSLPs) within 270 families from the Irish Study of High Density Schizophrenia Families (ISHDSF) [110] who had previously shown linkage to 6p24-21 [109, 111]. The 20 markers analysed at 6p22 spanned DTNBP1 (chr6:15,631,018-15,771,250), JARID2 (chr6:15,354,506-15,630,232) a gene 3' to DTNBP1, plus the 5' region proximal to DTNBP1. Significant evidence for association was observed between schizophrenia and a number of single markers within DTNBP1 ($p<0.01$). This evidence for association was further strengthened via haplotype analysis where several 3-marker haplotypes showed association ($0.001<p<0.008$) [112]. Although the association data was complex these results identified DTNBP1 as a susceptibility gene for schizophrenia which at least partly explained the linkage signal observed at 6p24-22. Reanalysis of this data, with the addition of extra markers, attributed the association signal to a single 8-marker haplotype spanning the DTNBP1 locus via a pedigree disequilibrium test (rs3213207, rs1011313, rs2619528, rs2005976, rs760761, rs2619522, rs1018381, rs1474605, GGAATGCG, PDT=0.002) [113].

Since the original report, 25 replication studies have been published to date. While some studies have failed to replicate the initial findings [114-128], further support for DTNBP1 as a schizophrenia susceptibility gene has been reported in Caucasian [117, 121, 127, 129-134], Asian [121, 135-138], African American [129] and Hispanic [117] populations. A summary of these positive associations are given in Table 1.1, while the more specific details of each study are described below.

The association between DTNBP1 and schizophrenia was first replicated in another family based association study [131] where six of the most significant SNPs reported in the initial study (rs3213207, rs1011313, rs2619528, rs760761, rs2619522, rs1018381) were genotyped in two independent samples, a sib-pair sample consisting of 78 families of German and Israeli origin and a sample of 127 German and Hungarian parent proband trios. Single marker association was reported in both samples separately and within the combined sample (rs3213207 p=0.0052, rs1011313 p=0.0092, rs2619528 p=0.140, rs760761 p=0.0007, rs2619522 p=0.030). Six-marker haplotype analysis revealed that 97% of the observed marker distributions could be explained with seven different haplotypes. Of these only one haplotype had a larger frequency of transmitted (T) rather than non-transmitted (NT) haplotypes (Trios: T=0.738, NT=0.594, Sib pairs: T=0.731, NT=0.566). This haplotype consisted of the common allele at each locus (AGGCTC). However as no formal association test was reported for this haplotype the single marker results are presented in Table 1.1.

A number of other family based studies have also observed an association between DTNBP1 and schizophrenia. Kirov and colleagues [130] genotyped 8 SNPs (rs2619539, rs3213207, rs1011313, rs2005976, rs2743852, SNP002901727/rs9476858, rs2619538 and rs909706) within a sample of Bulgarian trios (n=488) with schizophrenia or schizoaffective disorder. Over transmission of two markers (rs3213207 p=0.009 and rs2005976 p=0.0013) was observed and analysis of two, three and four marker haplotypes produced numerous positive results ($p<0.001$). The most significant of these was a two marker haplotype which included rs2619539 and rs3213207 (p=0.00027). However due to the number of different significant haplotypes,

a specific haplotype(s) carrying the susceptibility/protective alleles in the population could not be determined.

The first study to examine DTNBP1 and schizophrenia in an Asian population investigated seven SNPs previously typed in the Straub study (rs742105, rs760666, rs2619539, rs2619522, rs1018381, rs909706, rs3213207) in 233 Han Chinese trios [137]. Two of these SNPs (rs760666, rs3213207) were not included in any further analysis because both SNPs had a minor allelic frequency of <5% in the population sample. Although no significant association was found for any individual SNP, 4 marker haplotype analysis identified the haplotype GTCA (rs2619539, rs2619522, rs1018381, rs909706) which was significantly over-transmitted to affected individuals (p=0.00091).

Fallin *et al* genotyped 440 SNPs from 64 schizophrenia and bipolar candidate genes in a sample of 274 schizophrenic or schizoaffected Ashkenazi case-parent trios [139]. These candidate genes included DTNBP1 where 16 SNPs were selected for analysis on the basis of previous linkage or association and/or biological relevance. DTNBP1 met the criteria for a suggestive association (empirical p<0.05) to schizophrenia with both single marker and haplotypic associations observed. While SZgene [140] reports that the 16 markers genotyped were rs1047631, rs2619539, rs3213207, rs760666, rs766761, rs1011313, rs1018381, rs2619538, rs2619522, rs2619528, rs742106, rs1040410, rs2056942, rs760665, rs1018382 and DTNBP1_unknown , no specific details of the SNPs genotyped or the significant markers are given in the report. The results have therefore not been included in Table 1.1.

The most recent family based DTNBP1 association study involved association analysis of DTNBP1 polymorphisms in two USA based samples [129]. Duan and colleagues genotyped 26 DTNBP1 SNPs in a total of 646 subjects from 136 families, of which 72% were of European ancestry and 18% were African American. The SNPs examined included those previously reported as associated with schizophrenia (n=9), a number of coding SNPs (n=7), previously untested SNPs (n=7), plus three additional SNPs which

had previously shown no evidence of association but would allow the gene to be tagged at an r^2>0.8. rs7758659 showed single marker association in both the European (p=0.03) and the African American (p=0.045) samples. Both samples had the same overtransmitted allele (G) and therefore the combined sample showed even greater significance (p=0.004). This association was further strengthened by the addition of rs3213207 to form a two marker haplotype between rs7758659 and rs3213207 (GA, p=0.0015).

Although the initial association study plus a number of replication studies have been family based, a significant proportion of replication studies have been of case control design. The first case control study of DTNBP1 and schizophrenia was performed by Van den Bogaert and colleagues [127] who analysed five SNPs within three samples containing schizophrenic individuals and unaffected controls of German (418 cases, 285 controls), Polish (294 cases, 113 controls) and Swedish (142 cases, 272 controls) descent. While the German and Polish samples did not show a significant difference between cases and controls, a 5-marker haplotype was significantly associated in the Swedish sample (rs3213207, rs1011313, rs2005976, rs760761, rs1018381, AGATT, p=0.0098). This association became even more significant after a separate analysis of cases with a positive family history of schizophrenia was performed (p=0.00009).

Analysis of six previously tested DTNBP1 SNPs (rs909706, rs1018381, rs2619522, rs760761, rs2691528, rs1011313) in 524 individuals with schizophrenia or schizoaffective disorder and 573 controls also revealed evidence for association between DTNBP1 and schizophrenia [117]. Although no association was observed in the African American subset of the sample (215 cases, 74 controls), three SNPs (rs1018381, rs2619522 and rs2619528) were positively associated in the Caucasian subset (258 cases and 467 controls) as well as a smaller Hispanic subset (51 cases and 32 controls). Haplotypic analysis of the Caucasian sample identified a six-marker risk haplotype (rs909706, rs1018381, rs2619522, rs760761, rs2691528, rs1011313, GATGTG p=0.005). Combined analysis of both the Caucasian and Hispanic subsets

identified four significant markers (rs2619528, rs760761, rs2619522, rs1018381) with the greatest single marker association shown by rs1018381 (p=0.0006).

An association analysis between DTNBP1 and schizophrenia in a UK based case control sample (708 cases and 711 controls) [134] attempted to confirm DTNBP1 as a schizophrenia susceptibility gene by replicating the association observed in the initial Straub study [112] and/or by identifying other risk variant(s) through screening the exonic and putative promoter regions of DTNBP1. No evidence of single marker association between schizophrenia and a number of previously studied markers (rs2619539, rs3213207, rs1011313, rs2005976) was observed. The study also failed to replicate both the initial haplotypic association [112] and the association reported after reanalysis of the original data [113]. However Williams and colleagues did determine significant evidence for association in one new risk (rs2619539, rs3213207, rs2619538, CAT p=0.01) and two protective haplotypes (CAA p=0.006, GGT p<0.001). These haplotypes consisted of markers previously reported as associated as well as a novel SNP (rs2619538) which was identified through mutation detection of DTNBP1 and located within a putative promoter region. When this marker was typed in a sample that previously did not show association [123] the specific risk and protective haplotypes reported by Williams *et al* were also significantly associated with schizophrenia in the independent sample (CAT p=0.02, CAA p=0.047, GGT p=0.006) [134].

Two other case control studies of Caucasian origin have also provided evidence of an association between DTNBP1 and schizophrenia. A case control study of 80 cases and 106 controls of Italian origin identified significant haplotypes which also included rs2619538 (rs2619538, rs909706, AA p=0.048, rs2619538, rs909706, rs1018381, global p=0.040). However their most significant result was observed with the two marker haplotype containing rs760761 and rs2005976 (CA, p=0.034) [132].
A report into the association between DTNBP1 and schizophrenia in a Spanish sample (589 cases and 617 controls) identified both risk and protective haplotypes [133]. In addition to polymorphisms from other putative schizophrenia candidate genes, Vilella and colleagues genotyped 12 DTNBP1 SNPs, six of which had previously been typed

in other studies [112, 118]. Two single markers and four 3-marker haplotypes showed evidence for association. However, only one haplotype survived correction for multiple testing by permutation analysis. This 3-marker haplotype included three previously untyped markers rs875462, rs1040410 and rs6926401 (Risk haplotype TTT, corrected p=0.015). However rs875462 is in high LD (r^2=0.94) with rs7758659. rs7758659 is part of the 2-marker (rs7758659, rs3213207) risk haplotype identified in the Duan study (GA p=0.0015) [129].

In addition to the family based study by Tang and colleagues in Han Chinese trios [137], there have been four additional studies which have investigated DTNBP1 as a schizophrenia susceptibility gene in the Asian population. Three of these used a case control design [135, 138] and a further study analysed DTNBP1 in both an Asian family and Scottish case control sample [121].

Numakawa *et al* [135] genotyped six polymorphisms previously found to be associated with schizophrenia (rs2619539, rs3213207, rs1011313, rs760761, rs2619522, rs2619538) in a Japanese population of 670 cases and 588 controls. Four of these SNPs showed evidence of association (rs3213207 p=0.013, rs760761 p=0.027, rs2619522 p=0.022, rs2619538 p=0.025). The most significant of which, rs3213207, was further strengthened in haplotype analysis where a 2-marker haplotype containing rs3213207 showed the most significant association (rs3213207, rs1011313, GG p=0.00028).

Li and colleagues [121] examined 10 SNPs that were identified in the original report of association [112] plus rs2619538, in a sample of 638 Chinese trios and a Scottish sample consisting of 580 cases and 620 controls. Two polymorphisms (rs3213207 and rs2619528) showed single marker association with schizophrenia in the trio sample (p=0.02 and p=0.002 respectively). Further analysis also identified a 2-marker haplotype overtransmitted to affected individuals (rs2005976, rs2619528, GG corrected p<0.001). In contrast to the trio sample, no single marker was found to be significantly associated with schizophrenia in the Scottish case control sample. However one risk

haplotype (rs760761, rs2005976, CA p=0.0066) and one protective haplotype (rs2619538, rs909706, AT p=0.003) were identified.

In a response to the inconsistencies in the risk alleles reported in the two previous Asian studies, as well as earlier Caucasian studies, Tochigi *et al* examined 12 DTNBP1 SNPs in a Japanese sample of 314 cases and 314 controls [138]. Of the 12 SNPs studied, nine were common with the initial Straub study [112] and all the SNPs from the previous Asian studies [135, 137], except rs2619538, were included. No single marker association was observed however permutation tests detected a significant difference between cases and controls of several 10-marker haplotypes. In addition, haplotypic analysis of markers which had shown an association in previous Asian studies [135, 137], identified the risk haplotype TCTCG (rs742105, rs2619539, rs2619522, rs1018381, rs909706 p=0.0076). While this consisted of the same markers as the associated haplotype reported by Tang and colleagues [137] the risk haplotypes reported by each were different (Table 1.1).

The most recent study of DTNBP1 in an Asian population genotyped four of the six SNPs analysed in the Numakawa study [135] (rs3213207, rs1011313, rs760761, rs2619522) in a Korean sample of 908 cases and 601 controls [136]. Although no significant difference was observed in the allelic or genotypic distributions between cases and controls, haplotype analyses revealed a significant association with schizophrenia ($p<0.0001$). Two protective haplotypes were identified ACCC and ACTA ($p=0.00014$ and $p<0.001$ respectively). Sliding analysis showed that the markers rs760761 and rs2619522 exerted major contributions to this haplotype with the two protective haplotype (CC p=0.0003, TA p=0.0001) plus a significant associated risk haplotype (CA p=0.013).

dbSNP ID					rs875462	rs1040410	rs742105	rs6926401	rs7758639	rs2619539	rs3213207	rs1011313	rs2619528	rs2005976	rs760761	rs2619522	rs1018381	rs1474605	rs909706	rs2619538
Alternative name					--	--	P1333	--		P1655	P1635	P1325	P1765	P1757	P1320	P1763	P1578	P1792	P1583	SNP A
Chromosome Position (March 2006)					15646415	15655455	15681053	15693988	15701219	15728834	15736081	15741411	15757808	15758781	15759111	15761628	15765049	15766191	15768850	15773188
Alleles					T/C	C/T	T/C	T/G	G/A	G/C	A/G	G/A	G/A	G/A	C/T	T/G	C/T	A/G	G/A	T/A
Study	Year	Population	Sample	Association																
Straub/ Van den Oord	2002/3	Irish	Family	Haplotype							G	G	A	A	T	G	C	G		
Schwarb	2003	German and Israeli	Family	Single Marker							A	G	G		C	T				
Kirov	2004	Bulgarian	Family	Single Marker							A			G						
Li	2005	Scottish	Family	Haplotype										A	C					
Duan	2007	European and African American	Family	Haplotype					G		A									
Van den Bogaert	2003	Swedish	Case/Control	Haplotype							A	G		A	T		T			
Funke	2004	Caucasian and Hispanic	Case/Control	Single Marker									A		T	G	T			
		Caucasian	Case/Control	Haplotype								G	A		T	G	T		G	
Williams	2004	UK	Case/Control	Haplotype						C	A									T
Morris*	2003/4	Irish	Case/Control	Haplotype						C	A									T
Tosato	2007	Italian	Case/Control	Haplotype										A	C					
Vilella	2007	Spanish	Case/Control	Haplotype	T	T		T												
Tang	2003	Chinese	Family	Haplotype						G						T	C		A	
Li	2005	Chinese	Family	Haplotype									G	G						
Numakawa	2004	Japanese	Case/Control	Haplotype							G	G								
Tochigi	2006	Japanese	Case/Control	Haplotype			T			C						T	C		G	
Pae	2009	Korean	Case/Control	Haplotype											C	T				

Table 1.1. Summary of positive association studies between DTNBP1 and schizophrenia (studies that have reported an association with a specific phenotype are summarised in Table 1.2). SNPs genotyped by each study are given in grey. The alleles of SNPs reported as associated either on their own or as part of a multi-marker haplotype are given. For clarity the strand of the reported allele may have been changed. Adapted from Williams *et al* and Mutsuddi *et al* [134, 141]. *Significant haplotypic association only observed in the Irish population after additional genotyping of rs2619538.

1.3.1 Negative Association Studies

While the majority of publications have reported an association between DTNBP1 and schizophrenia, some studies have failed to observe an association [114-128]. These negative studies have mainly been performed with Caucasian samples and involved a case control design. Van den Bogart and colleagues were the first group to report a negative association between DTNBP1 and schizophrenia [127]. They examined 5 DTNBP1 SNPs (rs3213207, rs1011313, rs2005976, rs760761, rs1018381) in three samples containing schizophrenic individuals and unaffected controls of German (418 cases, 285 controls), Polish (294 cases, 113 controls) and Swedish (142 cases, 272 controls) descent. As mentioned previously, while a significant association was observed in the Swedish sample, no significant difference was observed between the cases and controls in either the German or Polish samples.

Other Caucasian case control samples which have failed to find an association between DTNBP1 and schizophrenia include a Scottish sample of 580 cases and 620 controls [121], a US sample of 451 schizophrenic or schizoaffective cases and 291 controls [128], a Dutch sample of 273 schizophrenic patients and 580 controls [114] and a UK sample of 450 cases and an equal number of controls [115]. A study of eight DTNBP1 SNPs in an Irish sample of 219 schizophrenia cases and 231 controls also failed to find an association [123]. Three of these SNPs were selected due to their significant association in the initial Straub study (rs3213207, rs2005976, rs760761). The remaining five SNPs (rs1047631, rs760666, rs2619539, rs2619542 and rs2619550) were identified from the NCBI dbSNP database. No association was observed with any of the markers analysed, even when cases positive for a family history of schizophrenia (n=65) were compared against controls. However as discussed in section 1.3, when an additional SNP s2619538 was genotyped within their sample, they were able to replicate the haplotypic associations reported by Williams *et al*[134].

Two studies have also failed to find an association after examining DTNBP1 SNPs in Caucasian family samples. The first by Hall *et al* investigated five SNPs in 210 US

trios plus 233 South African Afrikaner families [118]. The second study analysed DTNBP1 variants within 441 Finnish families which included 865 affected individuals [126].

Holliday and colleagues examined several DTNBP1 SNPs in both a case control (194 cases, 180 controls) and two family based samples (41 Australian and 197 Indian pedigrees but failed to find an association in any sample [119]. Two other studies also failed to observe an association between DTNBP1 and schizophrenia in the Asian population. The first examined DTNBP1 variants within a sample of 693 Taiwanese families [122]. The second was of case control design and analysed DTNBP1 in 194 Korean schizophrenic cases and 351 controls [120].

Finally, while DTNBP1 is associated with schizophrenia in one African American sample [129], analysis of DTNBP1 in another case control sample of African American origin (215 schizophrenic cases and 74 controls) failed to find an association [117].

It must be noted that the majority of these studies analysed very few DTNBP1 SNPs (nine out of the 15 negative studies genotyped <10 DTNBP1 polymorphisms). In order to avoid this caveat Peters and colleagues analysed 39 SNPs which tagged the DTNBP1 locus, in an Anglo-Irish case control sample of 336 schizophrenia cases and 172 controls [124]. However, again no association was observed.

Another possibility for the negative reports including the study by Peters *et al* [124] is the relatively low sample sizes analysed. One study which attempted to resolve the potential limitations of either low gene coverage or small sample sizes, investigated DTNBP1 along with another 13 of the best supported schizophrenia candidate genes within a large sample of 1870 cases and 2002 controls [125]. Sanders and colleagues genotyped a dense set of SNPs, which included previously associated polymorphisms as well as tagging markers, for each gene. In total 38 DTNBP1 SNPs were genotyped. Twelve of these had previously been reported as associated with schizophrenia, the

remainder tagged ~84% of the DTNBP1 locus (HapMap phase II, r^2>0.8, MAF>0.05). In addition to single marker analysis, haplotypic analysis was performed which included the examination of previously reported haplotypes. HapMap phase II SNPs that were not genotyped were imputed. No DTNBP1 SNPs showed a significant association with schizophrenia. However neither did any of the other 13 genes after experiment-wide correction for multiple testing. Sanders and colleagues suggest a number of reasons for their failure to find an association, including an atypical sample. However they also suggest that the odds ratios observed in their analysis are in a plausible range (1.10-1.23) for small susceptibility effects but below what would produce significant p values in their sample.

Despite these negative association results DTNBP1 markers show a significant association to schizophrenia after meta-analyses by two different groups [140, 142]. Li and colleagues analysed 9 SNPs from 12 DTNBP1 association studies [117, 121, 123, 127, 134, 135, 137, 143] which consisted of 3429 cases, 3376 controls and 721 trios [142]. Four out of the nine SNPs analysed showed a significant association to schizophrenia however, the authors note that the results from Funke and colleagues [117] appeared to account for the association signal at three of these variants and none survived correction for multiple testing. However, this study did not include all DTNBP1 association studies. In contrast the SZgene database contains an up to date collection of all published genetic association studies on which a continuously updated meta-analysis is performed on polymorphisms that have been genotyped in four or more independent case-control samples [140]. Genes that contain SNPs which posses a significant OR after meta-analysis, such as DTNBP1, are classed as "Top Results" and listed on the SZgene homepage.

1.3.2 DTNBP1 and Endophenotypes

In addition to reports of a general association between DTNBP1 and schizophrenia, several studies have also observed an association between DTNBP1 markers and specific schizophrenic disease correlates, often termed endophenotypes. These are summarised in Table 1.2.

The first report between DTNBP1 and a specific phenotypic variable was published by Williams and colleagues who observed that a protective haplotype (rs2619539, rs3213207, rs2619538, CAA p=0.006) was significantly associated with higher education achievement (corrected p=0.02) [134]. While higher education achievement would not meet the standard criteria for an endophenotype, it is often used as a proxy for IQ and cognitive function. In support of the link between DTNBP1 and cognition, Donohoe *et al* [144] reported that schizophrenic patients carrying the risk haplotype from the Williams study (CAT), which was also significant in their own case control sample, showed impaired spatial working memory (mean errors = 65) compared to patients without the risk haplotype (mean errors=50, p=0.009) [144]. Working memory has been consistently demonstrated to be impaired in schizophrenic patients and variations in working memory have been shown to be associated with general cognitive performance [145].

Risk haplotypes from other studies have also been found to be associated with endophentypes relating to cognition. Since the identification of their GATGAG (rs909706, rs1018381, rs2619522, rs760761, rs2619528, rs1011313) risk haplotype [117], Funke and colleagues have reported an association between this haplotype and a reduced general cognitive ability in both healthy volunteers (p=0.04) and schizophrenics (p=0.035) [146]. A more recent report by the same group found that carriers of the GATGAG risk haplotype also demonstrated a significantly greater decline in IQ than non-carriers (p=0.05) [147]. The individual alleles of DTNBP1 markers which form part of either the Williams [134] or Funke [117] risk haplotypes have also been shown to be significantly associated with IQ scores in a combined

sample of patients, siblings and controls (rs760761 p=0.026, rs2619522 p=0.025, rs2619538 p=0.038) [148]. A further study also found an association between rs2619528 and rs760761 and changes in prefrontal brain function [149]. However disturbed prefrontal brain function was observed in healthy individuals carrying the G and C alleles of rs2619528 and rs760761 respectively (p=0.0068) which are the opposite alleles to those found in the Funke risk haplotype. However these alleles were found to be associated with schizophrenia in the Schwab study [108].

In addition to comparing spatial working memory in schizophrenia patients with and without the CAT risk haplotype [144], Donohoe and colleagues have also reported significantly reduced P1 amplitudes in individuals carrying the CAT haplotype (n=14) than those not (n=12) [150]. PI performance is a measure of early visual processing. Other studies have also suggested a more generalised role for dysbindin in brain function than solely cognitive function.

Kircher *et al* analysed DTNBP1 genotypes and cognitive function as well as personality traits in a sample of healthy volunteers (n=521) [151]. Individuals carrying the common C allele of rs1018381 had a significantly higher total schizotypical personality questionnaire (SPQ-B) scale score compared to individuals carrying the T allele (p=0.0005). The SPQ-B is a 22 item self report screening instrument for schizophrenia spectrum disorder according to DSM-IV [152]. On the subscale score, although no association between rs1018381 and cognitive deficits were observed, the C allele carriers did show a significantly higher interpersonal deficit factor score (p=0.0005). This score describes negative symptoms such as social anxiety, no close friends, blunted affect and paranoid ideation. As the T allele of rs1018381, which tags the GATGAG risk haplotype, scored lower on the SPQ-B sum and the interpersonal deficit sub-score this suggests a protective effect by this allele. Although this is in the opposite direction to what would be expected it must be noted that the C allele of rs1018381 has been also been reported as associated with schizophrenia, including within the original Straub study [113, 137, 138].

In contrast the Kircher study, the Funke GATGAG risk haplotype has been shown to be associated with negative symptoms by DeRosse and colleagues [153]. In this study the risk haplotype was tested for association with the lifetime history of negative symptoms in 181 Caucasian patients with schizophrenia. *Post hoc* t test analysis revealed that carriers of the risk haplotype (n=26) had significantly higher ratings than non-carriers on avolition ($p<0.04$), alogia ($p<0.02$) and flattened affect ($p<0.02$). This resulted in a significantly higher overall negative symptom rating in carriers ($p=0.001$).

The risk haplotype demonstrated to account for all the association reported in the initial study by Straub *et al* (rs3213207, rs1011313, rs2619528, rs2005976, rs760761, rs2619522, rs1018381, rs1474605, GGAATGCG) has also been shown to be associated with high levels of negative symptoms in patients with schizophrenia ($p=0.004$) [154]. Furthermore, Pae *et al* [155] reported that a haplotype consisting of rs3213207 and rs1011313 is associated with higher positive and negative syndrome scale (PANSS) scores in 240 Korean schizophrenic inpatients. However subsequent analysis of the association identified that the major contributor to the association was a higher positive subscale score (AG $p=0.009$).

Several other studies have also reported an association between DTNBP1 and positive symptoms. Stefanis *et al* [156] examined 18 SNPs within a number of schizophrenia susceptibility genes, including DTNBP1, in 2243 male military conscripts. They observed a single marker association with the minor alleles of both rs2619522 and rs760761 with positive and paranoid schizotypy scores. These two alleles (G and T respectively) are on the GATGAG risk haplotype background. In addition they observed an association between these alleles and lower attention capacity supporting the link between DTNBP1 and cognitive deficits.

Three other studies have investigated variation at the DTNBP1 locus with schizophrenia psychotic symptoms [157-159]. Firstly, rs11558324 has been reported associated with child-onset psychosis (risk allele A, $p=0.014$) in a cohort of 102 children with onset of psychosis before age 13. The same report also showed an

association between several DTNBP1 SNPs (rs1047631, rs6924627, rs760761, rs2619522, rs11558324) and a number of endophenotypes measured by the premorbid adjustment scale (PAS) ($0.001<p<0.046$) suggesting DTNBP1 may contribute to early neurodevelopmental impairment [158]. Secondly Kishimoto and colleagues analysed the three markers that constitute the CAT risk haplotype in a Japanese cohort of 197 individuals with methamphetamine psychosis (MP) and 243 controls [159]. The psychotic symptoms shown in MP are close to those observed in schizophrenia and MP is considered a pharmacological model of schizophrenia [160]. In addition to a significant allelic association at rs3213207 ($p=0.00003$), the three marker haplotype was also significantly associated with MP ($p= 0.0005$). Specific haplotype analysis identified CAA as a protective haplotype ($p= 0.0013$). This is consistent with the Williams study which identified CAA as a schizophrenia protective haplotype [134].

Corvin and colleagues also investigated the impact of the CAT haplotype on clinical symptomatology in psychosis within 262 schizophrenic and schizoaffective patients of Irish descent [157]. Means scores were compared for risk haplotype carriers ($n=70$) against non-risk carriers ($n=123$) for principle components determined from 30 PANSS items. A significant difference was observed between groups for the hostility/excitability score ($p=0.004$) with risk haplotype carriers producing lower scores than non-risk carriers. Hostility/excitability has been identified as a mania-like factor by studies of functional psychosis [161] and in a recent study was shown to be the best factor at distinguishing between the schizophrenia and affective psychotic disorders [14]. It has therefore been suggested that the lower score observed in risk haplotype carriers indicates that DTNBP1 could contribute susceptibility to a "prototypical" schizophrenia, as described by Craddock *et al* [162], rather than affective symptoms on the mood psychosis spectrum.

An association between DTNBP1 and schizophrenia endophenotypes could suggest possible mechanism(s) through which pathogenesis could be mediated. However as with the general association data the studies described above are difficult to interpret as risk alleles and haplotypes have been reported to be associated with multiple

schizophrenia endophenotypes including cognition, positive and negative symptoms. A limitation of most studies that have attempted to investigate the relationship between DTNBP1 and specific endophenotypes relating to schizophrenia is that the sample sizes have been relatively small. In addition, a number of studies have only examined one risk haplotype often with only one endophenotype. These issues may have contributed to why the specific relationship between SNPs and haplotypes and certain symptoms remains unclear. A recent study by Lucinao and colleagues [163] tried to address some of these issues by investigating the association between several DTNBP1 polymorphisms and cognitive function in three large population samples (1054 Scottish, 1806 Australian, 745 English). Within each cohort a battery of well validated cognitive tasks were performed which included measurements of IQ, verbal ability and several memory functions. In total 12 markers were genotyped which included those within the Williams and Funke risk haplotypes (CAT and GATGAG respectively) plus six SNPs from the original Straub study. A number of single polymorphisms showed association with several cognitive abilities including executive function, processing speed, verbal declarative memory and freedom from distractibility. However the only association to survive correction for multiple testing was that of the CAT risk haplotype and verbal ability ($p<0.001$).

Overall the link between DTNBP1 risk alleles and haplotypes, particularly the Williams risk haplotype, appears to be strongest with an impaired cognitive ability. While the Funke risk haplotype shows association to both cognitive deficits and negative symptoms, these findings could be related, as negative symptoms and cognitive function have been shown to be highly correlated [164].

dbSNP ID					rs1047631	rs2619539	rs3213207	rs1011313	rs6924627	rs2619528	rs2005976	rs760761	rs2619522	rs1018381	rs1474605	rs909706	rs11558324	rs2619538
Alternative name					P3230	P1655	P1635	P1325	P3762	P1765	P1757	P1320	P1763	P1578	P1792	P1583	P3521	SNP A
Chromosome Pos					15631080	15728834	15736081	15741411	15741601	15757808	15758781	15759111	15761628	15765049	15766191	15768850	15771097	15773188
Alleles					A/G	G/C	A/G	G/A	G/A	G/A	G/A	C/T	T/G	C/T	A/G	G/A	A/G	T/A
Study	Year	Phenotype	Association	Original SZ Association Study														
Williams	2004	Higher Educational Attainment	Protective Haplotype	Williams 2004		C	A											A
Donohoe	2007	Impaired Spatial Working Memory	Risk Haplotype	Williams 2004		C	A											T
Donohoe	2008	Early Visual Processing	Risk Haplotype	Williams 2004		C	A											T
Corvin	2008	Hostility/Excitability	Risk Haplotype	Williams 2004		C	A											T
Lucinao	2009	Verbal Ability	Risk Haplotype	Williams 2004		C	A											T
Burdick	2006	General Cognitive Ability	Risk Haplotype	Funke 2004				G		A		T	G	T		G		
DeRosse	2006	Negative Symtoms	Risk Haplotype	Funke 2004				G		A		T	G	T		G		
Burdick	2007	Cognitive Decline	Risk Haplotype	Funke 2004				G		A		T	G	T		G		
Kircher	2009	Higher SPQ-B Total score and Interpersonal Defects	Risk Allele	Funke 2004										C				
Fanous	2005	Negative Symtoms	Risk Haplotype	Straub/Vanden Oord 2002/03			G	G		A	A	T	G	C	G			
Gornick	2005	Child onset psychosis	Sinle Marker														A	
		PAS Total	Single Marker		G				A			T	G				A	
Fallgatter	2006	Disturbed Prefrontal Brain Function	Risk Allele							G		C						
Stefanis	2007	Lower Attention Capacity	Risk Allele									T	G					
		Lower positive and paranoid schizotypy scores	Risk Allele									T	G					
Zinstock	2007	Full Sacle IQ	Risk Allele									T	G					T
Pae	2008	Psychotic Positive Symtoms	Haplotype				A	G										

Table 1.2. Summary of association studies between DTNBP1 and schizophrenia related phenotypes. The table has been divided into four sections. The first three give details of the associations observed with specific risk haplotype (Williams [134], Funke [117] and Straub [112, 113] respectively). The final section gives associations of individual risk alleles with different endophenotypes. Alleles or haplotypes are determined as risk if they have previously been reported as associated with schizophrenia (see Table 1.1 for details).

1.3.3 DTNBP1 and Bipolar Disorder

In light of the reports that DTNBP1 is associated with schizophrenia, a number of research groups have investigated variants at the DTNBP1 locus with bipolar disorder [139, 165-169]. The first association study of DTNBP1 markers and bipolar disorder was performed by Raybould and colleagues [169] who investigated the three marker schizophrenia risk haplotype reported by Williams *et al* [134] in a Caucasian case control sample of 726 bipolar cases and 1407 controls. Although they did not find any evidence that DTNBP1 influences risk to bipolar disorder in general, DTNBP1 was found to be nominally significantly associated in a subset of the bipolar cases (n=133) with predominantly psychotic episodes of mood disturbance. In these individuals rs2619538 was significantly associated with bipolar disorder with psychosis as well as the three marker schizophrenia risk haplotype CAT ($p=0.004$ and $p<0.042$ respectively). This is consistent with Corvin and colleagues who showed carriers of CAT had lower mania-like scores [157] and two studies that have observed significant association between DTNBP1 haplotypes and psychosis in schizophrenia patients [158, 159]. However it must be noted that these findings by Raybould *et al* were not significant after correction for multiple testing.

Breen and colleagues investigated 11 DTNBP1 SNPs within 213 bipolar patients and 197 controls [165]. Ten of these polymorphisms were those genotyped by the Straub study. The 11^{th} SNP chosen for analysis was rs2619538. Although single marker association analysis revealed nominal significance, the greatest association was observed with haplotypic analysis. While a number of three marker haplotypes were identified that were significantly associated with bipolar ($p=0.006$-0.008), the most significant result was observed with the four marker haplotype (rs2619522, rs760761, rs2005976, rs2619528, $p=0.005$). Within this 4 marker haplotype, the allele combination TCGG was found to be protective ($p=0.028$). No association, either allelic or haplotypic, was observed with rs2619538, the only SNP which was also genotyped in the Raybould study [169].

The most recent case control study of DTNBP1 and bipolar disorder in a Caucasian sample was performed in a sample of 515 bipolar patients and 1316 control subjects[166]. A total of seven DTNBP1 SNPs were analysed directly and rs2619522 was imputed. The majority of these SNPs were chosen based on their inclusion in either the Straub [112] or Williams [134] studies. A number of allelic and genotypic associations were observed, however apart from the allelic association for rs3213207 ($p=0.006$) none survived correction for multiple testing. The most significant result was produced via haplotypic analysis where a 4 marker haplotype consisting of rs16876571, rs2619539, rs3213207 and rs760761 was significantly associated with bipolar (GCGT, $p=0.02$).

In addition to case control studies there has been one family based study of DTNBP1 and bipolar disorder in a Caucasian sample [170]. This study investigated several genes previously associated with schizophrenia in 379 bipolar patient-affected offspring trios. This included DTNBP1, where 19 markers were genotyped which tagged 87% of the DTNBP1 locus at an $r^2>0.8$. These SNPs also captured 8 out of the 11 SNPs previously reported as associated with schizophrenia at an $r^2=1$. However analysis using gene-based tests revealed no significant evidence for association between DTNBP1 and bipolar disorder.

A second family based study analysed DTNBP1 with bipolar disorder in a population isolate sample [139]. In addition to analysing DTNBP1 for an association with schizophrenia, Fallin and colleagues also genotyped several DTNBP1 polymorphisms in 323 bipolar disorder Ashkenazi case-parent trios. As with the schizophrenia trio sample analysed in the same study, DTNBP1 met the criteria for a suggestive association. Several single markers were reported to show an association ($0.01<p<0.05$) although no specific details are given. In light of the Raybould study [169] it may be of note that 67% of the bipolar affected cases had psychosis.

Along with publications using Caucasian and population isolate samples there have been two studies which have investigated DTNBP1 and bipolar disorder in the Asian population. Pae and colleagues [168] analysed five SNPs (rs3213207, rs1011313, rs2005976, rs7607961, rs2619522) in a Korean sample of 155 bipolar cases and 478 controls. Although no single marker association was observed, the 5 marker haplotype was found to be significantly associated (global p=0.009) which appeared to be in part due to the protective effect of the ACGTA haplotype (p=0.00016). Additional sliding window analysis identified the GTA alleles of the haplotype to be the major contributors (p=0.00006).

Joo and colleagues genotyped rs2619538, rs2619522 and rs760761 in a sample of 163 patients and 350 controls again of Korean descent [167]. A genotypic association was observed for both rs2619522 and rs760761 (p=0.014 and p=0.026 respectively). However these did not remain significant after Bonferroni correction for multiple testing.

1.3.4 DTNBP1 and Major Depressive Disorder

A number of studies have examined DTNBP1 variants with major depressive disorder (MDD) [171-175]. Kim *et al* analysed 4 SNPs (rs3213207, rs1011313, rs760761 and rs2619522) in a Korean sample of 188 MDD patients and 350 controls [172]. A significant association (p=0.0014) was observed via 4-marker haplotype analysis with the protective haplotype ACTA showing the greatest association (p=0.002). Sliding window analysis revealed that the major contribution was due to the rs761761 and rs2619522 haplotype (p=0.000026). Although neither of these SNPs showed single marker association with MDD, rs760761 had previously been reported by the same group to show allelic association, along with rs2005926, with antidepressant response (p=0.00055 and p=0.0058 respectively) [173].

Three studies have analysed DTNBP1 in MDD case control samples of Caucasian origin [171, 174, 175]. None of these observed a significant association either at a single marker or haplotype level. However a number of factors may explain why the Caucasian studies reported a negative association with MDD. Two of the studies [171, 174] did not genotype all markers from the associated 4 marker haplotype reported by Kim *et al*. In addition the protective haplotype ACTA previously reported was not observed in the case control sample analysed by Zill *et al* [175]. Nevertheless, it must be noted that all studies analysing DTNBP1 and MDD have examined only a limited number of DTNBP1 variants (range=4-6 markers) and the majority of studies [171, 172, 174, 175] are likely to be underpowered due to small sample sizes.

1.4 Possible Explanations for Inconsistent Association Results

While the association studies reported in sections 1.3 and 1.3.2 provide strong support for DTNBP1 as a schizophrenia susceptibility gene, this data is by no means conclusive. Even allowing for the fact that many of the studies reporting a positive association have not analysed the same genetic markers, there is considerable disagreement between publications with respect to the associated markers and haplotypes reported. For example the initial replication study by Schwab and colleagues [131], which genotyped 6 SNPs from the 8-marker risk haplotype described by Van den Oord *et al* [113], determined that the 6-marker haplotype consisting of the common alleles at each locus (rs3213207, rs1011313, rs2619528, rs760761, rs2619522, rs1018381, AGGCTC) was the only 6-marker haplotype to occur at a larger frequency in transmitted than non-transmitted haplotypes. Although this observation was consistent with their single marker results (See Table 1.1), it was in contrast to the results reported in the Van den Oord reanalysis [113] of the Straub study [112] where the rare alleles were preferentially associated with schizophrenia in the 8-marker risk haplotype (gGantgCn, rare alleles are depicted in lower case and those not typed in the Schwab study are given as N). Even more generally, as can be seen from Table 1.1, the associated alleles reported tend to "flip-flop" between studies.

One potential explanation for the differences observed is that the design of the majority of the association studies to date makes it improbable that the significant SNPs reported have a direct pathogenic role in the function of DTNBP1. The original study selected SNPs from public databases and relied on these SNPs to be in sufficient LD with the actual functional variant(s) for an association signal to be detected. Most replication studies have not attempted to identify these functional variants but have typed SNPs previously reported as associated. Therefore these studies also rely on the fact that the SNPs examined are in LD with the functional variant(s) in their sample.

One study that did attempt to identify DTNBP1 causal variants undertook a direct approach by screening all exons and putative promoters for sequence polymorphisms [134]. However even here haplotypic analysis yielded a more significant association than single markers, suggesting the true causal variants have yet to be identified.

It should therefore be noted that the degree of correlation between two SNPs is highly population dependant. In addition the effect size of a single variant can also vary between populations. The potential differences in the genetic architecture of a sample could possibly explain the patterns of association observed. However some of the differences reported, such as those between the Straub/Van den Oord and Schwab studies are within relatively related populations. Furthermore, a recent report argues against the hypothesis of sample variation [141]. Mutsuddi *et al* identified the single marker or haplotype that best captured the association signal in six studies which had investigated DTNBP1 in samples of European ancestry [112, 117, 127, 130, 131, 134]. From this, the tag SNP(s) for each risk haplotype were determined and genotyped through the HapMap CEU sample. Mutsuddi and colleagues observed that the haplotypes frequencies reported in each study were broadly similar to their frequency in the CEU population. This suggests that the European samples studied are genetically similar and that population stratification cannot explain the differences observed.

Consequently, one of the more likely explanations for the disagreement between studies is that there are multiple risk variants at the DTNBP1 locus (known as allele heterogeneity). This hypothesis is supported by the Mutsuddi study [141] which, using a high density reference map also demonstrated that the associated haplotypes reported in each study was inconsistent with one common causal variant contributing to schizophrenia at the DTNBP1 locus as all of the five supposed replication studies [117, 127, 130, 131, 134] defined a different associated haplotype than the haplotype originally reported [112, 113]. Another possibility is that there is a single susceptibility allele, which has yet to be identified and is carried on a remarkable diversity of haplotypes even in closely related populations [5].

1.5 Altered Expression of Dysbindin in Schizophrenia

Of all the studies to date between DTNBP1 and schizophrenia, no non-synonymous alleles have been shown to account for any of the association reported. This suggests that the causal variant(s) within DTNBP1 may either be as yet unidentified non-synonymous SNPs or polymorphisms which function by a mechanism other than altering protein structure. The later possibility is supported by molecular genetic studies into other complex disorders where several susceptibility genes have been identified but no obvious pathogenic mutations have been found [112, 176-179]. Possible alternative mechanisms by which DTNBP1 could confer susceptibility to schizophrenia include enhancing or inhibiting alternative spicing, altering the expression of dysbindin or a particular dysbindin transcript, or affecting mRNA stability or processing. A number of recent studies have provided evidence that altered dysbindin expression may be the mechanism by which DTNBP1 confers susceptibility to schizophrenia. These have included observations of a reduction in dysbindin expression in schizophrenic patients compared with controls [180-183] and evidence that variation in dysbindin expression within the brain is the result of genetic variation [184-186]. Details of these findings are summarised in section 1.5.1 below.

1.5.1 Dysbindin Expression in Schizophrenic Patients

Brain imaging and cognitive assessments consistently show a dysfunction in the dorsolateral prefrontal cortex (DLPFC) of patients with schizophrenia [187, 188]. In addition the poor performance on working memory tasks observed in schizophrenia have been associated with reduced activation of the DLPFC [144]. As a result, Weickert and colleagues investigated the levels of dysbindin mRNA within the DLPFC of schizophrenic patients [183]. Comparison of 7 schizophrenic cases and 13 controls revealed a 15-20% reduction in dysbindin mRNA levels within the patient group ($p=0.04$).

Corroborative results for this finding have been reported in similar studies which have measured dysbindin mRNA and protein levels in the hippocampal formation (HF). The HF has an important role in mediating cognitive and behavioural aspects of schizophrenia and shows frequent synaptic abnormalities in patients with the disorder [189, 190]. The first study to look at dysbindin levels in the HF compared dysbindin protein levels within post-mortem tissue from two sets of schizophrenic patients (n=17 and 15 respectively) plus matched non-psychiatric controls [181]. A presynaptic reduction in dysbindin protein was reported at hippocampal formation sites of both sets of schizophrenic patients (average reductions of 18-42%, p=0.027-0.0001). The schizophrenic cases showed no loss or shrinkage of DGh neurons, no loss of DGiml thickness and no loss of the synaptic markers synaptophysin or synapsin-1 in the DGiml, thereby suggesting that the reduced presynaptic dysbindin observed within the HF was associated with schizophrenia rather than general differences in the HF [181]. Corroborative evidence for these findings was reported in a follow up study by the same group which observed reduced dysbindin mRNA in the HF of schizophrenics (n=10) compared with matched controls (20-40% reductions, p=0.04) [182].

While great care is taken when sample matching, measuring gene expression using post-mortem tissue is susceptible to confounding influences arising from characteristics of the tissues themselves (for example RNA quality, cellular heterogeneity) as well as demographic and environmental confounds (such as the effects of medication). Although the reduction in dysbindin expression detailed above was not significantly correlated with age, sex, PMI, or antipsychotic exposure, it's worth noting that a reduction in dysbindin RNA, similar to that observed in the post mortem studies, has also been observed in peripheral blood lymphocytes (PBL) of 12 schizophrenic patients compared to controls (28% reduction, p=0.02) [180]. The measurement of gene expression in PBL removes the influence of neurohormonal, medication and environmental factors by several passages of cell culture [180].

1.5.2 Cis-acting Variation of DTNBP1

If DTNBP1 does confer susceptibility through altered expression then this could be facilitated in a number of ways. Expression of human protein-encoding genes can be regulated at a number of different stages including the control of chromatin structure, initiation of transcription, mRNA processing, transport of mRNA to the cytoplasm, mRNA stability, translation of mRNA and protein activity [191, 192]. Altered dysbindin expression in schizophrenics has been detected at both the mRNA and protein level and in some instances reduction of these two forms has been shown in the same region of the brain, for example the dentate gyrus of the HF [181, 182]. This suggests that the mechanism by which dysbindin expression is reduced involves a reduction in dysbindin transcript levels or mRNA stability as opposed to degradation of the protein.

Factors that affect the transcription or mRNA stability of DTNBP1 can be genetic or environmental. Genetic influences on gene expression fall into two categories, cis or trans-acting. Cis-acting factors are located on the same chromosome as the target gene and although they are more likely to be close to their target i.e. within the promoter, cis-acting regulatory regions can be some distance from the coding sequence [191]. Trans-acting factors can be situated anywhere in the genome and influence the expression of a target gene through interaction with cis-regulatory sequences. While the post-mortem and PBL analyses described in section 1.5.1 cannot distinguish between these genetic influences, one method that can determine whether a gene is under the influence of cis-acting regulation is the allelic expression assay.

1.5.2.1. Relative Allelic Expression Analysis

Allelic expression (AE) analysis can determine whether the expression of a gene is under the influence of either polymorphisms within regulatory elements or other cis-acting phenomena such as epigenetic modification. In the absence of cis-acting influences that differentially effect the transcription or stability of the two copies of mRNA of an autosomal gene, both mRNA copies will be equally expressed. However if an individual is heterozygous for any cis-acting polymorphisms, or where one copy of a gene is systematically affected by epigenetic modification, then the mRNA from each chromosome will be expressed at different levels.

The relative expression of each copy of a specific gene can be measured with quantitative allele discrimination of a coding SNP within the assayed gene. This coding SNP is used as a copy-specific tag for any cis-acting regulatory SNPs (also known as eSNPs). Individuals heterozygous for the coding SNP are analysed. If any individuals are also heterozygote for a regulatory polymorphism, the presence of cis-acting regulatory variation will be detected in a significant departure from the expected 1:1 allele ratio in a sample (Figure 1.4) [193]. This 'within-subject' approach means that each allele acts as an internal control for the other. Only cis-acting factors will be detected and other influences, such as differences in tissue preparation, mRNA quality, environmental factors and trans-acting regulatory influences are controlled for [193].

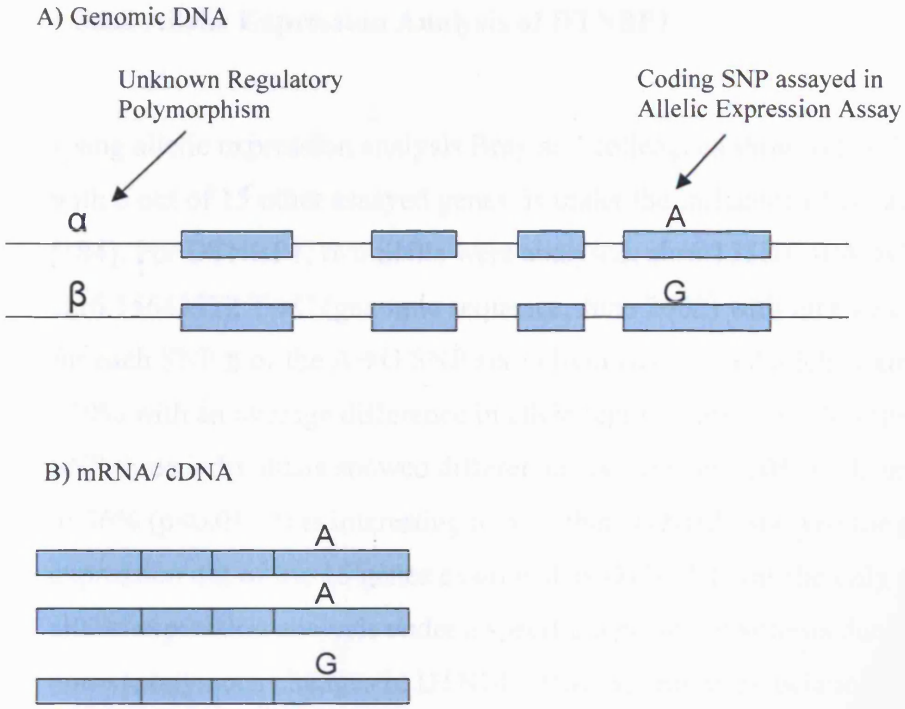


Figure 1.4 Principle of relative allelic expression analysis. A) An individual who is heterozygote for an exonic A/G polymorphism is also heterozygous for an unknown regulatory SNP denoted α/β. B) One allele of the regulatory polymorphism (denoted β) results in lower expression of the gene copy containing the G allele, with which it is in phase. C) The relative under expression of the G allele (and thus the presence of the regulatory polymorphism) is detected by applying a quantitative method of allele discrimination to the cDNA and comparing this with the relative representation of the two alleles observed in genomic DNA. Adapted from [193].

1.5.2.2 Allelic Expression Analysis of DTNBP1

Using allelic expression analysis Bray and colleagues showed that DTNBP1, along with 6 out of 15 other assayed genes, is under the influence of cis-acting variation [184]. For DTNBP1, two SNPs were analysed, chr6:15580740A→G and chr6:15643772 T→C (genomic sequence, June 2002) with nine heterozygotes assayed for each SNP. For the A→G SNP six individuals showed allelic expression differences >20% with an average difference in allele representation of 66% ($p<0.01$). For the T→C SNP three individuals showed differential expression >20% with an average difference of 36% ($p<0.01$). It is interesting to note that DTNBP1 showed the greatest differential expression out of the 15 genes examined as DTNBP1 was the only gene selected for allelic expression analysis under a specific *a priori* hypothesis due to the absence of non-synonymous changes in DTNBP1 that explain its association with schizophrenia.

A more in depth analysis by the same group confirmed the influence of cis-acting variation on DTNBP1 [186]. Allelic expression analysis of DTNBP1 was performed using rs1047631, an expressed SNP located within the 3'UTR of the majority of predicted DTNBP1 transcripts (Figure 1.5) [134]. 149 Caucasian individuals were genotyped for rs1047631. 31 of these individuals were heterozygous and therefore informative for allelic expression analysis. Primer extension analysis using SNaPshot chemistry, was performed on two cDNA samples and one corresponding gDNA sample for each individual. Bray and colleagues reported a significant reduction in expression of the cDNA carrying the common A-allele of rs1047631 relative to the cDNA carrying the G-allele ($p<0.0001$) with a mean A/G ratio of 0.86 (Figure 1.6). This deviation from a 1:1 ratio provided evidence that unknown DNA variants in cis to the DTNBP1 gene are causing variation in DTNBP1 expression.

Although a statistically significant reduction was observed, there was considerable variability in the cDNA ratios between the individuals assayed which ranged from 0.64, a 36% reduction of allele A, to 1.07, a 7% increase in allele A. Sample variance was

discounted as firstly repeat assays showed good reproducibility between individual cDNA ratios (SD/mean = 0.05) and secondly 15 extra samples were analysed from two brain regions and within subject comparisons between these regions indicated no significant differences (paired t test p=0.61). It was therefore suggested that the spread of data could potentially reflect multiple cis-acting variants influencing expression, interactive effects between a cis-acting variant and other trans-acting factors or environmental effects.

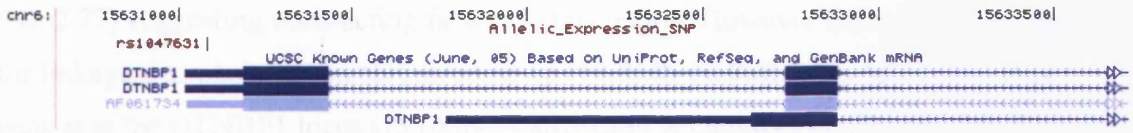


Figure 1.5. Chromosomal location of the allelic expression assay SNP rs1047631. rs1047631 is located in the 3'UTR of three DTNBP1 transcripts, DTNBP1a, DTNBP1c and AF061734. Differential expression of the transcript DTNBP1b was not captured by this assay.

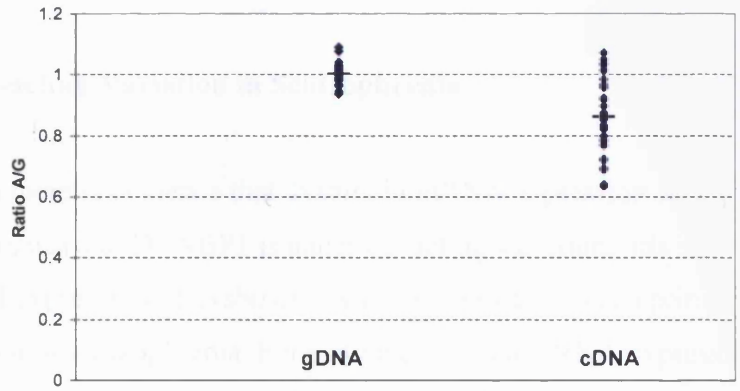


Figure 1.6. Comparison between corrected genomic and cDNA allele ratios in heterozygotes for SNP rs1047631. Data are represented as a ratio of A/G alleles. Significant under-representation of the A allele is observed in cDNA ($p < 0.0001$).

1.5.2.3 Other Evidence of Cis-acting Variation

The allelic expression results of DTNBP1 have been further corroborated by a recent study whose primary focus was to identify trans-acting loci on DTNBP1 expression [185]. This analysis involved using gene expression as a quantitative trait for linkage analysis to map polymorphic regulatory loci. Real time quantitative PCR measures of lymphoblastoid DTNBP1 expression were obtained from 200 individuals, consisting of 26 CEPH sibships, which were then used for genome-wide quantitative linkage analysis. A linkage peak was observed on chromosome 8p (28.1Mb-32.2Mb, max lod=2.77) suggesting trans-acting factors in this region. However the strongest evidence for linkage was observed on chromosome 6p (14.8Mb-16.9Mb, max lod=3.2). This peak is at the DTNBP1 locus (15.6Mb-15.8Mb) and is therefore consistent with cis-acting influences. 56 out of the 200 individuals that constituted the linkage sample were heterozygous for the coding SNP rs1047631 and allelic expression analysis on the lymphoblastoid cDNA from these individuals confirmed variable cis-acting influences on DTNBP1 expression operating in the lymphoblastoid cells [185].

1.5.3 Cis-acting Variation in Schizophrenia

Although there is evidence that dysbindin mRNA expression is reduced in schizophrenics and DTNBP1 is under cis-acting variation, this is not enough to infer decreased expression of dysbindin by cis-acting variants is a primary aetiological mechanism in schizophrenia. For example reduced mRNA expression could be a compensatory mechanism for enhanced dysbindin function. Ideally any schizophrenia risk SNPs and/or haplotypes need to be shown to affect the expression of DTNBP1. However given the complex pattern of associations in the literature as well as the difficulty in replicating complex systems such as the human brain *in vitro*, direct analysis is difficult. In an effort to prove that the reduction of dysbindin expression is relevant to schizophrenia pathophysiology two studies attempted to link the observations of low dysbindin expression with genetic associations reported.

A preliminary study by Weickert and colleagues [183] looked into whether the altered expression in dysbindin seen in the DLPFC could be linked to DTNBP1 genotypes. When individuals were stratified by genotype a significant difference in dysbindin expression was observed with 4/11 SNPs examined (rs1047631, p=0.048, rs2743864, p=0.04, rs11558324, p=0.004, rs2619537, p=0.05). These four SNPs were located in the promoter, 3'UTR and 5'UTR of DTNBP1 and, although the study involved relatively few individuals, the results suggested that variation in dysbindin mRNA levels in schizophrenics may be due to altered transcription rates caused by cis-acting factors.

To investigate whether cis-acting influences are relevant to the aetiology of a particular disease, allelic expression data can be stratified by the genotypes of an associated SNP/haplotype to determine whether this DNA variant is correlated with differential expression. Allelic expression ratios of individuals with one copy of a risk allele/haplotype are compared against individuals with zero or two copies of the risk allele/haplotype. Individuals with two copies of the risk allele/haplotype are grouped with individuals with no copies due to the nature of allelic expression data. As results are given in the form of a ratio even if an individual has two copies of a variant that alters mRNA expression they would give an allelic expression ratio of 1. Hence it is only individuals heterozygous for a risk allele/haplotype that will show a difference in expression if that risk allele/haplotype is also associated with altered expression.

In previous studies this strategy has been used to establish that a COMT haplotype associated with schizophrenia, is correlated with reduced COMT expression (p= 0.003) [194]. It has also been reported that mRNA carrying the ε4 allele of APOE, a risk allele for Alzheimer's disease, has significantly increased expression compared to mRNA carrying either the e3 or e2 alleles ($p<0.0001$) [195]. Consequently, this approach was applied to DTNBP1 allelic expression ratios using schizophrenia association data previously reported [134]. Firstly, Bray and colleagues carried out a comprehensive re-analysis of previous case control genetic association data [134] to identify the risk

haplotype that included phase information for rs1047631. As described above, this coding SNP had been used by the group to perform the allelic expression analysis. As a result a 3-marker haplotype, TAA (T allele of rs2619538, A allele of rs3213207 and A allele of rs1047631) was identified that maximally differentiated the cases and controls (5.2% increase in the cases). Individuals assayed in the allelic expression analysis were genotyped for this 3-marker haplotype and the allelic expression data stratified by individuals heterozygous for the risk haplotype and individuals carrying no copies of the risk haplotype. The relative expression of the A-allele of rs1047631 was shown to be significantly lower when it was carried on the risk haplotype (TAA AE ratio=0.79) compared with when it was carried on the non-risk haplotypes (Non risk AE ratio=0.92, p=0.002, see Figure 1.7).

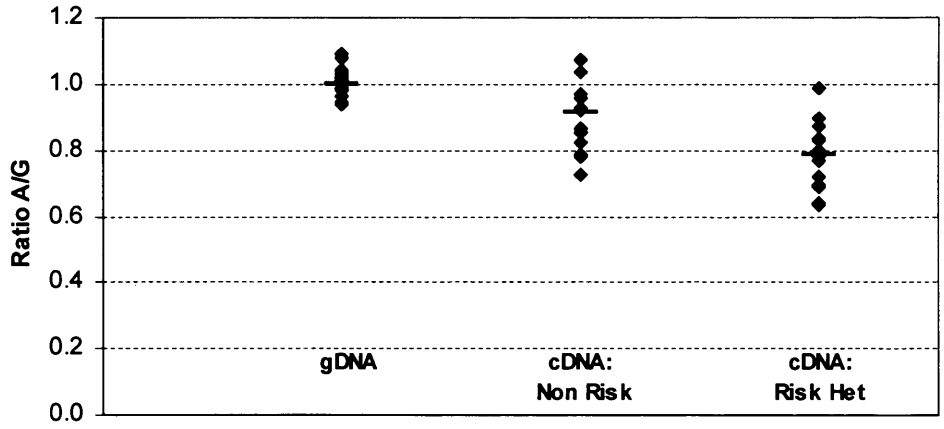


Figure 1.7. Allele ratios at SNP rs1047631, stratified by heterozygosity for the defined 3-marker schizophrenia risk haplotype. Data represented as a ratio of A/G alleles of rs1047631.

This observation suggests that variants associated with schizophrenia could be causing differential expression of DTNBP1. Nevertheless, Bray and colleagues note that it cannot be assumed that any of the SNPs within the risk haplotype analysed have direct effects on DTNBP1 expression. This is illustrated by the fact that the risk haplotype does not account for all the cis-acting variation observed (Figure 1.7). Individuals who do not carry the risk haplotype have an average allelic expression ratio of 0.95 and three individuals show a relative decrease in expression of greater than 20%. There may therefore be functional variants unidentified in the original case control study.

1.6 The Dysbindin Protein and Putative Function

When DTNBP1 was first identified as a schizophrenia susceptibility gene little was known about the dysbindin protein. However although its specific function is still unclear, evidence from a number of sources, such as interacting proteins and animal models, has provided some clues as to the function of dysbindin and how reduced expression of the protein could confer susceptibility to schizophrenia.

The gene DTNBP1 encodes dysbindin, a ubiquitously expressed protein with the highest levels of expression detected in the testis, liver, kidney, brain, heart and lungs [196]. Immunohistochemical studies have shown that dysbindin is expressed in multiple regions in the brain including the hippocampal formation and frontal cortex. Within these regions dysbindin is located in a diverse set of neuronal populations, within cell bodies as well as both pre and post synaptic sites. Multiple protein isoforms of dysbindin have been reported, reflecting alternative splicing within the gene. At present three mRNA transcripts (NM_032122, NM_183040 and NM_183041) have been validated according NCBI's Reference Sequence (RefSeq) [197] which encode the proteins dysbindin-1A, dysbindin-1B and dysbindin-1C respectively. Bioinformatic analysis suggests that many more mRNAs exist and 13 other alternative splice variants of DTNBP1 are listed on the AceView database (www.ncbi.nlm.nih.gov/AceView) [198]. However whether these encode active proteins remains to be seen. The three validated isoforms appear to be commonly expressed and have a molecular mass of 48kDA, 36kDA and 32kDA [199]. Computer aided sequence analysis of the three protein isoforms shows all contain a predicted coiled-coil region which spans 89 amino acids (See Figure 1.8) [196]. Coiled coils are bundles of intertwined alpha-helices that provide protein-protein interaction sites for the dynamic assembly and disassembly of protein complexes [200].

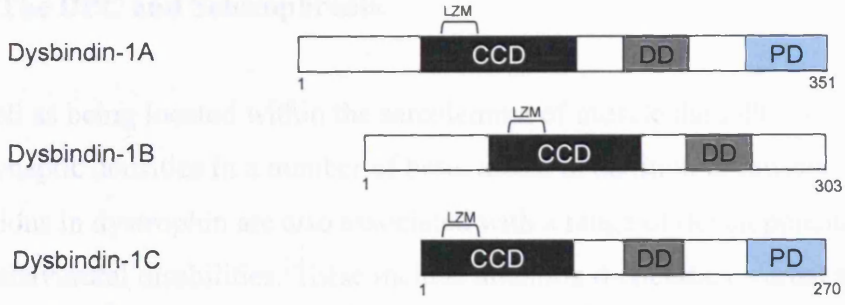


Figure 1.8. Structural comparison of the three major dysbindin isoforms. Numbers below each variant gives the amino acid sequence location beginning at the C-terminus. CCD = coiled coil domain. DD = dysbindin domain. PD = PEST domain. LZM = leucine zipper motif. Adapted from [199].

1.6.1 Dysbindin and the Dystrophin-associated Protein Complex

Dysbindin is known to interact with the dystrophin-associated protein complex (DPC), specifically the α- and β-dystrobrevins [196]. In fact, dysbindin was originally discovered by a yeast-two hybrid screen designed to identify β-dystrobrevin interacting proteins, hence dysbindin's full name dystrobrevin binding protein. The DPC, also known as the dystrophin glycoprotein complex (DGC), is a multi-protein complex comprised of three distinct components: the sarcoglycans, the dystroglycans and the cytoplasmic complex which includes the sytrophins and the dystrobrevins [201]. The correct assembly of the DPC is vital for the normal function of muscle and disruption of the complex causes muscular dystrophy, a disorder characterised by progressive muscle wasting [202]. Duchenne muscular dystrophy (DMD) is the most common and severe form of the disease. Patients are usually confined to a wheelchair by the age of 12 and die in their late teens or early 20s due to respiratory failure. DMD is caused by mutations in the DMD gene which encodes the 427-kDa cytoskeletal protein dystrophin. A milder form of the disease, Becker muscular dystrophy (BMD) is also caused by mutations in the DMD gene but the disorder has a much later onset and a longer lifespan [203].

1.6.2 The DPC and Schizophrenia

As well as being located within the sarcolemma of muscle the DPC is also found in postsynaptic densities in a number of brain areas. In addition to muscle disorders, mutations in dystrophin are also associated with a range of developmental cognitive and behavioural disabilities. These include attention difficulties, verbal short term memory deficits and problems in phonological language processing [203-205]. These observations are consistent with the DPC having a functional role within the brain. Furthermore, it has suggested that DMD has a neuropathology reminiscent of schizophrenia i.e. frontal temporal distribution, cortical heteropias and reduced dendritic arborisation of pyramidal neurons [46, 206].

Nonetheless, although there is some evidence that dysbindin could confer susceptibility to schizophrenia through the interaction with the DPC, additional studies do not support this hypothesis. Firstly, while the power to detect an association was relatively low, a study which analyzed of a number of DPC genes for an association to schizophrenia, including β-dystrobrevin (DTNB), δ-sarcoglycan (SGCD) and ε-sarcoglycan (SGCE) produced negative results [207]. Secondly, although dysbindin and β-dystrobrevin co-localise in the mossy fibres of brain neurons, β-dystrobrevin is located postsynaptically [196]. *In vitro* studies have shown that dysbindin can be detected presynaptically in regions of the HF known to receive intrinsic glutamate input and importantly where β-dystrobrevin is not expressed [181]. As discussed previously, Talbot and colleagues reported a significant reduction in *presynaptic* dysbindin expression in several hippocampal regions [181], suggesting that dysbindin might influence schizophrenia risk through presynaptic mechanisms that are independent of the DPC.

1.6.3 Dysbindin and the BLOC-1 Complex

In addition to the DPC, dysbindin is also a component of the biogenesis of lysosome related organelles complex 1 (BLOC-1) [208]. This is a ubiquitously expressed complex, located at both post and crucially presynaptic sites. The BLOC-1 complex is approximately 200kDa and consists of at least eight proteins; dysbindin, muted, pallidin, cappuccino, snapin and BLOC-1 subunits 1, 2 and 3 [208-211]. Within BLOC-1, dysbindin has been shown to interact directly with pallidin, muted and snapin [212] (See Figure 1.9). This interaction has been shown to be facilitated through the coiled coil region of the dysbindin protein [212].

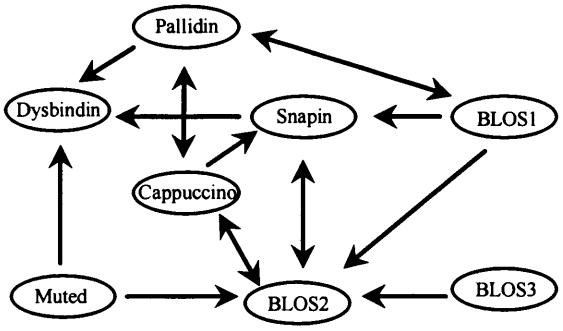


Figure 1.9. Illustration of the reported interactions between BLOC-1 proteins. Adapted from[211].

Although the exact molecular function of BLOC-1 remains unknown, evidence suggests that the complex is involved in the formation, development or maintenance of lysosome related organelles. Mutations in a number of the BLOC-1 genes, including DTNBP1, have been shown to cause various forms of Hermansky-Pudlak syndrome (HPS). HPS is the general term given to a collection of related autosomal recessive disorders which are characterised by the hypopigmentation of hair, skin and eyes and prolonged bleeding time [213]. These symptoms are caused primarily by defects in intracellular protein trafficking which results in the dysfunction of melanosomes and platelet dense granules. Melanosomes and platelet dense granules are intracellular organelles which act as catabolic compartments within the cell [214]. While melanosomes are involved in melanin synthesis and storage, platelet dense granules

have a role in the activation of platelet aggregation. Both organelles are referred to as lysosome-related organelles due to genetic and morphological evidence that supports the idea that their biogenesis pathway is analogous to that of lysosomes [211, 214].

Other lysosome-related organelles include azurophil granules (in neutophils), lytic granules (in cytotoxic T lymphocytes and natural killer cells) and lamellar bodies (in epithelial cells) [211]. Interestingly additional manifestations, such as pulmonary fibrosis [215] and defective lysosomal enzyme secretion [216], which can be ascribed to defects in lysosome or related organelles other than melanosomes or platelet dense granules [211], have been observed in subsets of HPS patients and HPS mouse models. This suggests that BLOC-1 genes may function in a more general way than simply regulating the biogenesis of melanosomes and platelet dense granules. This hypothesis is supported by the fact that BLOC-1 genes are expressed in a wide variety of cell types [208, 217].

Apart from BLOC-1 there are four other distinct complexes which are thought to be involved in the biogenesis of lysosome-related organelles, the AP-3 complex, BLOC-2, BLOC-3 and the HPS (homotypic vacuolar protein sorting) complex. Mutations of genes within these complexes cause similar phenotypic defects to the BLOC-1 genes and analysis of these complexes, in particular AP-3, has provided further evidence of the cellular function of the BLOC-1 complex. In fibroblasts the AP-3 complex traffics lysosome-associated membrane proteins (LAMPs) from early-associated tubules to late endosomes and lysosomes. In melanocytes AP-3 mediates the trafficking of the key melanogenic enzyme, tyrosinase, to maturing melanosomes. BLOC-1 has been shown to reside on microvesicles that contain AP-3 [218]. In addition the BLOC-1 complex can facilitate the trafficking of known AP-3 cargoes through interaction with AP-3 and BLOC2 [218, 219]. It has therefore been suggested that, like AP-3, BLOC-1 may function in membrane protein sorting.

The analysis of another BLOC-1 protein SNAPAP suggests that dysbindin may also have a specific role within the brain. SNAPAP (chr1:153,631,145-153,634,325) encodes the 15kDa protein snapin which is a binding protein of SNAP25 (synaptosomal associated protein). SNAP25 belongs to the family of soluble N-ethylmaleimide-sensitive-factor attachment protein receptor (SNARE) proteins. The primary role of SNARE proteins is to mediate the fusion of cellular transport vesicles with the cell membrane. SNAP25 is one the best studied SNARE proteins and is involved in the fusion or exocytosis of synaptic vesicles with the plasma membrane which results in the release of neurotransmitters. SNAREs can be divided into two categories, vesicle or v-SNARES, which are incorporated into the membranes of transport vesicles during budding and target or t-SNARES, which are located in the membrane of target compartments and the plasma membrane. SNAP25 is a t-SNARE and along with syntaxin (another t-SNARE) and a v-SNARE synaptobrevin, also known as VAMP (vesicle associated membrane protein), form the SNARE complex. Correct assembly of this complex is required for Ca^{2+} dependent exocytosis and Ca^{2+} dependent regulation of the SNARE fusion machinery is provided by the synaptotagmins. Investigations into the function of SNAPAP provided evidence that the gene may be an important regulator of neurosecretion as snapin has been shown to enhance the association between synaptotagmin and the SNARE complex [220]. Corroborative evidence of this has been found in SNAPAP mutant mice, where the association between synaptotagmin-1 and SNAP25 in brain homogenates was markedly decreased compared to the wild type mice, and in cultured cells where the knockdown of SNAPAP led to a significant reduction of vesicles residing in releasable pools [221].

1.6.4 The Sandy (sdy) Mouse

Alterations in the expression levels of dysbindin shown in schizophrenic patients [181, 183], could potentially disrupt the assembly and function of BLOC-1, thereby affecting intracellular lysosomal trafficking and possibly neurotransmitter release. To date, no major psychiatric manifestations have been documented in HPS patients, including type 7 which is caused by mutant dysbindin [208]. However behavioural analyses of the sandy mouse mutant, which lacks dysbindin owing to a deletion in the DTNBP1 gene, report a number of behavioural abnormalities in the mouse which have been associated with schizophrenia [222-226]. Morphological analyses of the sandy mouse [224, 225, 227, 228], plus several *in vitro* studies [135, 229-231] have also provided evidence of potential functional mechanisms.

The sandy (sdy) mouse mutant, which originally arose on the DBA12J inbred strain, carries a spontaneously occurring deletion of the DTNBP1 gene. This large deletion (38,129bp) spans intron 5 to intron 7 and results in the loss of 51 amino acids from the dysbindin protein and the complete knockdown of dysbindin expression (see Figure 1.10).

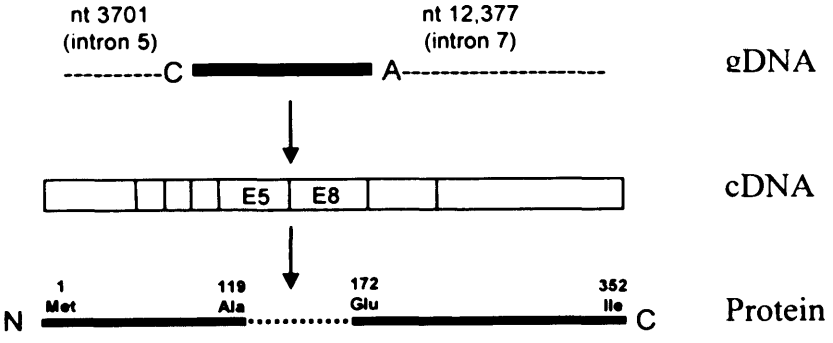


Figure 1.10. Dysbindin deletion in the Sdy mouse. nt=nucleotide. E=exon. Adapted from [208].

Recent behavioural assessments of the sdy mouse have observed a number of schizophrenia-like characteristics including cognitive and negative symptoms. Two studies which analysed the social interaction of sdy mice found they displayed reduced

social contact compared to wild type mice [224, 225]. This suggests that knockout of the DTNBP1 gene leads to social withdrawal which mimics negative symptoms exhibited in schizophrenia. Furthermore Hattori *et al* observed anxiety-like symptoms in the sdy mouse including decreased arm entries into an elevated maze and less time spent in the centre of an open field [225]. However other studies using similar tests failed to find a significant difference between sdy mutants and control mice [223, 224, 226]. Conflicting results have also been observed for locomotor activity. Hyperactivity is considered to be a proxy for positive symptoms displayed in schizophrenia such as delusions and hallucinations [225]. While studies by Bhardwaj and Cox [222, 223] both observed hyperactivity in sdy mice, two other studies reported a decrease in locomotor activity [225, 226] and one study found no significant difference between sdy mice and controls [224].

Although there are some disagreements between publications, all studies which tested for habituation found that sdy mice showed reduced habituation compared to control mice [222, 223, 225, 226]. It has been suggested that the impaired habituation of sdy mice resembles the decreased habituation to diverse stimuli reported in schizophrenia [223].

Studies which tested for cognitive function in sdy mice have also all reported deficits including reduced working memory [226], impaired long term recognition [222, 224] and impaired spatial memory [223, 226]. Deficits in cognitive functions related to the prefrontal cortex and hippocampus/medial temporal lobe such as recognition memory and other forms of declarative memory have been described in schizophrenia subjects. In addition, mice heterozygous for the DTNBP1 deletion also displayed significant impairments in some behavioural tests such as recognition memory [222] indicating that not only complete absence but also smaller variations in the expression of dysbindin, as seen in schizophrenia patients, can generate behavioural deficits related to schizophrenia and other psychotic disorders.

1.6.5 Putative Function of Dysbindin

The characteristics observed in the sdy mouse, such as hyperactivity, reduced habituation plus spatial learning and memory deficits are consistent with dysfunction of the hippocampal formation [223]. For example spatial learning deficits similar to those observed in the sdy mouse have also been shown in mice with hippocampal legions [232]. More specifically, evidence suggests that the dentate gyrus and mossy fibre terminus play important roles in working memory and long term memory retention [233]. The dentate gyrus is activated in spatial working memory tasks [234] and studies have reported a highly positive correlation between spatial working memory performance and the size of the mossy fibre terminals [235, 236]. Although abnormalities in the dentate gyrus have not been reported in the sdy mouse, dysbindin could play a critical role in memory disturbance in schizophrenia via these regions of the hippocampal formation. Dysbindin is expressed at high levels in both the dentate gyrus and mossy fibres [181, 237]. Moreover, as previously discussed, dysbindin levels have been shown to be reduced in both the dentate gyrus and the mossy fibre terminus of patients with schizophrenia [181-183].

Further evidence suggests that the behavioural deficits observed in sdy mice and potentially schizophrenic patients could, at least in part, be due to a dysfunction in neurotransmitter release. Dysbindin is located in synaptic vesicles [237] which share common features with lysosome related organelles [238]. As discussed above, SNAPAP, another BLOC-1 gene is involved in synaptic vesicle priming through its interaction with the SNARE complex [220, 221]. A recent study by Chen *et al* [227] reported that sdy mice exhibit abnormal neurotransmitter release and that this is likely to be caused by abnormal vesicle priming and fusion. A follow-up study by the same group [224] observed a 25% reduction of the steady state levels of snapin in the sdy mouse ($p<0.05$). In contrast overexpression of dysbindin in primary cortical neuronal cells induces the expression of two members of the SNARE complex, SNAP25 and synapsin1 [135]. As dysbindin has been shown to directly interact with snapin both in

humans [211, 212, 237] and in mice [224] it has been hypothesised that dysbindin has an upstream regulatory role on neurotransmitter release via an interaction with snapin which facilitates correct vesicle priming and stabilises ready releasable pools [224].

In vitro studies suggest that dysbindin, through this role in neurotransmitter release, may modulate the secretion of glutamate and/or dopamine. Abnormal transmission of both of these neurotransmitters has already been implicated in the neuropathology of schizophrenia [51, 239, 240] and the decreased habituation shown in mice and schizophrenic patients may indicate a problem with dopaminergic or glutamate processing [222].

Talbot and colleagues were first to provide evidence of a role for dysbindin in glutamate neurotransmission [181]. They reported that the presynaptic reduction in dysbindin protein, observed in patients with schizophrenia, was specifically located within regions of the HF known to receive intrinsic glutamate input. In addition, the reduction in dysbindin was inversely correlated with increased expression of vesicular glutamate transporter 1 (VGlut-1), the main vesicular glutamate transporter present in the HF [241]. This inverse relationship suggests an effect of dysbindin on VGlut-1 expression, synthesis or degradation within the HF.

The hypothesis of a role for dysbindin in glutamate neurotransmission is also supported circumstantially by the demonstration that overexpression of dysbindin increases extracellular basal glutamate levels and glutamate release while the knockdown of endogenous dysbindin protein by siRNA results in a reduction of glutamate release [135]. Deficiency in dysbindin could therefore cause glutamatergic dysfunction in the dentate gyrus (DG) and mossy fibres which in turn could underpin the cognitive deficits related to the DG in both schizophrenia and sdy mice.

In addition to being located within glutamatergic regions of the HF, dysbindin is also highly expressed in dopaminergic nuclei [183, 231]. Cell culture studies have shown

that knockdown of DTNBP1, or its BLOC-1 binding partner MUTED, can affect dopamine D2 receptor internalisation and signalling [229]. It is therefore interesting to note that like sdy, mice overexpressing D2 receptors show working memory deficits [242]. Consequently it been hypothesised that dysbindin, via its role in the BLOC-1, may regulate the recycling of dopamine D2 receptors in post synaptic targets of dopaminergic synapses [226]. However the role of dysbindin in dopamine neurotransmission still remains unclear. Although sdy mice have been shown to have lower levels of dopamine in the cerebral cortex, hippocampus and hypothalamus compared to wild type mice [225, 228], cell culture experiments in PC12 cells suggests a decrease in dysbindin expression causes an increase in dopamine release [231].

Overall cell culture studies and biochemical analysis of the sdy mouse mutant suggests several putative functional mechanisms by which dysbindin could confer susceptibility to schizophrenia. While these hypotheses are supported by behavioural analysis of the sdy mouse, it must be noted that similar behavioural effects can be seen in other mice mutants following the disruption of genes thought to have little to do with schizophrenia [226]. Notwithstanding the phenotypes of the sdy mice and cell culture studies provide evidence that dysbindin may have a role in neurotransmitter release and more specifically dopamine and glutamate processing.

1.7 Thesis Aims and Objectives

Identifying DTNBP1 causal variants would confirm DTNBP1 as a schizophrenia susceptibility gene and allow more accurate predictions as to the nature of the pathogenic function of these variants. It would also help determine the true magnitude of effect of DTNBP1 on the susceptibility to schizophrenia.

As previously discussed, current data suggests that dysbindin could promote susceptibility to schizophrenia through altered gene expression. In an attempt to link the genetic association data with the post mortem findings, Bray and colleagues identified a DTNBP1 risk haplotype correlated with reduced allelic expression [186]. However this risk haplotype does not account for all of the reduced expression and it is therefore highly possible that some or all of the cis-acting functional variants that cause altered dysbindin expression have not been identified. It is plausible that SNPs within the correlated risk haplotype may not affect dysbindin expression at all and are in LD with the actual functional variants or that there are unidentified functional variants that work in conjunction with SNPs within the risk haplotype to alter DTNBP1 expression.

Consequently one aim of the research presented within this thesis was to refine the findings of Bray *et al* and determine DTNBP1 causal variants by identifying and characterising both DTNBP1 schizophrenia risk variants and putative regulatory polymorphisms. This included identifying putative DTNBP1 regulatory regions and the sequence variants within these regions. Detected polymorphisms were then subjected to association analyses with both schizophrenia and DTNBP1 expression differences. Potential regulatory variants were verified using an *in vitro* luciferase expression assay.

In addition to identifying DTNBP1 functional variants, the biological pathway by which dysbindin could confer susceptibility to schizophrenia was also investigated further. Firstly, an attempt was made to replicate any previous positive associations reported of the BLOC-1 genes. Secondly this research set out to determine whether any other BLOC-1 genes like DTNBP1 are under the influence of cis-acting variation. If

evidence of differential expression was observed these genes were subjected to association analysis with schizophrenia.

Chapter 2: Materials and Methods

2.1 DNA samples

2.1.1 UK Schizophrenia Case Control Association Sample

All subjects were unrelated, Caucasian, resident in the British Isles, and had provided written informed consent to participate in genetic studies. Protocols and procedures were approved by relevant ethical review panels including the UK Wales Multi-centre Research Ethics Committee (Cardiff, Wales).

All 709 cases met DSM-IV criteria for schizophrenia and consisted of 483 males and 226 females. The mean age at interview was 44.5 years ± 14.6 years. Diagnosis was made by two raters from all available information following a semi-structured interview, Schedules for Clinical Assessment in Neuropsychiatry (SCAN) or Present State Examination (PSE) [17, 243] plus examination of case notes. Formative team reliability meetings took place weekly throughout recruitment.

716 control individuals (482 males, 234 females, mean age 41.5 years ± 11.5 years) were group matched to cases for age, sex, and ethnicity from more than 1400 blood donors from the British Blood Transfusion Service. Controls were not specifically screened for psychiatric illness but individuals were not taking regular prescribed medications. In the UK, blood donors are not remunerated even for expenses and are not over-represented for indigents or the socially disadvantaged in whom the rate of psychosis might possibly rise above a threshold that would influence power [88, 244].

2.1.2 Caucasian Brain Samples

Genomic DNA and total RNA was extracted from post-mortem brain tissue (frontal, temporal or parietal cortex) of 149 unrelated, anonymised Caucasians (86 males, 63 females; mean age = 58, standard deviation (SD)=19). Of these, 86 had received no psychiatric or neurological diagnosis at the time of death, 22 had a diagnosis of Alzheimer's disease, 12 had a diagnosis of schizophrenia, 14 had a diagnosis of bipolar disorder and 15 had a diagnosis of major depression. Details such as the cause of death and post-mortem interval were also recorded. These samples were obtained from four sources (The MRC London Neurodegenerative Diseases Brain Bank, UK; The Stanley Medical Research Institute Brain Bank, Bethesda, USA; the Mount Sinai School of Medicine, New York, USA and The Karolinska Institute, Stockholm, Sweden).

2.1.3 Mutation Screening Sample

Mutation screening was performed using a sample of 14 unrelated schizophrenic subjects selected from the UK case control sample described in section 2.1.1. Individuals were chosen that met DSM-IV criteria for schizophrenia and had at least 1 affected sibling. 14 unrelated individuals from the same population allows 95% power to detect alleles with a MAF>0.1 and 80% power to detect alleles with a MAF of 0.05.

2.2 DNA/RNA Extraction and Quantification

2.2.1 DNA Extraction and Storage

All DNA samples used in this study were extracted by members of the Department of Psychological Medicine within Cardiff University. For allelic expression assays (section 2.9) genomic DNA (gDNA) was extracted from neuronal tissue by Dr. Nicholas Bray.

All other gDNA samples were obtained as high molecular weight DNA fractions from either lymphocytes in venous whole blood or from buccal cavity epithelial cells via saline mouthwash. Each sample was prepared via standard phenol/chloroform DNA extraction followed by ethanol precipitation. Stock and diluted samples were stored in water at -20°C.

2.2.2 RNA Extraction and cDNA synthesis from Total RNA

Neuronal tissue RNA extraction was performed by Dr. Nicholas Bray using the Ambion RNAqueous-Midi kit. cDNA was synthesised from total RNA (DNase treated) by reverse transcription (RT) using a RETROscript kit (Ambion). All RTs were performed by either Dr. Nicholas Bray or Dr. Liam Carroll as described in the RETROscript guidelines. In order to detect any genomic DNA contamination RNA samples underwent PCR amplification prior to complementary DNA (cDNA) synthesis.

2.2.3 Spectrophotometer Quantification

Extracted DNA was initially quantified using a Beckman DU 640B spectrophotometer (Beckmann Instruments). Each DNA sample was diluted to a 5% solution in sterile water (i.e. 5µl DNA sample in 95µl of water). The absorbance (A) of UV light at 260nm and 280nm wavelengths (λ) was measured, and DNA concentrations were

calculated on the assumption that an A_{260nm} value of 1 was equivalent to 50μg of DNA. A ratio of A_{260nm} to A_{280nm} above a value of 1.8 indicated a suitable level of DNA and the absence of contaminating protein.

RNA concentrations were measured by Dr. Nicholas Bray using an Amersham Ultrospec 2100 pro UV/Visible spectrophotometer.

2.2.4 Pico Green DNA Quantification

A more accurate quantification of DNA samples was performed using a Fluoroskan Ascent fluorometer (Thermo Labsystems) and pico green (Invitrogen). Samples were first diluted to less than 50ng/μl based on spectrophotometer readings in a 100μl volume. Aliquots of the samples were then diluted with 1X TBE in a white 96 well cliniplate so that the final DNA concentration was 0.6-1.2ng/ul. A pico green working solution was prepared in parallel by adding 5μl of pico green to 995μl of 1X TE.

In order to measure DNA concentration, 100μl of the pico green working dilution was dispensed into each diluted sample. The fluorometer measures the concentration of a sample using an UV excitation wavelength of 485nm and an emission wavelength of 538nm. A standard curve (prepared by Dr. Nadine Norton) was then used to calculate the concentration of DNA for each sample. The original samples were then adjusted so that each was at a concentration of 4ng/ml ± 0.5ng.

2.3 Polymerase Chain Reaction

The polymerase chain reaction (PCR) is an enzymatic *in vitro* cycling technique for the amplification of a specific region of DNA that lies between two regions of known sequence. Thermostable Taq polymerase enzyme was used which synthesises a complementary strand from the DNA template in the presence of suitable buffers and a mix of adenine (abbreviated A), cytosine (C), guanine (G) and thymine (T)

deoxyribonucleotide triphosphates (dNTPs). Two oligonucleotide primers are designed to flank the specific region of DNA to be amplified. These primers provide the double stranded starting point for Taq polymerase to begin 5' to 3' synthesis. A PCR reaction is comprised of three steps, a denaturation step which produces a single stranded DNA template, a primer annealing step where the primers bind their complementary sequence and an elongation step when the synthesis of DNA occurs. Each step is accompanied by controlled temperature changes and there are typically 30-45 cycles per reaction.

2.3.1 PCR Primer Design

PCR primers were designed *in silico* using the Primer 3 web resource (http//www-genome.wi.mit.edu/cgi-bin/primer/primer3-www.cgi). If possible PCR primers were designed using the default Primer 3 settings of an average length of 20bp, an annealing temperature of ~60°C and a GC content less than 80%. In general, and specifically where PCR products were required to be sequenced, amplimeres were restricted to <500bp.

2.3.2 PCR Optimisation

All PCRs were performed on MJ thermocyclers. Three types of Taq polymerase were used in this study: HotStarTaq (Qiagen), Titanium Taq (BD Biosciences) and Expand High Fidelity Taq (Roche). Titanium Taq was used for Amplifluor based genotyping (see section 2.7.3), Expand High fidelity Taq was used for amplification of regions for cloning (section 5.2.2.2). HotStarTaq (Qiagen) was used in PCRs for mutation detection (section 2.5), sequencing (section 2.6) and genotyping (section 2.7). As the exact PCR conditions and reagents were dependant on the specific methodology, the optimised PCR conditions are given in the relevant sections below.

2.4 Agarose Gel Electrophoresis

The negative phosphate groups within DNA allow DNA fragments to be separated electrophoretically. When a potential difference is applied through a porous substance such as an agarose gel, DNA will move towards the anode, at a rate dependent on the fragment size. Analysis of pre or post PCR samples was performed using 1-2% agarose gels, depending on the fragment size and resolution required.

To construct a 1% gel, 1g of agarose (Sigma-Aldritch) was dissolved in 100ml 0.5x TBE buffer (Ultra pure electrophoresis grade, National Diagnostics). The solution was heated until it became clear. Once the solution had cooled slightly, 1µl Ethidium Bromide solution (10mg/ml) was added. This solution was then poured into a gel-former with appropriate gel combs added and left to cool until it formed a solid.

In order to run a specific sample in a gel, each PCR product was mixed with loading buffer. 6x loading buffer was made by creating a solution of 15% ficoll, 0.25% bromophenol blue, 0.25% xylene cyanel in water. An appropriate volume of PCR product was mixed with loading buffer and pipetted into a formed well. 3µl of size standard (for example 1kb plus DNA ladder, Invitrogen) was also run alongside samples to allow size comparison. Each gel was run at between 100-120V in an electrophoresis tank for the appropriate amount of time needed to see the DNA size expected. Samples were visualised using a UV transilluminator (UVP) and photographs taken using Kodak Electrophoresis Gel analysis system.

2.5 Mutation Detection

Mutation detection was performed using either Denaturing High Performance Liquid Chromatography (dHPLC) or High Resolution DNA Melting Analysis (HRMA). Details of both are given below. The PCR products of any individuals showing alternative profiles detected by either dHPLC or HRMA were then sequenced in order to identify DNA variants.

2.5.1 Denaturing High Performance Liquid Chromatography

Denaturing High Performance Liquid Chromatography (dHPLC) is an automated technology based on the separation of heteroduplex PCR products from their corresponding homoduplexes via an ion-pair reversed-phase liquid chromatography system. The hydrophobic stationary phase consists of alkylated nonporous poly(strene/divinylbenzene) particles and the mobile phase consists of triethylammonium acetate (TEAA) and acetonitrile (ACN). The column is maintained at a set temperature to partially denature DNA molecules. Under these conditions, heteroduplexs attributable to mismatch pairing will form weaker interactions with the hydrophobic matrix. Due to the use of a linear acetronitrile gradient, heteroduplexes are eluted earlier than homoduplexes. Samples containing a heterozygous mutation form both homoduplexes and heteroduplexes, and (theoretically at least) two peaks are observed (Figure 2.1B). In reality, resolution is not optimal and it is more common to see one peak with a "shoulder" (Figure 2.1C). Samples that give a single peak (Figure 2.1.A) could either be wild type sequence or contain a homozygous mutation.

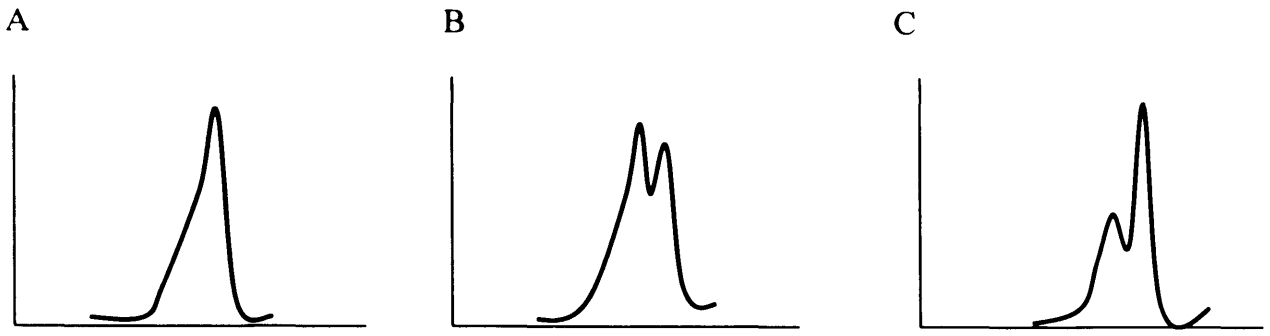


Figure 2.1. Representation of dHPLC traces (Absorbance vs. Retention Time) for A. homozygote B. heterozygote (theoretical) and
C. heterozygote (realistic) samples.

2.5.1.1. Heteroduplex Formation and Analysis Parameters

For each fragment the 14 DNA samples of the mutation screening sample were PCR amplified in a 24µl PCR reaction. This consisted of 4µl of genomic DNA (4ng/µl), 1.12µl of each primer (5pmol), 1.92µl dNTPs (5mM each), 2.4µl 10x buffer, 0.12µl HotStarTaq (10U/µl) and 14.44µl of water. If required 1.44µl of Dimethyl sulfoxide (DMSO) was added to the PCR reaction and the volume of water reduced. DMSO is used in PCR to inhibit secondary structures in the DNA template or the DNA primers. Each DNA fragment amplification was optimised to one of the four PCR cycling conditions outlined below (A,B,D,E).

A

1. 95°C for 15 minutes
2. 94°C for 5 seconds
3. 56°C for 5 seconds
   -0.5 per cycle
4. 72°C for 10 seconds
5. Go to step 2 for 11 cycles
6. 94°C for 5 seconds
7. 50°C for 5 seconds
8. 72°C for 10 seconds
9. Go to step 6 for 22 cycles
10. 72°C for 10 minutes
11. 15°C forever

B

1. 95°C for 15 minutes
2. 94°C for 5 seconds
3. 61°C for 5 seconds
   -0.5 per cycle
4. 72°C for 10 seconds
5. Go to step 2 for 11 cycles
6. 94°C for 5 seconds
7. 57°C for 5 seconds
8. 72°C for 10 seconds
9. Go to step 6 for 22 cycles
10. 72°C for 10 minutes
11. 15°C forever

D

1. 95°C for 15 minutes
2. 94°C for 20 seconds
3. 50°C for 30 seconds
4. 72°C for 20 seconds
5. Go to step 2 for 34 cycles
6. 72°C for 10 minutes
7. 15°C forever

E

1. 95°C for 15 minutes
2. 94°C for 20 seconds
3. 55°C for 30 seconds
4. 72°C for 20 seconds
5. Go to step 2 for 34 cycles
6. 72°C for 10 minutes
7. 15°C forever

To form heteroduplexes, PCR products were heated to 94°C and then gradually reannealed by cooling at a rate of 1°C for 40minutes. Optimal temperatures and corresponding elution gradients for each PCR fragment were selected using DHPLC Melt (http://insertion.stanford.edu/melt.html). In addition to the temperature suggested by the software (n°C), each fragment was also run at n+2°C to ensure maximum sensitivity.

2.5.1.2. dHPLC Analysis

dHPLC analysis was performed using the WAVE™ DNA Fragment Analysis System (Transgenomic) [245]. 5µl of heterodulpexed PCR product was injected onto a DNASep column. Hetero- and homoduplexes were then eluted with a linear acetonitrile gradient formed by mixing buffer A (0.1 TEAA, pH 7.0) and buffer B (0.1 TEAA, pH 7.0, containing 25% ACN) at a constant flow rate of 0.9ml/min. DNA was detected at 260nm. The analytical gradient was 4 minutes long with the concentration of buffer B increased at 2% per/min. For each fragment the initial and final concentrations of buffer B were adjusted to obtain a retention time between 3 and 5minutes. Between samples, the column was cleaned with 100% buffer B for 30seconds and equilibrated at starting conditions for 2 minutes. When all samples had been processed, the resultant chromatograms were compared, with a shift in trace pattern indicative of a heteroduplex.

2.5.2 High Resolution DNA Melting Analysis

High Resolution DNA Melting Analysis (HRMA) is based on the observation that the melting temperature (Tm) of a PCR amplimere can be largely dependent on its specific sequence composition [246]. By slowly melting a PCR amplimer in the presence of a suitable fluorescent dye, which binds specifically to double stranded DNA (dsDNA), it is possible to monitor the amplimere's melting curve via the change in fluorescence as the dye is released. When compared with a wild type sequence, the presence of both homo and heteroduplexes (caused by PCR products with heterozygous loci) can generate detectable changes in the shape of the melting curve [247]. The fluorescent

dye LC green is particularly suited to HRMA because it can be used at concentrations high enough to saturate the dsDNA binding sites during PCR without inhibiting Taq polymerase [248]. Saturation of the dsDNA reduces the potential of dye molecules released during HRMA being redistributed to dsDNA. This increases the sensitivity of the HRMA to detect subtle changes in fluorescence and LC green can efficiently detect single nucleotide variants in PCR products [249].

2.5.2.1. HRMA PCR

PCRs were performed in a 12µl reaction using 4µl of genomic DNA (4ng/µl), 0.56µl of each primer (5pmol), 0.96µl dNTPs (5mM each), 1.2µl of 10x LCgreen Plus (Idaho Technologies) and 1.2µl of 10X LCgreen Plus PCR buffer (20mM MgCl, Idaho Technologies) and 0.06µl of HotStarTaq polymerase (10units/µl, Quiagen). Fragments with a GC content of >60% were amplified in the presence of 10% (1.2µl) DMSO.

The PCR cycling parameters were as follows:

1. 95°C for 10 minutes
2. 94°C for 20 seconds
2. 56-66°C for 30 seconds (depending on amplimer)
3. 72°C for 1 minute
4. Go to step 2 for 44 cycles
5. 72°C for 10 minutes
6. 15°C forever

2.5.2.2 Mutation Detection by HRMA

HRMA was performed according to the manufactures instructions: each 12ul sample was denatured by increasing the temperature to 98°C at a rate of 0.1°C/s with fluorescent data points being acquired continuously at a rate of 14 points/°C.

Melting profiles were analysed using a semi-automated analysis [250]. This involved normalising the melting curves by manually defining the temperature interval before and after the major change in fluorescence that corresponds to 100% and 0% fluorescent respectively. The samples were then analysed using the Lightscanner HRMA software Call-IT™ (Idaho Technologies) using the high sensitivity setting. The automatic calls of the software were inspected by the user and manually clustered according to the similarity of the 'difference curve' plots.

2.6 Sequencing

PCR products from individuals showing alternative melt profiles, either by dHPLC or HRMA (and therefore suggestive of heteroduplex formation) were sequenced in both directions using the fluorescent Sanger sequencing method via Big-Dye termination chemistry.

The fluorescent sequencing reaction involves the incorporation of four fluorescently labelled dideoxynucleotides (ddATP, ddCTP, ddGTP, ddTTP) in addition to unlabelled dNTPs. Unlike dNTPs, ddNTPs terminate after extending one base during a primer extension reaction. After an appropriate number of cycles, such a reaction produces a series of DNA fragments which have been terminated at each successive base position. When these fragments are electrophoresed in a capillary sequencer, such as the ABI3100, each base of the sequence will be fractionated by size and under laser detection, fluoresce according to the base at that site.

In order to reduce errors and improve consistency an Agencourt semi-automated protocol was employed for the clean up of PCR and Sequencing products using the Beckman-Coulter NX liquid handler.

2.6.1.1 PCR Clean-up

PCR clean-up is needed when the product to be sequenced has been amplified via a PCR reaction as it removes unincorporated dNTPs, primers, DNA polymerase and salts. If a mini-prep product is to be sequenced this step can be omitted. The PCR product (10µl) is mixed with 21.6µl of AMPure reagent (Agencourt). This reagent contains magnetic beads which adhere to the DNA. The products not fixed to the beads are removed by successive 85% ethanol wash steps. The PCR amplimeres are then eluted in 195µl H20 in a new 96-well skirted plate.

2.6.1.2 Sequencing Reaction

5µl of the cleaned PCR or mini-prep product ws added to a 5µl sequencing reaction mix which consists of: 1.917µl 5X BigDye sequencing buffer, 0.116µl BigDye termination mix, 1µl of either the forward or reverse PCR primer (4pmol/µl) and 1.917µl H20. The BigDye reaction mix contains the four fluorescently labelled ddNTPs, unlabelled dNTPs and a Sequenase enzyme. The sequencing reaction was performed on a MJ thermocycler using the following conditions:

1. 96°C for 2 minutes
2. 96°C for 30 seconds
3. 55°C for 15 seconds
4. 60°C for 4 minutes
5. Go to step 2 24 times
6. 4°C for 4 minutes.

2.6.1.3 Post-sequencing Clean-up

The post-sequencing clean-up removes any unwanted impurities from the sequencing reaction such as unincorporated ddNTPs. The post-sequencing clean-up involves a CleanSEQ chemistry protocol which like the AmPure reagent used in the PCR clean-up contains magnetic beads. 10µl of sequencing product was added to 7.5µl of CleanSEQ reagent along with 36.39µl of 85% ethanol. As with the AMPure protocol the sequencing product binds to magnetic beads and the non-bound contaminants are removed by successive 85% ethanol wash steps. The cleaned sequencing product is eluted in 75µl of H_20 which can be read directly via a capillary sequencer.

2.6.1.4 Sequencing Analysis

Samples were run on the ABI3100 PRISM through a 36cm capillary using polyacrylamide POP6 (Applied Biosystems). The ABI3100 PRISM genetic analyser automatically analyses the raw data generated through electrophoresis using its Sequence Analysis Software (Applied Biosystems). This software calls each nucleotide based on the fluorescence at each base.

A combination of the sequence analysis software packages Sequencher (Gene Codes) and NovoSNP [251] were used to identify any polymorphisms within the amplimeres. Both packages align multiple sequencing traces and will highlight differences between the traces and/or a reference sequence. The user can then manually inspect these differences to judge whether a polymorphism exists.

2.7 Genotyping

The majority of genotyping was performed using the MassARRAY genotype platform. Where this failed or did not pass the stringent quality control (QC) checks SNaPshot or Amplifluor were attempted.

2.7.1 Sequenom MassARRAY

The Sequenom MassARRAY genotyping system allows the highly accurate genotyping of simple polymorphisms by combining iPlex GOLD primer extension chemistry with MALDI-ToF (Matrix Assisted Laser Desorption Ionisation – Time of Flight) Mass Spectrometry (MS). iPlex GOLD involves primer extension over the polymorphism of interest and the examination of the mass of the extended product to discern the genotype of the sample. Results are stored and analysed using the software Typer (Sequenom). The main advantage of this genotyping system is the high accuracy combined with a high multiplexing level (up to a 40-plex).

The initial step of MassARRAY genotyping involves the design of a multiplex assay using Sequenom Design Assay software. For each polymorphism the flanking DNA sequence is obtained and additional features of the sequence that may confound any assay are highlighted (for example known SNPs or repetitive sequence) to prevent assay design over these regions. From this information the Sequenom Assay Design software designs PCR and extension primers to the highest multiplex level possible. Details of these are provided in the design output file. In order to ensure extension peaks are detected at the MALDI-Tof MS stage, the design software may add a non-specific sequence to the extension primer. A 10bp non-specific tag sequence is also added to the 5' end of the PCR primers to ensure they are detected in the MADLI-ToF MS spectrum.

2.7.1.1 Sequenom PCR and Extension

The software designs the PCR primers so as to create the shortest amplimere possible that will allow efficient PCR at an annealing temperature of 56°C. This allows a universal PCR condition to be used. Each PCR is performed with 3µl of dried genomic DNA (4ng/µl) in a 384 microtitre plate (ABgene) with the addition of a 5µl PCR mix. This PCR mix consists of: 3.35µl water, 0.625µl 10X PCR buffer, 0.323µl MgCl2 (25mM), 0.1µl dNTPs (25mM), 0.1µl HotStarTaq and 0.5µl PCR primer mix (forward and reverse PCR primers at 1pmol/µl)

The following PCR is then performed:

1. 95°C for 15 mins
2. 94°C for 20s
3. 56°C for 30s
4. 72°C for 1 min
5. Repeat steps 2-6 for 44 cycles
6. 72°C for 3 mins
7. 15°C for 10 minutes

A number of genomic DNA positive control samples and negative control samples were electrophoresised on a 2% gel to check for both PCR efficiency and contamination. Should the assay pass this QC, a 2µl Shrimp Alkaline Phosphatase (SAP) mix was added to the 5µl PCR reaction. The SAP mix consists of:

SAP	0.3µl
SAP Buffer	0.17µl
Water	1.53µl

This 7μl reaction mix then undergoes the following thermocyclic conditions:

1. 37°C for 30 minutes
2. 85°C for 10 minutes
3. 95°C for 5 minutes
4. 15°C for 10 minutes

The extension reaction involves the addition of optimised concentrations of unextended extension primers, along with ddNTPs, to the 7μl PCR and SAP reaction product. The extension primer mix containing a mix of all unextended extension primers is defined by an optimisation procedure involving a small number of DNA samples. The extension primers are split into four groups dependent upon their mass (lowest to highest mass) which are diluted initially to final concentrations of 0.938μM, 1.17μM, 1.425μM and 1.875μM. The extension primers are divided in this way as lower mass products generate a lower signal to noise ratio when detected by MALDI-ToF. After an initial test run the extension primer concentrations are adjusted according to their peak height. For example if a peak height is low then the final concentration is increased. At the optimisation stage failed or abnormal assays (for example self priming assays) are also removed.

The 2μl extension mix consists of:

iPlex GOLD reaction buffer	0.2μl
iPlex GOLD termination mix	0.2μl
iPlex GOLD enzyme	0.041μl
Adjusted unextended primer mix	1.559μl

The extension reaction was then carried out as follows:

1. 94°C for 30 seconds
2. 94°C for 5 seconds
3. 52°C for 5 seconds
4. 80°C for 5 seconds
5. Go to step 3, 4 times
6. Go to step 2, 39 times
7. 72°C for 3 minutes
8. 4°C for 10 minutes

After the extension reaction, desalting of the solution using Clean Resin (Sequenom) was performed by the addition of 6mg of the resin using the Sequenom dimple plate followed by 25µl of water to the reaction mix. The reaction sample is then mixed on a rotor for ~1 hour. The resin removes all ions that may alter the spectra of the sample and therefore affect the subsequent analysis. After mixing the samples are spun in centrifuge for 15minutes at 3000rpm to separate the resin from the solution.

2.7.1.2 Sequenom Analysis

Samples are automatically spotted onto the Sequenom MassARRAY SpectroCHIP using a nanodispenser liquid handler (Sequenom). Each chip contains 384 spots which are composed of a combustible matrix (3-hydroxypicolinic acid) that allows ionisation of the product when excited by a laser [252]. Each ionised extended and unextended MassEXTEND primer product differs in mass and is therefore amenable to MALDI-ToF MS analysis using MassARRAY RT software (SpectroAcquire, Sequenom). The software estimates genotypes for each sample based upon the assay design output and certain parameters such as the peak heights (intensity of mass signal) of each allele and also the extension primer yield (successful extension of the unextended primer compared to residual unextended primer). These genotypes can then be viewed and manually revised by the user using the Typer software (Figure 2.2).

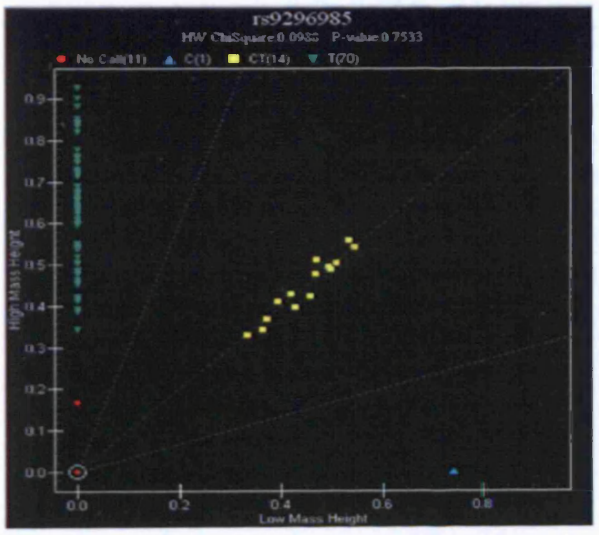


Figure 2.2. Screenshot from Typer analysis software (Sequenom) for rs9296985. TT homozygotes are shown in green, C homozygotes in blue and CT heterozygotes in yellow. No Calls are given in red.

2.7.1.3 Accurate Genotyping

All assays were initially optimised by genotyping DNA from 30 CEPH parent-offspring trios. All plates for genotyping contained a mixture of cases, controls, blanks, and 46 CEU samples. "Double-genotyping", where another experienced user of the Sequenom genotyping system and Typer software checks the genotypes for every assay, was used. Genotypes were called blind to sample identity, affected status, and blind to the other rater. Genotypes of CEU samples were compared to those available on the HapMap to provide a measure of genotyping accuracy. Genotyping assays were only considered suitable for analysis if a) during optimisation, genotypes for CEU individuals were the same as those in the HapMap when available and b) all subsequent duplicate genotypes from the CEU samples were consistent with the HapMap data.

2.7.2 SNaPshot

A polymorphism which varies at one particular nucleotide can be genotyped via oligonucleotide primer mediated extension of a single fluorescently labelled ddNTP using SNaPshot chemistry (Applied Biosystems). The SNaPshot reaction consists of the PCR of a sample of interest, which is then cleaned before primer extension by a single fluorescent ddNTP (ddATP, ddCTP, ddGTP, ddTTP) corresponding to the next 3' base (the polymorphic site of interest). This is followed by another clean-up step to remove excess ddNTPs and analysis using an ABI3100 PRISM Genetic Analyser (Applied Biosystems). Genotyping was performed by manual inspection of the extension peaks using Genotyper software (Appied Biosystems). Oligonucleotide extension primers were designed using the internet based algorithm FP primer designed by Dobril Ivanov (http://m034.pc.uwcm.ac.uk/FP_Primer.html.).

For individual SNaPshot genotyping, each sample underwent a standard 12µl PCR reaction using 3µl of genomic DNA (4ng/µl), 0.28µl of each primer (5pmol), 0.96µl dNTPs (5mM each), 1.2µl 10x buffer, 0.06µl HotStarTaq (10U/µl) and 6.22µl of water. The PCR cycling conditions are outlined below:

1. 94°C for 15 minutes
2. 94°C for 20 seconds
3. Tm°C for 20 seconds
4. 72°C for 30-45 seconds
5. Repeat steps 204 for 35-45 cycles
6. 72°C for 10 minutes
7. 15°C for ever

Tm°C was determined by the annealing temperature of each primer set.

Shrimp Alkaline Phosphatase (SAP) (Amersham) and exonuclease I (Amersham) was then added to each PCR product to degrade unincorportated dNTPs and unextended primers. The reaction involved the addition of a 5µl Sap mix described below to the 12µl PCR product.

SAP	0.5µl
Exonuclease I	0.1µl
Water	4.4µl

The reaction conditions were as follows:

1. 37°C for 1 hour
2. 80°C for 15mins
3. 15°C for ever

For SNaPshot primer extension an 8µl reaction mix consisting of 1.25µl SNaPshot reagent (containing fluor-labelled ddNTPs and a sequenase), 3.75µl reaction buffer, 2µl water and 1µl of extension primer diluted to the appropriate concentration, was added to 2µl of the cleaned PCR product.

Typically extension primers were used at a 0.5pmol/µl concentration, although this was altered in same cases to obtain optimum peak heights using the equation

$$\text{Concentration} = Y'/(YX)$$

where Y' is the required peak height (typically 3000 fluorescence intensity units as displayed by the Genotyper software (Applied Biosystems), Y is the initial peak height and X is the initial primer concentration [253].

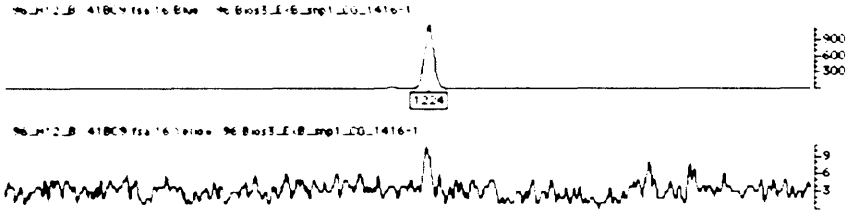
The following reaction was then performed:

1. 96°C for 2 minutes
2. 96°C for 5 seconds
3. 43°C for 5 seconds
4. 60°C for 5 seconds
5. Repeat steps 2-4 for 24 cycles
6. 15°C for ever

A further stage of SAP clean-up was then performed to degrade the unincorporated ddNTPs. A 5µl reaction mix comprising of 0.5µl SAP and 4.5µl water was added to the SNaPshot reaction product and the same conditions as the SAP PCR clean-up were performed. After this reaction, 3µl of the product was added to 10µl of HiDi formamide. The samples were then run through a 36cm capillary using POP4 polyacrylamide (Applied Biosystems). The raw data was analysed using the Genescan Analysis v3.7 software (Applied Biosystems) and imported into Genotyper software (Applied Biosystems)

The Genotyper software allows firstly the discrimination of correct genotypes and secondly the amount of each allele present in a sample. The latter is indirectly measured via the peak height of the fluorescence. The amount of each allele present is given as a numerical value (arbitrary absorbance units) and can be exported to an Excel file for further analysis. Individual samples are then genotyped based on the presence of fluorescence for the corresponding nucleotide (Figure 2.3). Along with genomic DNA, negative controls are added at the PCR and SNaPshot stages to check for contamination.

A



B

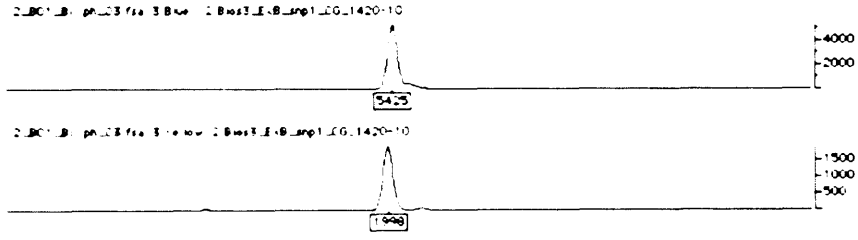


Figure 2.3. Individual genotyping via Snapshot using Genotyper software (Genecodes). A GG homozygote B GC heterozygote.

2.7.3 Amplifluor

The Amplifluor Uniprimer assay involves an allele specific PCR using a common anti-sense reverse PCR oligonucleotide primer and two sense forward primers which differ at their 3' ends corresponding to the complimentary base for each allele of a SNP. The forward primers also differ at the 5' end of the oligonucleotide, where each allele specific primer has a ~20bp stretch of nucleotides complimentary to two universally fluorescently labelled primers (Uniprimers). Each of the uniprimers is labelled either with a green or red fluorophore. However these do not fluoresce due to a hairpin structure that brings a quencher into contact with the fluorophore.

The unlabelled forward primers initiate a competitive allele-specific PCR. These allele specific amplimeres serve as templates for the binding of the labelled uniprimers. The incorporation of the uniprimer into the allele-specific amplimere displaces the hairpin structure of the uniprimer, releasing the fluorphore from its quencher and generating fluorescence. The resulting levels of red and green fluorescence can distinguish the levels of each allele in an individual sample and therefore the genotype.

The assays are designed using the internet based software Amplimfluor AssayArchitect (https://apps.serologicals.com/AAA). The design process requires the input of the SNP alleles with flanking DNA sequence. The software then designs two forward and one reverse primer.

For each Amplifluor reaction a SNP specific PCR primer mix was made which was added to the general PCR reaction mix. The primer mix consisted of:

Forward Primer (Allele 1)	100pmol/µl	2.5µl
Reverse Primer (Allele 2)	100pmol/µl	2.5µl
Reverse Primer	100pmol/µl	25µl
Water		470µl

This primer mix was then used in the PCR reaction mix as follows. All reagents except the primer mix are supplied by BD Biosciences.

10x Titanium Taq Buffer	0.5µl
dNTPs (2.5µl each)	0.4µl
Primer mix	0.07µl
SR labelled primer	0.07µl
FAM	0.07µl
Titanium Taq Polymerase	0.05µl
Water	3.84µl

For some assays 0.625µl of Reaction Mix S can be added (water reduced to 3.215µl) to improve the assay such as genotype clusters.

The total reaction volume of 5µl was added to 12ng of dried genomic DNA in a black 96 or 384 well microtitre plate (ABGene). The cycling conditions, for the Amplifluor reaction are as follows:

1. 96°C for 4 minutes
2. 96°C for 10 seconds
3. 58°C for 5 seconds
4. 72°C for 10 seconds
5. Repeat steps 2-4 for 20 cycles
6. 96°C for 10 seconds
7. 55°C for 20 seconds
8. 72°C for 40 seconds
9. Repeat steps 6-8 for 18-30 cycles depending on assay
10. 68°C for 7 minutes
11. 15°C for 10 minutes

The first stage of the reaction conditions (steps 1-5) involves the denaturation of the target DNA sample, the annealing of the allele specific primers and the elongation of the fragments which include the complimentary tails for the universal uniprimer. The second stage (6-11) involves the denaturation of the PCR product, the annealing of the fluorescently labelled primers to the complimentary sequence in the PCR product. When the fluorescently labelled primers bind to the complimentary sequence within the allele specific PCR products fluorescence occurs.

The fluorescence of each sample was analysed using an Analyst HTS Assay Detection Platform (LJT Biosystems) at the wavelengths shown below for each fluorophore:

FAM	Excitation at 485nm	Emission at 520nm
SR	Excitation at 580nm	Emission at 620nm

The results are given as the signal intensity for the two fluorophores which can be plotted on a graph program (https://apps.serologicals.com). The clusters of individual fluorescent points correspond to three genotype classes which can be assigned manually by the user (Figure 2.4). The output of the graph program is in the form of 11, 12, 22 where 1 corresponds to the FAM allele and 2 to the SR allele.

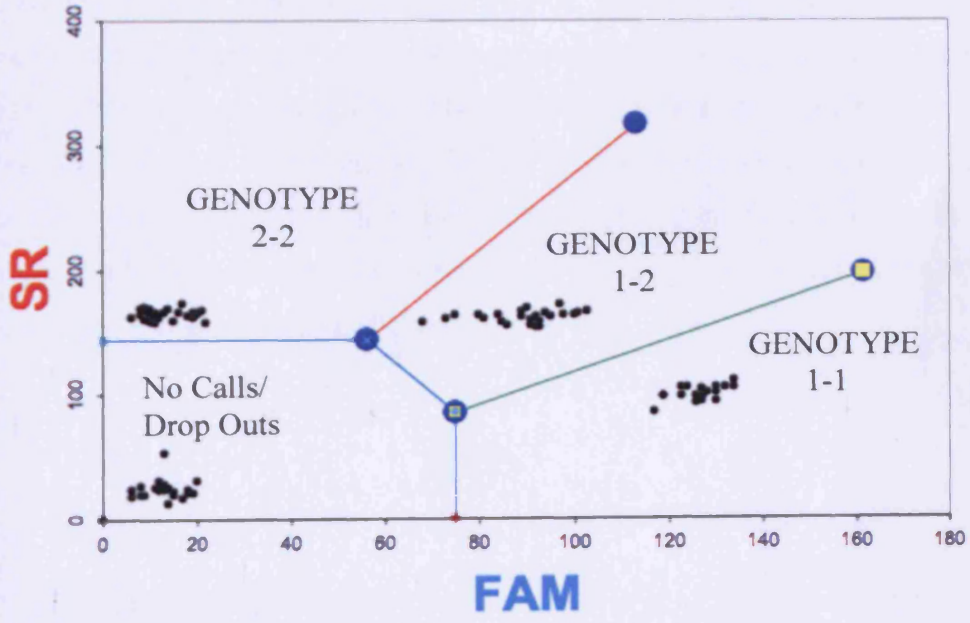


Figure 2.4. Cluster plot of individuals from an Amplifluor assay.

2.8 Sample Processing

For large scale PCR and post-PCR reactions involving many DNA samples, reagent master mixes and samples were aliquoted using robotic liquid handling systems. DNA samples were typically stored within shallow well DNA boxes (ABgene). These are compatible with the Beckman-Coulter FX and NX microdispensers. DNA samples were aliquoted into suitable (96 or 384 well) microtitre plates (ABgene). Both machines were also used to dispense reaction master mixes in the same manner. All programs for use with the Beckman-Coulter FX and NX were written by Sam Dwyer.

2.9 Relative Allelic Expression Assay

To assay gene transcripts for factors regulating steady-state mRNA levels in an allelic-specific manor (such as cis-acting factors) a quantitative assay of allelic expression has been developed in house [184]. The principle of the assay is that in any given tissue the relative level of steady state mRNA (as transcribed from each chromosome) should have a 1:1 ratio unless factors regulating the levels of the mRNA differentially affect the two autosomal transcripts. A transcribed heterozygote marker can be used to measure for such differences. PCR primers are designed which will encompasses the variant in question, and which will amplify the genomic DNA (gDNA) and complimentary DNA (cDNA) equally (i.e. primers are within an exon boundary). Any divergence of the 1:1 allelic ratio of gDNA observed in cDNA samples is indicative of an allele-specific regulation of steady-state mRNA levels in the tissue examined, such as a cis-acting regulatory polymorphism regulating transcription.

To study specific genes for allelic expression differences, polymorphisms were selected that were located within predicted mRNA sequence and had a high minor allele frequency (to maximise the number of heterozygotes studied). All markers were assayed using "universal" primers based on single exonic sequence capable of amplifying genomic and complementary DNA. The same analytical conditions were used for both genomic and cDNA. This enabled the average of the ratios observed from genomic DNA (representing a 1:1 ratio of the two alleles) to be used to correct allele ratios obtained from cDNA analysis for any inequalities in allelic representation specific to each assay [184]. As a result a divergence from the 1:1 allelic ratio by cDNA samples could be indicative of an allele-specific regulation of steady-state mRNA levels in the tissue examined, (Figure 2.5).

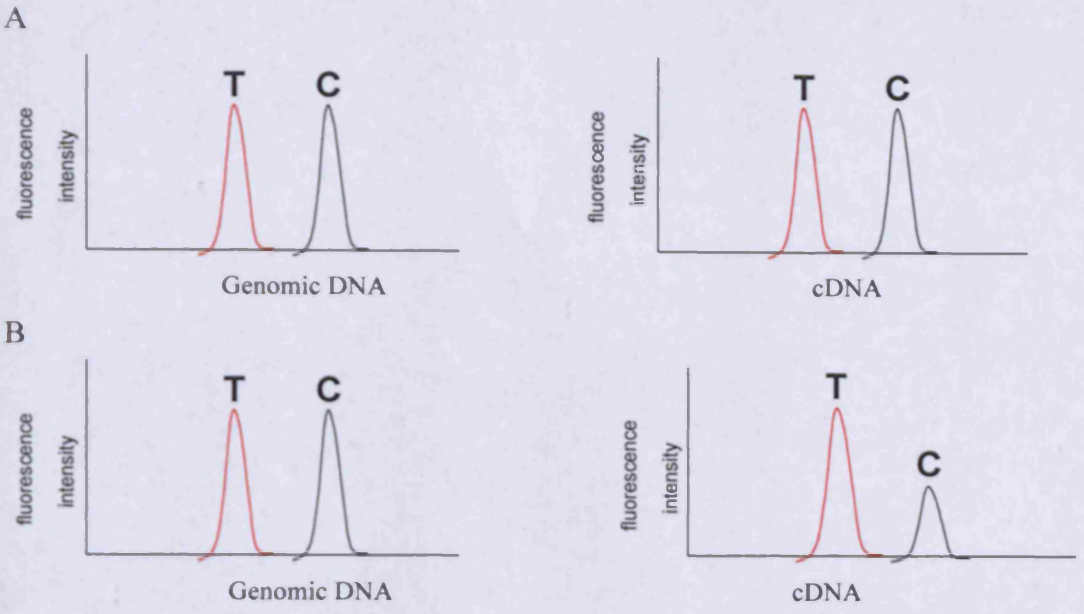


Figure 2.5. Illustration of a SNaPshot assay result for a heterozygote individual who A) shows no cis-acting variation and B) is under the influence of cis-acting variation. gDNA ratios are 1:1 however where cis-acting variation is influencing transcription of one mRNA strand the intensity of each allele will show a divergence from the 1:1 ratio in the cDNA. The florescence intensity readings are taken and a ratio of Major allele/Minor allele is then calculated.

All allelic expression assays were performed using the Caucasian Brain samples (described in section 2.1.2). Each cDNA sample was assayed alongside the corresponding heterozygous genomic DNA to ensure uniformity of conditions.

PCR amplification and primer extension was carried out with SNaPshot chemistry (Applied Biosystems) as described in section 2.7.2. Aliquots of 3µl SNaPshot reaction product were combined with 8µl HiDi formamide and loaded onto an ABI3100 PRISM Genetic Analyser (Applied Biosystems). Peak heights representing allele-specific extended primers were determined by using Genotyper software (Applied Biosystems) and were used to calculate the ratio of allelic representation for each sample. All peak heights were between 550-6000 fluorescence intensity units and assays showing genotypes for blank samples were excluded.

2.9.1 Allelic Expression Analysis

The following protocol was used to analyse relative differential allelic mRNA expression. gDNA and cDNA from individuals heterozygous for the assay SNP were analysed in parallel. Duplicate cDNA samples (independent reverse transcription (RT) reactions for the same sample) were used to increase assay confidence. Raw data from Genotyper software was collated for each gDNA and cDNA sample. To obtain an allele ratio the fluorescence intensity value for the common allele was divided by the fluorescence intensity value of the rare allele. The mean of all the gDNA ratios was taken and each gDNA and cDNA ratio corrected by this mean (thereby adjusting the gDNA average ratio to 1). Samples showing a standard deviation of ±0.2 between duplicate RTs were removed. The average corrected cDNA ratio from two duplicate RTs was then calculated. The average standard deviation over all corrected duplicate RT standard deviations was recorded as a quality score for the assay.

Evidence for differential allelic expression was deduced if an assay met the following criteria. Firstly, one of more samples had to show a differential expression of greater than 20% after correction using the average genomic ratio. Yan and colleagues [254] analysed allelic expression of the APC tumour suppressor gene in 17 HapMap CEU individuals plus a familial adenomatous polyposis (FAP) patient, previously shown to have decreased expression of one allele. As no significant variation in expression of APC was detected in the 17 CEU individuals, they concluded that there was little common variation in APC expression and that APC could thereby be used to model analysis of other genes. Analysis of the 95% confidence intervals of the APC allelic expression ratios in the 17 individuals resulted in Yan *et al* estimating that a variation in expression could be confidently identified when the expression of the two mRNA copies, and therefore the corrected cDNA ratio, differed by more than 20% (outside the corrected ratio of 0.8-1.2).This criterion has been used in several subsequent allelic expression analyses [184, 186, 194, 195, 255] and therefore was also used in this study. Secondly, a statistically significant difference ($p<0.05$) needed to be observed. To determine the statistical significance of any differences observed, all ratios (corrected genomic or cDNA) were normalised via a natural log transformation. After

confirmation by the Kolmogorov-Smirnov test that the transformed gDNA and cDNA data followed a normal distribution, gDNA and cDNA values were compared by a 2 independent sample t-test. To account for the possibility of weak LD between the assayed SNP and a potential regulatory SNP and/or multiple cis-acting variants, the spread of the cDNA ratios and gDNA ratios were compared using the Levene's test.

2.10 Bioinformatic and Statistical Analysis

Several types of statistical analyses were performed in this thesis. The majority were carried out using the analysis software PLINK (http://pngu.mgh.harvard.edu/purcell/plink) [256]. This included tests for single marker association in cases and controls (allelic and genotypic) and haplotypic analysis. Analysis of variants for deviations from Hardy-Weinberg were also performed using PLINK. Statistical analyses that did not use the PLINK software are outlined below or are found within the methods section of the relevant chapter.

2.10.1 LD Estimation and Tag SNP Determination

Determination of the LD between markers and tag SNP identification was performed using Haploview (http://www.broad.mit.edu/haploview/haploview), a software program designed for genetic association studies [257]. The program allows the user to import marker genotype data such as a CEU HapMap dataset or case-control genotype data. The quality of this data can then be assessed via a display of the percentage of individuals genotyped, Hardy-Weinberg equilibrium p values and non-Mendelisations. The LD (r^2 and D') between markers can also be identified. The Tagger function within Haploview uses the LD values to select "tag" markers for an association study via user defined parameters (r^2, MAF and the type of analysis to be performed; pairwise, haplotype). In addition to LD analysis Haploview can also be used test for marker or haplotype association.

2.10.2 Sample Size Power Calculations

Power calculations for the mutation detection samples were determined using the equation: $1-(1-f)^n$, where f = minor allele frequency and n = number of chromosomes examined (i.e. 2x number of individuals).

The power of an association sample to detect susceptibility variant(s) with a specific MAF and OR at a given p value were calculated using the software PS Power and Sample Size Calculations [92].

Chapter 3: The Identification of Putative DTNBP1 Regulatory Regions and Association Analysis of SNPs with Schizophrenia

3.1 Introduction

Allelic expression analysis has determined that DTNBP1 is under the influence of cis-acting regulatory variant(s). As dysbindin expression has been shown to be altered in schizophrenic patients compared to controls it has been suggested that this cis-acting variation may also confer susceptibility to schizophrenia. However while determining that DTNBP1 is under the influence of cis-acting variation is relatively straightforward, identifying the cis-acting variants provides a greater challenge.

Although some studies have screened predicted promoter regions in addition to the analysis of DTNBP1 coding regions [113, 134, 258], there has been no comprehensive analysis of genomic sequences that may have a cis-acting regulatory influence on dysbindin expression. This chapter describes the identification of putative DTNBP1 cis-regulatory regions, the screening of these regions for polymorphisms and the subsequent analysis of variants for association with schizophrenia.

While determining regulatory variants, particularly those distal to the target gene remains a challenge, there are a number of different approaches that can be used to identify genomic sequence that has a potential regulatory function. These, plus a brief overview of transcription and the different types of cis-regulatory elements, are described below.

3.1.1 Categories of Cis-acting Regulatory Elements

The initiation and regulation of transcription is a complex process, which involves the assembly of a number of proteins known as transcription factors (TFs). TFs bind to specific DNA sequences which are organised into a series of regulatory elements. Consequently the molecular basis for the transcriptional regulation of gene expression involves the binding of trans-acting proteins (transcription factors) to cis-acting sequences (binding sites) [192].

Cis-acting regulatory elements can be split into two groups based primarily on their distance from the transcription start site (TSS) of a gene (See Figure 3.1). The promoter is usually directly upstream of the TSS and contains core and proximal promoter elements. This group of elements provide the basal transcription levels for a gene. Distal regulatory elements often modulate this basal transcription. They contain enhancers and silencers which act independently of their distance from, and orientation to, the target gene.

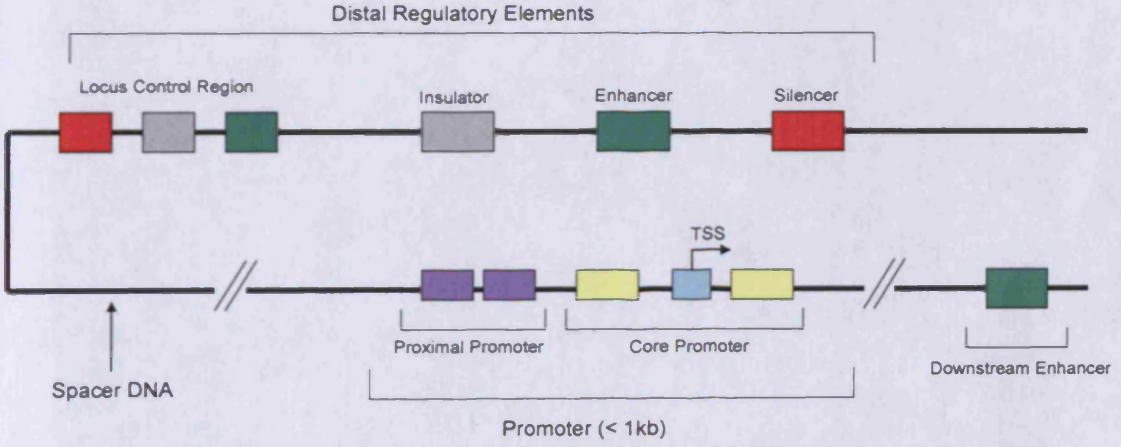


Figure 3.1. The complex arrangement of the regulatory elements for a typical gene. Regulatory elements include the promoter, which is composed of core promoter and proximal promoter elements and typically spans less than 1kb. Distal (upstream or downstream) regulatory elements, which can include enhancers, silencers, insulators and local control regions, can be located up to 1Mb from the promoter. Adapted from [191, 192, 259].

3.1.1.1 The Promoter: Core and Proximal Elements

The core promoter, also known as the minimal promoter, describes the region at the start of a gene that serves as the docking site for the basic transcriptional machinery. In the case of protein encoding genes this includes general transcription factors (GTFs) and RNA polymerase II. GTFs assemble in an ordered fashion on the core promoter to form the preinitiation complex (PIC). The PIC recruits RNA polymerase II to the promoter and therefore defines the start and direction of transcription. The first step in PIC assembly is the binding of transcription factor TFIID; a multi-subunit consisting of the TATA box binding protein (TBP) and a set of tightly bound TBP associated factors (TAFs). The TBP subunit binds to the consensus DNA sequence TATA(A/T)A(A/T), known as the TATA-box. In addition to the TATA box, metazoan core promoters can be composed of a number of recognition elements including initiator element (Inr), downstream promoter element (DPE), downstream core element (DCE), TFIIB-recognition element (BRE) and motif ten element (MTE). Except for BRE, which binds TFIIB, all other known core promoter elements bind TFIID [259]. It is therefore likely that many features of the core promoter have yet to be discovered. It is also possible that known core promoter elements are not as common as previously thought. Statistical analysis of ~10,000 predicted human promoters for the four core promoter elements TATA, BRE, Inr and DPE found that TATA boxes were present in only 1/8 of promoters and over a quarter of the promoters analysed did not include any of the four elements [260].

The proximal promoter is found upstream of the core promoter and typically contains recognition sites for sequence specific ubiquitous transcription factors. These recognition sites can include GC boxes, as well as CCAAT boxes. GC boxes bind the TF Sp1 and are often within 100bp of the transcription initiation site. CCAAT boxes are typically located at -75bp and are recognised by CTF and CBF (also known as Nuclear Factor I (NFI) and NFY respectively) [261].

The transcription factors which bind to both core and proximal promoter sequences are traditionally thought to promote basal transcription. However recent experiments [262-

264] have suggested that recognition sites in these regions may also bind proteins that are used as tethering elements to recruit distal regulatory elements to the core promoter.

3.1.1.2 Distal Elements

Distal cis-acting regulatory elements are so called as they can be located several kilobases away from their target gene [191, 192, 259]. Some distal elements have even been found at distances of up to 1Mb from the gene they regulate [265, 266]. Unlike promoter elements, distal regulatory elements function independently of both distance and orientation from the gene and as a result can be 3', 5' or within an intron of the regulated gene or even a neighbouring gene [259]. It is commonly thought that distal regulatory elements typically regulate transcription in a spatial and temporal specific manner. They come in the form of enhancers, silencers and insulators and the average gene is likely to have several of these regulatory elements.

As the name suggests enhancers are involved in the increased expression of their target gene. They are recognised by transcription factors known as activators. Activators can be distinguished by their different DNA binding domains. Specific activator families contain domains such as a cysteine-rich zinc finger, homeobox, helix-loop-helix (HLH) or basic leucine zipper (bZIP) [261]. Each binding domain has its own specific enhancer DNA sequence with which to bind. Enhancer sites are generally small (6-12bp) and specificity is usually dictated by no more than 4-6 positions [259].

Silencers are bound by TFs known as repressors which confer a negative effect on transcription. They can confer a negative effect through direct competition with an enhancer for the binding of a transcription factor [267] or by binding a repressor which blocks the binding of an activator to a nearby enhancer [268]. Consequently silencer sequences are often found with enhancers within clusters of transcription factor binding sites (TFBSs). Enhancers and silencers can also be found within proximal promoter regions and therefore the distinction between promoter and distal elements can become blurred.

There have been a number of theories as to how distal elements could affect transcription over such large physical distances. Recent *in vivo* experiments on D*rosophila melanogaster* [263, 269] have provided evidence that distal regulatory elements could function via chromatin looping whereby the distal regulatory unit is brought into close proximity to the promoter by looping out the intervening DNA (see Figure 3.2). In addition *in vivo* footprinting and chromatin immunoprecipitation has shown RNA polymerase II binding to distal enhancer elements which suggests the PIC may initially form at a distal enhancer rather than at the core promoter [270]. This could occur where rapid gene activation is required as it would allow precise regulation of transcription initiation.

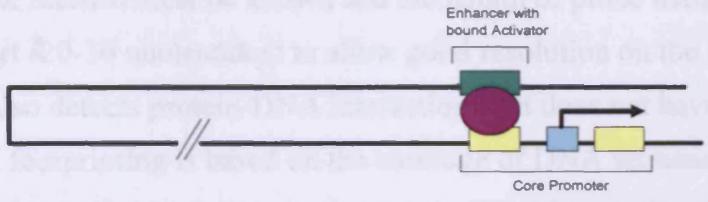


Figure 3.2. DNA looping model for distal regulatory element action. Distal (upstream or downstream) regulatory elements, which can include enhancers, silencers, insulators and local control regions, can be located up to 1Mb from the promoter. These distal regulatory regions may gain contact with the core or proximal promoter through a mechanism that involves looping out the intervening DNA. Adapted from [269].

Due to the long distances involved in enhancer and silencer action, they have the potential to activate several neighbouring genes. Although a number of enhancers have been reported to show a preference for core and proximal promoter elements of specific genes this does not always appear to be the case [262, 271]. Therefore the action of enhancers and silencers needs to be restricted to prevent the activation/inactivation of non-target genes. This role is performed by a third class of distal cis-acting regulatory elements known as insulators. They can be between 500bp and 3kb in length and evidence suggests they can function in two ways. An insulator can be located between enhancers and promoters to block the enhancer-promoter communication [272] or they can flank a gene and prevent the spread of repressive chromatin [273].

3.1.2 The Identification of Putative Regulatory Regions

Several techniques have been developed in order to identify genomic regulatory regions. These include electrophoretic gel-mobility shift assays (EMSA), DNA footprinting, chromatin immunoprecipitation (ChIP), DNase hypersensitivity analysis and reporter gene assays [274, 275].

EMSA allows the analysis of protein-DNA interactions *in vitro*. Radio- or fluorescent-labelled oligonucleotide probes are used as bait for transcription factors. The binding of transcription factors to a probe is detected by gel electrophoresis as bound probes will migrate slower through a gel than unbound probes. A limitation of EMSA is that the DNA region of interest must be known and the length of probe used needs to be relatively short (20-30 nucleotides) to allow good resolution on the gel. DNA footprinting also detects protein-DNA interactions but does not have the limitations of EMSA. DNA footprinting is based on the cleavage of DNA sequence by either a chemical or enzyme. In the presence of a protein-DNA interaction the DNA will be protected from cleavage and these uncut regions or "footprints" can be observed by gel electrophoresis.

Unlike *in vitro* binding assays such as EMSA and DNA footprinting, chromatin immunoprecipitation (ChIP) allows the identification of endogenous DNA-protein complexes. Cells are treated with formaldehyde in order to "fix" the DNA-protein interactions via cross-linking. These DNA-protein complexes are subsequently immunoprecipitated from nuclear extracts by an antibody specific to the protein of interest [276]. After reverse cross-linking, the DNA with which the protein of interest has been bound can be determined.

A further method of identifying regulatory regions is the reporter gene assay. In these experiments a genomic sequence of interest is inserted into a reporter gene construct. The expression of the reporter gene can then be compared between reporter gene constructs with and without the inserted sequence.

The mapping of DNase hypersensitive sites has also been used to identify the location of regulatory elements [277, 278]. DNA is packaged in chromatin which generally adopts a highly condensed structure that is relatively inaccessible to proteins. As a result, at any given time, most of the human genome is a poor template for biochemical reactions such as transcription. However the loss or remodelling of one or more nucleosomes, the basic repeating unit of chromatin, at a given genomic location produces hotspots of accessibility around regions important for gene regulation such as promoters. The increased availability of the DNA to transcription factors and other regulatory proteins also increases its sensitivity to digestion by nucleases such as DNase1. This is especially true of locus control regions which are often marked by clusters of DNase1 hypersensitive sites [259].

While these "wet-laboratory" methods have been successful in identifying regulatory sequence, they often require significant optimisation, are labour intensive and time consuming. *In silico* analysis can be used to identify putative regulatory regions which can then be investigated further. *In silico* methods either analyse a particular sequence for certain features, such as TFBS motifs, or utilise previous wet-laboratory results, for example DNase1 hypersensitivity experiments. The former of these methods are described below.

3.1.2.1 Promoter Regions

The simplest and most logical starting point for the identification of polymorphisms that could have an affect on transcription is to analyse the sequence immediately 3' and 5' of the TSS of the gene of interest. Reporter gene analysis of over 700 unique gene promoters revealed that sequence variants which altered gene expression by more than 1.5 fold were strongly biased towards the core and proximal promoter regions with 75% within the first 200 bases 5' to the TSS [279]. Therefore genomic sequence immediately 5' and 3' of a gene's TSS is an ideal region to screen for eSNPs, particularly if the sequence contains transcription factor binding sites (TFBSs) or is near a CpG island. A CpG island is a relatively short stretch of DNA (500bp-2kb) which has a high G+C content and a high frequency of CpGs (a dinucleotide consisting

of a cytosine nucleotide next to a guanine separated by a phosphate group). CpGs are scattered throughout the genome and are methylated at the 5th carbon position of the cytosine base. Approximately 60% of all human gene promoters fall near a CpG island and methylation status (the default of which is unmethylated) can be correlated with gene expression [280, 281].

3.1.2.2 Transcription Factor Binding Site Clusters

Individual TFs generally bind to DNA with relatively low specificity. Thus, the precise control of gene transcription requires a higher degree of specificity than that typically afforded by the binding if a single transcription factor to a single DNA recognition site [282]. Regulatory elements accomplish this desired control by being composed of relatively closely grouped clusters of TFBSs. A cluster of several recognition sites would rarely be encountered in the genome whereas a single recognition site of only 4-8bp could potentially be quite common. These clusters are also beneficial as multiple TFs bound to a cis-regulatory cluster typically function synergistically and activate transcription more strongly than a single factor alone [282, 283].

Other areas of the genome which contain clusters of TFBSs are locus control regions (LCRs). LCRs are a feature of long distance cis-acting regulation and are involved in the regulation of a group of genes [284]. They are often composed of multiple cis-acting elements including enhancers, silencers as well as insulators. LCRs are typically located upstream of the target gene(s). However they can also be found within introns of a target gene [285], downstream [286, 287] or even within an intron of a neighbouring gene [288].

3.1.2.3 Evolutionary Conserved Regions

Identifying regions that contain clusters of TFBSs can be a successful way of finding regulatory regions [289-292]. However a limitation in identifying clusters of TFBSs is that it relies on the knowledge of the transcription factor binding sequence. It has been shown that the majority of sequence variants within promoter regions that alter expression are not within predicted TFBSs [279]. This suggests there may be many recognition sites, or other unknown factors that may affect transcription, yet to be identified. One method of identifying putative regulatory sequence that does not require prior knowledge of transcription factor binding site sequence motifs is the identification of evolutionary conserved sequence.

It is reasonable to hypothesise that conserved sequence would be more likely to contain functional elements than the rest of the genome. A total of 5% of bases within the genome can be confidently identified as under evolutionary constraint in mammals [293]. However coding regions only comprise ~1.5% of the genome and comparisons of mammalian and non-mammalian genomes demonstrate that ~3% of non-coding sequence is strongly conserved. Therefore, these conserved non-coding regions (CNCs), also known as conserved non-genic regions (CNGs) [294], could potentially be functionally relevant.

By comparing distant species such as *D. melanogaster* and *D. virilus* scientists have been able to identify regulatory regions in the Drosophila genome [295, 296]. Consequently several attempts have been made in mammals, in particular the mouse, to identify regions of regulatory function by means of sequence conservation. Initial studies which have attempted to determine regulatory regions in the human genome characterised a region as conserved if the sequence contained at least 100bp of ungapped genomic sequence which was at least 70% identical between the human and mouse genome [297-300]. These criteria are above the average level of neutral sequence conservation between human and mouse genomes where only 40% of the mouse genome can be aligned to the human genome, of which an average of 67.2% of nucleotides are identical [301].

To increase the sensitivity of selecting functionally relevant sequence more recent studies have compared either a third species, usually the dog [302-305], or several mammalian species [306-308] simultaneously. These regions are often referred to as ultraconserved regions [309]. Sequence identified by comparing the human and mouse genomes appears to show a uniform distribution within intergenic sequences whereas ultraconserved regions appear to cluster around genes [266, 309]. Therefore, if only a minority of CNCs are cis-transcriptional regulators then it is likely to be these ultraconserved sequences.

While the most common hypothesis for the role of CNCs within the human genome is that this conserved sequence contains cis-acting regulatory sequence but it is possible that they have a different function, especially CNCs within so-called gene deserts. A recent study has suggested that up to 10% of CNCs may be matrix-attachment regions [310] which regulate the conformation of chromatin through the binding of particular proteins. They might also participate in inter-chromosomal interactions that are mediated through protein bridges and bring chromosomes together in the nucleus [311, 312].

3.1.3. Aims of This Chapter

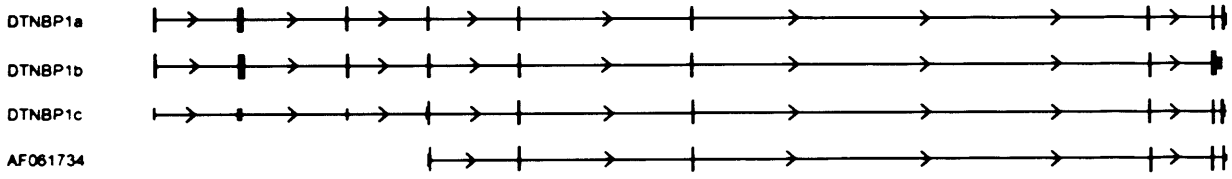
The objective of this chapter was to identify DTNBP1 putative functional SNPs associated with schizophrenia. This involved the identification of putative regulatory regions at the DTNBP1 locus. As knowledge of gene regulatory mechanisms is incomplete and each approach to identifying cis-acting regulatory regions (and therefore potential regulatory polymorphisms) has numerous caveats, a combination of a number of *in silico* methods was used to identify putative regulatory sequence. The regions identified were subsequently screened for variants. The SNPs detected, plus any additional SNPs required to tag the DTNBP1 gene itself, were subjected to association analysis with schizophrenia.

3.2 Methods

3.2.1 Putative Regulatory Region Identification

Evidence suggests that there are at least four transcription variants of DTNBP1 (UCSC, May 2004). Details of these transcripts are given in Figure 3.3. Three of these isoforms have supportive evidence of one GenBank RNA sequence plus at least one additional line of evidence from RefSeq, CCDS or Uniport. DTNBP1a (NM_032122) encodes the longest isoform (chr6:15631018-15771250) and contains 10 exons. Compared to DTNBP1a, DTNBP1b (NM_183040) contains additional coding sequence in exon 9 but has a shorter and distinct C terminus. DTNBP1c (NM_183041) contains an alternative splice site in the 5' coding region, uses a downstream start codon and has a shorter N-terminus compared to DTNBP1 variants a and b [313]. Another transcript AF061734, often called the alternative transcript or short isoform, has less conclusive experimental evidence but has supporting data of a GenBank mRNA [314]. Its mRNA is 104kb compared to the 138-140kb of DTNBP1a, b and c and contains an alternative transcription start site.

A



B

Transcript	mRNA ID	Transcript sequence	mRNA Length	Total exon count	Coding sequence	Number of AAs	Number of coding exons
DTNBP1a	NM_032122	chr6:15631018-15771250	140233	10	chr6: 15631185-15771079	352	10
DTNBP1b	NM_183040	chr6:15632635-15771250	139464	9	chr6: 15632635-15771079	304	9
DTNBP1c	NM_183041	chr6:15631018-15771250	140233	10	chr6:15631185-15735644	271	6
AF061734	AF061734	chr6:15631020-15735583	104563	6	chr6:15631185-15735578	165	6

Figure 3.3. A Alternative DTNBP1 transcripts. Exons in blue are coding, grey regions are non-coding. B Transcript specifics (May 2006 Freeze). AA = amino acids. Sources [198, 313].

Due to limitations on the number of regions that could be screened for putative eSNPs, a "window" of genomic sequence around the longest DTNBP1 variant (DTNBP1a) was determined which was then subjected to *in silico* analysis to identify putative regulatory elements. Using this "window" maximised the chance of discovering DTNBP1 regulatory regions while reducing the number of false positives identified (for example, regions with no regulatory effect or no regulatory effect on DTNBP1 expression). In total, the derived genomic sequence from 75kb 5' to 10kb 3' of DTNBP1a (chr6:15,621,018-15,844,250) was analysed *in silico* in order to identify putative cis-acting regulatory regions. A cut-off point of 75kb upstream of DTNBP1 was chosen although the nearest gene 5' to DTNBP1 is 465kb upstream, due to time and funding constraints, analysis of this entire region was unfeasible within this study. Another gene, JARID2 is situated only 1kb 3' to DTNBP1a. As regulatory elements within the JARID2 gene are more likely to regulate JARID2 expression rather than DTNBP1 only 10kb of genomic sequence 3' to DTNBP1a was selected. Details on the four different strategies for identifying putative regulatory regions are given below.

3.2.1.1 Core and Proximal Promoter

Most core and proximal promoter elements are located adjacent to the TSS of their target gene [279]. The transcription start sites of the four main DTNBP1 transcripts were identified using the genome browser UCSC (May 2004 freeze). The mRNA transcript of DTNBP1a, DTNBP1b and DTNBP1c starts at the same point (chr6:15771250). As these isoforms are likely to share the same core promoter region, 5kb 5' and 3' of transcription start site of these three isoforms (chr6: 15,765,956-15,776,344) was included for polymorphism screening.

The TSS of the alternatively spliced transcript AF061734 is situated at the beginning of exon 5 of the other DTNBP1 isoforms and consequently AF061734 is likely to be regulated by a separate promoter region. A smaller 1kb window 5' and 3' of the TSS of AF061734 (chr6:15,763,611-15,736,610) was screened for putative regulatory polymorphisms.

3.2.1.2 Transcription Factor Binding Sites Clusters

Regions containing multiple transcription factor binding sites (TFBSs) were identified *in silico* using the web based algorithm Cluster Buster (http://zlab.bu.edu/cluster-buster/cbust.html). Cluster Buster employs a probabilistic model to search for regions in a sequence that greater resembles a transcription factor binding site motif cluster more than background DNA [315]. Sequences that pass the likelihood ratio threshold (default =5) are displayed in an overview diagram where potential individual TFBSs are listed.

To identify clusters of TFBSs at the DTNBP1 locus the genomic sequence chr6:15,621,018-15,844,250 was analysed for the 16 TFBS motifs listed on the Cluster Buster website (TATA, Sp1, CRE, ERE, NF-1, E2F, Mef-2, Myf, CCAAT, Ap-1, Ets, Myc, GATA, LSF, SRF, Tef). Default parameters were used (gap parameter=5, cluster score threshold=5, motif score threshold=6) so as to balance the sensitivity between tight clusters with weak motifs and loose clusters with strong motifs [315].

3.2.1.3 Regions of High Conservation

Non-coding regions of high conservation between species (CNCs) were identified for chr6:15,621,018-15,844,250 *in silico* using two approaches originally described by Drake *et al* [316]. Firstly CNCs were identified using ECR (Evolutionary Conserved Regions) browser [317] (www.ECRbrowser.com). This database displays conservation profiles between multiple genomes such as human, rodent or fish (See Figure 3.4).

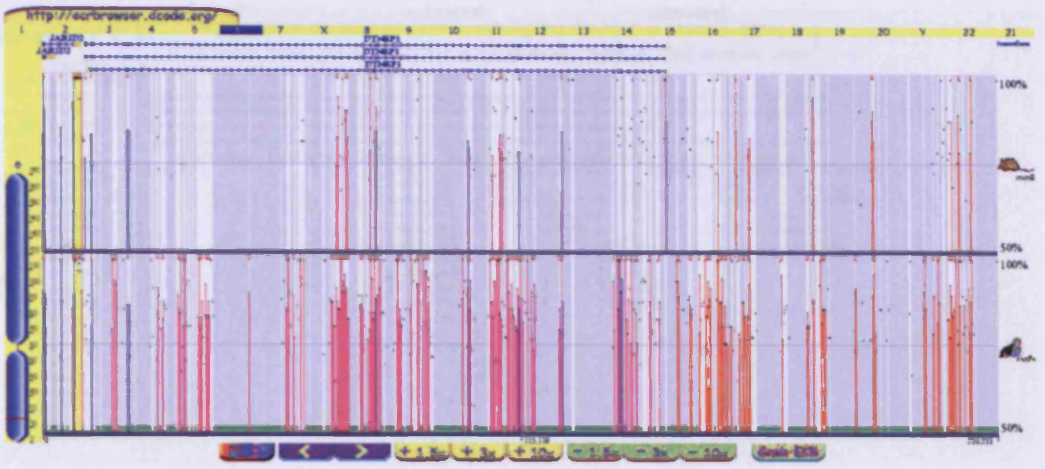


Figure 3.4. Conservation profile from ECR browser which depicts conserved regions between the human and the mouse genome plus the human and the dog genome at the DTNBP1 locus. Annotated genes are depicted as a horizontal blue line above the graph. Conserved regions within coding exons are coloured blue and conserved regions within UTRs are coloured yellow. Evolutionary conserved regions that do not correspond to transcribed sequences are highlighted in red if they are intergenic or pink if they lie within an intron.

Drake *et al* used a variety of different analysis parameters to detect CNCs across the entire genome by identifying conserved regions between the human and mouse genome and the human and dog genomes. As this study aimed to identify regulatory regions solely at the DTNBP1 locus, a less stringent set of criteria was used than that recommended by Drake *et al*. CNCs were defined as non-coding sequence with a minimum length of 100 bases and a minimum identity of 80% between the human genome and both the mouse and dog genome.

Secondly, CNCs were determined using the UCSC track "most conserved", May 2004 freeze (See Figure 3.5). This track catalogues conserved non-coding regions identified through analyses designed to extract the top 5% of the conserved genome. This analysis examined the conservation between vertebrate, insect, worm and yeast genomes [318]. For the DTNBP1 locus 145 coding and non-coding conserved regions were stored on this track. In order to prioritise these regions for further analysis, non-coding sequence with a lod threshold level of >40 was selected.

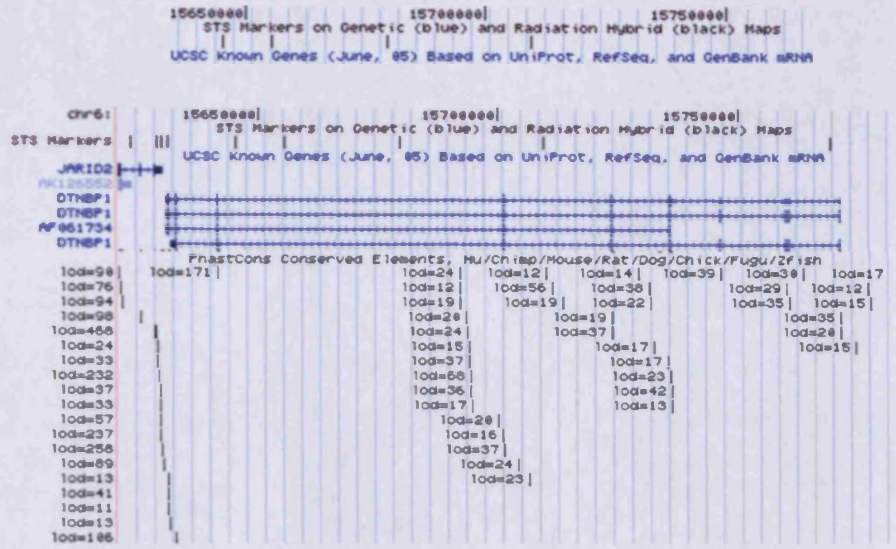


Figure 3.5. Conserved regions around the DTNBP1 locus identified from the comparison of vertebrate, insect, worm and yeast genomes. Non-coding regions with a lod score >40 were subsequently identified for further analysis.

3.2.1.4 DNase Hypersensitive Regions

DNase HS regions were determined using experimental data previously reported [319]. Conventionally, DNase1 hypersensitive sites (HS) have been detected by subjecting isolated nuclei to DNase treatment with subsequent digested genomic DNA detected via Southern blotting. This can reveal the extent to which DNase1 cleavage occurs in a specific genomic region near a selected probe. Crawford and colleagues [320] published a protocol that dramatically scaled up this method by using high throughput sequencing to sequence regions flanked by DNAseI hypersensitive sites. The results of which were deposited on the UCSC track NHGRI DNaseI-HS track (May 2004). The data from this track has since been merged with a follow up study by the same group [319] which aimed to create a more accurate, high throughput system to identify valid DNaseI HS sites across the genome by improving the signal to noise ratio. This merged data can be found on a new UCSC track Duke/NHGRI DNaseI-Hypersensitivity (May 2004). Both tracks were mined for DNase 1 HS regions around the DTNBP1 locus.

3.2.2 Samples

3.2.2.1 Mutation Screening Sample

Mutation screening of putative regulatory regions was performed in 14 unrelated schizophrenics (described in chapter 2.1.3) plus 13 other individuals. These 13 individuals included 7 individuals previously reported to have DTNBP1 allelic expression differences >20% [186]. The remaining six samples comprised of three sets of parents whose offspring had been included in a previous study which measured total DTNBP1 expression by real time PCR using Taqman Gene Expression Assays [185]. The parents of these individuals were chosen rather then the offspring themselves as the genetic variation of multiple siblings who showed altered total expression could be analysed by including just the two parents. Of the three sets of parents two pairs were included as their offspring showed the lowest DTNBP1 expression in the previous study and two individuals (one set of parents) were chosen as their offspring had the highest total DTNBP1 expression.

3.2.2.2 Schizophrenia Case Control Sample

The schizophrenia case control sample used in this chapter consisted of 709 schizophrenic patients and 716 controls. Specific details of this sample are given in chapter 2.1.1.

3.2.3 Mutation Screening of Putative Regulatory Regions

All putative regulatory regions, identified by the methods described above, were screened for polymorphisms using the Lightscanner software (Idaho Technologies) and BigDye Sanger sequencing. The protocols for these methods are described in chapter 2.5.2 and 2.6 respectively.

3.2.4 SNP Selection

All DNA variants detected were genotyped in the 30 CEPH parent-offspring trios that constitute the HapMap CEU sample. Detected SNPs were combined with all SNPs (n=216) of HapMap phase II (Jan 06) that span the *DTNBP1* locus (Chr6:15621211-15780522, March 2006 freeze). Pairwise tagging via the Tagger function of Haploview was then used to select a non-redundant set of SNPs which captured all the alleles at the DTNBP1 locus at a MAF >0.001 and at an r^2>0.95. In addition any putative regulatory variants with a MAF>0.001 were captured with an r^2=1.

3.2.5 Genotyping

All SNPs were genotyped by primer extension using Sequenom MassARRAY as described in the chapter 2.7.1. Genotyping was performed either by myself or Dr. Liam Carroll. If genotyping failed using Sequenom Mass Array, genotyping was attempted using Amplifluor (see chapter 2.7.3).

3.2.6 Statistical Analysis

3.2.6.1 Association Analysis

All statistical analysis was performed using plink software [256]. This included association, Hardy-Weinberg equilibrium (HWE) and haplotypic analysis. Each marker was tested for allelic association using the Armitage trend test. Genotypic association was performed using χ^2 2degrees of freedom (df) tests. For SNPs whose genotype counts were less than 5 genotypic association analysis was performed using CLUMP where tests of significance are achieved by permutation. In this instance 10000 permutations were performed [321]. Goodness of fit tests for HWE were performed in the cases and controls separately. Analysis of all combinations of 2 and 3 marker haplotypes was also performed using Plink as was the specific analysis of previously reported risk and protective haplotypes [134] in this extended case control sample.

3.2.6.2 Logistic Regression of Association Signal

Logistic regression analysis of the association signal was performed in SPSS. The genotypes of each SNP showing an allelic association were coded as 0, 1 or 2 based on the number of copies of the risk allele. For the logistic regression analysis case control status was selected as the dependant variable and the genotypes of all associated SNPs were selected as covariates.

Logistic regression was performed three times using the enter method plus the forward and backward stepwise logistic regression methods. Probability of stepwise thresholds used in this analysis were entry=0.05 and removal=0.08.

The enter method determines whether any SNPs show a significant association after p values are adjusted for the correlation between the test SNP and all other SNPs in the model. The forward stepwise logistic regression identifies the most associated SNP using the enter method then determines whether any other SNP remains significant after taking the correlation between the test SNP and the most significant SNP into account. If they still show a significant association they are added to the model. The backward stepwise logistic regression begins with all SNPs within the model. SNPs with non-significant adjusted P values are removed in a stepwise manner with the logistic regression repeated after each SNP is removed.

3.3 Results

3.3.1 Identification of Putative Regulatory Regions

As no one conclusive *in silico* method to identify cis-acting regulatory sequences has been described, a complimentary set of different methods was devised for determining putative regulatory regions at the DTNBP1 locus. This consisted of determining putative promoter regions, regions containing clusters of TFBSs, DNase HS sites plus evolutionary conserved sequence. Analysis of chr6:15,621,018-15,844,250 identified 21 putative regulatory regions (Figure 3.6). The exact chromosomal position of the regions identified and primers designed for mutation screening are given in Appendix Tables 9.1.1 and 9.1.2.

Two putative promoter regions for the four main DTNBP1 transcripts were identified (Figure 3.6). Of the remaining 19 regions, 5 were highly conserved regions determined using either ECR browser (n=3), USCS "most conserved" track (n=1) or both (n=1). Analysis using the web based program Cluster Buster predicted 13 regions to contain clusters of TFBSs and one DNase hypersensitive region was identified using the track NHGRI DNaseI-HS (May 2004).

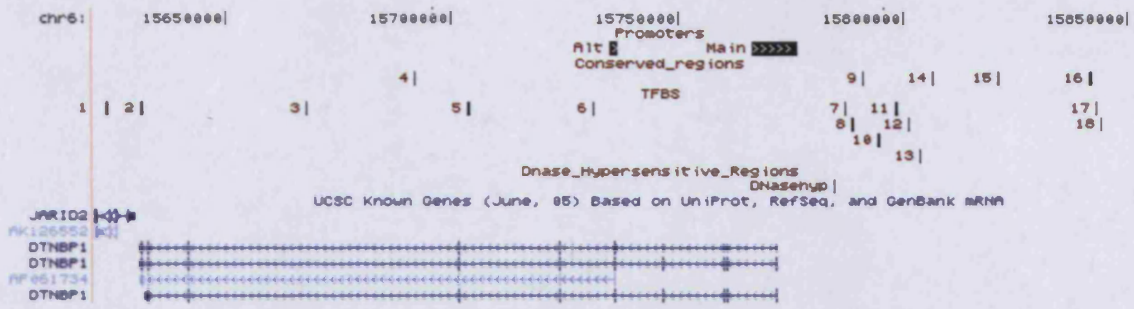


Figure 3.6. Chromosomal location of putative DTNBP1 regulatory regions.
Genomic sequence identified as putative regulatory regions by a combination of methods are shown. Sequence immediately surrounding the transcription start site of the three of the main DTNBP1 isoforms plus AF061734 are illustrated under the track named "Promoters". Evolutionary conserved regions, as well as DNase hypersensitive regions, are given within the "Conserved Regions" and "DNase Hypersensitive Regions" tracks respectively. Regions identified by Cluster Buster to contain clusters of TFBSs are shown under the track name "TFBS".

3.3.2 Comparison of Regulatory Region Detection Methods

Although ECR browser and UCSC use different methods to identify regions of conservation (see section 3.2.1.3) it could be expected that there would be some overlap in the regions identified by the two methods. While this is the case for the sequence chr6:15820928-15821304 (region 15 on Figure 3.6) which is identified by both ECR browser and UCSC to be a highly conserved region, there is no overlap in the other four regions identified by these methods.

It could also be hypothesised that there would be additional overlap in the genomic regions identified as conserved, or regions containing putative promoter regions, TFBSs or DNase HS sites. For example, both putative promoter regions and regions of high conservation could be expected to contain TFBSs or promoter regions could show high conservation between species. Furthermore methods that identify sequence containing TFBSs and DNase HS sites could potentially identify the same sequence. However as can be seen it Figure 3.6 this is not the case, as each method identifies unique sequence not determined by another *in silico* analysis.

3.3.3 Polymorphism Detection within Regulatory Regions

Screening all putative DTNBP1 regulatory regions identified 56 SNPs, 24 of which were novel (Table 3.1.) 32 of these SNPs are located within the core promoter region of either the main (n= 26) or alternative transcripts (n= 6). Five DNA variants were identified in regions of high conservation, 17 within sequence predicted to contain clusters of TFBSs and two polymorphisms were observed in the area of DNase hypersensitivity.

Of the 56 SNPs identified, 39 had not been previously genotyped as part of the HapMap project. Therefore an attempt was made to genotype these 39 SNPs through the HapMap CEU sample in order to determine the MAF and the LD structure of these variants. However 8 of these 39 polymorphisms (rs12194321, DTNBP1-C13_SNP1, DTNBP1-C24b_SNP1, DTNBP1-C26_SNP1, rs10949310, DTNBP1-R16_microsat, rs5874525, rs3070184) failed to be genotyped either by Sequenom or SNaPshot or Amplifluor. Therefore these have not been captured in the following association analysis.

SNP No	SNP ID	Chromosomal Position	SNP Major/Minor	MAF	Region
1	rs1076636	15623763	C/T	0.24	TFBS cluster
2	rs1076635	15624057	A/G	0.22	TFBS cluster
3	rs17470454	15631177	G/A	0.05	TFBS cluster
4	rs2743548	15691803	T/G	0.08	CNC
5	rs9464802	15703383	A/G	0.08	TFBS cluster
6	rs13217513	15731386	G/A	0.20	TFBS cluster
7	Alt_trans_SNP_p1	15735080	C/A	0	Alt Transcript Putative Promoter
8	rs16876738	15735532	G/C	0.17	Alt Transcript Putative Promoter
9	rs12525702	15735750	C/T	0.09	Alt Transcript Putative Promoter
10	rs3213207	15736081	A/G	0.12	Alt Transcript Putative Promoter
11	Alt_trans_SNP_p4b1	15736141	C/T	0.03	Alt Transcript Putative Promoter
12	Alt_trans_SNP_p5	15736261	T/G	0.09	Alt Transcript Putative Promoter
13	rs1474605	15766191	A/G	0.19	Main Transcript Putative Promoter
14	DTNBP1-C1_SNP1	15766308	T/C	0	Main Transcript Putative Promoter
15	DTNBP1-C2_SNP1	15766584	C/T	0.09	Main Transcript Putative Promoter
16	rs12196958	15766584	T/A	0.01	Main Transcript Putative Promoter
17	rs1997679	15766884	C/T	0.28	Main Transcript Putative Promoter
18	rs13192791	15767518	C/T	0.20	Main Transcript Putative Promoter
19	rs909706	15768850	C/T	0.37	Main Transcript Putative Promoter
20	rs2619516	15768930	A/G	0.08	Main Transcript Putative Promoter
21	DTNBP1-C9b_SNP1	15769372	G/C	0.09	Main Transcript Putative Promoter
22	rs9476886	15769440	C/T	0.25	Main Transcript Putative Promoter
23	rs12194321	15769507	C/T	N/A	Main Transcript Putative Promoter
24	DTNBP1-C9b_SNP2	15769573	T/C	0.03	Main Transcript Putative Promoter
25	DTNBP1-C12_SNP1	15770552	G/C	0	Main Transcript Putative Promoter
26	DTNBP1-C13_SNP1	15770555	G/C	N/A	Main Transcript Putative Promoter
27	DTNBP1-C13_SNP2	15770558	C/G	0	Main Transcript Putative Promoter
28	rs9476887	15770870	C/T	0	Main Transcript Putative Promoter
29	rs11558324	15771097	T/C	0.22	Main Transcript Putative Promoter
30	DTNBP1-C16_SNP1	15771705	C/G	0	Main Transcript Putative Promoter
31	rs2619536	15771826	T/C	0.09	Main Transcript Putative Promoter
32	DTNBP1-C16_SNP2	15771883	C/T	0.09	Main Transcript Putative Promoter
33	rs2619537	15772392	A/G	0.16	Main Transcript Putative Promoter
34	rs12204704	15773184	G/A	0.16	Main Transcript Putative Promoter
35	rs2619538	15773188	T/A	0.39	Main Transcript Putative Promoter
36	DTNBP1-C23_SNP1	15774468	C/G	0	Main Transcript Putative Promoter
37	DTNBP1-C24b_SNP1	15774812	G/T	N/A	Main Transcript Putative Promoter
38	DTNBP1-C26_SNP1	15775663	G/A	N/A	Main Transcript Putative Promoter
39	DTNBP1-dhyp_SNP1	15784456	T/C	0.01	DNase HS
40	rs17407828	15784456	G/C	0.29	DNase HS
41	DTNBP1-R10_SNP1	15794357	T/C	0.22	TFBS cluster
42	DTNBP1-R10_indel1	15794358	C/T	0.28	TFBS cluster
43	DTNBP1-R10_indel2	15794409	A/G	0.29	TFBS cluster
44	rs9296989	15794678	A/G	0.43	TFBS cluster
45	rs6906528	15798293	G/A	0.28	TFBS cluster
46	rs10949310	15800916	C/G	N/A	TFBS cluster
47	rs7450621	15806527	C/T	0.45	CNC
48	rs12212477	15820927	C/A	0.43	CNC
49	DTNBP1-R16_microsat	15841789	ACn	N/A	CNC
50	DTNBP1-R16_SNP1	15841944	C/T	0.02	CNC
51	DTNBP1-R17_SNP1	15843067	C/T	0.02	TFBS cluster
52	rs2025926	15843264	G/T	0.03	TFBS cluster
53	DTNBP1-R17_SNP2	15843267	G/A	0.02	TFBS cluster
54	rs5874525	15843366	G/-	N/A	TFBS cluster
55	rs3070184	15843367	G/A	N/A	TFBS cluster
56	rs2619514	15843987	T/C	0.06	TFBS cluster

Table 3.1. Sequence variants identified within putative DTNBP1 regulatory regions.
SNP positions are according to UCSC human genome chromosome 6 reference sequence (March 2006 freeze). Minor allele frequencies (MAF) in the HapMap CEU sample are shown. 24 novel polymorphisms were identified during this study. N/A indicates not available and is given where a DNA variant failed to be typed through the CEU population.

3.3.4 Tag SNP Selection

Tag SNP selection was performed in two stages. Firstly, a non-redundant set of SNPs was selected which captured the 48 variants (with CEU genotype data) identified within the putative regulatory regions at a MAF >0.001 and at an r^2=1. Secondly, extra SNPs were selected that would tag the core DTNBP1 locus (Chr6:15621211-15780522) at a MAF >0.001 and at an r^2>0.95. This combined tagging was employed due to the potential ambiguity of the regulatory region identification methods. It was hoped that tagging the core DTNBP1 locus in addition to the putative regulatory variants would maximise the chance of identifying DTNBP1 risk variant(s) and more specifically functional variant(s). Supplementing tagging of the core DTNBP1 locus with the variants identified by direct screening also has a number of advantages over solely selecting tag SNPs for the core DTNBP1 locus using the HapMap database alone. Firstly, direct sequencing of putative regulatory regions identified variants not deposited in the HapMap database. Secondly, the "window" of genomic sequence screened for putative regulatory regions covered a much larger area than the tagged DTNBP1 region (Figure 3.7). While tagging the larger region at the high density employed for the core DTNBP1 locus would have greatly increased the number of tag SNPs to be genotyped, including the putative regulatory variants increased coverage and enriched the SNPs analysed for putative regulatory variants while keeping the number of tag SNPs to be genotyped to a realistic level.

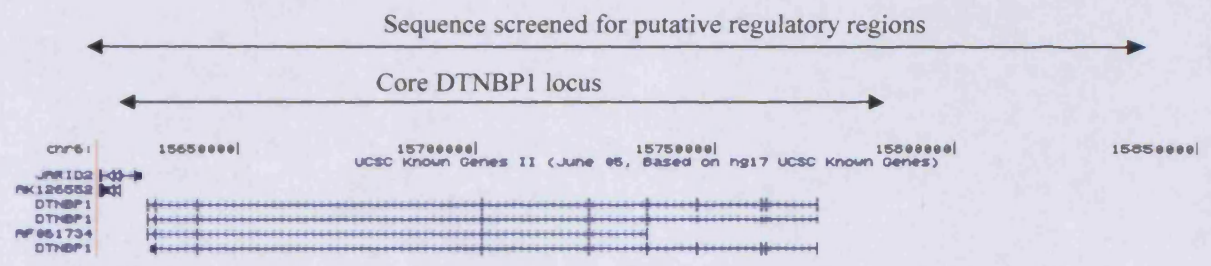


Figure 3.7. Comparison of the genomic sequence screened for putative regulatory regions (chr6:15,621,018-15,844,250) and the core DTNBP1 locus (chr6:15621211-15780522) tagged at an r^2<0.95 MAF<0.001.

Pairwise tagging selected a set of 66 tag SNPs which captured 100% of the variation at chr6:15621211-15780522 with an r^2>0.95 and MAF>0.001 plus all putative regulatory variants at an r^2=1. However due to genotyping failure (using both Sequenom and Amplifluor) a final set of 64 tag SNPs were genotyped which captured 94% of the variation at an r^2>0.95 and MAF>0.001. Of the 48 putative regulatory SNPs, 37 were either genotyped themselves or tagged at an r^2=1 with a SNP typed through the association sample. Of the remaining of 11 putative regulatory variants, 7 had MAF<0.001 and 4 SNPs (rs1076635, alt_trans_SNP_p4b1, DTNBP1-C9b_SNP2, DTNBP1-R10_SNP1) failed to be typed by Sequenom or Amplifluor and no proxies were available.

3.3.5 Association Analysis

Of the 64 tag SNPs genotyped across the DTNBP1 locus and putative regulatory regions, all had a call rate of >95%. Individuals were only included in analysis if they had genotypes for over 80% of the 64 SNPs analysed, therefore 18 individuals were removed from the sample prior to association analysis. All markers were in HW equilibrium at p>0.001 for both cases and controls. However to compensate for even small deviations in HW, the Armitage trend test was used to test for allelic association (Table 3.2).

Three polymorphisms showed nominal allelic association with schizophrenia (rs4715984 p=0.001, rs2619538 p=0.024, rs9296989 p=0.017), although rs4715984 was the only SNP to survive correction for multiple testing after 10,000 permutations (p=0.04). All three of these polymorphisms also showed genotypic association (rs4715984 corrected p=0.003, rs2619538 p=0.016, rs9296989 p=0.038). One SNP rs3778651 showed genotypic (corrected p=0.011) but not allelic association.

No	rsID	Chrom Pos	Alleles	Freq Cases	Freq Controls	Arm Trend P	Arm Trend Emp P (10000)	Counts cases	Counts Controls	HW Cases & Controls	HW Cases	HW Con	Genotypic p	Clump P 10000
1	rs2235258	15621461	G/A	0.2301	0.2323	0.8901	1	45/210/397	38/253/417	0.127	0.026	1.000	0.250	N/A
2	rs9396589	15621723	T/G	0.2709	0.2876	0.3248	0.9999	45/266/346	55/299/357	0.383	0.555	0.522	0.616	N/A
3	rs9654600	15622092	A/T	0.0699	0.0777	0.4378	1	7/78/573	2/107/605	0.552	0.032	0.302	NA	0.052
4	rs9396590	15622478	G/A	0.2753	0.2922	0.3247	0.9999	45/269/338	59/294/352	0.505	0.433	0.856	0.546	N/A
5	rs1076636	15623763	T/C	0.2096	0.2202	0.5186	1	36/195/406	36/227/416	0.072	0.056	0.502	0.546	N/A
6	rs9396591	15624264	A/G	0.2733	0.2894	0.3403	0.9999	44/270/341	55/301/354	0.229	0.376	0.466	0.627	N/A
7	rs742102	15624612	A/G	0.0320	0.0358	0.5979	1	2/38/616	2/47/664	0.067	0.137	0.224	NA	0.870
8	rs1474587	15624688	T/C	0.1508	0.1474	0.8063	1	21/155/477	20/169/520	0.025	0.066	0.178	0.913	N/A
9	rs909626	15625663	T/C	0.1368	0.1299	0.6071	1	14/152/492	19/147/546	0.046	0.619	0.029	0.467	N/A
10	rs3778651	15626511	T/C	0.0578	0.0721	0.1306	0.9731	5/66/586	1/101/612	0.826	0.058	0.165	NA	0.011
11	rs13213814	15627355	C/T	0.2656	0.2588	0.6843	1	49/251/357	45/279/389	1	0.617	0.627	0.693	N/A
12	rs13198512	15627864	T/C	0.5046	0.4846	0.2927	0.9999	169/323/163	161/368/183	0.705	0.755	0.369	0.385	N/A
13	rs13201824	15628757	T/G	0.2647	0.2556	0.594	1	49/244/353	46/270/392	0.621	0.479	1.000	0.736	N/A
14	rs1047631	15631080	C/T	0.1415	0.1421	0.963	1	13/158/479	9/183/515	0.266	1.000	0.123	0.479	N/A
15	rs17470454	15631427	A/G	0.0631	0.0497	0.1308	0.9736	2/79/577	3/65/646	0.613	1.000	0.410	NA	0.206
16	rs742106	15632459	A/G	0.3548	0.3622	0.6835	1	78/306/267	83/349/279	0.099	0.549	0.105	0.741	N/A
17	rs2056943	15632542	C/T	0.0464	0.0337	0.0855	0.913	1/59/597	0/48/664	0.721	1.000	1.000	NA	0.141
18	rs16876571	15632658	A/G	0.0130	0.0179	0.3005	0.9999	0/17/636	0/25/673	1	1.000	1.000	NA	0.349
19	rs16876575	15633136	T/C	0.1438	0.1273	0.2031	0.9961	15/159/483	8/165/538	0.729	0.635	0.310	0.215	N/A
20	rs6937379	15633976	A/G	0.2344	0.2525	0.2604	0.9991	36/236/385	34/292/387	0.107	1.000	0.029	0.156	N/A
21	rs4712253	15634396	T/C	0.4085	0.3787	0.1098	0.95	115/306/235	96/348/269	1	0.375	0.340	0.114	N/A
22	rs9464795	15638143	A/G	0.0809	0.0848	0.7192	1	6/93/550	6/106/584	0.285	0.298	0.622	0.893	N/A
23	rs9370822	15652715	C/A	0.3941	0.3827	0.5318	1	102/313/241	86/373/253	0.041	1.000	0.004	0.099	N/A
24	rs12527121	15654192	T/C	0.0968	0.0850	0.2791	0.9996	6/115/535	4/113/595	0.869	1.000	0.809	NA	0.560
25	rs9370823	15658637	G/A	0.2874	0.2956	0.6293	1	49/275/325	51/313/338	0.056	0.443	0.070	0.714	N/A
26	rs9296983	15663405	A/G	0.2266	0.2464	0.2195	0.9977	32/232/389	36/272/390	0.259	0.823	0.221	0.384	N/A
27	rs4715984	15669870	A/G	0.1107	0.0756	0.0013	0.0471	6/133/516	2/103/603	0.330	0.552	0.417	NA	0.003
28	rs9358063	15673010	T/C	0.4671	0.4698	0.8865	1	141/315/183	136/382/178	0.088	0.812	0.008	0.124	N/A

Table 3.2. Case/Control association analysis of 64 tag SNPs spanning the DTNBP1 locus and putative regulatory regions. SNP positions are according to UCSC human genome chromosome 6 reference sequence (March 2006 freeze). Alleles given as Minor/Major alleles. N/A= not analysed.

No	rsID	Chrom Pos	Alleles	Freq Cases	Freq Controls	Arm Trend P	Arm Trend Emp P (10000)	Counts cases	Counts Controls	HW Cases & Controls	HW Cases	HW Con	Genotypic p	Clump P 10000
29	rs7771339	15677985	A/G	0.04497	0.0407	0.5823	1	1/57/598	1/56/655	1	1.000	1.000	NA	0.636
30	rs2743548	15691803	G/T	0.07840	0.0898	0.2895	0.9999	6/91/560	6/116/591	0.384	0.276	0.820	0.457	N/A
31	rs9296985	15697994	C/T	0.0804	0.0878	0.4959	1	6/93/554	6/110/579	0.376	0.295	0.641	0.716	N/A
32	rs12203173	15705548	G/A	0.1476	0.1239	0.0695	0.8675	16/161/477	7/163/544	0.729	0.537	0.226	0.070	N/A
33	rs12199640	15706859	T/C	0.0356	0.0437	0.2953	0.9999	3/39/590	1/59/638	0.153	0.039	1.000	NA	N/A
34	rs7752070	15712898	G/A	0.0810	0.0915	0.3539	1	9/88/557	11/107/587	0.002	0.029	0.037	0.630	0.224
35	rs3829893	15723616	A/G	0.1549	0.1377	0.2087	0.9966	19/159/458	10/175/523	0.913	0.287	0.343	0.134	N/A
36	rs2619539	15728834	C/G	0.4765	0.4601	0.3916	1	150/307/180	135/376/191	0.381	0.384	0.048	0.084	N/A
37	rs3213207	15736081	C/T	0.1133	0.1075	0.6289	1	7/131/502	4/143/555	0.164	0.843	0.118	NA	N/A
38	Alt_trans_p5_SNP1	15736261	A/C	0.0338	0.0459	0.1203	0.9633	2/40/609	3/58/637	0.059	0.162	0.171	0.120	0.593
39	rs1011313	15741411	T/C	0.1000	0.0821	0.1080	0.9487	9/113/533	6/105/602	0.248	0.277	0.462	0.268	0.273
40	rs13212086	15742115	T/C	0.2395	0.2518	0.4533	1	35/238/370	39/276/388	0.304	0.746	0.316	0.677	N/A
41	rs2619528	15757808	T/C	0.1888	0.1923	0.8186	1	22/203/429	23/227/460	0.483	0.800	0.471	0.931	N/A
42	rs760761	15759111	A/G	0.1940	0.1987	0.753	1	22/209/421	24/235/453	0.303	0.616	0.410	0.931	N/A
43	rs2619520	15764134	C/A	0.1107	0.0906	0.0857	0.9126	12/121/522	6/117/589	0.228	0.113	0.822	0.154	N/A
44	rs1018381	15765049	A/G	0.0790	0.0845	0.6033	1	6/91/555	5/108/585	0.467	0.279	1.000	0.682	N/A
45	rs1474605	15766191	C/T	0.1948	0.1959	0.9411	1	22/212/423	23/233/456	0.230	0.535	0.342	0.979	N/A
46	rs12196958	15766584	A/T	0.0084	0.01471	0.1313	0.9766	1/9/648	0/21/693	0.167	0.041	1.000	NA	N/A
47	rs1997679	15766884	A/G	0.3208	0.3411	0.2604	0.9991	72/275/306	76/331/301	0.903	0.419	0.316	0.212	0.054
48	DTNBP1_c9b_SNP1	15769372	C/G	0.0898	0.0870	0.7877	1	4/110/543	1/122/590	0.063	0.810	0.033	NA	N/A
49	rs9476886	15769440	T/C	0.2676	0.2433	0.1484	0.9843	52/245/355	40/263/402	0.617	0.317	0.838	0.217	0.396
50	rs11558324	15771097	C/T	0.2393	0.2661	0.1069	0.9469	35/242/375	49/273/375	0.773	0.668	1.000	0.259	N/A
51	rs2619536	15771826	C/T	0.0919	0.1023	0.3656	1	9/102/542	6/133/570	0.535	0.101	0.685	0.213	N/A
52	DTNBP1_c16_SNP2	15771883	T/C	0.0395	0.0413	0.8170	1	3/46/609	3/53/658	0.021	0.074	0.113	NA	N/A
53	rs2619537	15772392	G/A	0.1322	0.1457	0.3117	0.9999	14/146/498	14/180/520	0.736	0.395	0.880	0.419	0.924
54	rs12204704	15773184	A/G	0.1816	0.1548	0.0630	0.8369	26/187/445	17/187/510	0.384	0.293	1.000	0.133	N/A
55	rs2619538	15773188	A/T	0.4698	0.4264	0.0238	0.5134	151/305/190	121/361/225	0.869	0.180	0.282	0.016	N/A

Table 3.2 cont. Case/Control association analysis of 64 tag SNPs spanning the DTNBP1 locus and putative regulatory regions. SNP positions are according to UCSC human genome chromosome 6 reference sequence (March 2006 freeze). Alleles given as Minor/Major alleles. N/A= not analysed.

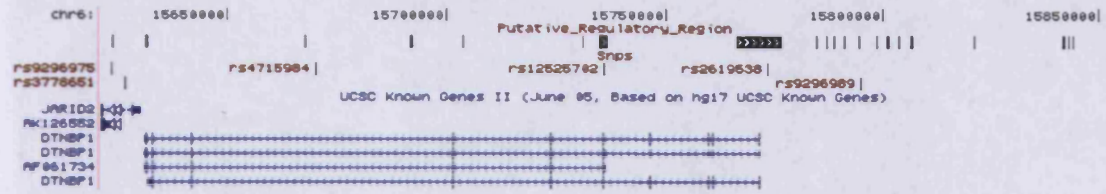
No	rsID	Chrom Pos	Alleles	Freq Cases	Freq Controls	Arm Trend P	Arm Trend Emp P (10000)	Counts cases	Counts Controls	HW Cases & Controls	HW Cases	HW Con	Genotypic p	Clump P 10000
56	DTNBP1_Dhypsnp1	15784456	C/T	0.0129	0.0098	0.4549	1	1/15/642	0/14/699	0.158	0.099	1.000	NA	N/A
57	rs17407828	15784456	C/G	0.3113	0.3379	0.1439	0.982	69/268/315	77/317/303	0.664	0.315	0.735	0.194	0.377
58	DTNBP1_r10_SNP1	15794357	C/T	0.1370	0.1419	0.7086	1	9/162/486	12/178/522	0.259	0.324	0.541	0.879	N/A
59	rs9296989	15794678	G/A	0.4648	0.4195	0.0175	0.4253	145/317/191	120/354/234	1	0.530	0.537	0.038	N/A
60	rs7450621	15806527	T/C	0.4977	0.4651	0.0918	0.9222	165/318/168	151/352/200	0.744	0.583	0.940	0.209	N/A
61	rs12212477	15820927	A/C	0.4363	0.4105	0.1791	0.9933	128/312/211	120/338/246	0.540	0.524	0.876	0.385	N/A
62	DTNBP1_r16_SNP1	15841944	T/C	0.0626	0.0509	0.1876	0.9946	2/77/568	2/68/637	1	1.000	0.701	NA	0.424
63	rs2025926	15843264	A/C	0.0123	0.00945	0.4701	1	0/16/633	0/13/675	1	1.000	1.000	NA	0.576
64	rs2619514	15843987	G/A	0.0965	0.0765	0.0660	0.8511	8/110/535	3/101/595	0.729	0.368	0.789	NA	0.129

Table 3.2 cont. Case/Control association analysis of 64 tag SNPs spanning the DTNBP1 locus and putative regulatory regions. SNP positions are according to UCSC human genome chromosome 6 reference sequence (March 2006 freeze). Significantly associated SNPs highlighted in red. The most significantly associated SNP rs4715984 is in bold and highlighted in red. Alleles given as Minor/Major alleles. N/A= not analysed.

3.3.6 Location of Associated SNPs

The chromosomal location of all significant polymorphisms and their proxies (r^2>0.95) are given in Figure 3.8A. The SNP showing the greatest association to schizophrenia, rs4715984, is an intronic SNP not located within a previously identified putative regulatory region. However rs12525702, a proxy of rs4715984, is located within the putative promoter region of the alternative transcript AF061734 and only 200bp upstream of the TSS (Figure 3.8B). Therefore rs12525705 could potentially have a regulatory effect on this DTNBP1 transcript. Both rs2619538 and rs9296989 are also within putative regulatory regions. As can be seen from Figure 3.8A, rs2619538 is located ~1.8kb upstream of the TSS of the three main DTNBP1 transcripts. Although not as close to the TSS as rs12525702, this SNP could also potentially have a regulatory effect on one or all of these transcripts. The polymorphism rs9296989 is within a region determined by the Cluster Buster program. However reanalysis showed that rs9296989 is not located within any of the predicted TFBSs within this region. rs3778651 which shows a genotypic association to schizophrenia is not located with putative regulatory regions, nor is its proxy rs9296975.

A



B

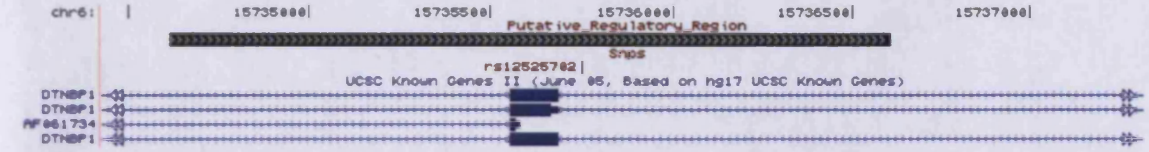


Figure 3.8.
A. Chromosomal position of significant SNPs and proxies (r^2>0.95) plus putative regulatory regions previously identified. rs12525705, a proxy for rs4715984 (p=0.001) is located within the putative promoter region of the alternative transcript AF061734. rs2619538 and rs9296989 are also located within putative regulatory regions. rs4715984 is not located with any putative regulatory regions, nor is rs3778651 or its proxy rs9296975. Based upon May 2004 UCSC freeze.
B. Chromosomal location of rs12525702 on a narrowed scale. rs12525702 is a proxy for rs4715984 (p=0.001) and, is ~200bp upstream of the alternative transcript AF061734. Based upon May 2004 UCSC freeze.

3.3.7 Further Analysis of the Allelic Association Signal

A high-density tagging strategy was used for this study (r^2>0.95). As a result the significant SNPs may be reporting the same association signal. LD analysis, illustrated in Table 3.3, shows that rs2619538 and rs9296989 are in a high LD (D'=1, r^2=0.92). However the LD between rs4715984 and rs2619538/rs9296989 is lower (D'=0.46/0.43, r^2=0.04/0.02).

	rs4715984	rs2619538	rs9296989
rs4715984		0.04	0.02
rs2619538	0.46		0.92
rs9296989	0.43	1	

Table 3.3. LD analysis of all significantly associated SNPs. The top half of the table shows r^2 values and D' values are given in the bottom half of the table.

In order to discern whether rs2619538 or rs9296989 show an association over and above rs4715984 (the most associated SNP) logistic regression was performed. Firstly the p value of each SNP was adjusted for the correlation between itself and all other associated SNPs. As can be seen from Table 3.4 only rs4715984 remains significant after this analysis (Enter Logistic Regression p=0.011).

Forward and backward stepwise logistic regression confirmed that the association observed by rs9296989 and rs2619538 shows no evidence for independence from rs4715984 as neither remain significant after adjustment for their correlation with rs4715984 (Stepwise Logistic Regression p=0.095 and p=0.107 respectively). Further details of the enter and stepwise logistic regression results are given in the Appendix (Tables 9.1.3-9.1.5).

SNP ID	Arm Trend P Value	Enter LR P Value	Stepwise LR P Value
rs4715984	0.001	0.011	-
rs9296989	0.018	0.656	0.095
rs2619538	0.024	0.919	0.107

Table 3.4. Logistic Regression Analysis of SNPs showing an allelic association. LR = Logistic Regression. Arm = Armitage. Enter LR p value refers to p values adjusted for the correlation between the test SNP and all other significant SNPs. Stepwise LR p values are adjusted for the correlation between the test SNP and rs4715984.

3.3.7 Further Analysis of the Allelic Association Signal

A high-density tagging strategy was used for this study (r^2>0.95). As a result the significant SNPs may be reporting the same association signal. LD analysis, illustrated in Table 3.3, shows that rs2619538 and rs9296989 are in a high LD (D'=1, r^2=0.92). However the LD between rs4715984 and rs2619538/rs9296989 is lower (D'=0.46/0.43, r^2=0.04/0.02).

	rs4715984	rs2619538	rs9296989
rs4715984		0.04	0.02
rs2619538	0.46		0.92
rs9296989	0.43	1	

Table 3.3. LD analysis of all significantly associated SNPs. The top half of the table shows r^2 values and D' values are given in the bottom half of the table.

In order to discern whether rs2619538 or rs9296989 show an association over and above rs4715984 (the most associated SNP) logistic regression was performed. Firstly the p value of each SNP was adjusted for the correlation between itself and all other associated SNPs. As can be seen from Table 3.4 only rs4715984 remains significant after this analysis (Enter Logistic Regression p=0.011).

Forward and backward stepwise logistic regression confirmed that the association observed by rs9296989 and rs2619538 shows no evidence for independence from rs4715984 as neither remain significant after adjustment for their correlation with rs4715984 (Stepwise Logistic Regression p=0.095 and p=0.107 respectively). Further details of the enter and stepwise logistic regression results are given in the Appendix (Tables 9.1.3-9.1.5).

SNP ID	Arm Trend P Value	Enter LR P Value	Stepwise LR P Value
rs4715984	0.001	0.011	-
rs9296989	0.018	0.656	0.095
rs2619538	0.024	0.919	0.107

Table 3.4. Logistic Regression Analysis of SNPs showing an allelic association. LR = Logistic Regression. Arm = Armitage. Enter LR p value refers to p values adjusted for the correlation between the test SNP and all other significant SNPs. Stepwise LR p values are adjusted for the correlation between the test SNP and rs4715984.

3.3.8 Haplotype Analysis

Haplotype analysis was performed on the 64 markers genotyped at this locus. Analysis of all 2 and 3 marker haplotypes (45,760 phased haplotypes) revealed the strongest evidence for association with a three marker haplotype composed of the SNPs rs17470454, rs4715984 and rs2619520 (global p=0.00025). Analysis of the specific 3 marker haplotypes formed by these SNPs identified both a risk haplotype (GAA p=0.0014, case frequency = 0.11, control frequency = 0.076) plus a protective haplotype (GGA p=0.00002, case frequency = 0.71, control frequency = 0.78). However the risk haplotype does not refine the allelic association observed with rs4715984 alone (p=0.001, cases frequency (A) = 0.11, control frequency = 0.076). In addition as ~45,000 two and three marker haplotypes were analysed, the best p values shown here would not survive a Bonferroni correction for multiple comparisons (p>0.05).

3.3.8.1 Analysis of Previous Risk Haplotype

The association sample used in this chapter is an extended but largely overlapping version of the UKCC sample analysed by Williams *et al* [134]. Therefore in addition to identifying new risk variants the significant results reported in the previous study were analysed in this extended sample. Williams *et al* determined significant evidence for association in one risk haplotype (rs2619538, rs3213207, rs2619539 TAC, p=0.01) and two protective haplotypes (AAC p=0.006, TGG p<0.001). Analysis of these three risk haplotypes in this extended case control sample are given in Table 3.5. While the rare protective haplotype is no longer associated with schizophrenia (TGG p=0.758) the other protective haplotype still shows an association in this sample (AAC p=0.016). Furthermore the original risk haplotype is also associated with schizophrenia in this extended sample (TAC, p=0.048).

rs4715984 the most significantly associated SNP in this study was not genotyped in the original publication by Williams and colleagues. Haplotype analysis was therefore performed with the previous three markers and rs4715984. As can be seen in Table 3.5

the four marker haplotype combination (rs2619538, rs3213207, rs2619539, rs4715984) has a globally significant p value (p=0.016). The addition of rs4715984 splits individuals carrying the risk TAC haplotype into carriers of the TACT or TACC haplotypes. The associated allele of rs4715984 (T) occurs on the same haplotype background as the risk haplotype TAC and refines the association (TACT p=0.005). The protective haplotype also shows a refined signal as the protective allele of rs4715984 is found on the AAC haplotype (AACC p=0.003).

However it must be noted that while addition of rs4715984 refines the risk haplotype association, only 2% of the 3% difference observed for the TAC haplotype is seen for the TACT haplotype. The remaining 1% difference, while not great enough for a significant association, is shown by the TACC haplotype.

A

Markers	Haplotype	Case	Control	P Value
rs2619539-rs3213207-rs2619538	Global	N/A	N/A	0.100
rs2619539-rs3213207-rs2619538	**TAC**	**0.19**	**0.16**	**0.048**
rs2619539-rs3213207-rs2619538	AAC	0.34	0.38	0.016

B

Markers	Haplotype	Case	Control	P Value
rs2619538-rs3213207-rs2619539-rs4715984	Global	N/A	N/A	0.016
rs2619538-rs3213207-rs2619539-rs4715984	**TACT**	**0.09**	**0.07**	**0.005**
rs2619538-rs3213207-rs2619539-rs4715984	TACC	0.10	0.09	0.588
rs2619538-rs3213207-rs2619539-rs4715984	AACT	0.02	0.01	0.303
rs2619538-rs3213207-rs2619539-rs4715984	AACC	0.32	0.37	0.003

Table 3.5. A Haplotype analysis results for the previously identified risk haplotype [134] with the extended association sample. B. The refinement of the risk haplotype association by the addition of rs4715984. Given are the marker IDs, haplotypes, case and control frequencies and an association p value. Risk haplotypes are given in red, protective haplotypes in blue.

3.4 Discussion

Recent studies suggest that complex diseases may be caused by the variation in expression of multiple genes and that pathogenic mutations will function not by altering the protein structure, but by affecting the levels of specific susceptibility genes [192, 322, 323]. The gene dysbindin (DTNBP1) shows all the features of a gene that could promote susceptibility to schizophrenia through altered expression. Firstly, while DTNBP1 has been widely reported as associated with schizophrenia, no obvious pathogenic mutations have been found [324]. Secondly dysbindin expression, both at the mRNA and protein level, is reduced in schizophrenic brains [181-183]. Finally, reductions in dysbindin mRNA have been shown to be associated with a DTNBP1 schizophrenia risk haplotype [186]. This evidence suggests that the causal variants within DTNBP1 could be located in cis-acting regulatory elements.

To date there has been no comprehensive study of the DTNBP1 locus for genomic sequence that may have a regulatory effect on the expression of dysbindin. This chapter describes the screening of a "window" around the DTNBP1 gene for functional variants. This involved the initial identification of putative regulatory regions followed by polymorphism detection within these regions. *In silico* analysis for regulatory sequence identified 21 putative regulatory regions and mutation screening of these regions detected 56 polymorphisms. These were combined with 216 SNPs at the DTNBP1 locus (Chr6:15621211-15780522) which had been genotyped in HapMap phase II. A defined a set of highly informative SNPs that captured the majority of common variation at this locus was then genotyped in a relatively ethnically homogeneous schizophrenia case control sample.

Allelic and genotypic analysis identified four SNPs associated with schizophrenia. However logistic regression analysis determined that the allelic association observed with rs2619538 and rs9296989 is not independent of the association signal detected at rs4715984. While this indicates a single association signal in this case control sample, it is worth noting that rs9296989 shows a trend towards a signal independent of rs4715984 (p=0.095). rs9296989 may therefore be in partial LD with another SNP, not

typed in this analysis, which shows an association independent of rs4715984. However, identifying this hypothetical SNP is beyond the scope of this thesis.

rs4715984 itself survives correction for multiple testing. It has also been found to refine the risk haplotype previously reported by Williams *et al [134]*. While rs4715984 does not reside within a putative regulatory element, analysis of the LD structure across DTNBP1 identified a proxy, rs12525702, which is found within the putative promoter region of the DTNBP1 isoform AF061734. rs12525702 is only 200bp upstream of the transcription start site of AF061734 and therefore an ideal candidate to cause cis-acting variation. As well as altering the binding of a transcription factor rs12525702 could cause expression differences by other mechanisms. Sequence variants that lie outside of transcription factor binding site may exert their effect by either altering the bending of DNA towards (or away from) its optimum configuration or by altering the flexibility of the DNA [322]. If this is the case it would make it equally likely that rs4715984 may be the causal variant.

While it is promising that a SNP located within a predicted regulatory region has been found to be associated with schizophrenia, there are many question marks over whether the determination of putative regulatory regions by the methods described, will be successful in identifying functional SNPs. The overlap in the regulatory regions identified by the different methods described in this chapter was relatively low. This may suggest that one or all of the methods used are not exceptionally efficient at determining regulatory elements and are producing a high number of false positives and/or negatives. However the pattern of regions could also be explained if each method is detecting a different type of regulatory element.

Another caveat of this study is that the analysis of the DTNBP1 locus for regulatory variants has only screened a small fraction of sequence that could potentially affect DTNBP1 expression. For example while it was decided to limit the sequence analysed 3' to DTNBP1 to 10kb, due to the presence of the gene JARID2 1kb 3' to DTNBP1, there have been reports of a number of regulatory elements for specific genes have been found within neighbouring genes [288]. Secondly, cis-acting variants have been

observed over 1Mb away from the target gene [265], however screening this amount of genomic sequence for putative DTNBP1 regulatory regions would have been unfeasible.

In order to establish whether identifying putative regulatory elements via the *in silico* analysis used in this chapter is a successful way of determining cis-acting variants, and also whether the risk polymorphisms identified in this chapter are causal variants, further investigation is needed. Therefore all informative SNPs examined in this chapter have been subjected to additional analysis which was designed to verify whether any of these SNPs could affect dysbindin expression. This consisted of association analysis with the DTNBP1 allelic expression data previously described [186]. In addition SNP(s) showing association with the allelic expression data, and deemed most likely to be influencing the expression differences observed, were examined in a functional reporter gene assay. These analyses are described in chapters 4 and 5 respectively.

Chapter 4: Association Analysis of DTNBP1 Putative Regulatory SNPs with cortical Dysbindin mRNA Levels

4.1 Introduction

As described in Chapter 1, Bray *et al* previously provided evidence that DTNBP1 is under the influence of cis-acting variation [184]. As reduced DTNBP1 expression had also been reported in post mortem brains of schizophrenic patients [181-183], it was suggested that cis-acting variation may be relevant to schizophrenia aetiology and that unidentified sequence variants in regulatory regions of DTNBP1, might promote susceptibility to schizophrenia by altering expression of the gene. In an attempt to link the observations of cis-acting variation with the genetic association previously reported between DTNBP1 and schizophrenia [134], Bray and colleagues showed that the relative expression of the A-allele of the exonic SNP rs1047631, used in the allelic expression assay, was significantly lower when it was carried on the schizophrenia risk haplotype TAA (T allele of rs2619538, A allele of rs3213207 and A allele of rs1047631, p=0.01) compared with when it was carried on the non-risk haplotypes [186].

However the report also noted that the risk haplotype did not account for all the cis-acting variation in DTNBP1 expression as a proportion of samples that did not carry the risk haplotype showed allelic distortion (Figure 4.1). In light of this it was suggested that the risk haplotype was unlikely to have a direct effect on DTNBP1 expression and may be in LD with the actual functional variant(s).

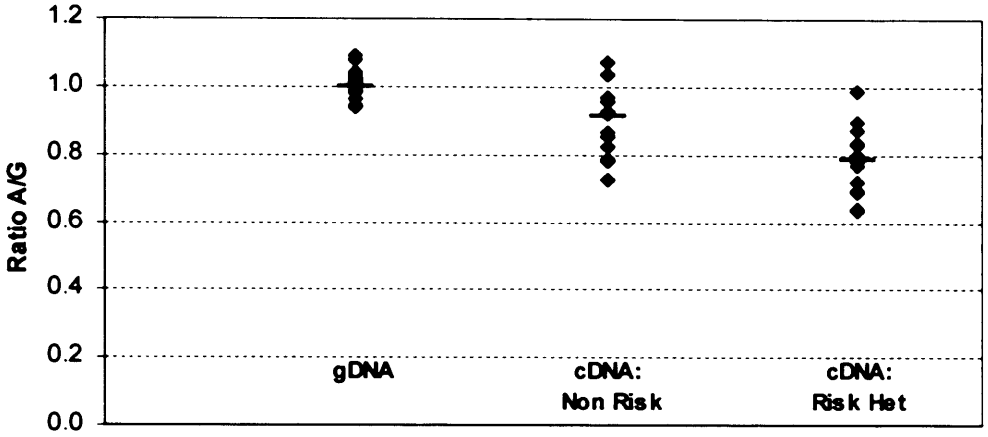


Figure 4.1. Allele ratios at SNP rs1047631, stratified by heterozygosity for the defined 3-marker schizophrenia risk haplotype. Data represented as a ratio of A/G alleles. The risk haplotype includes the reduced expression A-allele of rs1047631. cDNA samples from individuals who are heterozygous for the risk haplotype show lower expression of the A-allele than those from individuals who carry no copies of the risk haplotype (p=0.002). However a number of individuals homozygous for the risk haplotype also show allelic expression differences (average allelic expression ratio =0.95) including three individuals with a relative decrease in expression of greater than 20% [186].

Chapter 3 describes the identification of a set of SNPs that would allow the high density mapping of the DTNBP1 locus and putative DTNBP1 regulatory regions. The initial aim of this chapter was to identify potential cis-acting variants by determining which of these SNPs best correlate with the allelic expression data reported by Bray *et al* [186]. The second phase of the analysis was to determine whether any of these SNPs were also associated with schizophrenia. Of particular interest was rs4715984 which was observed in chapter 3 to show the greatest association with schizophrenia (p=0.0013) and survived correction for multiple testing (p=0.0471). Moreover, this SNP was found to refine the association of the schizophrenia risk haplotype, originally identified by Williams and colleagues [134], which was subsequently shown to also be associated with reduced DTNBP1 expression. In light of this the refined risk haplotype was also analysed for correlation with allelic expression differences.

4.2 Methods

4.2.1 Caucasian Brain Samples

Details of the 149 individuals which form the Caucasian brain sample used in this chapter are given in chapter 2.1.2. The extraction protocols used for these samples are described in chapter 2.2.

4.2.2 Identification of Putative Regulatory Regions, Polymorphism Detection and Tag SNP Identification

The identification of putative regulatory regions and the screening of these regions is described in chapter 3. The criteria used to determine the tag SNPs analysed in this chapter is also given. Briefly SNPs were selected that would tag variants located within the putative regulatory regions at an $r^2=1$. In addition extra polymorphisms were chosen which would tag the DTNBP1 locus (Chr6:15621211-15780522) at an $r^2>0.95$. Both criteria had a MAF>0.001.

4.2.3 Genotyping

Tag SNPs were genotyped through the Caucasian brain sample using the Sequenom MassARRAY genotyping system. Details of this protocol are given in chapter 2.7.1.

4.2.4 Allelic Expression Analysis

Allelic expression analysis of DTNBP1 was performed by Dr. Nicholas Bray using the coding SNP rs1047631, the results of which have previously been published [186]. Details of allelic expression protocol can be found in the chapter 2.9.

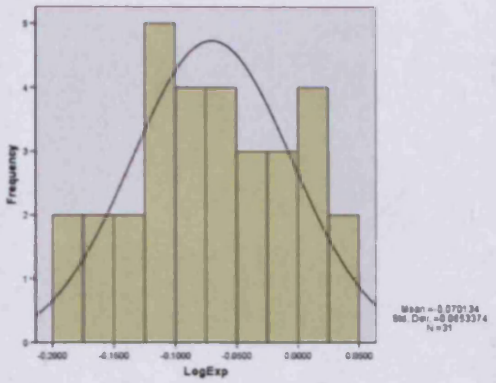
DNA samples from heterozygous subjects were assayed twice, with each assay consisting of two separate cDNA samples and one gDNA sample for each heterozygote. As a result, the genomic DNA ratios reported in this chapter are the average of two measurements and cDNA ratios given for each sample are the averages of four measurements. Samples were amplified using primers based on exonic sequence capable of amplifying both genomic DNA or cDNA:
5'-GTGGTGAGGACAGCGACTCT-3' and
5'-GCTGTTCTTTAAGTTTCTCACACA-3'. Allele representation was measured by primer extension and SNaPshot chemistry (Applied Biosystems) using the extension primer 5'-TTCTCACACATTATTGGCAATTA-3'.

4.2.5 Normalisation of Allelic Expression Data

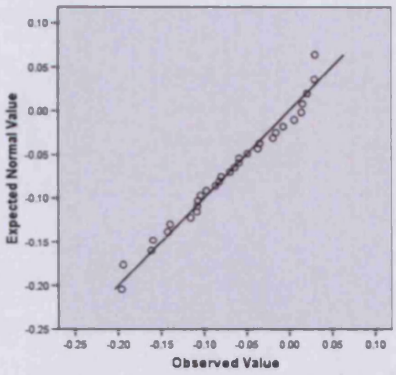
All prior statistical analysis of the allelic expression data has been performed using non-parametric tests for example Mann-Whitney U [186]. Non-parametric tests are used when a numerical outcome is produced but the dataset is not normally distributed. In these situations, non-parametric analysis, such as the Mann-Whitney U, can be more robust than parametric tests as they are less affected by extreme observations. However on normally distributed data parametric analysis has greater power as all information within a dataset is conserved, unlike non-parametric tests where numerical variables are converted into ranks. For this reason steps were taken to normalise the allelic expression data so subsequent analyses could be performed using parametric tests. Allelic expression ratios (n=31) were transformed by taking the natural log of each ratio. Histogram and Q-Q plots of the transformed ratios as well as statistical tests for normality (Figure 4.2) illustrate that the log transformation of the allelic expression

ratios follows a normal distribution. Subsequent analysis of stratified ratios was therefore achieved via parametric tests.

A.



B.



C.

Tests of Normality

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
LogExp	.075	31	.200*	.964	31	.374

*. This is a lower bound of the true significance.

a. Lilliefors Significance Correction

Figure 4.2. Normality tests on transformed (logarithmic) allelic expression ratios determined by analysis of individuals heterozygous for the SNP rs1047631. The results shown are A) Histogram B) Normal Q-Q plot of LogExp C) Statistical test results.

4.2.6 Statistical Analysis

Association analyses between cortical dysbindin mRNA levels and DTNBP1 tag SNP genotypes were performed on normalised allelic expression ratios (n=31) previously reported [186]. A statistically significant correlation was determined in two phases using the software package SPSS. Details of these analyses are given below.

4.2.6.1 Independent Samples T-test

Firstly, for each test SNP the relative expression of mRNA from each chromosome were compared between heterozygous and homozygous individuals. This was performed by comparing the normalised allelic expression ratios using an independent samples t-test. To determine whether equal or unequal variance should be assumed the Levene's test for homogeneity of variance was used. Homozygotes (whether major or minor alleles) were grouped together in this analysis as it would be expected that all individuals homozygous for a variant, whether functional or not, would show a relative allelic expression ratio of 1.

4.6.2.2 Direction of Effect Analysis

Where a significant difference was observed, the direction of effect (whether an allele is associated with reduced or increased expression) was established. This was determined by calculating the most probable diplotype between the test SNP and the allelic expression SNP rs1047631 for each individual. The probability of an individual carrying a particular diplotype was estimated using a diplotype probability algorithm designed by Dr. Valentina Moskvina. An example of the output from the diplotype probability program is given in Table 4.1. This program determines diplotype probabilities by summing the probabilities of the two relevant haplotypes. The probability of each haplotype is calculated by dividing the frequency of the given haplotype by the sum of the frequencies for all possible haplotypes, given the observed genotypes at each locus [195]. Haplotype frequencies were predicated with EH+ [325],

using genotypes from all 149 individuals of the allelic expression sample. Genotypes were entered into EH+ in a numerical format (i.e. common allele = 1, minor allele=2).

	Frequencies
Haplotype	Cases+Controls
AG	0.124
AA	0.413
TG	0.006
TA	0.457

Genotypes		Individual haplotype Probabilities (rs2619538, rs1047631)				Diplotype Probabilities			
rs1047631	rs2619538	TC	TT	AC	AT	Dip 1	Probability	Dip 2	Probability
AG	AA	0.5	0.5	0	0	AA,AG	1		
AG	AA	0.5	0.5	0	0	AA,AG	1		
AG	N/A	N/A	N/A	N/A	N/A	N/A			
AG	AA	0.5	0.5	0	0	AA,AG	1		
AG	AA	0.5	0.5	0	0	AA,AG	1		
AG	AA	0.5	0.5	0	0	AA,AG	1		
AG	TA	0.4796	0.0204	0.0204	0.4796	TA,AG	0.9592	AA,TG	0.0408
AG	AA	0.5	0.5	0	0	AA,AG	1		
AG	AA	0.5	0.5	0	0	AA,AG	1		
AG	AA	0.5	0.5	0	0	AA,AG	1		
AG	AA	0.5	0.5	0	0	AA,AG	1		
AG	TA	0.4796	0.0204	0.0204	0.4796	TA,AG	0.9592	AA,TG	0.0408
AG	TA	0.4796	0.0204	0.0204	0.4796	TA,AG	0.9592	AA,TG	0.0408
AG	TA	0.4796	0.0204	0.0204	0.4796	TA,AG	0.9592	AA,TG	0.0408
AG	AA	0.5	0.5	0	0	AA,AG	1		
AG	AA	0.5	0.5	0	0	AA,AG	1		
AG	TA	0.4796	0.0204	0.0204	0.4796	TA,AG	0.9592	AA,TG	0.0408
AG	TA	0.4796	0.0204	0.0204	0.4796	TA,AG	0.9592	AA,TG	0.0408
AG	AA	0.5	0.5	0	0	AA,AG	1		
AG	TA	0.4796	0.0204	0.0204	0.4796	TA,AG	0.9592	AA,TG	0.0408
AG	AA	0.5	0.5	0	0	AA,AG	1		
AG	TA	0.4796	0.0204	0.0204	0.4796	TA,AG	0.9592	AA,TG	0.0408
AG	AA	0.5	0.5	0	0	AA,AG	1		
AG	TA	0.4796	0.0204	0.0204	0.4796	TA,AG	0.9592	AA,TG	0.0408
AG	TA	0.4796	0.0204	0.0204	0.4796	TA,AG	0.9592	AA,TG	0.0408
AG	AA	0.5	0.5	0	0	AA,AG	1		
AG	TA	0.4796	0.0204	0.0204	0.4796	TA,AG	0.9592	AA,TG	0.0408
AG	TA	0.4796	0.0204	0.0204	0.4796	TA,AG	0.9592	AA,TG	0.0408
AG	TA	0.4796	0.0204	0.0204	0.4796	TA,AG	0.9592	AA,TG	0.0408
AG	TA	0.4796	0.0204	0.0204	0.4796	TA,AG	0.9592	AA,TG	0.0408

Table 4.1. Data output from the diplotype probability program for rs2619538. The original output of numerical genotypes has been amended to give actual nucleotides. Haplotype frequencies are provided from EH+. Individuals heterozygous for rs2619538 and rs1047631 are 96% likely to carry the TA/AG diplotype and only 4% likely to carry the AA/TG diplotype.

4.6.2.3 Linear Regression Analysis

By grouping all heterozygotes together the independent t-test method assumes that double heterozygote individuals (those heterozygous for the test SNP and the allelic expression SNP rs1047631) all carry the same diplotype. The most probable diplotype is then calculated to determine the direction of effect. However some test polymorphisms may not be in high linkage disequilibrium with the allelic expression SNP and therefore the two possible diplotypes may have similar probabilities. For example, an individual heterozygous at marker A and B (Aa, Bb) may have diplotype probabilities of AB, ab = 45% and Ab, aB = 55%. In these instances grouping heterozygotes together may not be suitable as some individuals may carry one diplotype (i.e. AB, ab) and other individuals may carry the alternative diplotype (i.e. Ab, aB).

In order to account for the possibility of nominal linkage disequilibrium between a tag SNP and rs1047631, and therefore the potential switching of phase in double heterozygotes, stepwise linear regression analysis was performed on the probability estimates for all haplotype combinations between the specified test SNP and rs1047631.

4.3 Results

4.3.1 Single Marker Correlation Analysis

To assess the correlation between DTNBP1 SNPs and cortical dysbindin mRNA levels, tag markers were genotyped in the 149 individuals that constitute the Caucasian brain sample. In total 60 of the 64 tag SNPs were successfully genotyped which captures 93% of the genetic variation of DTNBP1 locus and putative regulatory regions at r^2>0.95 MAF>0.001. All SNPs had a called rate >90% in the whole Caucasian brain sample (n=149) and a call rate >80% in the individuals with allelic expression data (n=31). Single marker analysis of association for these 61 SNPs and allelic expression ratios was subsequently performed using independent samples t-test and logistic regression. The results of this analysis are given in Table 4.2.

In total 7 SNPs were significantly associated with differential allelic expression. All markers, except DTNBP1_R16snp1 were non-significant for the Levene's test for homogeneity of variance, therefore equal variance was assumed for 59/60 t-tests. Five SNPs showed significant evidence for association by means of the independent t-test analysis (rs9370822 p=0.03, rs9358063 p=0.003, rs2619539 p=0.001 and rs2619538 p=0.001, rs9296989 p=0.003). Three of these were also significantly associated with allelic expression differences when variation in phase was taken into account using linear regression (rs9370822 p=0.03 and rs2619538 p=0.001, rs9296989 p=0.003). Two markers were significantly associated using the linear regression only (rs2235258 p=0.045 and rs13198512 p=0.007). The allelic expression ratios stratified by the genotypes of SNPs associated with allelic expression differences can be seen in Figure 4.3.

No	SNP	T-Test P	Levene's Test	LR P
1	rs2235258	0.167	0.766	0.045
2	rs9396589	0.334	0.466	N/S
3	rs9654600	0.890	0.777	N/S
4	rs9396590	0.086	0.979	N/S
5	rs1076636	0.582	0.955	N/S
6	rs9396591	0.114	0.992	N/S
7	rs742102	Non Poly	N/A	N/A
8	rs1474587	0.149	0.639	N/S
9	rs909626	0.628	0.577	N/S
10	rs3778651	0.832	0.787	N/S
11	rs13213814	0.128	0.493	N/S
12	rs13198512	0.198	0.942	0.007
13	rs13201824	0.153	0.219	N/S
14	rs1047631	Non Poly	N/A	N/A
15	rs17470454	0.103	0.060	N/S
16	rs742106	0.905	0.337	N/S
17	rs2056943	Non Poly	N/A	N/A
18	rs16876571	0.355	0.832	N/S
19	rs16876575	0.769	0.268	N/S
20	rs6937379	0.107	0.451	N/S
21	rs4712253	0.888	0.196	N/S
22	rs9464795	0.655	0.795	N/S
23	rs9370822	0.030	0.482	0.030
24	rs12527121	0.157	0.853	N/S
25	rs9296983	0.372	0.509	N/S
26	rs4715984	0.769	0.268	N/S
27	rs9358063	0.003	0.171	N/S
28	rs7771339	Non Poly	N/A	N/A
29	rs2743548	0.611	0.780	N/S
30	rs12203173	0.769	0.268	N/S
31	rs12199640	Non Poly	N/A	N/A
32	rs3829893	0.961	0.278	N/S
33	rs2619539	0.001	0.580	N/S
34	rs3213207	0.076	0.315	N/S
35	Alt_trans_snp_p5	0.185	0.862	N/S
36	rs1011313	0.324	N/A (Only 1 het)	N/S
37	rs2619528	0.251	0.884	N/S
38	rs2619520	0.305	N/A (Only 1 het)	N/S
39	rs1018381	0.655	0.795	N/S

Table 4.2. Single marker allelic expression correlation results. Shown for each SNP is chromosomal position (March 2006), independent t-test p value (T-test P), Levene's test from homogeneity of variance (Levene's Test), stepwise linear regression p value (LR P). Significantly associated SNPs are highlighted in red. Non-poly - the SNP was non-polymorphic through the 31 informative individuals. N/S – non significant. N/A = Not analysed.

40	rs1474605	0.251	0.884	N/S
41	rs12196958	Non Poly	N/A	N/A
42	rs1997679	0.268	0.613	N/S
43	DTNBP1_C9b_SNP1	0.176	0.821	N/S
44	rs9476886	0.113	0.650	N/S
45	rs11558324	0.377	0.774	N/S
46	rs2619536	0.325	0.667	N/S
47	DTNBP1_C16_SNP2	0.718	0.921	N/S
48	rs2619537	0.895	0.475	N/S
49	rs12204704	0.197	0.390	N/S
50	rs2619538	0.001	0.839	0.001
51	DTNBP1_Dhyp_SNP1	Non poly	N/A	N/A
52	rs17407828	0.415	0.395	N/S
53	DTNBP1_R10_SNP1	0.829	0.879	N/S
54	rs9296989	0.003	0.926	0.003
55	rs7450621	0.325	0.312	N/S
56	DTNBP1_R16_SNP1	0.207	0.012	N/S
57	rs2025926	0.839	0.787	N/S
58	rs2619514	0.086	0.154	N/S
59	rs9296985	0.655	0.778	N/S
60	rs7752070	0.611	0.750	N/S
40	rs1474605	0.251	0.884	N/S

Table 4.2 cont. Single marker allelic expression correlation results. Shown for each SNP is chromosomal position (March 2006), independent t-test p value (T-test P), Levene's test from homogeneity of variance (Levene's Test), stepwise linear regression p value (LR P). Significantly associated SNPs are highlighted in red. Non-poly - the SNP was non-polymorphic through the 31 informative individuals. N/S – non significant. N/A = Not analysed

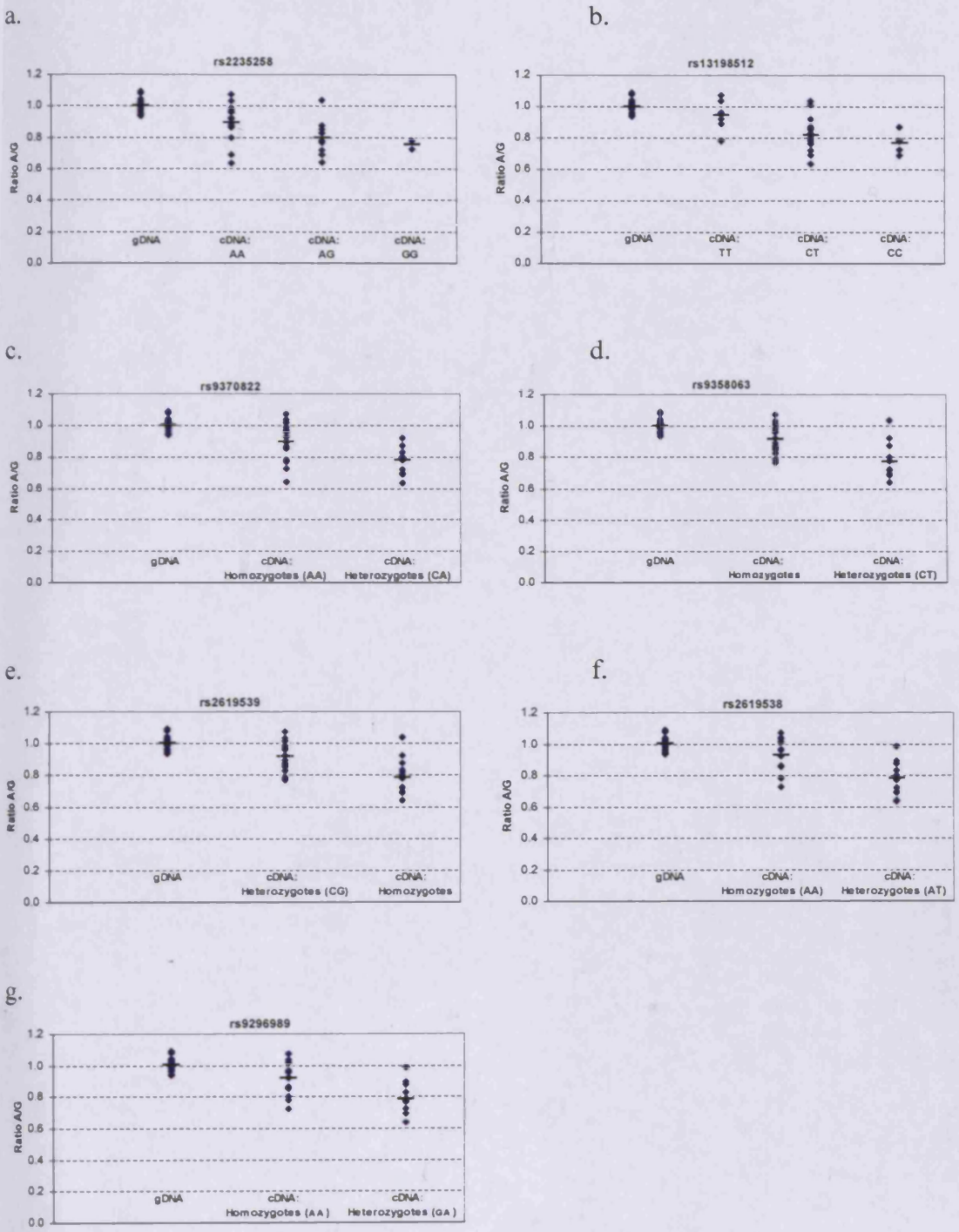


Figure 4.3. Allelic expression ratios (A/G) at SNP rs1047631 stratified by the genotypes of a) rs2235258 (Logistic Regression (LR) p=0.045) b) rs13198512 (LR p= 0.007) c) rs9370822 (independent samples t-test and LR p = 0.03) d) rs9358063 (t-test p =0.003) e) rs2619539 (t-test p = 0.001) f) rs2619538 (t-test and LR p= 0.001) g) rs9296989 (t-test and LR p = 0.003). Ratios are stratified by all genotypes unless statistical analysis of a SNP was only significant via t-test. In these cases ratios are stratified by heterozygosity.

4.3.2 LD between Significantly Associated SNPs

The LD between the SNPs significantly correlated with differential allelic expression is given in Table 4.3. The greatest LD between correlated SNPs is shown between rs2619538 and rs9296989 (r^2=0.92, D'=1). There is also relatively high LD between rs2619539, rs9370822 and rs9350863 ($0.51 \leq r^2 \leq 0.78$, $0.95 \leq D' \leq 1$).

	rs2235258	rs13198512	rs9370822	rs9358063	rs2619539	rs2619538	rs9296989
rs2235258		0.22	0	0.01	0.02	0	0
rs13198512	0.84		0.37	0.09	0.15	0.03	0.04
rs9370822	0.03	0.8		0.51	0.62	0.17	0.21
rs9358063	0.17	0.32	1		0.78	0.12	0.15
rs2619539	0.25	0.42	0.95	1		0.06	0.07
rs2619538	0.03	0.22	0.43	0.46	0.29		0.92
rs9296989	0.05	0.26	0.52	0.5	0.3	1	

Table 4.3. LD between SNPs significantly associated with differential allelic expression. r^2 values are given above the grey boxes, D' values are given below.

4.3.3 Location of Significant SNPs and Proxies

As this analysis uses a tag SNP approach it is important to note the proxies of the significantly associated SNPs. These are given in Table 4.4. Five of the 7 SNPs do not have any proxies (r^2>0.95), however rs9370822 and rs2619539 have five and four proxies respectively.

rsID	Chromosomal Position	Proxy	Chromosomal Position	r^2
rs9370822	15652715	rs909706	15768850	1
		rs7383568	15725911	1
		rs9396592	15646989	1
		rs6909929	15712434	0.963
		rs12207867	15699482	0.961
rs2619539	15728834	rs4715988	15726088	0.967
		rs2743868	15733787	0.967
		rs9476864	15719806	1
		rs4236167	15641930	0.967
rs2619538	15773188	No Proxy r^2>0.95		
rs9296989	15794678	No Proxy r^2>0.95		
rs2235258	15621461	No Proxy r^2>0.95		
rs13198512	15627864	No Proxy r^2>0.95		
rs9358063	15673010	No Proxy r^2>0.95		

Table 4.4 Proxies of SNPs significantly associated with differential expression.

The location of the SNPs significantly associated with allelic expression differences and their proxies is given in Figure 4.4. The locations of the putative regulatory regions identified in chapter 3 are also given.

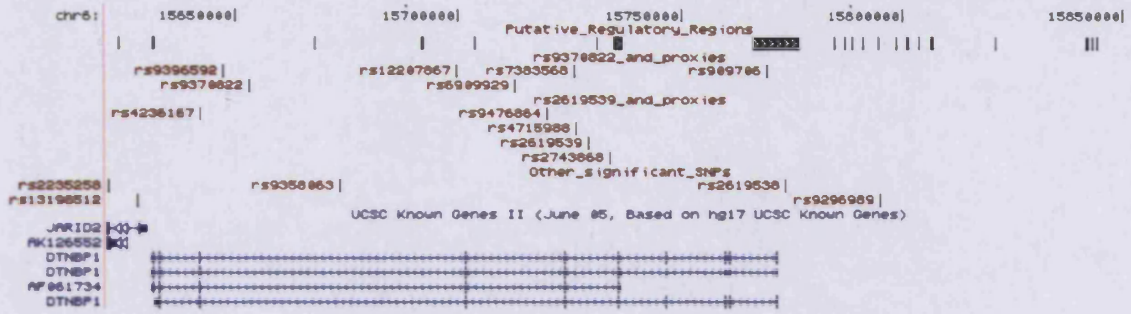


Figure 4.4 Location of significantly associated SNPs and proxies

Of the seven SNPs significantly correlated with allelic expression data and their nine proxies only three are located within putative regulatory regions identified in chapter 3. rs2926989 is within a region predicted to contain a cluster of TFBSs. However as described in chapter 3, rs9296989 is not predicted to disrupt any of the TFBSs detected by Cluster Buster within this region. rs2619538 is within the putative promoter region of three main DTNBP1 transcripts (DTNBP1a, DTNBP1b and DTNBP1c). This is also the case for rs909706 which is in an $r^2=1$ with rs9370822 ($p=0.030$).

4.3.4. Further Analysis of rs2619538 and rs2619539

If a SNP is a cis-acting variant, which causes the altered DTNBP1 expression observed in the rs1047631 allelic expression assay, it would be expected that this SNP would show the greatest association to allelic expression differences. It would also follow that individuals who are heterozygous for this SNP would display greater relative differences in the expression of each gene copy than those who are homozygous. The two SNPs most significantly associated with allelic expression differences were rs2619538 and rs2619539 (independent samples t-test p=0.001). However when allelic expression ratios are plotted against the genotypes of rs2619539 (Figure 4.3e), it is individuals homozygous for rs2619539 which show the greatest allelic distortion (homozygote mean cDNA ratio = 0.782, heterozygote mean cDNA ratio = 0.93). It is therefore unlikely that rs2619539 itself is the SNP causing the variation in DTNBP1 expression captured by this assay.

In contrast, individuals heterozygous for rs2619538 show the greatest differential expression (Figure 4.3f). Individuals heterozygous for rs2619538 are predicted to carry the diplotype TA/AG with a diplotype probability of 0.96 (see Table 4.1). Consequently, the mean reduction in expression of mRNA carrying the TA haplotype compared to mRNA carrying the AG haplotype is 22% (mean cDNA ratio = 0.78, range: 0.99-0.64). Therefore, in addition to being one of the two SNPs showing the greatest association to differential expression, unlike rs2619539, the pattern of the association of rs2619538 is consistent with the SNP having a cis-acting influence on expression.

4.3.5 Multi-locus Analysis

Although rs2619538 appears the most likely to be a cis-acting variant based on the correlation data, it is of note that the allelic variation at rs2619538 does not account for all the cis-acting variation observed. As with the schizophrenia risk haplotype initially shown to be correlated with reduced expression, some individuals homozygous for rs2619538 show differences in allelic expression (Figure 4.3f). This observation suggests a number of possibilities. There could be a single functional variant, as yet unidentified, that rs2619538 is in partial LD with. However for this to be the case the polymorphism would have to have been missed by our screening described in chapter 3 or be distal to the DTNBP1 locus (over 73kb 5' or 10kb 3' to the DTNBP1 gene). However as five of the seven SNPs associated with differential expression are outside the putative regulatory sequence identified in chapter 3, and high density tagging the DTNBP1 locus only included ±10kb of the gene, this is a possibility.

A second possibility is that the combination of rs2619538 and another DNA variant is causing the allelic expression differences observed. Since Bray and colleagues note that the spread of the allelic expression data could potentially reflect multiple cis-acting variants [186] the allelic expression data was reanalysed to determine whether rs2619538 could be affecting expression of mRNA in combination with other polymorphism(s). If rs219538 is working in conjunction with another cis-acting variant captured in this study, then it would be likely that this other variant would show some degree of correlation with allelic expression differences. Therefore haplotypes between rs2619538 and each of the other six significant SNPs were subjected to association analysis with allelic expression differences.

All haplotypes between rs2619538 and the other significantly correlated SNPs (rs2235258, rs13198512, rs9370822, rs9358063 and rs2619539, rs9296989) were systematically analysed. Phase information for the allelic expression SNP rs1047631 was also included in the analysis so a direction of effect could be determined. As this analysis created multiple diplotype combinations, a two sample independent t-test

analysis could not be performed. Therefore the resulting haplotype probabilities were subjected to stepwise linear regression analysis. The results of this analysis, along with the single marker association results, are summarised in Table 4.3.

The most significant linear regression result is shown for the combination of rs2619538 and rs13198512 (p=0.001). The genotypes of these two markers accounts for 74% of the variation in allelic expression. This is 21% greater than rs2619538 on its own which accounts for only 53% of the variation. The 2-marker combination of rs2619538 and rs2619539 is also calculated to account for 74% of the variation although the association shown for these two SNPs is less significant (p=0.003) than the haplotypes involving rs13198512. Further details on these two 2-marker combinations with phase information for rs1047631 are given below.

	Single Marker T-Test P Value	Single Marker Regression P Value	Single marker R value	Multilocus Linear Regression P value	rs2619538 haplotype R value
rs2235258	0.167	0.045	0.412	0.001	0.648
rs13198512	0.198	0.007	0.532	0.0001	0.741
rs9370822	0.030	0.030	0.428	0.0041	0.543
rs9358063	0.003	N/S	N/A	0.001	0.699
rs2619539	0.001	N/S	N/A	0.0003	0.743
rs2619538	0.001	0.001	0.560	N/A	N/A
rs9296989	0.003	0.003	0.537	0.002	0.58

Table 4.3. Results of both the single marker and mulilocus analysis for rs2619538 with each of the other SNPs associated with allelic expression differences. R values are also given which show the proportion of the data (1=100%) explained by genotypes of each SNP/haplotype. N/S = Non-significant ($p>0.05$). N/A = Not analysed.

4.3.5.1 rs2619538 and rs13198512

The allelic expression data stratified by the most probable diplotypes for rs2619538, rs1047631 and rs13198512 are given in Figure 4.5. As can be seen, there is no one haplotype that accounts for all the allelic variation; therefore, if both rs2619538 and rs13198512 are functional variants, the stratified ratios suggest that the two polymorphisms have independent effects on mRNA expression. This is further supported by the location of the two polymorphisms. rs13198512 is located ~3kb 3' to the four main DTNBP1 transcripts whereas rs2619538 is ~1.8kb 5' to the gene (Figure 4.6). As discussed in section 4.3.3 neither rs2619538 nor rs13198512 have any proxies at an r^2>0.95.

As suggested from the single marker correlation analysis, rs2619538 appears to have the greatest influence on allelic expression differences as individuals heterozygous for rs2619538 (diplotypes TAA/AGA, TAG/AGG, TAG/AGA) show the largest deviation from the 1:1 ratio (Combined average ratio =0.75, independent t-test against gDNA p=0.0000015). rs13198512 also appears to influence allelic expression differences as individuals homozygous for rs2619538, but heterozygous for rs13198512 (AAG/AGA, n=6), show a difference in allelic expression (average A/G ratio =0.91). Although the individuals carrying this diplotype are not significantly different from the corrected genomic DNA ratios they do show a trend (independent t-test p=0.08). Furthermore, individuals heterozygous for both rs2619538 and rs13198512 (TAG/AGA, n=8) show the greatest allelic expression differences (average A/G ratio =0.74). Also of note is that, unlike individuals homozygous for rs2619538 alone, individuals homozygous for rs2619538 and rs13198512 (AAA, AGA, n=5) show no expression differences except one individual carrying the AAG/AGG haplotype which shows an AE ratio of ~0.72. Therefore, the addition of rs13198512 accounts for more of the allelic variation than rs2619538 on its own. This is reflected in the linear regression analysis (rs2619538 linear regression model R= 0.560, rs2619538 and rs13198512 model R=0.741).

If these two SNPs are two independent cis-acting variants then it is the T allele of rs2619538 and the G allele of rs13198512 which cause a relative reduction in mRNA expression compared to the mRNA carrying their alternate alleles.

Figure 4.5. Allelic ratios at SNP rs1047631 stratified by the predicted 3-marker diplotype for SNPs rs2619538, rs1047631and rs13198512. The combination of markers rs13198512 and rs2619538 shows the greatest correlation with allelic expression ratios (LR P= 0.0001, R=0.741). The probabilities of all diplotypes shown are >73%.

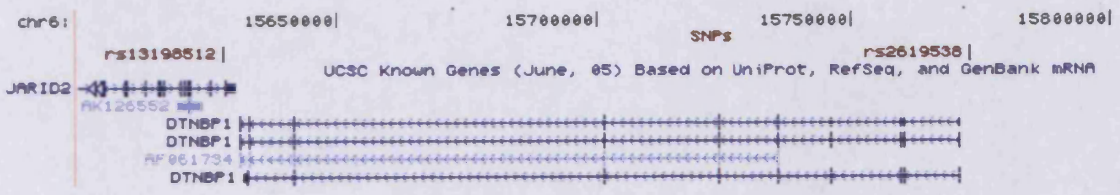


Figure 4.6. Chromosomal position of rs2619538 and rs13198512 (UCSC May 2004).

4.3.5.2 rs2619538 and rs2619539

Linear regression analysis of rs2619538 and rs2619539 calculated that this two marker haplotype combination accounted for the allelic expression data as well as the two marker haplotype containing rs2619538 and rs13198512 (R=0.74 respectively). However single marker analysis of rs2619539 suggested that variation at this locus is unlikely to have a cis-acting influence of DTNBP1 expression (section 4.3.4). Stratifying the allelic expression data by the most probable diplotypes for rs2619538, rs2619539 and rs1047631 (shown in Figure 4.7) further advocates this hypothesis.

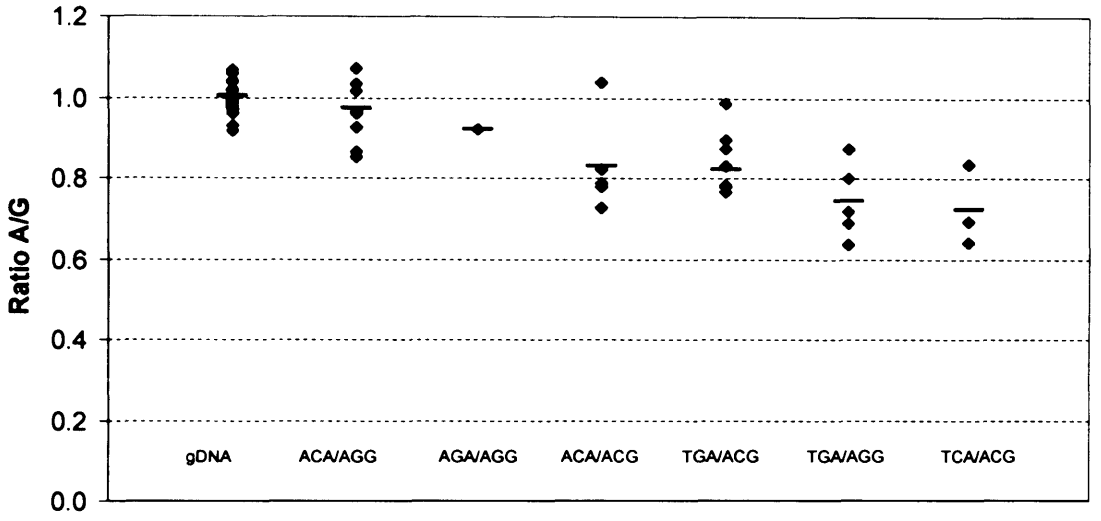


Figure 4.7 Allelic ratios at SNP rs1047631 stratified by the predicted 3-marker diplotype for SNPs rs2619538, rs2619539 and rs1047631. Average diplotype probability = 83%.

In the three marker combination described in section 4.3.5.1, individuals homozygous for rs2619538, but heterozygous for rs13198512, were found to show allelic expression differences. This was not the case for rs2619539. Individuals homozygous for rs2619538 but heterozygous for rs2619539 (ACA/AGG, n=7) show the least differential expression (average A/G ratio = 0.97) whereas individuals homozygous for both SNPs (ACA/ACG, n=4) do show allelic expression differences (average A/G ratio = 0.83). Furthermore, unlike rs13198512 and rs2619538, the double heterozygotes of rs2619538 and rs2619539 (TGA/ACG, n=7) do not show the greatest allelic expression differences. Therefore, detailed analysis of the allelic expression data stratified by

rs2619538 and rs2619539 indicates that the genotypes of rs2619538 and rs2619539 do not explain the spread of the data any better than rs2619538 alone.

4.3.6 UKCC Association Data and Correlation Analysis

In order to determine whether possible DTNBP1 cis-acting variants are relevant to schizophrenia aetiology, the UKCC association data, described in chapter 3, was reviewed for the SNPs showing association to differential expression and vice versa. All SNPs showing a significant association to either schizophrenia or allelic expression differences are given in Table 4.4.

SNP	Chromosomal Position	T-Test P	LR P	Arm Trend P	Genotypic P
rs2235258	15621461	0.167	0.045	0.890	0.250
rs3778651	15626511	0.832	N/S	0.131	0.011*
rs13198512	15627864	0.198	0.007	0.293	0.385
rs9370822	15652715	0.030	0.030	0.532	0.099
rs4715984	15669870	0.769	N/S	0.001	0.003*
rs9358063	15673010	0.003	N/S	0.887	0.124
rs2619539	15728834	0.001	N/S	0.392	0.084
rs2619538	15773188	0.001	0.001	0.024	0.016
rs9296989	15794678	0.003	0.003	0.017	0.038

Table 4.4. SNPs with a significant association to either allelic expression differences or schizophrenia are shown. The t-test and linear regression (LR) results for the association with differential expression are given. The Armitage trend (Arm Trend) and genotypic p values are given for the schizophrenia association analysis. *designates where a genotypic p value has been determined using CLUMP with 10000 permutations.

Four SNPs showed a significant association with schizophrenia (rs3778651, rs4715984, rs2619538 and rs9296989). Of these two do not show an association with allelic expression differences (rs3778651 p=0.832, rs4715984 p=0.769). Two schizophrenia risk variants, rs2619538 and rs9296989 also show an association to differential expression (p=0.001 and p=0.003 respectively). However the association to schizophrenia observed by these two SNPs was not found to be independent of rs4715984, which was found to show the greatest association to schizophrenia and was the only SNP to survive correction for multiple testing.

Of the seven SNPs significantly associated with allelic expression difference only rs2619538 and rs9296989, mentioned above, are also associated with schizophrenia.

rs13198512, which along with rs2619538, appears to explain the majority of the altered mRNA expression, is not associated with schizophrenia (p=0.311). Neither is the 3-marker haplotype between rs2619538, rs1047631 and rs13198512 (TAG p=0.229, Table 4.5) which shows the greatest relative reduction in DTNBP1 expression.

Haplotpye	Case Frequency	Control Fequency	P Value
Global	N/A	N/A	0.154
TAA	0.188	0.166	0.126
TAG	0.275	0.255	0.229
AGA	0.096	0.091	0.634
AGG	0.036	0.043	0.342
AAA	0.221	0.229	0.642
AAG	0.183	0.216	0.029

Table 4.5. Association analysis of 3-marker haplotype rs2619538, rs1047631, rs2619538 within the schizophrenia case control sample. N/A = Not applicable.

4.3.7. Analysis of the Refined Schizophrenia Risk Haplotype

Bray and colleagues have shown that a schizophrenia risk haplotype previously reported by Williams *et al* [134] (TAC: rs2619538, rs3213207, rs2619539, p=0.04163) is also significantly associated with reduced DTNBP1 expression [186]. The high density mapping of the DTNBP1 locus, described in chapter 3, illustrated that rs4715984 refines the association of this risk haplotype with schizophrenia in the extended UK case control sample (TACT: rs2619538, rs3213207, rs2619539, rs4715984, p=0.005).

In the initial investigation between the schizophrenia risk haplotype TAC and DTNBP1 expression, a *post-hoc* reanalysis was performed on the association data to identify the risk haplotype that included phase information with respect to rs1047631 and that maximally differentiated cases and controls. This was achieved by a 3-marker haplotype comprising of the T-allele of rs2619538, the A-allele of rs3213207 and the A-allele of rs1047631 (TAA frequency in cases = 45.6%, frequency in controls = 40.4%). It was subsequently shown that the relative expression of the A-allele of rs1047631 was significantly lower when it was carried on the risk haplotype compared with when it was carried on the non-risk haplotypes (Figure 4.8).

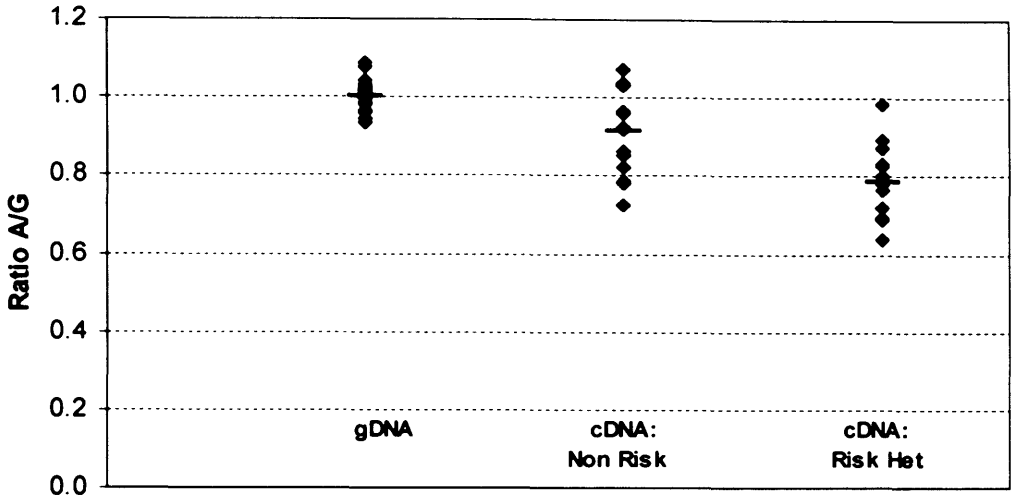


Figure 4.8. Allele ratios at SNP rs1047631, stratified by heterozygosity for the defined 3-marker schizophrenia TAA risk haplotype. Data represented as a ratio of A/G alleles. The risk haplotype includes the reduced expression allele of rs1047631. cDNA samples from individuals who are heterozygous for the risk haplotype show lower expression of the A-allele than those from individuals who carry no copies of the risk haplotype (p=0.002).

To determine whether the refined association signal is still associated with reduced expression, the refined risk haplotype which contains phase information for the allelic expression SNP (TATA: T-allele of rs2619538, the A-allele of rs3213207 and the T-allele of rs4715984, A-allele of rs1047631) was analysed.

To assess the effect of the refined TATA haplotype on allelic expression of DTNBP1, diplotype probabilities were determined for all 31 individuals informative for the allelic expression assay. 15 of the 31 individuals were originally predicted to be heterozygous for the 3-marker risk haplotype TAA (average diplotype probability =0.97, range 0.73-1). Following genotyping of rs4715984 through the allelic expression sample, 12 of these 15 individuals were predicted to be heterozygous for the non-associated TACA haplotype (T-allele of rs2619538, the A-allele of rs3213207 and the C-allele of rs4715984, A-allele of rs1047631 p=0.405) with an average diplotype probability of 0.88 (range = 0.6-0.99), while two were predicted to be heterozygous for the associated (p=0.0037) TATA haplotype (average diplotype probability = 0.63). One of the 31 informative individuals failed genotyping for rs4715984. Subsequently the allelic expression ratios of individuals heterozygous for the TAA haplotype were stratified into individuals carrying the TATA haplotype and the TACA haplotype (Figure 4.9). If

the association signal is linked to reduced expression, we would expect individuals carrying the TATA risk haplotype to still show a reduction in allelic expression. However as Figure 4.9 shows, although a significant association with allelic expression differences is still observed (p=0.0003), the relative expression of the A-allele of rs1047631 is significantly lower when it is carried on the TACA non-risk haplotype.

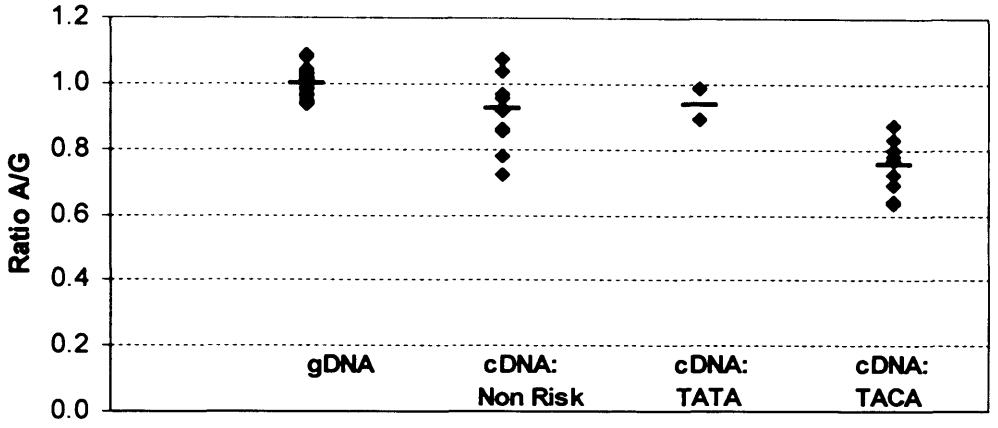


Figure 4.9. Individuals carrying the original TAA risk haplotype are further separated into those heterozygous for the associated TATA (p=0.0037) haplotype and those heterozygous for the non-associated TACA haplotype (p=0.405). The 4-marker haplotype includes the reduced expression allele of rs1047631. cDNA samples from individuals who are heterozygous for the non-associated haplotype TACA show lower expression (p=0.0003) than both the individuals carrying no copies of the original risk haplotype and those heterozygous for the new refined risk haplotype TATA.

Although this suggests that our refined schizophrenia association signal is not associated with reduced expression, the relatively low diplotype probability estimates mean the data are not conclusive. For example the most likely diplotype for two individuals who show allelic expression differences (allelic expression ratios 0.781 and 0.639) is TACA/AATG (probability = 0.6). However these same individuals also have a 39% chance of carrying the TATA/AACG diplotype and therefore the associated haplotype (See Table 4.6).

Nevertheless, in order for the TATA risk haplotype to be associated with reduced expression one or both of these two individuals would have to be TATA/AACG and two individuals most likely to carry TATA/AGCG (63%) that do not show allelic expression differences would have to actually carry the TACA/AGTG (probability

36%). The probability of all four individuals carrying the less likely diplotype is 2% and the probability of one of the individuals showing allelic expression differences carrying TATA and the other three carrying TACA is 5%. Therefore although the data is not conclusive, it is highly unlikely that the new refined risk haplotype is associated with reduced expression.

		Genotypes				Diplotype Probabilities			
ID	AE Ratio A/G	rs2619538	rs3213207	rs4715984	rs1047631	Diplotype 1	Probability	Diplotype 2	Probability
A283/96	1.073	AA	GA	CT	GA	AACA AGTG	0.82		
A74/90	1.070	AA	GA	N/A	GA				
G44	1.050	N/A	GA	N/A	GA				
A248/97	1.037	AA	AA	CC	GA	AACA AACG	1		
G30	1.034	AA	GA	CC	GA	AACA AGCG	0.99		
G39	1.014	AA	N/A	CC	GA				
S-9	0.985	TA	GA	CT	GA	TATA AGCG	0.63	TACA AGTG	0.36
A009/99	0.965	AA	GA	CT	GA	AACA AGTG	0.82		
G35	0.958	AA	GA	CC	GA	AACA AGCG	0.99		
G53	0.924	AA	GA	CC	GA	AACA AGCG	0.99		
G26	0.920	AA	GA	CC	GA	AACA AGCG	0.99		
A272/93	0.894	TA	GA	CT	GA	TATA AGCG	0.63	TACA AGTG	0.36
G54	0.874	TA	AA	CC	GA	TACA AACG	0.82		
S-24	0.873	TA	GA	CC	GA	TACA AGCG	0.99		
G42	0.863	AA	GA	CC	GA	AACA AGCG	0.99		
7dx	0.852	AA	GG	CC	GA	AACA AGCG	0.99		
A285/95	0.833	TA	AA	N/A	GA				
G37	0.829	TA	AA	CC	GA	TACA AACG	0.82		
A136/90	0.820	AA	AA	N/A	GA				
71dx	0.799	TA	GA	CC	GA	TACA AGCG	0.99		
A145/90	0.787	AA	AA	N/A	GA				
G21	0.781	TA	AA	CT	GA	TACA AATG	0.6	TATA AACG	0.3946
A17/90	0.780	AA	AA	CC	GA	AACA AGCG	1		
A94/93	0.779	TA	GA	CC	GA	TACA AGCG	0.99		
A190/94	0.766	TA	GA	CC	GA	TACA AGCG	0.99		
G55	0.725	AA	AA	CT	GA	AACA AATG	0.93		
A206/90	0.720	TA	GA	CC	GA	TACA AGCG	0.99		
A389/94	0.693	TA	AA	CC	GA	TACA AACG	0.82		
A215/90	0.689	TA	GA	CC	GA	TACA AGCG	0.99		
A98/89	0.639	TA	AA	CT	GA	TACA AATG	0.6	TATA AACG	0.3946
A99/89	0.636	TA	GA	CC	GA	TACA AGCG	0.99		

Table 4.6. Refined Risk Diplotype Probabilities
31 individuals heterozygous for rs1047631, and therefore informative for the allelic expression sample are shown with the individual genotypes for the four markers that constitute the refined schizophrenia risk haplotype TATA (T allele rs2619538, A allele rs3213207, T allele rs4715984, A allele rs1047631) as well as possible diplotypes and probabilities. Four individuals carry the genotypes which make the associated TATA (shown in yellow) possible. However the two individuals which show allelic expression differences are predicted to carry the non associated TACA (shown in purple) haplotype.

4.4. Discussion

The aim of this chapter was to identify putative DTNBP1 cis-acting variants by determining whether any SNPs which tag the DTNBP1 locus and putative regulatory regions were associated with reduced cortical dysbindin mRNA levels. Seven polymorphisms were found to be associated with the allelic expression differences first reported by Bray and colleagues [186]. Of these SNPs the two most significantly associated were rs2619538 and rs2619539 (p=0.001). Further analysis of the pattern of association shown by these two SNPs suggests that rs2619538 is the SNP most likely to have a cis-acting influence on DTNBP1 expression. The mean reduction of mRNA carrying the TA haplotype (T-allele rs2619538, A-allele rs1047631) compared to mRNA carrying the AG haplotype was 22%. However some individuals showing allelic expression differences were homozygous for rs2619538 therefore variation at this locus does not account for all of the expression differences observed. The spread of the data itself suggests that multiple cis-acting variants may be working in combination to cause the expression differences illustrated. In investigating this possibility the combination of two polymorphisms, rs2619538 and rs13198512, were found to show the most correlation with allelic expression differences (p=0.0001), accounting for 74% of the expression differences. rs2619538 is ~1.9kb upstream of the TSS of three of the main DTNBP1 transcript variants. rs13198512 is ~3kb downstream of DTNBP1, within an intron of JARID2.

If reduced DTNBP1 expression is a primary aetiological mechanism in the physiology of schizophrenia, it is likely that these cis-acting variants would also be associated with schizophrenia. However further investigation indicates that the reduction of DTNBP1 expression through cis-acting variation may not be a primary aetiological factor in schizophrenia. While rs2619538 is weakly associated with schizophrenia, previous analysis has determined that the genetic association signal shown by this SNP is not independent of the more associated SNP rs4715984. rs4715984 shows no association with allelic expression differences (p=0.769). Furthermore neither s13198512 nor the TAG haplotype (T allele rs2619538, A-allele of rs1047631, G allele of rs13198512)

which is associated with the greatest relative reduction in DTNBP1 expression, are associated with schizophrenia (p=0.2927 and p=0.229 respectively).

The hypothesis that DTNBP1 cis-acting variation may not cause susceptibility to schizophrenia is further supported by the reanalysis of a schizophrenia risk haplotype previously shown to be associated with significantly reduced DTNBP1 expression. The addition of rs4715984 to this risk haplotype refines the original association to schizophrenia. However when this refined haplotype is analysed in phase with rs1047631 the risk haplotype TATA is no longer correlated with reduced expression.

As with all studies of this nature there are a number of limitations to the analysis presented. The correlation analysis performed in this chapter consisted of a two step approach employing both independent t-tests and linear regression. By analysing all haplotype probabilities, linear regression analysis was included in order to take into account instances of nominal linkage disequilibrium between the test SNP and the allelic expression SNP which could result in the "flipping" of phase within double heterozygotes. While this two step analysis accounts for linkage disequilibrium differences to some extent, it is possible that the logistic regression might not be able to compensate for very low LD between a test SNP and the allelic expression SNP. In these instances the level of false negatives may be inflated.

Another caveat of this analysis is that the allelic expression assay described analyzes multiple DTNBP1 transcripts simultaneously and therefore findings relate to net effects of those transcripts. Consequently the effects of SNPs/haplotypes on specific transcripts may have gone undetected. This is potentially especially true of DTNBP1b as the allelic expression assay used in the chapter does not assay variation from this transcript.

Another caveat of the analysis reported in this chapter is that it has been performed on a relatively small number of individuals. As a result although a relatively high genotyping call rate was employed for the 31 informative individuals (>80%), even one missing genotype has the potential to affect the association results. It is also worth noting that the diplotype probabilities calculated are an estimate based on an estimate of

population haplotype frequencies. Therefore the diplotype probabilities (especially where relatively low) should be taken with caution. This is especially relevant in relation to the refined risk haplotype correlation analysis where the most probable diplotypes are as low as 60%.

This analysis also suggests that the identification of putative regulatory sequence by the methods described in chapter 3 may not be an effective way of detecting cis-acting variants. Five of the seven SNPs associated with allelic expression differences are not found within the putative regulatory sequence determined in chapter 3. This includes rs13198512, one of two SNPs most likely to cause cis-acting variation. However the most significantly correlated SNP rs2619538 was identified through the screening of the predicted promoter region for three of the main DTNBP1 transcripts DTNBP1a, DTNBP1b and DTNBP1c.

While correlation with differential expression suggests a SNP may be a cis-acting variant, it is possible that the SNP is in LD with the actual functional variant. In order to determine whether rs2619538 or rs13198512 have a direct effect on expression, both SNPs were subjected to functional analysis by luciferase reporter gene assay. The results of this assay are described in chapter 5.

Chapter 5: Functional Analysis of rs2619538 and rs13198512

5.1 Introduction

The previous chapter determined that the polymorphisms rs2619538 and rs13198512 are associated with differences in allelic expression at the DTNBP1 locus, implying that they may have a cis-acting regulatory influence on dysbindin expression. Alternatively, it is possible that one or both of rs2619538 and rs13198512 are not functional variants but are in strong LD with the actual regulatory polymorphisms.

A functional assay could help establish whether rs2619538 and/or rs13198512 do have a cis-acting regulatory influence on gene expression. In addition, prior to correlation analysis with the allelic expression data, rs2619538 was identified as a potential regulatory variant (due to its location within a putative regulatory region) while rs13198512 was not. Consequently, a functional assay may also help determine whether identifying putative regulatory variants, via the methods described in chapter 3, is a useful tool in identifying polymorphisms likely to have a regulatory influence.

Currently a common approach used to investigate the effects of a specific SNP on gene expression is the reporter gene assay. This *in vitro* technique utilises a plasmid which contains the coding sequence of a reporter gene (Figure 5.1). Putative regulatory sequence(s) are cloned into this reporter vector and the transcriptional activity of the inserted sequence can be indirectly estimated through quantification of the reporter gene protein.

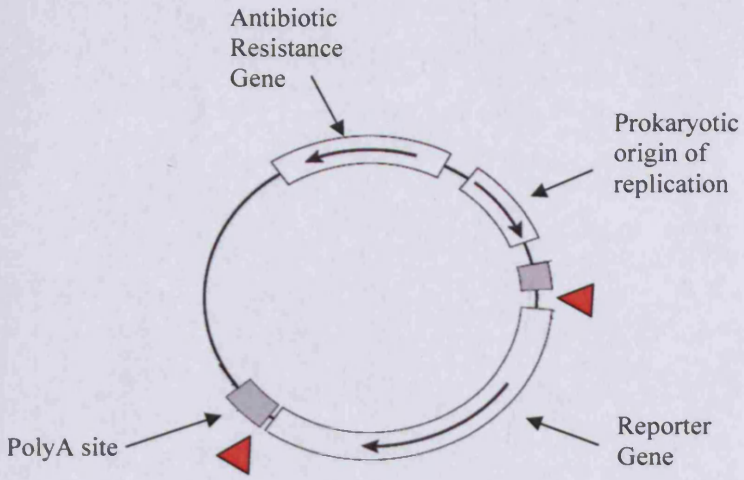


Figure 5.1. Principles of a reporter vector. The antibiotic resistance gene and prokaryotic origin of replication permit propagation and selection of the vector in *E. coli*. Putative regulatory sequences can be cloned either 5' or 3' to the reporter gene (depicted by ▲) Expression of the reporter gene is detected by enzymatic or immunological means. PolyA = polyadenylated. Adapted from Promega technical manual TM033..

Although the reporter gene assay is an *in vitro* experiment, it presents a number of advantages. Firstly, it is a relatively simple and rapid experiment. In addition, genetic and environmental factors, which may complicate the results of an *in vivo* assay, can be controlled. Finally, as well as determining whether a region of sequence has a regulatory effect on transcription, the reporter gene assay is sensitive enough to detect variation in expression caused by a single base change within the inserted sequence.

Over 100 SNPs, associated with a range of behavioural traits, physiology and disease susceptibility within several organisms, have been shown to affect gene expression using the reporter gene assay [326]. These include variants within the Duffy blood group [327], the lactose gene [328] and the 5-HTT gene. The 5-HTT gene modulates serotoninergic neurotransmission by the active reuptake of serotonin from the synaptic cleft [329] and is associated with a variety of different behaviours, psychological processes (such as social behaviour, aggression, anxiety) and psychiatric disorders [330]. A number of allelic variants that differentially modulate 5-HTT transcription have been identified within the promoter of the 5-HTT gene using luciferase reporter

gene assay. This altered expression has been shown to affect the rate of serotonin (5-HT) uptake [331].

The luciferase reporter gene assay is a frequently used reporter gene assay. The firefly luciferase gene (*Photinus pyralis*) produces a 61kDA monomeric protein whose enzymatic activity oxidises beetle luciferin in a luminescent reaction (Figure 5.2). The light emitted from the reaction is quantified using a luminometer [332] which allows for an extremely sensitive assay [333].

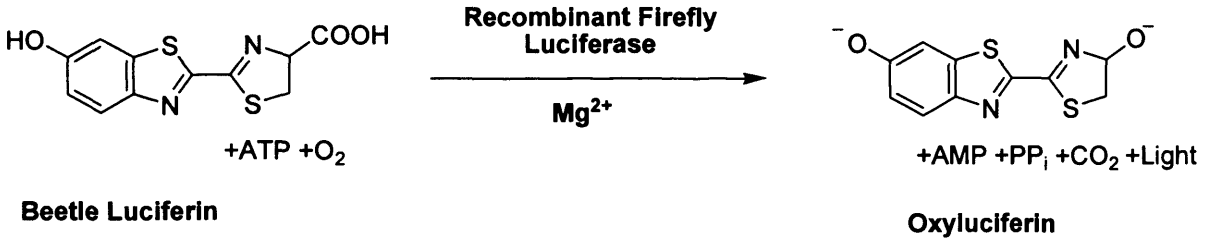
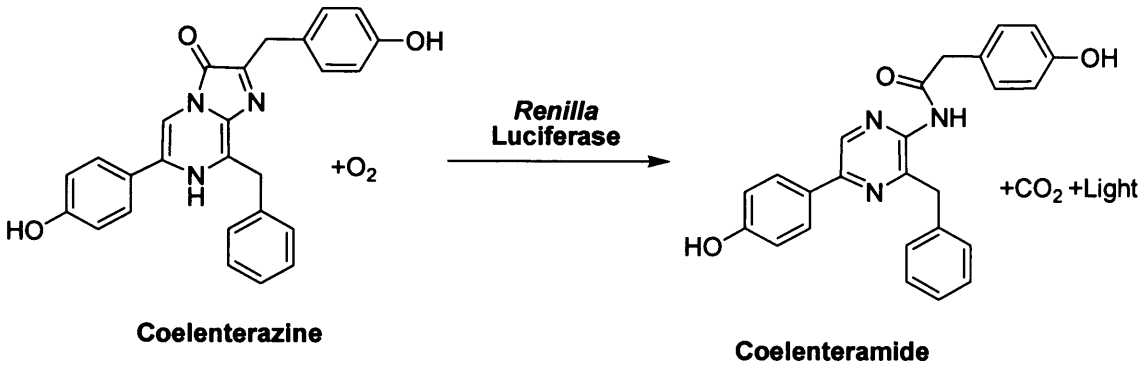


Figure 5.2. The oxidation of beetle luciferin by firefly luciferase.

One problem often encountered with reporter gene assays is that variation in transfection efficiency can cause a high degree of variability in results. To normalise transfection efficiency within and between experiments, co-transfection of a control reporter gene vector can be used. *Renilla* luciferase (*Renilla reniformis*) is employed as control gene for the firefly luciferase assay in Promega's dual-luciferase® reporter assay system. The *Renilla* luciferase gene produces a 36kDA protein which utilises O_2 and coelenterazine (Figure 5.3) [334]. Although both quantifiable by light emission, *Renilla* luciferase has a dissimilar enzyme structure and substrate requirement to firefly luciferase. These differences make it possible to selectively discriminate between the two bioluminescence assays.



Coelenterazine **Coelenteramide**

Figure 5.3. The oxidation of coelenterazine by *Rennila* luciferase.

The luciferase reporter vector system allows the analysis of both 5' and 3' flanking regions/SNPs [333] and therefore is an ideal assay to determine whether rs2619538 or rs13198512 have a cis-acting regulatory influence on transcription *in vitro*. As a result, the aim of this chapter was to determine whether rs2619538 or rs13198512 have a functional affect on transcription through the use of the dual-luciferase® reporter assay system (Promega). Any differences in transcription were compared with the results identified in the previous chapter in order to establish whether the two sets of analyses depict the same allele(s) associated with a reduction in dysbindin/luciferase expression.

5.2. Methods

5.2.1 Cell Strains and Media

5.2.1.1 Bacterial Strains

For the general expression of recombinant DNA plasmids, *Escherichia coli* XL1-Blue cells were used (Stratagene) [335].
Genotype: *recA1 endA1 gyrA96 thi-1 hsdR17 supE44 relA1 lac [F′ proAB lacI^{q}ZΔM15 Tn10 (Tet^r)].*

5.2.1.2 Mammalian Strains

Functional analysis was performed using HEK 293 (Human embryonic kidney) cells [336]. These were obtained from the European Collection of Cell Cultures (ECACC). Although of epithelial origin, the biochemical chemistry of HEK293 cells is capable of carrying out most of the post-translational folding and processing required to generate functional mature protein from a wide spectrum of both mammalian and non-mammalian DNA [337]. The HEK293 cell line also has a number of other advantages that make it a popular choice for a variety of expression studies. Firstly, HEK293 cell lines have low maintenance costs and rapid cell division (24-36h). In addition, HEK293 cells have a high transfection efficiency using a variety of methods which leads to a high protein production. The use of HEK293 cells also produces faithful translation.

5.2.1.3 Bacterial Media

All solutions were prepared and sterilised as appropriate. Media were autoclaved for 30 mins at 115°C and cooled to below 50°C before use. Antibiotics were dissolved in water, filtered sterilised (0.2μM) and stored at -20°C.

All bacterial *Escherichia coli* (*E. coli*) strains were grown in Luria-Bertani (LB) Media containing 10g/l tryptone, 5g/l yeast extract and 10g/l NaCl. Solid media plates were

prepared by adding 18.75g/l Bacto-agar prior to autoclaving. Antibiotic selection was provided by adding carbenicillin to either LB media or LB-agar solution at a final concentration of 50μg/ml.

5.2.1.4 Growth and Storage of Bacterial Cell Lines

E. coli strains were initially grown on LB solid media for 18-24h and stored at 4°C for up to one month. Before specific colonies were picked, agar plates were removed from storage two hours before use and allowed to dry. Selected *E. coli* strains were grown up in LB for 16-20 hours at 37°C with shaking at 200rpm. Longer term storage of *E. coli* was achieved by storing cells in 15% v/v glycerol at -80°C.

5.2.1.5 Growth of Mammalian Cell Lines

HEK 293 cells were grown in 25cm^2 tissue culture flasks containing MEM Glutamax media supplemented with 10% FBS, 1% glutamine and 1% NEM (non-essential amino acids). Cells were incubated at 37°C in 5% CO_2 and passaged at 70% confluence to maintain healthy cultures. Passaging was performed by first removing the cell culture media. Cells were then gently washed twice with phosphate buffered saline (PBS) and detached from the flask by incubation with 1ml of trypsin/EDTA for 1 minute at 37°C. This was followed by gently tapping the flask to dislodge the cells. To re-suspend the cells 15mls of pre-warmed media were added to the flask. From this stock solution 2mls were used to re-seed a new 25cm^2 culture flask containing 14mls of fresh pre-warmed media.

5.2.2 Molecular Biology

5.2.2.1 Primer Design for Cloning

PCR primers for the desired genomic region were designed to include the amplification sequence plus a specific restriction enzyme (RE) recognition sequence. The RE recognition sequence was incorporated into the primer in order to allow cloning of the putative regulatory region into a pGL3 luciferase vector. The restriction enzyme site was chosen based on a number of factors. Firstly the RE recognition site needed to be located only once within the required pGl3 luciferase vector and in an appropriate position for cloning, for example immediately 5', upstream or downstream to the luciferase gene. Secondly, the RE recognition site could not occur within the putative regulatory sequence to be inserted. As each cloning strategy required a combination of two restriction enzymes, it was also advantageous if the two restriction enzymes were functional with the same reagents and reaction conditions. Where there was the potential for subcloning (multiple sequences inserted into the same plasmid) all sequences were checked for recognition sites for each selected restriction enzyme.

5.2.2.2 PCR Conditions

In order to ensure sequence specificity high fidelity taq was used for all PCR reactions. A typical PCR reaction was as follows:

	Volume
High fidelity buffer	1.2µl
Forward primer (100ng/µl)	0.5µl
Reverse primer (100ng/µl)	0.5µl
dNTP mix (10mM)	1µl
DNA Template (4ng)	3µl (dried)
High fidelity Taq polymerase	0.2µl
Sterile molecular grade water	5.60µl

The PCR cycling parameters were as follows with annealing temperature dependant on the amplimer:

	X 40			
94°C	94°C	56°C-66°C	72°C	72°C
2 minutes	15 s	30 s	1min/kb	7 min

Where the fragment generated was >3kb an extension temperature of 68°C was used rather then 72°C.

5.2.2.3 Site-directed Mutagenesis

In order to introduce specific mutations into the inserted sequence site directed mutagenesis (SDM) was performed. SDM involved a standard PCR reaction except PCR primers were designed to incorporate the desired base change. Primers were designed by Primer 3 web resource (http//www-genome.wi.mit.edu/cgi-bin/primer/primer3-www.cgi). For increased specificity the polymerase Pfu-Ultra (Stratagene) was used. A typical mutagenesis reaction was as follows:

	Volume (µl)
10 x polymerase chain buffer	5µl
Forward primer (100ng/µl)	1.5µl
Reverse primer (100ng/µl)	1.5µl
Deoxynucleotide mix (10mM)	2.5µl
DNA Template (5-50ng)	3µl
Pfu DNA polymerase	2.5µl
Sterile molecular grade water	34µl

The PCR cycles were as follows:

	X 20			
95°C	95°C	55°C	68°C	68°C
2 minutes	30 s	1 min	1min/kb	7 mins

10µl of PCR product was run on a 1.5% agarose gel to confirm DNA amplification. 2µl of *DpnI* was then added to the remaining 40µl of PCR mix and incubated at 37°C for 1.5hrs to digest the template super-coiled double stranded DNA.

5.2.2.4 Gel Extraction of DNA

For the extraction of DNA from agarose gels, the QIAquick Gel Extraction Kit (Qiagen) was used according to the manufacturers' instructions. Samples were eluted either in 30µl-50µl of either elution buffer or molecular biology grade water. Prior to gel extraction, gels were blotted with Whatman Grade 1 (3mm) paper in order to avoid contamination of the extracted product.

5.2.2.5 Mutation Screening

Regions of sequence were screened for polymorphisms using the Lightscanner software (Idaho Technologies) and BigDye Sanger sequencing. The protocols for these methods are described in chapter 2.5.2 and 2.6 respectively.

5.2.3 Cloning

5.2.3.1 Plasmids

Plasmid	Description	Selection	Supplier
pGEM-T	'T' overhang cloning vector that allows direct insertion of PCR products.	Ampicillin	Promega
pGL3-basic	Luciferase reporter vector lacking any eukaryotic promoter or enhancer sequences. Allows for quantitative analysis of factors that could potentially regulate expression.	Ampicillin	Promega
pGL3-promoter	pGL3-basic backbone with the addition of SV40 promoter for the analysis of putative regulatory elements outside of the promoter region.	Ampicillin	Promega

Table 5.1 Description of Plasmids used in this chapter for luciferase reporter gene assays.

5.2.3.2 Cloning into a pGEM-T® Vector

In order to ensure precise cloning of putative regulatory regions/SNPs into a pGL3 luciferase vector, PCR products were first cloned using a pGEM-T® vector (Promega). The pGEM-T® vector consists of a pGem-57f(+) plasmid DNA which has been cleaved at the EcoR V site and had a single 3' terminal deoxy-thymidine added to both ends. As Taq DNA polymerase adds a deoxy-adenosine to the 3'-termini of any amplification product [338], ligation between a desired insert and the pGEM-T® vector can take place via TA ligation.

3µl of purified PCR product was ligated with 1µl of pGEM-T® vector (50ng) using T4 DNA ligase and ligase buffer. A control insert was ligated in parallel as a positive control. A negative control was also performed with water. All reactions were incubated at 4°C overnight. Products were then visualised on a 1.5% agarose gel to check ligation.

5.2.3.3 Preparation of Chemically Competent *E. coli* XL1-Blue Cells

Competent cells were prepared from XL1-blue *E. coli* by either Dr Caroline Tinsley or Matthew Lock using a standard protocol [339].

5.2.3.4 Preparation of Electrocompetent *E. coli* XL1-Blue Cells

Electrocompetent cells were prepared from XL1-blue *E. coli* by either Dr Caroline Tinsley or Matthew Lock.

5.2.3.5 Transformation of Bacteria by Heat-shock

50µl of chemically competent cells were thawed on ice and mixed thoroughly. 5µl of ligation/SDM mix or 1µl of control plasmid pUC18 (10ng/µl) were then mixed with the competent cells in a chilled eppendorf tube and incubated on ice for 30 mins. Cells were then heat-shocked for 45-50s at 42°C exactly and cooled on ice for a further 2 mins. After the addition of 950µl of LB media to each ligate/cell mix, cells were incubated at 37°C for 30-60 mins to allow recovery. The transformed cells were spun for 1 min at 13000rpm with the resulting pellet resuspended in 200µl of LB media. The resuspended cells were plated onto LB agar plates supplemented with carbenicillin, left to dry for 5-10 minutes, and subsequently incubated overnight at 37°C.

5.2.3.6 Transformation of Bacteria by Electroporation

5µl of ligation product was mixed with 40µl of electrocompetent cells on ice for 5 mins before transferring to an ice cold cuvette (0.2 cm electrode gap). After which cells were transformed by a pulse of electricity, 2.5kV, 10µF, 600Ω, 5msec (BioRad MicroPulser™ electroporator machine) and resuspended immediately in 1ml LB in a Falcon 2059 polypropylene tube. The cells were allowed to recover at 37°C for 30-60 minutes, spun for 1min at 13,000rpm and the resulting pellet resuspended in 200µl of LB media. The resuspended cells were then plated onto LB agar plates supplemented with carbenicillin, left to dry for 5-10 minutes and subsequently incubated overnight at 37°C.

5.2.3.7 Blue/white Screening of pGEM-T® Vector

Successful cloning of an insert into a pGEM-T® vector interrupts the coding sequence of the lacZ gene which produces β-galactosidase. β-galactosidase cleaves X-gal to produce galactose and 5-bromo-4-chloro-3-hydroxyindole. 5-bromo-4-chloro-3-hydroxyindole then is oxidized into 5,5'-dibromo-4,4'-dichloro-indigo, an insoluble blue product. As a result clones that contain PCR products, produce white colonies while blue colonies correspond to reformed pGEM-T® vectors. Recombinant clones can therefore be identified by colour screening on indicator plates containing Xgal. Indicator plates were prepared by spreading 20µl Xgal and 10µl Isopropyl β-D-1-thiogalactopyranoside (IPTG) across a solid media plate.

5.2.3.8 Preparation of Plasmid DNA

Desired colonies were picked from the solid media plate and suspended in 5mls of LB media plus 5µl of carbenicillin. The cells in solution were incubated overnight at 37°C. For all plasmid preparations DNA was then isolated using the QIAGEN Spin Miniprep kit according to the manufacturers' instructions. DNA was quantified using a spectrophotometer (BioPhotometer, Eppendorf) at 260nm/280nm (A260).

5.2.3.9 Restriction Digestion

The restriction enzymes used in this chapter, plus their reaction conditions are given in Table 5.2.

RE	Recognition Sequence	Buffer	Incubation Temperature
Mlu	ACGCGT	NEBuffer 3	37°C
NcoI	CCATGG	NEBuffer 3	37°C
BamHI	GGATCC	NEBuffer 3 + BSA	37°C
SalI	GTCGAC	NEBuffer 3 + BSA	37°C
FseI	GGCCGGCC	NEBuffer 4 + BSA	37°C

Table 5.2 Restriction Enzymes used in this chapter plus reaction conditions. All restriction enzymes were purchased from New England Biolabs (NEB).

Typical reactions contained 250ng-2μg DNA in a final reaction volume of 50μl. These reactions contained 10% buffer, 1x BSA (if required), the desired restriction enzyme (<10% total volume) plus sterile molecular biology grade water. Typically 10 units of enzyme were added per μg of DNA. The mixture was incubated for 4h at 37°C, then heat inactivated at 65-80°C depending of restriction enzyme used. For reactions comprising of more than one restriction enzyme, the buffer suggested by the manufacturer to be compatible with both enzymes was used. Enzymes were either added together at the start of the reaction or sequentially to improve efficiency.

After endonuclease digestion of the pGL3 vector, dephosphorylation of the 5' DNA termini was performed in order to prevent the relegation of the 5' and 3' ends of the vector. Vector DNA was treated with 1μl (1unit/μl) of calf intestinal alkaline phosphate (CAP) for 1 h at 37°C.

All digested products were then run on a 2% agarose gel in order to separate digested bands. Digests of a different nature, such as pGL3 vector and PCR product digests, were run in separate tanks to avoid cross contamination. The appropriate DNA fragment was then gel extracted as previously described (chapter 5.2.2.4).

5.2.3.10 Ligation of Digested Insert into pGL3

Ligation of digested DNA and digested pGL3 vector was performed using T4 DNA ligase (Promega). A typical reaction was carried out in a skirted 96-well plate using a range of ratios of PCR insert to pGL3 vector (3:1-14:1). A 10µl or 20µl reaction was set up using 1µl or 2µl 10X ligation buffer (Promega) plus 1-2 units of T4 DNA ligase. Negative ligations were also carried out simultaneously to confirm the complete digestion and dephosphorylation of the vector. Reactions were mixed thoroughly and centrifuged briefly before incubating either at room temperature for 10mins, 16°C for 4 hours or overnight at 4°C.

5.2.3.11 Digestion and Sequencing of Construct

Confirmation of a successful ligation between insert and vector was two fold. Firstly plasmid DNA extracted using a miniprep kit, as described in 5.2.3.8, was subjected to digestion by the appropriate restriction enzymes. Digested products were then run on an agarose gel and the product size(s) checked against those expected. Secondly, the junctions between insert and vector were sequenced using sequencing primers designed approximately 100bp away, 5' and 3' respectively, from the two cloning sites.

5.2.4 Dual-luciferase® Assay

5.2.4.1 Plating of HEK 293 Cells

The plating density of HEK 293 cells for the reporter assay was 10,000 cells per well in 100µl volume. This was achieved by counting cells using a haemocytometer (Figure 5.4).

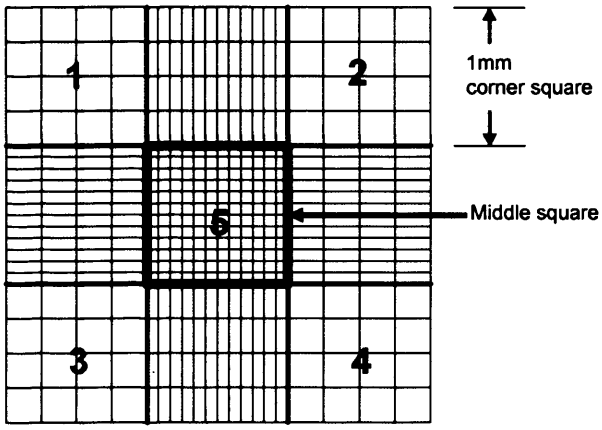


Figure 5.4. Standard Haemocytometer chamber.

To calculate the number of cells present the middle square was counted (N). When the coverslip is in the correct position the depth of the chamber is 0.1mm. Therefore the volume of one square is $0.1mm^3$ (depth of the chamber 0.1mm x area of square $1mm^2$) and the number of cells in $1cm^3 = Nx10^4$. If the number of cells in the square (N) was less than 100 then the number of cells within all five squares were counted. In this case:

$$\text{No of cells/ml} = \frac{\text{Total no counted}}{5} \times 10^4$$

5.2.4 Dual-luciferase® Assay

5.2.4.1 Plating of HEK 293 Cells

The plating density of HEK 293 cells for the reporter assay was 10,000 cells per well in 100µl volume. This was achieved by counting cells using a haemocytometer (Figure 5.4).

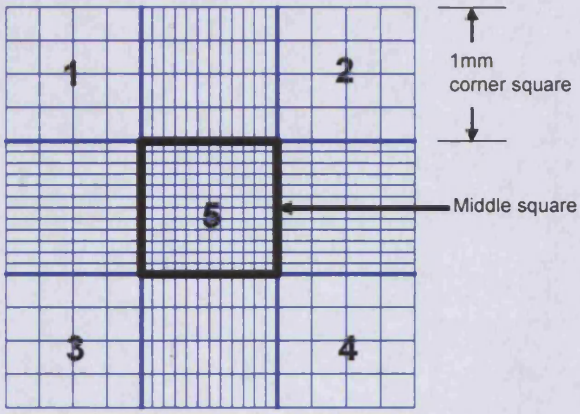


Figure 5.4. Standard Haemocytometer chamber.

To calculate the number of cells present the middle square was counted (N). When the coverslip is in the correct position the depth of the chamber is 0.1mm. Therefore the volume of one square is $0.1mm^3$ (depth of the chamber 0.1mm x area of square $1mm^2$) and the number of cells in $1cm^3 = Nx10^4$. If the number of cells in the square (N) was less than 100 then the number of cells within all five squares were counted. In this case:

$$\text{No of cells/ml} = \frac{\text{Total no counted}}{5} \times 10^4$$

5.2.4.2 Transfection of HEK 293 Cells

HEK293 cells were plated out at 10,000 cells/ml into a 96-well flat bottomed tissue culture treated sterile plate (Greiner Bio One) and incubated overnight at 37°C with 5% CO_2. Transfection was carried out with 500ng of transformed pGL3 reporter vector and 5ng of the control reporter vector. For each set of 6 replicates a transfection master mix was made which consisted of 1.5µl of fugene and 97µl of MEM glutaMAX media. This solution was briefly vortexed then incubated at room temperature for 5 minutes. 5µl of the control reporter vector was added to each master mix at a concentration of 1ng/µl. Apart from the negative control, 5µl of pGL3 luciferase reporter construct (100ng/µl) was also added. All tubes were vortexed briefly and incubated at room temperature for 30 minutes. 10µl of this solution was then added directly to the cells of each replicate within the 96 well plate. The plate with lid was swirled gently and returned to the incubator overnight.

5.2.4.3 Detection of Reporter Proteins

Reporter gene expression of pGL3 recombinants was performed using the dual-luciferase® reporter assay system (Promega). In this system both firefly and *Renilla* luciferase are detected. The dual-luciferase reporter assay system was used following the manufacturers instructions. Briefly, media was gently aspirated from the transformed HEK cells. Cells were then washed in 50µl of PBS and lysed in 20µl 1x passive lysis buffer (PLB). The plate was swirled gently on a bell-dancer platform for 15-30 minutes and then assayed using a dual injecting plate reading lumininometer (MicroLumat Plus LB96V, Berthold technologies). The firefly and *Renilla* luciferase catalysed reactions were activated through the addition of 100µl of LAR II to detect firefly luciferase activity followed by the addition of 100µl stop and glo to detect the *Renilla* activity. An overview of the chemistry is given in chapter 5.1 and a comprehensive explanation supplied in Part #TM040 (Promega).

5.2.4.4 Statistical Analysis of Luciferase Reporter Assay

Light emission (Relative light units, RLU) proportional to the expression of firefly and *Renilla* luciferase expression was collected using a dual injecting plate reading luminometer ((MicroLumat Plus LB96V, Berthold technologies) at interval time points 1 and 2 respectively. In order to account for the variation in transfection efficiency between samples, a ratio of firefly expression/*Renilla* expression was taken for each sample.

To allow for comparison between assays, the average ratio for all control (pGL3-Promoter) replicates from one assay were taken and ratios from the assay normalised with respect to this average. In each assay six technical replicates were performed for each vector construct, so an average from the six normalised ratios was taken for each construct.

Initially, in order to confirm that any differences detected were reproducible and not due to experimental variability, 18 technical replicates were performed for each construct (6 replicates per assay, 3 assays). A second set of experiments were then carried out with 1-2 biological replicates of each recombinant to determine whether any differences observed were due to a specific clone artefact.

For comparative analysis of constructs, all technical replicates of the same construct were averaged and regarded as one biological replicate. To determine whether there were any differences in luciferase expression between constructs, a one way ANOVA was performed on the luciferase expression levels of all biological replicates of the recombinant constructs. Where appropriate an independent samples t-test was then performed. A Levene's test for homogeneity of variance was used to establish whether unequal or equal variance should be assumed. Standard error of the mean (STDEV($\sqrt{}$(n)) was also calculated.

5.3. Results

In order to examine whether genomic variation at rs2619538 or rs13198512 affects gene expression, an attempt was made to clone each SNP plus surrounding sequence into a suitable firefly luciferase reporter gene vector (pGL3). The cloning of each SNP involved a number of steps including; determining the appropriate amount of surrounding sequence to clone for each SNP, identifying individuals homozygous for all variants within the region, PCR amplification and ligation of the region into a pGEM-T® vector, digestion of the recombinant pGEM-T® vector with suitable restriction enzymes, ligation of the digested insert with a suitable pGL3 vector, site directed mutagenesis of the pGL3 construct to create all common haplotype combinations within the inserted sequence and finally, quantification of luciferase expression for each construct. Details of each of these steps are given below.

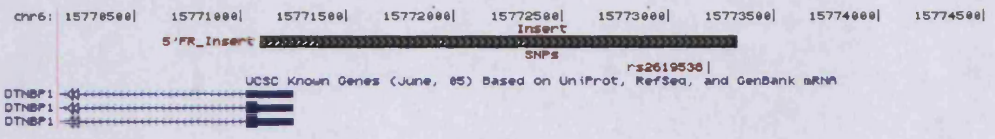
5.3.1 Primer Design

PCR primers were designed for both rs2619538 and rs13198512 which would amplify the sequence spanning each SNP. If a SNP was near or within a potential regulatory region, for example the sequence immediately 5' to the DTNBP1transcription start site, then this sequence was also included in the amplification product. Each set of primers also incorporated recognition sites for restriction enzymes (RE) that would allow the cloning of the PCR product into a pGL3 vector. REs selected were dependent on the luciferase cloning vector used and the designated position of cloning within the vector. Both these aspects were determined by the genomic location of rs2619538 and rs13198512 relative to the DTNBP1 gene.

rs2619538 is 1868bp upstream of the DTNBP1 coding sequence. Therefore PCR primers were designed to amplify a 2.2kb region which included rs2619538 plus the sequence immediately 5' to DTNBP1 (chr6: 15771090-15773323, see Figure 5.5A). These primers also included RE sites (MluI and NcoI in the forward and reverse primers respectively) which would allow the insertion of the DTNBP1 putative promoter region immediately upstream of the luciferase gene within a pGL3-Basic

reporter vector (Promega). The pGL3-Basic vector lacks any eukaryotic promoter or enhancer sequences, therefore any luciferase expression detected will be due to the inserted sequence. Consequently the pGL3-Basic vector is commonly used for the insertion of putative promoter regions. The cloning of the 2.2kb region 5' of DTNBP1 will allow the effect of SNPs such as rs2619538 plus any other potential variants subcloned into the vector to be analysed with a native DTNBP1 promoter rather than an artificial simian virus (SV) promoter. The position of cloning and the major features of the pGL3-Basic vector are given in Figure 5.5B.

A.



B.

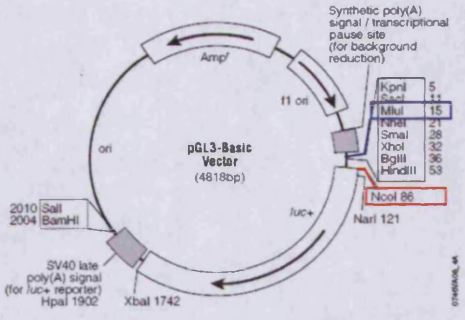


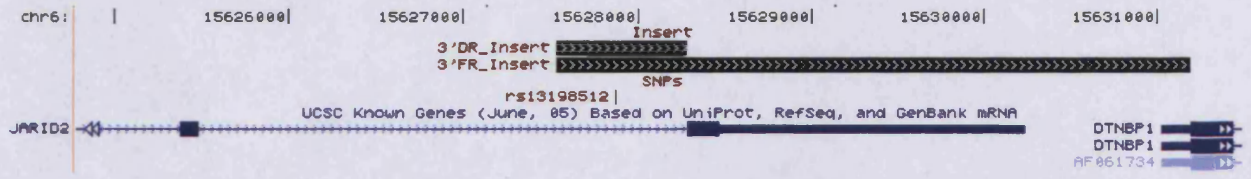
Figure 5.5. A. Chromosomal location of rs2619538 and the putative promoter sequence to be inserted into a pGL3-Basic luciferase vector. B. Basic structure of the pGL3-Basic luciferase vector (Promega technical manual TM033). The DTNBP1 5' flanking region will be inserted 5' to luciferase vector between the NcoI and Mlu sites (shown in red and blue respectively).

rs13198512 is 3154bp downstream of the DTNBP1a coding sequence. PCR primers were designed to amplify a 3.6kb region which included rs13198512 plus the 3'UTR of DTNBP1 (chr6: 15,627,534-15,631,184, see Figure 5.6A). PCR primers were designed to include RE recognition sites (FseI and BamHI in the forward and reverse primers respectively) which would allow this 3'flanking region to be inserted downstream of the luciferase gene within a pGL3-Promoter (pGL3-P) reporter vector (Promega). The pGL3-Promoter vector contains a SV40 promoter upstream of the luciferase gene. As a result this vector can be used to study potential regulatory elements away from the sequence immediately 5' of the gene of interest. Digestion of the pGL3-P vector with

FseI and BamHI, followed by the subsequent ligation of the DTNBP1 3' flanking region, will remove the native pGL3 polyA signal and replace it with the DTNBP1 3'UTR (Figure 5.6B).

It is generally suggested that the regions selected for PCR amplification and subsequent cloning should be kept under 3kb as regions larger than this have a higher rate of failure at the ligation stage [340]. As the DTNBP1 3'flanking region is >3kb, PCR primers were also designed to amplify a smaller, distal region immediately spanning rs13198512 (chr6: 15,627,534-15,628,278, see Figure 5.6A). PCR primers for this 750bp region were designed to include REs that would allow cloning of the region 3' to the luciferase gene again within a pGL3-Promoter vector. However as this region does not contain the DTNBP1 3'UTR, RE sites were chosen downstream of the polyA signal (BamHI and SalI in the forward and reverse primers respectively, Figure 5.6B).

A.



B.

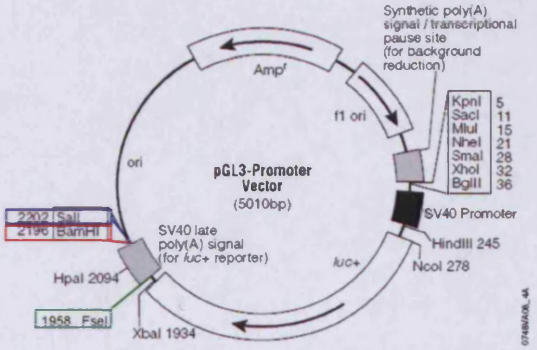


Figure 5.6. A. Chromosomal location of the two regions, 3' Flanking region (FR) and 3'distal region (DR), which include rs13198512, to be inserted into pGL3-Promoter luciferase vectors (Promega). B. Basic structure of the pGL3-Promoter luciferase vector (Promega technical manual TM033). The DTNBP1 3' flanking region will be inserted 3' to luciferase gene between the FseI (green) and BamHI (red) sites replacing the native polyA signal. The 3'distal region will be inserted 3' to luciferase gene between the SalI (blue) and BamHI (red) sites.

A detailed overview of the positions of cloning and the restriction enzyme recognition sites within the pGL3-Basic/Promoter reporter vectors discussed are given in Figure 5.7. The PCR primers designed and the reaction conditions for the incorporated restriction enzymes are given in Table 5.3.

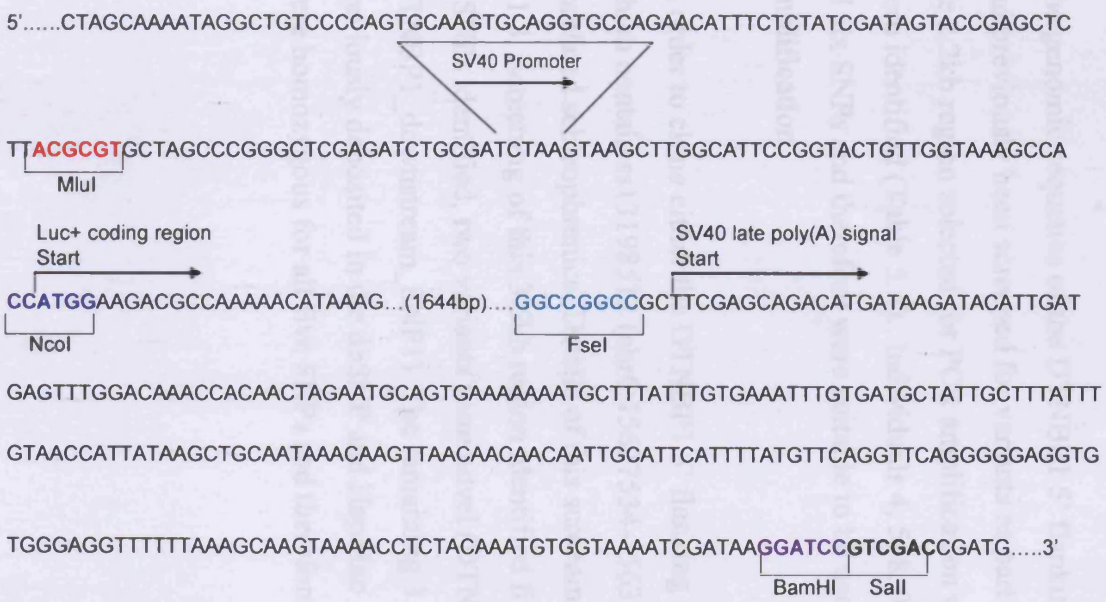


Figure 5.7. Restriction enzyme sites within pGL3-Basic luciferase vector. DTNBP1 5' flanking region including rs2619538 incorporates the RE sites MluI and NcoI. DTNBP1 3' flanking region incorporates the RE sites FseI and BamHI.

Region	Chromosomal Position	Size	PCR Primer	Sequence	RE site
5' Flanking Region	chr6:15,771,090-15,773,323	2233	Forward	ACGCGTcccccagctacaagctaag	Mlu
			Reverse	CCATGGCCggtctcctctcctca	NcoI
3' Flanking Region	chr6:15,627,534-15,631,184	3650	Forward	GGATCCtgtttgggggtgaaagga	BamHI
			Reverse	GTCGACcgaaggcacctttaactt	SalI
3' Distal Region	chr6:15,627,534-15,628,278	744	Forward	GGCCGGCCattgggacatgggcgttg	FseI
			Reverse	GGATCCgaaggcacctttaacttg	BamHI

Table 5.3. PCR primer sequence for 5' flanking region, 3' flanking region and 3'distal region. RE recognition sites are also given in capitals. DTNBP1 3' distal region including rs13198512 incorporates the RE sites BamHI and SalI.

5.3.2 Genomic Screen of rs2619538 and rs13198512 Flanking Sequence

In order to perform a luciferase reporter assay comparing the regulatory affect of each allelic variant of a SNP, constructs need to prepared which contain all haplotypic combinations within the inserted sequence. Production of these constructs can be done in one of two ways. Either, individuals with unique haplotype combinations are PCR amplified and cloned into separate pGL3 luciferase vectors or alternatively DNA from an individual, homozygous for all SNPs within the desired region, is cloned into a relevant pGL3 vector, after which site directed mutagenesis is performed on the construct to create all possible haplotype combinations. In this study the later strategy was used.

The genomic sequence of the DTNBP1 5' flanking region (chr6:15771090-15773323) had previously been screened for variants as part of the analysis for chapter 3. Within the 2.2kb region selected for PCR amplification with rs2619538, five additional SNPs were identified (Table 5.4). Individuals 4, 5, 8, 9, 11, 12 and 13 were homozygous for all six SNPs and therefore were suitable to be used as a template for initial PCR amplification.

In order to clone either the DTNBP1 3' flanking region or the 3'distal region, both of which contain rs13198512 (chr6:15627534-15631184) was screened through 14 unrelated schizophrenics. Details of this screening sample can be found in chapter 2.1.3. Screening of this 3.6kb region identified five polymorphisms (Table 5.5). Of the 5 SNPs identified, two variants were novel (DTNBP1_downstream_indel and DTNBP1_downstream_SNP1). The remaining 3 polymorphisms had all been previously deposited in the dbSNP and HapMap databases. Individuals 1, 2, 7, 9, 11 were homozygous for all five SNPs and therefore suitable for initial PCR amplification.

				Screening Set Individual													
SNP ID	Chromosomal Position	Alleles	CEU MAF	1	2	3	4	5	6	7	8	9	10	11	12	13	14
DTNBP1-C16_SNP1	15771705	C/G	0	CC	CC	CC	CC	CC	CC	CG	CC	CC	CC	CC	CC	CC	CC
rs2619536	15771826	T/C	0.09	TT	TT	TC	TT	TT	TT	TT	TT	TT	TT	TT	TT	TT	TT
DTNBP1-C16_SNP2	15771883	C/T	0.09	CT	CC	CC	CC	CC	CC	CC	CC	CC	CT	CC	CC	CC	CC
rs2619537	15772392	A/G	0.16	AG	AA	AG	AA	AA	AA	AA	AA	AA	AG	AA	AA	AA	AA
rs12204704	15773184	G/A	0.16	GA	GA	GG	GG	GG	GG	GA	GG	GG	GA	GG	GG	GG	GA
rs2619538	15773188	T/A	0.39	AT	AA	AA	AA	AA	AT	AA	AA	AA	AT	AA	AA	AA	AA

Table 5.4. Screening results for chr6:15771090-15773323. Six SNPs were identified. Individual genotypes for the 14 subjects screened are given with heterozygous genotypes indicated in red. Chromosomal position is based on March 2006. Individuals 4, 5, 8, 9, 11, 12 and 13 are homozygous for all SNPs within this region. MAF = Minor Allele Frequency.

				Screening Set Individual													
SNP ID	Chromosomal Position	Alleles	CEU MAF	1	2	3	4	5	6	7	8	9	10	11	12	13	14
rs1047631	15631080	A/G	0.2	AA	AA	AA	AG	AG	AG	AA	AA	AA	AA	AA	AA	AA	AA
DTNBP1_Downstream_indel1	15630460	Unknown	N/A	WT	WT	WT	WT	WT	ID	WT	WT	WT	WT	WT	ID	WT	WT
DTNBP1_Downstream_snp1	15629000	C/T	N/A	CC	CC	CC	CC	CC	CC	CC	CT	CC	CC	CC	CC	CC	CC
rs13198533	15627893	A/G	0.255	AA	GG	AG	AA	AA	AG	AA	GG	GG	AG	GG	GG	AG	AG
rs13198512	15627864	A/G	0.492	AA	GG	AG	AA	AA	AG	AA	AG	GG	AG	GG	AG	AG	AG

Table 5.5. Screening results for chr6:15627534-15631184. Five SNPs were identified. Individual genotypes for the 14 subjects screened are given with heterozygous genotypes indicated in red. Chromosomal position is based on March 2006. MAF = Minor Allele Frequency. ID = Insertion/Deletion. WT = Wild type. Individuals 1, 2, 7, 9 and 11 are homozygous for all SNPs within this region.

5.3.3 PCR Amplification and Gel Extraction

All three genomic regions (DTNBP1 5'flanking region, DTNBP1 3'flanking region and DTNBP1 3'distal region) were successfully amplified using the primers listed in Table 5.3. As each high fidelity PCR reaction produced multiple bands when visualised on an agarose gel (Figure 5.8) the correct size product for each PCR (noted by the arrows) was gel extracted.

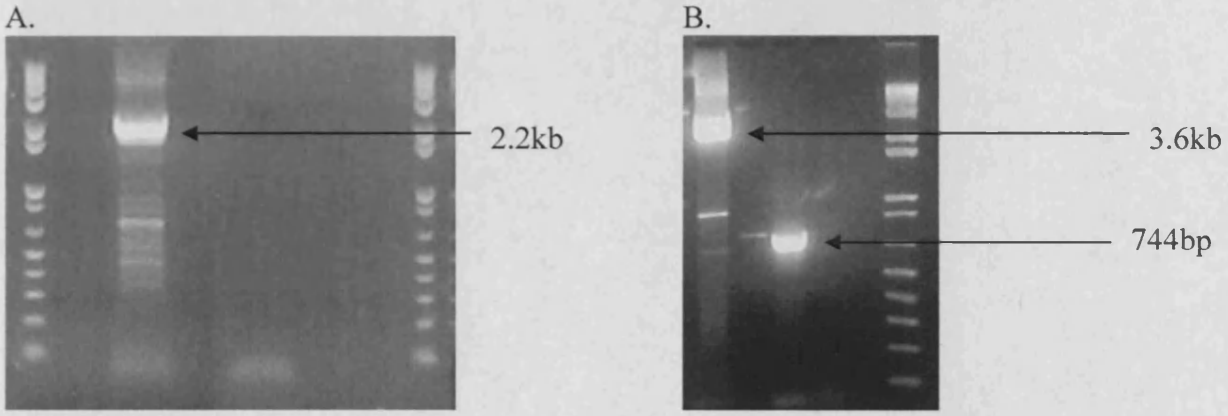


Figure 5.8. High Fidelity PCR of three regions (5'FR, 3'FR, 3'DR). A. 5' flanking region PCR (2.2kb) B. 3'flanking region (3.6kb) and 3' distal region (744bp). Arrows indicate PCR products of the correct size which were subsequently gel extracted. Lanes immediately to the right of each product are negative control runs.

5.3.4 pGEM-T® Vector Cloning

RE sites close to the end of a sequence are often ineffectively cleaved by their respective enzymes [341]. For example SalI needs 3bp each side of the recognition sequence to cut efficiently [342]. As the restriction enzyme sites incorporated into the PCR primers are located at the 3' and 5' ends of each PCR fragment, direct enzymatic treatment of the gel extracted PCR product would be likely to be inefficient.

As the pGEM-T® vector system (Figure 5.9) facilitates the direct cloning of PCR products without requiring prior enzymatic treatment, TA ligation of the gel extracted PCR products into pGEM-T® was used as an intermediate step prior to restriction

enzyme digestion and ligation into the appropriate pGL3 vector. Potential pGEM-T® recombinants were selected by an initial blue/white colony screen. In order to determine that the correct constructs had been produced, prospective recombinants (white colonies) were subjected to RE digest plus sequencing across the ligation boundary.

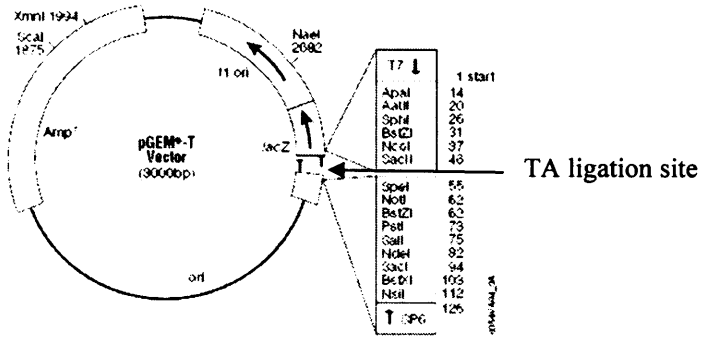


Figure 5.9. pGEM-T® vector. Promega technical manual TM042.

While a control insert was used to determine that the ligation conditions were correct and a range of ligation ratios were attempted (insert:vector 1:1- 3:1) no white colonies were detected for the ligation reaction between pGEM-T® and DTNBP1 3' flanking region.

As it is possible that the recombinant vector may not cause a frame shift in the lacZ gene and therefore white colonies will not be produced, a random selection of blue colonies were subjected to RE treatment and the predicted ligation boundaries sequenced. However all colonies analysed were re-ligated pGEM-T® vectors. This ligation failure may be due to the fact that the 3'flanking region (3.6kb) is larger than the pGEM-T® vector (3kb).

In contrast, the DTNBP1 3' distal region was successfully cloned into the pGEM-T® vector (pGEMT-3'DR). This recombination vector was therefore used in all further experiments investigating the effect of rs13198512. The DTNBP1 5' flanking region

was also successfully cloned into a pGEM-T® vector (pGEMT-5'FR). Sequence traces of boundaries of all recombinants can be found in Appendix Figures 9.2.1 and 9.2.2.

5.3.5 Digested Promoter Insert and pGL3-Basic Cloning

After the successful cloning of the DTNBP1 5' flanking region (chr6:15771090-15773323) into a pGEM-T® vector, an attempt was made to clone the 5'flanking region into a pGL3-Basic luciferase vector. Both the recombinant pGEMT® vector (pGEMT-5' FR) and the pGL3 luciferase vector were digested with restriction enzymes NcoI and MluI. In order to reduce the chance of religation, the digested pGL3-Basic product was treated with CAP to dephosphorylate the 5' DNA termini. Both the digested products of pGEMT-5' FR and pGL3 were run on an agarose gel where bands of the expected size were observed and gel extracted (Figure 5.10).

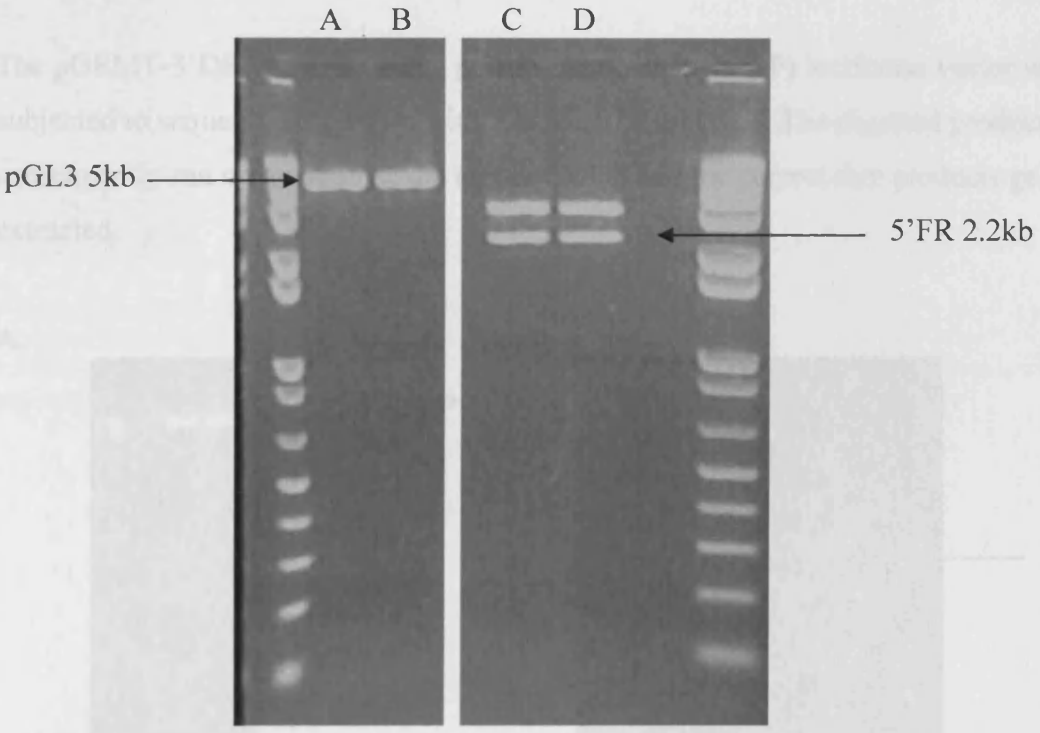


Figure 5.10. Visualisation of RE digest products. The sequential digestion of pGL3-Basic and pGEMT-5' FR are shown. A pGL3-Basic NcoI digest followed by MluI. B pGL3-Basic MluI digest followed by NcoI digest. C. pGEMT-5' FR NcoI digest followed by MluI. D. pGEMT-5' FR MluI digest followed by NcoI digest.
pGL3 5kb bands and 5'FR 2.2kb bands from A, B, C and D were gel extracted and subjected to a ligation reaction.

5µl of each gel extracted product was run on a second agarose gel to ensure a successful extraction had occurred. Ligation reactions were then performed with the remaining gel extracted products. A number of variations in ligation conditions were attempted including insert:vector ratios of between 1:1 and 20:1, as well as three different ligation temperatures (10 minutes at room temperature, 4 hours at 16°C and 12-16 hours at 4°C). Furthermore, transfection of all ligated products was attempted via heat shock and electroporation. Although a number a colonies were observed after incubation, digest and sequencing analysis determined that all colonies produced were reformed pGL3-Basic vector. Due to the failure of the 5' flanking region to be cloned into the pGL3-Basic vector functional analysis of rs2619538 via reporter gene assay could not be performed.

5.3.6 Digested 3'Insert and pGL3 Promoter Cloning

The pGEMT-3'DR construct and a pGL3-Promoter (pGL3-P) luciferase vector were subjected to sequential digestion with REs BamHI and SalI. The digested products were subsequently run on an agarose gel (Figure 5.11) and the correct size products gel extracted.

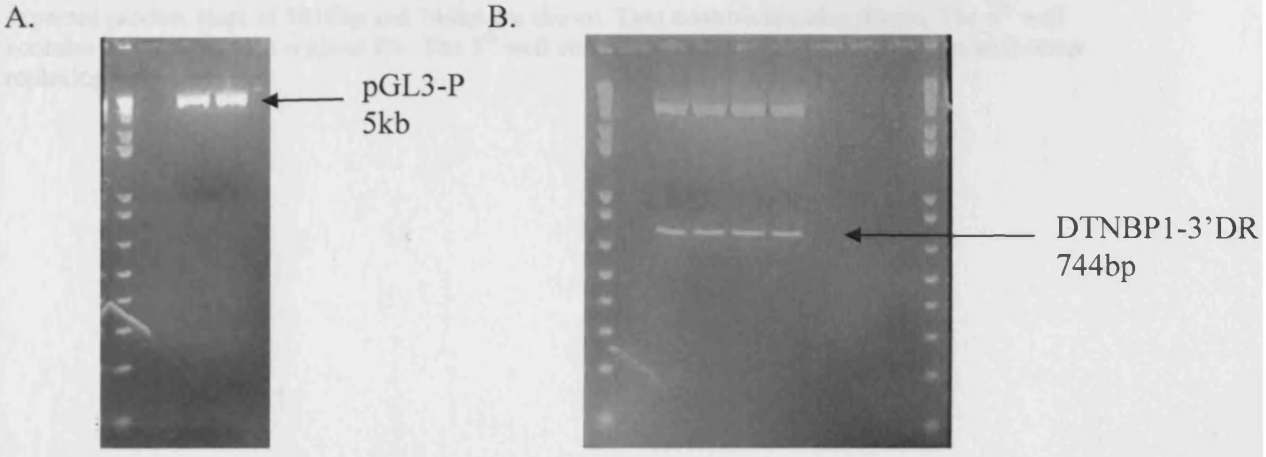


Figure 5.11. Product from BamHI and SalI sequential digestion of A. pGL3-Promoter plasmid and B. pGEMT-3'DR. Expected product sizes were 5010bp for pGl3-P digestion and 3000bp and 744bp for pGEMT-3'DR digestion. Digested products were run on separate gels and in separate tanks so as to avoid contamination. Bands at 5010bp and 744bp (depicted by arrows) were gel extracted, subjected to ligation and transformation into *E.coli*.

A ligation reaction between the digested products was then performed. The DTNBP1-3'DR was successfully cloned into the BamHI-SalI site of the pGL3-Promoter vector using a 14:1 insert:vector ratio. After transformation recombination was confirmed through digestion of the extracted DNA with BamHI and SalI (Figure 5.12) and sequencing across the ligation boundaries (Appendix Figure 9.2.3). Thus the 744bp sequence downstream from DTNBP1, containing rs13198512, was cloned 3' to the luciferase gene within a pGL3-Promoter vector. The resultant plasmid was designated pGL3-P_3'DR.

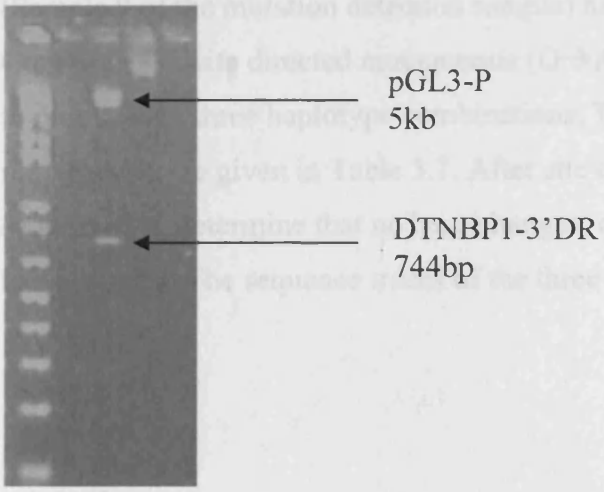


Figure 5.12. Result of BamHI and SalI digestion of extracted DNA from transformed colony. The expected product sizes of 5010bp and 744bp are shown. Two controls are also shown. The 4th well contains a digest reaction without RE. The 5th well contains a negative control (RE digest with water replacing DNA).

5.3.7 Site Directed Mutagenesis

In order to perform a comprehensive analysis of the potential regulatory effect of a SNP, all haplotypes with the inserted region were considered. Previous sequencing results found that in addition to rs13198512, the DTNBP1 3'DR amplimer also contains rs13198533 (Figure 5.13). Therefore the haplotype frequencies of these SNPs were determined using the HapMap database.

Within the CEU population these two SNPs form one of three haplotype combinations (Table 5.6). The individual used for initial PCR amplification of the DTNBP1 3'DR (Sample 9 of the mutation detection sample) has the genotype GG for both SNPs. Consequently, site directed mutagenesis (G→A) was performed on both sites in order to produce the three haplotype combinations. The primers used for site directed mutagenesis are given in Table 5.7. After site directed mutagenesis all vectors were sequenced to determine that no base changes, other than rs13198512 and rs13198533, had occurred. The sequence traces of the three constructs are given in Appendix 9.2.4.

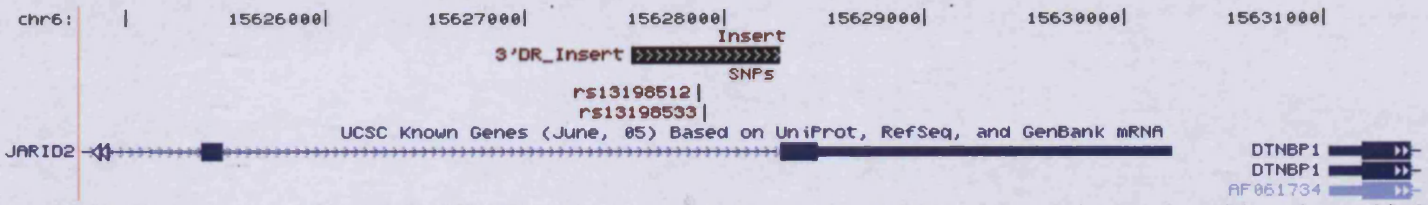


Figure 5.13. Chromosomal position of the DTNBP1 3'distal region insert sequence (3'DR) and the relative location of the polymorphic SNPs rs13198512 and rs13198533 within this region.

Haplotype	rs13198512	rs13198533	Freq CEU	Plasmid
1	G	G	0.50	pGL3-P_3'DR_GG
2	A	G	0.26	pGL3-P_3'DR_AG
3	A	A	0.24	pGL3-P_3'DR_AA

Table 5.6. Haplotypic analysis of rs13198512 and rs13198533 within the CEU population.

Primer Name	Sequence change	Primer Sequence (5' to 3')
rs13198512_149GA_F	cactgcctctaccacggagcactcagcaaagagacccagggg ctttgttacatcctggatggtgacacaaatacccaagctacttga ctcactttagctaccRgccactgaacctggcc	5'-cctggccatgcaagtaaatacaaacacttctcaacct-3'
rs13198512_149GA_R	atgcaagtaaat[G>A]caaacacttctcaacctatacatatcc ttgtaattcactcaagagaatcatttatctgtcggcaccaaaaac	5'-aggttgagaagtgtttgtatttacttgcatggccagg-3'
rs13198533_120GA_F	cactgcctctaccacggagcactcagcaaagagacccagggg ctttgttacatcctggatggtgacacaaatacccaagctacttga ctcactttagctacc[G>A]gccactgaacctggcc	5'-ctcactttagctaccagccactgaacctggc-3'
rs13198533_120GA_R	atgcaagtaaatRcaaacacttctcaacctatacatatcc ttgtaattcactcaagagaatcatttatctgtcggcaccaaaaac	5'-gccaggttcagtggctggtagctaaagtgag-3'

Table 5.7. Site directed mutagenesis primers.

5.3.8 Luciferase Expression Analysis

In order to determine whether genetic variation at rs13198512 affects transcription, the expression of the firefly luciferase reporter gene was quantified using the dual-luciferase® reporter assay system (Promega). In total five different vector constructs were assayed pGL3-Basic (negative control), pGL3-Promoter vector (positive control), pGL3-P 3'DR_GG, pGL3-P 3'DR_AG, pGL3-P 3'DR_AA. Details of these constructs is given in Figure 5.14.

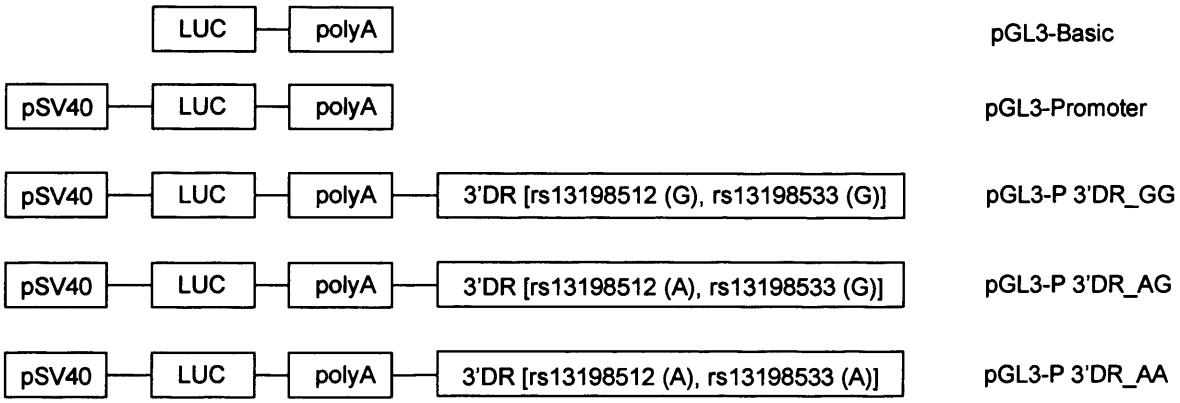


Figure 5.14. Plasmids constructs assayed. pSV40 – SV40 promoter, LUC = luciferase gene, polyA = SV40 polyadenylated signal, 3'DR = inserted sequence (SNP genotypes are given).

Results of the dual-luciferase® assay are illustrated in Figure 5.15. The luciferase/*Renilla* expression ratio shown for each vector construct is derived from the average of two-three biological replicates, each of which comprises of at least 6 technical replicates. The average ratios of each biological replicate are given in Table 5.8. ANOVA of the luciferase expression levels for the biological replicates of pGL3-P 3'DR_GG, pGL3-P 3'DR_AG, pGL3-P 3'DR_AA revealed a significant difference in expression between the constructs ($p=0.024$). *Post-hoc* independent sample t-test were therefore performed. The Levene's statistic for each t-test was non-significant therefore equal variance was assumed. All raw data (luciferase/*Renilla* ratios and normalised results) from each technical and biological replicate can be found in Appendix Table 9.2.5.

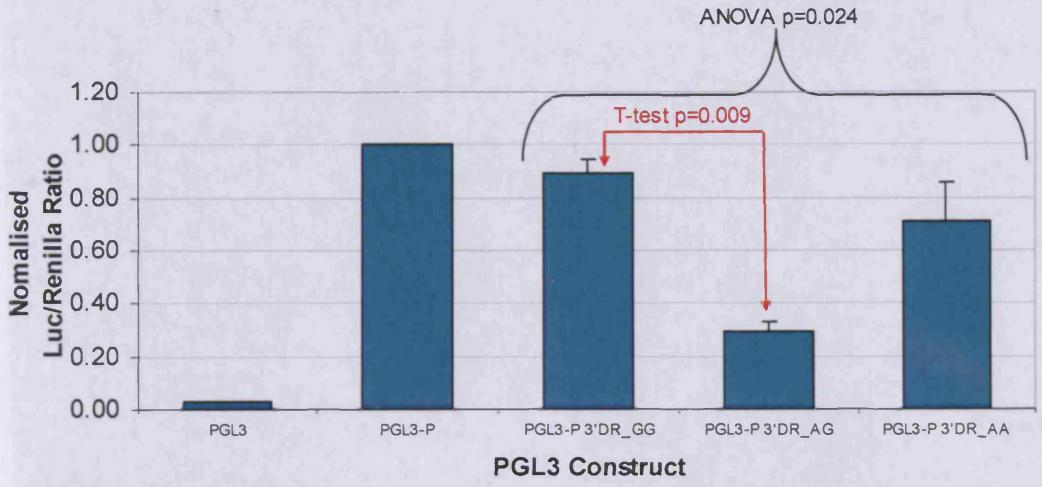


Figure 5.15. Graphical depiction of the results from the dual-luciferase® reporter assay comparing pGL3-P 3'DR vectors containing the haplotypic variants of rs13198512 and rs13198533 (G-G, A-G and A-A respectively). Negative (pGL3-Basic) and positive (pGL3-Promoter) controls are also given.

	Biological Rep 1*	Biological Rep 2	Biological Rep 3	Average	*Post-hoc* T-Test (P Value)
PGL3	0.03				
PGL3-Promoter	1				
PGL3-P 3'DR_GG	0.847	0.945		0.896	GG vs AG (0.009)
PGL3-P 3'DR_AG	0.367	0.269	0.252	0.296	AG vs AA (0.097)
PGL3-P 3'DR_AA	0.469	0.692	0.976	0.712	AA vs GG (0.339)

Table 5.8. Detailed results of the dual-luciferase® reporter assay. The average luciferase/*Renilla* ratio is given for each biological replicate of the different constructs. As an ANOVA test revealed a significant difference in luciferase expression level (p=0.024) *post-hoc* independent sample t–tests were performed. The p values for these t-tests (assuming equal variance) are given (pGL3-P 3'DR_GG vs pGL3-P 3'DR_AG p=0.0009, pGL3-P 3'DR_AG vs pGL3-P 3'DR_AA p=0.097, pGL3-P 3'DR_GG vs pGL3-P 3'DR_AA p=0.339).
* Average of three technical replication assays.

Post-hoc independent samples t-test between pGL3-P 3'DR_GG and pGL3-P 3'DR_AG indicated a significant difference in the luciferase expression levels (Table 5.8) with a three fold increase in expression of the pGL3-P 3'DR_GG construct (average expression of pGL3-P 3'DR_ GG/_AG=3.03). Previous reports have suggested that an increase in expression above 1.5 fold is greater than any predicted assay variation and is highly reproducible in independent replication studies using fresh clones [343]. Therefore, the magnitude of the difference detected plus the number of replicates performed strongly suggests that rs13198512 has a regulatory affect on transcription.

While the direct comparison of pGL3-P 3'DR_GG and pGL3-P 3'DR_AG indicates that rs13198512 is a regulatory variant, the results for pGL3-P 3'DR_AA are inconclusive. In order to analyse all common haplotypes within the insert, pGL3-P 3'DR_AA (rs13198512 A, rs13198533 A) was analysed in addition to pGL3-P 3'DR_AG (rs13198512 A, rs13198533 G). rs13213814, a proxy for rs13198533, was analysed in chapter 4 as part of the allelic expression correlation analysis. This SNP did not show an association with the allelic expression differences (p= 0.128) and therefore was not considered as a functional variant. Consequently, it was expected that pGL3-P 3'DR_AG and pGL3-P 3'DR_AA would show similar luciferase expression results. However although there is no significant difference between pGL3-P 3'DR_AA and pGL3-P 3'DR_AG (p=0.097), there is also no significant difference between pGL3-P 3'DR_AA and pGL3-P 3'DR_GG (p=0.339).

5.4 Discussion

Association analysis between cortical dysbindin mRNA levels and putative regulatory polymorphisms, described in chapter 4, identified that genetic variation at rs2619538 and rs13198512 is associated with reduced dysbindin expression. The aim of this chapter was to determine whether rs2619538 or rs13198512 are genuine functional variants which have a cis-acting regulatory influence on gene expression. In addition, it was hoped that functional characterisation of these SNPs would provide evidence as to whether correlation analysis between putative regulatory polymorphisms and allelic expression data is an effective technique for identifying cis-acting variants.

In order to allow functional characterisation of rs2619538 and rs13198512 an attempt was made to create pGL3 luciferase constructs which would determine the effect of each allelic variant of luciferase expression levels. Disappointingly, functional characterisation of rs2619538 could not be performed as cloning of the DTNBP1 5'FR, which contained rs2619538, into a pGL3 luciferase vector was unsuccessful. There are several reasons why this cloning attempt could have failed. These are summarised in Table 5.9. A number of steps were taken to resolve potential reasons for failure including ineffective RE digestion, relegation of vector, unoptimised ligation, inefficient transfection as well as individual error. However none of these resulted in successful cloning. This suggests that the DTNBP1 5'FR sequence may contain abnormalities, such as a complex secondary structure, that make it resistant to cloning.

Possible Reason	Steps taken
Ineffective digestion of vector/Insert by restriction enzymes	Initial TA ligation into pGEMT vector. Sequential RE digestion with heat inactivation after each reaction. Visualisation of bands of correct product size on agarose gel
Inhibitory complex in PCR product e.g. dATP	PCR product gel purified prior to ligation
Pyridamine dimers formed	UV exposure reduced to a minimum
Re-ligating vector	CIP treatment of vector prior to ligation
The optimum vector:insert not performed during ligation	A wide range of ratios used from 1:1 - 20:1
The optimum ligation conditions not performed	Ligation attempted at 4°C, 16°C and room tempertaure for 16hrs, 4hrs and 10 minutes respectively
Heat shock transfection not suitable	Electroportation attempted
Individual error	Cloning experiment also attempted by Dr Liam Carrol
Sequence abnormalities in PCR fragment such as complex secondary structure/ GC content	

Table 5.9. Summary of possible reasons that the ligation of DTNBP1 5'FR and pGL3-Basic luciferase vector was unsuccessful and the steps taken to try and resolve them.

It was hypothesised that analysis of rs2619538, and potentially rs13198512 through subcloning, in an expression vector with a native DTNBP1 promoter, rather than an SV40 promoter, would be more likely to produce expression levels representative of *in vivo* dysbindin. For this reason, in addition to incorporating rs2619538 into a pGL3 luciferase vector, a secondary objective of the cloning strategy was to insert the putative promoter of DTNBP1. An alternative approach, which may eliminate any incompatible structures, would be to clone ~1kb of sequence immediately surrounding rs2619538. Although this is likely to exclude the DTNBP1 promoter, analysis of this sequence cloned into a pGL3-Promoter vector would allow a preliminary examination of the effect of rs2619538 on transcription. However due to time constraints, this alternative cloning strategy could not be attempted as part of this study.

As analysis of the effect of rs2619538 on expression could not be performed using a luciferase assay, publicly available eQTL datasets were examined. eQTL databases are created by the association mapping of genotypes to gene expression profiles, measured using microarray data [344]. Out of three datasets examined [345-347] only one had analysed rs2619538. The GENe Expression VARiation (GENEVAR) project analysed the gene expression profiles of the 270 individuals genotyped in the HapMap Consortium to elucidate the detailed features of genetic variation underlying gene

expression [347]. However, rs2619538 was not significantly correlated with dysbindin expression in this dataset ($p>0.215$).

In contrast to rs2619538, rs13198512 plus surrounding sequence was cloned into a pGL3 luciferase vector. Although one haplotype construct was inconclusive, the comparison of the expression levels of two other pGL3 luciferase plasmids, differing only by rs13198512, revealed that genetic variation at this SNP had an affect on expression. The construct pGL3-P_3'DR_AG, which contains the A allele of rs13198512, showed a significant reduction in luciferase expression compared to pGL3-P_3'DR_GG, a construct containing the G allele ($p=0.009$). This result is supported by the GENEVAR project which also found the A allele of rs13198512 to be significantly associated with reduced dysbindin expression within CEU individuals ($p=0.03$) [347].

However while these two studies support the hypothesis that rs13198512 has a cis-acting influence on DTNBP1 expression, the effect shown by the reporter assay is in the opposite direction to the correlation analysis described in chapter 4 which predicted that the G allele of rs13198512 would show reduced expression compared to the A allele. This variation in results could be explained by a number of factors. Firstly rather than using definite haplotypes, the allelic expression correlation analysis was based on haplotype probabilities. These probabilities were derived from the genotypes of the relatively small allelic expression sample (149 individuals). As a result the haplotype probabilities and consequently the allele(s) associated with reduced relative expression may be inaccurate. Secondly an allelic expression assay identifies any expression differences that were present *in vivo* at the time of extraction. In contrast a luciferase reporter assay is an *in vitro* assay which analyses the affect on the putative regulatory sequence/SNP on a surrogate gene. In addition the luciferase vector used for the analysis of rs13198512 contains a SV40 promoter rather than the native DTNBP1 promoter. As discussed above one aim of this analysis was to replace the SV40 promoter with the native DTNBP1 promoter. However cloning of the putative DTNBP1 promoter (DTNBP1 5'FR) into a luciferase vector was unsuccessful. Finally activity of a reporter gene may fail to reproduce the expression pattern of its

endogenous equivalent owing to differences in chromatin context [348]. Naked DNA in a plasmid is highly unlikely to adopt a full chromatin structure. As chromatin modifications are important in modulating transcription, a reporter gene assay may miss true biological effects.

Finally, a difference between the reporter gene and allelic expression assays performed within this project, which could potentially explain the alternate alleles associated with reduced expression, is that the two assays were carried out within different tissues types. Luciferase expression was analysed using human embryo kidney cells (HEK). In contrast, allelic expression analysis was performed using RNA extracted from brain tissue. While advantageous for their transfection efficiency and ease of maintenance, HEK transformed cell lines do not provide the same sophisticated level of cellular architecture, subcellular organisation or biochemistry associated with native neuronal populations [337].

New research suggests that the majority of regulatory variants operate in an entirely cell-type specific manner [349]. Another group has determined that this cell-type specific regulation is particularly apparent in enhancer regions while promoter regions show a more consistent regulation across tissue types [350]. This may therefore be relevant to rs13198512 which is found distal to the DTNBP1 gene and could be located within an enhancer region.

As cell-type specific regulation is possible analysis of luciferase expression in neuronal cells may be needed to determine the true regulatory nature of both rs13198512 and rs2619538 in the brain. However again due to time constraints, this could not be performed as part of this project. While the GENEVAR project used lymphoblastoid cell lines derived from blood mononuclear cells, another study has performed expression profiling within brain tissue [346]. However this study did not genotype rs13198512 or rs2619538.

5.4.1 Possible Cis-acting Regulatory Mechanisms of rs13198512

Though differing in the functional allele, the results from the luciferase reporter assay and the correlation analysis both indicate that rs13198512 has a cis-acting regulatory influence on dysbindin expression. As noted in chapter 4, rs13198512 is ~3kb downstream of DTNBP1, within an intron of JARID2. A number of SNPs 3' to the target gene have been shown, via *in vitro* or *in vivo* techniques, to affect transcription however the majority of these are located within the 3'UTR of the gene [351]. This may be to be expected as recent findings suggest, in addition to post-translational modification and degradation [352, 353], the 3'UTR may also control mRNA activity [354] and affect several processes such as mRNA formation [351, 353], stability [355, 356], transport, localisation [357] and translation [358]. Nonetheless a number of SNPs downstream of the 3'UTR have been found to be associated with expression [359-366]. This includes rs356219, a SNP 9kb downstream of the SNCA gene which is associated with *in vivo* SNCA mRNA levels [359]. However the mechanism by which SNPs downstream of the 3'UTR (often known as the 3'downstream sequence or 3'DSS) could affect transcription is not clear. At present all 3'DSS SNPs with a characterised mechanism of function are found within TFBSs. It is therefore likely that these SNPs are within enhancer and/or silencer regulatory elements and affect the binding of transcription factors. While rs13198512 has not been found within a TFBS cluster, it may be possible that it disrupts a TFBS not predicted by the Cluster Buster program. As a result rs13198512 could alter the binding of an activator or a repressor. It is also possible that the sequence surrounding rs13198512 is involved in the initial formation of the pre-initiation complex (PIC). *In vivo* footprinting and chromatin immunoprecipitation has shown RNA polymerase II binding to distal enhancer elements suggesting the PIC may initially form at a distal enhancer rather than at the core promoter [270]. Finally, rs13198512 may alter transcription by affecting DNA curvature (the shape of the DNA sequence), flexibility (the ease with which the DNA can bend to allow interactions of different proteins bound to the DNA) or stability (the ease with which the DNA can become single stranded to allow transcription). Both curvature and flexibility have been shown to be important within DNA-promoter interactions [322]. These DNA attributes may therefore also have an affect on distal

regulatory regions and the "looping-out" mechanism by which a distal regulatory region can interact with the DTNBP1 promoter may be disrupted by variation at rs13198512.

The fact that rs13198512 is within an intron of one gene (JARID2) but appears to affect the expression of a neighbouring gene (DTNBP1) is not unprecedented. Adlam and Siu identified an enhancer within the gene ISOT (also known as USP5) that is critical for the expression of a neighbouring gene CD4 [288]. CD4 transcription is closely coupled with T-cell development. While the enhancer maps to the first intron of ISOT, which is located 3' to CD4, the enhancer does not appear to affect ISOT gene expression. Therefore, it is possible that rs13198512, a variant that maps to the first intron of JARID2, a gene 3' to DTNBP1, could have a regulatory effect on DTNBP1 but not JARID2. Alternatively, rs13198512 could also have a regulatory effect on both genes as other variants have been shown to have a regulatory effect on more than one gene at a locus [344].

5.4.2 Implications of Luciferase Assay Results

This chapter set out to determine whether rs13198512 and/or rs2619538 affect transcription *in vitro*. It was also hoped that this analysis would help establish whether the identification of putative regulatory SNPs and the correlation of these genotypes with allelic expression data is a viable way of identifying regulatory polymorphisms. Variation at rs13198512 has been shown to be associated with differential expression of dysbindin mRNA and to affect luciferase expression levels. Although these two findings cannot be directly linked, they both support the hypothesis that DNA variants associated with differential allelic expression are regulatory variants. In contrast, rs13198512 is not within a putative regulatory region identified by the methods described in chapter 3. It therefore appears that while initial screening of putative regulatory regions is advantageous in that it can substantially reduce the amount of screening and/or genotyping to be performed, it is likely to miss functional variants.

However these observations have been made from the analysis of just one polymorphism. Other SNPs, both positively and negatively correlated with allelic expression data, plus those located within and outside putative regulatory regions, need to be analysed using reporter gene assays before any definite conclusions can be drawn.

Chapter 6: An Examination of BLOC1S3 and MUTED as Schizophrenia Susceptibility Genes

6.1 Introduction

Although many studies have implicated DTNBP1 as a schizophrenia susceptibility gene [112, 113, 117, 121, 127, 129-138], the function of the dysbindin protein, specifically its role in schizophrenia pathology, remains largely unknown. Co-immunoprecipitation studies and yeast-two-hybrid analysis [208, 212] have shown that dysbindin is a member of the protein complex biogenesis of lysosome related organelles complex 1 (BLOC-1). It is therefore reasonable to postulate that DTNBP1 could confer a risk to schizophrenia through disruption of this complex and that other BLOC-1 genes could themselves contribute to schizophrenia susceptibility.

Morris and colleagues [367] investigated this hypothesis by performing association analyses on seven BLOC-1 genes (MUTED, PLDN, CNO, SNAPAP, BLOC1S1, BLOC1S2, and BLOC1S3). All exonic regions plus 1kb 5' to each of the genes were screened for polymorphisms and a non-redundant set of SNPs were genotyped through their Irish case control sample. A significant association was observed for rs12460985, a SNP 3' of the BLOC1S3 gene ($p=0.0028$). However this result would not survive Bonferroni correction for the number of SNPs analysed. No other single marker within the BLOC-1 genes showed a significant association to schizophrenia.

Members of the BLOC-1 complex could potentially contribute to schizophrenia susceptibility either independently or alternatively through interaction with other members of the complex. Under this epistatic model any given mutation needs to be considered in the context of other polymorphisms [368]. Therefore in addition to an independent association analysis, Morris and colleagues also performed a gene x gene interaction analyses for all combinations of the BLOC-1 genes including DTNBP1.

Although allelic and genotypic analysis did not show a significant association at the MUTED locus, gene based interaction analysis identified a significant interaction between SNPs at the DTNPB1 and MUTED loci (p=0.0009) [367]. Subsequent characterization of this gene x gene interaction using a standard logistic regression framework determined the strongest allele based signal to be between DTNBP1 rs2619539 (P1655) x MUTED rs10458217 (p=0.0094). A breakdown of the odds ratios (ORs) for each of the nine possible genotype combinations indicated that the majority of the significant interaction between these two SNPs could be accounted for by the genotype combination DTNBP1 GG x MUTED AA (Figure 6.1.).

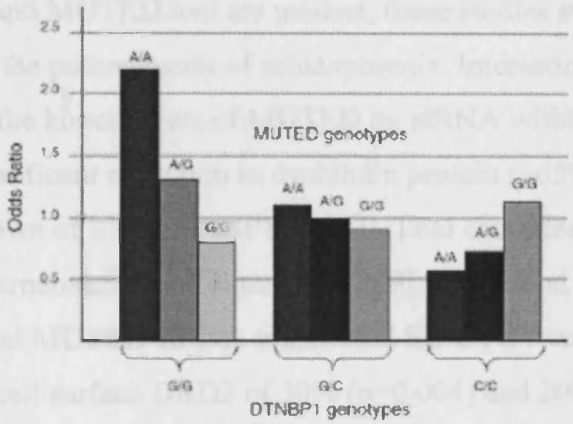


Figure 6.1. Breakdown of the odds ratios from the nine possible genotype combinations of the interacting SNPs DTNBP1 rs2619539 and MUTED rs10458217 reported by Morris *et al* [367]. The ORs are relative to the double heterozygote which has been fixed as the reference (OR=1). The majority of the significant interaction between these two SNPs can be accounted for by the genotype combination DTNBP1 GG x MUTED AA.

The fact that Morris and colleagues observed a significant interaction with DTNBP1, but not a main effect at the MUTED locus, highlights the possibility that interaction analysis may have an important role in finding susceptibility genes. Although inconsistent results across association studies can be caused by both genetic heterogeneity and phenocopy, it is also possible that the conflicting results observed for DTNBP1 and other complex disorder susceptibility genes may partly be explained by epistasis. In support of this, Moore and Williams hypothesised that the lack of replication for single locus results in studies of another complex disease, essential hypertension, may be because epistatic effects are more important than independent main effects [369].

In addition to the significant interaction between MUTED and DTNBP1 reported by Morris *et al* [367], significant evidence for association between markers at the MUTED locus and schizophrenia has been reported in abstract form. Straub and colleagues performed an association analysis on 17 MUTED SNPs. Of these, six were found to be significantly associated using the family base association test (FBAT) (rs10458217 p=0.0013, rs3734590 p=0.0003, rs3734591p=0.0002, rs11243223 p=0.0001, rs2815155 p=0.01, rs2743986 p=0.008) [207].

While the effect sizes for the significant association and interaction reported at the BLOC1S3 and MUTED loci are modest, these studies support the involvement of BLOC-1 in the pathogenesis of schizophrenia. Interestingly, cell culture studies have shown that the knock down of MUTED by siRNA within human neuroblastoma cells causes a significant reduction in dysbindin protein (~65%, p=0.002) [229]. In addition, the knockdown of both DTNBP1 and MUTED can affect dopamine D2 receptor (DRD2) internalisation and signalling [229]. Iizuka and colleagues demonstrated this in DTNBP1 and MUTED siRNA transfected SH-SYSY cell lines. These cells showed an increase in cell surface DRD2 of 30% (p=0.004) and 20% (p=0.027) respectively compared with cells transfected with a control siRNA. These results provide a possible functional link between two members of the BLOC-1 complex and a neurotransmitter system implicated in schizophrenia pathogenesis.

The aim of this chapter was to attempt to replicate the association reported by Morris and colleagues between schizophrenia and the BLOC1S3 variant rs12460985. In addition a detailed association analysis of the MUTED locus was performed in an attempt to replicate either the association reported by Straub and colleagues or Morris *et al.* This included genotyping a representative set of MUTED SNPs through our case control sample. In addition these markers were tested for evidence of interaction with 15 SNPs at the DTNBP1 locus that had been previously genotyped in the same sample [134]. This included the pair of markers reported to show evidence for interaction in the study of Morris *et al.*

6.2 Methods

6.2.1 Subjects

Mutation Screening was performed using a sample of 14 of unrelated schizophrenic subjects described in chapter 2.1.3. Individuals were chosen based on the criteria that they met the DSM-IV criteria for schizophrenia and each had at least 1 affected sibling. 14 unrelated individuals from the same population should allow 95% power to detect alleles with a minor allele frequency of >0.1 and 80% power to detect alleles with a minor allele frequency of 0.05.

Informative SNPs were genotyped through the UK case control sample consisting of 709 cases and 716 controls. The specific details of these individuals are described in chapter 2.1.1.

6.2.2 Identification of Putative Regulatory Regions

The genomic sequence of MUTED ±10kb (chr6:7949215-8019646) was analysed *in silico* to identify putative cis-acting regulatory regions. Three different strategies were used for identifying these regions; the protocol and the rationale for each is described in chapter 3. Briefly regions containing multiple transcription factor binding sites (TFBSs) were predicted using the web based programme Cluster Buster (http://zlab.bu.edu/cluster-buster/cbust.html) [315]. Evolutionary conserved sequence was identified using the ECR browser [317] (www.ECRbrowser.com) and the UCSC track "most conserved" [318]. Furthermore, as the sequence immediately 5' to the TSS of MUTED is highly likely to contain promoter elements, the genomic sequence 1kb 5' to exon 1 of the MUTED transcript AK02544 was also included for polymorphism detection.

6.2.3 Polymorphism Detection

The exonic structure of MUTED was ascertained according to AK02544 using the UCSC human genome reference sequence (March 2006 freeze). The genomic sequence of exonic and putative regulatory regions were subsequently derived and used to design PCR primers using Primer3 (http://www-genome.wi.mit.edu/cgi-bin/primer/primer3_www.cgi). Exonic and putative promoter PCR products were screened for sequence variation by Denaturing High Performance Liquid Chromatography (dHPLC) [245]. The protocol for which is described in chapter 2.5.1. Other putative regulatory regions were screened using HRMA (chapter 2.5.2).

6.2.4 Polymorphism Identification

PCR products from individuals showing chromatograms/melting profiles suggestive of heteroduplex formation were sequenced in both directions using Big-Dye terminator chemistry and the ABI3100 sequencer (Applied Biosystems, Foster City, USA), details of which are given in chapter 2.6.

6.2.5 SNP Selection

All DNA variants detected, as well as rs2743986, which had been reported to be associated to schizophrenia in previous studies [207, 367], were genotyped in the 30 CEPH parent-offspring trios that constitute the HapMap CEU sample and their genotypes combined with all SNPs (n=163) of HapMap phase II (Jan 06) that span the MUTED locus (chr6:7950869-8017990, March 2006 freeze). Aggressive tagging via the Tagger function of Haploview was then used to select a set of tag SNPs which captured all alleles at this locus with an MAF >0.05 at an r^2>0.8.

6.2.6 Genotyping

SNPs were genotyped using a Sequenom MassARRAY^{TM} system [370] as described in chapter 2.7.1.

6.2.7 Statistical Analysis

6.2.7.1 Association Analysis

Each marker was tested for allelic and genotypic association using χ^2 1df and 2df tests. Goodness of fit tests for Hardy-Weinberg equilibrium (HWE) were performed in the cases and controls separately. All association and permutation analysis was performed in PLINK [256] apart from genotypic permutation analysis which was performed using CLUMP [321]. Analysis of all combinations of 2 and 3 marker haplotypes were performed using UNPHASED [371].

6.2.7.2 Imputation Analysis

Imputation analysis allows the prediction of individual genotypes for a specific sample, in this case the schizophrenia case control sample, with reference to the genotypic and linkage disequilibrium data from the HapMap dataset. A model based imputation method operated using PLINK [256] was used to predict genotypes of SNPs at the MUTED locus (chr6:7950869-8017990) that were included in the HapMap phase II CEU data but were not genotyped in this study.

6.2.7.3 Interaction Analysis

Gene x gene interaction analysis was performed using logistic regression [67] with Gene Interaction Tool designed by Dr. Valentina Moskvina (http://x001.psycm.uwcm.ac.uk/GIT/GIT.html). Markers were coded in terms of additive and dominance components of the genotype, and then two models (main effects and main effects plus interaction terms) were fitted and compared by using the likelihood ratio test. For marker combinations where the full interactive model could not be calculated, either due to a high standard error for the coefficient estimates or the presence of singularities in the model, stepwise reduction of the number of interaction terms was then used to find the largest model. For each between-gene marker pairing an empirical estimate of the significance was made by permuting affection status and

determining whether and to what extent the model with interactive terms was significantly superior to that with main effects. Empirical p-values were then estimated by dividing the number of times the overall interaction effect p value was smaller in the simulated data set than in the specific marker-marker test by the number of simulated data sets (n=1000).

6.3 Results

6.3.1 BLOC1S3 Association Analysis

Genotyping rs12460985 through 709 cases and 719 controls produced a call rate of >95%. While HW equilibrium for both cases and controls showed $p>0.001$, the Armitage trend test was used to test for allelic association to compensate for even small deviations in HW. However, analysis of rs12460985 revealed no evidence for significant allelic or genotypic association (Table 6.1).

rsID	Chromosomal Position	Alleles	Frequency Cases	Frequency Controls	Arm Trend P-Value	Counts Cases	Counts Controls	HW Case and Controls	HW Cases	HW Controls	Genotypic P-Value
rs12460985	Chr19: 50382729	C/T	0.17 (C)	0.18	0.475	18/168/419	17/209/452	0.445	0.816	0.213	0.452

Table 6.1. Case/Control association analysis of rs1260985 at the BLOC1S3 locus. SNP position is according to UCSC human genome chromosome 6 reference sequence (March 2006 freeze). Minor allele frequencies (MAF) are shown for the specific allele in parentheses.

6.3.2 MUTED Putative Regulatory Region Identification

As no one comprehensive *in silico* method has been determined to detect cis-acting regulatory sequence, the identification of putative regulatory regions at the MUTED locus was performed using a complimentary set of analyses. This consisted of identifying putative promoter region(s) plus determining regions containing TFBSs and evolutionary conserved sequence. Analysis of chr6:7949215-8019646 identified 7 MUTED putative regulatory regions (Figure 6.3). Exact chromosomal positions and PCR primers for mutation screening of these putative regulatory regions and all exonic sequence determined are given in Appendix Table 9.3.1. One "core promoter" region for all MUTED transcripts was identified. Of the remaining 6 regions, 5 contained highly conserved sequence which was determined using either ECR browser (n=1), USCS "most conserved" track (n=2) or both (n=2). Analysis using the web based program Cluster Buster predicted one region to contain multiple transcription factor binding sites.

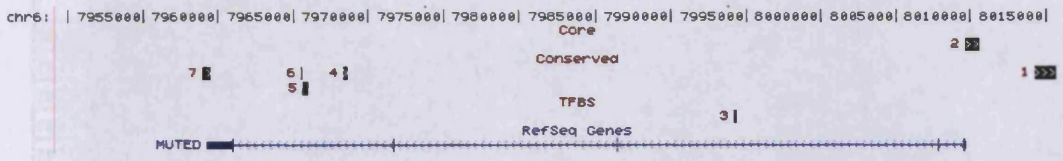


Figure 6.3. Analysis of the MUTED locus by a combination of methods identified 7 putative regulatory regions. The putative promoter sequence immediately 5' to the transcription start site of the MUTED isoforms is illustrated under the track named "Core". Evolutionary conserved regions are given within the "Conserved" track. Regions identified by Cluster Buster as containing multiple TFBSs are shown under the track name "TFBS".

6.3.3 MUTED Polymorphism Identification

PCR optimisation was achieved for all exonic and putative regulatory regions except the regulatory region 3 (chr6: 7994178-7994348). Attempted genotyping of three dbSNPs within this region (rs9379158, rs9379159, rs9406083) also failed. Therefore polymorphisms within this region have not been included in the subsequent analysis. Analysis of this region in UCSC identified the sequence to be highly repetitive with a 2 ALU elements plus a short repetitive sequence covering 91% of the region which may explain the problems encountered. dHPLC parameters for MUTED putative promoter and exonic regions are given in Appendix Table 9.3.2.

Screening all MUTED exons plus the remaining putative regulatory sequence identified 17 sequence variants (Table 6.2), three of which were upstream of the MUTED gene and seven were detected within the 3'UTR. Seven SNPs were intronic, four of which were located within conserved non-coding sequence (CNC). Four variants had not been previously deposited in dbSNP (Muted_Ex5_snp1, Muted_Reg6_snp1, Muted_Reg2_snp1 and Muted_Reg2_snp2).

Snp ID	Chromosomal Position	Alleles	Location	MAF
Muted_Ex5_snp1	7959444	C/G	3' UTR	0.011 (C)
rs2748375	7959703	G/T	3' UTR	Not typed
rs2815151	7959989	T/C	3' UTR	0.288 (T)
rs2748376	7960128	A/G	3' UTR	0.292 (A)
rs10458217	7960274	T/C	3' UTR	0.175 (T)
rs3734590	7960719	G/A	3' UTR	0.175 (G)
rs3734591	7960820	G/A	3' UTR	0.175 (G)
rs2057185	7961088	C/T	Intron 4	0.117 (C)
rs13202814	7961156	G/T	Intron 4	0.195 (G)
rs10458106	7965466	T/C	Intron 4 CNC	0.175 (T)
Muted_Reg6_snp1	7965521	A/G	Intron 4 CNC	0 (A)
rs2815128	7968461	G/T	Intron 4 CNC	0.314 (G)
rs13191023	7968556	A/G	Intron 4 CNC	0.175 (A)
rs9328452	7971576	A/G	Intron 4	0.305 (A)
Muted_Reg2_snp1	8009676	A/G	Upstream Core Promoter	Not typed
Muted_Reg2_snp2	8010014	A/G	Upstream Core Promoter	0.058 (A)
rs2815155	8010229	A/G	Upstream Core Promoter	0.458 (A)

Table 6.2. Sequence variants identified at the MUTED locus.
SNP positions are according to UCSC human genome chromosome 6 reference sequence (March 2006 freeze). Minor allele frequencies (MAF) in the HapMap CEU sample are shown for the specific allele in parenthesis. Muted_Ex5_snp1, Muted_Reg6_snp1, Muted_Reg2_snp1 and Muted_Reg2_snp2 are novel SNPs that were identified during this study.

6.3.4 MUTED Association Analysis

Aggressive tagging via the Tagger function of Haploview selected a set of 19 tag SNPs which captured 94% alleles at the MUTED locus with an MAF >0.05 at an r^2>0.8. Of the 19 tag SNPs genotyped in the UK case control sample, all had a call rate of >95%. Individuals were included in analysis if they had a genotype call rate >80%. This meant 46 individuals were dropped from the original sample.

All markers were in HW equilibrium at p>0.001 for both cases and controls. However to compensate for even small deviations in HW, the Armitage trend test was used to test for allelic association.

While an Armitage trend test of rs1002308 revealed no evidence for an allelic association (p=0.269), the test for genotypic association was nominally significant (p=0.022). However this would not survive Bonferroni correction for multiple testing (threshold p=0.002). Individual genotyping of the remaining 18 SNPs in the association sample of 709 schizophrenia cases and 716 controls revealed no evidence for significant allelic or genotypic association at any marker (Table 6.3). In reference to the previous association study by Straub *et al* [207], association analysis failed to replicate the allelic association previously observed at rs104758217 (p=0.75), rs2815145(r^2=1 with rs2815155, p=0.97) and rs2743986 (p=0.91).

As tag SNP selection was based on aggressive tagging, 2 and 3 marker haplotype analysis is also important. However this also failed to identify any 2 or 3 marker haplotypes that were globally associated with schizophrenia (p_{min}=0.08).

No	rsID	Chromosomal Position	Alleles	Frequency Cases	Frequency Controls	Arm Trend P (Emp P)	Counts Cases	Counts Controls	HW Case and Controls	HW Cases	HW Controls	Genotypic P (Emp P)
1	rs1002308	7955696	A/C	0.486 (A)	0.508	0.269 (0.268)	172/292/190	175/362/164	0.211	0.006	0.4063	0.020 (0.020)
2	Mutedex5_5snp1	7959444	G/C	0.014 (G)	0.017	0.486 (0.486)	1/16/638	0/24/678	0.275	0.112	1.000	NA (0.335)
3	rs2815151	7959989	T/C	0.343 (T)	0.351	0.669 (0.670)	91/262/294	85/323/294	0.118	0.011	0.8680	.115 (0.115)
4	rs10458217	7960274	T/C	0.166 (T)	0.162	0.756 (0.756)	23/173/462	19/190/495	0.279	0.204	0.890	0.683 (0.683)
5	rs2057185	7961088	C/T	0.180 (C)	0.190	0.502 (0.496)	30/174/447	30/205/462	0.015	0.023	0.267	0.545 (0.545)
6	rs13202814	7961156	G/T	0.139 (G)	0.154	0.276 (0.277)	9/164/481	16/183/500	0.447	0.325	1.000	0.392 (0.392)
7	rs2207720	7964196	C/T	0.492 (C)	0.473	0.353 (0.353)	156/296/166	157/324/193	0.164	0.296	0.354	0.646 (0.646)
8	rs10458106	7965466	T/C	0.167 (T)	0.162	0.702 (0.712)	22/174/456	18/190/491	0.488	0.323	1.000	0.683 (0.683)
9	rs2815128	7968461	G/T	0.366 (G)	0.373	0.702 (0.697)	89/298/264	106/307/283	0.198	0.736	0.144	0.679 (0.679)
10	rs13191023	7968556	A/G	0.136 (A)	0.156	0.138 (0.135)	9/157/479	16/185/496	0.511	0.403	0.8860	.281 (0.281)
11	rs9328452	7971576	T/C	0.365 (T)	0.369	0.837 (0.836)	88/300/264	102/311/285	0.349	0.866	0.256	0.792 (0.792)
12	rs9502675	7971924	T/C	0.192 (T)	0.186	0.719 (0.723)	26/199/430	24/213/464	0.790	0.614	1.000	0.866 (0.866)
13	rs13218209	7983148	A/G	0.306 (A)	0.317	0.530 (0.538)	63/273/317	63/316/319	0.568	0.712	0.255	0.439 (0.439)
14	rs2206098	7991601	G/A	0.149 (G)	0.159	0.477 (0.475)	16/163/477	19/184/497	0.531	0.643	0.671	0.774 (0.774)
15	rs2815145	7993302	T/C	0.489 (T)	0.489	0.973 (0.991)	156/325/170	175/335/191	0.414	1.000	0.257	0.735 (0.735)
16	rs2743986	8011954	T/C	0.501 (T)	0.499	0.910 (0.919)	162/329/161	179/334/181	0.586	0.876	0.324	0.689 (0.689)
17	rs2326975	8014321	C/T	0.056 (C)	0.052	0.687 (0.709)	2/69/586	2/69/630	1.000	1.000	0.711	NA (0.920)
18	rs9392958	8016561	T/C	0.338 (T)	0.342	0.802 (0.800)	69/299/279	75/325/294	0.201	0.430	0.312	0.961 (0.961)
19	rs9378519	8016671	A/T	0.258 (T)	0.266	0.598 (0.617)	41/256/359	46/281/373	0.401	0.683	0.562	0.864 (0.864)

Table 6.3. Case/Control association analysis of 19 tag SNPs spanning the MUTED locus. SNP positions are according to UCSC human genome chromosome 6 reference sequence (March 2006 freeze). Minor allele frequencies (MAF) are shown for the specific allele in parentheses. Allelic association results are shown with permutation (1000) p values in parentheses.

6.3.5 MUTED Imputation Analysis

Imputation analysis relies on genotypic data from the HapMap CEU sample dataset. Genotypic data for two of the 19 tag SNPs genotyped in this study were not available on the HapMap CEU sample database (Muted_ex5snp1 and rs2743986). These two SNPs were therefore not informative for imputation analysis. Imputation analysis of the remaining 17 tag SNPs allowed the genotypes of an additional 72 SNPs at the MUTED locus (chr6:7950869-8017990) to be predicted in our cases and controls with >80% probability (data not presented). Allelic and genotypic association analysis revealed that of these 72 SNPs, none were significantly associated with schizophrenia (imputed allelic p_{min}=0.100, imputed genotypic p_{min}=0.054).

6.3.6 MUTED and DTNBP1 Interaction Analysis

In an attempt to directly replicate the interaction reported by Morris *et al* [367] logistic regression was performed using the same markers reported to show a significant interaction (rs10458217 (MUTED) and rs2619539 (DTNBP1)). However this failed to reveal any evidence for a significant locus/locus interaction (p=0.273, Appendix Table 9.3.3A), nor at the specific allelic model that was reported (additive model p=0.078, Appendix Table 9.3.3B).

Subsequently all possible MUTED and DTNBP1 marker interactions were tested, using both additive and additive plus dominant models, with the 19 SNPs genotyped at the MUTED locus and 15 SNPs previously genotyped in the same sample at the DTNBP1 locus [134]. This identified nominally significant evidence for interaction at 26 different marker combinations using the full interaction model, five of which were significant at p<0.01 (rs2815151-SNPN p=0.0007, rs9328452-rs12204704 p=0.002, rs13218209-rs12204704 p=0.006, rs2206098-rs12204704 p=0.007, rs2815128-rs12204704 p=0.002) (Appendix Table 9.3.3A). Six marker combinations were significant using the additive model (0.01<p<0.05) and one significant marker combination at p<0.01 (rs2815151-SNPN, p=0.0001, Appendix Table 9.3.3B). Under

the full model no marker combinations were significant after multiple testing, however one significant result under the additive model was still associated after 1000 permutations (rs2815151-SNPN, p=0.026).

6.3.6.1 rs2815151 and SNPN Significant Interaction

rs2815151 (MAF=0.29) is located in the 3'UTR of the MUTED gene (chr6:7959989). SNPN is a rare intronic polymorphism (MAF=0.04) located near the 3' end of DTNBP1 (chr6:15658414). Further analysis of this SNP interaction (Table 6.4) indicates that the majority of the signal may be explained by genotypic combinations of very low frequency. The genotypic combination of MUTED rs2815151 (CC) x DTNBP1 SNPN (DEL/DEL) is the only combination which shows a difference between cases and controls where genotype frequencies are over 5% (14.5% in the cases and 10.5% in the controls, case/control ratio = 1.41). The remaining genotypic combination ratios whose case/control ratio significantly differs from 1 (TT:G/DEL ratio =1.63, CT:G/DEL ratio = 0.361, CC:G/DEL ratio=0.298) have frequencies in both the cases and the controls of <0.05.

	Case	DTNBP1_SNPN		
rs2815151		GG	G/DEL	DEL/DEL
	TT	0.00391 (2)	0.043 (22)	0.395 (202)
	CT	0 (0)	0.0156 (8)	0.395 (202)
	CC	0 (0)	0.0040 (2)	0.145 (74)
	Control	DTNBP1_SNPN		
rs2815151		GG	G/DEL	DEL/DEL
	TT	0 (0)	0.0263 (14)	0.386 (206)
	CT	0.00563 (3)	0.0432 (23)	0.422 (225)
	CC	0 (0)	0.0131 (7)	0.103 (55)
	Ratio Case/Control	DTNBP1_SNPN		
rs2815151		GG	G/DEL	DEL/DEL
	TT	INF	1.63	1.02
	CT	0	0.361	0.936
	CC	NA	0.298	1.41

Table 6.4. Breakdown of significant interaction between Muted SNP rs2815151 and DTNBP1 SNPN under an additive model. The majority of the interaction comes from very low frequency genotypes with the only common signal detected with CC:DEL/DEL which is 14.5% in cases and 10.3% in controls (ratio 1.41). Genotype counts are given in brackets.

6.4 Discussion

A large body of genetic association data implicates DTNBP1 as a schizophrenia susceptibility gene, but its exact role in the pathogenesis of schizophrenia is currently unknown. Evidence supporting the involvement of DTNBP1 through its role in the BLOC-1 complex has come from genetic association studies of genes that code for proteins within the complex [207, 367, 372], but that evidence is not yet compelling. Significant evidence for an association between MUTED and schizophrenia has been reported [207, 372] and while a study by Morris and colleagues failed to identify any evidence for a main effect at MUTED, they did report evidence for interaction between MUTED and DTNBP1. Morris and colleagues also reported a significant association between a SNP at the BLOC1S3 locus and schizophrenia [367].

This chapter set out to analyse the significant association observed by Morris and colleagues for rs12460985, as SNP 3' to BLOC1S3 in addition to performing a detailed association analysis between common variation at the MUTED locus and schizophrenia. Analysis of rs12460985 within 709 cases and 719 controls revealed no evidence for association (p=0.475). Therefore we failed to replicate the findings of Morris *et al.*

Mutation screening of MUTED exons and putative regulatory sequence identified 17 SNPs which were combined with 163 SNPs which spanned the MUTED locus (chr6:7950869-8017990) and had been genotyped in HapMap phase II. This defined a set of highly informative SNPs which captured the majority of common variation at this locus were then genotyped in a relatively ethnically homogeneous case-control sample. Analysis of allelic, genotypic plus 2 and 3 marker haplotype distributions in cases and controls failed to reveal evidence for association with schizophrenia after correction for multiple testing, as did imputation analysis of another 72 HapMap SNPs at the MUTED locus.

In order to test for evidence of a significant interaction between MUTED and DTNBP1 logistic regression analysis was performed between DTNBP1 and MUTED markers. The specific interaction analysis reported associated by Morris and colleagues [367] was not significant in our sample. Interaction analyses of all the other marker combinations identified one marker combination under the additive model which was still significant after multiple testing (rs2815151-SNPN p=0.026). However SNPN has a MAF=0.04 and the majority of the significant interaction appears to be accounted for by very low genotype counts (MAF<0.05).

In summary the data presented in this chapter provides no support for the hypothesis that either BLOC1S3 or MUTED are independent susceptibility genes for schizophrenia. However while we did not replicate the interaction found by Morris and colleagues [367], a weakly significant interaction was observed between DTNBP1 and MUTED with two other markers.

The presence of allelic heterogeneity at the DTNBP1 locus could dramatically influence the power of our study to replicate the locus-locus interaction analysis of Morris *et al* [367]. However given that the UK case-control sample used in this chapter is of similar ethnicity as the Irish sample used by Morris *et al* [367] and that the Irish sample has previously shown a significant association at the DTNBP1 locus with the same risk haplotype as the UK sample [134] this is not a particularly likely explanation.

Another potential explanation for the differing results is the fact that epistasis is difficult to detect and characterise using traditional parametric statistical methods. This includes logistic regression which has been used both in this chapter and in the Morris study [367]. This difficulty stems from the fact that when interaction between multiple polymorphisms, or polymorphisms of low frequency are considered, genotype combinations are produced that have very few or no data points. Logistic regression can also lead to either an increase in type I errors (i.e. false positives) due to the large standard errors, or an increase in type II errors (i.e. false negatives) and a decrease in power from attempting to fit the data to a regression model.

While our study provides no definitive support for BLOC1S3 or MUTED as schizophrenia susceptibility genes, and thereby no support for the hypothesis that DTNBP1 contributes to susceptibility to schizophrenia through the BLOC-1 complex, the latter hypothesis cannot be rejected. Firstly, we have not tested every SNP in every BLOC-1 related gene. Our analysis of the BLOC1S3 gene only included the analysis of rs12460985, a SNP reported as significantly associated with schizophrenia by Morris *et al* (p= 0.0028) [367]. Secondly, our sample, while relatively large in schizophrenia research and powered to identify association with common risk alleles at the BLOC1S3 or MUTED locus (power >0.8 to detect association with a relative risk of 1.4 to alleles with a population frequency of 0.1 at p=0.05) it is underpowered to detect small effects (OR<1.4). Finally, while the hypothesis that DTNBP1 contributes to susceptibility to schizophrenia through the BLOC-1 complex would certainly be strengthened by clear demonstration of the involvement of another gene in that complex, it is not a requirement of the hypothesis that a susceptibility variant exists in any other member.

Chapter 7: Allelic Expression Analysis of BLOC-1 Genes

7.1 Introduction

As discussed in chapter 6, Morris and colleagues [373] performed an association analysis on all known BLOC-1 genes other than dysbindin (CNO, MUTED, PLDN, SNAPAP, BLOC1S1, BLOC1S2 and BLOC1S3). Though nominal significance was observed for MUTED and BLOC1S3, no association was reported between schizophrenia and the five remaining genes. This could suggest that either the Morris study was underpowered to detect association in the other BLOC-1 genes or that only certain genes within the complex are involved in schizophrenia pathology. It is also possible that the actual causal variants were not subjected to association analysis or in sufficient LD with the SNPs that were genotyped. Apart from Morris *et al* who analysed polymorphisms within the sequence immediately 5' to each gene, no previous studies of BLOC-1 genes [207] have screened for DNA variants that have the potential to regulate expression. As discussed in previous chapters it is possible that dysbindin confers susceptibility to schizophrenia through altered expression. It is therefore also a valid hypothesis that alteration of expression of other BLOC-1 genes could confer susceptibility to schizophrenia. If this were true then risk variants may be located within cis-acting regulatory elements such as transcription factor binding sites.

Consequently allelic expression was performed to determine whether, in addition to dysbindin, any of the other known BLOC-1 genes are under the influence of cis-acting variation. Any genes showing differential expression were analysed for putative regulatory regions using the methods described in chapter 3. These regions plus all exonic sequence were screened for DNA variants which were subsequently subjected to association analysis with schizophrenia.

7.2 Methods

7.2.1 Subjects

7.2.1.1 Caucasian Brain Sample

Allelic expression analysis was performed using the Caucasian Brain sample described in chapter 2.1.2. The initial analysis utilised a sample of 60 individuals obtained from the Stanley Medical Research Institute Brain Bank, Bethesda, USA. The remaining 88 samples (which were assayed when a difference in allelic expression was detected in the first sample set) were obtained from three sources, the MRC London Neurodegenerative Diseases Brain Bank, London, UK; the Department of Clinical Neuroscience, Karolinska Institute, Sweden and the Mount Sinai School of Medicine, New York, USA.

7.2.1.2 UK Schizophrenia Case Control Association Sample

Informative SNPs for genes showing allelic expression differences were genotyped through the UK case control sample consisting of 709 cases and 716 controls. The specific details of these individuals are described in chapter 2.1.1.

7.2.2 BLOC-1 Gene Identification

After an initial literature search of the term "BLOC-1 complex", members of the BLOC-1 complex were sourced from a number a publications (CNO [209], PLDN and MUTED [210], SNAPIN, BLOC1S1, BLOC1S2 and BLOC1S3 [211]).

7.2.3 Relative Allelic Expression Assay

Allelic expression analysis was performed as described in chapter 2.9. The allelic expression analysis of CNO was performed by Dr. Nicholas Bray.

All ratios (corrected genomic or cDNA) were normalised via a natural log transformation. After conformation by the Kolmogorov-Smirnov test that the transformed gDNA and cDNA data followed a normal distribution, gDNA and cDNA values were compared by a 2 independent sample t-test. To account for the possibility of weak LD between the assayed SNP and a potential regulatory SNP and/or multiple cis-acting variants the spread of the cDNA ratios and gDNA ratios were compared using the Levene's test.

Evidence for differential allelic expression was deduced if an assay met the following criteria. Firstly, one or more samples had to show a differential expression of greater than 20% (0.8>ratio>1.2) after correction using the average genomic ratio. Secondly, a statistically significant difference ($p<0.05$) needed to be observed with either 2 independent samples t-test or Levene's test. More details of these criterion are given in chapter 2.9.1.

7.2.4 Power Detection

The calculation performed to determine the power of a sample to detect the effects of unknown regulatory variants is based on the binomial distribution. It assumes Hardy-Weinberg equilibrium at the regulatory SNP and no LD with the marker SNP [194]. The probability of an individual being homozygous at a putative regulatory locus with alleles in Hardy-Weinberg equilibrium is $p^2 + q^2$, where p and q are the two allele frequencies. Therefore the probability that, of n individuals assayed, all are homozygous for a regulatory polymorphism (meaning the regulatory variant will go undetected by the assay) is $(p^2 + q^2)^n$. The power to detect at least one heterozygote at an unknown regulatory locus is therefore $1 - (p^2 + q^2)^n$. If the marker and regulatory SNP are in LD, then a higher proportion of people selected for heterozygosity at the allelic expression marker will also be heterozygous for the regulatory SNP and the power will be increased.

7.2.5 Association Analysis

Any BLOC-1 genes showing common allelic expression differences (multiple individuals with ratio greater than 20% plus a significant independent sample t-test or Levene's test) were subjected to association analysis with schizophrenia. Putative regulatory regions were identified as described in chapter 3.2.1. These regions plus exonic sequence were then screened for polymorphisms using high resolution DNA melting analysis technology and Sanger sequencing as described in chapters 2.5.2 and 2.6 respectively.

All detected DNA variants were genotyped in the 30 CEPH parent-offspring trios that constitute the HapMap CEU sample. SNPs were then combined with all SNPs of the HapMap phase II (Nov 08) that spans the assayed gene (largest transcript ± 1kb). Pairwise tagging via the Tagger function of Haploview was then used to select a set of SNPs which captured all the alleles at the specified locus, plus any within distal putative regulatory regions identified, with a MAF >0.001 and at an r^2>0.95. Tag SNPs were genotyped by Sequenom Mass Array (see chapter 2.7.1). If genotyping failed using Sequenom Mass Array, genotyping was attempted using Amplifluor (see chapter 2.7.3).

7.2.5.1 Statistical Analysis for Association

Association analysis and Hardy-Weinberg equilibrium (HWE) tests were performed using plink software [256]. Each marker was tested for allelic association using the Armitage trend test. Genotypic association was performed using χ^2 2df tests. For SNPs whose genotype counts are less than 5 genotypic association analysis was performed using CLUMP with 10000 permutations [321]. Goodness of fit tests for HWE were performed in the cases and controls separately.

7.3 Results

7.3.1 SNP Selection

In order to determine whether any BLOC-1 genes are under the influence of cis-acting variation, polymorphisms were identified for each BLOC-1 gene that a) were located within one or more predicted mRNA transcripts, b) were able to be assayed by primers that would amplify both genomic and cDNA and c) had frequency information for the CEU population. For each BLOC-1 gene the SNP with the highest minor allele frequency was selected for allelic expression analysis so as to maximise the number of heterozygotes studied. In order to have power to infer whether a gene is under common, relatively rare or no cis-acting variation, a minimum MAF threshold of 0.05 was specified for the assay SNP. It was estimated that, for the sample size used (n=60), this should provide at least 5 heterozygotes for analysis.

The rs#, chromosomal position and minor allele frequency of each SNP selected for allelic expression analysis of PLDN, SNAPAP, CNO, MUTED and BLOC1S3 are given in Table 7.1.

Gene	SNP ID	SNP	Position	Frequency in CEU
PLDN	rs7181436	C>T	chr15:43687710	0.183
MUTED	rs3734590	T>C	chr6:7960719	0.175
CNO	rs3172604	T>G	chr4:6769676	0.392
SNAPAP	rs7345	C>A	chr1:151900682	0.467
BLOC1S3	rs758506	C>T	chr19:50374664	0.075

Table 7.1. The rs#, chromosomal position and MAF of each SNP selected for allelic expression analysis of PLDN, SNAPAP, CNO, MUTED and BLOC1S3. SNP positions are according to UCSC human genome chromosome 6 reference sequence (March 2006 freeze).

BLOC1S1 and BLOC1S2 could not be examined for differential expression. as while the exonic or untranslated regions of BLOC1S1 and BLOC1S2 contain 3 and 11 dbSNPs respectively, each of these SNPs either had no frequency information or a MAF<0.05 in the CEU population.

According to dbSNP BLOC1S3 contains 7 exonic SNPs. As with BLOC1S1 and BLOC1S2 none of these reached the MAF threshold of 5%. However as BLOC1S3 had previously been identified as associated with schizophrenia [373], the coding SNP rs758506 (which previously has no frequency information) was genotyped through HapMap CEU sample in order to assess its potential for allelic expression analysis. In this population the MAF of rs758506 was 0.075. Consequently BLOC1S3 was subjected the allelic expression analysis with this SNP.

7.3.2 Allelic Expression Analysis

In total five BLOC-1 genes, PLDN, SNAPAP, BLOS1S3, MUTED and CNO were subjected to allelic expression analysis. An outline of the results can be found in Table 7.2. A more detailed summary of each gene is given below.

Gene	SNP ID	Individuals genotyped	# of hets assayed	# of hets SD<0.2	SD/mean	cDNA average	Individuals >20%	T-test P	Levens P
PLDN	rs7181436	60	15	10	0.033	0.964	0	N/A	N/A
MUTED	rs3734590	148	40* in duplicate	36	0.081	1.010	4	0.872	0.012
CNO	rs3172604	60	20 in duplicate	18	0.045	1.227	11	2.423E-07	0.020
SNAPAP	rs7345	60	29	21	0.082	1.010	0	N/A	N/A
BLOS3	rs758506	60	7	6	0.050	0.941	0	N/A	N/A

Table 7.2. Summary of results allelic expression results for five BLOC-1 genes.
* Stanley and Extra samples genotyped. All other genes were analysed with just the Stanley sample.

7.3.2.1 PLDN

The SNP rs7181436 was selected for allelic expression analysis of PLDN (see Figure 6.1 for chromosomal location). Genotyping of SNP rs7181436 in the gDNA of 60 samples identified 15 heterozygotes. Allelic expression analysis was performed on cDNA generated from cerebral cortex post-mortem tissue of these 15 heterozygote individuals. Five samples showed a standard deviation >0.2 between duplicate RT reactions and so were removed from further analysis. The mean standard deviation for the remaining 10 samples was 0.033. Figure 7.2 shows the corrected genomic DNA ratios and the corrected cDNA averaged ratios (common allele/rare allele, C/T) for each sample.

No individuals showed differential expression of greater than 20% (1.2>C/T ratio>0.8), the predefined threshold to determine individuals showing allelic expression differences. No significant difference in variance between genomic and cDNA ratios was observed (Levene's test p=0.822) though a nominal difference was observed between the mean of the corrected allele ratios for the gDNA versus cDNA (independent t-test p=0.031). However this significant difference was observed between C/T ratios of 1 for gDNA and 0.96 for cDNA. Therefore it can be concluded that PLDN does not show any evidence that it is under the influence of cis-acting variation.

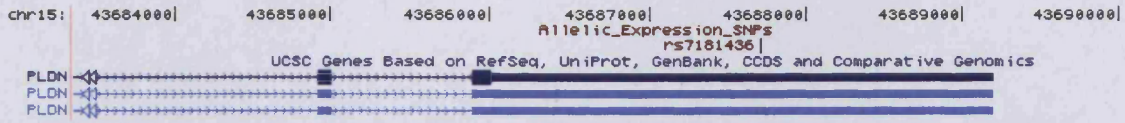


Figure 7.1. Chromosomal position of rs7181436. Allelic expression analysis of this SNP captures variation in expression of all predicted mRNA transcripts.

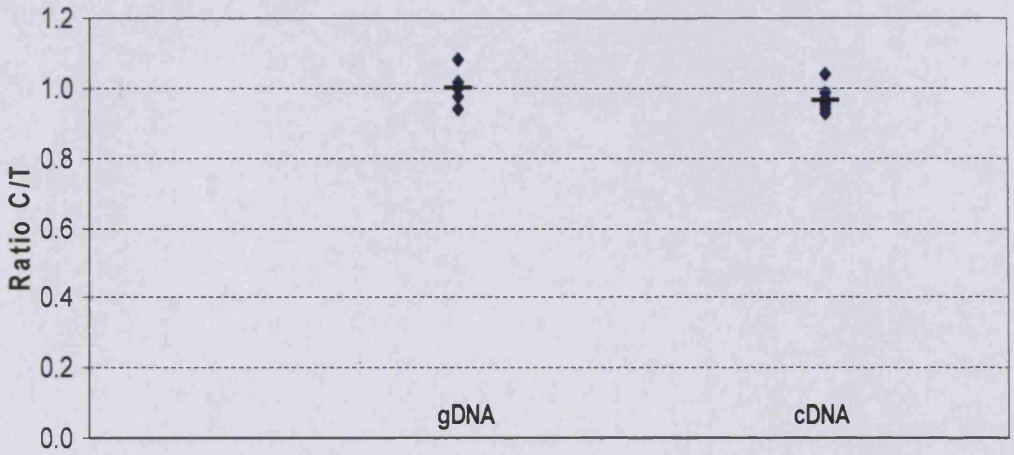


Figure 7.2. Relative allelic expression assay of rs7181436, a SNP within the 3'UTR of the PLDN gene. Both gDNA and cDNA were analysed in 10 heterozygote individuals using cDNA gained from the cerebral cortex. gDNA and cDNA allelic ratios (C/T) have been corrected by the mean gDNA ratio for all samples. Independent T-test p=0.031, Levene's test =0.822.

7.3.2.2 SNAPAP

The SNP rs7345 was selected for allelic expression analysis of SNAPAP (see Figure 7.3 for chromosomal location). Genotyping of this SNP in 60 individuals identified 29 heterozygotes. After allelic expression analysis was performed on these 29 samples, 8 samples showed a standard deviation >0.2 between duplicate RT reactions and so were removed from further analysis. The mean standard deviation between cDNA samples for the remaining 21 individuals was 0.082. Figure 7.4 shows the corrected genomic DNA ratios and the corrected cDNA averaged ratios (common allele/rare allele, C/A) for each sample.

No significant difference was observed between the mean of the corrected allele ratios for the gDNA versus cDNA (independent t-test p=0.786) though a nominal difference of the variance between genomic and cDNA ratios was observed (Levene's test p=0.017). However no individuals showed differential expression of greater than 20%. This could suggest that rs7345 is in relatively low LD with a weak regulatory SNP. Nevertheless, using the previously set threshold of 20% as an indication of differential expression, this assay does not provide any evidence that SNAPAP is under the influence of cis-acting variation.

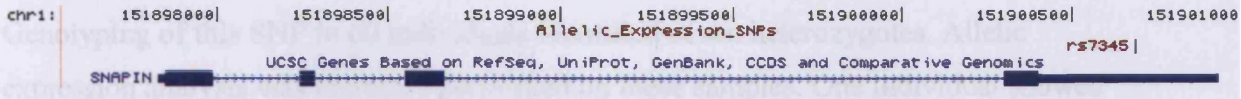


Figure 7.3. Chromosomal position of rs7345. Allelic expression analysis of this SNP will capture variation of all SNAPAP predicted mRNA transcripts.

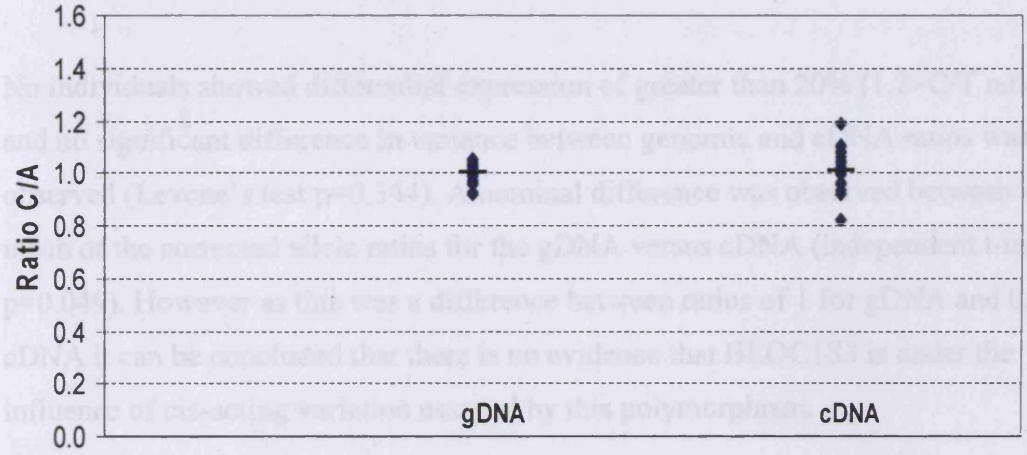


Figure 7.4. Relative allelic expression assay of rs7345, an exonic SNP within the SNAPAP gene. Both gDNA and cDNA were analysed in 21 heterozygote individuals. gDNA and cDNA allelic ratios (C/A) are corrected by the mean gDNA ratio for all samples. Independent T-test p=0.786, Levene's test =0.017.

7.3.2.3 BLOC1S3

rs758506 was selected for allelic expression analysis of BLOC1S3 (see Figure 7.5). Genotyping of this SNP in 60 individuals identified seven heterozygotes. Allelic expression analysis was therefore performed on these samples. One individual showed a standard deviation between duplicate RT reactions of >0.2 and so was removed from further analysis. The mean standard deviation for the remaining 6 samples was 0.050. Figure 7.6 shows the corrected genomic DNA ratios and the corrected cDNA averaged ratios (common allele/rare allele, C/T) for each sample.

No individuals showed differential expression of greater than 20% (1.2>C/T ratio>0.8) and no significant difference in variance between genomic and cDNA ratios was observed (Levene's test p=0.344). A nominal difference was observed between the mean of the corrected allele ratios for the gDNA versus cDNA (independent t-test p=0.049). However as this was a difference between ratios of 1 for gDNA and 0.94 for cDNA it can be concluded that there is no evidence that BLOC1S3 is under the influence of cis-acting variation assayed by this polymorphism.

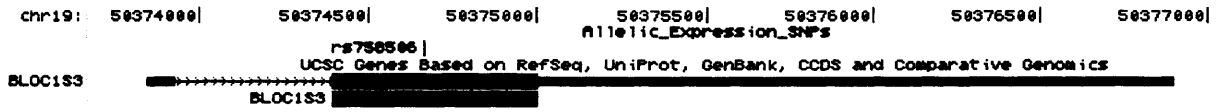


Figure 7.5. Chromosomal position of rs758506. Allelic expression analysis of this coding SNP captures variation at all BLOC1S3 predicted mRNA transcripts.

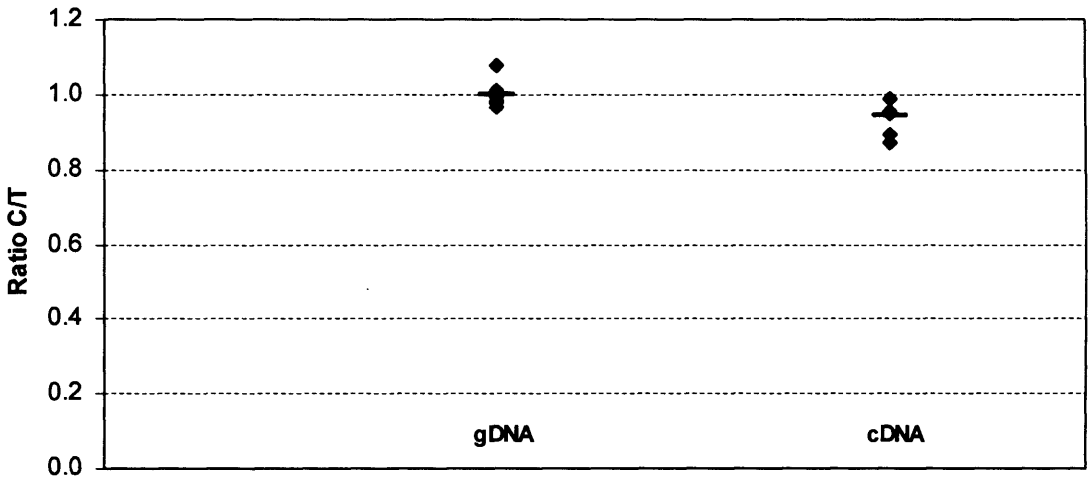


Figure 7.6. Relative allelic expression assay of rs758506, a coding SNP within the BLOC1S3 gene. Both gDNA and cDNA were analysed in 6 heterozygote individuals. Corrected gDNA and cDNA allelic ratios (C/T) are shown. Independent T-test p=0.049, Levene's test =0.344.

7.3.2.4 MUTED

Genotyping rs3734590, an exonic SNP within the MUTED gene (Figure 7.7) in 60 individuals identified 17 heterozygotes. Allelic expression analysis was performed on these 17 heterozygote samples. Two samples showed a standard deviation between duplicate RT reactions of >0.2 and so were removed from further analysis. The mean standard deviation for the remaining 15 samples was 0.075. Two of these individuals showed allelic expression differences of >20% (T/C = 0.66 and 0.75 respectively). As allelic expression differences had been identified in only 2/15 individuals, rs3734590 was genotyped in an additional 88 samples. This identified a further 23 heterozygotes which were also analysed for allelic expression differences. In total 40 individuals were analysed. Twelve of these showed a standard deviation between duplicate RT reactions of >0.2 and so were removed from further analysis. The mean standard deviation for the remaining 28 samples was 0.073. Four of these individuals showed allelic expression differences of >20%.

In order to ensure that these differences were due to cis-acting variation rather than, for example, experimental abnormalities, the assay was repeated for all heterozygotes. In total all 40 heterozygote cDNA samples were assayed four times and the corresponding genomic DNA twice. If a sample showed a standard deviation between duplicate RT reactions of >0.2 in one of the two assays, the allelic expression ratio was calculated from the one assay with the acceptable SD. After replication four samples showed a standard deviation between duplicate RT reactions of >0.2 in both assays and so were removed from further analysis. The mean standard deviation for the remaining 36 samples was 0.081. The average corrected genomic DNA ratios and the average of either four or two corrected cDNA ratios (common allele/rare allele, T/C) for each sample are given in Figure 7.8.

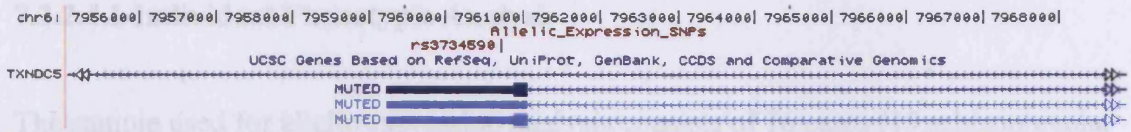


Figure 7.7. Chromosomal position of rs3734590. Allelic expression analysis of this coding SNP captures variation at all MUTED predicted mRNA transcripts.

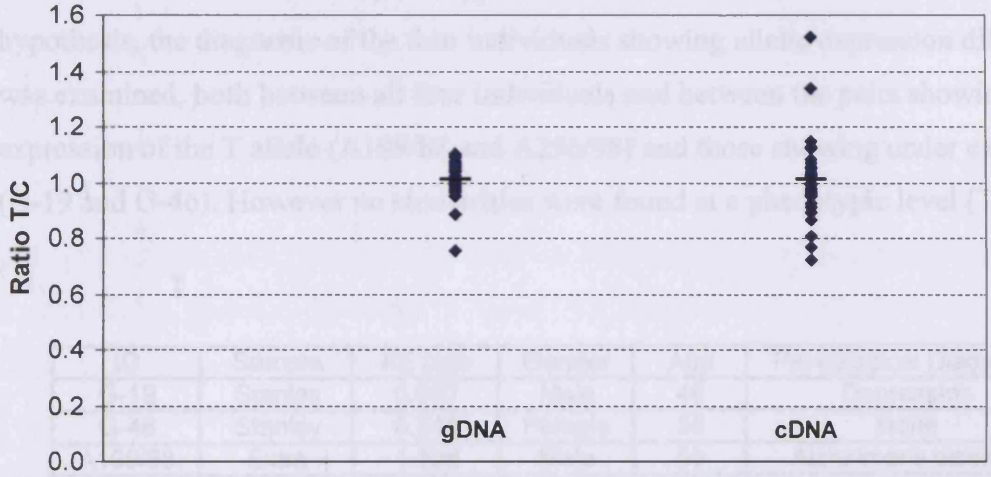


Figure 7.8. Relative allelic expression assay of rs3734590, an exonic SNP within the MUTED gene. Both gDNA and cDNA were analysed twice in 36 heterozygote individuals using cDNA gained from the cerebral cortex. Shown are the average gDNA and cDNA allelic ratios (T/C) corrected by the mean gDNA ratio for all samples. Independent T-test p=0.872, Levene's test =0.012.

Replication of the assay confirmed that four individuals were under the influence of cis-acting variation which caused differential expression of greater than 20% (1.2<C/T ratio<0.8).

Although differential expression was observed, the majority of individuals did not show allelic expression differences and no significant difference was observed between the mean of the corrected allele ratios for the gDNA versus cDNA (independent t-test p=0.872). A significant difference of the variance between genomic and cDNA ratios was observed however (Levene's test p=0.012). Of the four individuals showing differential expression of greater than 20%, two show an increase of the T allele compared to the C alleles (ratios>1.2) and two show a relative increase of the C allele over the T allele (ratios <0.8).

7.3.2.4.1 Individual Phenotypic Analysis

The sample used for allelic expression analysis consists of 76 cases (19 schizophrenia, 15 bipolar, 16 depression, 4 alcohol or drug dependence, 22 Alzheimer's disease) and 73 controls. Therefore the variability in allelic expression ratios between individuals could be attributable to these phenotypic differences. In order to examine this hypothesis, the diagnosis of the four individuals showing allelic expression differences was examined, both between all four individuals and between the pairs showing over expression of the T allele (A199/88 and A296/95) and those showing under expression (G-19 and G-46). However no similarities were found at a phenotypic level (Table 7.3).

ID	Sample	AE ratio	Gender	Age	Psycological Diagnosis
G-19	Stanley	0.657	Male	46	Depression
G-46	Stanley	0.748	Female	35	None
A199/88	Extra	1.306	Male	90	Alzheimer's disease
A296/95	Extra	1.386	Male	65	None

Table 7.3. Details of individuals showing allelic expression differences of >20% for rs3734590. The 4 samples do not show similarity on case/control status.

7.3.2.5 CNO

Genotyping of rs3172604, an exonic SNP within the CNO gene (Figure 7.9), in the gDNA of 60 individuals identified 20 heterozygotes. Allelic expression analysis was performed on these 20 heterozygotes. Two samples showed a standard deviation between duplicate RT reactions of >0.2 and so were removed from further analysis. The mean standard deviation for the remaining 18 samples was 0.045. Figure 7.10 shows the corrected genomic DNA ratios and the corrected cDNA averaged ratios (common allele/rare allele, T/G) for each sample.

A highly significant difference was observed between the mean corrected allele ratios for the gDNA versus cDNA with 11 out of 18 individuals showing allelic expression differences greater than 20% (independent t-test $p=2.42 \times 10^{-07}$). This is indicative of a common cis-acting allele influencing the levels of the CNO gene. Further analysis of this data (see Figure 7.10) identified a relative decrease in the expression of the mRNA carrying the rarer G allele compared to the mRNA carrying the T allele. As this difference was observed in a high percentage of the individuals analysed (61%) there was no need to perform allelic expression in further samples.

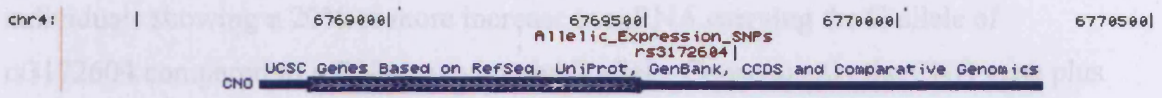


Figure 7.9. Chromosomal position of rs3172604. Allelic expression analysis of this exonic SNP will capture variation at all CNO predicted mRNA transcripts.

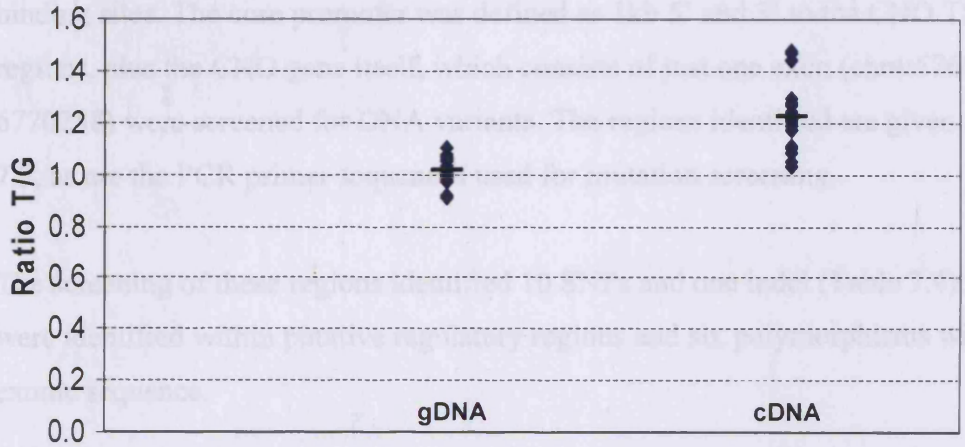


Figure 7.10. Relative allelic expression assay of rs3172604, an exonic SNP within the CNO gene. The gDNA and cDNA allelic ratios (T/G) of 18 heterozygotes were corrected by the mean gDNA ratio for all samples. Independent T-test p= 2.42E-07, Levene's test p =0.02.

7.3.3 CNO Association Analysis

CNO showed common allelic expression differences with over 60% of assayed individuals showing a 20% or more increase in mRNA carrying the G allele of rs3172604 compared to mRNA carrying the T allele. Consequently the CNO gene plus 10kb 5' and 3' (chr4:6,758,743-6,780,288) was screened for putative regulatory regions as described in chapter 3. While analysis for sequence of high conservation and DNase hypersensitivity did not identify any regions, Cluster Buster (http://zlab.bu.edu/cluster-buster) identified three regions predicted to contain multiple transcription factor binding sites. The core promoter was defined as 1kb 5' and 3' to the CNO TSS. These regions, plus the CNO gene itself, which consists of just one exon (chr4:6768743-6770288) were screened for DNA variants. The regions identified are given in Table 7.3, as are the PCR primer sequences used for mutation screening.

The screening of these regions identified 10 SNPs and one indel (Table 7.4). Five SNPs were identified within putative regulatory regions and six polymorphisms within the exonic sequence.

Name	Chromosomal Positon (March 2006)	Type of Region	Method	Relation to CNO	Forward	Reverse
CNO_Reg1	chr4:6766075-6766566	Regulatory	Cluster buster	5'	TCTTACATGAGGTTGCAGTCAA	ctcatctacccaagacctgga
CNO_5'region1	chr4:6767761-6768248	5'	Core Promoter	5'	actcccagttccaccatctg	aaggtgaagttgagatgtgacg
CNO_5'region2	chr4:6768160-6768541	5'	Core Promoter	5'	agtctaactgaatggtattctcagg	tgtgcacattgatgactttgt
CNO_exon1a	chr4:6768454-6768950	Exonic	N/A	N/A	gttcatgggtgctggacat	CTGTGGCTCTGGGAAACAGT
CNO_exon1b	chr4:6768843-6769312	Exonic	N/A	N/A	GGGTAGCTTTTCGGATGG	CCTTGGTGACCTGCTCCTC
CNO_exon1c	chr4:6769255-6769754	Exonic	N/A	N/A	GCCTTCGTGAGGATGGTG	GAGGTGGAAGGATCACTTGAG
CNO_exon1d	chr4:6769702-6770176	Exonic	N/A	N/A	TATGTTGCCCAGACTGTTCG	TGGACAAATGAAGACTGTGAGG
CNO_exon1e	chr4:6770052-6770449	Exonic	N/A	N/A	AACCTCCTATAGCGTGGATTG	gaaaagaaatataccatgctctgaaa
CNO_3'region1	chr4:6770330-6770848	3'	Core Promoter	3'	CGTACTGCTCATGTGCTTTTT	CCGCAGGAAAATCCTAATACC
CNO_3'region2	chr4:6770550-6771026	3'	Core Promoter	3'	cgttcacgccattctcct	TCAGATTCTTCACAACACACAGC
CNO_Reg2	chr4:6771114-6771509	Regulatory	Cluster buster	3'	ggaactctgaatgtagaagctgaa	ctcccgggttcacaccat
CNO_Reg3	chr4:6777653-6778154	Regulatory	Cluster buster	3'	ggcactgtcagagcaccac	tgcatgcatgtgtgatgtgt

Table 7.3. CNO exonic and putative regulatory regions screened for DNA variants. PCR primer sequences are also given.

SNP	Chromosomal Position	Alleles	Region	MAF CEU
CNO_Exon1_SNP1	chr4:6768674	G>C	Exonic	N/A
CNO_Exon1_SNP2	chr4:6769653	T>G	Exonic	N/A
rs3172604	chr4:6769676	G>T	Exonic	0.4
rs11550047	chr4:6769712	A>G	Exonic	0
CNO_Exon1_SNP3	chr4:6769908	C>T	Exonic	0.008
CNO_Exon1_INDEL	chr4:6770997-6770000	CAGT/-	Exonic	0.37
rs4689525	chr4:6771473	A>T	Regulatory2	N/A
rs4689526	chr4:6771482	T>A	Regulatory2	N/A
rs12503163	chr4:6777760	C>T	Regulatory3	0.57
rs12503199	chr4:6777877	C>T	Regulatory3	0.57
rs12505863	chr4:6777940	A>G	Regulatory3	0.69

Table 7.4. SNPs identified through screening of exonic and putative regulatory regions of CNO. Where frequency information was already known or where it has been possible to type these through the HapMap CEU sample, minor allele frequency (MAF) data is given.

Of the 11 polymorphisms identified, four variants were not referenced in dbSNP (CNO_exon1_snp1, CNO_exon1_snp2, CNO_exon1_snp3 and CNO_exon1_indel). Only two polymorphic SNPs within the CNO locus have been deposited in the HapMap database (March 2008), both of which were detected in this screening (rs3172604 and rs12503163). An attempt was made to genotype the remaining nine polymorphisms through the HapMap CEU sample. However four SNPs (See Table 7.4) detected through screening of the CNO gene and putative regulatory regions failed genotyping by both Sequenom MassARRAY and Amplifluor. All four of these SNPs fall in or near a repetitive element. These four SNPs have therefore not been captured in the subsequent study. Pairwise tagging of the remaining 7 polymorphisms selected a set of five tag snps which captured 100% of the variation at r^2>0.95 and MAF>0.001. However due to genotyping failure of rs3172604 through the UKCC sample a final set of 4 tag SNPs were typed which captures 83% at r^2>0.95 and MAF>0.001 of all DNA variants.

Of the 4 tag SNPs genotyped across the CNO locus and putative regulatory regions all had a call rate of >95%. Individuals were only included in analysis if they had genotypes for >2 of the 4 SNPs analysed. This meant 17 individuals were dropped from the original sample. All markers were in HW equilibrium at p>0.001 for both cases and controls. However to compensate for even small deviations in HW, the Armitage trend test was used to test for association (Table 7.5a). Of the four polymorphisms genotyped through the UKCC sample none showed allelic or genotypic association with schizophrenia (Table 7.5a and 7.5b).

rsID	Chromosome 6 Position	Alleles	Frequency Cases	Frequency Controls	Arm Trend P	Arm Trend Emp (10000)
CNO_Exon1_SNP3	6769908	C>T	0.002	0.001	0.948	0.5812
CNO_Exon1_INDEL	6770997-6771000	CAGT/-	0.366	0.387	0.248	0.2466
rs12503163	6777760	C>T	0.423	0.427	0.844	0.8468
rs12505863	6777940	A>G	0.311	0.315	0.828	0.8313

Table 7.5a. Allelic and genotypic results for CNO tag SNPs. Chromosomal positions are according to UCSC human genome chromosome 6 reference sequence (March 2006 freeze). Arm=Armitage.

rsID	Chromosome 6 Position	Alleles	Counts Cases	Counts Controls	HW Cases	HW Controls	Genotypic P	Clump P (10000)
CNO_Exon1_SNP3	6769908	C>T	0/2/662	0/2/707	1.000	1.000	NA	0.998
CNO_Exon1_INDEL	6770997-6771000	CAGT/-	84/308/258	90/365/249	0.673	0.017	0.218	0.218
rs12503163	6777760	C>T	122/314/223	137/330/240	0.524	0.219	0.903	0.903
rs12505863	6777940	A>G	64/283/313	72/303/334	1.000	0.794	0.960	0.960

Table 7.5b. Genotypic results for CNO tag SNPs. Chromosomal positions are according to UCSC human genome chromosome 6 reference sequence (March 2006 freeze). HW = Hardy-Weinberg P value.

7.4 Discussion

This chapter attempted to determine whether the BLOC-1 genes PLDN, SNAPAP, MUTED, CNO or BLOC-1 subunits 1, 2, or 3 are under the influence of cis-acting variation. However allelic expression analysis could not be performed on BLOC1S1 or BLOCS2 as neither gene contains any exonic SNPs suitable for use as a copy-specific tag.

Of the five BLOC-1 genes assayed, PLDN, SNAPAP and BLOC1S3 did not show any evidence for differential expression between the mRNA from each chromosome using the criteria previously defined. Therefore this study provides no evidence of polymorphic cis-acting effects on these genes. A potential limitation of allelic expression analysis to detect cis-acting variants can be inadequate power at a locus due to the relatively small number of heterozygotes studied. Considering linkage equilibrium between the assayed polymorphisms and potential regulatory variants, the subsets of 10, 21 and 6 heterozygotes assayed for PLDN, SNAPAP and BLOC1S3 loci respectively yielded a power of approximately 86%, 98% and 70% to detect the effect of a regulatory variant present in the general population at a MAF of 0.1 (Table 7.6). In the presence of linkage disequilibrium between the putative regulatory variant and the assayed SNP, the corresponding power to detect a cis-acting variant is greater as the number of individuals heterozygous for the regulatory variant would be higher than that expected by chance. It is therefore unlikely that the PLDN, BLOC1S3, or SNAPAP loci contain common cis-acting polymorphisms or cis-acting epigenetic mechanisms which have a significant impact on mRNA expression in the cerebral cortex.

		PLDN (10)	SNAPIN (21)	BLOS3 (6)	MUTED (35)	CNO (18)
MAF	0.05	63	88	45	97	83
	0.1	86	98	70	100	97
	0.15	95	100	83	100	100
	0.2	98	100	90	100	100

Table 7.7. Power (%) to detect a putative regulatory variant with varying MAF given the number of samples assayed. Test assumes linkage equilibrium with the marker polymorphism. The number of heterozygotes detected for each gene are given in brackets.

One potential caveat of this conclusion is that primers designed for the allelic expression analysis were not cDNA specific and therefore susceptible to genomic DNA contamination. Any gDNA contamination has the potential to mask any cis-acting effects by normalising the data towards the genomic ratio of 1:1. In order to minimise this possibility, all RNA samples were DNase treated prior to RT-PCR samples. Treated samples did not generate detectable PCR products with the "universal" primers in the absence of the reverse transcription cDNA synthesis step, suggesting that genomic contamination of cDNA was not an issue. Moreover, analysis of the two other BLOC-1 genes, MUTED and CNO, using the same gDNA and cDNA samples showed evidence for cis-acting influences and therefore also provided support against gDNA contamination.

Although allelic expression analysis of PLDN, BLOC1S3 and SNAPAP did not show any evidence for cis-acting variation, two BLOC-1 genes (MUTED and CNO) did show differential expression. Allelic expression of the MUTED gene demonstrated that the locus is likely to be under the influence of a rare cis-acting variant as, using a threshold previously defined by others [254], the majority of individuals do not show allelic expression differences. This possibility is further supported by the high number of heterozygotes assayed at this locus which provided a greater power to detect cis-acting effects (a power of approximately 97% to detect the effect of a regulatory variant present in the general population at a MAF of 0.1, Table 7.7). Furthermore, due to the spread of the data observed, the exonic polymorphism assayed rs3734590, appears to be in weak LD with the unknown regulatory polymorphism. In certain individuals the T allele is over-expressed and in others it is under expressed. This is suggestive of a switch in phase between the assayed SNP rs3734590 and the regulatory variant and therefore nominal linkage disequilibrium.

In comparison to MUTED, allelic expression analysis of CNO determined that the gene is under the influence of common cis-acting variation with 11/18 heterozygote individuals assayed showing differential expression of greater than 20%.

Although these genes have previously been analysed for association with schizophrenia [207, 373] no study has analysed putative regulatory SNPs for association. In order to test the hypothesis that CNO or MUTED could confer susceptibility to schizophrenia through altered expression, all putative regulatory polymorphisms plus exonic DNA variants at either locus were subjected to association analysis with schizophrenia. The results of the analysis of the CNO gene are described in this chapter while the analysis of MUTED is described in chapter 6. No allelic or genotypic association was observed for CNO, nor MUTED after correction for multiple testing. This suggests that while MUTED and CNO appear to be under the influence of cis-acting variation, this variation does not confer susceptibility to schizophrenia. Previous studies have shown that a high proportion of genes are likely to be under the influence of cis-acting variation and rather than being pathogenic, the majority of this cis-acting influence is responsible for population variation [184]. This could therefore be the case with CNO and MUTED.

On the other hand, it is possible the schizophrenia risk variants for MUTED or CNO and even the other BLOC-1 genes, have not yet been identified. Although an attempt was made to enrich for regulatory variants, only a small percentage of the genomic sequence surrounding both loci has been analysed. There may be other regularity variants not captured in the association analysis. It also must be noted that due to genotyping failure only 83% of the variation at the CNO locus has been captured. Furthermore, as discussed in chapter 6 the power of the schizophrenia UKCC to detect an association at a given locus (the sample is underpowered to detect effects with OR<1.4) must also be considered.

It is also possible that there may be risk variants at the MUTED or CNO loci which confer susceptibility through mechanisms not assayed by allelic expression or association analysis of coding regions. The allelic expression analysis described here does not address the possibility of post-transcriptional events such as translational efficiency.

Nevertheless, the data described in this chapter, along with the association analysis of the MUTED locus described in chapter 6, does not support a role for the BLOC-1 complex in schizophrenia pathology and suggests that dysbindin may confer susceptibility to schizophrenia via other pathways other than the BLOC-1 complex.

Chapter 8: General Discussion

The purpose of this thesis was two fold. The first objective was to investigate the hypothesis that DTNBP1 risk variants contribute susceptibility to schizophrenia through cis-acting regulation of dysbindin expression. The second section of this thesis investigated the hypothesis that DTNBP1 could cause susceptibility to schizophrenia through is role within the BLOC-1 complex.

In regards to the first objective, four schizophrenia risk variants were identified at the DTNBP1 locus, though logistic regression suggests that these polymorphisms constitute the same association signal. Seven SNPs were found to be associated with differential expression of DTNBP1 mRNA with the majority of the expression differences were accounted for by variation at two loc, one of which was subsequently shown to affect transcription *in vitro* within a cell based system. However comparison of the SNPs associated with schizophrenia and those associated with differential expression suggests that a reduction in dysbindin expression through cis-acting variation may not be the primary aetiological factor in schizophrenia pathogenesis. The most significantly associated risk variant, rs4715984, was not correlated with allelic expression differences. In addition, further analysis of a risk haplotype previously reported to be associated with allelic expression differences, determined that the refined haplotype was no longer correlated.

While these findings suggest cis-acting regulation of DTNBP1 does not cause susceptibility to schizophrenia, there are a number of other possibilities that explain the data observed. Firstly, the association shown between rs4715984 and schizophrenia may be a false positive. Alternatively, rs4715984 may be truly associated with schizophrenia but may not be the causal variant. For example rs4715984 could show the greatest association to schizophrenia as it is in LD with two causal variants, one of which could be rs2619538 or rs9296989. rs2619538 showed the greatest correlation

with allelic expression data and also showed association with schizophrenia. However unsuccessful cloning of rs2619538 into a luciferase vector meant that the SNP could not be assayed for an *in vitro* effect on expression. rs9296989 is also a candidate for the causal variant. Although it was less significantly associated with allelic expression differences than rs2619538 and located distal to DTNBP1, it was also significantly associated with schizophrenia.

Another potential explanation is that rs4715984 is the causal variant but the flipping of phase between rs4715984 and the allelic expression SNP rs1047631 may prevent the detection of a significant correlation by the statistical methods employed in this thesis. However even this could not explain the differential expression shown by rs4715984 homozygotes.

If rs4715984 is the causal variant, or in high LD with the causal variant, it may be that the allelic expression data used here is underpowered to assess the association between rs4715984 and differential expression. Another possibility is that cis-acting variation detected by rs1047631 does not cause susceptibility to schizophrenia. As described previously rs1047631 does not assay the expression differences of individual DTNBP1 transcripts. It may be the case that while DTNBP1 is generally under the influence of cis-acting variation, it is only the altered expression of a specific transcript that causes susceptibility to disease.

As well as transcript specific expression differences, cell type specific differences also need to be considered. Dimas and colleagues performed gene expression profiling followed by association with DNA variants within three cell types (primary fibroblasts, lymphoblastoid cell lines and primary T-cells) of 75 individuals [349]. Although regulatory polymorphisms were detected it was determined that up to 80% of these variants operated in a cell-type specific manner. Therefore a potential hypothesis could be that rs4715984 only regulates expression of DTNBP1 in certain tissues/cell types.

While these explanations are possible, the analyses performed in this study provide no evidence for cis-acting variation as a mechanism for the reduction in DTNBP1 mRNA expression observed in schizophrenic patients. Furthermore, the only previous evidence which linked these observations, that a schizophrenia risk haplotype was associated with reduced DTNBP1 expression, no longer appears to be the case. The reduction of DTNBP1 mRNA observed in schizophrenia patients [181-183] may therefore be caused by other non-cis processes. mRNA abundance is dependant on both transcription rates and mRNA stability. While both these aspects can be affected by cis-acting variation, other factors also exist. Epigenetic mechanisms, such as CpG methylation and histone modification can affect transcription rates [374]. In addition, trans-acting factors have been found to affect transcript stability either through mechanisms mediated by siRNAs or through protein-RNA interaction [375]. Finally, DNA variants at the DTNBP1 locus could affect alternative splicing and therefore the relative abundance of specific DTNBP1 transcripts.

8.1 Genome-wide Association Studies and Implications for DTNBP1

Until recently the evidence for association shown by DTNBP1 was unprecedented in schizophrenia and it was assumed that this was far beyond what could be attributed simply to chance. However results from recent studies, which have performed the first genome-wide association studies (GWAS) in schizophrenia, mean that DTNBP1 no longer shows the greatest evidence for association in schizophrenia research and the assumption that the association shown at DTNBP1 is not due to chance has been questioned. As the name suggests GWAS studies are able to survey the entire genome for common variants that might underlie a genetic disease in a single assay. These types of studies have become possible due to the completion of the human genome sequence, the deposition of millions of SNPs into public databases, the initiation of the HapMap project and the rapid improvements in SNP genotyping technology [77]. Currently the average number of SNPs that are analysed prior to QC is ~500,000. The scope of these studies is therefore unprecedented and will continue to increase with advances in SNP array technology.

To date there have been eight schizophrenia GWAS studies published [143, 376-382]. Although it must be noted that the eight GWAS studies that have been published are not totally independent, due to some overlap in the samples used. One of the common features of all schizophrenia GWAS is that while a number of novel loci have been identified as putative susceptibility genes, the top hits observed by each study are not within previous schizophrenia candidate genes. This includes those genes with the most compelling previous evidence such as DTNBP1. Shi *et al* do report a significant association with rs17619975, a SNP within JARID2, the gene adjacent to DTNBP1 [379]. However this SNP does not survive correction for the number of SNPs assayed in the study and the significance decreases after meta-analysis with two additional GWAS studies [378, 381].

While this is initially discouraging, a number of explanations are possible in which DTNBP1 is a schizophrenia susceptibility gene. Firstly, the majority of schizophrenia GWAS studies have analysed less than 1000 cases and have therefore been underpowered to detect an association after accounting for the large number of SNPs assayed (known as genome-wide significance). Genome-wide significance ($\sim p<10^{-8}$) has only been detected through meta-analysis of either several schizophrenia GWAS studies [378, 379, 381] or where schizophrenia and bipolar GWAS studies have been combined [377].

Due to multiple testing even a common SNP (MAF=0.2) with a relatively large effect size (OR=1.5) needs to be genotyped in a large sample (1500 cases and 1500 controls) to reach genome-wide significance and even this sample size would only have 80% power to detect the variant. As schizophrenia susceptibility is likely to be caused by multiple genes of weak effect (OR<1.5) even larger sample sizes will be needed to detect the majority of schizophrenia susceptibility variants, which may include DTNBP1.

Furthermore, while GWAS SNP arrays are designed to assay much of the common variation across the genome, the coverage of individual genes is unlikely to be 100%. The maximum coverage of DTNBP1 on the most technological advanced Affymetrix

and Illumina arrays (Affymetrix 6 and Illumina 1M) is still only 77% and 88% respectively (r^2>0.95, MAF>0.001). Moreover, these percentages do not take into account the removal of SNPs due to QC measures which would reduce this coverage further. It is also impossible to calculate the coverage of variants not genotyped as part of the HapMap project or variants that have yet to be identified. Therefore at present a negative or a non genome-wide significant association of a gene within a GWAS study is by no means conclusive.

8.2 Determining the Function of Susceptibility Variants Identified by GWAS

A major aim of this thesis was to identify putative DTNBP1 regulatory regions in which schizophrenia risk variants could be located. One advantage of GWAS studies is that due to the number of SNPs that can be assayed, the whole genome can be analysed in an unbiased way. Therefore no hypotheses are needed as to the function of susceptibility genes or the nature of the risk variants [77]. While a lack of functional annotation of the genome is not an initial hindrance in identifying associated polymorphisms, some of the strongest association signals detected in GWAS studies are within non-coding regions located in either large introns or far away from any annotated genes. As a result functional annotation is still, if not more, important than pre-GWAS.

Since the protocol which was used in this thesis for identifying putative regulatory regions was devised, significant advances have been made in the technology employed to identify putative regulatory sequence. The encyclopaedia of DNA elements (ENCODE) project aims to identify every sequence in the human genome with a functional role using a diverse set of techniques, including both computational analysis and "wet laboratory" technology [383]. The scale of the ENCODE project has meant several new technologies have been developed in order to generate high throughput data on functional elements. These techniques include several chromatin immunoprecipitation (ChIP) based methods. As discussed in section 3.1.2 chromatin immunoprecipitation allows the identification of protein-DNA interactions *in vivo*. By applying formaldehyde to cells, DNA-binding proteins, such as transcription factors,

are cross-linked to the DNA with which they are bound. The DNA is then fragmented by sonication prior to immunoprecipitation by an antibody specific to the protein of interest [276]. Historically, analysis of the immunoprecipitated DNA has taken the form of PCR and sequencing either directly or within a plasmid vector. However recent advances in technology in the form of ChIP-Chip and ChIP-Seq have allowed the identification of protein binding sites on a genome-wide scale. ChIP-Chip combines chromatin immunoprecipitation with microarray technology whereby immunoprecipitated DNA is labelled with a fluorescent tag then hybridized to a DNA microarray [384]. Hybridisation patterns can then be analysed to determine the DNA binding sites of the protein examined in the ChIP stage. ChIP-Seq uses high throughput sequencing rather than microarray technology to determine the protein-DNA interactions detected by ChIP [385]. ChIP-Seq is often thought of as a preferential method to Chip-Chip as it is not affected by the bias introduced when using arrays, which are restricted in the number of probes that can be fixed to them. Both ChIP-chip and ChIP-Seq can be used to detect TFBSs, promoter regions, histone modifications and methylation within the genome.

Another technique being used by ENCODE to identify regulatory regions is chromosome conformation capture-carbon copy (5C) [386]. This technique allows the identification of physical interactions between distant DNA segments and of chromatin loops which are formed as a consequence of these interactions [387]. As with ChIP, 5C involves the treatment of cells with formaldehyde. This is followed by restriction enzyme digest, intramolecular ligation and reverse cross-linking. The 3C library produced can then be determined via microarrays analysis or high throughput sequencing [386].

A number of promoter prediction programs are available for the *in silico* identification of promoters. These include PromoterInspector [388], FirstEF [389], McPromoter [390] and N-SCAN [391]. An assessment of these predictors [392], which used the ENCODE regions of the human genome as a point of reference, determined the best performing programs were those that combined promoter prediction with gene prediction, such as N-SCAN [391]. However the promoters predicted by these programs were inevitably

biased towards transcribed regions. Therefore promoter prediction software needs to be developed which also considers intergenic sequence.
Since the advent of the ENCODE project [383] and other independent research groups performing similar analysis the experimental data stored on the UCSC database relating to gene expression has grown substantially [313]. Therefore the challenge in functionally annotating a locus is likely to be what data to use rather than a lack of information. However, even if putative regulatory regions can be identified more confidently, the affect of a SNP on gene expression still needs to be determined. Therefore, it is essential to have powerful and reliable methods to easily test the influence of DNA variants on gene expression.

The method of identifying putative cis-acting variants used in this thesis, the allelic expression assay, is relatively low-throughput with only one gene analysed at a time. With the advent of GWAS studies more high-throughput methods have been devised. These methods are based on the association mapping of genotypes to multiple gene expression profiles, measured using microarray technology. In a similar approach to GWAS studies, microarrays allow the simultaneous assessment of the expression of the majority of genes in the genome [393]. The capacity of microarrays means that the expression of specific transcripts can also be measured in one assay. As expression can be considered a quantitative trait, the expression levels of a gene or a specific transcript have been named expression Quantitative Trait Locus' (eQTLs).

eQTL mapping is often performed in control individuals as this can significantly increase the power of an assay to detect expression differences relative to an assay which compares the expression of cases and controls. A number of eQTL studies are available publicly including the GENEVAR database [347] and the mRNA by SNP browser database [345]. While these two studies have performed expression profiling in lymphoblastoid cell lines, a number of eQTL studies have performed gene expression profiling in specific tissues including liver [394] and adipose tissue [395, 396]. However, potentially the most relevant eQTL study for researchers involved in schizophrenia is the study by Myers and colleagues [346]. This group performed gene expression analysis on RNA extracted from the cortex of 193 neuropathologically

normal human brains. In addition, DNA from these brains was genotyped on the Affymetrix genechip human Mapping 500k array. After QC 366,140 SNPs were analysed for association with the expression of 14,078 gene transcripts.

These eQTL databases are a useful tool for researchers attempting to determine a putative function for SNPs identified through GWAS studies. For example, a GWAS study of asthma identified a series of SNPs strongly associated with the disease [397]. These SNPs were in high LD and spanned more than 200kb of chr17q23. This region contained 19 genes, none of which were obvious candidates for disease. Examination of eQTL data derived from the same samples used in the GWAS showed that the disease-associated SNPs were highly significantly associated ($p<10^{-22}$) with the expression of one gene in the region (ORMDL3) and showed a borderline significance with another (GSDML) [345]. Further research determined that these SNPs were cis-acting regulatory variants for both genes [344]. Additional studies have therefore focused on investigating the biological function of these two genes and their relationship to asthma [398-402].

As well as determining putative functions for risk variants, the combination of GWAS and eQTL mapping can identify putative susceptibility genes that would not have been identified by GWAS alone. A GWAS into Crohn's disease (CD) identified markers on chromosome 5 which were strongly associated with CD. However as they resided within a 1.25Mb gene desert, a putative biological function could not be determined. Examination of a lymphoblastoid cell line eQTL database [344] showed that one or more of these associated SNPs act as long range cis-acting factors influencing expression of PTGER4, a gene over 250kb proximal to the associated region[403].

Another application of expression profiling is to compare the expression patterns of genes. By analysing the correlation in expression between genes, novel gene pathways or networks may be identified. This could result in the determination of additional functions for certain genes. At present DTNBP1 is known to function as part of the BLOC-1 and DPC complexes. While this thesis found no evidence to support the hypothesis that DTNBP1 causes susceptibility to schizophrenia through disruption of

the BLOC-1 complex and no association has been found with members of the DPC complex [207], it is possible that DTNBP1 functions in other pathways and/or complexes that are currently unknown. Expression profiling may be able to produce clues as to these alternative functions.

While eQTL mapping is advantageous in that gene expression arrays can analyse thousands of gene simultaneously, there are limitations to this approach. Systematic bias' can be introduced when using microarrays through differences in the hybridisation and measurement of expression, plus batch to batch variation in array manufacture and day to day variation in laboratory conditions [344].

The recent development of ultra-high throughput sequencing means that gene expression can be analysed without the limitations of microarray technology. Ultra-high throughput sequencing, also known as next generation sequencing, allows the parallel analysis of millions of sequence reads rather than the previous 96 which were possible with conventional capillary based systems. Next generation sequencers include the Illumina genome analyzer and Apllied Biosystems SOLiD sequencer. These systems can produce 13000-3000Mb of sequence per run. In contrast a single ABI 3730 can produce ~440kb of sequence per run [404].

The 1000 genome project has applied next generation sequencing technology to sequence ~1200 control human genomes (www.1000genomes.org). The project aims to discover >95 % of variants with minor allele frequencies as low as 1% across the genome and 0.1-0.5% in gene regions [405]. As with the International HapMap consortium [84], the 1000 genome project aims to estimate the population frequencies, haplotype backgrounds and linkage disequilibrium patterns of variant alleles identified. An alternative application for this technology is RNA-seq. RNA-seq uses high throughput sequencing technology to sequence cDNA in order to characterise the transcriptome and alternatively spliced transcripts [406, 407]. Sequence reads are individually mapped to the source genome and counted to obtain a density of reads corresponding to RNA from each known exon, splice event or new candidate gene [407]. However due to the size and complexity of the transcriptome and the low density

of transcribed heterozygous SNPs (one every ~3.3kb), most informative SNPs are not covered at the sequencing depth sufficient to make accurate allelic quantification using RNA-seq [408]. Digital RNA allelotyping combines the sensitivity and quantitative accuracy of RNA-seq with the efficiency of targeted sequencing. Digital RNA allelotyping is based on large scale synthesis of padlock probes [409]. Padlock probes contain the flanking sequence of a transcribed SNP plus linker DNA sequence which contains the primer sites for multiplex PCR amplification and sequencing. Padlock probes therefore allow sequencing efforts to be focused on the specific fraction of the transcriptome carrying SNPs. As with an allelic expression assay SNP capture and single-molecule sequencing is performed on both genomic DNA and cDNA of the same individuals. RNA allelic ratios are then calculated and corrected by the gDNA allelic ratios [408].

8.3 Tissue Specific Expression

One major factor that will need to be considered in allelic expression, total gene expression analysis and transcriptome sequencing is the possibility of tissue specific expression. Some eQTL studies of specific human tissues have been carried out, notably of the liver, adipose and brain tissue [346, 394-396]. Furthermore, this thesis has shown that allelic expression analysis assaying RNA extracted from specific tissues such as the brain is also possible. However a potentially vital resource for determining the function of risk variants identified within GWAS studies will be the Genotype-Tissue Expression (GTEx) project [410]. This project aims to produce a database containing eQTL data on up to 1000 samples each assayed across 30 different tissues. Researchers will therefore be able to analyse whether their risk variants affect expression levels of genes, not only within human tissue rather than lymphoblastoid cell lines, but also in tissue(s) relevant to their disease.

8.4 Biological Validation of Putative Regulatory Variants

While the technological advances described above are making the genome-wide analysis of regulatory variants a reality, it may still be necessary to determine whether a risk variant is a cis-acting variant and not just in high LD with functional polymorphism(s). In this thesis a luciferase reporter assay was used to determine whether SNPs correlated with allelic expression differences had an actual regulatory affect. However the development of luciferase reporter assay has not been as rapid as other methods described above such as association and expression analysis. While the ENCODE project is utilizing technology such as ChIP-chip and high throughput sequencing to identify putative regulatory regions, members of the consortium are using pGL3 luciferase constructs to test the transcriptional activity of putative promoters [383]. Details of the results of this analysis are available on the UCSC genome browser under the Stanford promoter activity track "Stanf Promoter". While the low throughput of luciferase reporter assays remains an issue, pGL4 luciferase vectors have been developed which contain both luciferase and *Renilla* reporter gene technology within a single construct, therefore removing the necessity for dual transfection in order to normalise transfection rates [411].

The results of another project which uses reporter gene assays to functionally validate putative regulatory regions are available on the VISTA enhancer browser database (http://enhancer.lbl.gov/) [412]. Pennacchio and colleagues first performed a comparative analysis of genomic sequence between the human genome and a wide range of available species (mouse, rat, chicken, frog, fugu, tetraodon and zebrafish) using the Gumby program [413] to determine evolutionarily conserved regions ($p < 0.001$). Regions identified were then filtered for transcribed sequence. In addition they utilized other comparative genomic datasets of extreme conservation such as non-coding "Ultra-conserved Elements" (defined as >200bp and 100% identical between human/mouse/rat) [309]. The enhancer activity of these regions has been assessed in transgenic mice using an *in vivo* lacZ reporter gene assay. As of September 2009 the group had tested nearly 2000 elements of which over 500 were observed to have enhancer activity.

8.5 General Conclusions

Since the research described in this thesis was started advances have been made in identifying DNA variants associated with disease. However the challenge still remains in determining the mechanisms by which these risk variants cause susceptibility to complex diseases and the biological pathways involved. While this thesis had limited success in identifying the DTNBP1 causal variants, it has indicated that correlation analysis of genotypes with expression data is a potential method of identifying cis-acting variants. eQTL mapping and high throughput allele specific expression are therefore likely to be essential tools as GWAS studies become routine. However a number of issues still need to be addressed. These include tissue and temporal specific expression, epigenetic mechanisms, gene interactions and gene-environment interactions. Consequently in order to improve our understanding of the biology of complex diseases such as schizophrenia a coordinated effort will be required from a broad range of geneticists, functional biologists and statisticians.

Appendix

9.1 Chapter 3 Appendices

Region	Chromosomal Position	Size	Subregion	Method	Forward	Reverse
1	15623636-15624128	493		Cluster buster	TTTCAAGCCATCCTTCAGAAA	GGACTACAGGCGCTCAACAT
2	15631372-15631829	458		Cluster buster	TCTGAGGGATTTGGAACCTG	CCGACTTTCTCAGCAGTGGT
3	15667548-15668039	492		Cluster buster	AGTGGGTGCCTTATGAGCTG	CAGTTGGCGAGAGGTCAAGT
4	15691653-15692140	488		ECR	ACAAACTCCATCCCAGTTGC	GGGGAATTGGCACTTTAACA
5	15703523-15704012	490		Cluster buster	GGAGTGCAGCGGTGTGAT	TGCTGTTCTGAAGTCTGTTTCC
6	15730941-15731413	473		Cluster buster	GCCTGACAGCTGTGCAAAA	GTCCAGGTTCCTTTCTGAGG
Alternative Transcript Promoter	15734658-15735135	478	P1	Putative promoter	GGCAATATTAAAAACAGGAGGAGA	GGATTGGAGATCAAACAAACCT
	15735038-15735511	474	P2		TCCTGCCTCAGTCTCCAGAA	AAACAACTTGGGCAGGGTTT
	15735451-15735934	484	P3		CCAGAATTTCATGTGTTCCTGA	TCATTTTTAAACTTTCTCTTCTGTGG
	15735760-15736001	242	P4a		TTTAATGTTATCTTTAACAACCCCTCT	TCTAGATTTAGCTTTTCAAATACATGG
	15736061-15736390	330	P4b		ACATAATTAACGGGTAATTA	AAAACAGAATTGTCTGGGAGAAA
	15736200-15736610	411	P5		GGAAACTGGCCAATTCCAGA	AAGAACCTTAATCTTGAGATGTACAAA
Main Transcript Promoter	15765956-15776344	10388		Putative promoter	SEE APPENDIX TABLE 3.2.	
19	15784237-15784635	399		DNase hyp	GGGGCAAAGCAAGCTCAC	AGTGACAGGAATGACCAAACG
7	15786747-15787164	418		Cluster buster	GAGAAAGGACTTCCCTAAAGAGG	TGAATCTATAAAACAAGGGCAAGA
8	15788437-15788915	479		Cluster buster	CCAGCCAATCTGACTAGGTAACA	CCCTTAAGACTGCAGGATGG
9	15790853-15791333	481		ECR	GGAGCAATAAAAATGAGCCATAA	GCCAACTGGGAAGTTGCTTA
10	15794240-15794722	483		Cluster buster	CATTTATGCATCCGTTGTCG	GGGTTGCCTCCATACTTTGA
11	15798238-15798703	466		Cluster buster	TTTCCACAGAGGTTGCATCA	TCTCCTTTGACGGTTTTTCC
12	15800898-15801321	424		Cluster buster	ACCAAGAGCTGATGGAGTGG	CCTGGCCAGTTCAGAATCTT
13	15803413-15803903	491		Cluster buster	TATGTGACCCGGGAACCTT	CTGTATGGAAGTCATAAATAGTGTCTG
14	15806380-15806876	497		ECR	CCAGTAAGTTGCAGGATTTCG	ACTTTCCAGCCACCAGAGTG
15	15820885-15821341	457		ECR/UCSC	TGTTGTGTTCTCAAAGCTTGC	GCAGCATCTGCCCTCTTATC
16	15841562-15841856	295	16a	UCSC	TGACATTCGAGAGATTTTCCTG	TGTTGACAATCCTGGCAGAC
	15841724-15842119	396	16b		TTGGTAAAAGAAAGAAGGTAGACAA	GCAAATGGTGGTGGTTCTTT
17	15842966-15843421	456		Cluster buster	TCATTTTACAGATGAGGCAACAA	AGCCTGGGCTACAGAGTGAA
18	15843888-15844152	265	18a	Cluster buster	TAGGCTCAACTGCGAGATGG	AAACGTGGGAGGAAATGATG
	15844048-15844446	399	18b		GAGCTCAGTGCCTAGCACATAAT	TTTCCACCCACCTCTGCTAC

Table 9.1.1. Putative DTNBP1 regulatory regions identified by the methods described in chapter 3.2.1 are shown in chromosomal order. Chromosomal positions are according to UCSC human genome chromosome 6 reference sequence (March 2006 freeze). Forward and reverse refer to the primers used for mutation screening. The PCR primers used for the screening of the putative promoter region of three of the main DTNBP1 transcripts are given in Appendix Table 9.2.

Region	Subregion	Chromosomal Position	Size	Method	Forward	Reverse
Main Transcript Promoter		chr6:15765956-15776344	10388	Putative Promoter		
	DYS_CIS_1		493		GCCTGCCACATAGTAAGCACT	TGCCAGTGCCAGACATTAGA
	DYS_CIS_2		493		TGAGCATATGGGCTCTCAAA	TTTAGGCATGCAAACCCTTC
	DYS_CIS_3		490		GCAGCTGTAAACCCGTCAGT	AAGGGCAGTGTAGAGATTACTGG
	DYS_CIS_4		491		CAAGGTTGACATAATTGTTACTACGC	ACACTTGGTGGCAGAGCAAT
	DYS_CIS_5		482		CTGAGGCCTTCCATTCCAT	TCACACCTGTAATCCCAGCA
	DYS_CIS_6		482		TAGTGGAGACGGGGTTTCAC	GAGACGTGAGCTGCATCACTA
	DYS_CIS_7		500		AAAACAAATCTGATTATGGACTGC	GGCCATTTTGGAGCTAGACA
	DYS_CIS_8		500		TTGTCAGCTGAAAGTGACCTG	TTGCATTTTGTTCTCCCTATTT
	DYS_CIS_9		497		CCATTTCAGCATCACACCAA	GGGATTACAGGCGTGTGC
	DYS_CIS_10		399		GGCAACAGAGCGAGAACTTG	AATCCTGTTGGAGGATGCAC
	DYS_CIS_11		471		GGGGGAGAAATTTAAAACAGTATG	CACCTGGGAACCTTTCAGAG
	DYS_CIS_12		457		CTTTAAAACGCCGTCTCCAG	GCCCTGGAGGGAAGTCAT
	DYS_CIS_13		392		CGCACGAGCAGGTGTCTG	CGACGAAAAGGGACCTGAG
	DYS_CIS_14		550		CCCGCAGGGACCTAAGTTAC	GATAGGAATGAGCCGAGGAG
	DYS_CIS_15		500		AGGTGAAATCCTGCTGCAC	TGTGCACATCGCCTATTGA
	DYS_CIS_16		481		TCCTTTGAGGGAAGTGTTGG	GAGGCTGGACTGTAGCCTTG
	DYS_CIS_17		497		TCTGGTTCCCCTCTTTTCCT	GGAGAAAAGTTTGTGAAAACACC
	DYS_CIS_18		500		ATCCACCAATTTGCTCAAGG	AACCTGGAATTTTCTAAGAACACA
	DYS_CIS_19		497		TTTTCTAGCACTGTATCCAAATTGA	TTGATATTTTGTTCATCACAGATTTTT
	DYS_CIS_20		496		GGGAGCCAAAAAGCTCAGTA	TGAATGCTCTTTGTTGGAAGTG
	DYS_CIS_21		500		GGAAGAGGTGACTACGATGATT	GCTTTGTTGGAATGTATGAAACTT
	DYS_CIS_22		472		TGAAAACCACAAGTACTGGGAAA	TTTTGCCAATCCATCTTCCT
	DYS_CIS_23		482		GGACCAAGAGTTGAAGCATGT	TGCCTGTAATCCCAGCATTT
	DYS_CIS_24		495		ATGTTGGACAGGCTGGTCTC	GTTCGCTCTGATGGTGGTTT
	DYS_CIS_25		471		TGAGGATTTGGCGGAGTTAG	AGGAATCGCCACACTGACTT
	DYS_CIS_26		491		AGGTGCTGGAGAGAATGTGG	GCCCCAGTGTATGATGTTCC
	DYS_CIS_27		489		CAAACACCGCATGTTCTCAC	TTGCTTCACGAGAAACATTTACA
	DYS_CIS_28		240		GCTTTGCAGGAACCTGACTT	CTTTTCAAAATGCTGTGCTGA

Table 9.1.2. PCR primer sequences for mutation detection within sequence 5kb 5' and 3' of the first exon of three of the main DTNBP1 transcripts (DTNBP1a, DTNBP1b and DTNBP1c). Chromosomal positions are according to UCSC human genome chromosome 6 reference sequence (March 2006 freeze). Forward and reverse refer to the primers used for mutation screening.

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1	rs4715984	.366	.144	6.446	1	.011	1.443
	rs9296989	-.111	.249	.198	1	.656	.895
	rs2619538	.026	.252	.010	1	.919	1.026
	Constant	-.047	.496	.009	1	.924	.954

Table 9.1.3. Logistic Regression of DTNBP1 SNPs showing allelic association using the enter method. P values of SNPs are adjusted for their correlation with all other SNPs in the model.

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1	rs4715984	.428	.138	9.631	1	.002	1.533
	Constant	-.159	.060	6.970	1	.008	.853

Variables not in the Equation

			Score	df	Sig.
Step 1	Variables	rs9296989	2.792	1	.095
		rs2619538	2.605	1	.107
		Overall Statistics	2.803	2	.246

Table 9.1.4. Forward Stepwise Logistic Regression of DTNBP1 SNPs showing allelic association. SNPs included and removed from the model are given. P values of SNPs removed are adjusted for their correlation with the other SNPs in the model.

Model if Term Removed

Variable		Model Log Likelihood	Change in -2 Log Likelihood	df	Sig. of the Change
Step 1	rs4715984	-921.922	6.518	1	.011
	rs9296989	-918.762	.198	1	.656
	rs2619538	-918.668	.010	1	.919
Step 2	rs4715984	-922.069	6.802	1	.009
	rs9296989	-920.064	2.792	1	.095
Step 3	rs4715984	-924.953	9.778	1	.002

Table 9.1.5. Backward Stepwise Logistic Regression of DTNBP1 SNPs showing allelic association. Variables entered in step 1: rs4715984, rs9296989 and rs2619538. P values of SNPs are adjusted for their correlation with all other SNPs in the model.

9.2 Chapter 5 Appendices

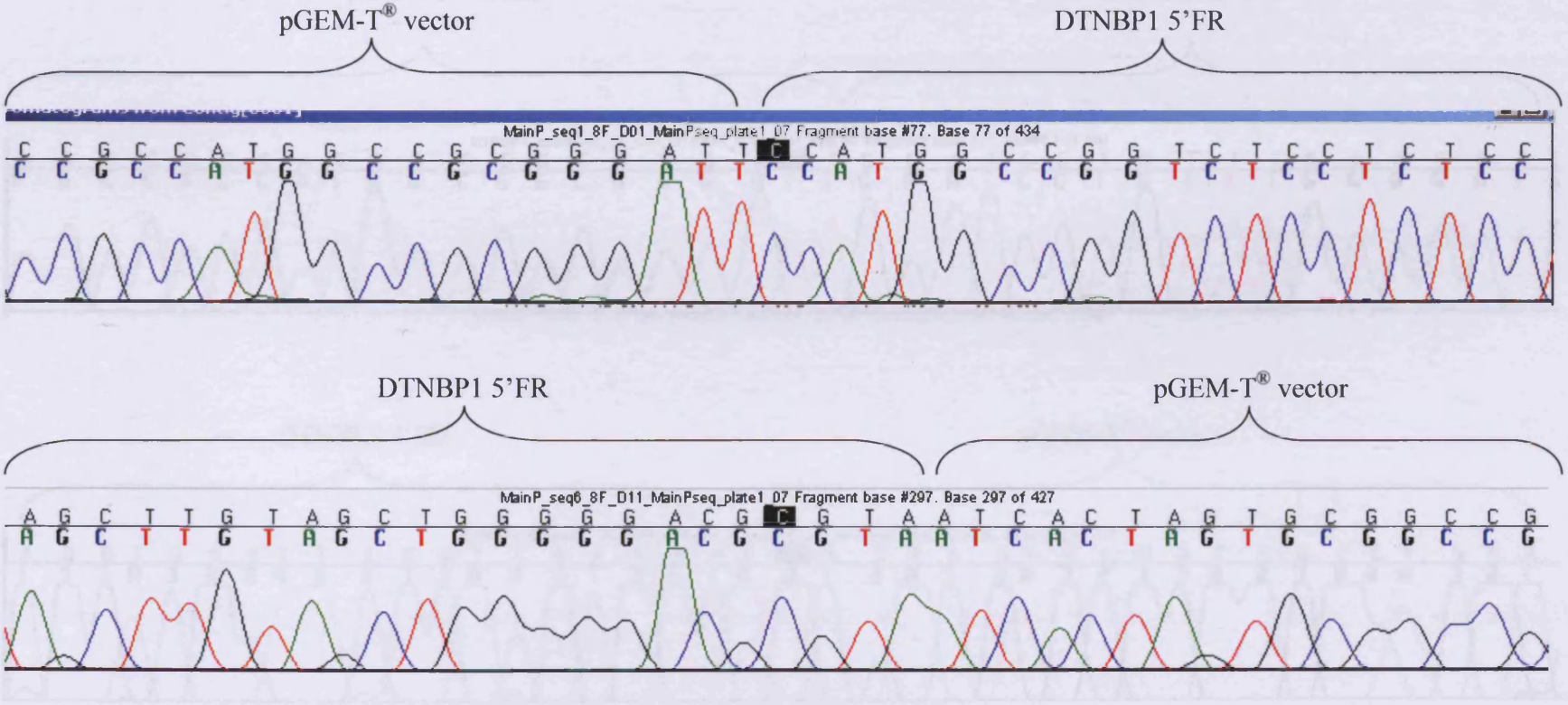


Figure 9.2.1. Sequencing traces of the ligation junction in the pGEMT-5'flanking region (FR) construct.

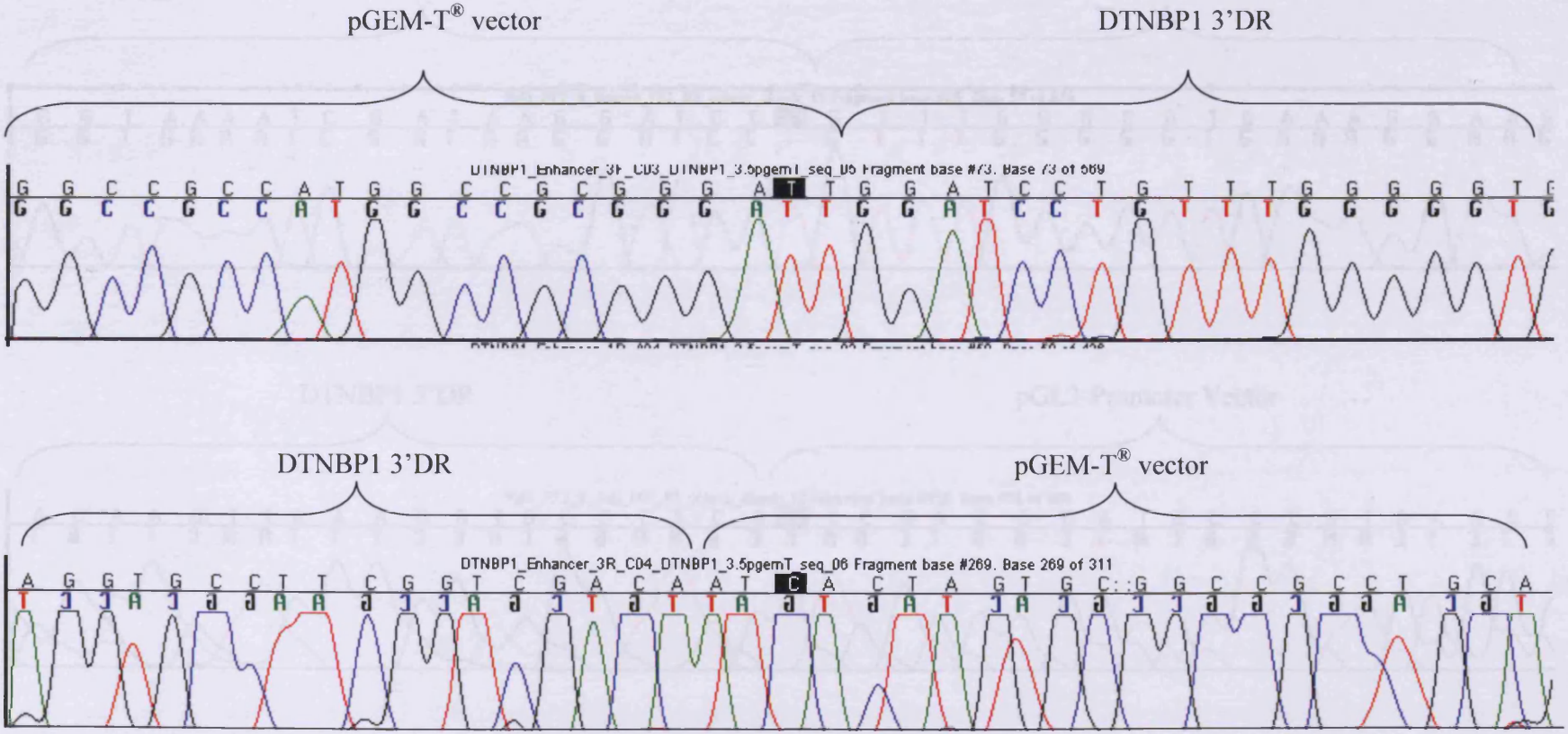


Figure 9.2.2. Sequencing traces of the ligation junction in the pGEMT-3'distal region (DR) construct.

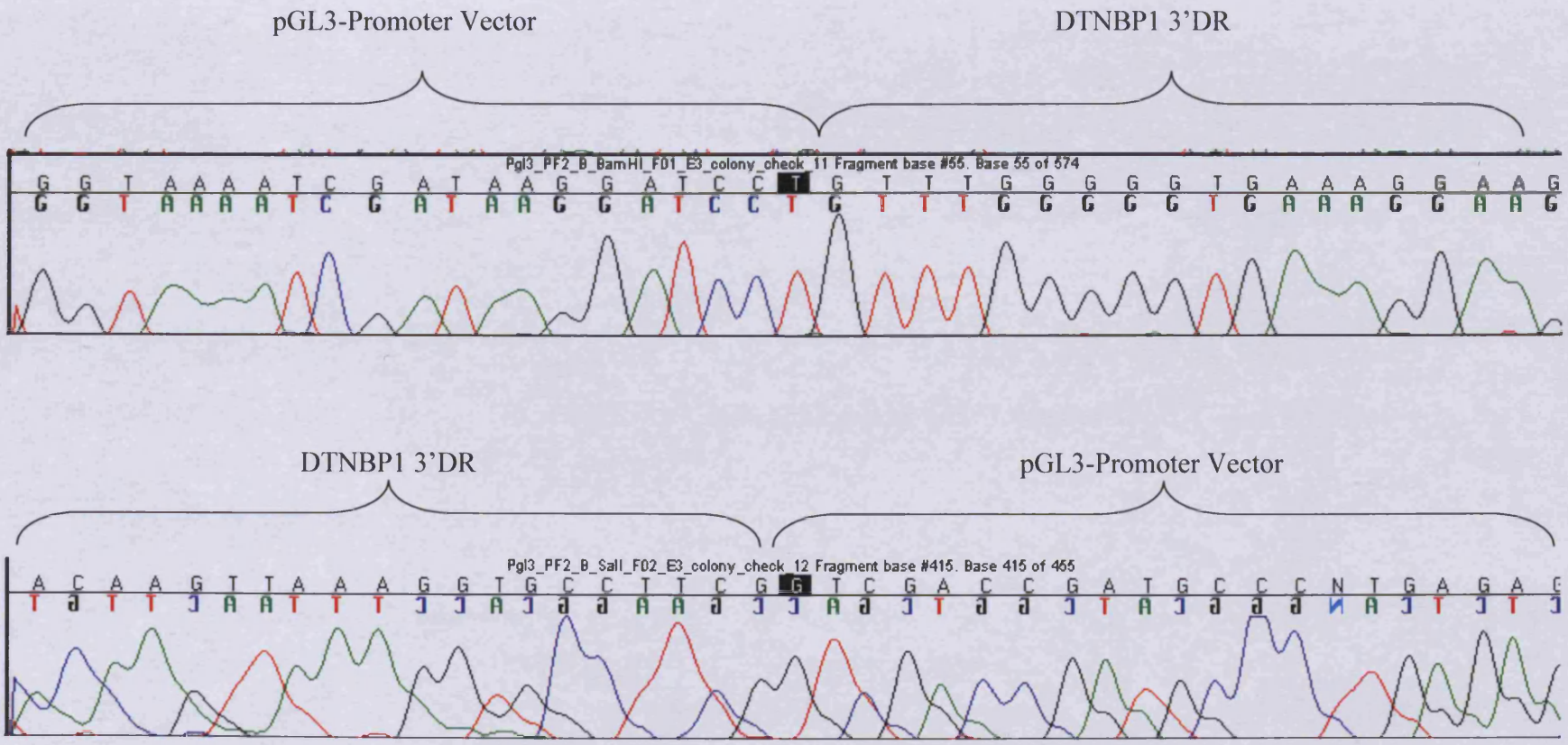


Figure 9.2.3. Sequencing traces of the ligation junction in the pGL3-P 3'DR construct.

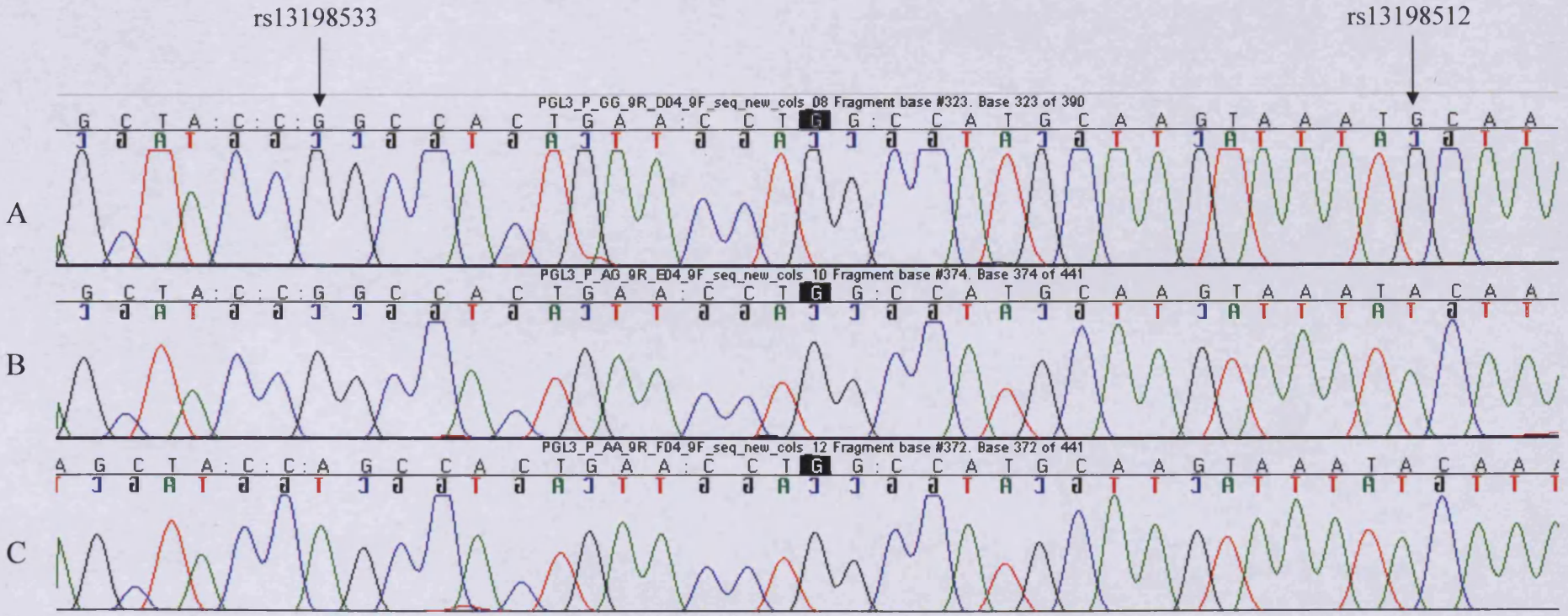


Figure 9.2.4. Sequencing traces of the three 3'DR inserted regions within their respective pGL3-P luciferase vectors after site directed mutagenesis. Genotypes for rs13198512 and rs13198533 are given for A. pGL3-P 3'DR_GG, B. pGL3-P 3'DR_AG, C. pGL3-P 3'DR_AA.

	Assay 1			Assay 2			Assay 3									All Data	
	Construct 1			Construct 1			Construct 1			Construct 2			Construct 3				
	Ratio	Normalised	Average	Ratio	Normalised	Average	Ratio	Normalised	Average	Ratio	Normalised	Average	Ratio	Normalised	Average	Average	SEM
	0.01			0.05			0.020										
	0.01			0.05			0.022										
	0.01			0.05			0.024										
	0.01			0.05			0.026										
	0.01			0.05			0.026										
PGL3	0.01		0.01	0.05		0.05	0.029		0.02							0.03	0.004
	0.42	0.91		1.99	1.00		1.075	0.932									
	0.43	0.93		1.89	0.95		1.215	1.054									
	0.44	0.96		1.96	0.99		1.007	0.874									
	0.46	1.01		1.94	0.98		1.182	1.026									
	0.51	1.11		1.95	0.98		1.184	1.027									
PGL3-P	0.49	1.07	1.00	2.21	1.11	1.00	1.254	1.088	1.00							1.00	0.016
	0.36	0.78		1.43	0.72		1.061	0.920		1.171	1.015						
	0.31	0.68		1.49	0.75		1.175	1.020		0.872	0.757						
	0.26	0.58		1.49	0.75		1.230	1.067		1.087	0.943						
	0.28	0.61		1.69	0.85		1.152	1.000		1.094	0.949						
	0.36	0.78		1.67	0.84		1.354	1.174		1.009	0.875						
PGL3-P 3'DR_GG	0.31	0.69	0.68	1.66	0.83	0.79	1.396	1.211	1.07	1.304	1.131	0.95				0.87	0.050
	0.03*	0.06*		0.36	0.18		0.311	0.270		0.281	0.244		0.269	0.234			
	0.11	0.24		0.31	0.16		0.252	0.219		0.274	0.238		0.252	0.218			
	0.24	0.53		0.58	0.29		0.299	0.259		0.290	0.251		0.227	0.197			
	0.29	0.64		0.65	0.32		0.303	0.263		0.288	0.250		0.314	0.272			
	0.38	0.83		0.94	0.47		0.347	0.301		0.345	0.299		0.335	0.291			
PGL3-P 3'DR_AG	0.19	0.41	0.53	0.69	0.35	0.30	0.382	0.332	0.27	0.381	0.330	0.27	0.348	0.302	0.25	0.32	0.027
	0.18	0.39		0.84	0.42		0.697	0.604		0.940	0.815		0.987	0.856			
	0.16	0.36		0.77	0.39		0.561	0.486		0.691	0.599		1.062	0.921			
	0.15	0.32		0.81	0.41		0.807	0.700		0.759	0.658		1.119	0.971			
	0.12	0.26		0.71	0.36		0.777	0.674		0.750	0.651		1.322	1.147			
	0.18	0.39		1.04	0.52		0.742	0.643		0.828	0.718		0.987	0.856			
PGL3-P 3'DR_AA	0.20	0.44	0.36	0.84	0.42	0.42	0.762	0.661	0.63	0.817	0.709	0.69	1.275	1.106	0.98	0.62	0.043

Table 9.2.5. Raw data for all biological and technical replicates of the luciferase gene reporter assay. Normalised values based on the average pGL3-P ratio for each assay are given to the right of each data point. SEM = Standard error of the mean. * Value not included in analysis

9.3 Chapter 6 Appendices

Name	Chromosomal Positon (March 2006)	Type of Region	Method	Relation to Muted	Forward	Reverse
MUTED_Reg1a	chr6:8014289-8014779	Regulatory	ECR/UCSC	5'	gatgggggagaaggagaact	tgcttccattatcatgattcctt
MUTED_Reg1b	chr6:8014689-8015188				gctggaaaagatgcaatgct	gcactgggtgaattttattgc
MUTED_Reg1c	chr6:8015117-8015479				ctctgcccagcaagagagat	tttctaacatttgacttaattccttttt
MUTED_Reg1d	chr6:8015338-8015694				gaaagaagaaatactgctctctgg	catgagagccctgctgaaat
MUTED_Reg2a	chr6:8010138-8010637	Regulatory	Promoter	5'	ccattatgtaggcatagttggtt	cacagaaagtgaaacctcagataa
MUTED_Reg2b	chr6:8009816-8010204				aacagcccccagctaattg	gcggatgctgctagaatagg
MUTED_Reg2c	chr6:8009603-8009894				caccagctttgagccaaata	actcatcccgaccagttcc
MUTED_Exon_1	chr6:8009318-8009714	Exonic	N/A	N/A	cgccctgtaatgacacacac	gttgtggtgggacgcattt
MUTED_Exon_2	chr6:8007685-8007929	Exonic	N/A	N/A	tggggaagggagagagtaaaa	aaaactggggaatttctcttctg
MUTED_Reg3	chr6: 7994178-7994348	Regulatory	Cluster buster	Intronic	Could not PCR	
MUTED_Exon_3	chr6:7986262-7986549	Exonic	N/A	N/A	aaaaaggaaaagacaccatatttatt	ctgagcttccctctccctct
MUTED_Exon_4	chr6:7971512-7971731	Exonic	N/A	N/A	cacactcctcccccaggtat	gagaattttctgatcaaaagcaa
MUTED_Reg4a	chr6:7968274-7968607	Regulatory	ECR/UCSC	Intronic	ggcttcaaaagttgacacca	tcttttatgtcaagacaatttgtgg
MUTED_Reg4b	chr6:7968243-7968689				cttcataattacttccagtctggtt	ccaccacacacggctaattt
MUTED_Reg5	chr6:7965568-7966064	Regulatory	UCSC	Intronic	gggtatggcatggacatctt	gcacaatgctcttataactttagcc
MUTED_Reg6	chr6:7965259-7965708	Regulatory	UCSC	Intronic	tcattccctaatgacatcacca	tttgggggagaagaactgg
MUTED_Exon_5a	chr6:7960793-7961192	Exonic	N/A	N/A	tttggatattataacacaacttttttcc	agcagaggctaaacggtctg
MUTED_Exon_5b	chr6:7960455-7960915				tggagaaggacctagcgaaa	aacagccaaggaggctatga
MUTED_Exon_5c	chr6:7960022-7960521				caatagtttattattgtggcttaatgg	gattactcattaaacagtcgaaacat
MUTED_Exon_5d	chr6:7959593-7960092				ctgttttgctgtgggtaagc	cctcctgagttcaggtgattct
MUTED_Exon_5e	chr6:7959165-7959661				tgttggtgcacgcttgtaat	gaaagttggcagaagttcagtg
MUTED_Exon_5f	chr6:7959094-7959487				tttttatttgccataaaccaagc	gcaagtgcgctttttagtcc
MUTED_Reg7	chr6:7958818-7959308	Regulatory	ECR	3'	ttcacatgagatgaacacaaact	gacaatgcctgcctgtgtaa

Table 9.3.1 All MUTED exons and putative regulatory regions identified by methods described in chapter 6.2.2 are shown. Chromosomal positions are according to UCSC human genome chromosome 6 reference sequence (March 2006 freeze). Forward and reverse refer to the PCR primers used for mutation screening.

		Assay 1			Assay 2			Assay 3			Assay 4		
Region	Chromosomal Position	Temp1	Grad1	Grad2	Temp2	Grad1	Grad2	Temp3	Grad1	Grad2	Temp4	Grad1	Grad2
Promoter_F1	chr6:8,010,138-8,010,637	55	33	25	60	40	32	62	38	30			
Promoter_F2	chr6:8,009,816-8,010,204	53	34	26	58	41	33	60	39	31			
Promoter_F3	chr6:8,009,603-8,009,894	63	46	38	65	44	36						
Exon1	chr6:8,009,318-8,009,714	64	42	34	66	40	32						
Exon2	chr6:8,007,685-8,007,929	54	44	36	56	42	34						
Exon3	chr6:7,986,262-7,986,549	52	47	39	57	44	36	59	42	34			
Exon4	chr6:7,971,512-7,971,731	63	38	30	58	45	37	60	43	35			
Exon5_F1	chr6:7,960,793-7,961,192	52	33	25	57	40	32	59	38	30			
Exon5_F2	chr6:7,960,455-7,960,915	56	39	31	58	37	29						
Exon5_F3	chr6:7,960,022-7,960,521	49	37	29	54	34	26	59	41	33	61	39	31
Exon5_F4	chr6:7,959,593-7,960,099	55	34	26	60	41	33	62	39	31			
Exon5_F5	chr6:7,959,165-7,959,661	56	38	30	60	43	35	62	41	33			

Table 9.3.2 Parameters for dHPLC analysis of MUTED exons and putative promoter. Temperature plus initial and final concentrations for buffer B are given for each assay. Optimal temperatures and corresponding elute gradients for each PCR fragment were selected using dHPLC Melt (http://insertion.stanford.edu/melt.html). In addition to the temperature suggested by the software (n°C), each fragment was also run at n+2°C to ensure maximum sensitivity.

	rs909706	rs2619539	rs12524251	rs2619538	SNPK	rs9476888	rs1047631	rs12525702	rs2743852	SNPN	rs3213207	rs12204704	rs12527496	rs3829893	SNPO
rs1002308	0.814 (0.809)	0.713 (0.716)	0.735 (0.734)	0.846 (.850)	0.955 (0.949)	0.849 (0.958)	0.969 (0.980)	0.091 (0.907)	0.032 (0.351)	0.077 (0.088)	0.059 (0.624)	0.931 (0.928)	0.870 (0.859)	0.846 (0.824)	N/A
Muted 5-5	N/A	N/A	N/A	N/A	0.088 (0.999)	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
rs10458217	0.646 (0.999)	0.273 (0.999)	0.820 (0.999)	0.607 (0.999)	0.988 (0.999)	0.332 (0.999)	0.508 (0.999)	0.783 (0.999)	0.188 (0.999)	0.112 (0.999)	0.205 (0.999)	0.535 (0.999)	0.956 (0.999)	0.964 (0.999)	N/A
rs2057185	0.707 (0.999)	0.217 (0.999)	0.313 (0.999)	0.961 (0.999)	0.679 (0.999)	0.670 (0.999)	0.183 (0.999)	0.171 (0.999)	0.824 (0.999)	0.179 (0.999)	0.456 (0.999)	0.0161 (0.882)	0.478 (0.999)	0.451 (0.999)	N/A
rs13202814	0.883 (0.999)	0.963 (0.999)	0.977 (0.999)	0.857 (0.999)	0.235 (0.999)	0.770 (0.999)	0.821 (0.999)	0.297 (0.999)	0.776 (0.999)	0.136 (0.999)	0.683 (0.999)	0.021 (0.884)	0.100 (0.999)	0.103 (0.999)	N/A
rs2207720	0.806 (0.999)	0.660 (0.999)	0.061 (0.950)	0.044 (0.975)	0.568 (0.991)	0.189 (0.999)	0.147 (0.965)	0.212 (0.996)	0.105 (0.999)	0.357 (0.999)	0.627 (0.999)	0.527 (0.999)	0.217 (0.999)	0.222 (0.999)	N/A
rs9328452	0.997 (0.999)	0.769 (0.999)	0.021 (0.830)	0.075 (0.996)	0.026 (0.999)	0.606 (0.999)	0.078 (0.999)	0.074 (0.999)	0.824 (0.999)	0.141 (0.999)	0.519 (0.999)	0.002 (0.289)	0.015 (0.798)	0.017 (0.831)	N/A
rs9502675	0.876 (0.999)	0.687 (0.999)	0.204 (0.999)	0.884 (0.999)	0.187 (0.999)	0.593 (0.999)	0.039 (0.999)	0.720 (0.999)	0.402 (0.999)	0.170 (0.999)	0.641 (0.999)	0.418 (0.999)	0.170 (0.999)	0.173 (0.999)	N/A
rs13218209	0.760 (0.999)	0.454 (0.999)	0.149 (0.999)	0.546 (0.999)	0.644 (0.999)	0.254 (0.999)	0.564 (0.999)	0.810 (0.999)	0.787 (0.999)	0.079 (0.996)	0.658 (0.999)	0.006 (0.561)	0.688 (0.999)	0.619 (0.999)	N/A
rs2206098	0.740 (0.999)	0.661 (0.999)	0.242 (0.999)	0.865 (0.999)	0.182 (0.999)	0.677 (0.999)	0.751 (0.999)	0.041 (0.971)	0.984 (0.999)	0.282 (0.999)	0.665 (0.999)	0.007 (0.569)	0.173 (0.999)	0.155 (0.999)	N/A
rs2815145	0.924 (0.999)	0.368 (0.999)	0.183 (0.999)	0.135 (0.999)	0.068 (0.999)	0.386 (0.999)	0.141 (0.999)	0.084 (0.999)	0.297 (0.999)	0.041 (0.971)	0.444 (0.999)	0.012 (0.737)	0.196 (0.999)	0.186 (0.999)	N/A
rs2743986	0.824 (0.999)	0.325 (0.999)	0.433 (0.999)	0.840 (0.999)	0.181 (0.999)	0.667 (0.999)	0.475 (0.999)	0.194 (0.999)	0.534 (0.999)	0.138 (0.999)	0.601 (0.999)	0.076 (0.998)	0.365 (0.999)	0.337 (0.999)	N/A
rs2326975	0.104 (0.999)	0.438 (0.999)	0.815 (0.999)	0.354 (0.999)	0.098 (0.999)	0.817 (0.999)	0.559 (0.999)	0.164 (0.999)	0.333 (0.999)	0.486 (0.999)	0.812 (0.999)	0.302 (0.999)	0.529 (0.999)	0.592 (0.999)	N/A
rs9392958	0.645 (0.999)	0.315 (0.999)	0.974 (0.999)	0.204 (0.999)	0.999 (0.999)	0.923 (0.999)	0.576 (0.999)	0.537 (0.999)	0.345 (0.999)	0.244 (0.999)	0.577 (0.999)	0.556 (0.999)	1.000 (0.999)	0.998 (0.999)	N/A
rs9378519	0.571 (0.999)	0.117 (0.999)	0.644 (0.999)	0.944 (0.999)	0.662 (0.999)	0.943 (0.999)	0.378 (0.999)	0.354 (0.999)	0.414 (0.999)	0.445 (0.999)	0.215 (0.999)	0.418 (0.999)	0.629 (0.999)	0.601 (0.999)	N/A
rs10458106	0.721 (0.999)	0.345 (0.999)	0.841 (0.999)	0.613 (0.999)	0.961 (0.999)	0.415 (0.999)	0.365 (0.999)	0.816 (0.999)	0.226 (0.999)	0.128 (0.999)	0.242 (0.999)	0.547 (0.999)	0.968 (0.999)	0.966 (0.999)	N/A
rs13191023	0.948 (0.999)	0.976 (0.999)	0.982 (0.999)	0.892 (0.999)	0.192 (0.999)	0.541 (0.999)	0.736 (0.999)	0.297 (0.999)	0.839 (0.999)	0.115 (0.999)	0.763 (0.999)	0.015 (0.799)	0.084 (0.996)	0.087 (0.997)	N/A
rs2815128	0.993 (0.999)	0.699 (0.999)	0.020 (0.856)	0.096 (0.997)	0.040 (0.980)	0.544 (0.999)	0.090 (0.997)	0.130 (0.999)	0.926 (0.999)	0.095 (0.997)	0.572 (0.999)	0.002 (0.207)	0.030 (0.945)	0.032 (0.954)	N/A
rs2815151	0.893 (0.999)	0.891 (0.999)	0.966 (0.999)	0.613 (0.999)	0.968 (0.999)	0.377 (0.999)	0.342 (0.999)	0.750 (0.999)	0.370 (0.999)	0.0007 (0.106)	0.020 (0.735)	0.515 (0.999)	0.961 (0.999)	0.956 (0.999)	N/A

Table 9.3.3A Interaction analysis between MUTED and DTNBP1 SNPs.
P-values are given uncorrected and amended for multiple testing in parentheses. Data was calculated using both the full additive model and dominant interactive model. Where the full interactive model could not be used a stepwise model was used to find the best term. These results are shaded in grey. N/A indicates that the marker pair could not be analysed as neither the specified interactive model nor any reduced model could be calculated. In these instances all interaction terms had either a high standard error for the coefficient estimates or singularities present. The ID's of the DTNBP1 SNPs are labelled as previously reported [134]. P values significantly associated are given in blue. Those still significant after permutation analysis are given in red.

	rs909706	rs2619539	rs12524251	rs2619538	SNPK	rs9476888	rs1047631	rs12525702	rs2743852	SNPN	rs3213207	rs12204704	rs12527496	rs3829893	SNPO
rs1002308	0.969 (0.999)	0.486 (0.999)	0.220 (0.999)	0.536 (0.999)	0.709 (0.999)	0.397 (0.999)	0.778 (0.999)	0.489 (0.999)	0.786 (0.999)	0.013 (0.774)	0.719 (0.999)	0.949 (0.999)	0.289 (0.999)	0.272 (0.999)	N/A
Muted 5-5	0.885 (0.999)	0.670 (0.999)	0.809 (0.999)	0.731 (0.999)	0.847 (0.999)	0.098 (0.998)	0.920 (0.999)	0.382 (0.999)	0.177 (0.999)	N/A	0.922 (0.999)	0.522 (0.999)	0.604 (0.999)	0.652 (0.999)	N/A
rs10458217	0.367 (0.999)	0.078 (0.997)	0.704 (0.999)	0.480 (0.999)	0.913 (0.999)	0.335 (0.999)	0.721 (0.999)	0.906 (0.999)	0.366 (0.999)	0.012 (0.757)	0.070 (0.995)	0.900 (0.999)	0.834 (0.999)	0.914 (0.999)	N/A
rs2057185	0.205 (0.999)	0.384 (0.999)	0.174 (0.999)	0.954 (0.999)	0.680 (0.999)	0.887 (0.999)	0.728 (0.999)	0.341 (0.999)	0.940 (0.999)	0.069 (0.995)	0.583 (0.999)	0.799 (0.999)	0.279 (0.999)	0.305 (0.999)	N/A
rs13202814	0.975 (0.999)	0.745 (0.999)	0.316 (0.999)	0.926 (0.999)	0.629 (0.999)	0.758 (0.999)	0.726 (0.999)	0.757 (0.999)	0.510 (0.999)	0.399 (0.999)	0.426 (0.999)	0.859 (0.999)	0.431 (0.999)	0.431 (0.999)	N/A
rs2207720	0.570 (0.999)	0.389 (0.999)	0.040 (0.968)	0.830 (0.999)	0.284 (0.999)	0.309 (0.999)	0.234 (0.999)	0.353 (0.999)	0.188 (0.999)	0.538 (0.999)	0.135 (0.999)	0.317 (0.999)	0.114 (0.999)	0.088 (0.998)	N/A
rs9328452	0.755 (0.999)	0.498 (0.999)	0.202 (0.999)	0.460 (0.999)	0.472 (0.999)	0.554 (0.999)	0.332 (0.999)	0.461 (0.999)	0.416 (0.999)	0.373 (0.999)	0.141 (0.999)	0.127 (0.999)	0.224 (0.999)	0.208 (0.999)	N/A
rs9502675	0.737 (0.999)	0.933 (0.999)	0.654 (0.999)	0.369 (0.999)	0.690 (0.999)	0.593 (0.999)	0.576 (0.999)	0.858 (0.999)	0.443 (0.999)	0.535 (0.999)	0.421 (0.999)	0.105 (0.998)	0.748 (0.999)	0.710 (0.999)	N/A
rs13218209	0.469 (0.999)	0.205 (0.999)	0.534 (0.999)	0.519 (0.999)	0.765 (0.999)	0.282 (0.999)	0.951 (0.999)	0.721 (0.999)	0.829 (0.999)	0.235 (0.999)	0.304 (0.999)	0.951 (0.999)	0.540 (0.999)	0.480 (0.999)	N/A
rs2206098	0.253 (0.999)	0.324 (0.999)	0.068 (0.995)	0.524 (0.999)	0.207 (0.999)	0.642 (0.999)	0.347 (0.999)	0.198 (0.999)	0.780 (0.999)	0.155 (0.999)	0.212 (0.999)	0.052 (0.984)	0.077 (0.997)	0.079 (0.997)	N/A
rs2815145	0.945 (0.999)	0.560 (0.999)	0.234 (0.999)	0.124 (0.999)	0.390 (0.999)	0.145 (0.999)	0.380 (0.999)	0.539 (0.999)	0.971 (0.999)	0.020 (0.868)	0.977 (0.999)	0.019 (0.860)	0.238 (0.999)	0.183 (0.999)	N/A
rs13207958	0.955 (0.999)	0.924 (0.999)	0.421 (0.999)	0.723 (0.999)	0.983 (0.999)	0.995 (0.999)	0.790 (0.999)	0.805 (0.999)	0.746 (0.999)	0.989 (0.999)	0.967 (0.999)	0.174 (0.999)	0.662 (0.999)	0.693 (0.999)	N/A
rs9392956	0.975 (0.999)	0.811 (0.999)	0.863 (0.999)	0.971 (0.999)	0.276 (0.999)	0.658 (0.999)	0.744 (0.999)	0.264 (0.999)	0.398 (0.999)	0.513 (0.999)	0.875 (0.999)	0.970 (0.999)	0.564 (0.999)	0.649 (0.999)	N/A
rs9392958	0.378 (0.999)	0.184 (0.999)	0.571 (0.999)	0.512 (0.999)	0.886 (0.999)	0.827 (0.999)	0.498 (0.999)	0.779 (0.999)	0.521 (0.999)	0.062 (0.994)	0.549 (0.999)	0.397 (0.999)	0.992 (0.999)	0.952 (0.999)	N/A
rs9378519	0.648 (0.999)	0.450 (0.999)	0.829 (0.999)	0.822 (0.999)	0.736 (0.999)	0.942 (0.999)	0.543 (0.999)	0.931 (0.999)	0.606 (0.999)	0.222 (0.999)	0.459 (0.999)	0.521 (0.999)	0.996 (0.999)	0.964 (0.999)	N/A
rs10458106	0.366 (0.999)	0.092 (0.998)	0.875 (0.999)	0.428 (0.999)	0.802 (0.999)	0.486 (0.999)	0.795 (0.999)	0.935 (0.999)	0.312 (0.999)	0.014 (0.797)	0.079 (0.996)	0.820 (0.999)	0.923 (0.999)	0.892 (0.999)	N/A
rs13191023	0.993 (0.999)	0.800 (0.999)	0.350 (0.999)	0.943 (0.999)	0.741 (0.999)	0.694 (0.999)	0.821 (0.999)	0.696 (0.999)	0.586 (0.999)	0.552 (0.999)	0.631 (0.999)	0.664 (0.999)	0.439 (0.999)	0.443 (0.999)	N/A
rs2815128	0.601 (0.999)	0.455 (0.999)	0.171 (0.999)	0.423 (0.999)	0.498 (0.999)	0.768 (0.999)	0.393 (0.999)	0.548 (0.999)	0.534 (0.999)	0.520 (0.999)	0.157 (0.999)	0.082 (0.998)	0.262 (0.999)	0.239 (0.999)	N/A
rs2815151	0.765 (0.999)	0.757 (0.999)	0.873 (0.999)	0.688 (0.999)	0.910 (0.999)	0.426 (0.999)	0.253 (0.999)	0.963 (0.999)	0.529 (0.999)	0.0001 (0.026)	0.917 (0.999)	0.741 (0.999)	0.832 (0.999)	0.868 (0.999)	N/A

Table 9.3.3B. Interaction analysis between MUTED and DTNBP1 SNPs.
P-values are given uncorrected and amended for multiple testing in parenthesis. Data was calculated the additive model. N/A indicates that the marker pair could not be analysed as neither the specified interactive model nor any reduced model could be calculated. In these instances all interaction terms had either a high standard error for the coefficient estimates or singularities present. The ID's of the DTNBP1 SNPs are labelled as previously reported [134]. P values significantly associated are given in blue. Those still significant after permutation analysis are given in red.

Bibliography

1. Gottesman, I. 1991. *Schizophrenia Genesis. The Origins of Madness*. New York: W.H. Freeman.

2. Andreasen, N.C. 1995. Symptoms, signs, and diagnosis of schizophrenia. *Lancet* 346(8973), pp. 477-81.

3. Maki, P. et al. 2005. Predictors of schizophrenia--a review. *Br Med Bull* 73-74, pp. 1-15.

4. WHO 1992. *The International Classification of Health Problems, Tenth Revision (ICD-10) - Section V. Mental and Behavioural Disorders*. World Health Organisation, Geneva.

5. Owen, M.J. et al. 2005. Schizophrenia: genes at last? *Trends Genet* 21(9), pp. 518-25.

6. Black, D.W. and R. Fisher 1992. Mortality in DSM-IIIR schizophrenia. *Schizophr Res* 7(2), pp. 109-16.

7. Mueser, K.T. and S.R. McGurk 2004. Schizophrenia. *Lancet* 363(9426), pp. 2063-72.

8. Knapp, M. and S. Kavanagh 1997. Economic outcomes and costs in the treatment of schizophrenia. *Clin Ther* 19(1), pp. 128-38; discussion 126-7.

9. 2001. *US Institute of Medicine, Neurological, psychiatric, and developmental disorders: meeting the challenges of the developing world.* Washington, DC: National Acadamy of Sciences.

10. Kraeplin, E. et al. 1919. *Dementia Praecox and paraphrenia*. Edinburgh: E&S Livingstone.

11. Wing, J.K. and N. Agarawal 2003. *Concepts and Classifications of Schizophrenia*. In: D.R. Weinberger ed. *Schizophrenia*. Second ed. Blackwell Science Ltd., pp. 3-14.

12. Fuller, R., L,M. et al. 2003. *The Symptoms of Schizophrenia*. In: D.R. Weinberger ed. *Schizophrenia*. Second ed. Blackwell Science Ltd., pp. 25-33.

13. Cutting, J. 2003. *Descriptive Psychopathology*. In: D.R. Weinberger ed. *Schizophrenia*. Second ed. Blackwell Science Ltd., pp. 15-24.

14. Dikeos, D.G. et al. 2006. Distribution of symptom dimensions across Kraepelinian divisions. *Br J Psychiatry* 189, pp. 346-53.

15. Jablensky, A. 2003. *The epidemiological horizon.* In: D.R. Weinberger ed. *Schizophrenia.* Second ed. Blackwell Science Ltd., pp. 203-231.

16. Kendell, R.E. et al. 1971. Diagnostic criteria of American and British psychiatrists. *Arch Gen Psychiatry* 25(2), pp. 123-30.

17. Wing, J.K. et al. 1974. *The Measurement and classification of psychiatric symptoms : an instruction manual for the PSE and Catego program* Cambridge, UK: Cambridge University Press.

18. Peralta, V. and M.J. Cuesta 2003. The nosology of psychotic disorders: a comparison among competing classification systems. *Schizophr Bull* 29(3), pp. 413-25.

19. McGrath, J. et al. 2004. A systematic review of the incidence of schizophrenia: the distribution of rates and the influence of sex, urbanicity, migrant status and methodology. *BMC Med* 2, p. 13.

20. 1979. *World Heath Organisation, Schizophrenia: an international follow-up study.* Chichester: John Wiley & Sons.

21. Saha, S. et al. 2005. A systematic review of the prevalence of schizophrenia. *PLoS Med* 2(5), p. e141.

22. Wyatt, R.J. et al. 1988. Schizophrenia, just the facts. What do we know, how well do we know it? *Schizophr Res* 1(1), pp. 3-18.

23. Seeman, M.V. 1981. Outpatient groups for schizophrenia--ensuring attendance. *Can J Psychiatry* 26(1), pp. 32-7.

24. Murray, R.M. and J. Van Os 1998. Predictors of outcome in schizophrenia. *J Clin Psychopharmacol* 18(2 Suppl 1), pp. 2S-4S.

25. Angermeyer, M.C. et al. 1990. Gender and the course of schizophrenia: differences in treated outcomes. *Schizophr Bull* 16(2), pp. 293-307.

26. Aleman, A. et al. 2003. Sex differences in the risk of schizophrenia: evidence from meta-analysis. *Arch Gen Psychiatry* 60(6), pp. 565-71.

27. McGrath, J.J. and R.M. Murray 2003. *Risk factors for schizophrenia: from conception to birth.* In: D.R. Weinberger ed. *Schizophrenia.* Second ed. Blackwell Science Ltd., pp. 232-250.

28. Boydell, J. et al. 2003. Incidence of schizophrenia in south-east London between 1965 and 1997. *Br J Psychiatry* 182, pp. 45-9.

29. Tandon, R. et al. 2008. Schizophrenia, "just the facts" what we know in 2008. 2. Epidemiology and etiology. *Schizophr Res* 102(1-3), pp. 1-18.

30. McGuffin, P. et al. 1994. The strength of the genetic effect. Is there room for an environmental influence in the aetiology of schizophrenia? *Br J Psychiatry* 164(5), pp. 593-9.

31. 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(1), pp. 12-7.

32. Cannon, T.D. et al. 1998. The genetic epidemiology of schizophrenia in a Finnish twin cohort. A population-based modeling study. *Arch Gen Psychiatry* 55(1), pp. 67-74.

33. Cardno, A.G. et al. 1999. Heritability estimates for psychotic disorders: the Maudsley twin psychosis series. *Arch Gen Psychiatry* 56(2), pp. 162-8.

34. Franzek, E. and H. Beckmann 1998. Different genetic background of schizophrenia spectrum psychoses: a twin study. *Am J Psychiatry* 155(1), pp. 76-83.

35. Klaning, U. et al. 1996. Increased occurrence of schizophrenia and other psychiatric illnesses among twins. *Br J Psychiatry* 168(6), pp. 688-92.

36. Tsujita, T. et al., *Twin concordence rates of DSM-IIIR schizophrenia in a new Japanese sample*, in *Abstracts of the 7th International congress of twin studies*. 1992: Tokyo, Japan.

37. Heston, L.L. 1966. Psychiatric disorders in foster home reared children of schizophrenic mothers. *Br J Psychiatry* 112(489), pp. 819-25.

38. Kety, S.S. et al. 1976. Mental illness in the biological and adoptive families of adopted individuals who have become schizophrenic. *Behav Genet* 6(3), pp. 219-25.

39. Rosenthal, D. et al. 1971. The adopted-away offspring of schizophrenics. *Am J Psychiatry* 128(3), pp. 307-11.

40. Wender, P.H. et al. 1974. Crossfostering. A research strategy for clarifying the role of genetic and experiential factors in the etiology of schizophrenia. *Arch Gen Psychiatry* 30(1), pp. 121-8.

41. Tienari, P. 1991. Interaction between genetic vulnerability and family environment: the Finnish adoptive family study of schizophrenia. *Acta Psychiatr Scand* 84(5), pp. 460-5.

42. Keshavan, M.S. et al. 2008. Schizophrenia, "just the facts": what we know in 2008 Part 3: neurobiology. *Schizophr Res* 106(2-3), pp. 89-107.

43. Glantz, L.A. et al. 2006. Apoptotic mechanisms and the synaptic pathology of schizophrenia. *Schizophr Res* 81(1), pp. 47-63.

44. Wenk, G.L. 2003. Neuropathologic changes in Alzheimer's disease. *J Clin Psychiatry* 64 Suppl 9, pp. 7-10.

45. Harrison, P.J. and D.A. Lewis 2003. *Neuropathology of schizophrenia.* In: D.R. Weinberger ed. *Schizophrenia.* Second ed. Blackwell Science Ltd., pp. 310-325.

46. Harrison, P.J. and D.R. Weinberger 2005. Schizophrenia genes, gene expression, and neuropathology: on the matter of their convergence. *Mol Psychiatry* 10(1), pp. 40-68; image 5.

47. McCarley, R.W. et al. 1999. MRI anatomy of schizophrenia. *Biol Psychiatry* 45(9), pp. 1099-119.

48. Cecil, K.M. et al. 1999. Proton magnetic resonance spectroscopy in the frontal and temporal lobes of neuroleptic naive patients with schizophrenia. *Neuropsychopharmacology* 20(2), pp. 131-40.

49. Moghaddam, B. 2003. Bringing order to the glutamate chaos in schizophrenia. *Neuron* 40(5), pp. 881-4.

50. Carlsson, A. 1978. Does dopamine have a role in schizophrenia? *Biol Psychiatry* 13(1), pp. 3-21.

51. Moghaddam, B.M. and J.H. Krystal 2003. *The neurochemistry of schizophrenia.* In: D.R. Weinberger ed. *Schizophrenia.* Second ed. Blackwell Science Ltd., pp. 349-364.

52. Benes, F.M. and S. Berretta 2001. GABAergic interneurons: implications for understanding schizophrenia and bipolar disorder. *Neuropsychopharmacology* 25(1), pp. 1-27.

53. Lewis, D.A. 2000. GABAergic local circuit neurons and prefrontal cortical dysfunction in schizophrenia. *Brain Res Brain Res Rev* 31(2-3), pp. 270-6.

54. Lewis, D.A. et al. 2005. Cortical inhibitory neurons and schizophrenia. *Nat Rev Neurosci* 6(4), pp. 312-24.

55. Owen, M.J. et al. 2002. *Schizophrenia.* In: P. McGuffin, M.J. Owen, and I. Gottesman eds. *Psychiatric Genetics and Genomics.* Oxford University Press, pp. 247-266.

56. Riley, B. et al. 2003. *Genetics and Schizophrenia.* In: D.R. Weinberger ed. *Schizophrenia.* Second ed. Blackwell Science Ltd., pp. 251-276.

57. Risch, N. 1990. Linkage strategies for genetically complex traits. III. The effect of marker polymorphism on analysis of affected relative pairs. *Am J Hum Genet* 46(2), pp. 242-53.

58. 2007. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* 447(7145), pp. 661-78.

59. Falconer, D.S. 1965. The inheritance of liability to certain diseases, estimated from the incidence among relatives. *Ann. Hum. Genet.* 29, pp. 51-76.

60. Gottesman, II and J. Shields 1967. A polygenic theory of schizophrenia. *Proc Natl Acad Sci U S A* 58(1), pp. 199-205.

61. Sham, P.C. and P. McGuffin 2002. *Linkage and association.* In: P. McGuffin, M.J. Owen, and I. Gottesman eds. *Psychiatric Genetics and Genomics.* Oxford University Press, pp. 55-76.

62. Strachan, T. and A.P. Reed 2003. *Human Molecular Genetics.* Third Edition ed. BIOS Scientific Publishers Ltd.

63. Kirov, G. et al. 2005. Finding schizophrenia genes. *J Clin Invest* 115(6), pp. 1440-8.

64. McGue, M. and Gottesman, II 1989. A single dominant gene still cannot account for the transmission of schizophrenia. *Arch Gen Psychiatry* 46(5), pp. 478-80.

65. McGue, M. et al. 1985. Resolving genetic models for the transmission of schizophrenia. *Genet Epidemiol* 2(1), pp. 99-110.

66. Frankel, W.N. and N.J. Schork 1996. Who's afraid of epistasis? *Nat Genet* 14(4), pp. 371-3.

67. Cordell, H.J. 2002. Epistasis: what it means, what it doesn't mean, and statistical methods to detect it in humans. *Hum Mol Genet* 11(20), pp. 2463-8.

68. Georgieva, L. et al. 2006. Convergent evidence that oligodendrocyte lineage transcription factor 2 (OLIG2) and interacting genes influence susceptibility to schizophrenia. *Proc Natl Acad Sci U S A* 103(33), pp. 12469-74.

69. Caspi, A. et al. 2005. Moderation of the effect of adolescent-onset cannabis use on adult psychosis by a functional polymorphism in the catechol-O-methyltransferase gene: longitudinal evidence of a gene X environment interaction. *Biol Psychiatry* 57(10), pp. 1117-27.

70. Cheng, J.Y. et al. 2008. Meta-regression analysis using latitude as moderator of paternal age related schizophrenia risk: high ambient temperature induced de novo mutations or is it related to the cold? *Schizophr Res* 99(1-3), pp. 71-6.

71. Hanninen, K. et al. 2008. Interleukin-1 beta gene polymorphism and its interactions with neuregulin-1 gene polymorphism are associated with schizophrenia. *Eur Arch Psychiatry Clin Neurosci* 258(1), pp. 10-5.

72. Tienari, P. et al. 2004. Genotype-environment interaction in schizophrenia-spectrum disorder. Long-term follow-up study of Finnish adoptees. *Br J Psychiatry* 184, pp. 216-22.

73. Zammit, S. et al. 2007. Genotype effects of CHRNA7, CNR1 and COMT in schizophrenia: interactions with tobacco and cannabis use. *Br J Psychiatry* 191, pp. 402-7.

74. Morton, N.E. 1955. Sequential tests for the detection of linkage. *Am J Hum Genet* 7(3), pp. 277-318.

75. Kruglyak, L. and E.S. Lander 1995. Complete multipoint sib-pair analysis of qualitative and quantitative traits. *Am J Hum Genet* 57(2), pp. 439-54.

76. Altmuller, J. et al. 2001. Genomewide scans of complex human diseases: true linkage is hard to find. *Am J Hum Genet* 69(5), pp. 936-50.

77. Hirschhorn, J.N. and M.J. Daly 2005. Genome-wide association studies for common diseases and complex traits. *Nat Rev Genet* 6(2), pp. 95-108.

78. Risch, N. and K. Merikangas 1996. The future of genetic studies of complex human diseases. *Science* 273(5281), pp. 1516-7.

79. Bell, G.I. et al. 1984. A polymorphic locus near the human insulin gene is associated with insulin-dependent diabetes mellitus. *Diabetes* 33(2), pp. 176-83.

80. Chiu, K.C. et al. 1993. Glucokinase gene variants in the common form of NIDDM. *Diabetes* 42(4), pp. 579-82.

81. Devlin, B. and N. Risch 1995. A comparison of linkage disequilibrium measures for fine-scale mapping. *Genomics* 29(2), pp. 311-22.

82. Mueller, J.C. 2004. Linkage disequilibrium for different scales and applications. *Brief Bioinform* 5(4), pp. 355-64.

83. Zondervan, K.T. and L.R. Cardon 2004. The complex interplay among factors that influence allelic association. *Nat Rev Genet* 5(2), pp. 89-100.

84. 2003. The International HapMap Project. *Nature* 426(6968), pp. 789-96.

85. Laird, N.M. and C. Lange 2006. Family-based designs in the age of large-scale gene-association studies. *Nat Rev Genet* 7(5), pp. 385-94.

86. Spielman, R.S. et al. 1993. Transmission test for linkage disequilibrium: the insulin gene region and insulin-dependent diabetes mellitus (IDDM). *Am J Hum Genet* 52(3), pp. 506-16.

87. Cardon, L.R. and L.J. Palmer 2003. Population stratification and spurious allelic association. *Lancet* 361(9357), pp. 598-604.

88. Owen, M.J. et al. 1997. Association studies in psychiatric genetics. *Mol Psychiatry* 2(4), pp. 270-3.

89. Salyakina, D. et al. 2005. Evaluation of Nyholt's procedure for multiple testing correction. *Hum Hered* 60(1), pp. 19-25; discussion 61-2.

90. Clark, A.G. et al. 2005. Determinants of the success of whole-genome association testing. *Genome Res* 15(11), pp. 1463-7.

91. Wang, W.Y. et al. 2005. Genome-wide association studies: theoretical and practical concerns. *Nat Rev Genet* 6(2), pp. 109-18.

92. Dupont, W.D. and W.D. Plummer, Jr. 1990. Power and sample size calculations. A review and computer program. *Control Clin Trials* 11(2), pp. 116-28.

93. Reich, D.E. and E.S. Lander 2001. On the allelic spectrum of human disease. *Trends Genet* 17(9), pp. 502-10.

94. Liu, P.Y. et al. 2005. A survey of haplotype variants at several disease candidate genes: the importance of rare variants for complex diseases. *J Med Genet* 42(3), pp. 221-7.

95. McClellan, J.M. et al. 2007. Schizophrenia: a common disease caused by multiple rare alleles. *Br J Psychiatry* 190, pp. 194-9.

96. Pritchard, J.K. 2001. Are rare variants responsible for susceptibility to complex diseases? *Am J Hum Genet* 69(1), pp. 124-37.

97. Terwilliger, J.D. and T. Hiekkalinna 2006. An utter refutation of the "Fundamental Theorem of the HapMap". *Eur J Hum Genet* 14(4), pp. 426-37.

98. Zhu, X. et al. 2005. Haplotypes produced from rare variants in the promoter and coding regions of angiotensinogen contribute to variation in angiotensinogen levels. *Hum Mol Genet* 14(5), pp. 639-43.

99. Pe'er, I. et al. 2008. Estimation of the multiple testing burden for genomewide association studies of nearly all common variants. *Genet Epidemiol* 32(4), pp. 381-5.

100. Stefansson, H. et al. 2002. Neuregulin 1 and susceptibility to schizophrenia. *Am J Hum Genet* 71(4), pp. 877-92.

101. Chumakov, I. et al. 2002. Genetic and physiological data implicating the new human gene G72 and the gene for D-amino acid oxidase in schizophrenia. *Proc Natl Acad Sci U S A* 99(21), pp. 13675-80.

102. Hennah, W. et al. 2003. Haplotype transmission analysis provides evidence of association for DISC1 to schizophrenia and suggests sex-dependent effects. *Hum Mol Genet* 12(23), pp. 3151-9.

103. Li, T. et al. 1996. Preferential transmission of the high activity allele of COMT in schizophrenia. *Psychiatr Genet* 6(3), pp. 131-3.

104. Chowdari, K.V. et al. 2002. Association and linkage analyses of RGS4 polymorphisms in schizophrenia. *Hum Mol Genet* 11(12), pp. 1373-80.

105. 1996. Additional support for schizophrenia linkage on chromosomes 6 and 8: a multicenter study. Schizophrenia Linkage Collaborative Group for Chromosomes 3, 6 and 8. *Am J Med Genet* 67(6), pp. 580-94.

106. Lewis, C.M. et al. 2003. Genome scan meta-analysis of schizophrenia and bipolar disorder, part II: Schizophrenia. *Am J Hum Genet* 73(1), pp. 34-48.

107. Moises, H.W. et al. 1995. An international two-stage genome-wide search for schizophrenia susceptibility genes. *Nat Genet* 11(3), pp. 321-4.

108. Schwab, S.G. et al. 2000. A genome-wide autosomal screen for schizophrenia susceptibility loci in 71 families with affected siblings: support for loci on chromosome 10p and 6. *Mol Psychiatry* 5(6), pp. 638-49.

109. Straub, R.E. et al. 2002. Genome-wide scans of three independent sets of 90 Irish multiplex schizophrenia families and follow-up of selected regions in all families provides evidence for multiple susceptibility genes. *Mol Psychiatry* 7(6), pp. 542-59.

110. Kendler, K.S. et al. 1996. Irish study on high-density schizophrenia families: field methods and power to detect linkage. *Am J Med Genet* 67(2), pp. 179-90.

111. Straub, R.E. et al. 1995. A potential vulnerability locus for schizophrenia on chromosome 6p24-22: evidence for genetic heterogeneity. *Nat Genet* 11(3), pp. 287-93.

112. Straub, R.E. et al. 2002. Genetic variation in the 6p22.3 gene DTNBP1, the human ortholog of the mouse dysbindin gene, is associated with schizophrenia. *Am J Hum Genet* 71(2), pp. 337-48.

113. van den Oord, E.J. et al. 2003. Identification of a high-risk haplotype for the dystrobrevin binding protein 1 (DTNBP1) gene in the Irish study of high-density schizophrenia families. *Mol Psychiatry* 8(5), pp. 499-510.

114. Bakker, S.C. et al. 2007. The PIP5K2A and RGS4 genes are differentially associated with deficit and non-deficit schizophrenia. *Genes Brain Behav* 6(2), pp. 113-9.

115. Datta, S.R. et al. 2007. Failure to confirm allelic and haplotypic association between markers at the chromosome 6p22.3 dystrobrevin-binding protein 1 (DTNBP1) locus and schizophrenia. *Behav Brain Funct* 3, p. 50.

116. De Luca, V. et al. 2005. Untranslated region haplotype in dysbindin gene: analysis in schizophrenia. *J Neural Transm* 112(9), pp. 1263-7.

117. Funke, B. et al. 2004. Association of the DTNBP1 locus with schizophrenia in a U.S. population. *Am J Hum Genet* 75(5), pp. 891-8.

118. Hall, D. et al. 2004. The contribution of three strong candidate schizophrenia susceptibility genes in demographically distinct populations. *Genes Brain Behav* 3(4), pp. 240-8.

119. Holliday, E.G. et al. 2006. Association study of the dystrobrevin-binding gene with schizophrenia in Australian and Indian samples. *Twin Res Hum Genet* 9(4), pp. 531-9.

120. Joo, E.J. et al. 2006. The dysbindin gene (DTNBP1) and schizophrenia: no support for an association in the Korean population. *Neurosci Lett* 407(2), pp. 101-6.

121. Li, T. et al. 2005. Identifying potential risk haplotypes for schizophrenia at the DTNBP1 locus in Han Chinese and Scottish populations. *Mol Psychiatry* 10(11), pp. 1037-44.

122. Liu, C.M. et al. 2007. No association evidence between schizophrenia and dystrobrevin-binding protein 1 (DTNBP1) in Taiwanese families. *Schizophr Res* 93(1-3), pp. 391-8.

123. Morris, D.W. et al. 2003. No evidence for association of the dysbindin gene [DTNBP1] with schizophrenia in an Irish population-based study. *Schizophr Res* 60(2-3), pp. 167-72.

124. Peters, K. et al. 2008. Comprehensive analysis of tagging sequence variants in DTNBP1 shows no association with schizophrenia or with its composite neurocognitive endophenotypes. *Am J Med Genet B Neuropsychiatr Genet* 147B(7), pp. 1159-1166.

125. Sanders, A.R. et al. 2008. No Significant Association of 14 Candidate Genes With Schizophrenia in a Large European Ancestry Sample: Implications for Psychiatric Genetics. *Am J Psychiatry.*

126. Turunen, J.A. et al. 2007. The role of DTNBP1, NRG1, and AKT1 in the genetics of schizophrenia in Finland. *Schizophr Res* 91(1-3), pp. 27-36.

127. Van Den Bogaert, A. et al. 2003. The DTNBP1 (dysbindin) gene contributes to schizophrenia, depending on family history of the disease. *Am J Hum Genet* 73(6), pp. 1438-43.

128. Wood, L.S. et al. 2006. Significant Support for DAO as a Schizophrenia Susceptibility Locus: Examination of Five Genes Putatively Associated with Schizophrenia. *Biol Psychiatry.*

129. Duan, J. et al. 2007. DTNBP1 (Dystrobrevin Binding Protein 1) and Schizophrenia: Association Evidence in the 3' End of the Gene. *Hum Hered* 64(2), pp. 97-106.

130. Kirov, G. et al. 2004. Strong evidence for association between the dystrobrevin binding protein 1 gene (DTNBP1) and schizophrenia in 488 parent-offspring trios from Bulgaria. *Biol Psychiatry* 55(10), pp. 971-5.

131. Schwab, S.G. et al. 2003. Support for association of schizophrenia with genetic variation in the 6p22.3 gene, dysbindin, in sib-pair families with linkage and in an additional sample of triad families. *Am J Hum Genet* 72(1), pp. 185-90.

132. Tosato, S. et al. 2007. Association study of dysbindin gene with clinical and outcome measures in a representative cohort of Italian schizophrenic patients. *Am J Med Genet B Neuropsychiatr Genet.*

133. Vilella, E. et al. 2007. Association of schizophrenia with DTNBP1 but not with DAO, DAOA, NRG1 and RGS4 nor their genetic interaction. *J Psychiatr Res.*

134. Williams, N.M. et al. 2004. Identification in 2 independent samples of a novel schizophrenia risk haplotype of the dystrobrevin binding protein gene (DTNBP1). *Arch Gen Psychiatry* 61(4), pp. 336-44.

135. Numakawa, T. et al. 2004. Evidence of novel neuronal functions of dysbindin, a susceptibility gene for schizophrenia. *Hum Mol Genet* 13(21), pp. 2699-708.

136. Pae, C.U. et al. 2009. Dysbindin gene (DTNBP1) and schizophrenia in Korean population. *Eur Arch Psychiatry Clin Neurosci.*

137. Tang, J.X. et al. 2003. Family-based association study of DTNBP1 in 6p22.3 and schizophrenia. *Mol Psychiatry* 8(8), pp. 717-8.

138. Tochigi, M. et al. 2006. Association study of the dysbindin (DTNBP1) gene in schizophrenia from the Japanese population. *Neurosci Res* 56(2), pp. 154-8.

139. Fallin, M.D. et al. 2005. Bipolar I disorder and schizophrenia: a 440-single-nucleotide polymorphism screen of 64 candidate genes among Ashkenazi Jewish case-parent trios. *Am J Hum Genet* 77(6), pp. 918-36.

140. Allen, N.C. et al. 2008. Systematic meta-analyses and field synopsis of genetic association studies in schizophrenia: the SzGene database. *Nat Genet* 40(7), pp. 827-34.

141. Mutsuddi, M. et al. 2006. Analysis of high-resolution HapMap of DTNBP1 (Dysbindin) suggests no consistency between reported common variant associations and schizophrenia. *Am J Hum Genet* 79(5), pp. 903-9.

142. Li, D. and L. He 2007. Association study between the dystrobrevin binding protein 1 gene (DTNBP1) and schizophrenia: a meta-analysis. *Schizophr Res* 96(1-3), pp. 112-8.

143. Kirov, G. et al. 2009. A genome-wide association study in 574 schizophrenia trios using DNA pooling. *Mol Psychiatry* 14(8), pp. 796-803.

144. Donohoe, G. et al. 2007. Variance in neurocognitive performance is associated with dysbindin-1 in schizophrenia: A preliminary study. *Neuropsychologia* 45(2), pp. 454-8.

145. Donohoe, G. et al. 2006. Are deficits in executive sub-processes simply reflecting more general cognitive decline in schizophrenia? *Schizophr Res* 85(1-3), pp. 168-73.

146. Burdick, K.E. et al. 2006. Genetic variation in DTNBP1 influences general cognitive ability. *Hum Mol Genet* 15(10), pp. 1563-8.

147. Burdick, K.E. et al. 2007. DTNBP1 genotype influences cognitive decline in schizophrenia. *Schizophr Res* 89(1-3), pp. 169-72.

148. Zinkstok, J.R. et al. 2007. Association between the DTNBP1 gene and intelligence: a case-control study in young patients with schizophrenia and related disorders and unaffected siblings. *Behav Brain Funct* 3, p. 19.

149. Fallgatter, A.J. et al. 2006. DTNBP1 (dysbindin) gene variants modulate prefrontal brain function in healthy individuals. *Neuropsychopharmacology* 31(9), pp. 2002-10.

150. Donohoe, G. et al. 2008. Early visual processing deficits in dysbindin-associated schizophrenia. *Biol Psychiatry* 63(5), pp. 484-9.

151. Kircher, T. et al. 2009. Association of the DTNBP1 genotype with cognition and personality traits in healthy subjects. *Psychol Med*, pp. 1-9.

152. Raine, A. and D. Benishay 1995. The SPQ-B: a brief screening instrument for schizotypal personality disorder. *Journal of Personality Disorder* 9, pp. 346-355.

153. DeRosse, P. et al. 2006. Dysbindin genotype and negative symptoms in schizophrenia. *Am J Psychiatry* 163(3), pp. 532-4.

154. Fanous, A.H. et al. 2005. Relationship between a high-risk haplotype in the DTNBP1 (dysbindin) gene and clinical features of schizophrenia. *Am J Psychiatry* 162(10), pp. 1824-32.

155. Pae, C.U. et al. 2008. DTNBP1 haplotype influences baseline assessment scores of schizophrenic in-patients. *Neurosci Lett* 440(2), pp. 150-4.

156. Stefanis, N.C. et al. 2007. Impact of schizophrenia candidate genes on schizotypy and cognitive endophenotypes at the population level. *Biol Psychiatry* 62(7), pp. 784-92.

157. Corvin, A. et al. 2008. A dysbindin risk haplotype associated with less severe manic-type symptoms in psychosis. *Neurosci Lett* 431(2), pp. 146-9.

158. Gornick, M.C. et al. 2005. Dysbindin (DTNBP1, 6p22.3) is associated with childhood-onset psychosis and endophenotypes measured by the Premorbid Adjustment Scale (PAS). *J Autism Dev Disord* 35(6), pp. 831-8.

159. Kishimoto, M. et al. 2008. The dysbindin gene (DTNBP1) is associated with methamphetamine psychosis. *Biol Psychiatry* 63(2), pp. 191-6.

160. Ujike, H. 2002. Stimulant-induced psychosis and schizophrenia: the role of sensitization. *Curr Psychiatry Rep* 4(3), pp. 177-84.

161. Lindenmayer, J.P. et al. 2004. An excitement subscale of the Positive and Negative Syndrome Scale. *Schizophr Res* 68(2-3), pp. 331-7.

162. Craddock, N. et al. 2006. Genes for schizophrenia and bipolar disorder? Implications for psychiatric nosology. *Schizophr Bull* 32(1), pp. 9-16.

163. Luciano, M. et al. 2009. Variation in the dysbindin gene and normal cognitive function in three independent population samples. *Genes Brain Behav* 8(2), pp. 218-27.

164. Pantelis, C. et al. 2004. Relationship of behavioural and symptomatic syndromes in schizophrenia to spatial working memory and attentional set-shifting ability. *Psychol Med* 34(4), pp. 693-703.

165. Breen, G. et al. 2006. Association of the dysbindin gene with bipolar affective disorder. *Am J Psychiatry* 163(9), pp. 1636-8.

166. Gaysina, D. et al. 2008. Association of the dystrobrevin binding protein 1 gene (DTNBP1) in a bipolar case-control study (BACCS). *Am J Med Genet B Neuropsychiatr Genet.*

167. Joo, E.J. et al. 2007. Dysbindin gene variants are associated with bipolar I disorder in a Korean population. *Neurosci Lett* 418(3), pp. 272-5.

168. Pae, C.U. et al. 2007. Effect of 5-haplotype of dysbindin gene (DTNBP1) polymorphisms for the susceptibility to bipolar I disorder. *Am J Med Genet B Neuropsychiatr Genet* 144B(5), pp. 701-3.

169. Raybould, R. et al. 2005. Bipolar disorder and polymorphisms in the dysbindin gene (DTNBP1). *Biol Psychiatry* 57(7), pp. 696-701.

170. Perlis, R.H. et al. 2008. Family-based association study of lithium-related and other candidate genes in bipolar disorder. *Arch Gen Psychiatry* 65(1), pp. 53-61.

171. Arias, B. et al. 2009. Dysbindin gene (DTNBP1) in major depression: association with clinical response to selective serotonin reuptake inhibitors. *Pharmacogenet Genomics* 19(2), pp. 121-8.

172. Kim, J.J. et al. 2008. Is there protective haplotype of dysbindin gene (DTNBP1) 3 polymorphisms for major depressive disorder. *Prog Neuropsychopharmacol Biol Psychiatry* 32(2), pp. 375-9.

173. Pae, C.U. et al. 2007. Dysbindin associated with selective serotonin reuptake inhibitor antidepressant efficacy. *Pharmacogenet Genomics* 17(1), pp. 69-75.

174. Wray, N.R. et al. 2008. Association study of candidate variants from brain-derived neurotrophic factor and dystrobrevin-binding protein 1 with neuroticism, anxiety, and depression. *Psychiatr Genet* 18(5), pp. 219-25.

175. Zill, P. et al. 2004. The dysbindin gene in major depression: an association study. *Am J Med Genet B Neuropsychiatr Genet* 129B(1), pp. 55-8.

176. Horikawa, Y. et al. 2000. Genetic variation in the gene encoding calpain-10 is associated with type 2 diabetes mellitus. *Nat Genet* 26(2), pp. 163-75.

177. Rioux, J.D. et al. 2001. Genetic variation in the 5q31 cytokine gene cluster confers susceptibility to Crohn disease. *Nat Genet* 29(2), pp. 223-8.

178. Suzuki, A. et al. 2003. Functional haplotypes of PADI4, encoding citrullinating enzyme peptidylarginine deiminase 4, are associated with rheumatoid arthritis. *Nat Genet* 34(4), pp. 395-402.

179. Ueda, H. et al. 2003. Association of the T-cell regulatory gene CTLA4 with susceptibility to autoimmune disease. *Nature* 423(6939), pp. 506-11.

180. Chagnon, Y.C. et al. 2008. Differential RNA expression between schizophrenic patients and controls of the dystrobrevin binding protein 1 and neuregulin 1 genes in immortalized lymphocytes. *Schizophr Res* 100(1-3), pp. 281-90.

181. Talbot, K. et al. 2004. Dysbindin-1 is reduced in intrinsic, glutamatergic terminals of the hippocampal formation in schizophrenia. *J Clin Invest* 113(9), pp. 1353-63.

182. Weickert, C.S. et al. 2008. Reduced DTNBP1 (dysbindin-1) mRNA in the hippocampal formation of schizophrenia patients. *Schizophr Res* 98(1-3), pp. 105-10.

183. Weickert, C.S. et al. 2004. Human dysbindin (DTNBP1) gene expression in normal brain and in schizophrenic prefrontal cortex and midbrain. *Arch Gen Psychiatry* 61(6), pp. 544-55.

184. Bray, N.J. et al. 2003. Cis-acting variation in the expression of a high proportion of genes in human brain. *Hum Genet* 113(2), pp. 149-53.

185. Bray, N.J. et al. 2008. Cis- and Trans- Loci Influence Expression of the Schizophrenia Susceptibility Gene DTNBP1. *Hum Mol Genet.*

186. Bray, N.J. et al. 2005. Haplotypes at the dystrobrevin binding protein 1 (DTNBP1) gene locus mediate risk for schizophrenia through reduced DTNBP1 expression. *Hum Mol Genet* 14(14), pp. 1947-54.

187. Weickert, T.W. et al. 2000. Cognitive impairments in patients with schizophrenia displaying preserved and compromised intellect. *Arch Gen Psychiatry* 57(9), pp. 907-13.

188. Weinberger, D.R. and K.F. Berman 1996. Prefrontal function in schizophrenia: confounds and controversies. *Philos Trans R Soc Lond B Biol Sci* 351(1346), pp. 1495-503.

189. Harrison, P.J. and S.L. Eastwood 2001. Neuropathological studies of synaptic connectivity in the hippocampal formation in schizophrenia. *Hippocampus* 11(5), pp. 508-19.

190. Schmajuk, N.A. 2001. Hippocampal dysfunction in schizophrenia. *Hippocampus* 11(5), pp. 599-613.

191. Levine, M. and R. Tjian 2003. Transcription regulation and animal diversity. *Nature* 424(6945), pp. 147-51.

192. Villard, J. 2004. Transcription regulation and human diseases. *Swiss Med Wkly* 134(39-40), pp. 571-9.

193. Bray, N.J. and M.C. O'Donovan 2006. Investigating cis-acting regulatory variation using assays of relative allelic expression. *Psychiatr Genet* 16(4), pp. 173-7.

194. Bray, N.J. et al. 2003. A haplotype implicated in schizophrenia susceptibility is associated with reduced COMT expression in human brain. *Am J Hum Genet* 73(1), pp. 152-61.

195. Bray, N.J. et al. 2004. Allelic expression of APOE in human brain: effects of epsilon status and promoter haplotypes. *Hum Mol Genet* 13(22), pp. 2885-92.

196. Benson, M.A. et al. 2001. Dysbindin, a novel coiled-coil-containing protein that interacts with the dystrobrevins in muscle and brain. *J Biol Chem* 276(26), pp. 24232-41.

197. Pruitt, K.D. et al. 2005. NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res* 33(Database issue), pp. D501-4.

198. Thierry-Mieg, D. and J. Thierry-Mieg 2006. AceView: a comprehensive cDNA-supported gene and transcripts annotation. *Genome Biol* 7 Suppl 1, pp. S12 1-14.

199. Tang, J. et al. 2009. Dysbindin-1 in dorsolateral prefrontal cortex of schizophrenia cases is reduced in an isoform-specific manner unrelated to altered dysbindin-1 gene expression *Nature Proceedings*.

200. Strauss, H.M. and S. Keller 2008. Pharmacological interference with protein-protein interactions mediated by coiled-coil motifs. *Handb Exp Pharmacol* (186), pp. 461-82.

201. Sillitoe, R.V. et al. 2003. Abnormal dysbindin expression in cerebellar mossy fiber synapses in the mdx mouse model of Duchenne muscular dystrophy. *J Neurosci* 23(16), pp. 6576-85.

202. Rees, M.L. et al. 2007. Dystrobrevins in muscle and non-muscle tissues. *Neuromuscul Disord.*

203. Blake, D.J. et al. 2002. Function and genetics of dystrophin and dystrophin-related proteins in muscle. *Physiol Rev* 82(2), pp. 291-329.

204. Bresolin, N. et al. 1994. Cognitive impairment in Duchenne muscular dystrophy. *Neuromuscul Disord* 4(4), pp. 359-69.

205. Hinton, V.J. et al. 2000. Poor verbal working memory across intellectual level in boys with Duchenne dystrophy. *Neurology* 54(11), pp. 2127-32.

206. Mehler, M.F. 2000. Brain dystrophin, neurogenetics and mental retardation. *Brain Res Brain Res Rev* 32(1), pp. 277-307.

207. Straub, R.E. et al. 2005. Muted, a protein that binds to dysbindin (dtnbp1), is associated with schizophrenia. *Am J Med Genet B Neuropsychiatr Genet* 138(1), pp. 3-147.

208. Li, W. et al. 2003. Hermansky-Pudlak syndrome type 7 (HPS-7) results from mutant dysbindin, a member of the biogenesis of lysosome-related organelles complex 1 (BLOC-1). *Nat Genet* 35(1), pp. 84-9.

209. Ciciotte, S.L. et al. 2003. Cappuccino, a mouse model of Hermansky-Pudlak syndrome, encodes a novel protein that is part of the pallidin-muted complex (BLOC-1). *Blood* 101(11), pp. 4402-7.

210. Falcon-Perez, J.M. et al. 2002. BLOC-1, a novel complex containing the pallidin and muted proteins involved in the biogenesis of melanosomes and platelet-dense granules. *J Biol Chem* 277(31), pp. 28191-9.

211. Starcevic, M. and E.C. Dell'Angelica 2004. Identification of snapin and three novel proteins (BLOS1, BLOS2, and BLOS3/reduced pigmentation) as subunits of biogenesis of lysosome-related organelles complex-1 (BLOC-1). *J Biol Chem* 279(27), pp. 28393-401.

212. Nazarian, R. et al. 2006. Reinvestigation of the dysbindin subunit of BLOC-1 (biogenesis of lysosome-related organelles complex-1) as a dystrobrevin-binding protein. *Biochem J* 395(3), pp. 587-98.

213. Huizing, M. et al. 2000. Hermansky-Pudlak syndrome and related disorders of organelle formation. *Traffic* 1(11), pp. 823-35.

214. Dell'Angelica, E.C. et al. 2000. Lysosome-related organelles. *Faseb J* 14(10), pp. 1265-78.

215. Lyerla, T.A. et al. 2003. Aberrant lung structure, composition, and function in a murine model of Hermansky-Pudlak syndrome. *Am J Physiol Lung Cell Mol Physiol* 285(3), pp. L643-53.

216. Swank, R.T. et al. 1998. Mouse models of Hermansky Pudlak syndrome: a review. *Pigment Cell Res* 11(2), pp. 60-80.

217. Zhang, Q. et al. 2003. Ru2 and Ru encode mouse orthologs of the genes mutated in human Hermansky-Pudlak syndrome types 5 and 6. *Nat Genet* 33(2), pp. 145-53.

218. Salazar, G. et al. 2006. BLOC-1 complex deficiency alters the targeting of adaptor protein complex-3 cargoes. *Mol Biol Cell* 17(9), pp. 4014-26.

219. Di Pietro, S.M. et al. 2006. BLOC-1 interacts with BLOC-2 and the AP-3 complex to facilitate protein trafficking on endosomes. *Mol Biol Cell* 17(9), pp. 4027-38.

220. Ilardi, J.M. et al. 1999. Snapin: a SNARE-associated protein implicated in synaptic transmission. *Nat Neurosci* 2(2), pp. 119-24.

221. Tian, J.H. et al. 2005. The role of Snapin in neurosecretion: snapin knock-out mice exhibit impaired calcium-dependent exocytosis of large dense-core vesicles in chromaffin cells. *J Neurosci* 25(45), pp. 10546-55.

222. Bhardwaj, S.K. et al. 2009. Behavioral characterization of dysbindin-1 deficient sandy mice. *Behav Brain Res* 197(2), pp. 435-41.

223. Cox, M.M. et al. 2009. Neurobehavioral abnormalities in the dysbindin-1 mutant, sandy, on a C57BL/6J genetic background. *Genes Brain Behav.*

224. Feng, Y.Q. et al. 2008. Dysbindin deficiency in sandy mice causes reduction of snapin and displays behaviors related to schizophrenia. *Schizophr Res* 106(2-3), pp. 218-28.

225. Hattori, S. et al. 2008. Behavioral abnormalities and dopamine reductions in sdy mutant mice with a deletion in Dtnbp1, a susceptibility gene for schizophrenia. *Biochem Biophys Res Commun* 373(2), pp. 298-302.

226. Takao, K. et al. 2008. Impaired long-term memory retention and working memory in sdy mutant mice with a deletion in Dtnbp1, a susceptibility gene for schizophrenia. *Mol Brain* 1(1), p. 11.

227. Chen, X.W. et al. 2008. DTNBP1, a schizophrenia susceptibility gene, affects kinetics of transmitter release. *J Cell Biol* 181(5), pp. 791-801.

228. Murotani, T. et al. 2007. High dopamine turnover in the brains of Sandy mice. *Neurosci Lett* 421(1), pp. 47-51.

229. Iizuka, Y. et al. 2007. Evidence that the BLOC-1 protein dysbindin modulates dopamine D2 receptor internalization and signaling but not D1 internalization. *J Neurosci* 27(45), pp. 12390-5.

230. Kubota, K. et al. 2009. Dysbindin engages in c-Jun N-terminal kinase activity and cytoskeletal organization. *Biochem Biophys Res Commun* 379(2), pp. 191-5.

231. Kumamoto, N. et al. 2006. Hyperactivation of midbrain dopaminergic system in schizophrenia could be attributed to the down-regulation of dysbindin. *Biochem Biophys Res Commun* 345(2), pp. 904-9.

232. D'Hooge, R. and P.P. De Deyn 2001. Applications of the Morris water maze in the study of learning and memory. *Brain Res Brain Res Rev* 36(1), pp. 60-90.

233. Kesner, R.P. 2007. A behavioral analysis of dentate gyrus function. *Prog Brain Res* 163, pp. 567-76.

234. Kubik, S. et al. 2007. Using immediate-early genes to map hippocampal subregional functions. *Learn Mem* 14(11), pp. 758-70.

235. Prior, H. et al. 1997. Dissociation of spatial reference memory, spatial working memory, and hippocampal mossy fiber distribution in two rat strains differing in emotionality. *Behav Brain Res* 87(2), pp. 183-94.

236. Schwegler, H. et al. 1990. Hippocampal mossy fibers and radial-maze learning in the mouse: a correlation with spatial working memory but not with non-spatial reference memory. *Neuroscience* 34(2), pp. 293-8.

237. Talbot, K. et al. 2006. Dysbindin-1 is a synaptic and microtubular protein that binds brain snapin. *Hum Mol Genet* 15(20), pp. 3041-54.

238. Sudhof, T.C. 2004. The synaptic vesicle cycle. *Annu Rev Neurosci* 27, pp. 509-47.

239. Coyle, J.T. 2006. Glutamate and schizophrenia: beyond the dopamine hypothesis. *Cell Mol Neurobiol* 26(4-6), pp. 365-84.

240. Sealfon, S.C. and C.W. Olanow 2000. Dopamine receptors: from structure to behavior. *Trends Neurosci* 23(10 Suppl), pp. S34-40.

241. Kaneko, T. et al. 2002. Immunohistochemical localization of candidates for vesicular glutamate transporters in the rat brain. *J Comp Neurol* 444(1), pp. 39-62.

242. Glickstein, S.B. et al. 2002. Mice lacking dopamine D2 and D3 receptors have spatial working memory deficits. *J Neurosci* 22(13), pp. 5619-29.

243. Wing, J.K. et al. 1990. SCAN. Schedules for Clinical Assessment in Neuropsychiatry. *Arch Gen Psychiatry* 47(6), pp. 589-93.

244. Moskvina, V. et al. 2005. Design of case-controls studies with unscreened controls. *Ann Hum Genet* 69(Pt 5), pp. 566-76.

245. Liu, W. et al. 1998. Denaturing high performance liquid chromatography (DHPLC) used in the detection of germline and somatic mutations. *Nucleic Acids Res* 26(6), pp. 1396-400.

246. Ririe, K.M. et al. 1997. Product differentiation by analysis of DNA melting curves during the polymerase chain reaction. *Anal Biochem* 245(2), pp. 154-60.

247. Graham, R. et al. 2005. Distinguishing different DNA heterozygotes by high-resolution melting. *Clin Chem* 51(7), pp. 1295-8.

248. Wittwer, C.T. et al. 2003. High-resolution genotyping by amplicon melting analysis using LCGreen. *Clin Chem* 49(6 Pt 1), pp. 853-60.

249. Reed, G.H. and C.T. Wittwer 2004. Sensitivity and specificity of single-nucleotide polymorphism scanning by high-resolution melting analysis. *Clin Chem* 50(10), pp. 1748-54.

250. Dwyer, S.C., L; Mantripagada, K; Owen, MJ; O'Donovan, MC; Williams, NM. 2009. Mutation screening of the DTNBP1 exonic sequence in 669 schizophrenics and 710 controls. *Neuropsychiatric Genetics*, p. Manuscript acepted.

251. Weckx, S. et al. 2005. novoSNP, a novel computational tool for sequence variation discovery. *Genome Res* 15(3), pp. 436-42.

252. Buetow, K.H. et al. 2001. High-throughput development and characterization of a genomewide collection of gene-based single nucleotide polymorphism

markers by chip-based matrix-assisted laser desorption/ionization time-of-flight mass spectrometry. *Proc Natl Acad Sci U S A* 98(2), pp. 581-4.

253. Norton, N. et al. 2002. Universal, robust, highly quantitative SNP allele frequency measurement in DNA pools. *Hum Genet* 110(5), pp. 471-8.

254. Yan, H. et al. 2002. Allelic variation in human gene expression. *Science* 297(5584), p. 1143.

255. Bray, N.J. et al. 2004. The serotonin-2A receptor gene locus does not contain common polymorphism affecting mRNA levels in adult brain. *Mol Psychiatry* 9(1), pp. 109-14.

256. Purcell, S. et al. 2007. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet* 81(3), pp. 559-75.

257. Barrett, J.C. et al. 2005. Haploview: analysis and visualization of LD and haplotype maps. *Bioinformatics* 21(2), pp. 263-5.

258. Liao, H.M. and C.H. Chen 2004. Mutation analysis of the human dystrobrevin-binding protein 1 gene in schizophrenic patients. *Schizophr Res* 71(1), pp. 185-9.

259. Maston, G.A. et al. 2006. Transcriptional Regulatory Elements in the Human Genome. *Annu Rev Genomics Hum Genet* 7, pp. 29-59.

260. Gershenzon, N.I. and I.P. Ioshikhes 2005. Synergy of human Pol II core promoter elements revealed by statistical sequence analysis. *Bioinformatics* 21(8), pp. 1295-300.

261. Strachan, T. and A.P. Read 2004. *Human Molecular Genetics*.3rd ed. New York : Garland Press.

262. Calhoun, V.C. et al. 2002. Promoter-proximal tethering elements regulate enhancer-promoter specificity in the Drosophila Antennapedia complex. *Proc Natl Acad Sci U S A* 99(14), pp. 9243-7.

263. Cleard, F. et al. 2006. Probing long-distance regulatory interactions in the Drosophila melanogaster bithorax complex using Dam identification. *Nat Genet* 38(8), pp. 931-5.

264. Su, W. et al. 1991. DNA looping between sites for transcriptional activation: self-association of DNA-bound Sp1. *Genes Dev* 5(5), pp. 820-6.

265. Lettice, L.A. et al. 2003. A long-range Shh enhancer regulates expression in the developing limb and fin and is associated with preaxial polydactyly. *Hum Mol Genet* 12(14), pp. 1725-35.

266. Nobrega, M.A. et al. 2003. Scanning human gene deserts for long-range enhancers. *Science* 302(5644), p. 413.

267. Li, L. et al. 2004. Gene regulation by Sp1 and Sp3. *Biochem Cell Biol* 82(4), pp. 460-71.

268. Harris, M.B. et al. 2005. Repression of an interleukin-4-responsive promoter requires cooperative BCL-6 function. *J Biol Chem* 280(13), pp. 13114-21.

269. Celniker, S.E. and R.A. Drewell 2007. Chromatin looping mediates boundary element promoter interactions. *Bioessays* 29(1), pp. 7-10.

270. Szutorisz, H. et al. 2005. The role of enhancers as centres for general transcription factor recruitment. *Trends Biochem Sci* 30(11), pp. 593-9.

271. Butler, J.E. and J.T. Kadonaga 2001. Enhancer-promoter specificity mediated by DPE or TATA core promoter motifs. *Genes Dev* 15(19), pp. 2515-9.

272. Gaszner, M. and G. Felsenfeld 2006. Insulators: exploiting transcriptional and epigenetic mechanisms. *Nat Rev Genet* 7(9), pp. 703-13.

273. West, A.G. and P. Fraser 2005. Remote control of gene transcription. *Hum Mol Genet* 14 Spec No 1, pp. R101-11.

274. Chorley, B.N. et al. 2008. Discovery and verification of functional single nucleotide polymorphisms in regulatory genomic regions: current and developing technologies. *Mutat Res* 659(1-2), pp. 147-57.

275. Knight, J.C. 2003. Functional implications of genetic variation in non-coding DNA for disease susceptibility and gene regulation. *Clin Sci (Lond)* 104(5), pp. 493-501.

276. Dedon, P.C. et al. 1991. A simplified formaldehyde fixation and immunoprecipitation technique for studying protein-DNA interactions. *Anal Biochem* 197(1), pp. 83-90.

277. Gross, D.S. and W.T. Garrard 1988. Nuclease hypersensitive sites in chromatin. *Annu Rev Biochem* 57, pp. 159-97.

278. Wu, C. et al. 1979. The chromatin structure of specific genes: II. Disruption of chromatin structure during gene activity. *Cell* 16(4), pp. 807-14.

279. Buckland, P.R. et al. 2005. Strong bias in the location of functional promoter polymorphisms. *Hum Mutat* 26(3), pp. 214-23.

280. Bird, A.P. et al. 1987. Non-methylated CpG-rich islands at the human alpha-globin locus: implications for evolution of the alpha-globin pseudogene. *Embo J* 6(4), pp. 999-1004.

281. Gardiner-Garden, M. and M. Frommer 1987. CpG islands in vertebrate genomes. *J Mol Biol* 196(2), pp. 261-82.

282. Kadonaga, J.T. 2004. Regulation of RNA polymerase II transcription by sequence-specific DNA binding factors. *Cell* 116(2), pp. 247-57.

283. Carey, M. et al. 1990. A mechanism for synergistic activation of a mammalian gene by GAL4 derivatives. *Nature* 345(6273), pp. 361-4.

284. Li, Q. et al. 2002. Locus control regions. *Blood* 100(9), pp. 3077-86.

285. Aronow, B.J. et al. 1992. Functional analysis of the human adenosine deaminase gene thymic regulatory region and its ability to generate position-independent transgene expression. *Mol Cell Biol* 12(9), pp. 4170-85.

286. Lang, G. et al. 1991. Deletion analysis of the human CD2 gene locus control region in transgenic mice. *Nucleic Acids Res* 19(21), pp. 5851-6.

287. Lee, G.R. et al. 2003. Regulation of the Th2 cytokine locus by a locus control region. *Immunity* 19(1), pp. 145-53.

288. Adlam, M. and G. Siu 2003. Hierarchical interactions control CD4 gene expression during thymocyte development. *Immunity* 18(2), pp. 173-84.

289. Berman, B.P. et al. 2002. Exploiting transcription factor binding site clustering to identify cis-regulatory modules involved in pattern formation in the Drosophila genome. *Proc Natl Acad Sci U S A* 99(2), pp. 757-62.

290. Frith, M.C. et al. 2001. Detection of cis-element clusters in higher eukaryotic DNA. *Bioinformatics* 17(10), pp. 878-89.

291. Markstein, M. et al. 2002. Genome-wide analysis of clustered Dorsal binding sites identifies putative target genes in the Drosophila embryo. *Proc Natl Acad Sci U S A* 99(2), pp. 763-8.

292. Wasserman, W.W. and J.W. Fickett 1998. Identification of regulatory regions which confer muscle-specific gene expression. *J Mol Biol* 278(1), pp. 167-81.

293. Consortium, E.P. 2007. Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature* 447(7146), pp. 799-816.

294. Dermitzakis, E.T. et al. 2005. Conserved non-genic sequences - an unexpected feature of mammalian genomes. *Nat Rev Genet* 6(2), pp. 151-7.

295. Culi, J. and J. Modolell 1998. Proneural gene self-stimulation in neural precursors: an essential mechanism for sense organ development that is regulated by Notch signaling. *Genes Dev* 12(13), pp. 2036-47.

296. Martinez-Cruzado, J.C. et al. 1988. Evolution of the autosomal chorion locus in Drosophila. I. General organization of the locus and sequence comparisons of genes s15 and s19 in evolutionary distant species. *Genetics* 119(3), pp. 663-77.

297. Dermitzakis, E.T. et al. 2002. Numerous potentially functional but non-genic conserved sequences on human chromosome 21. *Nature* 420(6915), pp. 578-82.

298. DeSilva, U. et al. 2002. Generation and comparative analysis of approximately 3.3 Mb of mouse genomic sequence orthologous to the region of human chromosome 7q11.23 implicated in Williams syndrome. *Genome Res* 12(1), pp. 3-15.

299. Duret, L. et al. 1993. Strong conservation of non-coding sequences during vertebrates evolution: potential involvement in post-transcriptional regulation of gene expression. *Nucleic Acids Res* 21(10), pp. 2315-22.

300. Loots, G.G. et al. 2000. Identification of a coordinate regulator of interleukins 4, 13, and 5 by cross-species sequence comparisons. *Science* 288(5463), pp. 136-40.

301. Hardison, R.C. et al. 2003. Covariation in frequencies of substitution, deletion, transposition, and recombination during eutherian evolution. *Genome Res* 13(1), pp. 13-26.

302. Dermitzakis, E.T. et al. 2004. Comparison of human chromosome 21 conserved nongenic sequences (CNGs) with the mouse and dog genomes shows that their selective constraint is independent of their genic environment. *Genome Res* 14(5), pp. 852-9.

303. Dubchak, I. et al. 2000. Active conservation of noncoding sequences revealed by three-way species comparisons. *Genome Res* 10(9), pp. 1304-6.

304. Frazer, K.A. et al. 2001. Evolutionarily conserved sequences on human chromosome 21. *Genome Res* 11(10), pp. 1651-9.

305. Kirkness, E.F. et al. 2003. The dog genome: survey sequencing and comparative analysis. *Science* 301(5641), pp. 1898-903.

306. Dermitzakis, E.T. et al. 2003. Evolutionary discrimination of mammalian conserved non-genic sequences (CNGs). *Science* 302(5647), pp. 1033-5.

307. Frazer, K.A. et al. 2004. Noncoding sequences conserved in a limited number of mammals in the SIM2 interval are frequently functional. *Genome Res* 14(3), pp. 367-72.

308. Thomas, J.W. et al. 2003. Comparative analyses of multi-species sequences from targeted genomic regions. *Nature* 424(6950), pp. 788-93.

309. Bejerano, G. et al. 2004. Ultraconserved elements in the human genome. *Science* 304(5675), pp. 1321-5.

310. Glazko, G.V. et al. 2003. A significant fraction of conserved noncoding DNA in human and mouse consists of predicted matrix attachment regions. *Trends Genet* 19(3), pp. 119-24.

311. Fraser, P. and W. Bickmore 2007. Nuclear organization of the genome and the potential for gene regulation. *Nature* 447(7143), pp. 413-7.

312. Muller, H.P. and W. Schaffner 1990. Transcriptional enhancers can act in trans. *Trends Genet* 6(9), pp. 300-4.

313. Karolchik, D. et al. 2003. The UCSC Genome Browser Database. *Nucleic Acids Res* 31(1), pp. 51-4.

314. Mao, Y.M., Xie,Y. and Ying,K 1998. Genbank Direct Submission: AF061734. *http://www.ncbi.nlm.nih.gov*.

315. Frith, M.C. et al. 2003. Cluster-Buster: Finding dense clusters of motifs in DNA sequences. *Nucleic Acids Res* 31(13), pp. 3666-8.

316. Drake, J.A. et al. 2006. Conserved noncoding sequences are selectively constrained and not mutation cold spots. *Nat Genet* 38(2), pp. 223-7.

317. Ovcharenko, I. et al. 2004. ECR Browser: a tool for visualizing and accessing data from comparisons of multiple vertebrate genomes. *Nucleic Acids Res* 32(Web Server issue), pp. W280-6.

318. Siepel, A. et al. 2005. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res* 15(8), pp. 1034-50.

319. Crawford, G.E. et al. 2006. DNase-chip: a high-resolution method to identify DNase I hypersensitive sites using tiled microarrays. *Nat Methods* 3(7), pp. 503-9.

320. Crawford, G.E. et al. 2006. Genome-wide mapping of DNase hypersensitive sites using massively parallel signature sequencing (MPSS). *Genome Res* 16(1), pp. 123-31.

321. Sham, P.C. and D. Curtis 1995. Monte Carlo tests for associations between disease and alleles at highly polymorphic loci. *Ann Hum Genet* 59(Pt 1), pp. 97-105.

322. Buckland, P.R. 2006. The importance and identification of regulatory polymorphisms and their mechanisms of action. *Biochim Biophys Acta* 1762(1), pp. 17-28.

323. Knight, J.C. 2005. Regulatory polymorphisms underlying complex disease traits. *J Mol Med* 83(2), pp. 97-109.

324. Williams, N.M. et al. 2005. Is the dysbindin gene (DTNBP1) a susceptibility gene for schizophrenia? *Schizophr Bull* 31(4), pp. 800-5.

325. Zhao, J.H. et al. 2000. Model-free analysis and permutation tests for allelic associations. *Hum Hered* 50(2), pp. 133-9.

326. Rockman, M.V. and G.A. Wray 2002. Abundant raw material for cis-regulatory evolution in humans. *Mol Biol Evol* 19(11), pp. 1991-2004.

327. Iwamoto, S. et al. 1996. Characterization of the Duffy gene promoter: evidence for tissue-specific abolishment of expression in Fy(a-b-) of black individuals. *Biochem Biophys Res Commun* 222(3), pp. 852-9.

328. Tishkoff, S.A. et al. 2007. Convergent adaptation of human lactase persistence in Africa and Europe. *Nat Genet* 39(1), pp. 31-40.

329. Amara, S.G. and T. Pacholczyk 1991. Sodium-dependent neurotransmitter reuptake systems. *Curr Opin Neurobiol* 1(1), pp. 84-90.

330. Murphy, D.L. et al. 2004. Serotonin transporter: gene, genetic disorders, and pharmacogenetics. *Mol Interv* 4(2), pp. 109-23.

331. Heils, A. et al. 1996. Allelic variation of human serotonin transporter gene expression. *J Neurochem* 66(6), pp. 2621-4.

332. de Wet, J.R. et al. 1987. Firefly luciferase gene: structure and expression in mammalian cells. *Mol Cell Biol* 7(2), pp. 725-37.

333. Sherf, B.A. et al. 1996. Dual-Luciferase Reporter Assay: An Advanced Co-Reporter Technology Integrating Firefly and Renilla Luciferase Assays. *Promega Notes* 57, pp. 2-9.

334. Matthews, J.C. et al. 1977. Purification and properties of Renilla reniformis luciferase. *Biochemistry* 16(1), pp. 85-91.

335. Bullock WO, F., J. and J.M. Stuart 1987. XL1-Blue: a high efficiency plasmid transforming Escherichia coli strain with beta-galactosidase selection. *Biotechniques* 5(376-379).

336. DuBridge, R.B. et al. 1987. Analysis of mutation in human cells by using an Epstein-Barr virus shuttle system. *Mol Cell Biol* 7(1), pp. 379-87.

337. Thomas, P. and T.G. Smart 2005. HEK293 cell line: a vehicle for the expression of recombinant proteins. *J Pharmacol Toxicol Methods* 51(3), pp. 187-200.

338. Clark, J.M. 1988. Novel non-templated nucleotide addition reactions catalyzed by procaryotic and eucaryotic DNA polymerases. *Nucleic Acids Res* 16(20), pp. 9677-86.

339. Alexander, D.C. 1987. An efficient vector-primer cDNA cloning system. *Methods Enzymol* 154, pp. 41-64.

340. Jones, L. Personal communication.

341. Kaufman, D.L. and G.A. Evans 1990. Restriction endonuclease cleavage at the termini of PCR products. *Biotechniques* 9(3), pp. 304, 306.

342. NEB http://www.neb.com/nebecomm/products/productR0138.asp.

343. Buckland, P.R. et al. 2004. A high proportion of polymorphisms in the promoters of brain expressed genes influences transcriptional activity. *Biochim Biophys Acta* 1690(3), pp. 238-49.

344. Cookson, W. et al. 2009. Mapping complex disease traits with global gene expression. *Nat Rev Genet* 10(3), pp. 184-94.

345. Dixon, A.L. et al. 2007. A genome-wide association study of global gene expression. *Nat Genet* 39(10), pp. 1202-7.

346. Myers, A.J. et al. 2007. A survey of genetic human cortical gene expression. *Nat Genet* 39(12), pp. 1494-9.

347. Stranger, B.E. et al. 2007. Population genomics of human gene expression. *Nat Genet* 39(10), pp. 1217-24.

348. de Vooght, K.M. et al. 2009. Management of gene promoter mutations in molecular diagnostics. *Clin Chem* 55(4), pp. 698-708.

349. Dimas, A.S. et al. 2009. Common Regulatory Variation Impacts Gene Expression in a Cell Type-Dependent Manner. *Science*.

350. Heintzman, N.D. et al. 2009. Histone modifications at human enhancers reflect global cell-type-specific gene expression. *Nature* 459(7243), pp. 108-12.

351. Chen, J.M. et al. 2006. A systematic analysis of disease-associated variants in the 3' regulatory regions of human protein-coding genes II: the importance of mRNA secondary structure in assessing the functionality of 3' UTR variants. *Hum Genet* 120(3), pp. 301-33.

352. Di Paola, R. et al. 2002. A variation in 3' UTR of hPTP1B increases specific gene expression and associates with insulin resistance. *Am J Hum Genet* 70(3), pp. 806-12.

353. Gehring, N.H. et al. 2001. Increased efficiency of mRNA 3' end formation: a new genetic mechanism contributing to hereditary thrombophilia. *Nat Genet* 28(4), pp. 389-92.

354. Kuersten, S. and E.B. Goodwin 2003. The power of the 3' UTR: translational control and development. *Nat Rev Genet* 4(8), pp. 626-37.

355. Lee, N.H. et al. 1994. Agonist-mediated destabilization of m1 muscarinic acetylcholine receptor mRNA. Elements involved in mRNA stability are localized in the 3'-untranslated region. *J Biol Chem* 269(6), pp. 4291-8.

356. Moseley, P.L. et al. 1993. Heat stress regulates the human 70-kDa heat-shock gene through the 3'-untranslated region. *Am J Physiol* 264(6 Pt 1), pp. L533-7.

357. Reddy, K.K. et al. 2005. Perinuclear localization of slow troponin C m RNA in muscle cells is controlled by a cis-element located at its 3' untranslated region. *Rna* 11(3), pp. 294-307.

358. Spassov, D.S. and R. Jurecic 2003. The PUF family of RNA-binding proteins: does evolutionarily conserved structure equal conserved function? *IUBMB Life* 55(7), pp. 359-66.

359. Fuchs, J. et al. 2008. Genetic variability in the SNCA gene influences alpha-synuclein levels in the blood and brain. *Faseb J* 22(5), pp. 1327-34.

360. Hao, K. et al. 2005. Single-nucleotide polymorphisms of the KCNS3 gene are significantly associated with airway hyperresponsiveness. *Hum Genet* 116(5), pp. 378-83.

361. Moi, P. et al. 1992. Delta-thalassemia due to a mutation in an erythroid-specific binding protein sequence 3' to the delta-globin gene. *Blood* 79(2), pp. 512-6.

362. Morgan, K. et al. 1993. Point mutation in a 3' flanking sequence of the alpha-1-antitrypsin gene associated with chronic respiratory disease occurs in a regulatory sequence. *Hum Mol Genet* 2(3), pp. 253-7.

363. Morgan, K. et al. 1997. Mutation in an alpha1-antitrypsin enhancer results in an interleukin-6 deficient acute-phase response due to loss of cooperativity between transcription factors. *Biochim Biophys Acta* 1362(1), pp. 67-76.

364. Sato, M. et al. 1999. Genetic polymorphism of drug-metabolizing enzymes and susceptibility to oral cancer. *Carcinogenesis* 20(10), pp. 1927-31.

365. Shin, H.D. et al. 2003. Polymorphisms in fatty acid-binding protein-3 (FABP3) - putative association with type 2 diabetes mellitus. *Hum Mutat* 22(2), p. 180.

366. Zhernakova, A. et al. 2005. CTLA4 is differentially associated with autoimmune diseases in the Dutch population. *Hum Genet* 118(1), pp. 58-66.

367. Morris, D.W. et al. 2007. Dysbindin (DTNBP1) and the Biogenesis of Lysosome-Related Organelles Complex 1 (BLOC-1): Main and Epistatic Gene Effects Are Potential Contributors to Schizophrenia Susceptibility. *Biol Psychiatry*.

368. Moore, J.H. 2003. The ubiquitous nature of epistasis in determining susceptibility to common human diseases. *Hum Hered* 56(1-3), pp. 73-82.

369. Moore, J.H. and S.M. Williams 2002. New strategies for identifying gene-gene interactions in hypertension. *Ann Med* 34(2), pp. 88-95.

370. Jurinke, C. et al. 2004. MALDI-TOF mass spectrometry: a versatile tool for high-performance DNA analysis. *Mol Biotechnol* 26(2), pp. 147-64.

371. Dudbridge, F. 2003. Pedigree disequilibrium tests for multilocus haplotypes. *Genet Epidemiol* 25(2), pp. 115-21.

372. Straub, R.E. and D.R. Weinberger 2006. Schizophrenia genes - famine to feast. *Biol Psychiatry* 60(2), pp. 81-3.

373. Morris, D.W. et al. 2005. Association analyses of the bloc-1 genes suggest the involvement of bloc-1 in Schizophrenia etiology. *Am J Med Genet B Neuropsychiatr Genet* 138(1), pp. 3-147.

374. Sperling, S. 2007. Transcriptional regulation at a glance. *BMC Bioinformatics* 8 Suppl 6, p. S2.

375. Grosshans, H. and W. Filipowicz 2008. Molecular biology: the expanding world of small RNAs. *Nature* 451(7177), pp. 414-6.

376. Lencz, T. et al. 2007. Converging evidence for a pseudoautosomal cytokine receptor gene locus in schizophrenia. *Mol Psychiatry* 12(6), pp. 572-80.

377. O'Donovan, M.C. et al. 2008. Identification of loci associated with schizophrenia by genome-wide association and follow-up. *Nat Genet* 40(9), pp. 1053-5.

378. Purcell, S.M. et al. 2009. Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. *Nature* 460(7256), pp. 748-52.

379. Shi, J. et al. 2009. Common variants on chromosome 6p22.1 are associated with schizophrenia. *Nature* 460(7256), pp. 753-7.

380. Shifman, S. 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(2), p. e28.

381. Stefansson, H. et al. 2009. Common variants conferring risk of schizophrenia. *Nature* 460(7256), pp. 744-7.

382. Sullivan, P.F. et al. 2008. Genomewide association for schizophrenia in the CATIE study: results of stage 1. *Mol Psychiatry* 13(6), pp. 570-84.

383. Birney, E. et al. 2007. Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature* 447(7146), pp. 799-816.

384. Hudson, M.E. and M. Snyder 2006. High-throughput methods of regulatory element discovery. *Biotechniques* 41(6), pp. 673, 675, 677 passim.

385. Wei, C.L. et al. 2006. A global map of p53 transcription-factor binding sites in the human genome. *Cell* 124(1), pp. 207-19.

386. Dostie, J. et al. 2006. Chromosome Conformation Capture Carbon Copy (5C): a massively parallel solution for mapping interactions between genomic elements. *Genome Res* 16(10), pp. 1299-309.

387. Simonis, M. et al. 2007. An evaluation of 3C-based methods to capture DNA interactions. *Nat Methods* 4(11), pp. 895-901.

388. Scherf, M. et al. 2000. Highly specific localization of promoter regions in large genomic sequences by PromoterInspector: a novel context analysis approach. *J Mol Biol* 297(3), pp. 599-606.

389. Davuluri, R.V. et al. 2001. Computational identification of promoters and first exons in the human genome. *Nat Genet* 29(4), pp. 412-7.

390. Ohler, U. et al. 2000. Stochastic segment models of eukaryotic promoter regions. *Pac Symp Biocomput*, pp. 380-91.

391. Gross, S.S. and M.R. Brent 2006. Using multiple alignments to improve gene prediction. *J Comput Biol* 13(2), pp. 379-93.

392. Bajic, V.B. et al. 2006. Performance assessment of promoter predictions on ENCODE regions in the EGASP experiment. *Genome Biol* 7 Suppl 1, pp. S3 1-13.

393. Gorlov, I.P. et al. 2009. GWAS meets microarray: are the results of genome-wide association studies and gene-expression profiling consistent? Prostate cancer as an example. *PLoS One* 4(8), p. e6511.

394. Schadt, E.E. et al. 2008. Mapping the genetic architecture of gene expression in human liver. *PLoS Biol* 6(5), p. e107.

395. Chen, Y. et al. 2008. Variations in DNA elucidate molecular networks that cause disease. *Nature* 452(7186), pp. 429-35.

396. Emilsson, V. et al. 2008. Genetics of gene expression and its effect on disease. *Nature* 452(7186), pp. 423-8.

397. Moffatt, M.F. et al. 2007. Genetic variants regulating ORMDL3 expression contribute to the risk of childhood asthma. *Nature* 448(7152), pp. 470-3.

398. Bouzigon, E. et al. 2008. Effect of 17q21 variants and smoking exposure in early-onset asthma. *N Engl J Med* 359(19), pp. 1985-94.

399. Duan, S. et al. 2008. Genetic architecture of transcript-level variation in humans. *Am J Hum Genet* 82(5), pp. 1101-13.

400. Galanter, J. et al. 2008. ORMDL3 gene is associated with asthma in three ethnically diverse populations. *Am J Respir Crit Care Med* 177(11), pp. 1194-200.

401. Sleiman, P.M. et al. 2008. ORMDL3 variants associated with asthma susceptibility in North Americans of European ancestry. *J Allergy Clin Immunol* 122(6), pp. 1225-7.

402. Tavendale, R. et al. 2008. A polymorphism controlling ORMDL3 expression is associated with asthma that is poorly controlled by current medications. *J Allergy Clin Immunol* 121(4), pp. 860-3.

403. Libioulle, C. et al. 2007. Novel Crohn disease locus identified by genome-wide association maps to a gene desert on 5p13.1 and modulates expression of PTGER4. *PLoS Genet* 3(4), p. e58.

404. Mardis, E.R. 2008. The impact of next-generation sequencing technology on genetics. *Trends Genet* 24(3), pp. 133-41.

405. Siva, N. 2008. 1000 Genomes project. *Nat Biotechnol* 26(3), p. 256.

406. Cloonan, N. et al. 2008. Stem cell transcriptome profiling via massive-scale mRNA sequencing. *Nat Methods* 5(7), pp. 613-9.

407. Mortazavi, A. et al. 2008. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat Methods* 5(7), pp. 621-8.

408. Zhang, K. et al. 2009. Digital RNA allelotyping reveals tissue-specific and allele-specific gene expression in human. *Nat Methods* 6(8), pp. 613-8.

409. Nilsson, M. et al. 1994. Padlock probes: circularizing oligonucleotides for localized DNA detection. *Science* 265(5181), pp. 2085-8.

410. http://nihroadmap.nih.gov/GTEx/. [cited.

411. Dougherty, D.C. and M.M. Sanders 2005. Comparison of the responsiveness of the pGL3 and pGL4 luciferase reporter vectors to steroid hormones. *Biotechniques* 39(2), pp. 203-7.

412. Pennacchio, L.A. et al. 2006. In vivo enhancer analysis of human conserved non-coding sequences. *Nature* 444(7118), pp. 499-502.

413. Prabhakar, S. et al. 2006. Close sequence comparisons are sufficient to identify human cis-regulatory elements. *Genome Res* 16(7), pp. 855-63.