The Molecular Genetics of Developmental Dyslexia

Jade Chapman

Thesis submitted for the degree of Doctor of Philosophy

Department of Psychological Medicine and Neurology
Cardiff University

Supervisors: Prof. Julie Williams, Prof. Michael O'Donovan and
Dr. Denise Harold

2011

Contents

Appendix on CD

Acknowledgements

I would first like to thank my supervisors, Prof. Julie Williams, Prof. Mick O'Donovan and Dr. Denise Harold for giving me the opportunity to do a PhD in the department and for their help and supervision. I would also like to thank Denise for mentoring me and patiently answering my constant stream of questions over the last four years. Thanks must also go to the other members of Psychological Medicine, particularly to the other members of my office, Sarah, Amy, Didi, Evie, Debbie and Irina for providing sympathetic ears and ensuring the office was well stocked with chocolate and cake. I would also like to thank Dr. George Kirov for his help and guidance with the copy number variant analysis.

Many thanks have to go to the children and families who took part in this research; it would not have been possible without their participation.

I would also like to thank my friends, particularly Ann, Jenna, Mike, Mark, Lisa, and Laura for providing much needed distraction with food and wine and for ensuring that I left my flat now and again.

Finally I would like to thank my family, particularly my parents, Ken and Hazel, for their unfaltering love, support and encouragement throughout my education. This thesis is dedicated to them.

Summary

Developmental Dyslexia (DD) is a complex, cognitive disorder which is characterised by an impairment in reading despite adequate educational, motivational and intellectual opportunities. Family and twin studies have shown that this common neurodevelopmental disorder has a highly heritable component.

The aim of this thesis was to identify novel susceptibility variants for DD using several approaches.

A candidate gene study was conducted, testing variants within the genes *CDC42, PRTG*, *KIAA0319L*, *DCDC2b* and *RIOK3* for association with DD in the Cardiff case-control sample. None of the variants within these genes showed a significant association with DD.

A genome-wide association study (GWAS) was conducted in collaboration with other groups as part of the NeuroDys consortium, using DD cases from Europe and population controls. 27 of the most significant SNPs identified were selected and genotyped in a larger replication sample. 8 of these showed a significant association with DD, with the most interesting SNPs within the gene *SNX29*.

An additional GWAS was conducted by the NeuroDys consortium in the form of a pooling study using a larger array. 38 of the most significant SNPs identified were selected for individual genotyping after which 14 remained significant, with the most interesting within the genes *TMC1* and *WDR78.*

Another GWAS was conducted in the form of a pooling study using the Cardiff cases and screened controls only. 57 of the most significant SNPs identified were selected for individual genotyping of which 54 remained significant. This study highlighted a number of interesting genes and demonstrated the effect of using a homogeneous case-control sample when conducting pooling studies.

Analysis of copy number variants (CNVs) was also conducted using data from the initial NeuroDys GWAS. This study highlighted the technical issues that can affect the outcome of such studies. As such, the CNVs in this study need to be validated before these results can be relied upon.

To conclude, some interesting variants have been identified in this thesis but further work is required to confirm these findings.

Chapter 1: Introduction

1.1 Developmental Dyslexia

1.1.1 Definition

The World Health Organisation (WHO 2003) define dyslexia as:

"...an unexpected specific and persistent failure to acquire efficient reading skills despite conventional instruction, adequate intelligence and sociocultural opportunity."

Dyslexia is also referred to as specific reading disability (SRD) or developmental dyslexia (DD). Throughout this thesis I use the term DD as this refers to individuals who fail to develop competent reading skills, rather than those that have lost their ability to read competently as a result of brain damage (i.e. 'acquired dyslexia').

1.1.2 Epidemiology

Epidemiological data has suggested that reading ability occurs along a continuum, where individuals with DD form the lower tail end of the normal distribution and no 'cut point' can be used to distinguish individuals with DD from typically normal readers (Shaywitz et al. 1992). It is the most common of the learning disabilities, affecting between 5 and 10% of children of a school age (Shaywitz et al. 1990). Typically, these epidemiological studies of DD have been conducted using Western populations, many of which had English as their first language. English is considered to be a non-transparent language and the prevalence rates of DD may vary when studying more transparent languages such as German, Italian and Spanish. DD has also been found in non-Western populations at varying prevalence. A study of Chinese children from Hong Kong estimated prevalence rates of DD to be between 9.7% - 12.6% (Chan et al. 2007) and a large study of Egyptian children with DD estimated the prevalence to be 1.3% (Farrag et al. 1988).

Epidemiological studies have observed that DD is often found in males at a greater rate than in females, with a ratio of ~2:1 (Flannery et al. 2000). Some studies have suggested that this increase in prevalence in boys reflects an ascertainment bias rather than a gender difference, due to teachers rating boys as being significantly more inattentive and having more behaviour, language and academic problems than girls, resulting in more boys being referred for help (Shaywitz et al. 1990; Wadsworth et al. 1992; Katusic et al. 2001). However, large studies conducted on unselected samples have made this argument difficult to justify, with twice as many more boys affected by DD than girls in a large US sample of 32,223 children (Flannery et al. 2000), and a study of reading ability in nearly 200,000 children finding that girls outperformed boys on reading tests in every one of the 43 countries examined (Chiu & McBride-Chang 2006). Geschwind and Behan (1982) have hypothesised that gender differences could be explained by an excess of, or sensitivity to, androgens such as testosterone. This may delay left-hemisphere maturation which could result in abnormalities of neuronal migration and/or connections during gestation and ultimately DD.

1.1.3 Theories of DD

The exact processes involved in learning to read are unknown. Learning to read an alphabetic system requires learning the grapheme-phoneme correspondence. A grapheme is a written representation of a sound. For example, in Table 1.1, the word ‘book’ is spelt using the graphemes, ‘b’, ‘oo’ and ‘k’. Phonemes represent the smallest discernible segments of speech. For example, in Table 1.1, the word ‘scarf’ is spoken using the phonemes /s/, /k/, /ahr/ and /f/.

Item	**Examples**						
Word	Book			Scarf			
Grapheme	B	OO	K	S	C	AR	F
Phoneme	/b/	/oo/	/k/	/s/	/k/	/ahr/	/f/

Table 1.1: Terminology used to describe units of written and spoken language. Adapted from Paracchini et al. (2007).

Skills involved in processing phonemes include phonological coding/decoding, and phonological or phonemic awareness. Phonological coding is the ability to identify, discriminate and manipulate the phonological structure of a written word (Snowling 1995). Phonological awareness is the ability to be aware of and manipulate the sound structure of spoken words (Olson et al. 1994).

Orthographic processes are also important when learning to read, as these involve exploiting whole-word information, in particular, the visual appearance or shape (orthography) of a written word (Olson et al. 1994). Orthographic coding is the ability to recognise spelling patterns to establish the proper meaning of words. These skills are of particular importance when reading the English language for example, as the English language has many instances where identical sounding words differ in their meaning depending on how they are spelt (e.g. 'there' versus 'their'). In addition, the English language contains words whose sound do not represent their spelling (e.g. 'yacht').

There have been a number of theories put forward to explain the deficits observed in DD individuals but none of these theories are able to account for all cases of DD on their own.

1.1.3.1 The Phonological Core Deficit Theory

The phonological core deficit theory suggests that DD may be caused by a problem in representing, storing or retrieving phonemes from long-term memory, impairing an individual's ability to segment a word into its phonological units, resulting in poor or ineffectual reading (Snowling 1981). The phonological deficit is believed to be independent of non-phonologic abilities and so higher-order cognitive and linguistic functions (such as IQ and vocabulary) remain intact.

At the neurological level, it is usually assumed that the origin of the disorder is a congenital dysfunction of left-hemisphere perisylvian brain areas underlying phonological representations, or connecting between phonological and orthographic representations (Ramus et al. 2003b). Anatomical work (Galaburda et al. 1985) and functional brain images have supported this theory (Paulesu et al. 1996; Paulesu et al. 2001; Shaywitz et al. 1998; Brunswick et al. 1999; McCrory et al. 2000; Pugh et al. 2000; Temple et al. 2001; Shaywitz et al. 2002).

The phonological core deficit theory is the most widely accepted theory of DD. Castles and Coltheart (2004) have criticised this theory as no study has been able to show unequivocal evidence that there is a causal link for ability in phonological awareness and success in reading and spelling acquisition. It also suffers from its inability to explain the sensory and motor disorders that occur in a significant proportion of individuals with DD (Ramus et al. 2003b). However, supporters of the theory tend to dismiss these disorders as not being part of the core features of dyslexia, considering them to be potential markers of DD rather than playing a causal role (Snowling 2000).

1.1.3.2 The Double Deficit Hypothesis

The double-deficit hypothesis proposes that DD arises from deficits in both phonological processes and the rapid naming of simple stimuli (Wolf & Bowers 1999). Studies have shown that naming-speed deficits can cause variance in reading, independent of the variance contributed by phonological awareness measures (Blachman 1984; Bowers & Swanson 1991; Olson et al. 1994).

Wolf and Bowers (1999) have suggested that deficits in the processes underlying naming-speed hinder lower level perceptual requirements that result in non-fluent word identification and hinder comprehension. They also suggest that deficits in naming-speed represent a broader system of rate or efficiency-based difficulties that affect orthographic and phonological routes and representations.

1.1.3.3 The Rapid Auditory Processing Theory

This theory suggests that DD develops from an auditory deficit that inhibits the perception of short or rapidly varying sounds (Tallal 1980; Tallal et al. 1993). Dyslexics show poor performance on a number of auditory tasks, including frequency discrimination (McAnally & Stein 1996; Ahissar et al. 2000) and temporal order judgement (Tallal 1980; Nagarajan et al. 1999).

A failure to correctly identify short sounds and fast transitions can cause difficulties when such acoustic events are the cues to phonemic contrasts. For example, it may result in an inability to distinguish between the phonemes /ba/ and /da/. There has been evidence that individuals with DD may have poor categorical perceptions of sound contrasts, indicating that the auditory deficit may be the direct cause of phonological deficits which result in a failure to read (Mody et al. 1997; Adlard & Hazan 1998; Serniclaes et al. 2001).

1.1.3.4 The Visual Theory

This theory suggests DD is caused by a visual impairment giving rise to difficulties in processing letters and words in text (Lovegrove et al. 1980; Livingstone et al. 1991; Stein & Walsh 1997). It does not exclude a phonological deficit, but emphasizes a visual contribution to reading problems in some individuals with DD. The visual theory suggests that the magnocellular pathway is selectively disrupted in certain dyslexic individuals and this leads to deficiencies in visual processing, and, via the posterior parietal cortex, to abnormal binocular control and visuospatial attention (Stein & Walsh 1997; Hari et al. 2001). Support for this theory comes from anatomical studies showing abnormalities of the magnocellular layers of the lateral geniculate nucleus (Livingstone et al. 1991) and psychological studies have demonstrated decreased sensitivity in the magnocellular range, i.e. low spatial frequencies and high temporal frequencies, in individuals with DD (Lovegrove et al. 1980; Cornelissen et al. 1995). However, criticism for this theory has come from failures to replicate findings of a visual deficit (Victor et al. 1993; Johannes et al. 1996) or from findings that such a deficit exists in only a subgroup of individuals with DD (Cornelissen et al. 1995; Witton et al. 1998; Amitay et al. 2002).

1.1.3.5 The Cerebellar Theory

The cerebellar deficit theory attempts to tie in the motor deficits often associated with DD by recognising that the cerebellum is important in both movement controls and the automation of overlearned tasks, such as driving, typing and reading (Nicolson et al. 2001; Stoodley et al. 2005; Haslum & Miles 2007). A weak capacity to automatise would affect the learning of grapheme-phoneme correspondences. Support for this theory has come from evidence of individuals with DD performing poorly in a large number of motor tasks (Fawcett et al. 1996), in dual tasks demonstrating impaired automatisation of balance (Nicolson & Fawcett 1990) and in time estimation, which is a non-motor cerebellar task (Nicolson et al. 1995). Brain imaging studies have also shown anatomical, metabolic and activation differences in the cerebellum of dyslexics (Rae et al. 1998; Nicolson et al. 1999; Brown et al. 2001; Leonard et al. 2001).

This theory does not account for sensory disorders observed in individuals with DD, such as auditory and visual deficits, but supporters of the cerebellar theory have suggested that there may be two distinct subtypes of DD, one involving the cerebellum, the other the magnocellular pathways (Fawcett & Nicolson 2001). However, it is uncertain how many individuals with DD have motor problems as some studies have failed to find such problems (Wimmer et al. 1998; van Daal & van der Leij 1999; Kronbichler et al. 2002) whereas others only find them in subgroups of DD (Yap & van der Leij 1994; Ramus et al. 2003a).

1.1.3.6 The Magnocellular Theory

This theory attempts to integrate all of the findings mentioned above. The magnocellular (auditory and visual) theory suggests that a general impairment in magnocellular pathways will affect visual, auditory and tactile sensory modalities (Stein & Walsh 1997). The cerebellum is also thought to be affected by the general magnocellular deficit because it receives a large amount of input from various magnocellular systems in the brain (Stein & Walsh 1997). This theory therefore manages to account for all aspects of DD, including visual, auditory, motor, tactile and phonological difficulties (Ramus et al. 2003b). Evidence specifically relevant to the magnocellular theory includes magnocellular abnormalities that have been observed in the medial and lateral geniculate nuclei of dyslexic brains (Livingstone et al. 1991; Galaburda et al. 1994), poor performance of individuals with DD in the tactile domain (Grant et al. 1999; Stoodley et al. 2000) and the co-occurance of visual and auditory problems in certain individuals with DD (Witton et al. 1998; Cestnick 2001; Van Ingelghem et al. 2001).

Many supporters of the auditory and visual theories now agree that visual and auditory disorders in dyslexia are part of a more general magnocellular dysfunction (Ramus et al. 2003b). However, this theory still fails to describe why not all deficits are observed in all individuals with DD. For example, there have been a number of failures to replicate findings of auditory disorders in dyslexia (Heath et al. 1999; Hill et al. 1999; McArthur & Hogben 2001). Other studies have found auditory deficits in individuals with DD, but only in a subgroup (Tallal 1980; Reed 1989; Manis et al. 1997; Mody et al. 1997; Adlard & Hazan 1998; Lorenzi et al. 2000; Marshall et al. 2001; Rosen & Manganari 2001). In addition, results have shown inconsistencies with the theory that auditory deficit lies in 'rapid' auditory processing, and therefore with magnocellular function , with 'rapid' auditory processing remaining intact with some tasks while 'slow' auditory processing is found to be impaired with others (Reed 1989; McAnally & Stein 1996; Adlard & Hazan 1998; Schulte-Körne et al. 1998b; Witton et al. 1998; Nittrouer 1999; Lorenzi et al. 2000; Rosen & Manganari 2001; Share et al. 2002). It has also been argued that auditory deficits do not predict phonological deficits (Mody et al. 1997; Schulte-Körne et al. 1998a; Bishop et al. 1999; Marshall et al. 2001; Rosen & Manganari 2001; Share et al. 2002). The visual side of the magnocellular theory has also been criticised because visual impairments observed in individuals with DD tend to be in a whole range of stimuli rather than those that specifically tap into the magnocellular system (Skottun 2000; Amitay et al. 2002; Farrag et al. 2002). Kronbichler and colleagues (2002) found significant differences between DD cases and controls in phonological tests but not in visual, auditory or motor tasks, supporting the idea of a phonological deficit but not a deficit in the magnocellular system.

1.1.3.7 The Attentional Deficit theory

The neural process which allows for the processing of stimuli relevant to a particular task, while inhibiting irrelevant stimuli, is referred to as attention. When reading text, the attention needs to be shifted between individual letters and letter groups, as well as rapid and accurate integration of visual and auditory (if reading aloud) cues. Accurate reading also requires filtering out information from the periphery. Individuals with DD have been found to have deficits in suppressing interfering peripheral information and focusing their attention on the text (Geiger et al. 1994; Steinman et al. 1998; Facoetti et al. 2000). Individuals with DD have also been found to be easily distracted and often have problems maintaining attention on one task for prolonged periods (Keogh & Margolis 1976). Further support for this theory comes from the observation that attentional deficit hyperactivity disorder (ADHD) is often found to be comorbid with DD (see Section 1.1.5.1)

It is possible that all these theories are true, with different individuals having partially overlapping subtypes of DD (Ramus et al. 2003b). However, it could be that one theory accounts for every case of DD and that the other manifestations observed are different markers for DD (i.e. they are associated with DD but are not the cause of DD) (Ramus et al. 2003b).

1.1.4 Neurobiology of Developmental Dyslexia

Evidence for a neurobiological basis for DD comes from post-mortem examinations and brain imaging of individuals with DD.

1.1.4.1 Post Mortem Studies of Individuals with DD

The planum temporale is located within the sylvian fissure on the superior-posterior surface of the temporal lobe and it functions in auditory comprehension and possibly phonologic processing (Frank & Pavlakis 2001). Generally the planum temporale is larger on the left side and this asymmetry forms as early as 33 weeks gestation. Post mortem examinations of 7 brains from individuals with DD revealed that there was an increase in abnormalities of the left hemisphere, particularly around the perisylvian region, and the planum temporale was nearly symmetrical which has been suggested to be the result of enlargement of the right side (Galaburda & Kemper 1979; Galaburda et al. 1985; Humphreys et al. 1990). The abnormalities consisted of neuronal ectopias, dysplasias and vascular micro-malformations. The ectopias were often found in layer I of the left inferior frontal and superior temporal gyri, suggesting that these occurred at a time of peak neuronal migration during embryonic development (Galaburda et al. 1985). It should be noted however, that some individuals showing these abnormalities reported oral language delay as well as DD (Cohen et al. 1989).

Subsequent visual processing and auditory processing experiments suggested hypotheses which led to re-examination of these brains, revealing disorganisation of the magnocellular layers of the lateral geniculate nuclei (LGN) (Galaburda & Kemper 1979; Galaburda et al. 1985; Humphreys et al. 1990; Livingstone et al. 1991). These observations are consistent with the visual processing deficiencies observed in individuals with DD as this region of the brain forms part of the primate visual system (Livingstone et al. 1991). It was also noticed that the cell bodies making up the magnocellular layers of the LGN from the brains of individuals with DD appeared smaller than in control brainș. As the medial geniculate nuclei (MGN) are involved in the auditory processing system, these regions of the dyslexic brains were also examined (Galaburda & Kemper 1979; Galaburda et al. 1985; Galaburda et al. 1994; Humphreys et al. 1990). The dyslexic brains showed greater asymmetry between the left and right

MGN than in control brains, and in general, the left MGN had a higher number of smaller neurons and less larger neurons (Galaburda et al. 1994).

In addition to these studies on DD brains, studies on mice and rats with ectopia and microgyri have found that they exhibit a number of learning deficits (Denenberg et al. 1991; Schrott et al. 1992; Rosen et al. 1995; Balogh et al. 1998). Mice with ectopia in layer I of the cortex were found to learn differently compared to non-ectopic mice and since the ectopia that were observed are structurally similar to those of dyslexic individuals, this suggest that individuals with DD may learn differently to those without the learning disability. The specific location of the cortical disruption (e.g. in the pre-fontal cortex or motor cortex) also influenced the type of learning disability exhibited by the mouse and this may reflect the variability in the extent of learning disability found in individuals with DD (Hyde et al. 2001).

1.1.4.2 Brain Imaging Studies of Individuals with DD

Functional neuroimaging of brains unaffected by DD has taught us more about the processes behind reading. These studies suggest that two posterior pathways are involved in reading, the dorsal and ventral pathways, along with an anterior component. The dorsal pathway is centred in the left temporoparietal regions and includes the angular and supramarginal gyri as well as the left posterior end of the superior temporal gyrus (Simos et al. 2000b). This pathway deals with linking graphemes of a visual word with phonemes and an underactivation in this pathway is considered to be linked with a phonological deficit. The ventral pathway is centred on the left inferior occipitotemporal region, including the posterior fusiform gyrus. It is thought to be required for the quick automatic processing of familiar words or frequent letter strings within words and an underactivation of this pathway in individuals with DD is thought to be linked with slow and inaccurate word recognition. The anterior component is centred on the left inferior frontal gyrus and is involved in the articulation of speech sounds (Richlan et al. 2009).

Functional neuroimaging studies on individuals with DD have shown altered activity of these regions (Démonet et al. 2004). Corina and colleagues (2001) found that phonological and lexical tasks resulted in activation of the left inferior temporal gyrus in most controls brains, but very few of the DD showed any activation in this region. Other studies have also shown reduced activity of the left temporoparietal regions on tasks of word reading, non-word reading and letter rhyming (Simos et al. 2000a; Simos et al. 2000b; Temple et al. 2001) and left occipitotemporal regions on tasks of letter matching (Temple et al. 2001).

Shaywitz and colleagues (2002) conducted a neuroimaging study comparing 70 individuals with DD to 74 controls and found that individuals with DD had reduced activity in the left inferior frontal, left superior temporal, left occiptotemporal and left temporoparietal regions on several reading related tasks. They also found a correlation between reading skill and activity in left posterior regions (Shaywitz et al. 2002).

Neuroimaging studies have also identified that DD brains have greater asymmetry and less grey matter content of the cerebellum (Brown et al. 2001; Eckert et al. 2003; Leonard et al. 2001). Leonard and colleagues (2001) also found that phonological deficits correlate with a smaller right anterior lobe.

As well as this decreased activity in the left hemisphere of the brain in individuals with DD, increased activity in the right hemisphere is often observed, possibly as part of a compensatory mechanism. For example, studies have found that the right temporoparietal regions of the brains of individuals with DD showed greater activity in response to both word and non-word reading (Simos et al. 2000a; Simos et al. 2000b). Corina and colleagues (2001) also demonstrated that there was increased activity in the right inferior temporal gyrus relative to the left in the brains of individuals with DD during a phonological task.

These neuroimaging studies have indicated that there seems to be a reduction in activity of the left temporal areas which is consistent with the results from post-mortem brain studies of individuals with DD. Together, these results suggest that reading requires a number of brain areas, including posterior (phonological processes) and anterior brain regions (syntactic processing).

It is important to note that many of these neurobiological studies involve small numbers of participants which often included only adults and controls which were not well matched in terms of sex, handedness, intelligence or educational experience. Schlaug and colleagues (1995) have suggested that intensive training in language skills can modify the symmetry observed in individuals with DD, resulting in 'normalisation' of the brain and so studies need to be conducted on larger, homogenous samples of children, rather than adults.

1.1.4.3 Neuronal Migration and DD

A number of studies have suggested that impaired neuronal migration may be a functional cause of DD. The developmental progression of the cerebral cortex is unique to mammals and is fairly conserved throughout species (Ayala et al. 2007). Neuronal migration is a key step in the development of the neocortex, resulting in the organisation of neurons in to six specialised laminar layers (Hatten 1999). It is achieved through the rearrangement of cytoskeletal components in response to extracellular cues and is mediated by numerous intracellular pathways (Ayala et al. 2007).

Cortical neurones arise from proliferative pseudostratified epithelium at the margin of the embryonic cerebral vesicles (Rakic 1982). During each cell cycle, progenitor cells undergo a pattern of oscillation in the ventricular zone, termed interkinetic nuclear migration, shown in Figure 1.1. Projection neurons are born from radial glial cells in the ventricular zone. These migrate radially along radial glial fibres towards the pial surface. The S phase of the cell cycle occurs at the basal surface of the ventricular zone, with mitosis occurring at the apical surface. Once a cell has exited the cell cycle, it must migrate out of the ventricular zone to its final resting place in the developing neocortex. The first set of neurons that migrate out of the ventricular zone make up the preplate with the next wave of migration splitting the preplate into two layers: the marginal zone and the deeper subplate. The development of the cerebral cortex progresses with successive waves of migration that position neurons within the different layers in the cortical plate (Hatten 1999). These layers are established according to an inside-out pattern where the earliest waves of neuronal migration will go on to form the deeper layers of the cortical plate while the last waves of neurons will be localised to the more peripheral layers (Marín & Rubenstein 2001).

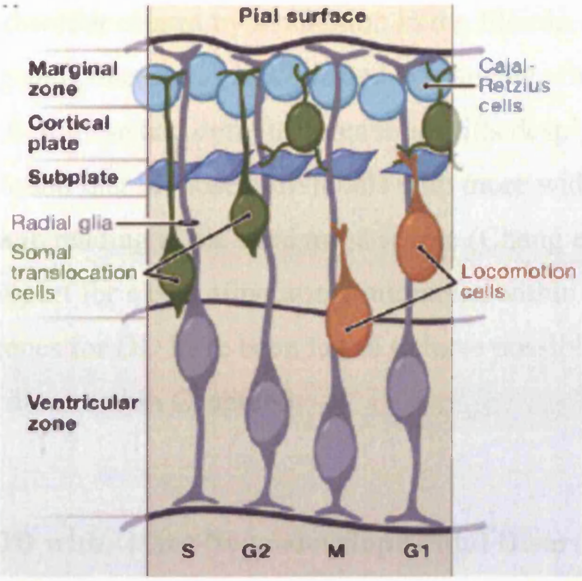


Figure 1.1: Diagram illustrating the migration of neurons from the ventricular zone, along glial fibres towards the pial surface, at different stages of the cell cycle. Taken from Ayala et al. (2007).

The general model for neuronal movement consists of three repetitive and highly regulated steps. In the first step, the cell extends a leading neurite which is preceded by a growth cone which extends and contracts as it explores the microenvironment. The growth cone is a very dynamic structure, showing extension and retraction of the filopodia as well as ruffling and transient advance or collapse of lamellipodia as it detects signals providing information on the direction that the cell should migrate in. The next step involves translocating the nucleus into the leading neurite, before the trailing process is retracted in the final step. This last step is still poorly understood (Ayala et al. 2007). The molecular mechanisms involved in regulating neuronal migration consist of extracellular guidance cues which are interpreted by receptors. These receptors then relay signals to a large network of intracellular signalling pathways, converging to the cytoskeleton. Both microtubule and actin networks are believed to operate synergistically to mediate migration (Ayala et al. 2007).

Defects in neuronal migration are the cause of several human syndromes with symptoms of epilepsy and mental retardation (Bielas et al. 2004; McManus & Golden 2005). As mentioned previously, Galaburda and colleagues (1979; 1985) undertook microscopic analysis of brain tissue, and they showed the presence of ectopia, dysplasia and vascular micro-malformations in cortical parietal regions of the brains from individuals with DD, suggesting impairments in neuronal migration. Further support for the involvement of impaired neuronal migration in DD has come from behavioural studies conducted on individuals with periventricular nodular heterotopias (PNH) which is a neuronal migration disorder caused by a mutation in the filamin A (*FLNA*) gene (Fox et al. 1998). Chang and colleagues (2005) tested 10 patients with PNH and epilepsy and found that 8 of these had deficits in reading skills despite normal intelligence. They also found that in those individuals with more wide-spread heterotopias, the deficits in reading skills were more severe (Chang et al. 2005).

Providing further support for a role of neuronal migration within DD, a number of putative susceptibility genes for DD have been found to have possible roles in neuronal migration and these are discussed in Chapter 3.

1.1.5 Comorbidity of DD with Other Neurodevelopmental Disorders

A number of disorders have been found to exist in individuals with DD more often than in matched controls.

1.1.5.1 Attentional Deficit Hyperactivity Disorder (ADHD)

ADHD is characterised by inattention, overactivity and impulsiveness. Studies have estimated that 15-40% of individuals diagnosed with DD also have ADHD (Gilger et al. 1992; Willcutt & Pennington 2000; Willcutt et al. 2007). Twin and family studies on DD and ADHD have suggested that these disorders have shared genetic underpinnings (Willcutt et al. 2000; Willcutt et al. 2007). When ADHD is subdivided into its symptom dimensions, twin studies also predict a stronger relationship between DD and symptoms of inattention compared with symptoms of hyperactivity/impulsivity (Willcutt et al. 2000).

Results from linkage studies of DD and genome scans of ADHD have indicated several regions that overlap. Willcutt and colleagues (2002) have suggested that the comorbidity between DD and ADHD may be partly due to the effects of the quantitative trait locus (QTL) on chromosome 6p (see section 1.4.4). Overlapping loci on 15q, 16p, 17p, 10q, 14q32, 13q32 and 20q11 have also been suggested by additional studies (Loo et al. 2004; Gayán et al. 2005; Wigg et al. 2004; Wigg et al. 2008).

1.1.5.2 Developmental Dyscalculia (DC)

DC is generally defined as a specific impairment in arithmetic abilities, despite any deficits in intelligence, socioeconomical background, general motivation, emotional stability, educational opportunity or sensory acuity. Studies have estimated that 25-37% of individuals with DD also have DC (Knopik et al. 1997; Lewis et al. 1994) and 17 - 70% of individuals with DC also have DD (Gross-Tsur et al. 1996; Knopik et al. 1997; Lewis et al. 1994).

1.1.5.3 Developmental Coordination Disorder (Dyspraxia)

Dyspraxia is an impairment in the development of motor coordination which is not attributable to a general medical condition or mental retardation. It has been estimated that there is an overlap of 30-50% between DD and dyspraxia (Kadesjö & Gillberg 1999; Richardson & Ross 2000). Whilst a relationship between lower motor ability, such as hand motor skill and DD has been observed, the genetic effects in motor skill are largely distinct from DD (Francks et al. 2003).

1.1.5.4 Specific Language Impairment (SLI) and Speech-Sound Disorder (SSD)

SLI is an impairment in the ability to acquire adequate language skills, despite normal intelligence and development. Studies have estimated that 20-60% of individuals with DD also have SLI (McArthur et al. 2000) and 40-80% of individuals with SLI also have DD (McArthur et al. 2000).

SSD (or phonological disorder) is characterised by speech-sound production errors associated with deficits in articulation, phonological processing and cognitive linguistic processing. There is not much evidence of increased co-morbidity between DD and SSD alone, but in conjunction with language impairments there is significant co-morbidity with DD, particularly with deficits in spelling (Bishop & Adams 1990; Lewis et al. 1994). Stein and colleagues (2004) found that 21.6 % of individuals with SSD also have DD and also found linkage of SLI to a region of chromosome 3, a region implicated in linkage studies of DD as discussed in section 1.4.5.

1.2 Genetic Basis of Developmental Dyslexia

1.2.1 Familiality of Developmental Dyslexia

The first step in determining whether or not a disorder may have a genetic aetiology is to establish that the disorder runs in families (familiality). This is measured by comparing the rate of a disorder in relatives of probands to the baseline rate found in the general population.

Familial clustering of DD was observed over 100 years ago (Hinshelwood 1907; Stephenson 1907; Thomas 1905). Since then, a number of large studies have confirmed DD familiality. An early study by Rutter and Yule (1975) found that 9% of control children had a sibling or parent with a reading problem, compared with 34% of children with DD. More recently, it has been shown that 20-33% of siblings of affected individuals, with unaffected parents, are themselves also affected (Gilger et al. 1996). This percentage increased to 54-63% if either parent was also affected (Gilger et al. 1996).

The probability of an individual being affected with DD given that a sibling is already affected (regardless of parental affection status) is estimated to be between 38% and 60% (Hallgren 1950; Finucci et al. 1976; Vogler et al. 1985; Gilger et al. 1991; Gilger et al. 1996). These sibling recurrence risk estimates are considerably higher than the general population risk (5 to 10%) (Pennington 1990).

Gilger and colleagues (1991) also studied the probability of a mother or father being affected if they had affected offspring in three population samples from Iowa and Colorado. They found that the risk of a father being affected if their son was affected was between 30% and 35% and between 12% and 15% for mothers. For those parents with an affected daughter, the risk was between 17% and 41% for fathers and 30% and 42% for mothers. Whereas the risk for parents of a normal control proband was 4% for fathers and 3% for mothers, which is close to the general population risk (Pennington 1990).

From these studies, it is clear that the risk of having DD is greater among first degree relatives than in the general population. However, as families tend to share their environments, this alone does not imply that DD is strictly influenced by genetic factors. Twin studies have therefore been used to try to differentiate the genetic and environmental factors that influence DD and estimate the heritability of this disorder.

1.2.2 Heritability of Developmental Dyslexia

Typically, twin studies consist of large sets of monozygotic (MZ) and same sex dizygotic (DZ) twins. The concordance rate for DD is then compared between the two sets of twins, with a higher concordance rate for DD in the MZ twins being suggestive of some genetic aetiology for DD.

Early twin studies of DD showed significantly greater MZ than DZ concordance for DD (Hallgren 1950; Hermann 1956; Bakwin 1973), but these suffered from methodological problems including ascertainment bias, inconsistent definitions of DD, failure to determine zygosity adequately and failures to limit DZ twin samples to same sex twins (Stevenson et al. 1987; Pennington 1989). The first compelling evidence that DD is influenced by genetic factors came from two large twin studies in the 1980s: the London Twin Study (Stevenson et al. 1987) and the Colorado Twin Reading Study (DeFries et al. 1987). Stevenson and colleagues (1987) examined the reading skills of 285 pairs of 13 year old twins identified from the general population on the basis of birth records and from the registers of schools in the London area. They did not find any evidence supporting the heritability of reading in general, but did find significant heritability for phonological coding (82%). In a later study of this twin sample, Stevenson (1991) observed significant heritability for impaired spelling (62%) and found that deficits in phonological processing were more heritable than deficits in orthographic coding.

DeFries and colleagues (1987) used multiple regression analysis of twin data in their analysis of 64 MZ and 55 DZ twin pairs in which at least one member of the pair (the proband) had DD. These twins were ascertained from schools in Colorado. The reading ability of both twins was then assessed by a standard battery of tests that had been previously shown to discriminate between individuals with DD and normal readers. This study found highly heritable components to reading (44%), spelling (62%) and deficits in phonological processing (75%), but did not identify a significant heritable component for orthographic processing (31%).

Gayán and Olson (1999) used a larger sample from the Colorado Twin Study and reported significant heritabilities of 61% for phonological decoding, 56% for phonological awareness and 58% for orthographic processing. This suggests that phonological and orthographic deficits in DD have similar levels of heritability, in contrast with the studies by Stevenson (1991) and DeFries and colleagues (DeFries et al. 1987) which observed lower heritability for orthographic deficits.

The concept of heritability is not a fixed one as the variance that can be attributable to genes is partly dependent on the variance in exposure to relevant (but unknown) environmental risk factors and on the characteristics of the population studied (Williams & O'Donovan 2006). In a further study of the Colorado twin sample, DeFries and colleagues (1997) demonstrated that reading had a higher heritability in younger compared with older children, whereas the heritability of spelling increased with age. Another study on this sample by Wadsworth and colleagues (2000) found that heritability estimates increase with increasing levels of IQ. They reported a heritability of 43% for a group of twins with the average IQ of each twin pair below 100 and a heritability of 72% for a group of twins with an average IQ above 100, suggesting that genetic factors may be more important in DD in individuals with higher IQs. However, they did not determine whether the genetic factors that influenced DD differed as a function of IQ. It may be that the same genetic factors are involved in DD, regardless of IQ but the proportion of variance accounted for by these genetic factors may differ because the degree and nature of environmental influences may vary as a function of IQ.

1.2.3 Mode of Transmission

Pennington and colleagues (1991) performed segregation analysis of DD in order to establish a mode of transmission. They examined the relatives of DD probands in four independent samples from Colorado, Washington and Iowa, producing a sample of 204 families and 1,698 individuals. Three of the four samples showed evidence for major locus transmission. In the first three samples, the estimates of penetrance of the AA, Aa and aa genotypes (where A is the risk allele) were, respectively, 1, 1 and 0.001-0.039 in males and 0.56-1, 0.55-0.897, and 0 in females. Therefore these samples were consistent with models of dominant or additive transmission, with sex-dependent penetrance. The fourth sample showed evidence of a multifactorial-polygenic transmission. Pennington and colleagues (1991) noted that they may have underestimated the multifactorial background of DD; the ascertainment procedures for two of the three samples showing evidence of major gene effect may have created bias. They also noted that the male penetrance estimates of 1.0 for homozygotes and heterozygotes were inconsistent with male MZ twin concordance rates, which are substantially less than 1.0 (Pennington et al. 1991).

Gilger and colleagues (1994) followed up this study by carrying out segregation analysis on a quantitative reading phenotype in 125 families which had been ascertained through normal probands. The results of this study suggested that a major gene with dominance was responsible for a significant amount of variance in reading scores (54%).

Wijsman and colleagues (2000) conducted segregation analysis on two phonological phenotypes (non-word memory and digit span) in 102 families with a DD affected proband. They found evidence in support of a Mendelian mode of inheritance with an intermediate heterozygous phenotype with a dominance of ~0.8 for non-word memory. The inheritance of digit span was best described by a dominant gene model.

As discussed in section 1.1.2, a higher proportion of males than females appear to be affected by DD, suggesting that there may be sex effects on the transmission of this disorder. However, none of the family studies or segregation analyses found evidence for X-linked transmission (Pennington et al. 1991) or for mitochondrial transmission as transmission rates from each parental sex were essentially equal.

There does not appear to be a consensus mode of transmission of DD as yet. The findings that the MZ concordance rates and heritabilities for DD from twin studies are substantially less than 1.0 makes it unlikely that there is a major locus transmission. There may be a mixed model of a small number of susceptibility loci operating against a multifactorial background, or there may be a polygenic multifactorial model with a small number of Mendelian subforms of the disorder also existing.

1.3 Methods for Identifying Susceptibility Variants

As there is strong evidence for a genetic component influencing susceptibility to DD, the next logical step is to attempt to identify genes which confer susceptibility to DD. The two main methods that have been used to locate such susceptibility genes for DD so far are linkage and association analysis. They can be used to test specific candidate genes, or in a systematic scan of a chromosome or the entire genome. Linkage is only able to detect genes of major effect, but allows scanning of the entire genome with only a few hundred markers (Sham & McGuffin 2002). Association analysis can detect genes of more moderate effect, although until recently, only relatively small regions could be analysed at one time due to the number of markers required as these needed to be much more densely spaced compared to linkage. The advent of genome-wide association studies (GWAS) have now enabled the whole genome to be systematically scanned for association with a disorder. In the last few years, analysis of structural variants has also been employed to identify susceptibility variants for complex disorders.

1.3.1 Linkage Analysis

Linkage analysis has been particularly successful at detecting genes that cause Mendelian traits and is based on the phenomenon known as genetic linkage. During meiosis, homologous pairs of chromosomes can exchange genetic material in a process known as recombination which occurs at crossover points and chiasma. This process generates increased diversity among the human species, restructuring genes and their alleles (Sham & McGuffin 2002). If two loci are on different chromosomes, the probability that their alleles will be inherited together is 0.5, a phenomenon which Mendel described as independent assortment. For loci on the same chromosome, the nearer that two genes are to one another, the less likely it is that a crossover point will occur between them and the more likely it is that they will be inherited together. This departure from the law of independent assortment is known as genetic linkage.

Linkage analysis tests for cosegregation of a genetic marker and disease phenotype within many independent families or over many generations in an extended pedigree. Although the marker itself may not be causing the disease or phenotype, genetic linkage indicates that a susceptibility locus causing the phenotype is within the same chromosomal region as the segregating marker. The traditional approach for calculating the statistical evidence for linkage is the LOD score (Morton 1955). This 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. It has been suggested that a cumulative LOD score exceeding 3 can be regarded as good evidence for linkage, while a cumulative LOD score below -2 should be regarded as strong evidence against linkage.

1.3.1.1 Parametric and Non-parametric Analysis

Classic linkage analysis (known as parametric analysis) involves the specification of a genetic model and is a powerful method for detecting loci segregating in a Mendelian fashion (Sham & McGuffin 2002). However, DD is a complex disease and as discussed in section 1.2.3, the mode of transmission in unknown. In the case of complex diseases, non-parametric linkage analysis is a more appropriate method, albeit a less powerful one. Parametric linkage analysis follows the cosegregation of markers and disease/phenotype over a number of generations in large multiplex families, whereas non-parametric linkage analysis usually examines allele sharing in affected relatives, such as affected sib-pairs. Allele sharing can either be defined by identity-by-state (IBS) or identity-by-descent (IBD). If two alleles have the same DNA sequence at the polymorphic site then they are characterised as IBS. If these alleles are also both descended from a recent common ancestor then they are said to be IBD. The IBD measure of allele sharing is more informative and less dependent on knowledge of the exact marker allele frequency (Sham & McGuffin 2002).

1.3.1.2 Limitations of Linkage Analysis

For most common diseases, linkage analysis has only achieved limited success (Altmüller et al. 2001; Hirschhorn & Daly 2005), which can be attributed to a number of factors. For linkage analysis to succeed, markers that flank the disease gene must segregate with the disease in families. Linkage studies have been successful for mapping genes which underlie Mendelian traits because variants that cause monogenic disorders are often rare and so each segregating disease allele will be found in the same 10-20cM chromosomal background within each family (Hirschhorn & Daly 2005). In addition, Mendelian diseases are caused by highly penetrant variants and so markers within 10-20cM of the disease causing alleles will co-segregate with disease status (Hirschhorn & Daly 2005). Advocates of the common disease, common variant hypothesis argue that many of the alleles affecting susceptibility to common complex traits (such as DD) will themselves be common (Reich & Lander 2001; Lohmueller et al. 2003). Most common diseases also have complex architectures in which the phenotype is determined by interactions between multiple genetic and environmental factors (Wang et al. 2005) and as such, any individual genetic variant will generally have a relatively small effect on disease risk (Hirschhorn & Daly 2005). Linkage analysis is less powerful at identifying common genetic variants which have modest effects and prohibitively large sample sizes would be needed to detect small effects (Risch & Merikangas 1996). In addition, the standard set of microsatellites used in linkage analysis are spaced ~10cM apart and are therefore unlikely to extract complete inheritance information (Hirschhorn & Daly 2005). Increasing the density of the marker map does not have a great effect on the resolution (Teare & Barrett 2005). Finally, whilst the linkage region identified may contain a susceptibility gene, such regions often contain hundreds of genes, many of which are biologically plausible candidates (Teare & Barrett 2005).

1.3.2 Association Studies

Genetic association studies aim to detect association between one or more genetic polymorphisms and a trait by looking for a significant difference in marker allele frequencies between a group of disease affected cases and unaffected controls. Association differs from linkage in that the same marker allele (or alleles) is associated with the trait in a similar manner across the whole population, while linkage allows different marker alleles to be linked with the trait in different families (Cordell & Clayton 2005). Association studies have greater power than linkage studies to detect small effects, but they require many more markers to be examined (Cordell & Clayton 2005).

Allelic association describes a significant difference in marker allele frequency between cases and controls, whereas genotypic association refers to a significant difference in genotype frequency. There are 3 reasons why an association between a genetic marker and a trait might exist in a population. The polymorphism may itself be the causal variant and this is referred to as direct association. Alternatively, the polymorphism may not have a causal role but is associated with a nearby causal variant and this is referred to as indirect association. Finally, the association may also be due to some underlying stratification or admixture in the population being studied.

1.3.2.1 Linkage Disequilibrium

Indirect association arises due to linkage disequilibrium (LD) which refers to the co-occurrence or correlation between two loci on the same chromosome. The reshuffling of genes in meiosis will tend to reduce the level of LD between all pairs of loci from one generation to the next. However, as discussed earlier, markers that are close together are less likely to be separated by recombination and so the degree of LD between such alleles will decay at a slower rate over time. The degree of LD existing between two loci is usually calculated by one of two measures, *D'* or r^2 (Devlin & Risch 1995). 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 frequencies of the two alleles (p_A and q_B) and the haplotype frequency (α_{AB}):

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

However, a limitation of using *D* as a measure of LD between two markers is that its possible value is constrained by the frequencies of each marker allele. In order to compare values of *D* between different pairs of markers with different allele frequencies, *D* is normalised to *D'* (Mueller 2004). Unlike *D, D'* lies on a scale of 0-1 and is calculated using the theoretical maximal and minimal values of *D* (D_{max} and D_{min}) (Devlin & Risch 1995; Mueller 2004):

$$\text{If } D \geq 0,\ D' = D/D_{max}$$
$$\text{If } D < 0,\ D' = D/D_{min}$$

A *D'* value of 0 indicates there is no correlation between two loci, while a value of 1 indicates complete LD, where all copies at one locus occur exclusively with one of the two possible alleles at the second marker. *D'* is an important measure for the identification of regions in which there has been little recombination and, therefore, in regions where there is the potential to map causal loci by indirect association studies (Cordell & Clayton 2005). However, a limitation of this measure 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 since it reflects the correlation only since the most recent mutation occurred (Zondervan & Cardon 2004). For example, allele A of one locus may always occur with allele B of a second locus. However, allele B of the second locus may occur with both allele A and a of the first locus.

The alternative measure, r^2, accounts for this. It is measured using D plus the product of the allele frequencies at the two loci:

$$r^2 = \frac{D^2}{f(A)f(a)f(B)f(b)}$$

As with *D'*, r^2 values lie between 0 and 1, but an r^2 of 1 indicates a perfect correlation between the genotypes of two markers, where the occurrence of an allele at one marker perfectly predicts the allele at a second locus (Zondervan & Cardon 2004).

1.3.2.2 Association Due to Population Stratification

As well as direct and indirect association, another possible reason for a marker showing association with a disease or trait may be due to some underlying stratification or admixture in the population being studied. This may arise if the cases and controls being studied are not ethnically comparable as differences in allele frequency may arise whether the alleles are causally related to a disease or not. Meta-analyses (Ioannidis et al. 2003) have indicated that causal variants for complex disease are likely to have small effect sizes and so large studies will be required to detect them (Dahlman et al. 2002). In this situation, even modest confounding by stratification and admixture could have a large effect on the results (Cordell & Clayton 2005). It is therefore vital to ensure that cases and controls are well matched for ethnicity and other confounding factors. One method to deal with this is to use unaffected family members as controls, such as parents or siblings. However, using siblings results in a loss of power as they are over-matched to the cases. In addition, family studies are difficult to conduct on a sufficiently large scale to detect associations reliably (Cordell & Clayton 2005). Therefore, well chosen unrelated controls may be the best option.

1.3.2.3 Association Study Design

To guarantee detection of all possible disease-associated variants at a given gene or locus, every base at which variation might conceivably alter gene function or expression would need to be examined in very large samples (Hattersley & McCarthy 2005). This is currently unrealistic and so research groups have used various strategies in the design of their association studies in order to retain as much power as possible to detect true causal variants for a disease, but without the study becoming prohibitively expensive.

Association studies can be conducted using a hypothesis-led candidate gene approach or using a more systematic approach which may involve testing polymorphisms in a chromosomal segment or conducting a genome-wide association study (GWAS). They can also be conducted using a case control sample, or using a family based sample.

1.3.2.3.1 Tag Marker Selection

Advocates of the common disease, common variant hypothesis argue that many of the alleles affecting susceptibility to common complex traits (such as DD) will themselves be common (Reich & Lander 2001; Lohmueller et al. 2003). If this is true, then typing of regional 'tag-markers' which are selected specifically to capture such common alleles should provide an efficient approach for detecting complex trait susceptibility alleles (Zondervan & Cardon 2004; Gabriel et al. 2002) .

The presence of LD in the human genome allows for the selection of these tag markers when conducting association studies. If an association signal is detected then the polymorphisms tagged by the significant marker are considered to be potential risk variants along with the significant marker itself. In order to identify these tag markers, it is first necessary to genotype all polymorphisms at a locus within either a subset of the association sample or a representative population. This allows estimation of the LD structure of the association sample and so tag markers can be selected. The HapMap database (www.hapmap.org) is a research tool which often allows the researcher to avoid this initial step. It is a publically available database created by the International Haplotype Map Consortium (The International HapMap Consortium 2007). 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 (CEU), West African (ASW), Han Chinese (CHB) and Japanese (JPT). The more recent phase III of the HapMap project now contains genotype information for individuals from populations across the world including Chinese in Colorado (CHD), Gujariti Indians in Texas (GIH), Luhya in Kenya (LWK), those with Mexican ancestry in California (MEC), Maasai in Kenya (MKK), Tuscan in Italy (TSI) and Yoruban in Nigeria (YRI).

The selection of tag SNPs is based on a number of factors. The researcher needs to decide how much of the genetic variation in a region they want to cover (e.g. common variation versus rare variation) and how thoroughly (i.e. the degree of LD between the genotyped marker and the tagged SNP). Due to the large sample sizes that are needed to detect an association with rare SNPs, a minor allele frequency (MAF) ≥ 0.05 is often chosen. The ideal tag SNP selection strategy would also genotype all markers where the alleles are not in an r^2 of 1 with any other genotyped marker. However, this could still result in a prohibitively large number of SNPs to genotype so this threshold is often reduced to $r^2 \geq 0.8$.

1.3.2.3.2 Case Control Studies

The simplest type of association study involves comparing individuals with a disease (cases) with unaffected subjects from the same population (controls) (Sham & McGuffin 2002). Because family data are not required, case control samples are relatively easy to collect and this allows for the collection of larger samples, giving the study greater power to detect true association variants (see section 1.3.2.4). As mentioned earlier, the one drawback of a case control study is that spurious association can arise in the presence of population stratification. To protect against this, it is important to ensure that the sample is ethnically homogenous and that the cases are matched to the controls for possible confounding factors such as age and sex. It is often not possible to obtain perfectly matched case control samples however, and often unknown confounding factors can remain. In these instances, epidemiological methods of adjustment such as stratified analysis or logistic regression can be used to correct for confounding factors as much as possible (Sham & McGuffin 2002). Logistic regression is particularly useful in also allowing the analysis of potential interaction between genotype and demographic or environmental factors.

1.3.2.3.3 Family Based Studies

As mentioned previously, one way to prevent population stratification producing confounding effects is to use family based samples, in which unaffected family members are used as controls, such as parents or siblings. The most popular family based association study design uses parent-proband trios in which both parents and their affected offspring are genotyped (Sham & McGuffin 2002). 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 than expected by chance, then the allele is said to be associated.

However, there are two additional costs to using this type of study design. Firstly, complete families will inevitably be more difficult to identify and recruit, especially for late-onset diseases such as Alzheimer's disease. Secondly, for equivalent power, the trio design is relatively more expensive than a case/control approach, requiring multiples of three individuals to be genotyped compared to two individuals of a case-control pair. A drawback of using unaffected siblings is the resulting loss of power as they are over-matched to the controls, as mentioned earlier.

1.3.2.3.4 Pooling Studies

Another option to reduce the cost of association studies is to carry them out in the form of a pooled study. These involve mixing equal amounts of DNA from each sample to form a pool of case samples and another of control samples. Thus, the allele frequencies in a sample of 200 cases and 200 controls can be measured from two pooled samples, rather than 400 individual samples. However, as allele frequencies can only be estimated from such studies, replicates of pooled samples are often run in order to calculate the average allele frequencies.

Unfortunately, the increase in efficiency gained by genotyping fewer samples is somewhat reduced due to a loss of detailed information that could have been obtained through individual genotyping (Sham et al. 2002). Pooling studies and their caveats are discussed in more detail in Chapters 5 and 6 of this thesis.

1.3.2.3.6 Candidate Gene Studies

This type of association study involves testing variants within interesting genes for an association with a disease or trait, as opposed to systematically testing the whole genome or a chromosomal segment. In these hypothesis based studies, genes are selected for further study on the basis of evidence that might affect disease risk.

Evidence from a range of sources can be used to identify candidate genes (Hattersley & McCarthy 2005). These may include: biological evidence which shows that the function of the protein encoded by the gene is implicated in the disease or trait; the gene may encode a protein which is implicated in the mechanism-of-action of a disease-modifying drug; animal homologues of the gene may be implicated in related traits in animal models; genome-wide scans for linkage or association could indicate regions with a high probability of containing a susceptibility gene. A major problem in selecting candidate genes is that if the precise biology underlying the trait is unknown, it is possible to find evidence connecting almost any gene to the disease of interest. Therefore a combination of evidence from different sources is often required when selecting candidate genes.

An additional caveat of candidate gene is that detection of a true association would not only require that the gene product is involved in pathways relevant to the development of the trait of interest, but also that the gene contains variants that are capable of influencing its regulation or function (Hattersley & McCarthy 2005).

1.3.2.3.5 Genome Wide Association Studies

Genome wide association studies (GWAS) involve systematically testing variants across the genome for association with a particular trait or disease. This enables a hypothesis-free approach to identifying susceptibility variants, and therefore does not require prior knowledge of the biology of the disease.

The increasing knowledge of common polymorphisms in the genome in different populations through the HapMap project (The International HapMap Consortium 2007) has enabled the development of commercial arrays which currently allow the researcher to genotype over 2.5 million markers across the human genome. The two main manufacturers of these arrays are Illumina (www.illumina.com) and Affymetrix. These companies use different methods to select markers for inclusion in their arrays. The probes on the Affymetrix arrays are spaced evenly throughout the genome without taking inter-SNP LD into account whereas Illumina selected tag SNPs to maximise genetic coverage. The earlier Illumina arrays (e.g. HumanHap300) contained probes for over 300,000 tag SNPs, whereas the newest arrays (e.g. HumanOmni2.5-Quad) contain probes for ~2.5 million markers, including probes for common ($MAF > 0.05$) as well as rare ($MAF > 0.025$) SNPs and non-polymorphic probes to capture common copy number variation (see section 1.3.3 and Chapter 7).

As the older arrays provide less coverage of the genome, a modest boost to their power can be achieved by using computational approaches to improve the detection of associations that are attributable to SNPs that are known but have not themselves been directly genotyped (Browning 2008). This is referred to as 'imputation'. Imputation can also enable results from two or more studies that have been genotyped on differing sets of markers to be compared or meta-analysed. Imputation is carried out using a reference panel of samples. Traditionally this reference panel has come from HapMap II but recently reference panels from the 1000 Genomes Project can be used as this panel contains most of the variation occurring at a population frequency > 1% (Via et al. 2010). Imputation is related to tagging, in that HapMap data can be used to infer LD between alleles at each SNP. If genotyped SNP(s) are correlated with another un-genotyped SNP, the missing genotype data can be imputed using the haplotype structure defined by the reference panel. A larger sample size is needed to achieve comparable power to genotyping the imputed SNPs directly because imputation is typically less accurate than genotyping (Anderson et al. 2008). A number of imputation algorithms have been developed to carry out imputation of missing genotype data (Browning 2008). The degree of accuracy that can be achieved when imputing un-genotyped markers varies greatly depending on the extent of LD between the un-genotyped marker and the nearby genotyped markers (Browning 2008). Imputation algorithms are able to estimate the accuracy of an imputed SNP and so the researcher can discard those that have low estimated accuracies before carrying out association analysis. However, it is important to bear in mind that there will still be some degree of inaccuracy in the remaining imputed SNPs (Browning 2008).

GWAS have produced strongly significant evidence that common polymorphisms influence genetic susceptibility in more than 40 different phenotypes (Manolio et al. 2008) and in the last 3 years, almost 1000 variants associated with a range of human traits and common diseases have been identified using genome-wide methods (Visscher & Montgomery 2009). One of the landmark GWAS to be conducted so far has been the Wellcome Trust Case Control Consortium's (WTCCC) study which scanned 17,000 individuals for seven diseases, including type 2 diabetes (T2D) and bipolar disorder (WTCCC 2007). This study was conducted using the Affymetrix GeneChip 500K Mapping Array, comparing ~2000 cases for each of the disease against a shared sample of ~3000 controls. 24 independent significant association signals were identified in the diseases with P-values < 1 x 10^{-7}, enabling the identification of a number of novel susceptibility genes.

1.3.2.4 Issues of Multiple Testing in Association Studies

In a test of statistical significance, a P-value of 0.05 or less is typically used to indicate that the null hypothesis of no association can be rejected (Sham & McGuffin 2002) as such a value is likely to occur by chance on only 5% of occasions. However, if testing more than 20 SNPs, such a P-value can be expected to occur by chance for at least one variant, assuming all SNPs are independent. In order to reduce the number of false positives, methods are used to adjust the probability estimate for multiple tests. The most commonly used methods are Bonferroni correction, experiment or gene wide adjustments and permutation methods (Hirschhorn & Daly 2005). However, it is likely that these methods are over conservative in the presence of weak, but true genetic effects (Salyakina et al. 2005). The Bonferroni correction, for example, assumes that all tests are independent. This is considered to be too conservative for genetic association tests as there is likely to be a certain degree of LD between some SNPs, and so the P-value should be adjusted for the number of independent SNPs.

Permutation testing provides an empirical method to correct P-values for multiple testing in a way that retains the correlation present in the actual data (Doerge & Churchill 1996). This approach creates a simulated dataset identical to the original except that the case/control labels are randomly permuted in the artificial dataset. By randomly permuting just the individual identifiers, the correlation among genotypes is preserved, as is the number of cases and controls, but any association between genotype and phenotype is broken. The complete set of association tests is then performed on the permuted data, and the permutation process is repeated a preset number of times. This generates a distribution of the best P-value expected in the entire experiment under the null model of no association between genotype and phenotype. For example, if an association has a P-value of 0.001 and a P-value 0.001 or lower is observed 60 times in 1000 permutations, then the corrected empirical experiment-wide P-value is 0.06.

Bayesian methods have also been proposed that can take into account pre-test estimates of the likelihood that a particular variant is truly associated with a phenotype (Wacholder et al. 2004). However, the mathematics behind these methods require knowledge of not only the prior probability of association, but also of the distribution of the size of effects that will be encountered (Cordell & Clayton 2005).

For GWAS, in which many hundreds of thousands of SNPs are tested, it has been suggested that an appropriate threshold for genome-wide significance is $P \leq 5 \times 10^{-8}$ (Pe'er et al. 2008). This is based upon an estimate of 1 million independent tests genome-wide in Europeans (i.e. $P = 0.05/1000000 = 5 \times 10^{-8}$).

However, even after correcting for multiple testing false positives may still remain and so the ideal method for assessing whether a reported association is a true effect or not is for the association to be replicated in an independent sample (Hattersley & McCarthy 2005).

1.3.2.5 Power of Association Studies

In an association study, the power of a study to detect a significant association is affected by the sample size, the significance level required, the effect size and the risk allele frequency in the general population. The more common a risk allele and/or the larger the effect size, the greater the power a study of a fixed sample size has to find a significant association with that variant at a fixed value of P. The odds ratio (OR) of a variant is a measure of its effect size and is defined as the odds of exposure to a susceptibility variant in cases compared to controls. For example, if a variant has an OR of 3, the odds of an individual with a copy of the risk allele being affected by the disease is three times higher than someone without the risk allele.

As discussed previously, DD is a common, complex disease. The common disease/common variant (CDCV) hypothesis has proposed that common diseases are the result of common variants (Reich & Lander 2001). 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. Due to their common nature, however, these variants are also likely to be of weak effect (Wang et al. 2005). This means that large sample sizes will be required within association studies of DD in order to be sufficiently powered to detect these variants of small effects.

As an example of how MAF and OR can affect the power of a study, a sample of 500 cases and 500 controls would be needed to have an 80% chance of detecting a risk allele with an OR of 1.5 and a MAF of 0.1 at a P-value < 0.05 (Dupont & Plummer

1990). However, if the variant had the same effect size but a MAF of 0.01, 4000 cases and 4000 controls would be required to achieve the same level of power. If the MAF remained the same but the OR was 1.3, over 1000 cases and 1000 controls would be required.

These may seem like reasonably manageable samples, however these figures assume a multiplicative model of risk and that either the disease variant is assayed itself or that there is perfect LD with the genotyped marker and the disease variant. If this is not the case then even larger sample sizes may be required to achieve the same power (Wang et al. 2005). Also, multiple testing requires far more stringent P-values to be confident that a true association is being observed, as mentioned previously. With GWAS, a P-value $< 5 \times 10^{-8}$ is deemed to be genome-wide significant (Pe'er et al. 2008). At this level of significance very large sample sizes of more than 5,000 cases and 5,000 controls would be required for 80% power to achieve convincing support for an association with a variant that has a MAF of 0.1 and an OR of 1.3 (Wang et al. 2005).

1.3.3 Identifying Structural Variants

SNPs identified through PCR-based sequencing methods were once thought to be the main source of genetic and phenotypic variation, but the advent of genome scanning technologies has enabled the identification of an unexpectedly large amount of structural variation (e.g. copy number variants or CNVs) in the human genome (Feuk et al. 2006; Stankiewicz & Lupski 2010). In the last few years, this has led to a number of studies investigating the association of CNVs with complex diseases. These types of association studies are discussed in detail in Chapter 7.

1.4 Linkage and Association Studies of Developmental Dyslexia

In order to identify the genes that underlie the genetic predisposition to DD, a number of linkage and association studies have been conducted. The earlier linkage studies have identified a number of regions which may harbour DD susceptibility gene(s). Regions showing replicated evidence of linkage to DD have been named *DYX1-DYX9* by the Human Gene Nomenclature Committee (www.genenames.org). The evidence for linkage and association of each of these regions with DD is discussed below.

1.4.1 *DYX1* (Chromosome 15)

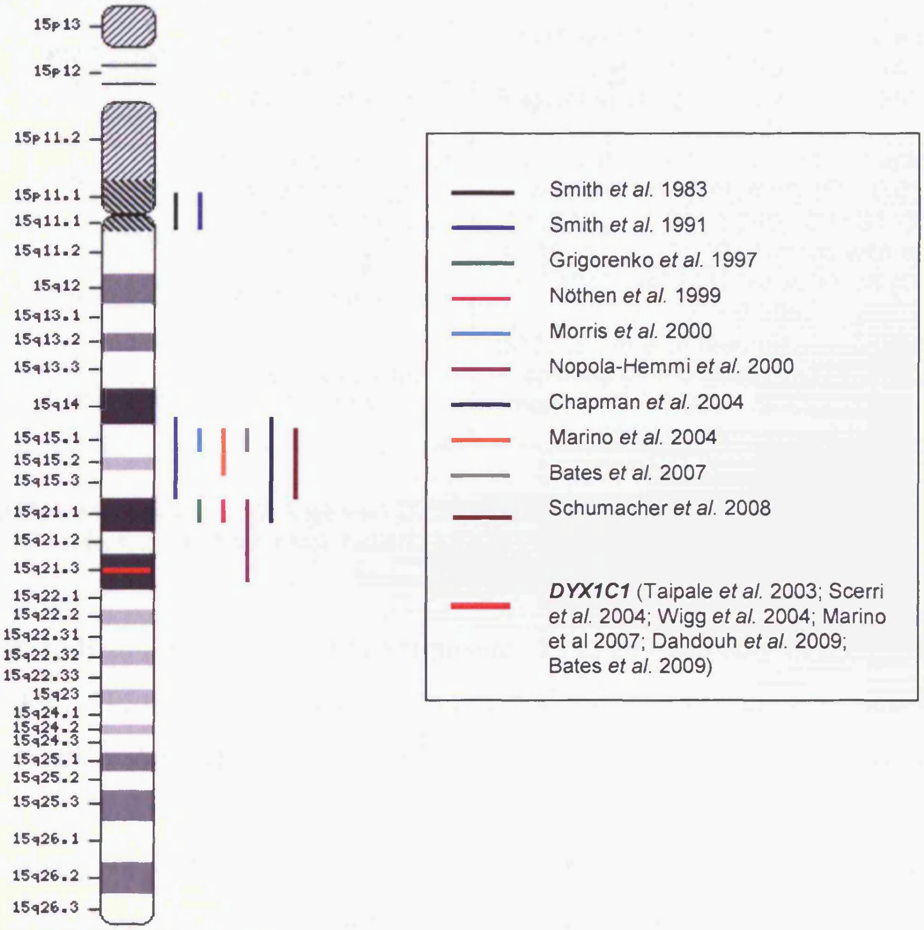


Figure 1.2: Regions on chromosome 15 showing evidence of linkage/association with DD.

Study	Type	Sample	Main Findings
Smith et al. (1983)	Linkage	9 American families (n = 84)[a]	Cen15 (LOD = 3.2)
Smith et al. (1991)	Linkage	9 American families (n = 84)[a]	Qualitative analyses: ynz90 (P = 0.0085)
Grigorenko et al. (1997)	Linkage	6 American Families (n = 94)	SWR (LOD = 3.15)
Nöthen et al. (1999)	Linkage	7 German Families	D15S143 (Parametric LOD = 1.26; NPL 2.19), D15S132 (Parametric LOD = 1.78)
Morris et al. (2000)	Association	101 (stage 1) and 77 (stage 2) UK trios	D15S994/D15S214/D15S146 haplotype (P<0.001 stage 1, P = 0.009 stage 2)
Taipale et al. (2003)	Association	109 cases vs. 195 controls, Finland	DYX1C1 SNPs -3A (P = 0.002) and 1249T (*P* = 0.006), -3A/1249T haplotype (*P* = 0.015)
Chapman et al. (2004)	Linkage	111 American families (n = 898)	Linkage with SWR (D15S143, LOD = 2.34)
Marino et al. (2004)	Association	158 Italian families[c]	D15S214/D15S508/D15S182 (P = 0.005)
Scerri et al. (2004)	Association	264 UK families (n = 1153)	DYX1C1 -3G/1249G haplotype (P = 0.015) with OC choice
Wigg et al. (2004)	Association	148 Canadian families	rs11629841 (P = 0.018, Corrected P = 0.036), -3G/1249G (P = 0.026)
Bates et al. (2007)	Linkage	403 Australian families (n = 980)[d]	Regular spelling linked with D15S994 (P = 0.002)
Marino et al. (2007)	Association	114 Italian probands and 50 sibings[c]	Association of the -3A/1249T haplotype with short term memory (P = 0.0114)
Schumacher et al. (2008)	Linkage	82 German families (n = 331)[b]	D15S182 (LOD = 1.246), D15S143 (LOD 1.310) and D15S1032 linked with spelling
Dahdouh et al. (2009)	Association	66 German trios[b]	rs3743205/rs3743204/rs600753 (G/G/G) (P = 0.006)
Bates et al. (2009)	Association	789 Australian families[d]	rs17819126 with reading (P = 0.0003) and spelling (P = 0.0086); rs3743204 with reading (P = 0.009); rs685935 with short term memory (P = 0.04)

Table 1.2: Evidence for association/linkage with DYX1 (SWR = single word reading, NPL = non-parametric LOD score, OC = orthographic). Letters (a,b,c,d) indicates overlapping samples.

The first study to report linkage of chromosome 15 to DD was carried out by Smith and colleagues (1983). Linkage analysis between reading disability and chromosomal heteromorphisms in American families produced a LOD score of 3.2 for the marker cen15. In an extension of this study, analyses were carried out using both qualitative and quantitative phenotype measures (Smith et al. 1991). Qualitative analyses showed significant linkage of DD with the marker ynz90 (P = 0.0085), however this was not replicated in the quantitative analyses. Grigorenko and colleagues (1997) went on to genotype 6 extended American families (n = 94) using markers in the 15pter-qter region. Five theoretically derived phenotypes were used in the linkage analysis: 1) phonological awareness (PA); 2) phonological decoding (PD); 3) rapid automatized naming (RAN); 4) single-word reading (SWR); 5) discrepancy between intelligence and reading performance (DISC). Significant linkage was found for a marker on chromosome 15q21.1 (D15S143, LOD = 3.15) with the SWR phenotype. Linkage of this particular phenotype to the same region was also found in an American sample by Chapman and colleagues (2004), with a single point LOD score of 2.34 for the D15S143 marker. A previous study also found linkage to this marker (Nöthen et al. 1999). This study used a spelling phenotype to diagnose dyslexic patients in a German sample and genotyped 13 microsatellite markers across chromosome 15. Significant linkage was found using both parametric (LOD = 1.78, P = 0.0042) and non-parametric analysis (LOD = 2.19, P = 0.03). Although different phenotypes have been used in these studies, it is perhaps not surprising that they have both shown linkage to the same region because spelling and reading disability have been shown to be strongly correlated (Malmquist 1958).

Morris and colleagues (2000) found association between DD and a 3-marker haplotype on chromosome 15q15.1 (D15S994/D15S214/D15S146; P <0.001, corrected P = 0.03) in a UK sample. This was then replicated in another sample from the UK (P = 0.0091). The marker D15S944 was also found to be linked to DD in a genome-wide study using an Australian sample (LOD = 1.89, P = 0.002) (Bates et al. 2007). Another 3 marker haplotype in this region (D15S214/D15S508/D15S182) was shown to be associated with DD in an Italian population (P = 0.005) (Marino et al. 2004). Although this study did not find any significant association with the marker D15S994, this marker lies within the region covered by the significant haplotype and therefore still provides further evidence for an association with DD within this region. D15S994 is within a phospholipase gene, Phospholipase C β 2 (*PLCB2*) and is 1.6 Mb from another phospholipase gene, Phospholipase A_2, group IVB (*PLS2G4B*). Morris and colleagues (2004) went on to test sequence variants within these genes for an association with DD in their UK case-control (164 cases vs 174 controls) and family-based samples (178 trios). In the case-control sample, one variant within *PLCB2* (PCLB2 no.9, P = 0.038) and two variants within *PLA2G4B* (PLA2G4B no.8, P = 0.049; PLA2G4B no. 26, P = 0.048) showed association with DD, but none of the variants showed association in the trios. This difference in results may have been due to low power of the trios sample to detect significant association. However as it was this sample that had previously shown replication for a significant association between DD and D15S944, it should have been powerful enough to detect the variants causing this association, therefore these variants cannot account for the association signal originally observed.

The region surrounding the marker D15S143 was linked with dyslexia through a study by Nopola-Hemmi and colleagues (2000). They identified 2 Finnish families with different balanced translocations, the breakpoints of which were both mapped to a 6-8 Mb region between markers D15S143 and D15S1029. In one family, the translocation t[2;15][q11;q21] was present in the father and two of his children who had DD. The mother and other child had normal karyotypes and normal reading abilities. However, it should be noted that one of the children diagnosed with DD also had low intelligence. The second family showed the translocation t[2;15][p13;q22] in the father and all three children. The evidence for a link between this translocation and DD is less convincing. The father and the oldest child had cornea plana, and the oldest child had reading disability whereas the two younger siblings carrying this translocation had no symptoms of DD. There was also no history of learning difficulties in the father. As the translocation does not completely segregate with the presence of DD, even in this family alone, it is unlikely to be linked to the disease.

The translocation in the first family (t[2;15][q11;q21]) was shown to disrupt the gene *DYX1C1* (Taipale et al. 2003). 8 SNPs within this gene were tested for an association with DD using 109 cases and 195 controls from Finland. Two of the SNPs were found to be significantly associated. The SNP rs3743205 (-3 G>A) gave an OR of 3.2 (95% CI 1.5 - 6.9, P = 0.002). This SNP is three bases 5' to the ATG translational start site and disrupts a predicted Elk-1 transcription factor binding site. The SNP rs57809907 (1249 G>T) gave an OR of 2.3 (95% CI 1.2 - 4.2, P = 0.006). This SNP introduces a premature stop codon and is predicted to truncate the protein by 4 amino acids. These SNPs also showed significant association when present as the haplotype -3A:1249T (P = 0.015). Further studies have provided conflicting results for the association of rs3743205 and rs57809907 with DD. They were also found to be associated with an orthographic choice (OC) measure by Scerri and colleagues (2004) in a UK sample, but as the haplotype -3G:1249G (P = 0.0158). This was again found to be the case in a Canadian sample by Wigg and colleagues (2004). Dahdouh and colleagues (2009) identified a 3 marker haplotype with these SNPs and rs3743204 (G/G/G, P = 0.006) in the female subset of their German trio sample. Four other studies tested these SNPs for an association with DD in Italian (Bellini et al. 2005; Marino et al. 2005), UK(Cope et al. 2005b), American (Meng et al. 2005a) and Australian (Bates et al. 2007) samples, but none of them found significant evidence. However, a later study by Marino and colleagues (2007) on the same Italian sample found association for the haplotype -3A:1249T (P = 0.0114) with a measure of short term memory, but did not find significant association with other subphenotypes of DD. The case control sample used by Taipale et al. (2003) came from just 23 families, 33 unrelated case/control pairs and 1000 population controls, but no adjustment for relatedness was carried out. The non-independence of alleles in different related individuals from the same family may have distorted the evidence for association and so this may be why this result has not yet been convincingly replicated in subsequent studies.

Wigg and colleagues (2004) also found significant association to a SNP in *DYX1C1* that had not been tested in the previous studies. The SNP rs11629841 showed association with DD both alone (P = 0.036 corrected for multiple testing) and as part of a haplotype with rs3743204 (haplotype C/G, P = 0.0089) and with rs692691 (T/T, P = 0.0058; G/T, P = 0.0389). However, this was not replicated by Cope and colleagues (2005b). They genotyped the marker rs11629841 in a sample of 247 UK parent proband trios and did not find any significant association with DD. This sample had previously shown association with a marker outside of *DYX1*, D15S994 (Morris et al. 2000), so Cope and colleagues (2005b) carried out LD analysis between this marker and the SNPs rs3743205, rs11629481 and rs57809907 within *DYX1C1*, but no significant LD was observed (P-values all >0.25). Together with the fact that D15S994 is 15 Mb away from *DYX1C1,* it is unlikely that the observed association between chromosome 15 and DD in this particular sample is due to *DYX1C1.*

Recently, Bates and colleagues (2009) looked for association with subphenotypes of reading ability in their sample of 789 Australian families that had not been selected for a DD phenotype. They found association of three other SNPs within *DYX1C1* with measures of irregular-word reading (rs17819126, P = 0.02), non-word reading (rs17819126, P = 0.0003; rs3743204, P = 0.0089), irregular-word spelling (rs17819129, P = 0.0086) and short term memory (rs685935, P = 0.04). The SNP rs17819129 has not been typed in previously reported studies and codes for a non-synonymous protein sequence alteration. This SNP is in complete LD with SNPs in two nearby genes, *RAB27* and *C15orf5* suggesting that these genes may also warrant further investigation within this linkage region.

The evidence for *DYX1C1* as a susceptibility gene for DD has not been convincingly replicated. It seems that this gene is only significantly associated in certain populations, such as Finland. Due to the small sample size studied and the large number of related individuals in that sample, it could be that the results were false positives. However, it is also possible that different causal variants within this gene exist in the different populations that have been studied so far and differences in language or variations in LD could be causing the differences in results. The association for the -3A/1249T haplotype was originally found in a Finnish population (Taipale et al. 2003). This population is a genetic isolate that was established around 10,000 years ago, with a limited number of founders, and has since gone through several bottlenecks (Arcos-Burgos & Muenke 2002). In such populations, regions of linkage disequilibrium will generally extend further. This could result in a causal genetic variant in a Finnish population being in significant LD with a particular genetic background (such as the -3A/1249T haplotype) that is relatively far away. Other more heterogeneous populations would have weaker LD in these regions and so would not necessarily pick up these associations. Another explanation is that these risk alleles may be in LD with a causal variant that has different founders, which could account for an association in different directions being observed in different populations.

Further evidence supporting *DYX1C1* as a possible susceptibility gene for DD comes from functional studies which have suggested that this gene may have a role within neuronal migration, as discussed in Chapter 3, section 3.1.1.

Two other neurodevelopmental disorders have been linked with chromosome 15q. Bakker and colleagues (2003) carried out a genome wide scan using 164 Dutch sib pairs diagnosed with ADHD which shares comorbidity with DD (see section 1.1.5.1). The most promising chromosome region was 15q, with the marker D15S944 producing the highest single point LOD score (3.37). This particular marker had been found to show significant association with DD in the study by Morris and colleagues (2000). Smith and others (2005) used a sample of 86 sib pairs from 65 families in a study that linked this region to phonological memory (D15S1017-D15S1029, $Z_{max} = 2.31$) and articulation (D15S1017-D15S1029, $Z_{max} = 2.719$). Speech-sound disorder (SSD) is a common childhood disorder which is characterised by 'developmentally inappropriate errors in speech production that greatly reduce intelligibility' (Smith et al. 2005). There is evidence to suggest that children with SSD also have phonological processing problems as discussed in section 1.1.5.5. Stein and colleagues (2006) also analysed the 15q14-q21 region for linkage with SSD, and obtained the most significant results at the marker D15S214 ($P = 0.0072$) which again showed significant association with DD in the study by Morris and others (2000). These results suggest that DD, ADHD and SSD share some genetic aetiology within this region and further investigation of the genes affected may be able to provide more information about why these disorders share a high level of comorbidity.

1.4.2 *DYX2* (Chromosome 6p)

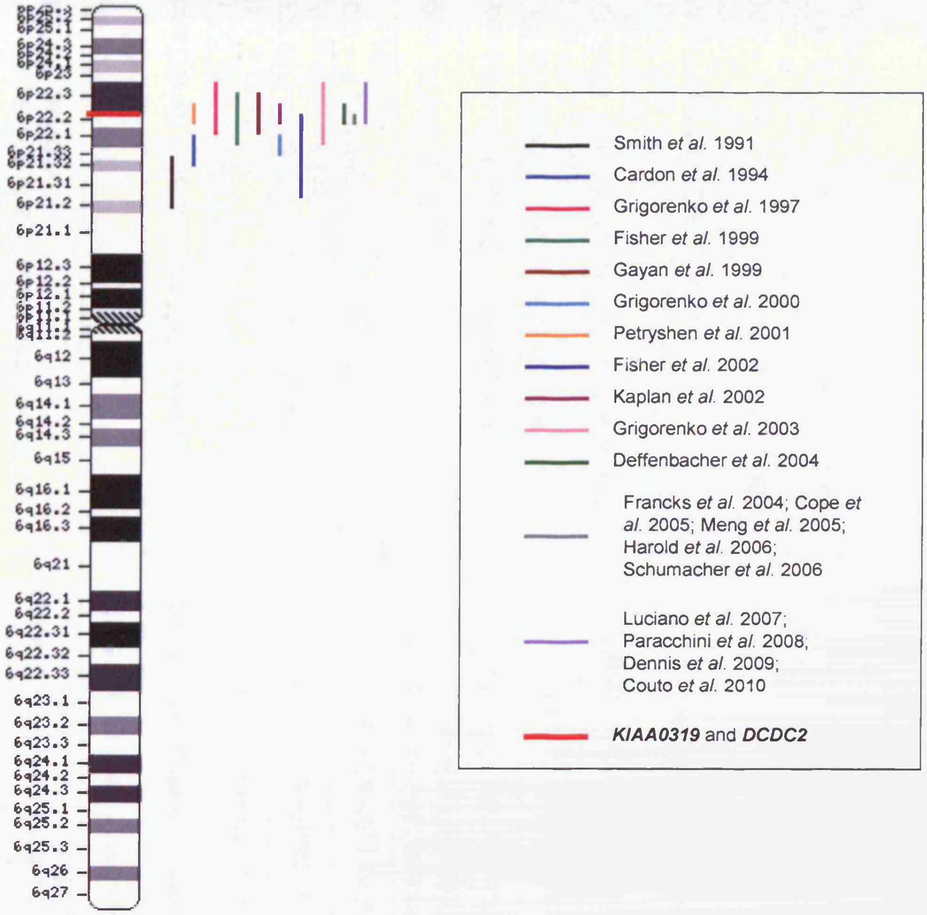


Figure 1.3: Regions on chromosome 6p showing evidence of linkage/association with DD

Study	Type	Sample	Main Findings
Smith et al. (1991)	Association	18 American families[a]	Qualitative analysis, GLO P = 0.0153; Quantitative analysis, BF P < 0.001
Cardon et al. (1994)	Linkage	Sibling sample from 19 American families (n = 358)[a], dizygotic twin sample from Colorado twin study (n = 50 pairs)[b]	QTL between D6S105 and TNFB (sibling P = 0.066; twin P < 0.00001; combined P <0.0001)
Grigorenko et al. (1997)	Linkage	6 extended American families (n = 94)[c]	Multipoint P < 10^{-6} for PA for markers D6S108, D6S461, D6S299, D6S464 and D6S306
Fisher et al. (1999)	Linkage	181 sib pairs from 82 nuclear UK families[d]	Linkage in D6S1660-D6S291 for OC and PC (P = 0.038 to 0.00035)
Gayán et al. (1999)	Linkage	79 American families (126 sib pairs)[b]	Linkage in region D6S276-D6S105 with OC (LOD = 3.10), PD (LOD = 2.42), PA (LOD = 1.46)
Grigorenko et al. (2000)	Linkage	8 American families (n = 171)[c]	Significant P-values within D6S464-D6S306 for the SWR, vocabulary and spelling (IBD analyses)
Fisher et al. (2002)	Linkage	89 UK families (195 sibling pairs)[d]	Linkage with PD (D6S276, singlepoint P = 0.00006; D6S1610, single point P = 0.00001)
		119 American families (180 sibling pairs) from Colorado twin study[b]	Linkage with PD (D6S276, singlepoint P = 0.002)
Kaplan et al. (2002)	Linkage	104 American families (n = 392)[b]	Linkage with OC and PD in JA04 region (P = 0.05 - 0.00049)
Grigorenko et al. (2003)	Linkage	8 American families (n = 176)[c]	Linkage with D6S299 (LOD = 2.01 for the PA/PD/SWR pathway) and D6S222 (LOD = 2.57 for PA)
Deffenbacher et al. (2004)	Linkage	349 American families (n = 1559)[b]	Linkage with 5 phenotypes (PA, PD, SWR, OC, DISC) over interval D6S1597 to D6S1571
	Association	114 American families[b]	VMP (P = 0.05-0.004), DCDC2 (P = 0.05-0.001), KIAA0319 (P = 0.03), TTRAP (P = 0.03-0.008) and THEM2 (P = 0.008).
Francks et al. (2004)	Association	89 UK families[e]	rs1061925 (P = 0.0269 for OC)
		175 UK families[e]	rs9467247 (P = 0.0006 for OC-irreg, P =0.0003 for READ); rs1061925 (P = 0.0005 for OC-choice, P = 0.0008 for READ)
		159 American families[b]	rs9467247 (P = 0.0038 for READ, P = 0.042 for PA); rs3033236 (P = 0.0023 for READ, P = 0.015 for SPELL)

Table 1.3: Evidence for association/linkage with *DYX2* (PA = phonological awareness, PD = phonological decoding, OC = orthographic coding, NWR = non word reading, SWR = single word reading, DISC = discrepancy between IQ and reading ability). N.B some of these studies have overlapping samples. Letters (a,b,c,d,e,f) indicate sample overlap.

Study	Type	Sample	Main Findings
Cope et al. (2005a)	Association	240 cases vs. 312 controls, UK[f]	15 SNPs associated in KIAA0319/TTRAP/THEM2 locus (P ≤ 0.05)
		223 cases vs. 273 controls, UK[f]	7 SNPs associated in KIAA0319/TTRAP/THEM2 locus (P ≤ 0.05)
		143 parent-proband trios and 223 cases vs 273 controls, UK[f]	KIAA0319 (rs4504469, P = 0.002; rs2173515, P = 0.007; rs6935076, P = 0.006), MRS2L (rs2793422, P = 0.003), THEM2 (rs3777664, P = 0.008), intergenic (rs1053598, P = 0.02)
Meng et al. (2005b)	Association	153 American families (n = 536)[b]	rs807724 in DCDC2 (P = 0.0003) associated with DISC
Harold et al. (2006)	Association	264 nuclear families and 350 cases vs. 273 controls, UK[e,f]	Association within intron 1 of KIAA0319 (rs2038137, P = 0.00002)
Schumacher et al. (2006a)	Association	137 triads, Germany	rs793862 in DCDC2 (P = 0.011) and D6S276 (P = 0.004)
		239 triads, Germany	Association with haplotype A-C at rs7938620-rs807701 (P = 0.001)
Luciano et al. (2007)	Association	440 Australian families unselected for DD (n = 858)	TTRAP (rs2143340; P = 0.009), KIAA0319 (rs6935076; P = 0.008), haplotype spanning KIAA0319 and TTRAP (rs4504469/rs2038137/rs214340)
Paracchini et al. (2008)	Association	10,261 UK children unselected for reading ability	rs2143340 with reading (P = 0.003), spelling (P = 0.008) and NWR (P = 0.03); rs4504469/rs2038137/rs2143340 (A/A/C) haplotype with reading (P = 0.009) and spelling (P = 0.03)
Couto et al. (2010)	Association	291 Canadian families	Qualitative analysis: VMP (rs9356928 P = 0.034; rs4285310, P = 0.048; rs3178 P = 0.043), KIAA0319 (rs6935076 P = 0.014), TTRAP (rs3181238 P = 0.031). Quantitative analysis: KIAA0319 (rs6935076, P = 0.025 for spelling), VMP (rs3178, P = 0.043 for PD) Haplotype analysis: KIAA0319, rs4504469-rs6935076 (G-A) P = 0.018
Dennis et al. (2009)	Association	264 UK families, 126 in severe subset[e]	Association of rs9461045, rs3212236 and rs9467247 with measures of OC, PD, reading and spelling

Table 1.3 Continued.

An association of chromosome 6p21 with DD was first identified by Smith and colleagues (1991). They used a sample of 18 American families and looked for linkage between four markers (BF, GLO, thh157 and 2C5) and DD using both quantitative and qualitative measures of the trait. For the qualitative analysis, the marker GLO (glyoxylase 1) showed significance (P = 0.0153). However, for the quantitative analyses, the marker BF (properdin factor) rather than GLO was significant (P < 0.0001).

Cardon and colleagues (1994) targeted the 6p21 region in their study which used an American kindred sample comprising 358 individuals from 19 families, and a twin sample consisting of 50 families from the Colorado twin study. They genotyped the markers used by Smith and colleagues (1991), as well as five other markers that were more informative due to higher heterozygosities. They found significant linkage in this region between the markers D6S105 and TNFB (tumour necrosis factor beta) in the twin sample (twin P < 0.00001; kindred P = 0.066). Combining both the twin and kindred samples reinforced this finding (P < 0.0001). Grigorenko and colleagues (1997) used six extended American families and genotyped markers in the 6p23-p21.3 region using the 5 phenotypes described earlier. Two point nonparametric analyses revealed significant P-values for five markers relatively close to each other (D6S109, D6S461, D6S464, D6S306, and D6S276). However, the results varied with the phenotype used and the most statistically significant results were obtained for the PA phenotype. In an extension of this study involving another 2 families (Grigorenko et al. 2000), IBD analyses showed significant P-values in the region D6S464-D6S306 but only for the phenotypes SWR, vocabulary and spelling with little evidence for PA and PD. Grigorenko and colleagues later went on to use this sample of 8 American families to identify linkage to 6 phenotypes; PA, PD, RAN and SWR as before and the PA/PD/SWR and PA/RAN/SWR pathways (Grigorenko et al. 2003). They observed the most significant linkage with the markers D6S299 (LOD = 2.01 for the PA/PD/SWR pathway) and D6S222 (LOD = 2.57 for PA). Overall, this study highlighted three interesting regions within 6p21.3, all of which have been replicated in other studies. The first is the D6S109-JA01 region (replicated by Turic et al. (2003)), the second being the D6S299-D6S1261 region (also implicated by Kaplan et al. (2002)) and the third is the D6S105-D6S265 (as found by Cardon et al. (1994), Fisher et al. (1999) and Grigorenko et al. (2000)).

Fisher and colleagues (1999) also used quantitative phenotypes in their study involving 181 sib pairs from 82 families in the UK. They found evidence of linkage in the D6S1660 - D6S291 region for OC and PC phenotypes (P-values between 0.038 and 0.00035), with the highest significance observed for the marker D6S276. Gayán and colleagues (1999) also found evidence for these phenotypes in a sample of 79 American families. They obtained large LOD scores for OC (LOD = 3.10), PD (LOD = 2.42) and PA (1.46) within the region D6S276-D6S105. An extension of these studies, involving a whole genome linkage study of both the US and UK samples, was carried out by Fisher and colleagues (2002). In the 6p21 region, the most significant results from the single point analyses were obtained for PD in the UK sample (D6S276, P = 0.00006; D6S1610, P = 0.00001). These were replicated in the US sample, although with less significance (D6S276, P = 0.002). Kaplan and colleagues (2002) used this US sample and 11 quantitative phenotypes in their analyses of 29 markers in the 6p21.3 – 6p22 region. All phenotypes yielded evidence for association to 6p ($P < 0.05$), however they obtained the most significant results for orthographic and phonological processes (P-values between 0.05 and 0.00049), with the most likely location of the quantitative trait locus being within a 4-Mb region surrounding the marker JA04 (P = 0.0021 for OC). Turic and colleagues (2003) used two separate samples of UK proband/parent trios in a two-stage study involving 21 microsatellite markers covering an 18 cM region on chromosome 6p. The three marker haplotypes D6S109/D6S422/D6S1665 and D6S506/D6S1029/D6S1660 showed the most association across both of the samples, with the most significant haplotype (D6S109/D6S422/D6S1665) showing association with subphenotype measures of SWR, spelling, PA, PD, orthographic accuracy and RAN. These results suggest a broad region of association spanning the markers D6S109 to D6S1260.

The region 6p21.3-22.3 has been widely replicated in American and UK samples (see Figure 1.3 and Table 1.3). Several of these studies have focused on quantitative sub-phenotypes of dyslexia, and those involving orthographic and phonological processes have shown the most amount of evidence for an association with this region. However, a number of other studies have been carried out that have not found significant association of this region with phenotypes of DD. Field and Kaplan (1998) used a sample of 79 Canadian families and a qualitative diagnosis of 'Phonological Coding Dyslexia' (PCD). Subjects were assigned to one of five categories: definitely affected, probably affected, uncertain, probably unaffected, and definitely unaffected. However no significant results were found. This may have been due to the qualitative nature of the phenotype they used, whereas other studies have used more quantitative analyses for subphenotypes. This group then extended their study and used four quantitative measures: PA, PD, RAN and spelling (Petryshen et al. 2000), but still no significant association was found. Chapman and colleagues (2004) attempted to confirm linkage in this region using a sample of 111 American families and continuous measures of PD and SWR in a genome wide scan. Their results showed only weak evidence of linkage of PD with chromosome 6p in a region that was ~10 cM distal to the regions previously reported. Studies in other populations have also failed to find an association of this region to DD, including a genome-wide study carried out using a Norwegian sample showing impaired orthographic and phonological processing abilities (Fagerheim et al. 1999), and a study using a German sample (Nöthen et al. 1999). These differences are again likely to be due to population heterogeneity and differences in diagnostic criteria, especially when considering the German sample which was selected on the basis of spelling ability rather than difficulties in reading.

There is still a great deal of evidence for this region however, and a number of studies have sought to identify possible candidate genes lying within this region. Deffenbacher and colleagues (2004) used a sample of 1,559 individuals from 349 nuclear families from the Colorado twin study to refine the region of linkage. Both single-point and multi-point analyses showed significant linkage with all of five phenotypes (PA, PD, SWR, OC and DISC) over the interval D6S1597 to D6S1571, with maximal linkage converging between markers D6S276 and D6S1554. Of 12 genes within this region, 10 were tested for an association with DD in a subset of 114 families. Five of these genes showed evidence of association: *VMP* (P = 0.05-0.004), *DCDC2* (P = 0.05-0.001), *KIAA0319* (P = 0.03), *TTRAP* (P = 0.03-0.008) and *THEM2* (P = 0.008). All of these genes are expressed in the central nervous system and so would make good candidates for further screening. *VMP* is a neuron-specific vesicular membrane protein thought to play a role in vesicular organelle transport and neurotransmission (Cheng et al. 2002). *DCDC2* is expressed ubiquitously and contains two doublecortin peptide domains that were originally described in the doublecortin gene (*DCX*) encoded on the X chromosome. *DCX* encodes a cytoplasmic protein that directs neuronal migration by regulating the organisation and stability of microtubules and is mutated in X-linked lissencephaly (a neuronal migration defect) (Dobyns et al. 1999) and double cortex syndrome, which is caused by arrested migration halfway to the cortex, producing a 'double cortex'. *KIAA0319* is highly expressed in the brain and codes for a novel protein of an unknown function. The predicted KIAA0319 protein contains four polycystic kidney disease (PKD) domains that have an immunoglobulin-like fold, originally found in the PKD1 protein (Bycroft et al. 1999). PKD domains have been implicated in cell-cell adhesion processes (Ibraghimov-Beskrovnaya et al. 2000). *TTRAP* encodes a tumour necrosis factor receptor-associated protein known to inhibit the activation of nuclear factor-kappa B (NF-κB) and subsequent down-stream activation of transcription (Pype et al. 2000). Activation of NF-κB transcription has been shown to play a role in long-term potentiation and synaptic plasticity associated with learning and memory. *THEM2* encodes a protein belonging to the thioesterase superfamily that catalyses the hydrolysis of long-chain fatty acyl-CoA thioesters. Abnormal fatty acid metabolism has been suggested to play a role in a spectrum of neurodevelopmental disorders, including dyslexia.

The Colorado sample (159 families) was also used by Francks and colleagues (2004), as were two samples from the UK (89 and 175 families). They first refined the linked region on chromosome 6p to 5.8Mb (LOD = 3.48). Within this region, 8 genes are expressed in the brain which separate into 4 clusters. Cluster 1 contains the genes *ALDH5A1, KIAA0319, TTRAP,* and *THEM2*. Cluster 2 contains the gene *C6orf32*, cluster 3 contains the gene *SCGN* and cluster 4 contains the genes *BTN3A1* and *BTN2A1*. Francks and colleagues (2004) analysed 15 SNPs within these genes using a sample of 89 UK families, with significant association being found for the SNP rs1061925 (P = 0.0269 for OC). 42 additional SNPs surrounding this SNP were then analysed with the combined UK samples, revealing evidence of association for 21 SNPs which were then typed in the Colorado sample. In these analyses, evidence for association was found when the samples were selected for more extreme ends of the phenotype. A specific haplotype (tagged by the SNPs rs4504469, rs2034469, and rs2143340) spanning the genes *KIAA0319, TTRAP* and *THEM2,* was found to be associated with DD in both the UK and US families. Francks and colleagues (2004) then tried to identify variants within these genes using 32 probands with severe DD, however the only SNP identified that had an effect on protein sequence was rs4504469 within exon 4 of *KIAA0319*. The minor allele frequency of this SNP was 0.47 in these samples and was not unique to the risk haplotype identified.

Cope and colleagues (2005a) also analysed this region, using 137 SNPs. A UK case control sample of 240 cases and 312 controls was used to identify significant markers and these were followed up in a sample of 143 parent-proband trios. In both samples, Cope and colleagues (2005a) found evidence for association with three SNPs in *KIAA0319* (rs4504469, P = 0.002; rs2173515, P = 0.007; rs6935076, P = 0.006), with one SNP in *MRS2L* (rs2793422, P = 0.003) and in *THEM2* (rs3777664, P = 0.008) and with an intergenic SNP (rs1053598, P = 0.02). They also found strong evidence that the association observed within this region was due to the *KIAA0319* SNPs rs4504469 and rs6935076. A haplotype of these two SNPs (A/G) was found to be highly significantly associated with DD in both the case-control sample (P = 0.00003) and the trio sample (P = 0.006). A recent study conducted by Couto and colleagues (2010) also found evidence for a significant association of these two SNPs with DD in their sample of 291 Canadian families, but as the haplotype G/A only (P = 0.018). This particular haplotype was also associated with DD in the study by Cope and colleagues (P = 0.02), but was not as significant as the A/G haplotype.

Luciano et al. (2007) genotyped 10 SNPs in or near to the *KIAA0319* gene using a sample of 440 Australian families that were unselected for DD but were tested using reading and spelling tasks. They found significant association with reading ability for SNPs within the *TTRAP* (rs2143340; P = 0.009) and *KIAA0319* (rs6935076; P = 0.008) genes, and for a three-SNP haplotype that spans *KIAA0319* and *TTRAP* (rs4504469/rs2038137/rs2143340, global P = 0.005). This is the same 3 marker haplotype that was found to be associated with DD by Francks and colleagues (2004). Franks and colleagues found the most significant association with the 1-1-2 individual haplotype and a measure of orthographic coding (P = 0.0007). Luciano and colleagues (2007) also found significant association with this haplotype (P = 0.04 for a bivariate analysis of whole word reading), but in the opposite direction. Luciano and colleagues (2007) found a higher level of significance with the 1/1/1 and 2/2/2 haplotypes with a univariate analysis of principle components of reading (P = 0.02 for both haplotypes). The 1/1/1 haplotype was found to be associated with phonological decoding in the UK sample studied by Francks and colleagues (2004) (P = 0.031) and was also associated in the study by Cope and colleagues (P = 0.03).

Another study conducted using individuals that had been not been selected for DD was carried out in the UK by Paracchini and colleagues (2008). They also found significant association of the *TTRAP* SNP rs2143340 with measures of reading (P = 0.003), spelling (P = 0.008) and non-word reading (P = 0.03) as well as association with the same three marker haplotype rs4504469/rs2038137/rs2143340 as the 1-1-2 haplotype with measures of reading (P = 0.009) and spelling (P = 0.03). The other forms of this three marker haplotype were not tested in this study. Whilst this three marker haplotype has shown association in a three independent samples, the individual risk haplotypes do not appear to be the same in all samples. The 1-1-2 risk haplotype had an opposite direction of effect in the Australian sample (Luciano et al. 2007) compared with the two UK samples (Francks et al. 2004; Paracchini et al. 2008). This may be indicative of a false positive in the Australian sample or it could be a consequence of the sample being ethnically heterogeneous as only ~82% of this sample was reportedly of Anglo-Celtic origin (Luciano et al. 2007). Despite this, Paracchini and colleagues (2006) showed that this risk-haplotype reduces the expression of *KIAA0319*, making it a good functional candidate for increasing an individual's susceptibility to DD.

Dennis and colleagues (2009) tried to identify variants upstream of *KIAA0319* within the region spanned by the *TTRAP/KIAA0319* haplotype that may affect the expression of *KIAA0319*. They identified 7 risk SNPs within this haplotype region and found significant association of 3 of these SNPs (rs9461045, rs3212236 and rs9467247) with orthographic choice and spelling in their sample of 264 families in the UK. When they selected for a subset of severe cases (126 families), the significance of these SNPs increased and they were also found to be significantly associated with measures of phonological decoding and reading. Dennis and colleagues (2009) produced luciferase-expressing constructs containing the region upstream of *KIAA0319* to demonstrate that the minor allele of rs9461045 confers reduced luciferase expression in both neuronal and non-neuronal cell lines. This suggests that the minor allele of this associated variant reduces the expression of *KIAA0319* and is therefore likely to be functionally relevant for the development of DD (Dennis et al. 2009).

Another study using subjects from the Colorado sample (153 nuclear families) found association with another gene in this region (Meng et al. 2005b). 147 SNPs in the 1.5 Mb region surrounding the marker JA04 (the marker that this group previously found to be associated with DD (Kaplan et al. 2002)) were genotyped. The strongest evidence was found for the SNP rs807724 located in intron 6 of the gene *DCDC2* (P = 0.0003). This gene is 500 kb away from the JA04 marker, so is within the replicated *DYX2* region.

Schumacher and colleagues (2006a) used a categorical definition of DD based on spelling abilities. Using two independent trio samples of German families, they found association with extreme spelling disability in two SNPs within the *DCDC2* gene (rs793862, P = 0.011; rs807701, P = 0.058), most significantly as a two-marker haplotype in intron 7 (P < 0.0001).

A collaboration of groups from Oxford and Cardiff attempted to replicate the association of *DCDC2* to DD using two UK samples (Harold et al. 2006). The sample from Oxford included 264 unrelated nuclear families, while the Cardiff sample included 350 cases and 273 controls. Both samples were used to genotype those polymorphisms that showed the most significant association in the previous US (Meng et al. 2005b) and German (Schumacher et al. 2006a) samples. In the Oxford sample, nominally significant associations were detected with several traits, with the strongest association being found between the marker rs1087266 and the PA phenotype (P = 0.005). However, when this sample was selected for more severe phenotypes, these associations were no longer significant. No association with DD was observed for any of the *DCDC2* polymorphisms in the Cardiff sample. These groups then went on to genotype new polymorphisms in or flanking the *KIAA0319* gene. In total, 5 SNPs were significantly associated with DD in both samples. After combining the P-values of these five SNPs, the most significant association with DD was with the SNP rs2038137 in intron 1 (P = 0.00002). Another 4 SNPs in the 5' flanking region or intron 1 showed association, but only in one of the two samples. Despite finding no association with *DCDC2* in their samples, these groups tested for statistical interactions between markers in this gene and the five SNPs in *KIAA0319* that showed association in both samples. They found significant interactions between variants of the two genes, the most significant being between rs793862 in *DCDC2* and rs761100 in *KIAA0319* (P = 0.007).

The association of variants in this region with reading ability in populations that have not been selected for DD suggests that variants within the genes in this region may influence reading ability in the general population. However, the findings that the significance of the associations within these genes increase when selecting a sub-set of samples containing the more severe cases of DD (Francks et al. 2004; Harold et al.

2006; Schumacher et al. 2006a; Dennis et al. 2009) suggests that there may also be specific functional variants within these genes that can cause DD.

Overall, the genes *KIAA0319* and *DCDC2* have received the most support in this region, being widely replicated in a number of studies. Evidence for an association between DD and *KIAA0319* has been found in UK, US and Australian samples, while evidence for an association to *DCDC2* has been found in UK, US and German samples. The differences in results across studies may be due to the different populations studied, differing ascertainment criteria, and different marker sets being used, making comparisons between them difficult. Even those studies that have used subsets from the same population in Colorado have obtained different results. Francks and colleagues (2004) found association with *KIAA0319,* Meng and colleagues (2005b) found association with *DCDC2,* while Deffenbacher and colleagues (2004) found association with both genes. These differences may have been due to different sampling criteria being used by these groups when deciding on which subsets to use.

KIAA0319 and *DCDC2* share a number of similarities; they are physically close on chromosome 6p, are both expressed in the brain and have similar putative functions. As these genes are so close together, it could be the case that the associations detected in both genes are due to a single mutation that has not yet been identified (Paracchini et al. 2007). This would explain why this particular region has been widely replicated in samples from different populations. However, Harold and colleagues (2006) reported that there is a significant lack of LD across and between these two genes, so this theory is unlikely to be true. It could be that different subgroups of individuals with DD are determined by the effects of either one or the other gene.

Recently, Couto and colleagues (2010) attempted to map acetylated histones in this region using chromatin immunoprecipitation coupled with genomic tiling arrays (ChIP-chip). Acetylated histones are frequently associated with accessible chromatin at genomic regions containing regulatory elements (Eberharter & Becker 2002; Kurdistani et al. 2004; Roh et al. 2005; Heintzman et al. 2007; Roh et al. 2007) and therefore identifying these regions provides functional clues as to the location of genomic sequences involved in gene regulation. Couto and colleagues (2010) identified several regions marked by acetylated histones that mapped near to associated markers in this locus, including intron 7 of *DCDC2* and the 5' region of *KIAA0319*.

Both genes have been implicated in neuronal migration (discussed in Chapter 3, section 3.1.1), making them good functional candidate genes and overall, these genes have received the most support as candidate genes for DD across the whole genome. At this stage no gene within this region can be ruled out and the consistent replication of findings in this region make it a worthwhile area of the genome to explore further.

1.4.3 *DYX3* (Chromosome 2)

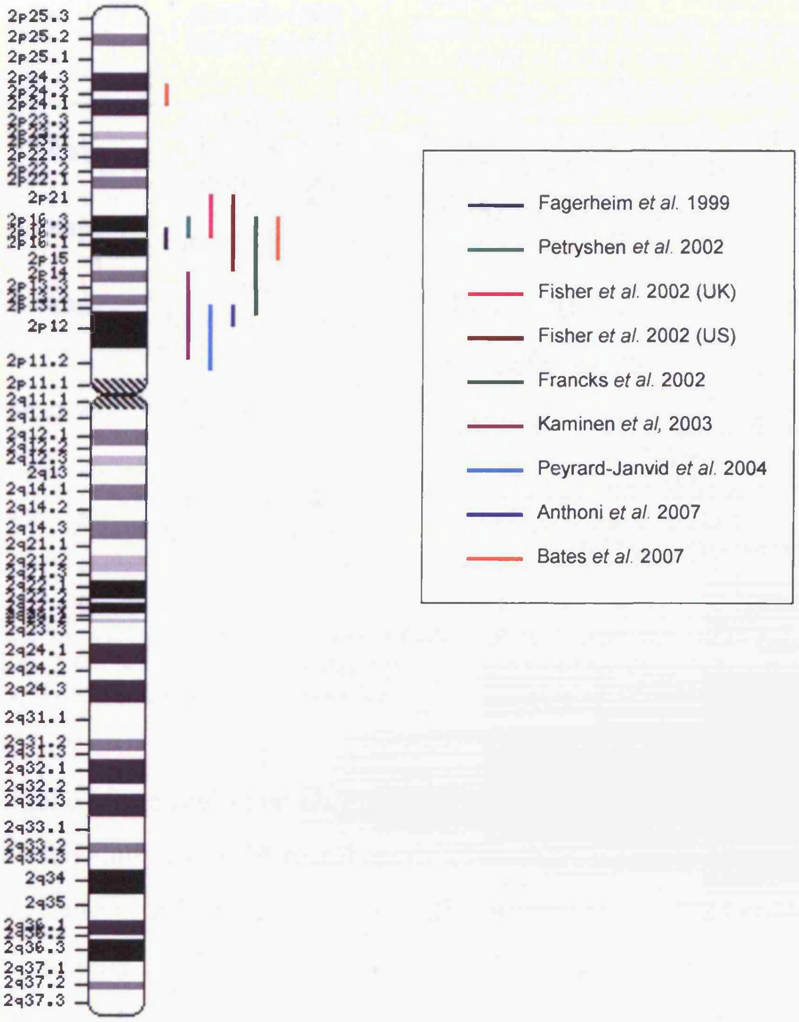


Figure 1.4: Region on chromosome 2p showing evidence of linkage/association with DD

Study	Type	Sample	Main Findings
Fagerheim et al. (1999)	Linkage	1 Norwegian family (n = 36)	D2S2183 (Z = 5.53, θ = 0.00), D2S393 (Z = 2.93, θ = 0.0) and D2S378 (Z = 4.32, θ = 0.0).
Petryshen et al. 2002	Linkage	96 Canadian families (n = 877)	Linkage with spelling (D2S2352-D2S378, LOD = 3.82), PC (D2S378, LOD = 1.13) and PA (D2S378, LOD = 1.01)
Fisher et al. (2002)	Linkage	89 UK families (195 sibling pairs)	OC (D2S2211, P = 0.001; D2S391, P = 0.0007) and reading (D2S391, P = 0.007) in single point analysis
		119 American families (180 sibling pairs)[a]	Reading (D2S2368, P = 0.013), PA (D2S2368, P = 0.001; D2S286 P = 0.0003) and OC (D2S2368, P = 0.005) in the single point analysis, 2p15 with reading (P = 0.006), PA (P = 0.001) and OC (P = 0.005) in multipoint analysis
Francks et al. (2002)	Linkage	119 American families (180 sibling pairs)[a]	OC (D2S2240, P = 0.003), SWR (D2S2378, P = 0.004) PD (D2S2378, P = 0.004)
Kaminen et al. (2003)	Linkage	11 Finnish families (n = 97)[b]	D2S2216 (LOD = 2.55, non-parametric); D2S286 (LOD = 3.01, parametric)
Peyrard-Janvid et al. (2004)	Linkage	11 Finnish families (n = 97)[b]	D2S2216 (LOD = 3.0, non-parametric)
Anthoni et al. (2007)	Association	11 Finnish families (n = 97)[b]	rs1000585/rs917235/rs714939 (GGG, P = 0.0076)
		251 German families	rs917235/rs714939/rs6732511 (GGC, P = 0.036)
Bates et al.(2007)	Linkage	403 Australian families (n = 980)	Non-word spelling (D2S1360, LOD = 0.83); SWR (D2S2972, LOD 1.04); spelling (D2S1360, LOD = 1.13)

Table 1.4: Evidence for association/linkage with *DYX3* (PA = phonological awareness, PD = phonological decoding, OC = orthographic coding, SWR = single word reading). N.B some of these studies have overlapping samples. Letters (a,b) indicate the same samples.

The first evidence for a region on chromosome 2 showing linkage with DD came from a genome wide study using 36 members of a Norwegian family (Fagerheim et al. 1999). At first the screen did not show any significant linkage, but marker D2S1356 on 2p15-p16 gave a slightly significant LOD score of 0.8. Fagerheim and colleagues (1999) then analysed 17 additional microsatellite markers around this region and found significant LOD scores for the markers D2S2183 (Z = 5.53, using a model of complete linkage: $\theta = 0.0$), D2S393 (Z = 2.93, $\theta = 0.0$) and D2S378 (Z = 4.32, $\theta = 0.0$).

Petryshen and colleagues (2002) replicated this linkage in a sample of 96 Canadian families, each containing two or more siblings diagnosed with phonological coding dyslexia (PCD). They used both categorical and quantitative definitions of PCD in their study. Using nonparametric analysis and a categorical diagnosis of PCD, evidence for linkage was found within the *DYX3* region (P = 0.009). Using variance components analysis and a quantitative definition, peak LOD scores were found for spelling (3.82 between D2S2352 and D2S378), PD (1.13 at D2S378) and PA (1.01 at D2S378).

Fisher and colleagues (2002) carried out a genome wide scan using a UK sample of 89 families and an American sample of 119 families from the Colorado twin study. The UK sample showed linkage on 2p25 with the OC phenotype (P = 0.001) and 2p16 with reading (P = 0.007) and OC (P = 0.0007) in the single point analysis, but gave little evidence with multipoint analysis. The US sample showed linkage on 2p15 with reading (P = 0.013), PA (P = 0.001) and OC (P = 0.005) in the single point analysis. In the multipoint analysis, the US sample showed evidence of 2p15 being linked to reading (P = 0.006), PA (P = 0.001) and OC (P = 0.005).

Francks and colleagues (2002) used the US sample to fine-map the 2p12-17 region using 21 microsatellite markers. They refined the linkage region to 12 cM between the markers D2S337 and D2S286 (see Figure 1.4), with the peak significance of linkage very similar to that reported by Fisher and colleagues (2002). This linkage was replicated in a genome-wide study by Bates and colleagues (2007), using a sample of 403 Australian samples that were not selected for reading ability. They identified nominally significant peaks within the region reported by Francks and colleagues (2004), with a LOD of 0.83 at the marker D2S1360 for nonword spelling and a second peak of 1.04 at marker D2S2972 for SWR. They also found linkage on chromosome 2p outside of the *DYX3* region (see Figure 1.4). Regular-word spelling was found to be linked to D2S1360 on chromosome 2p24.2 (LOD = 1.13, P = 0.030). This could represent an alternative DD locus on chromosome 2p, but as yet has not been reported by any other studies.

Kaminen and colleagues (2003) carried out a whole genome scan using 11 Finnish families, and obtained a linkage peak for the marker D2S2216 on 2p11 (NPL = 2.55, P = 0.004). The marker D2S2216 is about 34 cM centromeric from the *DYX3* locus implicated in other studies (Fagerheim et al. 1999; Francks et al. 2002; Petryshen et al. 2002). This result could represent a different locus, or the difference could be due to different populations, diagnostic criteria and markers being used in the studies. Peyrard-Janvid and colleagues (2004) used this same sample to fine map the region further. They used 24 markers in a 40 cM region. Their highest NPL score was 3.0 (P = 0.001) for the marker D2S2216, replicating the result found by Kaminen and colleagues

(2003). They refined their region of linkage to a 12 cM region between D2S2216 and D2S181, supporting other evidence of a DD susceptibility locus in this region. Anthoni and colleagues (2007) also used this sample of Finnish families in a study that used 8 microsatellites and 43 SNPs to refine the location of linkage to a 157 kb region on 2p12. This was then replicated in an independent set of 251 German families. Two overlapping risk haplotypes were identified in the two sample sets. The haplotype rs1000585/rs917235/rs714939 was found to be significantly associated in the Finnish sample (GGG, P = 0.0076), while the haplotype rs917235/rs714939/rs6732511 was significantly associated in the German sample (GGC, P = 0.036). In a joint analysis of the two sample sets, these risk haplotypes were still significant (P = 0.0049 for the Finnish haplotype, P = 0.0013 for the German haplotype). These haplotypes span a 16.6 kb region, located in an intergenic region between the hypothetical gene *FLJ13391* and the genes *MRPL19* (mitochondrial ribosomal protein 19) and *C2ORF3* (chromosome 2 open reading frame 3). The haplotype block structure of the region revealed a 62 kb block of strong LD containing *MRPL19* and *C2ORF3*. Both of these genes were shown to be co-expressed across a panel of tissues from regions of adult brains, as well as showing correlation of expression with four other putative dyslexia susceptibility genes (*DYX1C1, ROBO1, DCDC2* and *KIAA0319*). Anthoni and colleagues (2007) went on to sequence the coding exons and the flanking sequences of these two genes in one affected individual from each of the 19 Finnish families. Several non-synonymous variants were identified, but none of these were seen to be over-transmitted in affected individuals and so did not show significant association with DD. However, they did observe a reduction in the expression of both *MRPL19* and *C2orf3* from chromosomes carrying both rs917235(G) and rs714939(G) (Anthoni et al. 2007).

The *DYX3* region seems to be a promising region for harbouring DD susceptibility loci, being widely replicated in samples from Norway (Fagerheim et al. 1999), Canada (Petryshen et al. 2002), UK (Fisher et al. 2002), US (Fisher et al. 2002; Francks et al. 2002), Australia (Bates et al. 2007), Finland (Kaminen et al. 2003; Peyrard-Janvid et al. 2004; Anthoni et al. 2007) and Germany (Anthoni et al. 2007). However, further work on this region now needs to concentrate on identifying candidate genes for DD and testing these for an association in larger samples.

1.4.4 *DYX4* (Chromosome 6q)

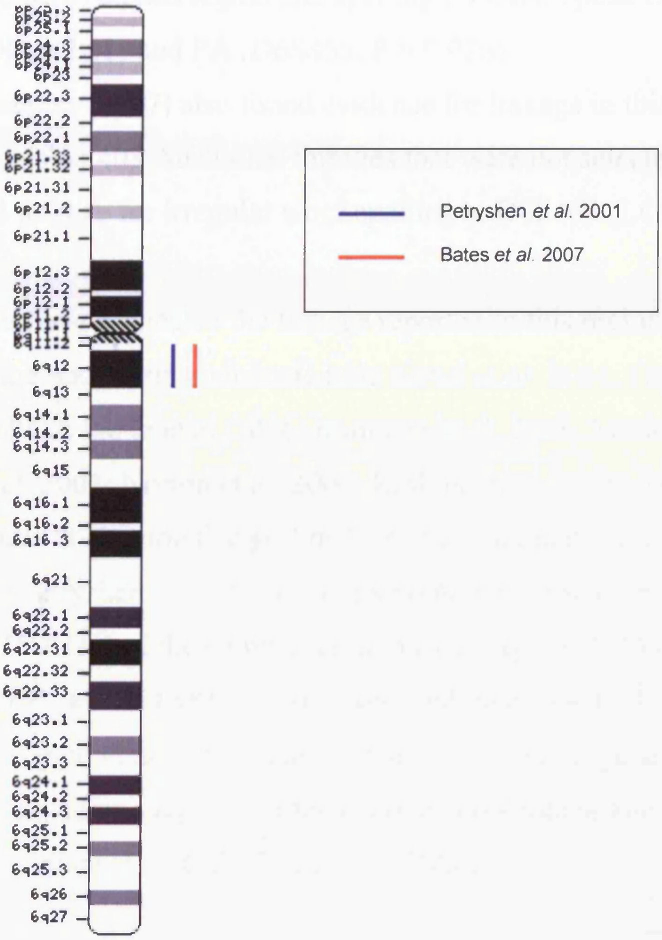


Figure 1.5: Region on chromosome 6q showing evidence of linkage/association with DD

Study	Type	Sample	Main Findings
Petryshen et al. (2001)	Linkage	96 Canadian families (n = 877)	PC (D6S965, LOD = 2.08), Spelling (D6S865, LOD = 3.34)
Bates et al. (2007)	Linkage	403 Australian families (n = 980)	Irregular word spelling at D6S462 (LOD = 1.59, P = 0.003)

Table 1.5: Evidence for association/linkage with *DYX4* (PC = phonological coding).

Petryshen and colleagues (2001) carried out a linkage study using a sample of 96 Canadian families and a qualitative PCD phenotype (affected, unaffected or uncertain) and found suggestive evidence of linkage with chromosome 6q11.2-12 (see Figure 1.5). Two-point parametric analyses found evidence for linkage between PCD and the markers D6S254, D6S965, D6S280 and D6S251 (LOD$_{max}$ scores = 2.4 to 2.8) across an 11 cM region. Multipoint parametric analysis supported this linkage with a peak LOD score of 1.6 between markers D6S20 and D6S286. Petryshen and colleagues (2001) then used separate reading measures (PA, PC, spelling and RAN speed) and found evidence of linkage between this region and spelling (D6S865, peak LOD = 3.34), PC (D6S965, peak LOD = 2.08) and PA (D6S455, P = 0.026).

Bates and colleagues (2007) also found evidence for linkage in this region in their genome wide study using 403 Australian families that were not selected for reading ability. They found linkage for irregular word spelling at D6S462 (LOD = 1.59, P = 0.003).

No other studies have replicated the linkage reported in this region, despite several genome-wide linkage scans being conducted (de Kovel et al. 2004; Fagerheim et al. 1999; Fisher et al. 2002; Igo Jr et al. 2006; Kaminen et al. 2003; Marlow et al. 2003; Nopola-Hemmi et al. 2001; Norton et al. 2000; Raskind et al. 2005). However, several genes are located near this region that still make strong candidates for a DD susceptibility gene (Petryshen et al. 2001). The serotonin neurotransmitter receptor genes *HTR1B* and *HTR1E* and the gamma-aminobutyric acid (GABA) receptor rho-subunit genes *GABRR1* and *GABRR2* make good candidates due to their involvement in brain development (Levitt et al. 1997). The cannabinoid receptor gene (CNR1) is also a good candidate since studies suggest that the endogenous cannabinoid system plays a role in neural development (Fernández-Ruiz et al. 2000).

1.4.5 *DYX5* (Chromosome 3)

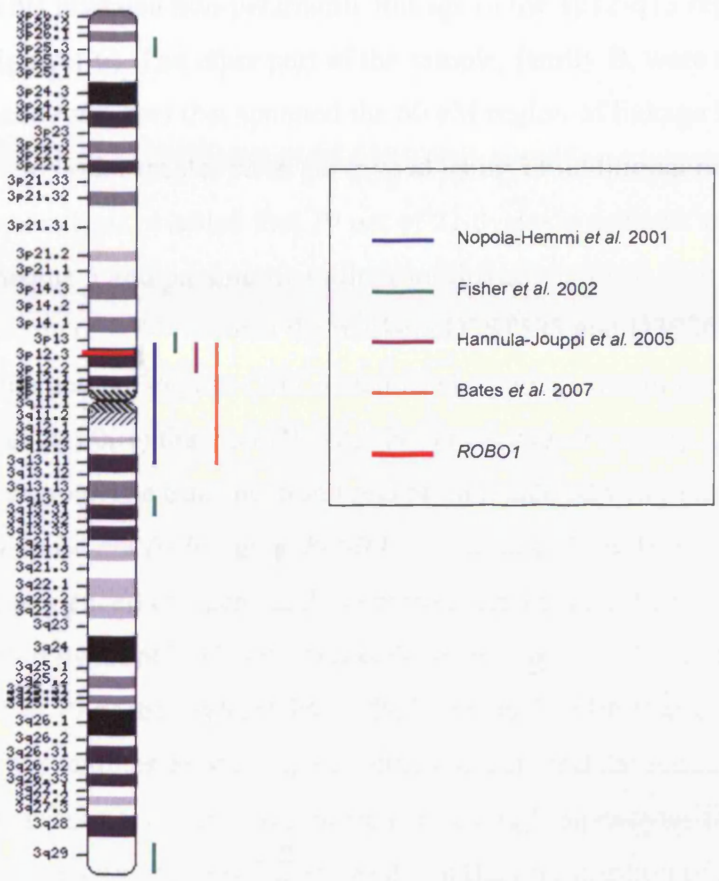


Figure 1.6: Regions on chromosome 3 showing evidence of linkage/association with DD

Study	Type	Sample	Main Findings
Nopola-Hemmi et al. (2001)	Linkage	1 Finnish family (n = 74)	D3S1595-D3S3655 (LOD = 3.84)
Fisher et al. (2002)	Linkage	89 UK families (195 sibling pairs)	PD (D3S1566, P = 0.044), OC-irreg (D3S1566, P = 0.001; D3S1311, P = 0.0008)
		119 American families (180 sibling pairs)	D3S1278 linked to reading P = 0.002, PA P = 0.097, PD P = 0.0004, and OC P = 0.026 (single-point); linked to reading P = 0.003, PA P = 0.072, PD P = 0.0003, and OC P = 0.025 (multipoint
Bates et al. (2007)	Linkage	403 Australian families (n = 980)	Irreg word spelling (D3S1292, LOD = 1.66, P = 0.003)

Table 1.6: Evidence for association/linkage with *DYX5* (PA = phonological awareness, PD = phonological decoding, OC = orthographic coding).

Linkage of DD to the 3p12-q13 region was first identified by Nopola-Hemmi and colleagues (2001). They carried out a genome-wide scan using 74 members of a four generation Finnish family, of which 21 were affected with dyslexia. In the first part of the study, part of the sample, family A, was genotyped using 320 markers spanning the whole genome. This revealed non-parametric linkage in the 3p12-q13 region (Z = 5.8, P = 0.0017) (see Figure 1.6). The other part of the sample, family B, were then genotyped using 7 microsatellite markers that spanned the 60 cM region of linkage implicated in the first sample, and both samples were genotyped using 11 additional markers in this region. Haplotype analysis revealed that 19 out of 21 dyslexic subjects shared identical copies of chromosome 3 and parametric multipoint linkage analysis resulted in a maximum LOD score of 3.84 between the markers D3S1595 and D3S3655.

Hannula-Jouppi and colleagues (2005) identified a patient within the Finnish family (Nopola-Hemmi et al. 2001) that had DD and the translocation t[3;8][p12;q11] which is within the *DYX5* region. The translocation breakpoint localised within the orthologue of the *Drosophila roundabout (robo)* gene *ROBO1*, disrupting *ROBO1* between exons 1 and 2. They then sequenced this gene and the region surrounding it in the original Finnish sample used by Nopola-Hemmi and colleagues (2001) and found that the haplotype present in 19 of the dyslexic individuals spanned *ROBO1* and was not detected in the other samples examined, including controls and the remaining family members. Gene expression analysis was carried out using lymphocytes from four affected members of the family and this showed that the transcription of *ROBO1* from this haplotype was absent or attenuated, suggesting that in this family, DD may be caused by a reduction in the expression of *ROBO1*. However, reduction in expression varied across the four individuals, so there does not seem to be a uniform phenotype. Also, as the expression analysis was only carried out in 4 individuals, it is not known if this reduction in expression occurs in all members carrying the haplotype. Another important point is that a sibling of a translocation carrier was diagnosed with severe DD but did not carry the translocation themselves, so there must also be other variants causing the disorder, perhaps on different chromosomes. Functionally, *ROBO1* is a good candidate gene for DD. It is widely expressed in the brain and the *robo* gene in *Drosophila* was found to be involved in controlling the decision by axons to cross the central nervous system midline (Kidd et al. 1998). In *robo* mutants, it was found that too many axons cross the midline (Seeger et al. 1993), and so *ROBO1* appears to play a role within axon guidance/neuronal migration (see Chapter 3).

In another genome-wide scan, Fisher and colleagues (2002) found linkage to several regions on chromosome 3 using quantitative trait analysis in US and UK samples. As shown in Figure 1.6, some of these regions were not close to the *DYX5* region (D3S1311 at 3q29 linked to OC of irregular words in UK sample, P = 0.0008; D3S1263 at 3p25 linked to reading, PA and OC in US sample, P = 0.016, 0.023 and 0.001, respectively) while others were just outside of the region. A marker on 3p13 was linked to PD and OC of irregular words in the UK samples (D3S1566, PD P = 0.044; OC P = 0.001), while the region 3q13 was linked to DD in the US samples using both single point analysis (D3S1278 linked to reading P = 0.002, PA P = 0.097, PD P = 0.0004, and OC P = 0.026) and multipoint analysis (linked to reading P = 0.003, PA P = 0.072, PD P = 0.0003, and OC P = 0.025). The *DYX5* region was also linked to DD in a genome-wide scan by Bates and colleagues (2007) using 403 Australian families unselected for reading ability. A linkage peak for irregular word spelling was found in the 3p12-q13 region for irregular word spelling (LOD = 1.66, P = 0.003). This is within 20 cM of the linkage peak reported by Nopola-Hemmi and colleagues (2001).

So far, the region on chromosome 3 that shows evidence of linkage with DD is quite broad. More studies need to be carried out on this region, perhaps with denser panels of markers and larger sample sets. However, *ROBO1* appears to be a promising candidate gene for DD.

1.4.6 *DYX6* (Chromosome 18)

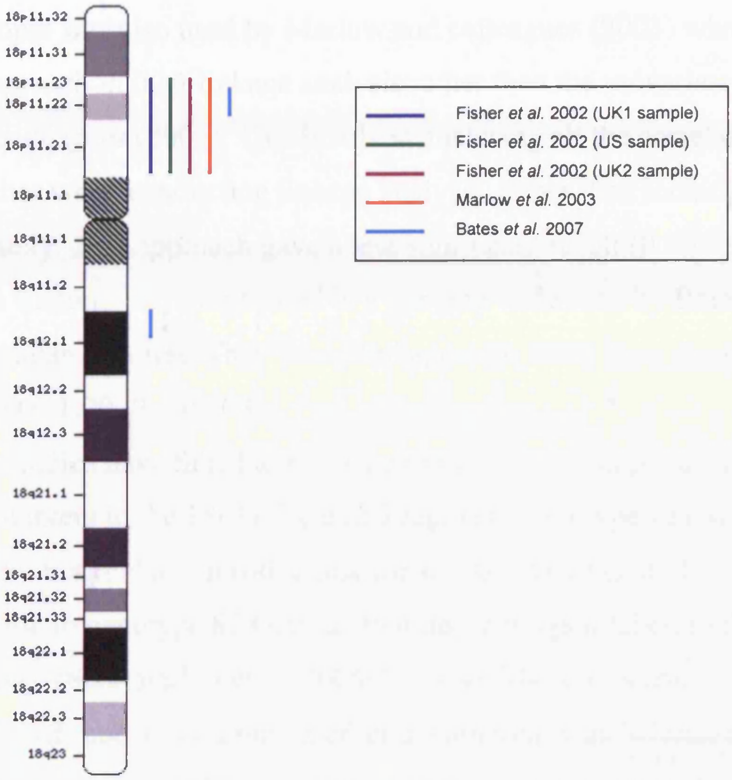


Figure 1.7: Regions on chromosome 18 showing evidence of linkage/association with DD

Study	Type	Sample	Main Findings
Fisher et al. (2002)	Linkage	89 UK families (195 sibling pairs)^{a}	SWR (Single point - D18S53, P = 0.0002; D18S464, P = 0.0006; Multipoint - P = 0.00001)
		119 American families (180 sibling pairs)	SWR (Single point - D18S53, P = 0.0004; D18S1102, P = 0.0005; Multipoint - P = 0.0004)
		84 UK families (143 sibling pairs)^{b}	PA (Singlepoint - D18S452, P = 0.005; D18S464, P = 0.0001; Multipoint - P = 0.0005)
Marlow et al. (2003)	Linkage	173 UK families^{a+b}	Multivariate analysis P = 0.0011
Bates et al. (2007)	Linkage	403 Australian families (n = 980)	Reading (D18S464, LOD = 1.70, P = 0.03)

Table 1.7: Evidence for association/linkage with *DYX6* (SWR = single word reading; PA = phonological awareness). Letters (a,b) indicate sample overlap.

The first study implicating the *DYX6* region in DD was the first whole-genome scan carried out by Fisher and colleagues (2002). The strongest evidence for linkage in this study was found in the 18p11 region (see Figure 1.7 and Table 1.7). Linkage was found in both the US and UK samples, and was replicated in a separate UK sample. However, the first UK sample had shown the strongest linkage to the SWR measure (P = 0.0001), whereas the second sample gave the most significant results for the PA measure (P = 0.0005), raising concerns over whether this finding represents a true replication. The second UK sample was also used by Marlow and colleagues (2003) who applied a multivariate approach in their linkage analysis rather then the univariate approach used by Fisher and colleagues (2002). This involved analysing all the correlated trait measures together when conducting linkage analysis, rather than looking at each measure separately. This approach gave a less significant result (P = 0.0011) for 18p11.2. These findings were replicated in a genome-wide scan by Bates and colleagues using 403 Australian families. They found linkage with SWR for a marker on 18p11.2 (D18S464, LOD = 1.70, P = 0.003).

Two other studies have failed to replicate these results. Chapman and colleagues (2004) used 8 markers in the 18p11.3 – q12.3 region to genotype 111 American families (n = 898), but did not find any positive linkage signals. Another study used 14 markers in the same region to genotype 82 German families, but again failed to find any significant linkage (Schumacher et al. 2006b). These differences could well be due to differing sample sets and ascertainment criteria, particularly as Schumacher and colleagues (2006b) used a spelling measure as their ascertainment criteria, whereas Fisher and colleagues (2002) measured reading ability. However, the American sample used by Chapman and colleagues (2004) is larger than that used by Fisher and colleagues (2002) and so would have more power to detect linkage. It could be that the putative disease gene in the *DYX6* locus confers a smaller risk to DD than originally suggested.

1.4.7 *DYX7* (Chromosome 11)

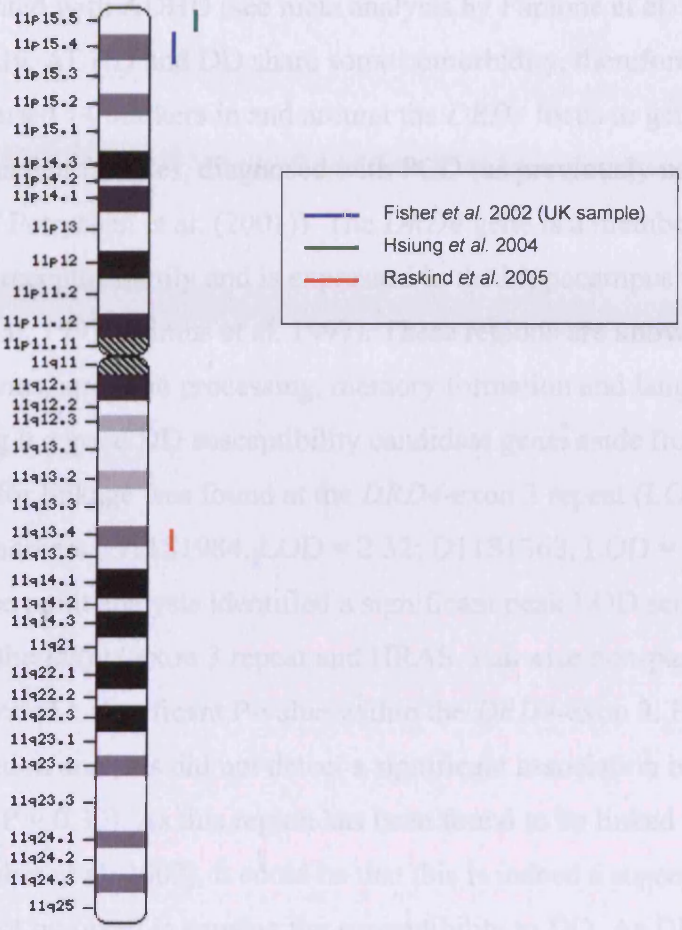


Figure 1.8: Regions on chromosome 11 showing evidence of linkage/association with DD

Study	Type	Sample	Main Findings
Fisher et al. (2002)	Linkage	89 UK families (195 sibling pairs)	PA (D11S1338 (P = 0.001)
Hsuing et al. (2004)	Linkage	100 Canadian families (n = 914)	Linkage between *DRD4*-exon 3 repeat and HRAS (LOD = 3.57, P = 0.00005)

Table 1.8: Evidence for association/linkage with *DYX7* (PA = phonological awareness).

Evidence for linkage of a region on chromosome 11 to DD was first identified in the genome-wide scan conducted by Fisher and colleagues (2002). In their UK sample, they found linkage with the marker D11S1338 on 11p15 with the PA measure. However, no significant linkage on this chromosome was identified in the US sample.

Hsiung and colleagues (2004) also studied this region for linkage with DD. They looked at this region due to the presence of an ADHD candidate gene, the dopamine D4 receptor (*DRD4*). This gene contains several polymorphic elements, including a 48-bp variable number tandem repeat (VNTR) in exon 3 which is in the region previously found to be associated with ADHD (see meta analysis by Faraone et al. (1999)). As discussed previously, ADHD and DD share some comorbidity, therefore Hsiung and colleagues (2004) used 14 markers in and around the *DRD4* locus to genotype their sample of 100 Canadian families, diagnosed with PCD (as previously used by Field and Kaplan (1998) and Petryshen et al. (2001)). The *DRD4* gene is a member of the dopamine D2-like receptor family and is expressed in the hippocampus and frontal cortex (Defagot et al. 1997; Primus et al. 1997). These regions are known to be involved in executive functions, attention processing, memory formation and language processing, making it a good DD susceptibility candidate gene, aside from the link with ADHD. Evidence for linkage was found at the *DRD4*-exon 3 repeat (LOD = 2.27) and at several nearby markers (D11S1984, LOD = 2.32; D11S1363, LOD = 2.13; HRAS, LOD = 2.68). Three point analysis identified a significant peak LOD score of 3.57 (P = 0.00005) between the *DRD4*-exon 3 repeat and HRAS. Pairwise non-parametric sib-pair analyses also generated a significant P-value within the *DRD4*-exon 3. However, subsequent association analysis did not detect a significant association between DD and the *DRD4* VNTR (P = 0.30). As this region has been found to be linked with DD in a previous study (Fisher et al. 2002), it could be that this is indeed a susceptibility locus, but that *DRD4* is not involved in causing the susceptibility to DD. As DD is common in the general population, another possibility is that there are multiple *DRD4* variants contributing to susceptibility (Hsiung et al. 2004).

Another genome-wide study conducted using 51 American families found suggestive evidence for linkage on a different region of chromosome 11. Raskind and others (2005) obtained a LOD score of 2.32 for the marker D11S1314 using two-point parametric analysis. However, this marker is on the other arm of chromosome 11 as shown in Figure 1.8. This region has not yet been found to be linked with DD in any other studies and the result from this study isn't highly significant, especially when considered on a genome-wide scale, so may not be a true finding. Alternatively, it could represent a second DD susceptibility locus on chromosome 11 that has only been highlighted in this particular sample so far.

1.4.8 *DYX8* (Chromosome 1)

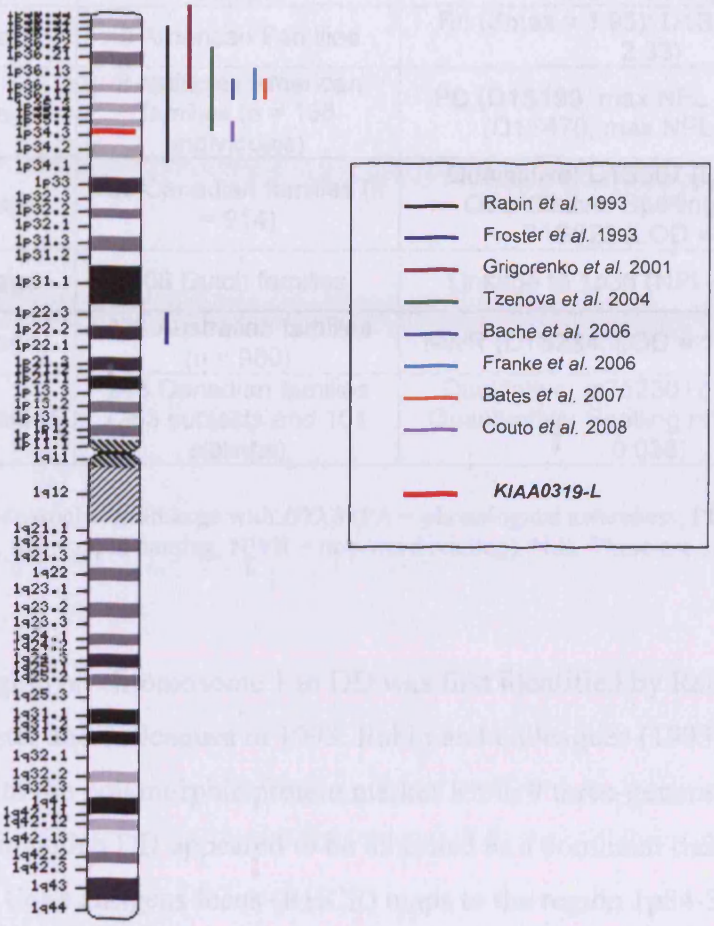


Figure 1.9: Regions on chromosome 1 showing evidence of linkage/association with DD

Study	Type	Sample	Main Findings
Rabin et al. (1993)	Linkage	9 American Families	Rh (Zmax = 1.95); D1S165 (Zmax = 2.33)
Grigorenko et al. (2001)	Linkage	8 multiplex American families (n = 165 individuals)	PD (D1S199, max NPL = 2.623); RN (D1S470, max NPL = 5.737)
Tzenova et al. (2004)	Linkage	100 Canadian families (n = 914)	Qualitative: D1S507 (LOD = 3.65). Quantitative: Spelling D1S552 - D1S622 (LOD = 4.01)
Franke et al. (2006)	Linkage	108 Dutch families	Linkage to 1p36 (NPL-LOD = 2.0_
Bates et al. (2007)	Linkage	403 Australian families (n = 980)	NWR (D1S234, LOD = 1.2, P = 0.009)
Couto et al. (2008)	Association	263 Canadian families (263 subjects and 101 siblings)	Qualitative: rs7523017 (P = 0.035). Quantitative: Spelling rs7523017 (P = 0.036)

Table 1.9: Evidence for association/linkage with *DYX8* (PA = phonological awareness, PD = phonological decoding, RN = rapid naming, NWR = non-word reading). N.B. These are all independent samples.

Linkage of a region on chromosome 1 to DD was first identified by Rabin and colleagues and Froster and colleagues in 1993. Rabin and colleagues (1993) carried out linkage analyses with the polymorphic protein marker Rh in 9 three-generation American families in which DD appeared to be inherited as a dominant trait. This rhesus blood group CcEe antigens locus (RHCE) maps to the region 1p34-36 and showed significant linkage in all the families (Zmax = 0.95, $\theta \leq 0.2$). Linkage was also found for the markers FUCA1 (Zmax = 0.950, $\theta = 0$) and D1S165 (Zmax = 2.33, $\theta =$ 0.2). Froster and colleagues (1993) identified a co segregation of severely delayed speech development and the reading and writing disability with a balanced translocation (t[1;2][1p22;2p31]) in a German family that had a history of learning difficulties. A father and two of his sons carried the translocation and showed evidence of learning difficulties as well as severely delayed speech development. The members of the family with normal karyotypes had normal phenotypes.

Bache and colleagues (2006) also found evidence for a translocation on chromosome 1 being linked with DD, but this translocation involved the 1p36.1 region of the chromosome. They found that the balanced translocation t[1;18][p36.1;q21] cosegregated with DD in 5 members of a Dutch family. However, diagnosis of DD was based on the family members self reporting their disease status, rather than carrying out any tests so this result is not completely reliable.

Grigorenko and colleagues (2001) found evidence for linkage of DD with chromosome 1p by examining the region around Rh (1p36-1q23) using eight extended American families who possessed at least four individuals with dyslexia (n = 165). Again, this group used five theoretical phenotypes for DD in their search for linkage: 1) PA; 2) PD; 3) RAN; 4) SWR; 5) vocabulary. A "lifetime" diagnosis (LD) was also used as a phenotype. An individual (adult or child) was diagnosed as having DD if it had been reported that they had difficulty acquiring initial reading skills (n = 68). If they still showed impairment in their reading or still required specialist instruction, they were classed as clearly impaired (n = 33 out of the 68 DD cases). If however they had managed to obtain literacy level or no longer required ongoing reading help, they were diagnosed as "borderline" (n = 35). Individuals were classed as normal if they had no reported history of difficulty with reading and if there was a deficiency in no more than one aspect of reading. Single point analysis indicated two broad regions of linkage: 1) D1S253 (6.3 Mb) to D1S478 (21.3 Mb); 2) MATN1 (30.9 Mb) to PPT (40.2 Mb). However, multipoint analyses gave an inconsistent pattern and may be invalid due to two of the markers used in the analysis being in the wrong order. The PD phenotype showed linkage to the marker D1S199 (NPL score = 2.623) and the RAN phenotype showed linkage to the marker D1S470 (NPL score = 5.737).

Tzenova and colleagues (2004) found further evidence for linkage of DD to the region 1p34-36. This involved a larger sample than the previous studies, with analyses being carried out on 100 Canadian families (n = 914). The group used both qualitative and quantitative definitions of DD in their linkage analysis. The qualitative definitions used were 'affected', 'unaffected' or 'uncertain' using the scores on phonological coding tasks as the primary determinant of affection status. Under these phenotypes, the strongest evidence for linkage was found at the marker D1S507 (max LOD = 3.65). This marker is ~5Mb away from D1S199, which had previously been found to be linked to DD by Grigorenko and colleagues (2001). For the quantitative phenotype analysis, psychometric tests were conducted to assess four components of DD: 1) PA; 2) PD; 3) spelling; 4) RAN. Using multipoint analysis, the maximum LOD score for spelling was 4.01 and occurred between D1S552 and D1S622. PD and RAN speed showed non-significant evidence for linkage to the same region (max LOD scores of 1.65 and 0.37 respectively), while the LOD scores for PA were close to zero throughout the region.

However, again these results are unreliable due to two markers being in the wrong order.

A linkage study carried out on a Dutch population also looked at DD as both a categorical trait and as a number of different quantitative traits (Franke et al. 2006). The categorical trait showed the strongest linkage to 1p36 (NPL-LOD = 2.0). The LOD scores for the quantitative traits: SWR, NWR and RAN were found to be correlated and peaked near the same location as the categorical trait.

A gene within the 1p34-36 region shows homology to the *KIAA0319* gene on chromosome 6 and is called *KIAA0319-Like (KIAA0319-L)* (see Figure 1.9). Couto and colleagues (2008) genotyped a sample of 156 Canadian families using 5 SNPs within this gene. Evidence for an association was found with the marker rs7523017 (P = 0.042) when DD was defined as a categorical trait. They also identified a significant haplotype with the markers rs1203138, rs1203148, rs12408030, and rs7523017 (C/A/A/A; P = 0.031).When using quantitative measures of DD and their whole sample of 291 families (the 156 previous families plus an additional 135 families that had been ascertained through a proband who had reading difficulties but did not meet the categorical criteria), this haplotype showed significant association with measures of word-reading efficiency (P = 0.032) and rapid object and colour naming (P = 0.047). However, as the authors point out, these results would not withstand correction for multiple testing. These results correlate with those of Tzenova and colleagues (2004) who also found linkage on chromosome 1p for spelling in a region that is 5 kb from rs7523017. In the genome wide study carried out by Bates and colleagues (2007) on 403 Australian families, the region 1p34-36 also showed some evidence of linkage to NWR (max LOD = 1.2).

Further evidence supporting linkage of the *DYX8* region to DD has come from studies focusing on SSD. Smith and colleagues (2005) tested whether SSD is linked to risk loci for DD, including those on chromosome 1p36. Although only suggestive evidence for linkage was found in this particular study (P = 0.053 at D1S620), another study on SSD supported this evidence by obtaining highly significant evidence for linkage (Miscimarra et al. 2007). This group obtained significant linkage signals for articulation (P = 0.0009) and listening comprehension (P = 0.0019) in two separate regions on chromosome 1.

However, studies using subjects sampled from other populations have failed to replicate linkage of DD to chromosome 1. For example, Cardon and colleagues (1994) typed markers in the Rh region in 358 individuals from the Colorado twin study and failed to find significant linkage. Several genome wide studies have also failed to identify linkage to DD on chromosome 1. Fagerheim and colleagues (1999) conducted a genome wide search with an average 20 cM marker density in a Norwegian population. This search included 12 markers on chromosome 1p, but no linkage was found to DD. A genome scan using a Finnish population of 140 families conducted by Nopola-Hemmi and colleagues (2001) used 320 microsatellite markers, but found no association with DD on chromosome 1. Fisher and colleagues (2002) carried out two complete quantitative trait locus-based (QTL-based) genome wide linkage studies in large samples from the United Kingdom (195 total sibling pairs) and United States (180 total sibling pairs). This group used over 400 microsatellite markers spaced at about 10 cM intervals throughout the genome. Again, this study failed to replicate an association of chromosome 1 with DD.

The evidence so far seems to suggest that a region on chromosome 1 is linked with DD, with the most widely replicated region being 1p34.2 to 1p36.13 as shown in Figure 1.9. However, while evidence for linkage in this region has been found in American (Rabin et al. 1993; Grigorenko et al. 2001), Canadian (Tzenova et al. 2004; Couto et al. 2006), Australian (Bates et al. 2007) and Dutch (Franke et al. 2006) populations, genome-wide scans carried out on populations in Norway (Fagerheim et al. 1999), Finland (Nopola-Hemmi et al. 2001), UK and a different American population (Fisher et al. 2002), have failed to support these results. It is possible that this region may be involved in DD, but only in certain populations.

1.4.9 *DYX9* (Chromosome X)

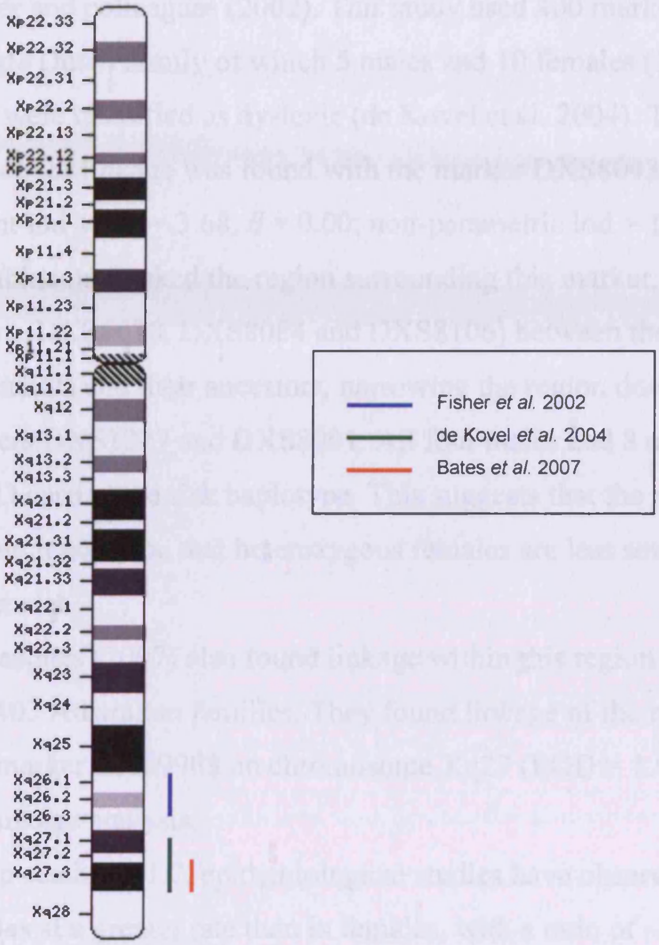


Figure 1.10: Regions on chromosome X showing evidence of linkage /association with DD.

Study	Type	Sample	Main Findings
Fisher et al. (2002)	Linkage	89 families (195 sib pairs), UK	Xq26 linked to reading (P = 0.001), PD (P = 0.018), OC-irreg (0.038)
de Kovel et al. (2004)	Linkage	1 Dutch family (*n* = 29)	DXS8043 (multipoint LOD = 3.68; NPL = 1.95, P = 0.0014)
Bates et al. (2007)	Linkage	403 Australian families (*n* = 980)	Non-word spelling (DXS9908 LOD = 1.09, P = 0.012)

Table 1.10: Evidence for association/linkage with *DYX9* (PD = phonological decoding, NPL = non-parametric LOD score)

The genome-wide study conducted by Fisher and colleagues (2002) was again the first study to identify linkage to DD on the X chromosome. They found linkage of the Xq26 region with measures of reading (P = 0.001), PD (P = 0.018) and OC using irregular words (P = 0.038) in their UK sample. However, no significant linkage was found between this chromosome and DD in the US sample.

Another genome-wide study found evidence of linkage 12 cM away from the region identified by Fisher and colleagues (2002). This study used 400 markers to genotype 29 members of a single Dutch family of which 5 males and 10 females (1 of which was an unrelated spouse) were classified as dyslexic (de Kovel et al. 2004). The most significant evidence for linkage was found with the marker DXS8043 on chromosome Xq27.3 (multipoint lod score = 3.68, $\theta = 0.00$; non-parametric lod = 1.95, P = 0.0014). As two key recombinants flanked the region surrounding this marker, the group tested three extra markers (DXS8028, DXS8084 and DXS8106) between the two breakpoints in the key recombinants and their ancestors, narrowing the region down to ~8 cM between the markers DXS1227 and DXS8091. All four males and 8 of the 9 women diagnosed with DD carried the risk haplotype. This suggests that the risk allele has a dominant effect, but it could be that heterozygous females are less severely affected than hemizygous males.

Bates and colleagues (2007) also found linkage within this region in their genome-wide study using 403 Australian families. They found linkage of the non-word spelling measure with the marker DXS9908 on chromosome Xq27 (LOD = 1.09, P = 0.012) using multipoint linkage analysis.

As discussed in section 1.1.2, epidemiological studies have observed that DD is often found in males at a greater rate than in females, with a ratio of ~2:1 (Flannery et al. 2000). A possible cause for this could be the involvement of X linked loci, backed up by evidence of a susceptibility locus on chromosome X as shown here. However, other explanations for this skewed ratio could involve male specific hormonal differences during development that may interact with an autosomal locus (James 1992). Geschwind and Behan (1982) have hypothesised that gender differences could be explained by an excess of, or sensitivity to, androgens such as testosterone.

1.4.10 Other Possible DD Susceptibility Loci

Other loci have also been found to be linked or associated with DD but have not been designated a DYX region by the Human Gene Nomenclature Committee (http://www.gene.ucl.ac.uk/nomenclature/). Two of these have shown replication and are worth mentioning.

1.4.10.1 Chromosome 2q

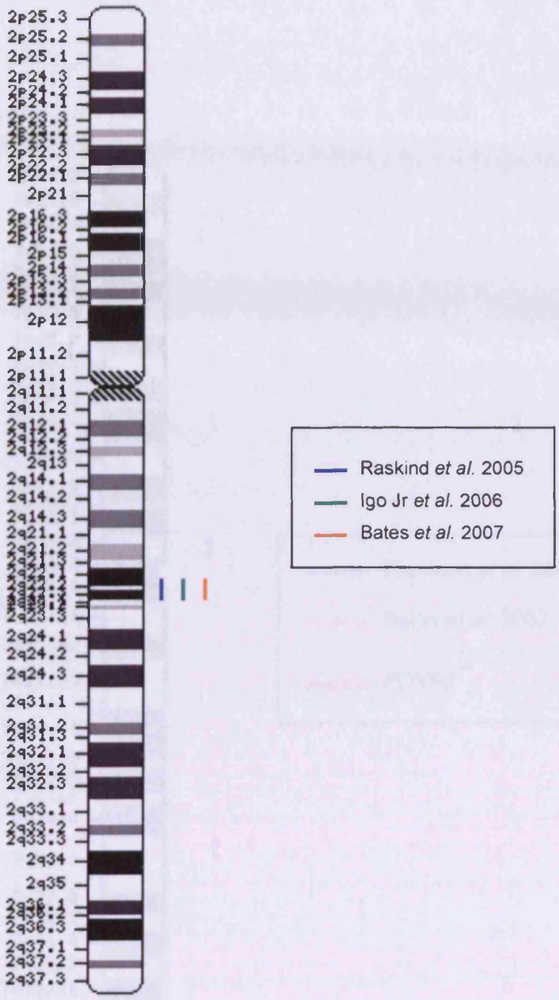


Figure 1.11: Regions on chromosome 2q showing evidence of linkage/association with DD.

Chromosome 2q22.3 first showed evidence of linkage with DD in a genome scan carried out on a total of 874 individuals from 108 American families (Raskind et al. 2005). This study found that a phonemic decoding efficiency measure (PDE) was significantly linked with the marker D2S1399 (LOD = 3.0). Igo Jr and others (2006) used this same sample in a genome scan using measures of single-word reading efficiency (SWE) and word identification (WID). The same region showed linkage with SWE for the three marker combination D2S1334-D2S1326-D2S1399 (LOD = 1.88). WID, which is a measure of accuracy alone, did not show any evidence of linkage to this region, suggesting that this locus is involved in the speed rather than accuracy of phonological decoding. Further evidence of linkage in this region comes from the genome-wide scan by Bates and others (2007) . The scan of 403 Australian families showed linkage with regular-word spelling for the marker XRCC5 (LOD = 2.18, P = 0.001).

1.4.10.2 Chromosome 7q

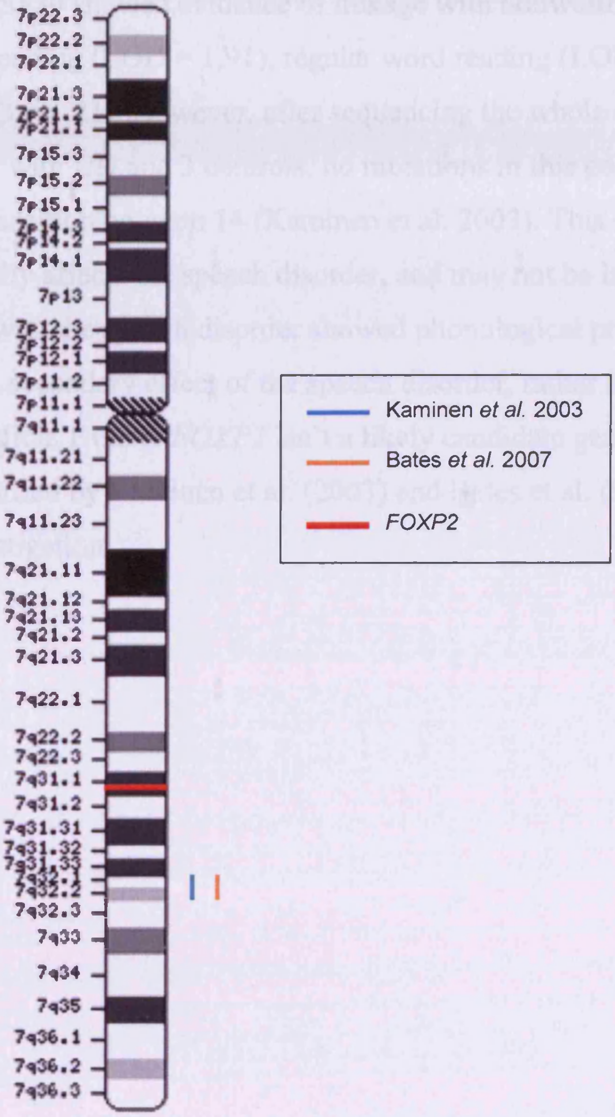


Figure 1.12: Regions on chromosome 7q showing evidence of linkage/association with DD.

The *FOXP2* gene on 7q31 was first associated with speech and language disorder in a study on 30 members from 4 generations of a family (KE) in which half of the family members were affected by a severe speech and language disorder (Pennington et al. 1991). Those affected mainly showed difficulties in articulation and grammar, but they also showed deficits in phonological processing (Vargha-Khadem et al. 1995). Lai and colleagues (2001) found that a G→A nucleotide transition in exon 14 of *FOXP2* cosegregated perfectly with the speech and language disorder in the KE family. A region 15 Mb from this gene showed linkage to DD in a genome wide scan using 88 subjects from 11 Finnish families (D7S530, NPL = 2.77, P = 0.003) (Kaminen et al. 2003). This was also replicated in a genome wide scan using 403 Australian families (Bates et al. 2007). The marker D7S530 showed evidence of linkage with nonword spelling (LOD = 2.05), irregular word reading (LOD = 1.91), regular word reading (LOD = 1.13) and nonword reading (LOD = 1.21). However, after sequencing the whole coding region of *FOXP2* in six subjects with DD and 3 controls, no mutations in this gene were found, including the G→A transition on exon 14 (Kaminen et al. 2003). This suggests that the *FOXP2* gene specifically affects the speech disorder, and may not be involved in DD. Although individuals with the speech disorder showed phonological processing deficits, these may have been a secondary effect of the speech disorder, rather than being related to an actual reading deficit. Even if *FOXP2* isn't a likely candidate gene for DD, the region of linkage identified by Kaminen et al. (2003) and Bates et al. (2007) is still worthy of further investigation.

1.5 Aim of Thesis

The aim of this thesis was to identify susceptibility variants for DD using several approaches:

- A candidate gene study was conducted selecting genes within the *DYX* linkage regions that either have a possible role within neuronal migration or share homology to putative DD susceptibility genes identified so far. This is presented in Chapter 3 of this thesis.
- The first GWAS of DD was carried out in collaboration with other DD research groups in Europe (NeuroDys collaboration) in an effort to identify new susceptibility variants for DD and this is presented in Chapters 4 and 5 of this thesis.
- An additional GWAS in the form of a pooling study was conducted using the Cardiff case-control sample in order to identify new variants that show association for DD in this homogeneous sample and this is presented in Chapter 6.
- Finally, a CNV analysis was conducted using data from the initial NeuroDys GWAS in order to identify if these types of structural variants may have a significant association with DD. This is presented in Chapter 7 of this thesis.

Chapter 2: Materials and Methods

2.1 Sample Ascertainment and Collection

2.1.1 Collection of the Cardiff Sample

The collection of DNA samples and phenotypic information for this study was undertaken by Dr Gary Hill and colleagues under the supervision of Prof. Julie Williams. The study has ethical approval obtained from local ethics committees in the UK and informed written consent was obtained for all participants in the study. Written consent for children under the age of 18 years was obtained from parental guardians.

DD-probands and their families were ascertained through contacts with Local Education Authorities (LEA) in South Wales and schools specialising in the education of children with reading difficulties in England. All English schools, with the exception of one, were members of Crested (the Council for the Registration of Schools Teaching Dyslexic Pupils).

A pro-rated full-scale IQ score was calculated using four subtests from the WISC III UK, including vocabulary, similarities, block design and picture completion (Weshler 1992). Reading disability was assessed using either the Neale Analysis of Reading Disability (NARA) (Neale 1989) or the British Ability Scale (BAS) single word reading test (Elliot 1983) depending on age and ability of the proband. Whilst NARA is based on prose reading, the BAS single word reading is based on reading a list of words (correlation coefficient between NARA and BAS tests, $r = 0.89$). The measure used to assess reading disability is comparable to other definitions of DD used in molecular genetic studies including those of Grigorenko and colleagues (1997) and others (Fisher et al. 1999; Gayán & Olson 1999). Probands were required to have an IQ of 85 or above and a reading age 2.5 years or more behind their chronological age. This criterion represents a severe degree of reading disability and is likely to represent the lower 5th percentile of children. English was the first language of all participants.

2.1.2 1958 Birth Control Cohort

Some studies presented in this thesis also used population controls from the 1958 British Birth Cohort (or the National Child Development Study (NCDS). This began as a study of Perinatal Mortality focussing on just over 17,000 births in the UK in a single week in 1958 (Power & Elliott 2006). Members of the cohort were then followed-up by parental interview and examination at ages 7, 11, and 16 years and by cohort member interview at 23, 33 and 42 years. The first biomedical assessment in adulthood was conducted at 44-45 years. Genetic data for this cohort has been made available to researchers, and the data presented in this thesis was based on genetic data from the cohort genotyped on the Illumina HumanHap500 array.

2.1.3 Ascertainment Tests

2.1.3.1 British Ability Scales (BAS) Single Word Reading

The BAS single word reading involves reading single words from a list rather than prose text. The test is divided into 9 blocks of 10 words and is discontinued when one block is failed. Each block contains two lines of five words. Individuals read the words across each line before proceeding to the next. Words get progressively more difficult. Failure of a block occurs when five words on one line have been read incorrectly. The test has a retest reliability of 0.96 and a heritability of 0.44 (Hohnen & Stevenson 1999).

2.1.3.2 Neale Analysis of Reading Ability (NARA)

The NARA reading test consists of a set of graded prose passages that allow the testing of rate, comprehension and accuracy of oral reading. Test material is presented in the form of a book, which consists of short, graded narratives, each constructed with a limited number of words and with a central theme, action and resolution. Pictures accompany the narrative however these set the scene rather than tell the story.

Within the book there are six passages of increasing difficulty. Comprehension questions are available after the oral reading of the passage, which tap into the child's use of contextual cues, pictures and prompts. They also test the immediate recall of the main idea of the narrative, the sequence of events and other details. In order to answer some questions inference is required.

The NARA tests are only suitable for children up to the age of 12 years. As a result, analysis undertaken using measures from this test are only done on children aged 12 years and under.

2.1.3.2.1 Reading Accuracy

Children start reading the passages but at a level below expected by their age (since they are poor readers), but which is not too low that they lose interest. Children move to the next level of passage until 16 mistakes have been made (20 on level 6). At this point the reading test is discontinued. The accuracy score is based on the number of words correct out of the number read. The accuracy score is converted to reading age based on population norms.

2.1.3.2.2 Reading Comprehension Task

Reading comprehension was measured on the number of correct questions answered by the child based on the number of passages they read and at what level. Like reading accuracy, a reading comprehension age is calculated and a discrepancy measure calculated between this and the child's chronological age.

2.1.3.2.3 Reading Rate Task

Whilst undertaking the prose reading task, the child is timed. Question answering is not included in the timing of individuals. A time is calculated and used in the following equation:

$$\text{Words per minute} = \frac{\text{Total number of words}}{\text{Total time (seconds)}} \times 60$$

From this equation a reading rate age is calculated based on child norms. An age discrepancy is calculated between the reading rate age and the child's chronological age.

2.2 Samples Used In Thesis

The Cardiff case-control sample consisted of 357 cases with DD and 269 screened controls. The cases consisted of 281 males and 76 females and the controls consisted of 124 males and 145 females. The sample demographics are described in Table 2.1.

	Cases				Controls			
Variable	Mean	SD	Minimum	Maximum	Mean	SD	Minimum	Maximum
Age (years)	13.93	2.17	7.8	17.49	12.2	2.2	5.08	16.67
IQ*	102.05	11.8	85	141	109.5	12.15	85	137
RD (years)	-4.58	1.56	-2.5	-9.13	1.64	1.55	-1.75	5.67

Table 2.1: Demographics of the Cardiff case-control sample. RD = Reading Disability, i.e. the discrepancy between their reading age and their chronological age at time of testing. * Not all controls were tested for IQ therefore demographics are based on the 102 participants tested for IQ

Different sample sets are used throughout this thesis. Figure 2.1 and Figure 2.2 explain these sample sets and show the terms which will be used when referring to these sample sets in this thesis.

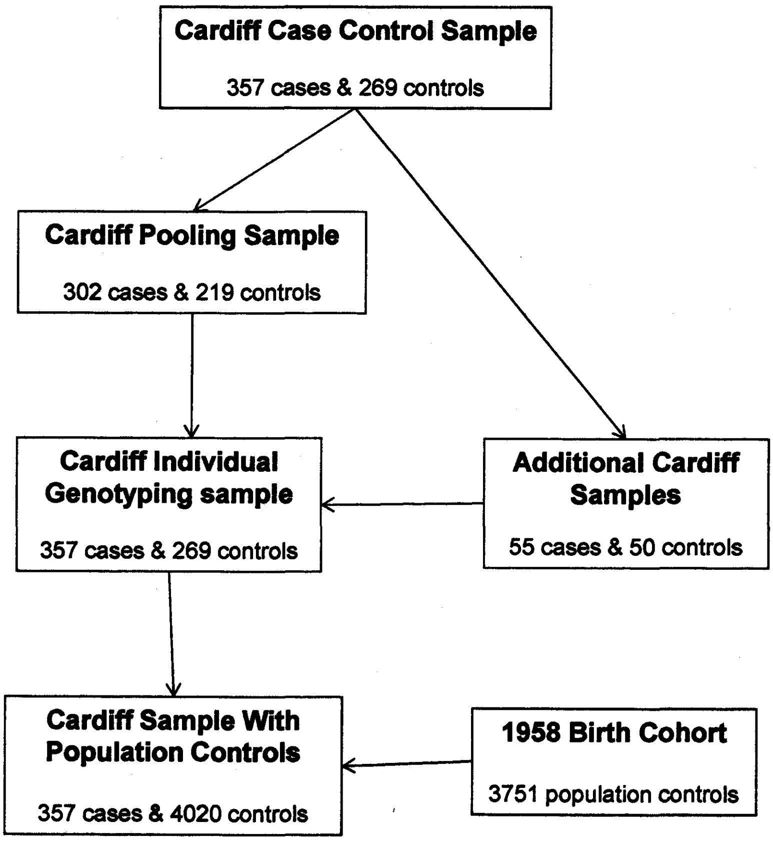


Figure 2.1: Flow diagram explaining the different sample sets of the Cardiff case-control sample that are used in this thesis.

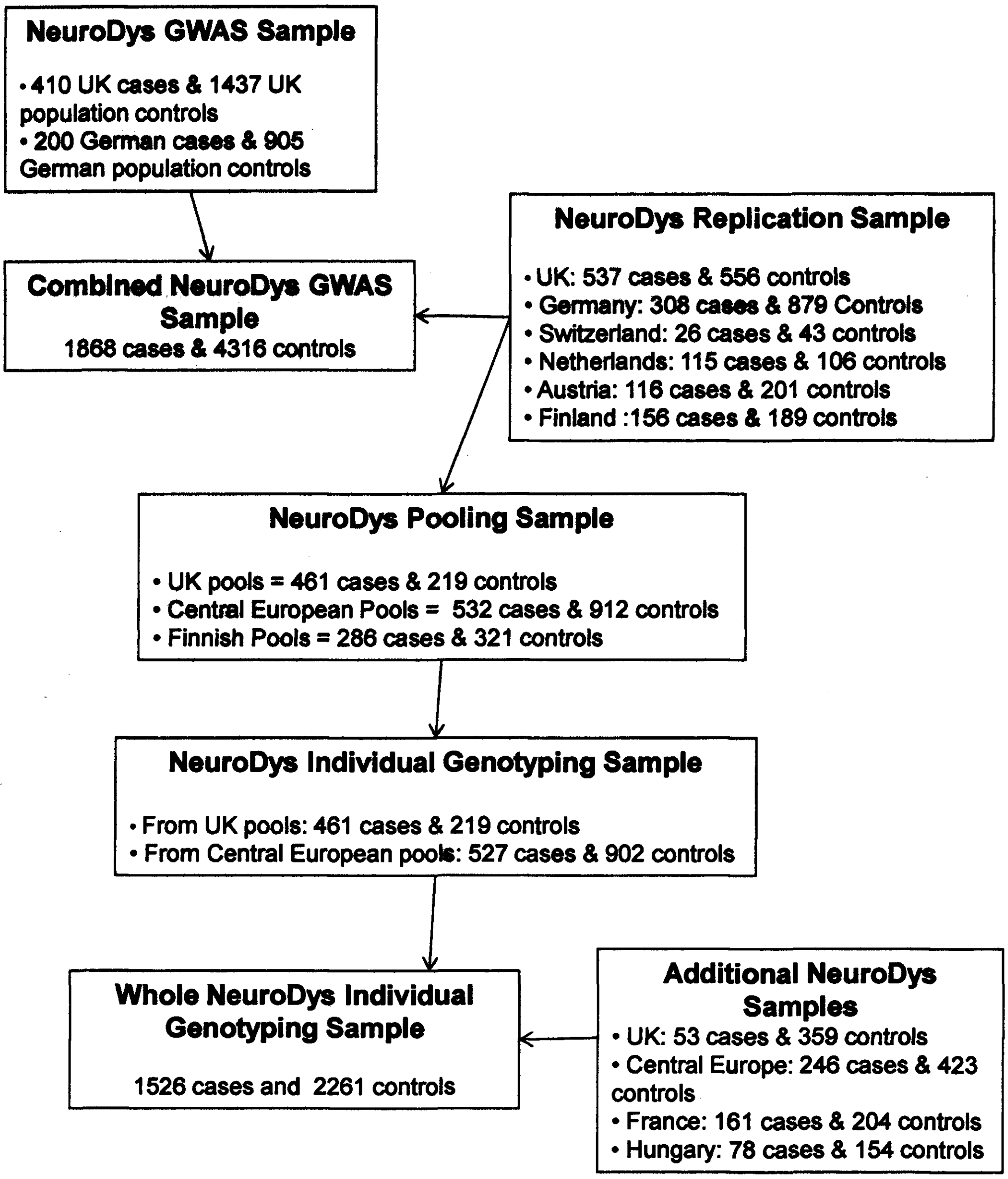

NeuroDys GWAS Sample
• 410 UK cases & 1437 UK population controls
• 200 German cases & 905 German population controls
NeuroDys Replication Sample
• UK: 537 cases & 556 controls
• Germany: 308 cases & 879 Controls
• Switzerland: 26 cases & 43 controls
• Netherlands: 115 cases & 106 controls
• Austria: 116 cases & 201 controls
• Finland :156 cases & 189 controls
Combined NeuroDys GWAS Sample
1868 cases & 4316 controls
NeuroDys Pooling Sample
• UK pools = 461 cases & 219 controls
• Central European Pools = 532 cases & 912 controls
• Finnish Pools = 286 cases & 321 controls
NeuroDys Individual Genotyping Sample
• From UK pools: 461 cases & 219 controls
• From Central European pools: 527 cases & 902 controls
Whole NeuroDys Individual Genotyping Sample
1526 cases and 2261 controls
Additional NeuroDys Samples
• UK: 53 cases & 359 controls
• Central Europe: 246 cases & 423 controls
• France: 161 cases & 204 controls
• Hungary: 78 cases & 154 controls

Figure 2.2: Flow diagram explaining the different sample sets used within the NeuroDys studies that are presented in this thesis.

2.3 DNA and Sample Extraction

2.3.1 DNA Extraction

Venous blood was taken from participants willing to give blood and DNA extracted from lymphocytes using standard procedures. If blood samples could not be given, 25ml saline mouthwashes were obtained. DNA was extracted from buccal cavity epithelial cells by centrifugation at 3000*g* for 15 minutes, followed by incubation with proteinase K (Sigma, USA), SET buffer (Qiagen, UK) and SDS (Invitrogen, UK) at 50°C for 12 hours. The DNA was then isolated by standard phenol-chloroform extraction. Stock and diluted samples were stored in water at -20°C.

2.3.2 DNA Quantification

2.3.2.1 Quantification via Spectrophotometer

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

All DNA stocks were kept at -20°C in individual Eppendorf tubes. Working dilutions of DNA, diluted to 8ng/μl, were kept in 96-deep-well plates at 4°C.

2.3.2.2 Pico Green DNA Quantification

A more accurate quantification of DNA samples was also performed using a Fluoroskan Ascent fluorometer (Thermo Labsystems) and PicoGreen (Invitrogen). Samples were first diluted to less than 50ng/μl based on spectrophotometer readings. Aliquots of the samples were then diluted by a constant factor with 1X TBE in a white 96 well cliniplate such that the DNA concentration was expected to lie in the range of 0.2-1.6ng/μl. A PicoGreen 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 PicoGreen working dilution was dispensed into 100μl of each diluted sample. The fluorometer measures the concentration of a sample using an UV excitation wavelength of 485nm and an emission wavelength of 538 nm. A standard curve was then used to calculate the concentration of DNA for each sample. All DNA stocks were kept at -20°C in individual Eppendorf tubes. Working dilutions of DNA, diluted to 8ng/μl, were kept in 96-deep-well plates at 4°C.

2.4 Constructing Pools

DNA pooling was used in order to allow the estimation of allele frequency differences between cases and controls using fewer genotyping reactions that would be necessary if individual samples were individually assessed. DNA concentration was determined using PicoGreen (as described in section 2.3.2.2). To make sure that high quality DNA samples were used, only those which had genotyping call rates > 98% in previous genotyping studies were included. To produce pools, water was added to produce a target DNA concentration of 40ng/μl. Samples were allowed to equilibrate at 4 °C for 48 hours before quantification using PicoGreen. Each sample was then further diluted to 10ng/μl, allowed to equilibrate at 4 °C for 48 hours, and then quantified again. Based on the final quantification, equimolar amounts of the DNA samples were combined to form the DNA pools.

2.4.1 Concentrating Pooled Samples

After construction, the concentration of the UK pools was 10ng/μl so they were concentrated using Microcon YM-100 Centrifugal Filter tubes (Millipore). The Microcon sample reservoirs were inserted into the vials and the 500μl DNA samples were pipetted into the sample reservoir. The tubes were the sealed and spun at 500g for ~40 minutes until the volume of remaining sample in the reservoir was ~100μl. The sample reservoirs were then placed upside down in a new vial and spun at 1000g for 3 minutes to transfer the concentrate to the vial. The concentrated samples were then quantified using PicoGreen and subsequently diluted to 50ng/μl.

2.4.2 Validating Pooled Samples.

To test the accuracy of the case and control pools in estimating allele frequency differences, the pools were genotyped using SNaPshot, as described in section 2.9.2.

2.5 Polymerase Chain Reaction (PCR)

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 anneal to the template and 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.

The methods given below were used for general PCR reactions. Specific PCR conditions for other techniques are described in the relevant sections.

2.5.1 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, amplimers were restricted to <500bp.

2.5.2 PCR Optimisation

All PCRs were performed on MJ thermocyclers (MJ Research, UK). To find the optimum temperature for primer annealing, optimisation reactions were undertaken for each primer-pair on control DNA using a temperature gradient. PCR reactions were carried out using the following mix shown in Table 2.2 (note: the amount of each reagent is given per single PCR reaction).

Reagent	Company	Volume (µl)
Buffer (10X containing 15mM $MgCl_2$)	Qiagen	1.2
dNTPs (2.5 mM)	Amersham	1.2
Forward Primer (10pmol/µl)	Eurogentec	0.6
Reverse Primer (10pmol/µl)	Eurogentec	0.6
ddH_2O		4.34
HotStarTaq Polymerase (5units/µl)	Qiagen	0.06
Genomic DNA (8ng/µl)		4

Table 2.2: Reagents required for PCR reactions

The PCR cycling conditions were as follows:

94°C for 15 minutes
94°C for 30 seconds
56-66°C for 30 seconds (depending on amplimer)
72°C for 45 seconds
Go to step 2 for 34 cycles
72°C for 10 minutes.

2.6 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 a Kodak Electrophoresis Gel analysis system.

2.7 Mutation Detection

Mutation detection was carried out using High Resolution DNA Melting Analysis (HRMA) on the Lightscanner (Idaho Technologies). HRMA is based on the observation that the melting temperature (Tm) of a PCR amplimer can be largely dependent on its specific sequence composition (Ririe et al. 1997). 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 amplimer'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 (Graham et al. 2005). 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 (Wittwer et al. 2003). 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 (Reed & Wittwer 2004).

2.7.1 HRMA PCR

PCRs were performed in a 12µl reaction using the reagents shown in Table 2.3.

Reagent	Volume (µl)
LCgreen Plus PCR buffer (10X containing 20mM $MgCl_2$)	1.2
LCgreen Plus (10X)	1.2
dNTPs (5 mM)	0.96
Forward Primer (5pmol/µl)	0.56
Reverse Primer (5pmol/µl)	0.56
ddH_2O	3.46
HotStarTaq Polymerase (5units/µl)	0.06
Genomic DNA (8ng/µl)	4

Table 2.3: Reagents and volumes required for HRMA PCR

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

2.7.2 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 (Dwyer et al. 2010). 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% fluorescence respectively. The samples were then analysed using the Lightscanner HRMA software Call-IT^{TM} (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.

2.8 Sequencing

PCR products from individuals showing alternative melt profiles by 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.

2.8.1 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. The PCR product (10µl) was 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 were removed by successive 85% ethanol wash steps. The PCR amplimeres were then eluted in 195µl H20 in a new 96-well skirted plate.

2.8.2 Sequencing Reaction

5µl of the cleaned PCR product was added to a 5µl sequencing reaction mix which was made as shown in Table 2.4.

Reagent	Volume (µl)
5x BigDye sequencing buffer	0.917
BigDye termination mix	0.116
Forward or reverse PCR primer (4pmol/µl)	1
H_2O	1.917

Table 2.4: Reagents and volumes required for the sequencing reaction mix.

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. Repeat steps 2-4 23 times
6. 4°C for 4 minutes

2.8.3 Post Sequencing Clean Up

The post-sequencing clean-up removes any unwanted impurities from the sequencing reaction such as unincorporated ddTTPs. 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 is 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 H20 which can be read directly via a capillary sequencer.

2.8.4 Sequencing Analysis

Samples were then 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. The sequence analysis software package Sequencher (Gene Codes) was used to identify any polymorphisms within the amplimers. This package aligns 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.

2.9 Genotyping

Genotyping of candidate genes and follow up SNP panels was conducted using Sequenom MassARRAY genotype platform. The GWAS were conducted using Illumina SNP arrays and SNaPshot genotyping was used to validate the pooling samples.

2.9.1 Genotyping Using 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 10 nucleotide 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.

2.9.1.1 Sequenom Reactions

The software designs the PCR primers so as to create the shortest amplimer possible that will allow efficient PCR at an annealing temperature of 56°C. This allows a universal PCR condition to be used. Each PCR was performed with 3µl of dried genomic DNA (8ng/µl) in a 384 microtitre plate (ABgene) with the addition of a 5µl PCR mix. The reagents for this mix and the volumes require per sample are shown in Table 2.5.

Reagent	Volume (µl)
10x PCR Buffer	0.625
$MgCl_2$ (25mM)	0.325
dNTPs (25mM)	0.1
HotStarTaq (5units/µl)	0.2
Forward and Reverse PCR primers (1pmol/µl)	0.5
Water	3.25

Table 2.5: Reagents and volumes required per sample for the Sequenom PCR

The following PCR was then performed:

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

A number of negative control samples and genomic DNA positive control samples were electrophoresised on a 2% gel to check for both PCR efficiency and contamination. If the assays passed this quality control (QC), a 2µl Shrimp Alkaline Phosphatase (SAP) mix was added to the 5µl PCR reaction. The SAP mix is shown in Table 2.6.

Reagent	Volume (µl)
SAP	0.3
SAP Buffer	0.17
Water	1.53

Table 2.6: Reagents and volumes required for the SAP mix.

This 7µl reaction mix then underwent 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 was defined by an optimisation procedure involving a small number of DNA samples. The extension primers were split into four groups dependent upon their mass (lowest to highest mass) which were diluted initially to final concentrations of 0.938µM, 1.17µM, 1.425µM and 1.875µM. The extension primers were 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 were adjusted according to their peak height. For example if a peak height was low then the final concentration was increased. At the optimisation stage, failed or abnormal assays (for example self priming assays) were also removed. The 2µl extension mix was made up as shown in Table 2.7

Reagent	Volume (µl)
iPLEX GOLD Reaction Buffer	0.2
iPLEX GOLD Termination Mix	0.2
iPLEX GOLD Enzyme	0.041
Adjusted Unextended Primer Mix	1.559

Table 2.7: Reagents and volumes required for the extension mix.

The 9µl reaction then underwent the following thermocyclic conditions:

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. Repeats steps 3 and 4, 4 times
6. Repeat steps 2 to 5 39 times
7. 72°C for 3 minutes
8. 15°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 was 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 were spun in centrifuge for 15minutes at 3000rpm to separate the resin from the solution.

2.9.1.2 Sequenom Analysis

Samples were 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. 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 primer compared to residual unextended primer). These genotypes can then be viewed and manually revised by the user using the Typer software, as shown in Figure 2.3.

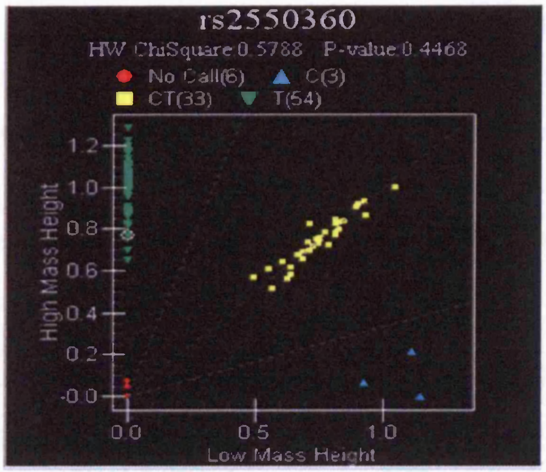


Figure 2.3: Screenshot from Typer analysis software (Sequenom) for rs2550360. TT homozygotes are shown in green, C homozygotes in blue and CT heterozygotes in yellow. No Calls are shown in red.

2.9.1.3 Accurate Genotyping

All assays were initially optimised by genotyping DNA from 30 CEPH parent-offspring trios from Utah with ancestry from northern and western Europe (CEU). 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 and affected status. 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.

2.9.2 Genotyping Using 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 strep 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 SNaPshot genotyping, each sample underwent a standard 12µl PCR reaction using the reagents and volumes shown in Table 2.8.

Reagent	Volume (µl)
10x Buffer	1.2
dNTPs (5mM)	0.96
Forward primer (5pmol/µ)	0.28
Reverse primer (5pmol/µ)	0.28
HotStarTaq (5units/µl)	0.06
dH_2O	6.22
Genomic DNA (8ng/µl)	3

Table 2.8: Reagents and volumes required for the SNaPshot PCR Reaction

The PCR cycling conditions are outlined below:

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

T_A°C is a function of the melting temperature of each primer set.

SAP (Amersham) and exonuclease I (Amersham) were then added to each PCR product to degrade unincorporated dNTPs and unextended primers. The reaction involved the addition of a 5µl SAP mix described in Table 2.9 to the 12µl PCR product.

Reagent	Volume (µl)
SAP	0.5
Exonuclease I	0.1
Water	4.4

Table 2.9: Reagents and volumes required for the SAP mix in the SNaPshot reaction.

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 shown in Table 2.10 was added to 2µl of the cleaned PCR product.

Reagent	Volume (µl)
SNaPshot Reagent	1.25
Reaction buffer	3.75
Extension primer	1
Water	2

Table 2.10: Reagents and volumes required for the SNaPshot extension mix. The SNaPshot reagent contains fluor-labelled ddNTPs and a sequenase.

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

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

Here, 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 (Norton et al. 2002).

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 were 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 2.4). Along with genomic DNA, negative controls are added at the PCR and SNaPshot stages to check for contamination.

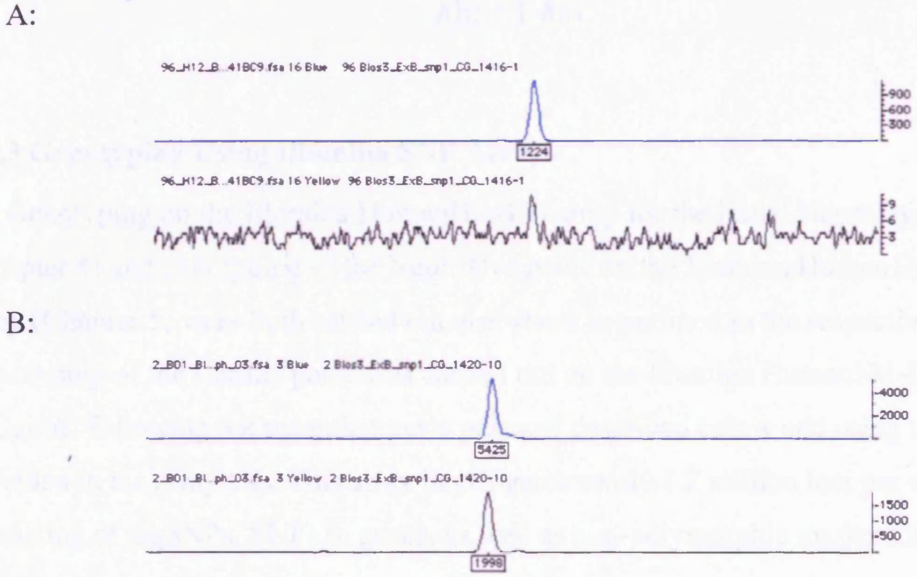


Figure 2.4: Individual genotyping via SNaPshot using Genotyper software (Genecodes). Part A shows a GG homozygote and part B shows a GC heterozygote.

In the analysis of bi-allelic markers in DNA pools, the primer extension products may not be represented with equal efficiency, thus not providing an accurate basis for the calculation of allele frequencies. To allow for the unequal representation of alleles, the estimated allele frequencies from pools were corrected by using the mean of the ratios obtained from measurements taken from heterozygous samples. Since heterozygous individuals contain one of each allele at a known polymorphism, fluorescence corresponding to each allele of a SNP should be the same (under perfect conditions) resulting in a 1:1 ratio of fluorescence units for each allele. Any deviation away from 1:1 fluorescence ratio can be determined and so pooled assays can be corrected for any unequal representation of the alleles from the known heterozygote ratio. The correction could be made using the following equation:

$$f(\mathrm{a}) = A/(A+kB)$$

Where A and B are the peak heights of the primer extension products representing alleles A and B in a pool and k is the mean of the replicates of A/B ratios observed in a heterozygote (Hoogendoorn et al. 2000). f(a) is the frequency of allele A. The frequency of allele B (f(b)) was then calculated from the formula:

$$f(\mathrm{b}) = 1\text{-}f(\mathrm{a})$$

2.9.3 Genotyping Using Illumina SNP Arrays

Genotyping on the Illumina HumanHap300 array for the initial NeuroDys GWAS (Chapter 4) and genotyping of the NeuroDys pools on the Illumina Human1M-Duo array (Chapter 5) were both carried out elsewhere as outlined in the respective chapters. Genotyping of the Cardiff pools was carried out on the Illumina Human1M-Duo array in Cardiff following the manufacturer's protocol described below and using the reagents provided in the array kits. This array interrogates nearly 1.2 million loci per sample, consisting of tagSNPs, SNPs in genes, as well as non-polymorphic markers in known CNV regions.

2.9.3.1 Preparation of DNA Samples for Array Genotyping

The first step involved denaturing the DNA samples using NaOH. 8µl of DNA at 50ng/µl was mixed with 8µl of 0.1 NaOH and incubated for 10 minutes at room temperature. 135µl the Illumina MP1 reagent was then added to neutralise the DNA. To perform uniform whole-genome amplification of the DNA, 150µl of the Illumina AMM was added to each sample before incubating at 37°C for 20-24 hours.

The samples were then split into two, each containing 150µl. 50µl of the Illumina FRG reagent was then added before heating at 37°C for an hour. This step enzymatically fragments the DNA, using an end-point method to avoid over fragmentation. The DNA was then precipitated by adding 100µl of the Illumina PA1 reagent, followed by incubation at 37°C for 5 minutes before the addition of 300µl of 100% 2-propanol. The samples were mixed and then incubated at 4°C for 30 minutes and then spun at 3000g at 4°C for 20 minutes. The supernatant was removed and the remaining DNA pellets were left to dry at room temperature for 1 hour. The DNA pellets were then resuspended by adding 46µl of the Illumina RA1 reagent before incubation at 48°C for 1 hour.

2.9.3.2 Hybridisation of DNA to BeadChips

This process involved hybridising the fragmented DNA to locus specific 50-mer oligos which were covalently linked to one of over 1,100,000 bead types on the surface of the array. These carefully designed 50-mer probes selectively hybridize to the loci of interest, stopping one base before the interrogated marker.

The samples were first denatured at 95°C for 20 minutes, before rejoining the previously split samples. The BeadChips were then placed into hybridisation chamber inserts and 84µl of each DNA sample was dispensed into each BeadChip inlet port (2 samples loaded for each BeadChip). The BeadChips were visually inspected to ensure that the DNA samples covered all of each of the chips. The chips and their hybridisation chamber inserts were then loaded into hybridisation chambers containing 400µl of the Illumina PB2 reagent in the humidifying buffer reservoirs. The chambers were then closed securely and incubated at 48°C for 16-24 hours.

The BeadChips were removed from the hybridisation chambers and the IntelliHyb seals were removed from the surface of the chips. The chips were then washed in two washes of the Illumina PB1 reagent in order to remove any unhybridised DNA. The chips were then immediately assembled into flow-through chambers in preparation for single base extension and staining.

2.9.3.3 Single-base Extension

The next process involved carrying out single-base extension of the oligos on the BeadChip using the captured DNA as a template. This incorporated detectable labels on the BeadChip in order to determine the genotype call of the sample. C and G nucleotides were biotin labelled (stained green) and A and T nucleotides were dinitrophenol labelled (stained as red).

The assembly of the chips within flow-through chambers allows for a series of reagents to be flowed evenly across the surface of the chip whilst the chip is incubated at certain temperatures in a chamber rack. Each chip within its flow-through chamber was placed in a chamber rack heated to 44°C. 150µl of the Illumina RA1 reagent was first added to the reservoir of each flow-through chamber to wash away unhybridised and non-specifically hybridised DNA sample. The chips were left to incubate at 44°C for 30 seconds before this step was repeated another 5 times. To prepare the chips for the extension reaction, 450µl of the Illumina XC1 reagent was then added before incubating for 10 minutes. 450µl of the Illumina XC2 reagent was added next followed by another 10 minute incubation. To extend the primers hybridised to the DNA on the BeadChip and incorporate labelled nucleotides into the extended primers, 200µl of the Illumina TEM reagent was added to the flow-through chamber and left to incubate for 15 minutes. 450µl of 95% formamide/1mM EDTA was then added twice at 1 minute intervals in order to remove the hybridised DNA. The chips were then left to incubate at 44°C for 5 minutes before adding 450µl of the Illumina XC3 reagent twice at 1 minute intervals to neutralise the chips.

2.9.3.4 Staining of the BeadChips

The next process involved dual-colour staining so that the nucleotides incorporated during the extension step could be detected by the Illumina iScan imaging system.

The temperature of the chamber rack was lowered to 37°C. The multi-layer staining process was carried out by adding a sequence of Illumina staining reagents (STM and ATM) and washing with Illumina XC3, with incubation steps in between, as described below:

250µl of STM, incubate for 10 minutes.
450µl of XC3, incubate for 1 minute. Repeat once and then wait 5 minutes.
240µl of ATM, incubate for 10 minutes.
450µl of XC3, incubate for 1 minute. Repeat once and then wait 5 minutes.
Repeat steps 1-4 twice more.

After staining, the flow-through chambers were immediately removed from the chamber rack and placed horizontally on the lab bench. The BeadChips were removed from the flow-through chambers and placed into a staining rack before being washed with PB1 reagent to thoroughly remove any remaining staining reagents. The staining rack was then submerged in XC4 and allowed to soak for 5 minutes in order to coat the surface of the BeadChips. The chips were then left to dry on a tube rack inside a vacuum dessicator at 508 mm Hg (0.68 bar) for 50-55 minutes. The undersides of the chips were then carefully cleaned with ethanol in order to remove any excess XC4 and ensure that the chips would lie flat in the iScan Reader tray.

2.9.3.5 Processing the BeadChips and Extracting Normalised Intensities

The BeadChips were then imaged using the iScan system. The iScan Reader uses a laser to excite the fluor of the single-base extension product on the beads of the BeadChip. Light emissions from these fluors are then recorded in high-resolution images. If fluorescence could not be detected from all sections of the chip then the samples hybridised to these chips were excluded from further analyses. Data from the images produced by the iScan was then exported into the BeadStudio Genotyping Module v3.2.

The BeadStudio Genotyping Module v3.2 enables the user to normalise the signal intensities obtained from the BeadChips. Because the performance of external controls can vary from sample to sample, Illumina have developed a self-normalisation algorithm that uses information contained within the array itself (www.illumina.com). This algorithm is designed to adjust for channel-dependant background and global intensity differences.

The normalised signal intensities of each SNP were then used to estimate allele frequencies in the pooled samples, as described in Chapters 5 and 6.

2.10 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 Sarah Dwyer.

2.11 Statistical/Bioinformatic Analysis

2.11.1 Tag SNP Determination

Tag SNP identification was performed using Haploview (http://www.broad.mit.edu/haploview/haploview), a software program designed for genetic association studies (Barrett et al. 2005). 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, minor allele frequency 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.

2.11.2 LD Estimation

The LD between markers was determined using the SNP Annotation and Proxy Search tool (http://www.broadinstitute.org/mpg/snap/ldsearch.php) (Johnson et al. 2008). SNAP finds proxy SNPs based on LD, physical distance and/or membership in selected commercial genotyping arrays (e.g. Illumina arrays). Pair-wise LD is pre-calculated based on phased genotype data from the International HapMap Project. SNPs were considered to be in highly correlated if the r^2 value between them was ≥ 0.8 and a value < 0.5 was considered to indicate low correlation.

2.11.3 Sample Size Power Calculations

Power calculations for the mutation detection sample was 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 odds ratio (OR) at a given P-value were calculated using the software PS Power and Sample Size Calculations (Dupont & Plummer 1990).

2.11.4 Statistical Analysis Using PLINK

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) (Purcell et al. 2007). These included tests for single marker association (additive and genotypic), and haplotype analysis. Analysis of variants for deviations from Hardy-Weinberg equilibrium were also performed using PLINK. Statistical analyses that did not use the PLINK software are found within the methods section of the relevant chapter.

2.11.4.1 Tests for Deviation from Hardy-Weinberg Equilibrium

The Hardy-Weinberg principle states that if an infinitely large, random mating population is free from outside evolutionary forces (i.e. mutation, migration and natural selection), the gene frequencies will not change over time and the allele frequencies of A and a in the next generation will be p^2 for the AA genotype, 2pq for the Aa genotype and q^2 for the aa genotype.

Departures from the Hardy-Weinberg equilibrium (HWE) can be due to inbreeding, population stratification, selection or they can be a symptom of disease association (Balding 2006). Deviations from HWE can arise in the presence of a common deletion polymorphism, because of a mutant PCR-primer site or because of a tendency to miscall heterozygotes as homozygotes. This has led to researchers testing for HWE as a data quality check. In this study, SNPs were tested for deviations from Hardy-Weinberg in the control samples using PLINK. PLINK tests for deviations from the Hardy-Weinberg equilibrium using the Exact test described by Wigginton and colleagues (2005). This test is more accurate for rare genotypes compared with the χ^2 goodness-of-fit test (Wigginton et al. 2005).

2.11.4.2 Association Tests

Genotypic tests were carried out in PLINK which uses a χ^2 test with 2 degrees of freedom to test the null hypothesis that there is no association between rows and columns of a 2 x 3 matrix containing the counts of the three genotypes among cases and controls as shown in Table 2.11.

	Genotypes			
	AA	Aa	aa	Total
Cases	n_{1AA}	n_{1Aa}	n_{1aa}	n_1
Controls	n_{2AA}	n_{2Aa}	n_{2aa}	n_2
Total	n_{AA}	n_{Aa}	n_{aa}	N

Table 2.11: An example of a 2 x 3 contigency table to present genotype data in cases and control.

As the χ^2 genotypic test cannot be reliably used when the counts for a particular genotype are less than 5, the CLUMP programme was also used to test for genotypic association in some instances using 1000 permutations (Sham & Curtis 1995). This method calculates the χ^2 based on the 2 x 3 contigency table but uses Monte Carlo simulations in order to test the significance.

For complex traits, it is widely thought that the contributions to disease risk from individual SNPs will often be roughly additive - i.e. the heterozygote risk will be intermediate between the two homozygote risks (Balding 2006). Additive tests were conducted using the Cochran-Armitage trend test in PLINK. This test was used instead of the χ^2 allelic test as it is more robust to departures from the HWE. The trend test tests the hypothesis of zero slope for a line that fits the three genotypic estimates best, as shown in Figure 2.5.

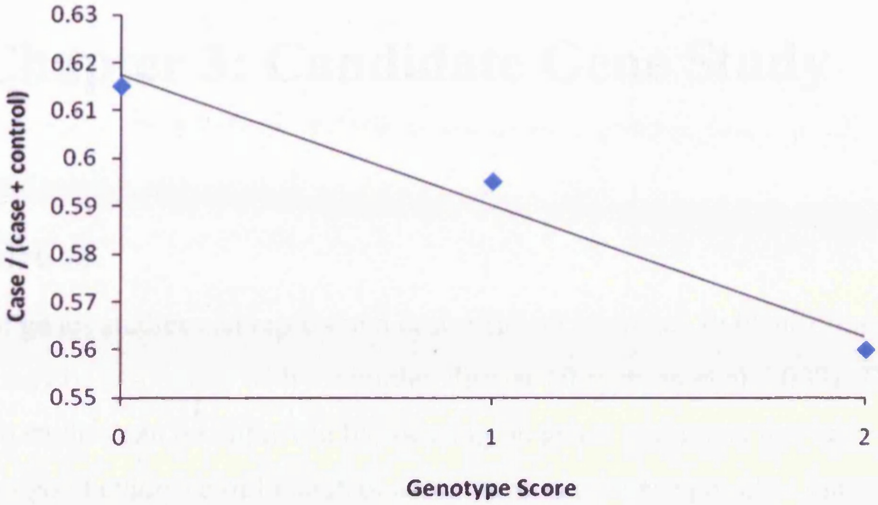


Figure 2.5: Cochran-Armitage trend test. The dots indicate the proportion of the cases, among cases and controls combined, at each of three SNP genotypes (coded as 0,1, and 2). The Cochran-Armitage trend test corresponds to testing the hypothesis that the line has zero slope. Adapted from Balding (2006).

Logistic regression within PLINK was also used to test for association in some instances so that covariates could be included. Both the additive and genotypic tests were performed within the logistic regression framework.

2.11.5 Imputation

Imputation of SNPs which had not been genotyped in the intial NeuroDys GWAS was carried out using PLINK. A reference panel of genotypes was obtained from Phase II of the International HapMap Project (The International HapMap Consortium 2007) and merged with the GWAS data, after ensuring that both sets of data were aligned on the positive strand. Only those imputed SNPs with an INFO score > 0.8 were included in further analyses, as recommended in the PLINK documentation.

Chapter 3: Candidate Gene Study

3.1 Introduction

Candidate genes studies can represent a cost-efficient approach to identifying variants that may be associated with a complex disease (Jorgensen et al. 2009). The power of such studies can be improved by focusing on genes in regions that have already shown good evidence of linkage or association and whose possible functions indicate a plausible role within the disease.

This candidate gene study focuses on a small number of genes that are in well replicated linkage regions for DD, have interesting functions in terms of neuronal migration and on those that show homology or correlated expression with previously identified DD candidate genes.

As discussed in Chapter 1, impaired neuronal migration has been previously linked with DD through the identification of neocortical malformations in the post-mortem brains of individuals with DD (Galaburda & Kemper 1979; Galaburda et al. 1985; Humphreys et al. 1990) and also through behavioural studies on patients with the neuronal migration disorder PNH (Chang et al. 2005). Neuronal migration involves the rearrangement of cytoskeletal components in response to extracellular cues, mediated by numerous intracellular signalling pathways resulting in the organisation of neurons into six specialised laminar layers (Ayala et al. 2007). Notably, four genes that have received support for an association with DD are thought to be involved in controlling neuronal migration.

As discussed in section 1.4.2, *KIAA0319* is in the *DYX2* linkage region on chromosome 6p21.2-6p22.3 and has shown association with DD in several studies (Deffenbacher et al. 2004; Francks et al. 2004; Cope et al. 2005a; Harold et al. 2006; Luciano et al. 2007; Paracchini et al. 2008). Functional studies on *KIAA0319* have highlighted a possible role for this gene in neuronal migration. Paracchini et al. (2006) showed that the risk haplotype spanning *TTRAP* and *KIAA0319* is associated with a reduction of *KIAA0319* gene expression in cell line models and demonstrated that knocking down *Kiaa0319* through *in utero* RNAi in the developing rat neocortex significantly reduces the distance migrated by neurons. Peschansky et al. (2009)

recently demonstrated that the disruption in neuronal migration caused by reduced expression of *KIAA0319* results in specific types of anomalies in the postnatal brain, such as periventricular heterotopias (PVHs) and abnormal laminar locations. These anomalies are similar to the anomalies identified by Galaburda and others in their post-mortem studies on the brains of individuals with DD (Galaburda & Kemper 1979; Galaburda et al. 1985; Humphreys et al. 1990).

DCDC2 is also within the *DYX2* region and, as discussed in section 1.4.2, has shown association with DD in many studies (Deffenbacher et al. 2004; Francks et al. 2004; Meng et al. 2005b; Schumacher et al. 2006a; Harold et al. 2006; Wilcke et al. 2009). *DCDC2* contains two doublecortin peptide domains. These protein domains were originally described in the doublecortin gene (*DCX*), which has a well-recognised role in neuronal migration. Defects in *DCX* have been shown to cause the neuronal migration disorders X-linked lissencephaly and double cortex syndrome (des Portes et al. 1998; Gleeson et al. 1998). In X-linked lissencephaly, *DCX* mutations in males result in a severe disruption to cortical neuronal migration, leading to a rudimentary four-layered cortex (Berg et al. 1998). Mutations in *DCX* heterozygous females lead to a less severe disease, double cortex, in which some neurons form a relatively normal cortex, while a second population of neurons apparently arrests, leading to a collection of neurons beneath the outgrowth cortex (Gleeson & Walsh 1997). Gleeson and colleagues (1999) have shown that DCX is expressed in migrating neurons throughout the central and peripheral nervous system and may direct neuronal migration by regulating the organisation and stability of microtubules. Meng et al. (2005b) used *in utero* RNAi in developing rat neocortex to test for a functional role of *DCDC2* in neuronal migration and found that a loss of function of *Dcdc2* resulted in abnormal migration during the prenatal period. Burbridge et al. (2008) found that knocking down *Dcdc2* resulted in neocortical malformations such as PVHs in the cerebral cortices of the brains of postnatal rats.

As the name suggests, *DYX1C1* lies within the *DYX1* region and has shown association with DD in a number of studies (Taipale et al. 2003; Scerri et al. 2004; Wigg et al. 2004) although this association was not supported by others (Bellini et al. 2005; Cope et al. 2005b; Marino et al. 2005; Meng et al. 2005a; Bates et al. 2007) (see section 1.4.1 for more detail on this region). Functional studies on *DYX1C1* have found that *in utero* RNA interference (RNAi) of *Dyx1c1* in the developing rat neocortex prevents correct neuronal migration and this results in heterogeneous malformations that appear to be associated with distinct impairments in auditory processing and spatial learning (Threlkeld et al. 2007; Wang et al. 2006). When examining these malformations in the adult rat brain, Rosen et al. (2007) found that they resemble malformations identified in the brains of post mortem dyslexics. Other functional studies have suggested that DYX1C1 interacts with the U-box protein CHIP (Carboxy terminus of Hsc70-Interacting Protein) (Hatakeyama et al. 2004), and regulates the levels of estrogen receptors alpha (ERα) and beta (ERβ) (Massinen et al. 2009). ER receptors been shown to be important in brain development and to be involved in cognitive processes and memory. This suggests that *DYX1C1* may affect neuronal migration via interactions with ERs.

ROBO1 is within the *DYX5* region and has shown association with DD, although only in one study (Hannula-Jouppi et al. 2005) (see section 1.4.5). The *Drosophila* homolog of this gene, *Robo*, is thought to be a neuronal axon guidance receptor gene involved in brain development and controls whether or not axons cross the central nervous system midline (Kidd et al. 1998).

This section of the thesis sought to identify genes encoding proteins with a possible role in neuronal migration that also lie within major DD linkage regions. The gene cell division cycle 42 (*CDC42*) lies on chromosome 1p36.12 within the *DYX8* region. This region has been linked with dyslexia in several studies (Rabin et al. 1993; Grigorenko et al. 2001; Tzenova et al. 2004; Franke et al. 2006; Bates et al. 2007). The protein encoded by this gene is a member of the Rho family of GTPases and has been shown to regulate the formation of filopodia. These occur at the leading edge of migratory neurons and sense environmental cues, enabling the neurone to migrate to the correct place (Kozma et al. 1995; Nobes & Hall 1995; Luo 2000). BioGPS (http://biogps.gnf.org) (Wu et al. 2009) shows that *CDC42* is highly expressed in the fetal brain and possesses a similar expression pattern to that of *KIAA0319* ($r = 0.913$).

Protogenin (*PRTG*) lies on chromosome 15q21.3 within the *DYX1* region. This region showed linkage to dyslexia in a large number of studies (Smith et al. 1983; Smith et al. 1991; Grigorenko et al. 1997; Schulte-Körne et al. 1998; Nöthen et al. 1999; Bakker et al. 2003; Chapman et al. 2004; Smith et al. 2005; Bates et al. 2007). Its name is based on the fact that it is expressed during early development of the nervous system ('proto') and is structurally similar to *Neogenin*. The protein encoded by this gene is most closely related to the *DCC-Neogenin* subclass of the Ig superfamily of proteins (Toyoda et al. 2005). Functional studies on *DCC* and *Neogenin* have revealed important roles in axonal guidance in a number of species (Culotti & Merz 1998). Expression studies on the chicken (Toyoda et al. 2005), murine and zebrafish (Vesque et al. 2006) homologues of *PRTG* have suggested that *Prtg* may have a conserved role in axis elongation. In a recent study by Wigg et al. (2008), 20 markers across *PRTG* were genotyped in 253 families with proband diagnosed with ADHD. They found association with this gene and ADHD as both a categorical trait and with symptoms of ADHD measured as quantitative traits, but failed to find evidence for association with two key components of reading. As discussed in section 1.1.5.1, previous twin studies have found evidence for shared genetic factors between ADHD and DD, particularly for inattention and reading phenotypes (Willcutt et al. 2007). Further support for genetic overlap between these disorders comes from overlapping chromosomal regions that have been identified for by linkage and association studies (Wigg et al. 2008). Therefore, a gene that has previously shown evidence for an association in one of the disorders may also by associated with symptoms of the other.

In addition to those genes that may be involved in neuronal migration, this study sought to identify candidate genes based on their homology with genes that have shown convincing evidence of association with DD previously. Two genes within the *DYX8* susceptibility locus show homology with *KIAA0319* and *DCDC2* which have both shown replicated evidence for a role within DD.

The protein sequence of *KIAA0319-like* (*KIAA0319L*) shows a 61% similarity to that of *KIAA0319.* It also contains PKD domains, which are thought to mediate cell-cell adhesion. Two of the four PKD domains in *KIAA0319L* show a 69% and 75% protein sequence similarity with the two PKD domains of *KIAA0319* suggesting a similar function for these genes. *KIAA0319L* has already shown nominal association with DD in a previous study. Couto and colleagues (2008) genotyped 5 SNPs within *KIAA0319L* in a sample of 156 Canadian families. Evidence for association was found with the marker rs7523017 (P = 0.042) when DD was defined as a categorical trait (≤ 1.5 standard deviations below the population mean on two of three standardised reading tests or ≤ 1 standard deviation on the average of the three tests). They also identified a significant haplotype with the markers rs1203138, rs1203148, rs12408030, and rs7523017 (C/A/A/A; P = 0.031). When using quantitative measures of DD and their whole sample of 291 families (the 156 previous families plus an additional 135 families that had been ascertained through a proband who had reading difficulties but did not meet the categorical criteria), this haplotype showed significant association with measures of word-reading efficiency (P = 0.032) and rapid object and colour naming (P = 0.047). However, as the authors point out, these results would not withstand correction for multiple testing.

DCDC2b is on chromosome 1p35.1 and shows a 34% protein sequence similarity to *DCDC2*. Its DCX domain shows 48% protein sequence similarity to the DCX domain of *DCDC2* and so may prove to have a similar role within neuronal migration.

Myers and colleagues (2007) conducted a survey of gene expression on 193 neuropathologically normal human brain samples using the Affymetrix Gene Chip Human Mapping 500K array set and Illumina HumanRefseq-8 Expression BeadChip platforms. This survey has provided a resource that allows for the assessment of genetic effects on normal human cortical gene expression. This study showed that the minor alleles of two SNPs within the *KIAA0319* gene were both associated with the expression of the gene *RIOK3* (Myers et al. 2007). This suggests that SNPs within *KIAA0319* may have a trans-acting effect on the expression of *RIOK3* and as SNPs within *KIAA0319* have shown association with DD previously, this makes *RIOK3* a potential candidate gene for DD. In addition, this gene is on chromosome 18 and lies just outside the *DYX6* linkage region. The specific function of this gene has not yet been determined and it shows ubiquitous expression.

3.1.2 Aims

As these genes are within or near to the *DYX* linkage regions and have putative roles in neuronal migration/axonal guidance (*CDC42* and *PRTG*), show homology to previously associated DD genes (*DCDC2b* and *KIAA0319L*) or because their expression appears to affected by SNPs in a convincing candidate gene for DD (*RIOK3*), variants within these genes were tested for association with DD in the Cardiff case-control sample.

3.2 Methods

3.2.1 Sample

Variants within the candidate genes were genotyped in the Cardiff case-control sample, consisting of 357 DD cases and 269 screened controls. As previously described in Chapter 2, the criteria for cases is an IQ $\geq$ 85 and a reading age that is $\geq$ 2.5 years below their chronological age. DNA for these samples were extracted from either blood or saliva samples using phenol/chloroform methodology as previously described (see Chapter 2). DNA quantification and dilution was also as described (Chapter 2), with a final sample dilution of 8ng/µl.

3.2.2 Genotyping

For the genes *CDC42, PRTG, KIAA0319L* and *RIOK3*, SNPs were identified using HapMap (Rel 22). The regions to be tagged were extended at either ends of the genes to reach the edges of blocks of LD using Haploview v4.01. For *CDC42* the region of chromosome 1 that was tagged extends from 22,218,596-22,309,033, for *PRTG* the tagged region on chromosome 15 extends from 53,676,324-53,834,838bp, for *KIAA0319L* the region of chromosome 1 that was tagged extends from 35,491,743-35,854,982 bp, and for *RIOK3* the tagged region on chromosome 18 extends from 19,278,296-19,325,923 bp (NCBI build 36.1). The Tagger function in Haploview was then used to select a panel of tagSNPs for each gene based on pairwise tagging, using an r^2 threshold of > 0.8 and a minor allele frequency (MAF) threshold of > 0.05. 12 tagSNPs were selected for *CDC42*, 35 for *PRTG*, 13 for *KIAA0319L* and 10 for *RIOK3*. One of the SNPs within *KIAA0319* (rs16889511) that had been shown to be associated with reduced expression of *RIOK3* (Myers et al. 2007) was also genotyped. An assay for the other SNP that showed an association with the expression of *RIOK3* could not be designed alongside the other SNPs in the panels, however this SNP was in perfect LD with rs16889511 ($D' = 1, r^2 = 1$).

At the time of designing this study, *DCDC2b* did not have any known SNPs with a minor allele frequency $\geq$ 0.05 in the HapMap CEU population (HapMap Release 22). Therefore this gene was screened for polymorphisms using high resolution melting analysis (HRMA). 23 primer pairs were designed using the Primer 3 software (Rozen &

Skaletsky 2000) to amplify the length of *DCDC2b*, plus an extra 2000 bp of sequence in the 5' direction and 300 bp in the 3' direction (chr1:32445282-32454285; NCBI Build 36.1) (see Table A.1 in Appendix for primer sequences). After amplification by PCR, each amplimer was screened for polymorphisms using HRMA using the LightScanner and a sample of 15 DD subjects (see Chapter 2 for protocol). Any polymorphisms were then characterised by DNA sequencing on the ABI3100 PRISM (see Chapter 2 for sequencing protocol). This method for mutation detection has been tested, optimised and automated in the department by Sarah Dwyer (Dwyer et al. 2010).

All SNPs were genotyped using the Sequenom MassARRAY iPlex GOLD system in accordance with the manufacturers' instructions (see Chapter 2 for description). PCR primers and extensions probes for each SNP were designed using the Assay Design v3.1 software and genotype calling was carried out using the Typer 3.4 Software (see Table A.2 in Appendix for primer sequences). All SNP assays were initially optimised by genotyping DNA from 30 CEPH parent-offspring trios. Cluster plots for all SNPs were inspected manually, and SNP assays that did not produce distinct clusters were excluded. 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 raters. 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.

After genotyping, samples were removed if their call rate was less than 70% in order to exclude poor quality samples. SNPs were tested for Hardy-Weinberg equilibrium in controls and their MAFs were calculated using PLINK v1.05 (Purcell et al. 2007). PLINK was also used to test for association using the 'model' function to perform genotypic and Cochran-Armitage trend tests. Haplotype analysis was performed using the Unphased programme (Dudbridge 2003).

3.3 Results

3.3.1 *CDC42* Analysis

Out of 22 markers at the *CDC42* locus, 12 SNPs tagging *CDC42* at $r^2 \geq 0.8$ and a MAF > 0.05 in the HapMap CEU population were selected. These were genotyped in the Cardiff case-control sample, of which 355 cases and 268 controls passed QC. One of the SNPs was dropped due to a poorly performing assay. No proxy for this tag SNP could be found at an r^2 value ≥ 0.8. The remaining 11 SNPs had a call rate above 98% and were all in HWE in the controls (P > 0.05) as shown in Table A.3 in the Appendix. As shown in Table 3.1 and Figure 3.1, none of these variants showed significant association with DD in this sample, with the lowest P-value being obtained for rs10917139 (additive test P = 0.145). Haplotype analysis was also carried out for all possible haplotype combinations, but none were significant (see Table A.4 in Appendix for the haplotype combinations with the lowest global haplotype P-values).

							P-values	
SNP	Position (bp)	Minor Allele	MAF Cases	MAF Control	OR	95% CI	Genotypic	Additive
rs11577378	22293036	T	0.137	0.110	1.279	0.91-1.81	0.370	0.159
rs2501275	22247652	C	0.223	0.195	1.180	0.90-1.56	0.421	0.227
rs2501276	22246211	T	0.094	0.084	1.142	0.77-1.70	0.780	0.527
rs1534949	22298774	C	0.410	0.380	1.131	0.90-1.42	0.566	0.294
rs10917139	22274125	A	0.138	0.111	1.289	0.91-1.82	0.345	0.145
rs2473317	22267838	G	0.145	0.127	1.161	0.84-1.62	0.647	0.375
rs17837965	22267212	G	0.048	0.045	1.078	0.63-1.84	0.917	0.785
rs12035094	22257300	T	0.023	0.024	0.933	0.45-1.96	0.983	0.852
rs2473323	22261914	C	0.062	0.062	0.998	0.63-1.60	0.352	0.993
rs2056975	22281114	A	0.062	0.061	1.011	0.64-1.61	0.347	0.964
rs2268177	22287997	T	0.185	0.185	1.002	0.75-1.34	0.924	0.988

Table 3.1: Results of *CDC42* genotyping. MAF – minor allele frequency, OR – odds ratio, CI – confidence interval.

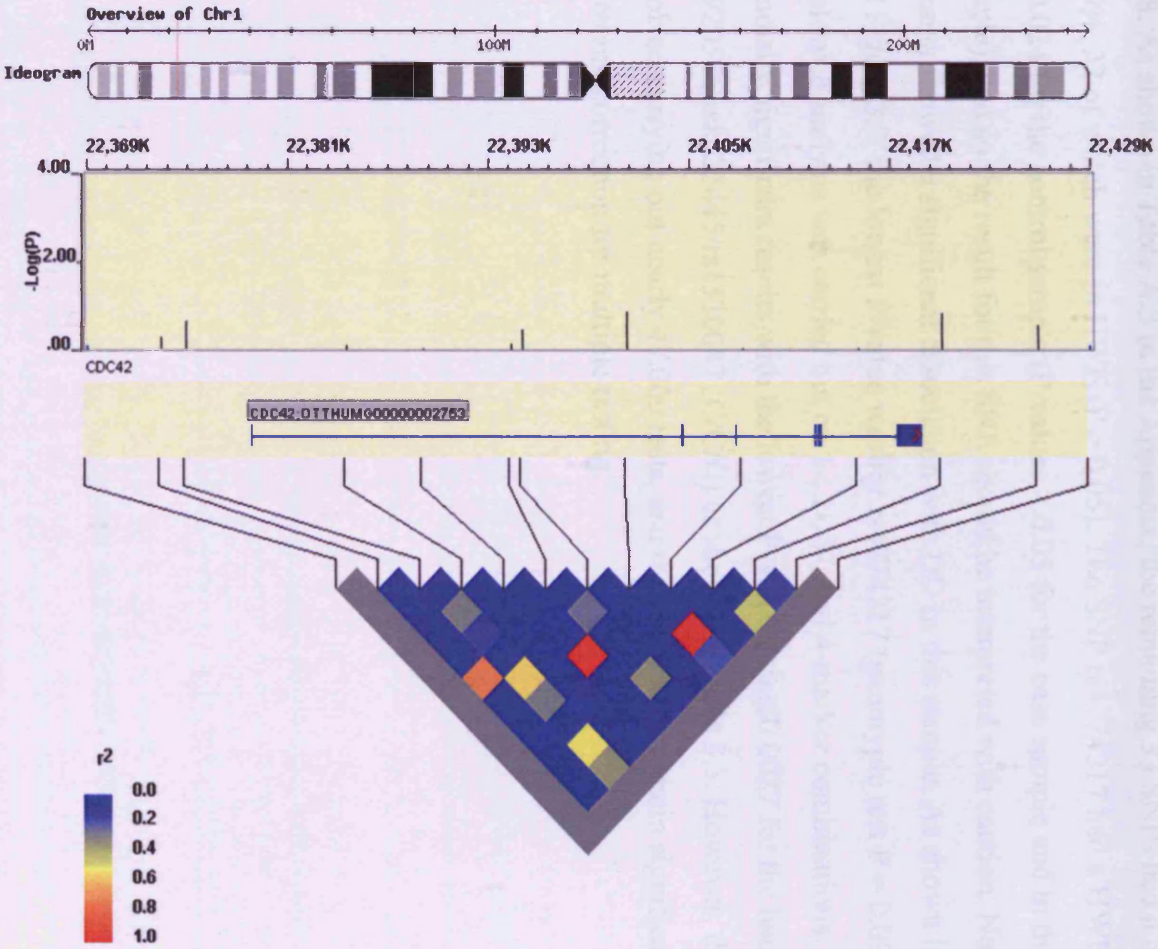


Figure 3.1: LD plot and results of tag SNPs in *CDC42* that were genotyped in the candidate gene study. The black bars show the position of the tag SNPs and the $-\text{Log}_{10}$ of their additive P-value.

3.3.2 *PRTG* Analysis

Out of 69 markers at the *PRTG* locus, 35 SNPs tagging *PRTG* at $r^2 \geq 0.8$ and a MAF > 0.05 in the HapMap CEU population were selected. These were genotyped in the Cardiff case control sample, of which 350 cases and 266 controls passed QC. Two were dropped due to poorly performing assays, and no proxies could be found at an r^2 value ≥ 0.8. As shown in Table A.3 in the Appendix, the remaining 33 SNPs had a call rate >89%, 32 of which were in HWE ($P > 0.05$). The SNP rs4774217 had a HWE P-value of 0.024 for the control sample (P-values > 0.05 for the case sample and in the whole sample), and so the result for this SNP should be interpreted with caution. None of the variants showed a significant association with DD in this sample. As shown in Table 3.2 and Figure 3.2, the lowest P-value was for rs4774217 (genotypic test $P = 0.099$). Haplotype analysis was carried out on 1-, 2-, 3-, and 4-marker combinations. This gave nominally significant results with the lowest P-value being 0.0027 for the haplotype rs9920546/rs8025445/rs1530087 (C/C/G) as shown in Table 3.3. However, this analysis involved carrying out nearly 47,000 tests, and thus no results remain significant following correction for multiple testing.

							P-Values	
SNP	**Position (bp)**	**Minor Allele**	**MAF Cases**	**MAF Control**	**OR**	**95% CI**	**Genotypic**	**Additive**
rs552292	53676324	T	0.330	0.316	1.066	0.84-1.36	0.764	0.609
rs4774217	53680603	A	0.460	0.491	0.881	0.69-1.12	0.099	0.283
rs7175728	53681850	T	0.384	0.344	1.185	0.93-1.50	0.390	0.170
rs617137	53691536	T	0.137	0.138	0.992	0.72-1.37	0.713	0.962
rs1438915	53699820	A	0.088	0.094	0.935	0.63-1.38	0.931	0.733
rs16976432	53703325	C	0.097	0.078	1.258	0.84-1.88	0.498	0.259
rs12591646	53706431	G	0.192	0.182	1.064	0.80-1.42	0.809	0.682
rs7165971	53708305	C	0.294	0.279	1.079	0.84-1.38	0.710	0.550
rs16976436	53711247	T	0.243	0.233	1.055	0.81-1.37	0.512	0.691
rs581287	53711901	T	0.343	0.331	1.055	0.83-1.34	0.703	0.653
rs7164393	53715424	T	0.104	0.099	1.054	0.73-1.53	0.961	0.788
rs1550326	53720491	A	0.435	0.413	1.090	0.86-1.38	0.230	0.475
rs492363	53730834	A	0.241	0.249	0.959	0.74-1.25	0.596	0.755
rs9920246	53732745	A	0.275	0.260	1.082	0.84-1.40	0.734	0.541
rs9920546	53734024	C	0.380	0.369	1.048	0.83-1.33	0.522	0.693
rs17819156	53742397	A	0.056	0.056	1.011	0.62-1.65	0.938	0.966
rs12903822	53745709	T	0.383	0.383	1.003	0.80-1.26	0.876	0.981
rs8025445	53753934	A	0.344	0.331	1.059	0.84-1.34	0.819	0.623
rs8030790	53754689	G	0.241	0.230	1.059	0.81-1.38	0.515	0.663
rs2118781	53755208	C	0.102	0.095	1.090	0.74-1.60	0.575	0.656
rs12373006	53755831	A	0.225	0.218	1.037	0.79-1.36	0.534	0.793
rs16976466	53760089	C	0.105	0.098	1.084	0.75-1.58	0.751	0.665
rs7176699	53761062	C	0.098	0.091	1.078	0.72-1.61	0.880	0.717
rs687128	53771897	A	0.406	0.402	1.017	0.81-1.28	0.108	0.881
rs9920680	53773093	C	0.085	0.086	0.987	0.66-1.48	0.888	0.952
rs1011061	53774790	A	0.499	0.494	1.017	0.81-1.27	0.791	0.882
rs1530087	53777321	A	0.129	0.121	1.080	0.77-1.52	0.837	0.666

Table 3.2: Results of *PRTG* genotyping. MAF – minor allele frequency, OR – odds ratio, CI – confidence interval.

							P-Values	
SNP	Position (bp)	Minor Allele	MAF Cases	MAF Control	OR	95% CI	Genotypic	Additive
rs11858195	53785887	A	0.326	0.318	1.037	0.82-1.32	0.881	0.766
rs2414424	53787355	T	0.149	0.134	1.136	0.82-1.57	0.750	0.457
rs4377101	53804814	C	0.079	0.081	0.978	0.63-1.51	0.958	0.921
rs7163167	53824295	C	0.349	0.339	1.045	0.82-1.33	0.842	0.716
rs2414433	53831481	T	0.179	0.173	1.042	0.78-1.40	0.917	0.789
rs8036481	53834838	C	0.459	0.472	0.948	0.76-1.19	0.597	0.626

Table 3.2 continued

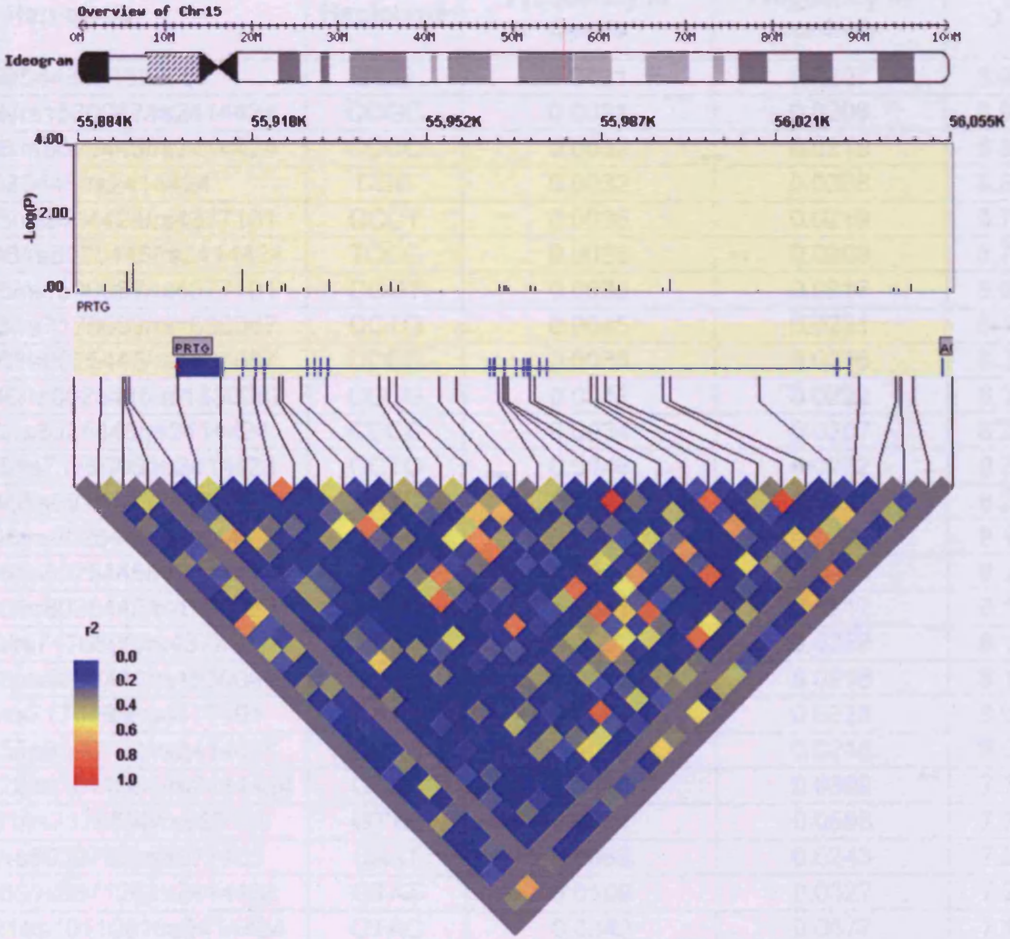


Figure 3.2: LD plot and results of tag SNPs in *PRTG* that were genotyped in the candidate gene study. The black bars show the position of the tag SNPs and the –Log_{10} of their additive P-value.

SNPs in Haplotype	Haplotype	Frequency in Cases	Frequency in controls	χ^2	Individual Haplotype P-value
rs9920546/rs8025445/rs1530087	CCG	0.0031	0.0207	8.97	0.0027
rs9920546/rs8025445/rs1530087/rs2414424	CCGC	0.0031	0.0208	8.96	0.0028
rs1438915/rs9920546/rs8025445/rs2414424	CCCC	0.0032	0.0210	8.84	0.0029
rs9920546/rs8025445/rs2414424	CCC	0.0032	0.0208	8.82	0.0030
rs9920546/rs8025445/rs2414424/rs4377101	CCCT	0.0036	0.0219	8.77	0.0031
rs16976432/rs9920546/rs8025445/rs2414424	TCCC	0.0033	0.0209	8.76	0.0031
rs9920546/rs8025445/rs1530087/rs4377101	CCGT	0.0036	0.0217	8.63	0.0033
rs9920546/rs8025445/rs7176699/rs1530087	CCTG	0.0045	0.0231	8.45	0.0037
rs7164393/rs9920546/rs8025445/rs1530087	CCCG	0.0038	0.0216	8.39	0.0038
rs16976436/rs9920546/rs8025445/rs1530087	CCCG	0.0041	0.0222	8.34	0.0039
rs581287/rs9920546/rs8025445/rs2414424	CCCC	0.0034	0.0207	8.28	0.0040
rs9920546/rs8025445/rs7176699/rs2414424	CCTC	0.0046	0.0232	8.25	0.0041
rs16976432/rs9920546/rs8025445/rs1530087	TCCG	0.0040	0.0217	8.22	0.0042
rs16976436/rs9920546/rs8025445/rs2414424	CCCC	0.0041	0.0221	8.18	0.0042
rs7164393/rs9920546/rs8025445/rs2414424	CCCC	0.0040	0.0219	8.12	0.0044
rs1438915/rs9920546/rs8025445/rs1530087	CCCG	0.0041	0.0217	8.12	0.0044
rs492363/rs8030790/rs7176699/rs4377101	AATT	0.0057	0.0258	8.11	0.0044
rs9920546/rs8025445/rs8030790/rs1530087	CCAG	0.0042	0.0218	8.10	0.0044
rs617137/rs492363/rs7176699/rs4377101	CATT	0.0094	0.0328	8.09	0.0044
rs9920546/rs8025445/rs8030790/rs2414424	CCAC	0.0042	0.0218	8.03	0.0046
rs16976436/rs12903822/rs1011061/rs2414424	CTAC	0.0088	0.0302	7.71	0.0055
rs9920246/rs2118781/rs7176699/rs687128	GTTA	0.0221	0.0508	7.30	0.0069
rs617137/rs492363/rs8030790/rs4377101	CAAT	0.0059	0.0243	7.25	0.0071
rs9920246/rs16976466/rs687128/rs2414433	GTAT	0.0109	0.0327	7.21	0.0073
rs16976436/rs2118781/rs1011061/rs2414424	CTAC	0.0342	0.0677	7.18	0.0074
rs492363/rs9920246/rs8030790/rs4377101	AGAT	0.0076	0.0267	6.83	0.0090
rs8030790/rs2118781/rs687128/rs2414424	ATAC	0.0258	0.0542	6.70	0.0096
rs492363/rs8030790/rs12373006/rs4377101	AAGT	0.0075	0.0261	6.68	0.0098

Table 3.3: Results of the haplotype analysis of *PRTG* SNPs for those haplotypes with P-values < 0.01.

3.3.3 *KIAA0319L* Analysis

Out of 44 markers at the *KIAA0319L* locus, 13 SNPs tagging *PRTG* at $r^2 \geq 0.8$ and a MAF > 0.05 in the HapMap CEU population were selected. These were genotyped in the Cardiff case control sample, of which 355 cases and 267 controls passed QC. Two SNPs were dropped due to poorly performing assays, and no proxies could be found at an r^2 value ≥0.8. The remaining 11 SNPs had a call rate >98% (as shown in Table A.3 in the Appendix). None of the variants showed a significant association with DD in this sample. As shown in Table 3.4 and Figure 3.3, the lowest P-value was for rs1188633 (genotypic test P = 0.065). Haplotype analysis was carried out on all possible haplotype combinations, but none were significant (see Table A.5 in Appendix for a list of the haplotypes with the lowest global P-values).

							P-Values	
SNP	Position (bp)	Minor Allele	MAF Cases	MAF Control	OR	95% CI	Genotypic	Additive
rs12122296	35547202	C	0.110	0.112	0.979	0.69-1.40	0.470	0.908
rs3814302	35673202	C	0.018	0.028	0.644	0.30-1.36	0.503	0.241
rs1188633	35676499	C	0.063	0.086	0.719	0.47-1.10	0.065	0.122
rs1203138	35723292	T	0.121	0.110	1.114	0.78-1.59	0.561	0.536
rs2486297	35734324	G	0.041	0.058	0.691	0.41-1.16	0.071	0.165
rs1203148	35741486	C	0.228	0.238	0.947	0.73-1.24	0.716	0.694
rs1635718	35747056	C	0.020	0.025	0.805	0.38-1.73	0.852	0.572
rs12408030	35783621	G	0.263	0.285	0.894	0.70-1.15	0.490	0.388
rs7523017	35787598	A	0.048	0.062	0.774	0.47-1.27	0.146	0.307
rs6425949	35821996	C	0.018	0.028	0.645	0.30-1.37	0.508	0.244
rs6668196	35841791	G	0.045	0.056	0.791	0.47-1.32	0.349	0.362

Table 3.4: Results of *KIAA0319L* genotyping. MAF – minor allele frequency, OR – odds ratio, CI – confidence interval.

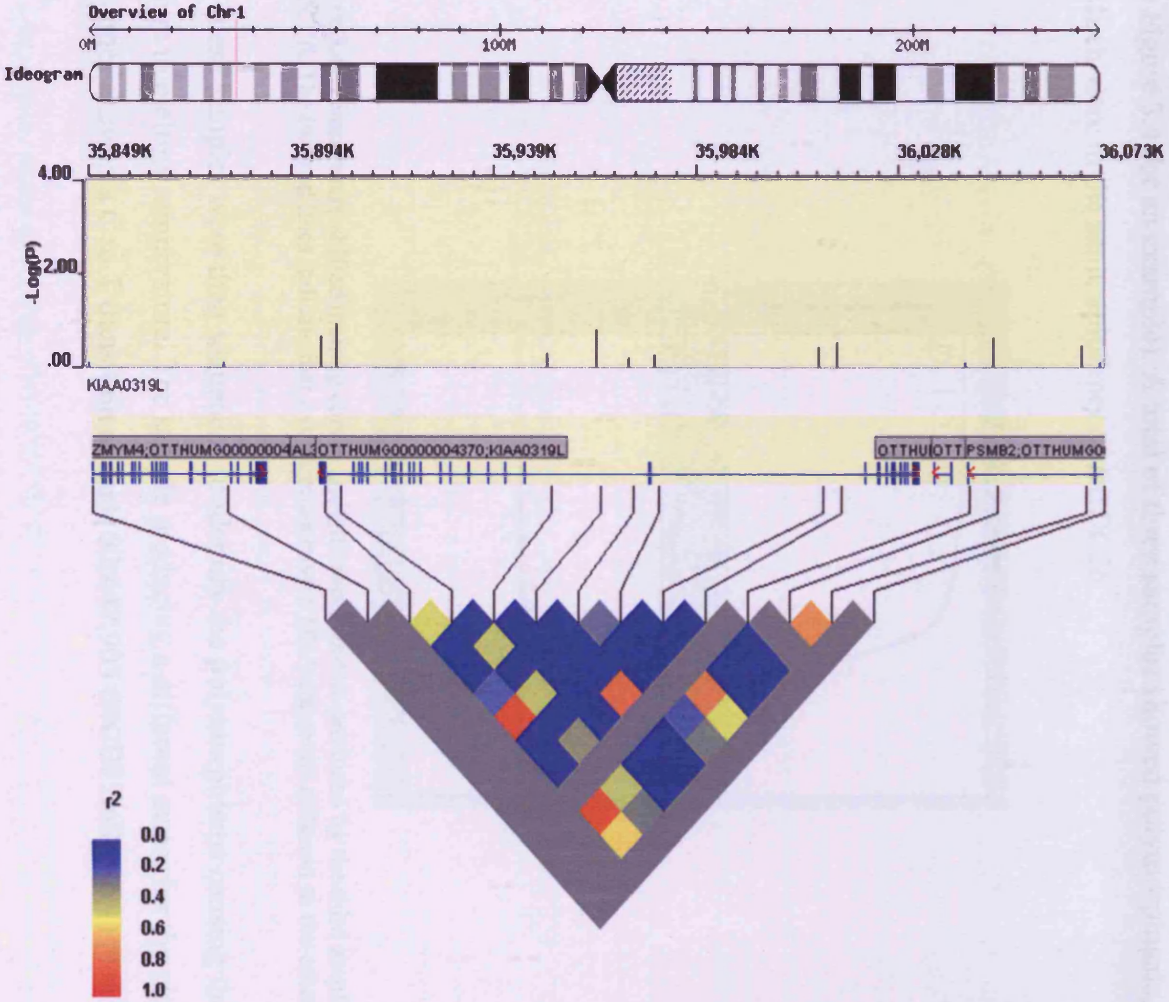


Figure 3.3: LD plot and results of tag SNPs in *KIAA0319L* that were genotyped in the candidate gene study. The black bars show the position of the tag SNPs and the $-\text{Log}_{10}$ of their additive P-value.

3.3.4 *DCDC2b* Analysis

As no HapMap SNPs with a MAF > 0.05 were identified in *DCDC2b,* high resolution DNA melting analysis on the LightScanner was used to detect polymorphisms within the *DCDC2b* region in 15 DD cases. A change in melting curve indicated a different DNA sequence in that particular sample compared to the others (see Figure 3.4 for an example). A total of three samples showed polymorphisms, two of which were in the same amplimer of *DCDC2b.*

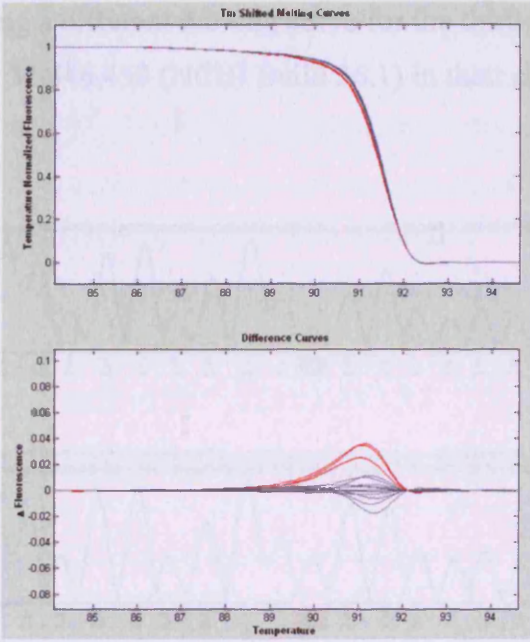


Figure 3.4: Temperature shifted melting curves and difference curves produced by the third amplimer of *DCDC2B.* The two red lines indicate that two samples have a DNA sequence different to the others

These samples were then sequenced to identify the polymorphisms causing the change in melting temperature. The sample producing a different curve for the eighth amplimer showed a C to T transition at chr1: 32,447,903 (NCBI Build 36.1) (see Figure 3.5).

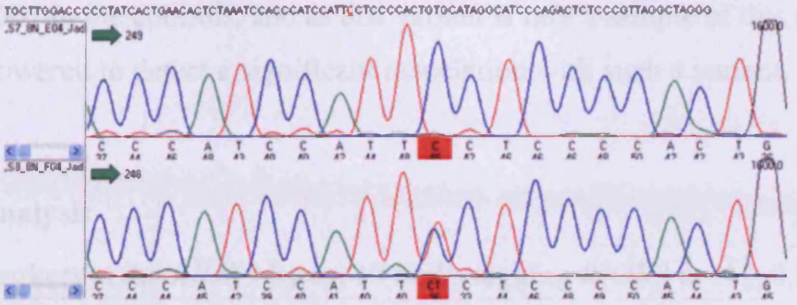


Figure 3.5: Section of the DNA sequence of the eighth amplimer. The bottom trace shows the C to T transition at chr 1: 32,447,903.

Two samples producing a different melting curve for the third amplimer showed a G to C transversion at chr1: 32,446,453 (NCBI Build 36.1) in their sequences (see Figure 3.6).

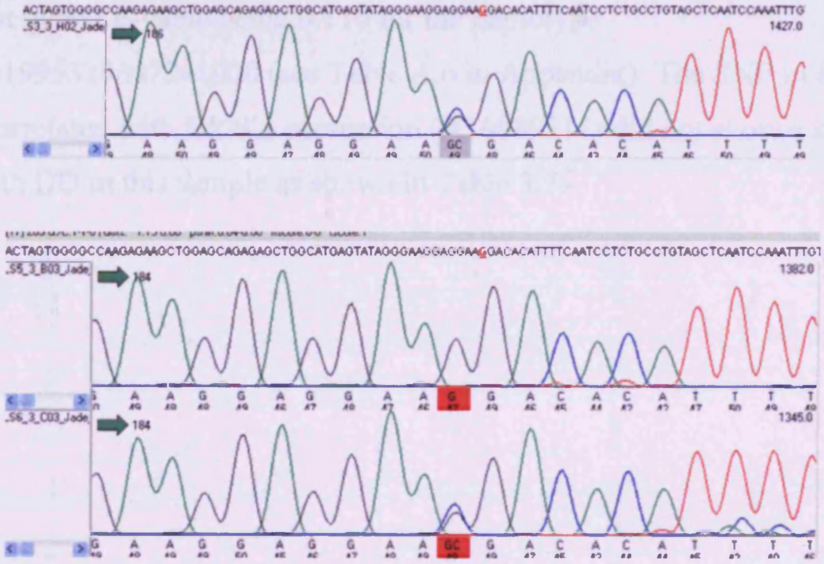


Figure 3.6: Section of the DNA sequence of the third amplimer in two samples. The traces at the top and bottom show the same G to C transversion at chr 1: 32,446,453 in both samples.

The G to C transversion shown in the third amplimer (-829G>C) is situated 829 bp upstream of the *DCDC2b* gene, and the C to T transition (IVS1+355C>T) in the eighth amplimer occurs within intron 1 of the *DCDC2b* gene. Both of these SNPs were genotyped in the case control sample and had a call rate > 99%. 355 cases and 267 controls passed QC. As shown in Table 3.5, none of these variants showed a significant association with DD in this sample, with the lowest P-value being obtained for -829G>C (genotypic test P = 0.214). However, the SNP IVS1+355C>T shows a very low MAF of 0.007 in the controls, and as this variant is rare a sample of this size is insufficiently powered to detect a significant association with such a variant.

3.3.5 *RIOK3* Analysis

Out of 22 markers at the *RIOK3* locus, 10 SNPs tagging *RIOK3* at $r^2 \geq 0.8$ and a MAF > 0.05 in the HapMap CEU population were selected. These were genotyped in the Cardiff case control sample of which 354 cases and 265 controls passed QC. All SNPs had a call rate > 96%. As shown in Table 3.6, only one variant showed nominally significant association (rs11659196, genotypic test P = 0.046), however this association would not remain significant after correction for multiple testing. This SNP is in the 3' UTR of the gene, as shown in Figure 3.7. Haplotype analysis was carried out on all possible haplotype combinations. No haplotypes were significantly associated with DD with the lowest global P-value being 0.110 for the haplotype rs11659196/rs1995329/rs7241000 (see Table A.6 in Appendix). The SNP in *KIAA0319* shown to be correlated with *RIOK3* expression (rs16889511) did not show a significant association with DD in this sample as shown in Table 3.7.

							P-Values	
SNP	Position (bp)	Minor Allele	MAF Cases	MAF Controls	OR	95% CI	Genotypic	Additive
-829G>C	32446253	C	0.055	0.058	0.943	0.58-1.53	0.214	0.816
IVS1+355C>T	32447903	T	0.003	0.007	0.375	0.07-2.05	0.499	0.238

Table 3.5: Results of *DCDC2b* genotyping. MAF – minor allele frequency, OR – odds ratio, CI – confidence interval.

							P-Values	
SNP	Position (bp)	Minor Allele	MAF Cases	MAF Control	OR	95% CI	Genotypic	Additive
rs1995329	19291516	C	0.381	0.351	1.140	0.90-1.44	0.500	0.292
rs8083331	19295223	C	0.086	0.095	0.894	0.60-1.32	0.461	0.559
rs17202653	19296516	T	0.151	0.155	0.964	0.71-1.32	0.211	0.817
rs17187071	19297687	G	0.047	0.032	1.501	0.83-2.77	0.354	0.179
rs2291993	19298605	A	0.323	0.342	0.922	0.73-1.17	0.104	0.513
rs2047683	19303801	C	0.066	0.066	1.011	0.64-1.59	0.503	0.963
rs11663375	19312080	C	0.462	0.485	0.912	0.73-1.14	0.378	0.440
rs2270885	19312965	A	0.087	0.096	0.906	0.61-1.34	0.490	0.604
rs11659196	19316693	T	0.124	0.140	0.875	0.63-1.22	**0.046**	0.430
rs7241000	19318372	T	0.200	0.188	1.077	0.81-1.43	0.383	0.605

Table 3.6: Results of *RIOK3* genotyping. MAF – minor allele frequency, OR – odds ratio, CI – confidence interval.

							P-Values	
SNP	Position (bp)	Minor Allele	MAF Cases	MAF Controls	OR	95% CI	Genotypic	Additive
rs16889511	24714015	G	0.2044	0.2271	0.8745	0.66-1.16	0.631	0.3373

Table 3.7: Results of the SNP in *KIAA0319* that is associated with expression of *RIOK3.* MAF – minor allele frequency, OR – odds ratio, CI – confidence interval.

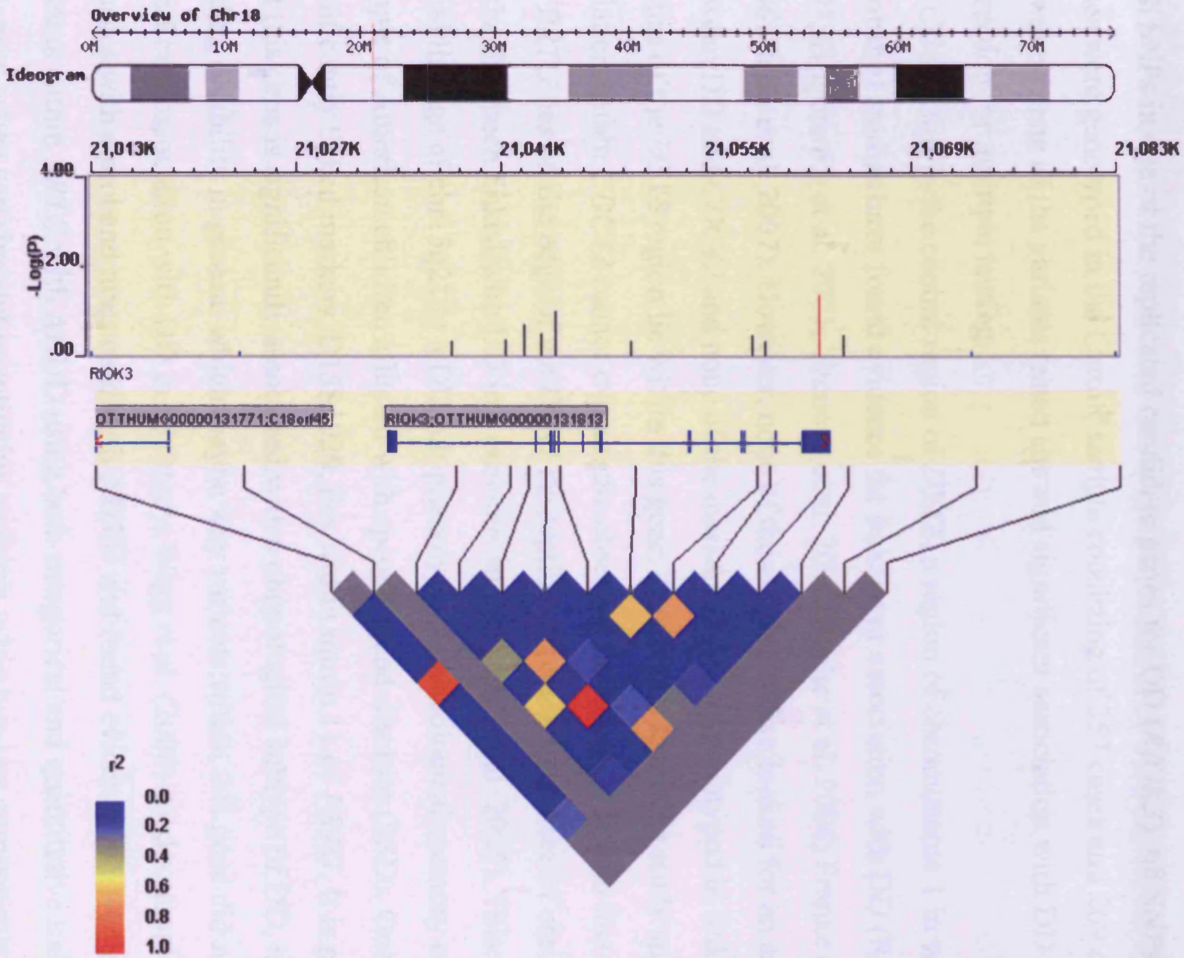


Figure 3.7: LD plot and results of tag SNPs in *RIOK3* that were genotyped in the candidate gene study. The black bars show the positions of the tag SNPs and the –Log_{10} of their genotypic P-value. The red bar shows the significant SNP rs11659196 which had a genotypic P-value of 0.046.

3.4 Discussion

Variants within 5 genes were tested for an association with DD due to their hypothesised roles within neuronal migration (*CDC42* and *PRTG*), their homology with genes that have shown convincing evidence for association with DD previously (*DCDC2b* and *KIAA0319L*) or because their level of expression has been associated with SNPs in one of the replicated candidate genes for DD (*RIOK3*). 68 SNPs in these genes were genotyped in the Cardiff sample consisting of 357 cases and 269 controls. However, none of the variants tested showed significant association with DD after correction for multiple testing.

CDC42 lies in the central region of *DYX8*, a region of chromosome 1 in which a number of studies have found evidence for linkage or association with DD (Rabin et al. 1993; Grigorenko et al. 2001; Tzenova et al. 2004; Bache et al. 2006; Franke et al. 2006; Bates et al. 2007). However, none of these studies have looked for an association between DD and *CDC42* and none of the markers that were genotyped in linkage studies of the *DYX8* region lie within this gene. The results from this study suggest that variation within *CDC42* cannot explain the observed linkage of DD with this region.

PRTG lies on the edge of the *DYX1* susceptibility region in an area of chromosome 15 that has been linked with DD in a previous study (Smith et al. 2005). This study found linkage of chr15q21 to a DD sub-phenotype of phonological memory using a sample of American children affected with speech sound disorder (SSD). One of the significantly linked markers, D15S1029, lies within intron 11 of *PRTG*. It is possible that this gene is significantly associated with a phonological subtype of DD, rather than reading disability in general which maybe why variants within this gene did not show significant association with DD in this study. Wigg et al. (2008) used a sample of 253 families with a proband diagnosed with ADHD and found evidence of association of variants within *PRTG* with ADHD using both categorical and quantitative trait analysis. However, when carrying out quantitative analyses using two key components of reading (word identification and decoding), none of the SNPs genotyped were found to be significantly associated. Out of the 20 SNPs genotyped in that study, 8 SNPs were typed directly in this study, and a further 6 were typed indirectly via proxies. Thus, it appears that although variation within PRTG may be associated with a phonological sub-phenotype of DD in a SSD sample, it does not appear to be associated with reading ability when defined categorically in this study, or through reading components in an ADHD sample.

KIAA0319L lies within the *DYX8* susceptibility locus and has been previously tested for an association with DD by Couto and colleagues (2008). All of the SNPs that made up the significant haplotype identified in that study (rs1203138/rs1203148/rs12408030/rs7523017; C/A/A/A haplotype, P = 0.03; global P = 0.339) were genotyped here, but failed to show a significant association (C/A/A/A haplotype P = 0.3069; global P = 0.266). The SNP showing a significant association by Couto and colleagues (rs7523017; P = 0.042) was also not significant in this study (P = 0.307). The difference in results may be due to the different populations studied. However, 68.1% of the families used in the sample in Couto and colleagues' study described their descendants' ethnicity as being European or British, with another 26% describing them as "Caucasian Canadians" so the populations between the two studies should not differ dramatically. Both studies also used similar tests and inclusion criteria to define subjects as 'affected'. Further studies in other independent samples will need to be carried out to determine if an association between variants within *KIAA0319L* and DD exists or not.

DCDC2b is on chromosome 1p35.1 within the well replicated *DYX8* region. Although this region has shown linkage to DD in a number of studies, no variants within *DCDC2b* have been tested for association with the disorder previously. Two novel variants were identified in *DCDC2b* through high resolution DNA melting analysis but neither of these variants were significantly associated in this sample. The mutation detection sample used in this study had a power of 96% to detect a variant with a MAF > 0.1 and 79% power to detect a variant with a MAF > 0.05. It is therefore possible that variants within *DCDC2b* may exist that are associated with DD but are either too rare and/or do not have large enough effect sizes in order to be identified in this study. The SNP IVS1+355C>T showed a very low MAF of 0.007 in the controls, and as this variant is not highly polymorphic, a sample of this size is unlikely to pick up a significant association with such a rare variant. As there are no SNPs in the HapMap CEU population within this gene that have a MAF $\geq$ 0.05, it is likely that any variants within *DCDC2b* are too rare for association with DD to be detected in small samples.

RIOK3 is on chromosome 18q11.2, just outside of the *DYX6* linkage region. One of the two SNPs within the *KIAA0319* gene that were shown to be correlated with expression of *RIOK3* (Myers et al. 2007) was also genotyped in this study but did not show a significant association with DD (rs16889511, P = 0.3373). The other SNP that showed a correlation with the expression of this gene (rs16889506), has also been genotyped in this sample in a previous study but failed to show a significant association (P = 0.8673) (Harold et al. 2006), which supports the results in this current study as these SNPs are in perfect LD with each other. As the *KIAA0319* SNPs that are correlated with the expression of *RIOK3* are not significantly associated with DD in this sample, it implies that the expression of this gene may not affect an individual's susceptibility to this disorder. Neither of the *KIAA0319* SNPs are present in more recent eQTL databases (Dixon et al. 2007; Dimas et al. 2009). Although the *KIAA0319* SNP was not significantly associated with DD in this study, variation within *RIOK3* may contribute to DD susceptibility, i.e. an association of rs11659196 with DD was observed (genotypic P = 0.046). This result does not survive correction for multiple testing however, and genotyping in a larger, independent sample is required to either confirm or refute the association.

However, no single study of this size can exclude a gene from involvement in disease susceptibility and there could be a number of reasons why no association was found for these genes in this particular study. It is possible that these genes do have a role within DD, but their effect sizes are too small for association to be detected in a sample of this size. For example, for a susceptibility variant with an OR of 1.3 and a MAF of 0.4, this sample had 62% power to detect a significant association ($P < 0.05$).

Another possibility is that while *CDC42* and *PRTG* may have roles within neuronal migration (Luo 2000; Wang et al. 2006), the roles they play in this pathway do not have an effect on reading ability and so are not associated with DD. These genes are far from an exhaustive list of neuronal migration genes that lie within DD susceptibility regions. Neuronal migration is a complex process involving a large network of pathways and it is still a biological process that is worth exploring further within the context of reading ability.

In summary, variants within five candidate genes which lie within the *DD* susceptibility regions were genotyped in the Cardiff case control sample but did not show a significant association with DD. However, further studies in independent samples need to be carried out before these genes can be confidently discounted as DD susceptibility genes.

Chapter 4: NeuroDys Genome-wide Association Study

4.1 Introduction

Genome-wide association studies (GWAS) involve testing a large number of common genetic variants across the human genome for association with a disease or trait. Because no hypotheses are made about the location of the susceptibility variants or their biology, this method provides an unbiased approach to identifying new susceptibility loci for disease (Hirschhorn & Daly 2005), as opposed to candidate gene studies which are often based on an imperfect understanding of biological pathways and can yield associations that are difficult to replicate (Manolio et al. 2009). GWAS have become possible through the development of commercial arrays that can capture most of the known single nucleotide polymorphism (SNP) variation with a minor allele frequency (MAF) greater than 5% in the general population (The International HapMap Consortium 2007). This has been achieved by exploiting the linkage disequilibrium that exists between SNPs in the human genome. For example, the Illumina HumanHap300 array consists of 317,503 SNPs which capture 76% of the SNPs with MAF > 0.05 in Phase II of HapMap at an r^2 threshold > 0.8 (in the Caucasian population) (Mägi et al. 2007). The Illumina HumanHap550 array consists of >550,000 SNPs which capture 86% of the SNPs in Phase II of HapMap (MAF > 0.05, $r^2 > 0.8$) (Mägi et al. 2007).

As discussed in section 1.3.2.4, multiple testing issues in GWAS require stringent P-values in order for an association to be considered to be genome-wide significant. It has been suggested that an appropriate threshold for genome-wide significance is $P \leq 5 \times 10^{-8}$ (Pe'er et al. 2008). In order to be sufficiently powerful enough to identify common genetic variants with small effect sizes at this genome-wide level of significant association, GWAS of complex diseases require thousands of case and control samples in the initial stages in order to be powerful enough to identify common genetic variants with small effect sizes at genome-wide levels of significant association (Risch 2000; Cardon & Bell 2001). This results in high costs that are often beyond the budgets of many research groups. GWAS are most commonly carried out in the form of case-control studies, as it is often more time consuming and expensive to collect large family-based samples (Craddock et al. 2008). In addition, case-control based studies can take advantage of data from large population based samples that have become publicly available by using such samples as controls against cases for a disease of interest. For example, whole-genome genotype data for participants that were recruited as part of the National Child Development Study (otherwise known as the 1958 birth cohort) has been made available. Using such population data in a study means that only the case samples need to be genotyped, thus reducing costs.

Taking a multi-stage approach to the design of a GWAS can also improve the efficiency of the study by reducing the amount of genotyping required without sacrificing too much power (Hirschhorn & Daly 2005). In the first stage, the full set of genome-wide SNPs (i.e. all SNPs on a genome-wide SNP array) is genotyped in a subset of the samples (often referred to as the 'discovery sample'). A P value threshold (e.g. $< 1 \times 10^{-4}$) is used to identify a subset of SNPs with putative associations, which are then re-tested in a larger independent sample (the 'follow-up' or 'replication sample') in the second or sometimes third stages. This allows researchers to distinguish the true-positive associations identified in the first stage from the false-positives which may be identified in the first stage by chance.

In recent years, GWAS have improved our understanding of complex diseases and have resulted in the identification of a large number of novel susceptibility loci in a range of human diseases, including type 1 and type 2 diabetes mellitus, prostate cancer and breast cancer (McCarthy et al. 2008). For example, five separate GWAS of type 2 diabetes (T2D) identified six new gene regions on top of the five that were already known (Frayling 2007). This a remarkable result when taking into account the weak genetic component of this disease compared with many other common diseases - the sibling relative risk is at the most 3 - 4 for T2DM, compared with 15 for type 1 diabetes (WTCCC 2007). In comparison, DD has an estimated sibling relative risk of 4-6 (Ziegler et al. 2005). The success of these studies can be partly attributed to the large sample sizes that were used, with a total of approximately 55,000 cases and controls in the combined discovery and follow up samples across the five studies (Frayling 2007).

As discussed in Chapter 1, a total of 19 independent linkage studies have been carried out in an effort to identify susceptibility variants for DD and these have identified nine DD susceptibility regions (*DYX1* to *DYX9*). Eight of these linkage studies were genome-wide screens (de Kovel et al. 2004; Fagerheim et al. 1999; Fisher et al. 2002; Igo Jr et al. 2006; Kaminen et al. 2003; Marlow et al. 2003; Nopola-Hemmi et al. 2001; Raskind et al. 2005). Another genome-wide linkage screen for general reading and spelling ability has been carried out using samples that were not specifically selected for DD (Bates et al. 2007). However, these linkage studies have only achieved limited success. The limited success of linkage studies in complex diseases in general can be attributed to their low power and resolution for variants with small effect sizes (Hirschhorn & Daly 2005), as discussed in section 1.3.1.2. A powerful GWAS of DD may have more success in identifying the common susceptibility variants underlying this complex disease.

While no GWAS of DD have been published as yet, a GWAS of general reading ability has been performed. Meaburn and colleagues (2008) used samples from the UK that had been recruited as part of the Twins Early Development Study (TEDS) (n = 3043, one member from each twin pair). They selected two subsets from the 3043 individuals based on their reading scores, with one subset in the lower 25% of the reading distribution (n = 755) and the other in the top 25% of the distribution (n = 747). They formed pools with these two subsets and genotyped them on the Affymetrix GeneChip Human Mapping 100K Array Set. They then selected another 4258 individuals from the TEDS cohort and individually genotyped 75 of the top SNPs in a subset of individuals in the lower 10% of the reading distribution (n = 452) and another subset in the higher 10% (n = 452). 9 of these SNPs showed a significant association with reading ability and another 14 showed low versus high allele frequency differences in the predicted direction. Out of these 23 SNPs, 10 showed significant association with reading ability when genotyped in the remaining 3,408 individuals from the second TEDS cohort that had not been selected for reading ability. However, these SNPs were only nominally significant at the 0.05 level, and would not survive correction for multiple testing. Seven SNPs within the *DYX1* and *DYX2* linkage regions were among the 300 most significant SNPs in the first stage of this study of which five showed association with a low reading ability, but they were not selected for individual genotyping. However, this study was carried out using the 100K Affymetrix GeneChip array and testing 100,000 SNPs (~30% coverage of the genome) cannot be considered to be a comprehensive genome-wide scan and many of the susceptibility variants for reading ability may have been missed in these studies.

The NeuroDys consortium (www.neurodys.com) have recently undertaken the first GWAS of DD involving research groups from the UK and Germany in the first stage and then additional groups from Switzerland, Netherlands, Austria, Finland, France and Hungary in the replication stages. The overall aim of the NeuroDys project is to gain a better understanding of DD by investigating the correlations between candidate genes and brain functions that are found to be relevant for learning to read and to spell, such as grapheme-phoneme association. The genotyping and quality control for the NeuroDys GWAS was conducted by the consortium before this PhD project was started, but I was involved in subsequent follow-up studies.

4.1.2 Aims

The aim for this section of the thesis was to use the data from the NeuroDys GWAS to identify new susceptibility variants for DD as well as compare the results with previous findings. The first stage of the study tested over 300,000 SNPs and highlighted a number of interesting variants which were then selected for follow up in an independent, larger sample in an effort to replicate these findings.

4.2 Materials and Methods

4.2.1 First Stage of GWAS

4.2.1.1 Sample Ascertainment and Criteria

The first stage of the GWAS involved 410 cases from the UK (made up of 200 cases from the University of Oxford and 210 cases from Cardiff University) and 200 cases from Germany (see Table 4.1). The criteria for determining affection status differed slightly between centres. The Cardiff case criteria is an IQ ≥ 85 and a reading age that is ≥ 2.5 years below their chronological age (see Chapter 2). Samples from Oxford were identified from the dyslexia clinic at the Royal Berkshire Hospital in Reading. Participants were classed as cases if they had an IQ ≥ 90 and discrepancy of > 1.5 SD between the average score obtained from British Abilities Scales (BAS) II matrices and similarities subtests (a test for IQ) compared with BAS II Word Reading Test, provided reading age was no higher than chronological age (Marlow et al. 2001; Fisher et al. 2002). German cases were recruited from the Departments of Child and Adolescent Psychiatry and Psychotherapy at the Universities of Marburg and Würzburg. Participants were classed as cases if they had observed spelling scores that were ≥ 1 SD below that predicted based on an assumed correlation between IQ and spelling of 0.40 (Schulte-Körne et al. 2001; Schumacher et al. 2006a).

DNA for these case samples were extracted from either blood or saliva samples using phenol/chloroform methodology as previously described (see Chapter 2). DNA quantification and dilution was also as described (Chapter 2), with a final sample dilution of 50ng/μl.

The UK control sample consisted of 1437 population controls recruited as part of the 1958 British Birth Cohort (Power & Elliott 2006) (see Chapter 2 for more information on this sample) and the German control sample consisted of 905 samples from the Heinz Nixdorf Recall Study (an epidemiological study with a focus on risk factors for coronary heart disease (Stang et al. 2005; Kröger et al. 2006)) and the Munich Antidepressant Response Signature Project (Hennings et al. 2009).

4.2.1.2 Genotyping and Analysis

All German cases and 102 of the Cardiff cases were genotyped on the Illumina HumanHap300 array according to manufacturer's instructions at the University of

Bonn. The Oxford cases and the rest of the Cardiff case sample were genotyped at Oxford University. The Oxford cases and 32 of the Cardiff cases were genotyped on the Illumina HumanHap550 array, with the remaining Cardiff cases (76) genotyped on the Illumina HumanHap300 array. The population controls had previously been genotyped on the Illumina HumanHap550 arrays.

QC filtering was performed by Andrew Morris from Oxford University and Bertram Müller-Myhsok from the Max Planck Institute of Psychiatry in Munich. Individuals were excluded if they had a SNP call rate < 98%, if their autosomal heterozygosity was < 33.5 % or > 36.5%, or if they showed non-European ancestry. Potential duplicate individuals were removed by calculating identity by state (IBS) distances for all possible pairs of individuals in PLINK v1.04 (Purcell et al. 2007) using those SNPs that passed QC filters (see below), and removing one of each pair with an IBS distance > 98%. After QC filters were applied, a total of 585 cases and 2326 controls remained, as shown in Table 4.1.

	Cases	Controls
UK Sample	389	1421
German Sample	196	905
Total	585	2326

Table 4.1: Sample genotyped in the first stage of the GWAS

As samples were genotyped on both the Illumina HumanHap300 and HumanHap550 arrays, only those autosomal SNPs that were common to both arrays were used in the analysis. Additional SNPs were excluded if their call rate was < 98%, if their minor allele frequency was < 0.05 and if their Hardy-Weinberg P-value was < 1 x 10^{-5} in cases or controls. 297,650 SNPs passed these QC filters.

SNPs were tested for an association with DD using logistic regression carried out in PLINK v1.05 (Purcell et al. 2007). To correct for possible population stratification, the genome-wide average IBS distance was calculated in PLINK between each pair of individuals in the resulting dataset using those SNPs that passed the QC filters mentioned above. Multidimensional scaling (MDS) analysis was then performed on the resulting matrix of IBS distances to extract four components. MDS is a method of analysis that provides a visual representation of the pattern of similarities between datasets. For example, given a matrix of IBS distances between various individuals, those individuals that are perceived to be similar are plotted close together, whereas those who are perceived to be different are plotted far apart from each other.

Components can then be extracted which are able to explain the genetic variation occurring between individuals, with the first component explaining most of the variation and the remaining components explaining the rest of the variation. The components can then be used as covariates in order to control for differences between individuals that may be the result of population stratification. The impact of including the components as covariates was evaluated by calculating the genomic control inflation factor (λ). Including the first component as a covariate with centre of origin (shown in Table 4.3) had the maximum impact on λ as shown in Table 4.2 and Figure 4.1. Therefore, these covariates were used when carrying out logistic regression. Both the additive and genotypic tests of association were performed within the logistic regression framework.

Component with Country of Origin	λ
0	1.042
1	1.041
2	1.042
3	1.044
4	1.041

Table 4.2: Table shows the effect of varying the number of components extracted from the MDS analysis, on the genomic control inflation factor (λ). These values are based on analysis of 297,650 SNPs that passed QC filters.

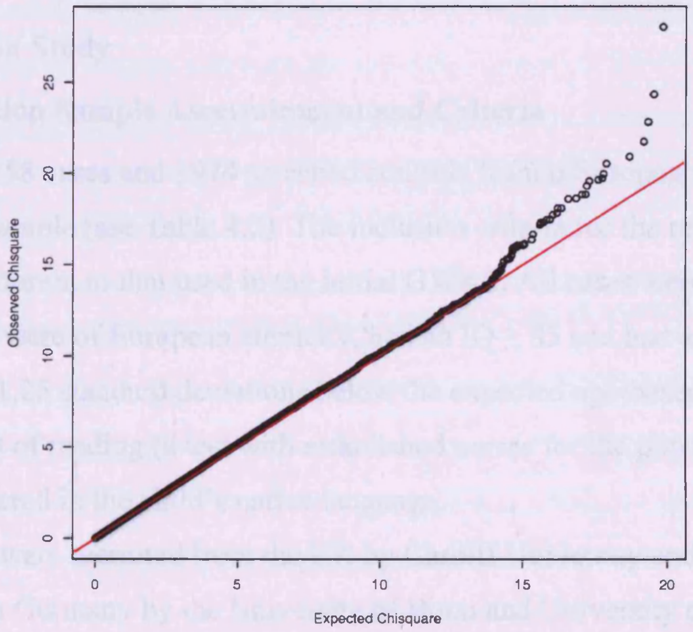


Figure 4.1: Quantile-quantile (QQ) plot of 297,650 observed genome wide association χ^2 test statistics (y-axis) against those expected under the null expectation (x-axis) using country of origin and MDS component 1. The line of equality is coloured red (λ = 1.041).

Set-based analysis was also carried out on SNPs within the DD linkage regions, *DYX1-DYX8*, that had been genotyped in the GWAS. Set-based analysis may offer a number of possible advantages over single locus tests (Neale & Sham 2004). For example, if there is more than one independent association signal within a gene or set of markers, such as where there is more than one functional variant, combining these into a single statistic might offer enhanced power over single SNP analysis (Moskvina et al. 2009). All SNPs located within the regions *DYX1* to *DYX8* that had been genotyped in the GWAS were identified, and logistic regression using the additive model was carried out as before. Two region-wide tests were then performed using PLINK v1.05 (Purcell et al. 2007). The first was based on the most significant single P-value within each of the *DYX* regions, correcting this value based on the number of independent SNPs within the region. The second analysis was based on the product of the P-values within each region. The significance in both tests was obtained by comparing the test statistic in the observed data to that obtained when disease status was randomly permuted among individuals, thereby accounting for inter-SNP LD. For each permutation, the smallest P-value and the product of P-values were obtained as performed in the original data set. The final empirical P-value was determined by the number of times the permuted P-value exceeded the original value. 1000 permutations were performed.

4.2.2 Replication Study

4.2.2.1 Replication Sample Ascertainment and Criteria

A total of 1258 cases and 1974 screened controls from 6 European countries formed the replication sample (see Table 4.3). The inclusion criteria for the replication sample were slightly different to that used in the initial GWAS. All cases were between 8 and 12 years of age, were of European ethnicity, had an IQ ≥ 85 and had a performance level of at least 1.25 standard deviations below the expected age-based norms on a standardised test of reading (a test with established norms for the population being tested) administered in the child's native language.

Participants were recruited from the UK by Cardiff University and Oxford University, from Germany by the University of Bonn and University of Munich, from Switzerland by the University of Zürich, from the Netherlands by the University of

Maastricht, from Austria by the University of Salzburg and from Finland by the University of Jyvăskylă.

DNA was extracted from saliva or blood samples using phenol/chloroform methodology as mentioned previously. It was then quantified using PicoGreen (see Chapter 2) and diluted to 5ng/μl.

Country/Centre	Cases	Controls	Centre
UK:			
Cardiff	209	268	1
Oxford	328	288	2
Germany:			
Bonn	200	685	3
Munich	108	194	4
Switzerland	26	43	5
Netherlands	115	106	6
Austria	116	201	7
Finland	156	189	8
Total	**1258**	**1974**	

Table 4.3: Replication sample

4.2.2.2 Genotyping of Replication Panel and Analysis

SNPs were chosen based on their minimum P-value from the additive and genotypic association tests in the GWAS. The top 65 hits were put through the Sequenom MassARRAY Assay Design 3.1 software in order to design a multiplex panel of SNPs that contained most of the top hits from the GWAS. A panel of 29 SNPs was designed which included 17 SNPs from the top 25 hits, and another 12 that were in the top 65 hits (P-min $< 1 \times 10^{-4}$). Table 4.8 shows all SNPs entered into this panel (see Table B.2 in Appendix for primer sequences).

This panel of SNPs was genotyped in the replication sample using the Sequenom MassARRAY iPlex GOLD system as described in Chapter 2. Genotype calling was carried out using the Typer 3.4 software. All SNP assays were initially optimised by genotyping DNA from 30 CEPH parent-offspring trios. Cluster plots for all SNPs were inspected manually, and SNP assays that did not produce distinct clusters were excluded. 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 rates. 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.

After genotyping, SNPs were tested for Hardy-Weinberg equilibrium in controls and their MAFs were calculated using PLINK v1.05 (Purcell et al. 2007). Samples with a call rate < 70% were excluded.

SNPs were tested for an association with DD using logistic regression carried out in PLINK v1.05 (Purcell et al. 2007) using both the additive and genotypic association tests. To correct for possible population stratification, a covariate was applied as shown in Table 4.3. It was not possible to carry out MDS analysis at this stage because genome-wide SNP data was not available for the replication sample. Additional logistic regression analysis was carried out combining the genotypes for the SNPs in both the replication and initial GWAS samples. Due to the change in ascertainment criteria and to control for artefacts that may be present as a result of the samples being collected and extracted at different times and being genotyped in different centres, the UK cases and controls in the GWAS sample were assigned a different covariate to the UK cases and controls in the replication sample. The German cases and controls in each sample set were also analysed in this way. The samples from the other countries were all assigned a covariate according to their centre as these were not included in the initial GWAS.

4.3 Results

4.3.1 First Stage of GWAS

A total of 585 cases and 2326 controls from the UK (389 cases, 1421 controls) and Germany (196 cases, 905 controls) were included in further analyses (see Table 4.1) following sample QC. A total of 297,650 SNPs passed QC. The results from the logistic regression analyses are shown in Figure 4.2 for the additive test and Figure 4.3 for the genotypic test (see Table B.1 in Appendix for a list of the 200 most significant SNPs). No SNPs achieved genome-wide significance ($P < 5 \times 10^{-8}$). Alternatively, the Bonferroni corrected P-value based on the number of SNPs analysed in this study is $P < 1.7 \times 10^{-7}$ and the top hit had a P-value that was just below this threshold (rs10513829, additive $P = 1.2 \times 10^{-7}$, OR = 0.68), as shown in Table 4.4. Another 35 SNPs showed a suggestive level of significance ($P < 5 \times 10^{-5}$) based on their minimum P-value when using either the additive or genotypic test.

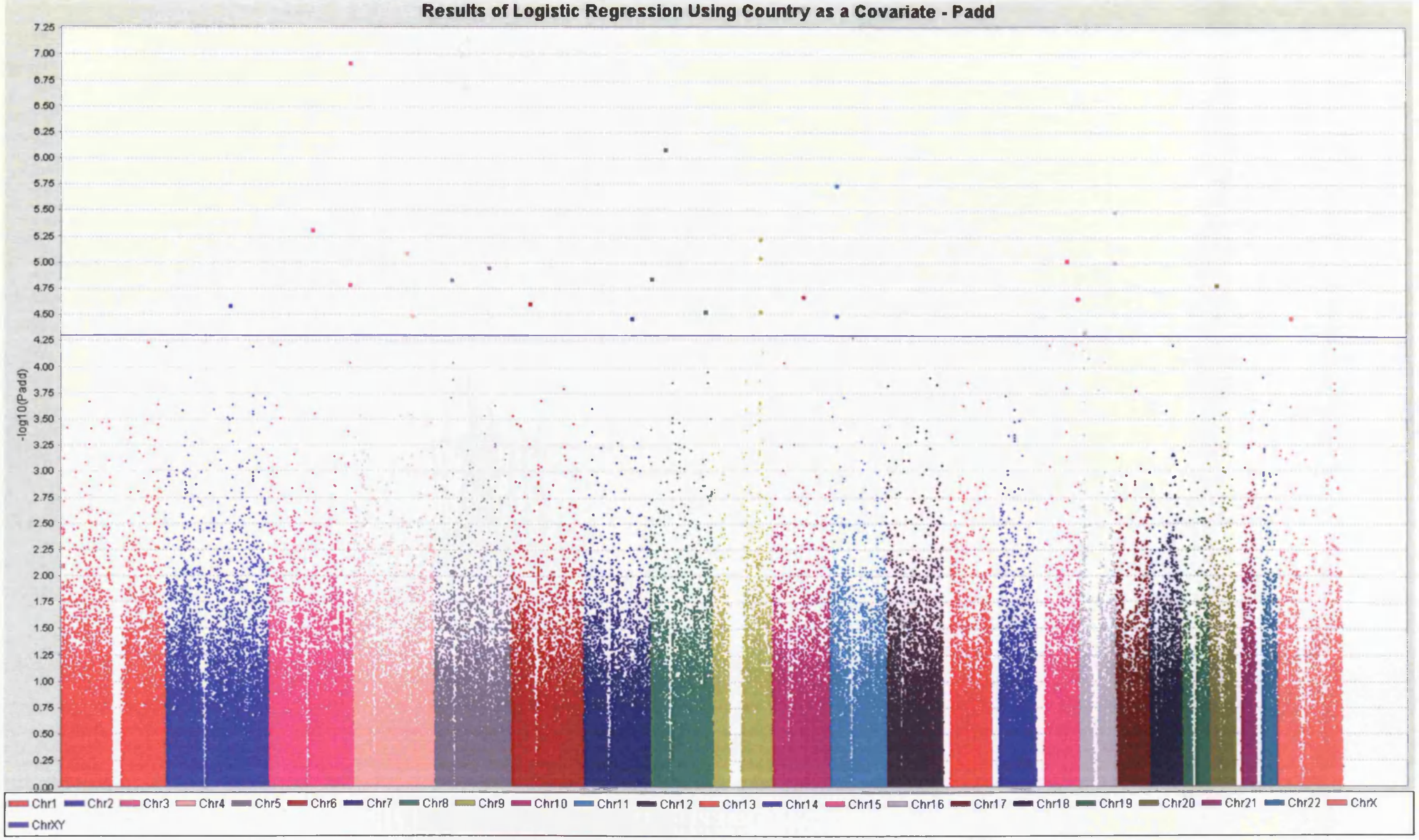


Figure 4.2: Manhattan plot of the additive P-values from the logistic regression analysis on all samples using country and MDS component 1 as a covariates. The blue line indicates a P-value of 5 x 10^{-5}.

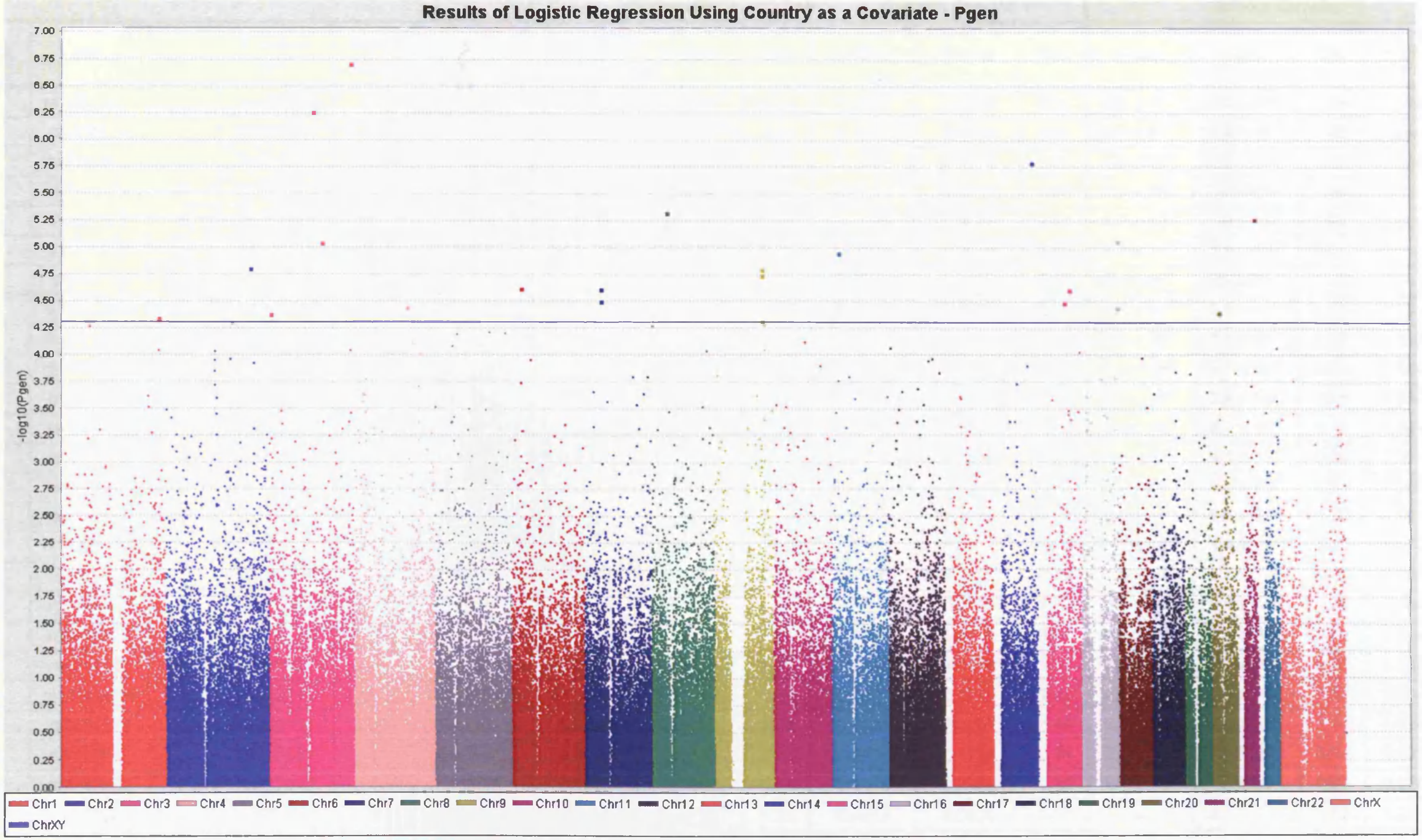


Figure 4.3: Manhattan plot of the genotypic P-values from the logistic regression analysis on all samples using country and MDS component 1 as a covariates. The blue line indicates a P-value of 5 x 10^{-5}

						UK Sample			German Sample			Whole Sample			
Rank	**SNP**	**Chr**	**Position (BP)**	**Minor Allele**	**Closest RefSeq Gene**	**P-Gen**	**P-Add**	**OR**	**P-Gen**	**P-Add**	**OR**	**P-Gen**	**P-Add**	**OR**	**95% CI**
1	rs10513829	3	189971026	C	*LPP*	**3.9×10^{-6}**	**4.5×10^{-6}**	0.66	**0.0241**	**0.0063**	0.71	**2.0×10^{-7}**	**1.2×10^{-7}**	0.68	0.59-0.79
2	rs6796074	3	103316739	T	*LOC152225*	**5.7×10^{-9}**	**2.8×10^{-9}**	1.74	0.5842	0.9352	0.99	**5.6×10^{-7}**	**4.8×10^{-6}**	1.45	1.24-1.71
3	rs7840675	8	34075413	C	*DUSP26*	**1.3×10^{-6}**	**6.7×10^{-7}**	1.86	0.1212	0.1334	1.33	**4.8×10^{-6}**	**8.1×10^{-7}**	1.67	1.36-2.04
4	rs3742673	14	89805459	T	*PSMC1*	**0.0070**	0.9254	0.99	**3.5×10^{-5}**	0.9512	0.99	**1.6×10^{-6}**	**0.9597**	1.00	0.88-1.13
5	rs747783	11	15670131	T	*SOX6*	**7.1×10^{-4}**	**1.8×10^{-4}**	1.39	**0.0095**	**0.0027**	1.44	**1.1×10^{-6}**	**1.8×10^{-6}**	1.41	1.22-1.62
6	rs11117425	16	84529771	T	*IRF8*	**9.3×10^{-4}**	**3.1×10^{-4}**	0.72	**0.0101**	**0.0036**	0.69	**8.8×10^{-6}**	**3.1×10^{-6}**	0.71	0.61-0.82
7	rs2836341	21	38656626	A	*ERG*	**5.2×10^{-5}**	0.2107	1.11	0.0526	0.8496	0.98	**5.4×10^{-6}**	**0.3474**	1.06	0.94-1.21
8	rs10123957	9	110900238	C	*C9orf5*	**7.4×10^{-4}**	**1.9×10^{-4}**	1.36	**0.0134**	**0.0071**	1.35	**1.6×10^{-5}**	**5.9×10^{-6}**	1.35	1.19-1.54
9	rs10518444	4	125945653	G	*ANKRD50*	**5.8×10^{-6}**	**9.0×10^{-7}**	1.97	0.4366	0.4280	1.18	**3.7×10^{-5}**	**7.9×10^{-6}**	1.67	1.33-2.08
10	rs10816767	9	110822490	A	*C9orf5*	**0.0015**	**5.9×10^{-4}**	1.33	**0.0088**	**0.0034**	1.39	**1.9×10^{-5}**	**8.9×10^{-6}**	1.34	1.18-1.53
11	rs4678029	3	123391123	C	*CASR*	**7.3×10^{-4}**	**0.0202**	1.25	**0.0117**	0.0757	1.25	**9.1×10^{-6}**	**3.4×10^{-3}**	1.25	1.08-1.45
12	rs4887111	15	71815337	G	*LOC388135*	**7.3×10^{-4}**	**6.7×10^{-4}**	0.75	**0.0115**	**0.0027**	0.70	**2.5×10^{-5}**	**9.3×10^{-6}**	0.73	0.64-0.84
13	rs7202472	16	84535002	T	*IRF8*	**0.0050**	**0.0037**	0.74	**0.0036**	**5.7×10^{-4}**	0.59	**3.6×10^{-5}**	**9.7×10^{-6}**	0.68	0.58-0.81
14	rs1181841	5	128580604	G	*ISOC1*	**0.0025**	**6.4×10^{-4}**	1.32	**0.0325**	**0.0089**	1.35	**6.2×10^{-5}**	**1.1×10^{-5}**	1.34	1.18-1.53
15	rs4327894	8	1740903	T	*MIRN596*	**0.0121**	**0.0038**	0.75	**0.0033**	**7.8×10^{-4}**	0.59	**5.4×10^{-5}**	**1.4×10^{-5}**	0.70	0.59-0.82
16	rs10512712	5	39728088	C	*DAB2*	**0.0017**	**3.9×10^{-4}**	1.34	**0.0489**	**0.0214**	1.29	**8.2×10^{-5}**	**1.4×10^{-5}**	1.33	1.17-1.52
17	rs1429411	2	197852246	C	*ANKRD44*	**3.4×10^{-4}**	**0.0075**	0.80	**0.0358**	**0.0448**	0.79	**1.6×10^{-5}**	**7.8×10^{-4}**	0.80	0.70-0.91
18	rs7623540	3	189972233	C	*LPP*	**0.0010**	**2.4×10^{-4}**	1.38	**0.0412**	**0.0195**	1.33	**9.0×10^{-5}**	**1.6×10^{-5}**	1.36	1.18-1.56
19	rs6136213	20	17798339	G	*SNX5*	**4.8×10^{-4}**	**1.0×10^{-4}**	0.69	0.0530	**0.0489**	0.77	**4.1×10^{-5}**	**1.6×10^{-5}**	0.72	0.62-0.83
20	rs4747165	10	72969307	G	*CDH23*	**0.0264**	**0.0108**	1.22	**7.2×10^{-4}**	**1.4×10^{-4}**	1.53	**7.7×10^{-5}**	**2.1×10^{-5}**	1.32	1.16-1.49
21	rs11855844	15	96462744	A	*LOC728459*	**2.9×10^{-5}**	**1.3×10^{-6}**	1.64	0.2937	0.2019	1.22	**9.4×10^{-5}**	**2.1×10^{-5}**	1.47	1.23-1.76
22	rs2894536	6	43909855	C	*VEGFA*	**0.0026**	**6.1×10^{-4}**	1.43	**0.0296**	**0.0091**	1.49	**1.1×10^{-4}**	**2.4×10^{-5}**	1.44	1.22-1.70
23	rs9465637	6	20222086	T	*MBOAT1*	**0.0028**	**0.0331**	0.83	**0.0079**	0.1720	0.85	**2.4×10^{-5}**	**0.0125**	0.84	0.73-0.96
24	rs2727822	7	36627947	T	*AOAH*	**0.0072**	0.7097	0.97	**0.0019**	0.7006	1.05	**2.5×10^{-5}**	**0.9239**	0.99	0.87-1.14
25	rs1465234	2	151493615	A	*RBM43*	**6.0×10^{-4}**	**1.2×10^{-4}**	1.75	0.0777	0.0659	1.47	**1.1×10^{-4}**	**2.5×10^{-5}**	1.65	1.31-2.09

Table 4.4: Top 25 hits from the GWAS according to their minimum P-value. P values < 0.05 are in bold. Chr – Chromosome; UTR – Untranslated Region; P-Gen – Genotypic P-value; P-Add – Additive P-value; OR - Odds Ratio with respect to the minor allele; CI - confidence interval.

Table 4.4 shows the top 25 hits from the first stage of the GWAS, based on their minimum P-value. Two of these top hits are in the gene LIM domain containing preferred translocation partner in lipoma (*LPP*) on chromosome 3, as shown in Figure 4.4. Both of these SNPs lie close together within intron 8 of the gene and are in high LD with each other, although the r^2 between them is low ($r^2 = 0.2$, $D' = 0.92$). While these SNPs were significantly associated in the German sample set alone (rs10513829 P-add = 0.0063, OR = 0.71; rs7623540 P-add = 0.0195, OR = 1.33), they showed much stronger association in the UK sample set (rs10513829 P-add = 3.9 x 10^{-6}, OR = 0.66; rs7623540 P-min = 2.4 x 10^{-4}, OR = 1.38). Both of these SNPs had slightly higher effect sizes in the UK sample. The UK sample was also larger than the German sample, therefore it would have had greater power to detect a significant association with these SNPs. Four other SNPs within this gene also had P-values < 0.05 in the combined GWAS sample, indicated by the red bars in Figure 4.4 and shown in Table 4.5. None of the significant SNPs in this gene are in a high level of LD with each other, with the highest level of LD existing between the two top SNPs, rs10513829 and rs7623540.

Two other SNPs in the top hits shown in Table 4.4 are downstream of the gene interferon regulatory factor 8 (*IRF8)* as shown in Figure 4.5. These SNPs and are in LD with each other ($D' = 1$), although the r^2 between them is low ($r^2 = 0.57$). The SNP rs11117425 (P-min = 8.8 x 10^{-6}) lies ~16 Kb downstream of *IRF8* and rs7202472 (P-min = 9.7 x 10^{-6}) lies ~21 Kb downstream and both SNPs are outside of the LD blocks of *IRF8*. The SNP rs11117425 showed a higher level of significance in the UK sample (P-add = 3.1 x 10^{-4}, OR = 0.72) compared with the German sample (P-add = 0.0036, OR = 0.69), while the opposite is true of rs7202472 (UK sample: P-add 0.0037, OR = 0.74; German sample: P-add = 5.7 x 10^{-4}, OR = 0.59). Both of these SNPs had larger effect sizes in the German sample. The reason that rs1117425 showed a higher level of significance in the UK sample even though it had a smaller effect size is likely to be due to the greater amount of power that the larger UK sample had. Three other SNPs in this region had P-values < 0.05 (see Table 4.6), and these were in LD with other ($D' > 0.9$) although the correlation between them all was low ($r^2 < 0.57$).

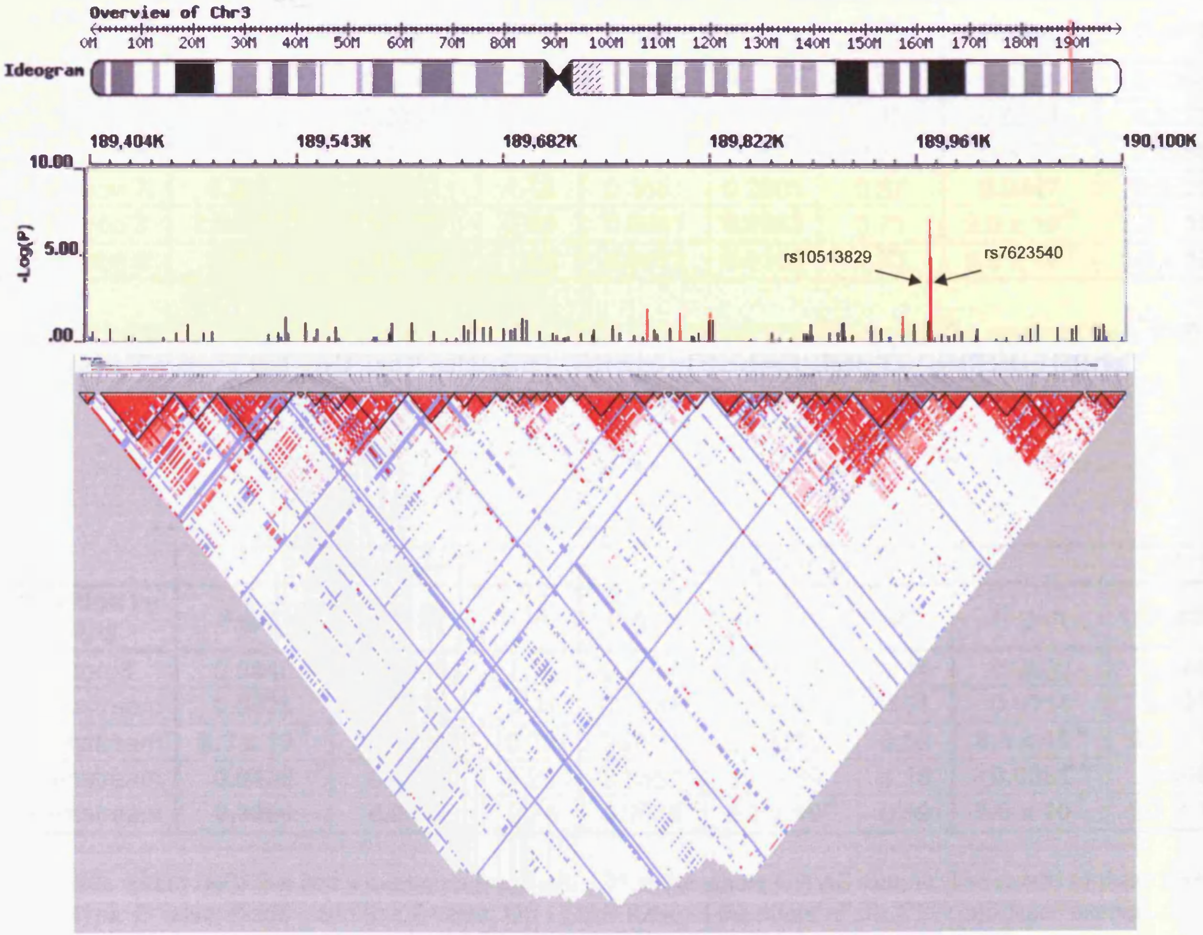


Figure 4.4: LD plot and results of SNPs in *LPP* that were genotyped in the GWAS. Red bars denote P-values <0.05. The two most significant SNPs are not in high LD with each other.

			UK Sample			German Sample			All Samples			
SNP	Position (bp)	Position in *LPP*	P-gen	P-add	OR	P-gen	P-add	OR	P-gen	P-add	OR	95% CI
rs10937357	189780599	Intron 5	**0.0225**	0.1113	0.82	0.5275	0.6901	0.93	**0.0165**	0.1288	0.86	0.71-1.04
rs9837401	189802757	Intron 5	0.0770	**0.0351**	1.20	0.6373	0.3544	1.12	0.0874	**0.0273**	1.17	1.02-1.34
rs4322991	189822630	Intron 6	0.3141	0.1411	0.87	0.0769	0.0679	0.77	0.0671	**0.0268**	0.84	0.71-0.98
rs13314127	189952657	Intron 7	**0.0217**	0.1597	1.14	0.3851	0.2601	0.87	**0.0447**	0.6628	1.03	0.89-1.19
rs10513829	189971026	Intron 8	**3.9×10^{-6}**	**4.5×10^{-6}**	0.66	**0.0241**	**0.0063**	0.71	**2.0×10^{-6}**	**1.2×10^{-6}**	0.68	0.59-0.79
rs7623540	189972233	Intron 8	**0.0045**	**2.4×10^{-4}**	1.38	**0.0412**	**0.0195**	1.33	**9.0×10^{-5}**	**1.6×10^{-5}**	1.36	1.18-1.56

Table 4.5: Table of results from all SNPs within *LPP* that had a minimum P-value < 0.05 in the whole GWAS sample. The results of these SNPs in the UK and German samples are also shown. P-gen – genotypic P-value; P-add – additive P-value; OR - odds ratio of the minor allele; CI - confidence interval.

			UK Sample			German Sample			All Samples			
SNP	Position (bp)	Position in *IRF8*	P-gen	P-add	OR	P-gen	P-add	OR	P-gen	P-add	OR	95% CI
rs391525	84501939	Intron 3	0.0840	0.0842	1.16	**0.0327**	**0.0103**	1.44	0.0630	**0.0401**	1.15	1.01-1.32
rs366078	84522063	Downstream	**0.0435**	0.0618	0.81	0.2541	**0.0041**	0.61	0.0014	**0.0012**	0.73	0.61-0.88
rs11117425	84529771	Downstream	**9.3×10^{-4}**	**3.1×10^{-4}**	0.72	**0.0101**	**0.0036**	0.69	**8.8×10^{-6}**	**3.1×10^{-6}**	0.71	0.61-0.82
rs305061	84533159	Downstream	**0.0438**	**0.0211**	1.21	0.3387	0.1489	1.18	**0.0051**	**0.0060**	1.21	1.06-1.38
rs7202472	84535002	Downstream	**0.0050**	**0.0037**	0.74	**0.0036**	**5.7×10^{-4}**	0.59	**3.6×10^{-5}**	**9.7×10^{-6}**	0.68	0.58-0.81

Table 4.6: Table of results from all SNPs within *IRF8* that had a minimum P-value < 0.05 in the whole GWAS sample. The results of these SNPs in the UK and German samples are also shown. P-gen – genotypic P-value; P-add – additive P-value; OR - Odds Ratio of the minor allele; CI - confidence interval.

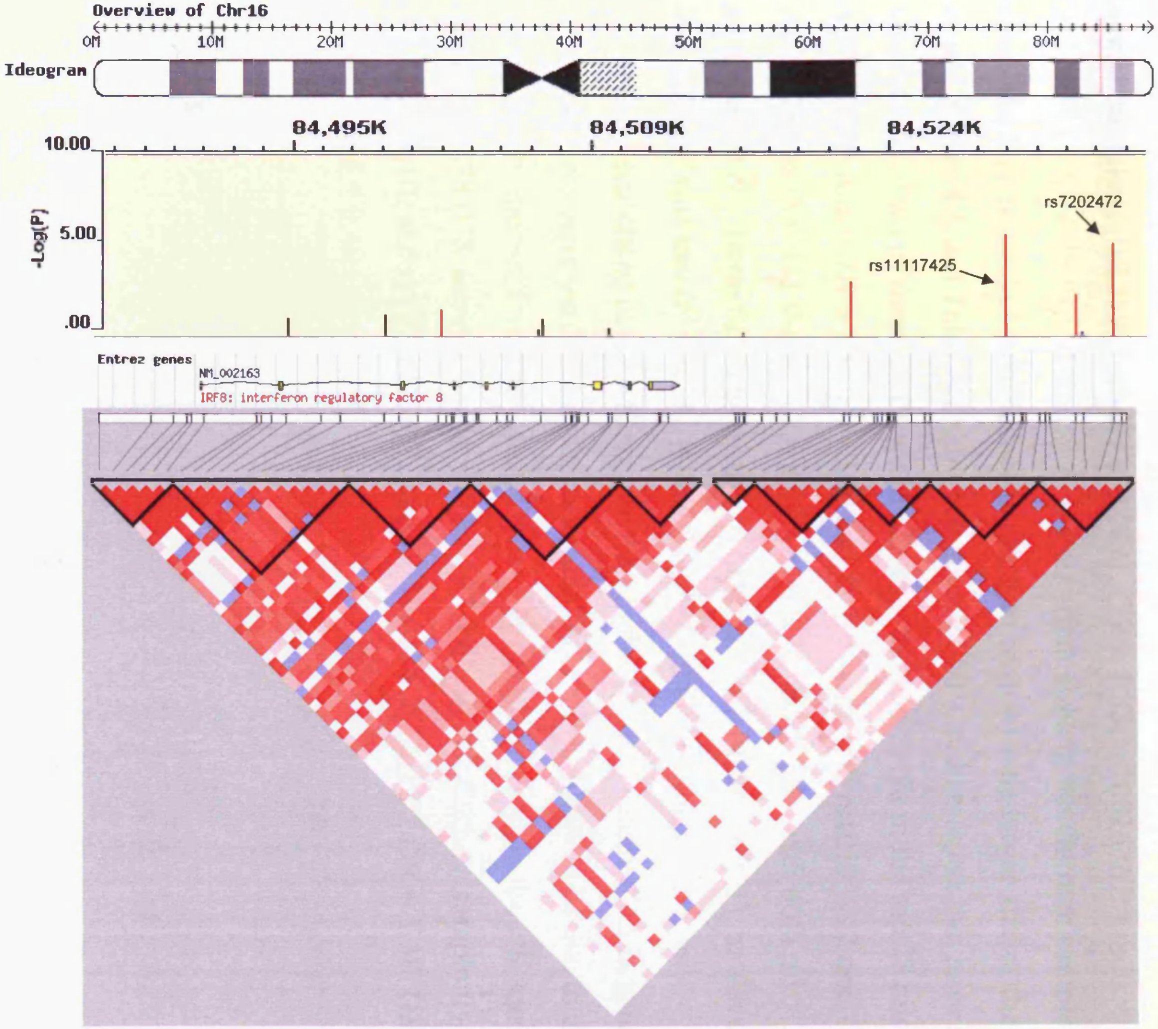


Figure 4.5: LD plot and results of SNPs in *IRF8* that were genotyped in the GWAS. Red bars denote P-values <0.05. The two most significant SNPs are not in high LD with each other. This diagram shows that both of these SNPs lie outside of the LD block of *IRF8*.

Another gene with multiple SNPs in the top hits is *C9orf5*. These SNPs are rs10123957 (P-min = 5.9 x 10^{-6}) and rs10816767 (P-min = 8.9 x 10^{-6}) and are in complete LD with each other in the HapMap CEU population (r^2 = 1, *D'* = 1). These SNPs both showed a higher level of significance in the UK sample (rs10123957: P-add = 1.9 x 10^{-4}, OR = 1.36; rs10816767: P-add = 5.9 x 10^{-4}, OR = 1.33) than in the German sample (rs10123957: P-add = 0.0071, OR = 1.35; rs10816767: P-add = 0.0034, OR = 1.39). These SNPs had similar or higher effect sizes in the German sample, so as before, the higher level of significance is likely to be due to the larger size of the UK sample. Figure 4.6 and Table 4.7 show the other 7 SNPs in this region that have P-values < 0.05. Two of these SNPs are in LD with the two SNPs in the top 25 hits (rs1003346 P-min = 2.9 x 10^{-5}, r^2 = 0.76 and *D'* = 1 with both rs10123957 and rs10816767; rs1537431 P-min = 2.6 x 10^{-4}, r^2 = 0.76 and *D'* = 1 with both rs10123957 and rs10816767). These SNPs were also in LD with each other (*D'* = 1), however the r^2 between them was low (r^2 = 0.58). All of the SNPs in this region were more significantly associated in the UK sample than in the German sample, but in general their effect sizes were very similar in both samples so this is likely to be due to the power of each sample. Two SNPs that did have larger effect sizes in the UK sample were rs2271878 (UK sample: P-add = 2.7 x 10^{-4}, OR = 1.38; German sample: P-add = 0.42, OR = 1.10) and rs7879057 (UK sample: P-add = 2.2 x 10^{-4}, OR = 1.38; German sample: P-add = 0.49 OR = 1.09).

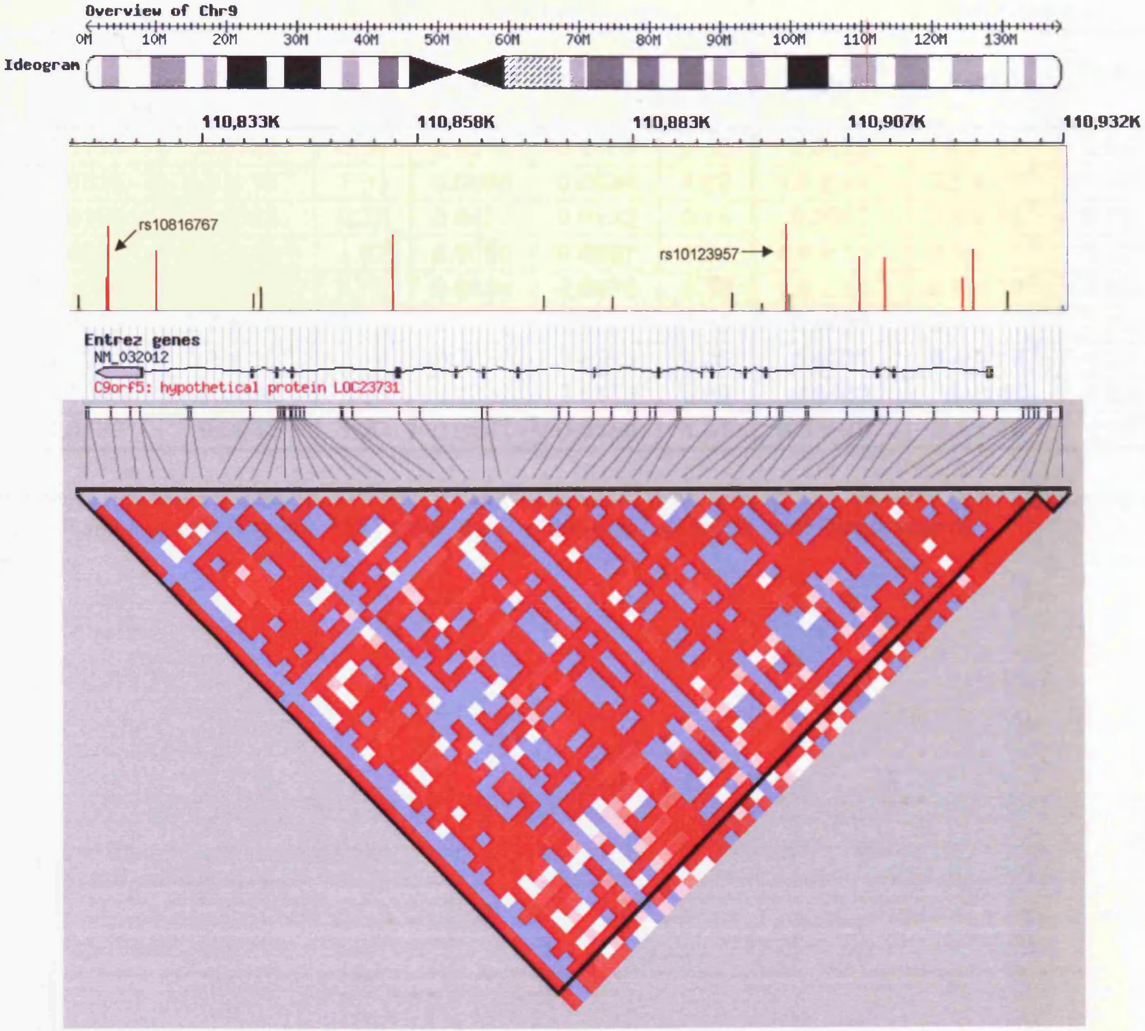


Figure 4.6: LD plot and results of SNPs in *C9orf5* that were genotyped in the GWAS. Red bars denote P-values <0.05. The two most significant SNPs are in perfect LD with each other with an $r^2 = 1$.

			UK Sample			German Sample			All Samples			
SNP	Position (bp)	Position in *C9orf5*	P-gen	P-add	OR	P-gen	P-add	OR	P-gen	P-add	OR	95% CI
rs7776	110822351	3' UTR	**0.0134**	0.0722	0.84	0.4556	0.2108	0.85	**0.0129**	**0.0318**	0.85	0.73-0.99
rs10816767	110822490	3' UTR	**0.0015**	**5.9×10^{-4}**	1.33	**0.0088**	**0.0034**	1.39	**1.8×10^{-5}**	**8.9×10^{-6}**	1.34	1.18-1.53
rs1537431	110828048	Intron 16	**0.0168**	**0.0046**	0.79	**0.0472**	**0.0132**	0.75	**0.0012**	**2.6×10^{-4}**	0.78	0.69-0.89
rs1003346	110855160	Intron 12	**0.0031**	**8.9×10^{-4}**	1.32	**0.0080**	**0.0087**	1.34	**4.9×10^{-5}**	**2.9×10^{-5}**	1.32	1.16-1.50
rs10123957	110900238	Intron 3	**7.4×10^{-4}**	**1.9×10^{-4}**	1.36	**0.0134**	**0.0071**	1.35	**1.6×10^{-5}**	**5.9×10^{-6}**	1.35	1.19-1.54
rs2271878	110908705	Exon 3	**0.0013**	**2.7×10^{-4}**	1.38	0.7072	0.4240	1.10	**0.0028**	**6.4×10^{-4}**	1.28	1.11-1.47
rs7870597	110911684	Intron 1	**0.0011**	**2.2×10^{-4}**	1.38	0.7657	0.4893	1.09	**0.0031**	**6.4×10^{-4}**	1.28	1.11-1.47
rs2805888	110920556	Intron 1	0.1043	**0.0396**	0.81	0.1339	0.1466	0.82	**0.0463**	**0.0141**	0.82	0.70-0.96
rs6825	110921747	Exon 1	**0.0136**	**0.0042**	1.27	0.0541	**0.0158**	1.31	**9.5×10^{-4}**	**2.3×10^{-4}**	1.28	1.12-1.46

Table 4.7: Table of results from all SNPs within *C9orf5* that had a minimum P-value < 0.05 in the whole GWAS sample. The results of these SNPs in the UK and German samples are also shown. P-gen – genotypic P-value; P-add – additive P-value; OR - Odds Ratio of the minor allele; CI - confidence interval.

4.3.2 Replication Study

A panel of 29 SNPs in the top hits based on their minimum P-values were chosen for genotyping in the replication sample. The nearest genes and positions of the SNPs relative to these genes are shown in Table 4.8. These SNPs were genotyped in the replication sample in which a total of 1244 cases and 1955 controls passed sample QC. Two of the SNPs failed the optimisation stage of the genotyping (see Table B.3 in Appendix). Of the remaining 27, all SNPs had a call rate > 80% and all had a MAF > 0.05. Two SNPs were out of Hardy-Weinberg equilibrium in the controls (rs4887111, P = 0.011; rs958877, P = 0.002). These were not excluded from the association analyses but any association found with these SNPs should be treated with caution.

SNP	Rank in GWAS	Proxy	Chr	Position (bp)	Closest RefSeq Gene	Position Relative to Gene
rs10513829	1		3	189971026	*LPP*	Intronic
rs7840675	3		8	34075413	*DUSP26*	Intergenic
rs747783	5		11	15670131	*SOX6*	Intergenic
rs11117425	6	rs11648084	16	84529771	*IRF8*	Intergenic
rs10123957	8		9	110900238	*C9orf5*	Intronic
rs10518444	9	rs2271081	4	125945653	*ANKRD50*	Intergenic
rs10816767	10	rs7034615	9	110822490	*C9orf5*	3' UTR
rs4887111	12		15	71815337	*LOC388135*	Downstream
rs7202472	13		16	84535002	*IRF8*	Intergenic
rs1181841	14		5	128580604	*ISOC1*	Intergenic
rs4327894	15		8	1740903	*MIRN596*	Intergenic
rs10512712	16		5	39728088	*DAB2*	Intergenic
rs1429411	17		2	197852246	*ANKRD44*	Intronic
rs7623540	18		3	189972233	*LPP*	Intronic
rs6136213	19		20	17798339	*SNX5*	Intergenic
rs4747165	20		10	72969307	*CDH23*	Intronic
rs9465637	23	rs13191158	6	20222086	*MBOAT1*	Intronic
rs1003346	26	rs11792635	9	110855160	*C9orf5*	Intronic
rs6984900	27		8	128373450	*POU5F1B*	Intergenic
rs1872285	29		11	15621627	*SOX6*	Intergenic
rs902025	31		15	61019453	*TLN2*	Intergenic
rs6498274	35		16	12273876	*SNX29*	Intronic
rs7541094	40		1	68536861	*WLS*	Intergenic
rs7411544	41		1	206428792	*PLXNA2*	Intronic
rs2077268	43		15	31661042	*RYR3*	Exonic
rs4940802	45		18	54861101	*ZNF532*	Intronic
rs958877	48		2	356409	*FAM150B*	Intergenic
rs3821173	49		2	207186404	*ADAM23*	Intronic
rs905950	64		16	12265706	*SNX29*	Intronic

Table 4.8: Table showing all SNPs in the replication panel, their rank in the GWAS, their closest RefSeq gene and their position relative to that gene.

Table 4.9 shows the results of the logistic regression using country and centre as a covariate for both additive and genotypic tests. SNPs are listed in order of each SNP's rank in the original GWAS. The minimum P-values for these SNPs in the replication sample alone ranged from 0.0066 to 0.9530, with only eight of the SNPs giving P-values < 0.05 (P-min = 0.0066 – 0.0444). The most significant of these is the SNP rs10512712 on chromosome 5 (P-add = 0.0066, OR = 1.15; P-gen = 0.0164). In the initial GWAS sample, this SNP had an additive P-value of 1.42 x 10^{-5} (OR = 1.33), ranking it as the 16th most significant hit. None of the SNPs achieved a high level of significant association in any of the sample groups individually, with only one SNP with P-values < 0.05 in more than one sample group, although showing the opposite direction of effect (rs7202472: German sample additive P = 0.019, OR = 1.31, Austrian sample additive P = 0.048, OR = 0.66) as shown in Table B.4 of the Appendix. It is clear that in the replication sample, these SNPs did not approach the level of significant association that was seen in the GWAS.

The genotypes for these SNPs when genotyped in the replication sample were combined with the genotypes from the discovery GWAS sample, making a total of 1828 cases and 4274 controls which passed QC. The results from the logistic regression analyses on this combined sample are shown in Table 4.9.

			GWAS Sample			Replication Sample			Whole Sample			
SNP	Rank in GWAS	Proxy	P-gen	P-add	OR	P-gen	P-add	OR	P-gen	P-add	OR	95% CI
rs10513829	1		**1.98×10^{-7}**	**1.2×10^{-7}**	0.68	0.7363	0.5533	1.03	**0.0154**	**0.0111**	0.90	0.83-0.98
rs7840675	3		**4.75×10^{-6}**	**8.11×10^{-7}**	1.67	0.9631	0.9530	0.99	**0.0131**	**0.0033**	1.21	1.07-1.38
rs747783	5		**1.12×10^{-5}**	**1.78×10^{-6}**	1.41	0.2723	0.1685	1.09	**2.25×10^{-4}**	**4.23×10^{-5}**	1.20	1.10-1.31
rs11117425	6	rs11648084	**8.78×10^{-6}**	**3.12×10^{-6}**	0.71	0.8594	0.6041	1.03	**0.0163**	**0.0070**	0.89	0.82-0.97
rs10123957	8		**1.61×10^{-5}**	**5.88×10^{-6}**	1.35	0.3077	0.1282	0.93	0.2955	0.1459	1.06	0.98-1.14
rs10518444	9	rs2271081	**3.66×10^{-5}**	**7.89×10^{-6}**	1.67	0.1481	0.3085	0.90	0.1056	**0.0413**	1.16	1.01-1.35
rs10816767	10	rs7034615	**1.85×10^{-5}**	**8.88×10^{-6}**	1.34	0.4168	0.1860	0.93	0.2681	0.1558	1.06	0.98-1.14
rs4887111	12		**2.46×10^{-5}**	**9.32×10^{-6}**	0.73	0.9126	0.9774	1.00	**0.0072**	**0.0022**	0.88	0.81-0.96
rs7202472	13		**3.64×10^{-5}**	**9.67×10^{-6}**	0.68	0.0682	**0.0280**	1.15	0.4116	0.1898	0.94	0.85-1.03
rs1181841	14		**6.22×10^{-5}**	**1.11×10^{-5}**	1.34	0.7085	0.9030	0.99	**0.0446**	**0.0200**	1.10	1.02-1.20
rs4327894	15		**5.39×10^{-5}**	**1.41×10^{-5}**	0.70	0.8580	0.9513	1.00	**0.0408**	**0.0114**	0.89	0.81-0.97
rs10512712	16		**8.24×10^{-5}**	**1.42×10^{-5}**	1.33	**0.0164**	**0.0066**	1.15	**1.79×10^{-5}**	**3.78×10^{-6}**	1.20	1.11-1.30
rs1429411	17		**1.57×10^{-5}**	**7.79×10^{-4}**	0.80	0.5838	0.3489	0.95	**0.0017**	**0.0018**	0.88	0.81-0.95
rs7623540	18		**9.02×10^{-5}**	**1.59×10^{-5}**	1.36	0.9644	0.7888	0.98	**0.0453**	**0.0183**	1.12	1.02-1.22
rs6136213	19		**4.06×10^{-5}**	**1.61×10^{-5}**	0.72	0.0832	**0.0413**	1.12	0.7208	0.4265	0.97	0.89-1.05
rs4747165	20		**7.71×10^{-5}**	**2.10×10^{-5}**	1.32	**0.0294**	0.5281	1.03	**1.63×10^{-4}**	**5.77×10^{-4}**	1.15	1.06-1.24
rs9465637	23	rs13191158	**2.43×10^{-5}**	**0.0125**	0.84	0.1269	**0.0444**	0.90	**4.00×10^{-4}**	**0.0028**	0.88	0.82-0.96
rs1003346	26	rs11792635	**4.88×10^{-5}**	**2.91×10^{-5}**	1.32	0.3857	0.1675	0.93	0.2606	0.1674	1.06	0.98-1.14
rs6984900	27		**9.32×10^{-5}**	**2.93×10^{-5}**	0.68	0.6355	0.3469	0.94	**0.0038**	**0.0010**	0.84	0.76-0.93
rs1872285	29		**1.74×10^{-4}**	**3.17×10^{-5}**	1.38	0.4199	0.3404	1.06	**0.0018**	**4.95×10^{-4}**	1.18	0.08-1.30
rs902025	31		**3.27×10^{-5}**	**0.0077**	0.78	0.7487	0.4531	1.06	**0.0186**	0.1302	0.92	0.83-1.03
rs6498274	35		**2.46×10^{-4}**	**4.47×10^{-5}**	1.31	0.0566	**0.0169**	1.14	**1.32×10^{-4}**	**2.65×10^{-5}**	1.19	1.10-1.28
rs7541094	40		**5.37×10^{-5}**	**2.15×10^{-4}**	1.27	0.6017	0.3136	1.05	**0.0025**	**0.0013**	1.13	1.05-1.23
rs7411544	41		**3.02×10^{-4}**	**5.84×10^{-5}**	1.31	**0.0109**	0.6942	0.98	**0.0097**	**0.0226**	1.10	1.01-1.19

Table 4.9: Comparison of results for the SNPs in the replication panel when genotyped in the GWAS sample, the replication sample and in the total sample. Significant P-values ($P < 0.05$) are in bold. P-add – additive P-value; P-gen – genotypic P-value; OR – odds ratio; CI – confidence interval.

			GWAS Sample			Replication Sample			Whole Sample			
SNP	Rank in GWAS	Proxy	P-gen	P-add	OR	P-gen	P-add	OR	P-gen	P-add	OR	95% CI
rs2077268	43		**2.62 x 10^{-4}**	**6.09 x 10^{-5}**	0.60	Failed Optimisation						
rs4940802	45		**1.45 x 10^{-4}**	**6.13 x 10^{-5}**	0.58	Failed Optimisation						
rs958877	48		**3.26 x 10^{-4}**	**6.42 x 10^{-5}**	1.30	0.1273	0.1101	0.92	0.4726	0.2525	1.05	0.97-1.13
rs3821173	49		**1.20 x 10^{-4}**	**6.47 x 10^{-5}**	0.76	0.7360	0.8729	1.01	**0.0119**	**0.0231**	0.91	0.84-0.99
rs905950	64		**5.19 x 10^{-4}**	**9.86 x 10^{-5}**	1.30	**0.0446**	**0.0134**	1.14	**2.04 x 10^{-4}**	**4.52 x 10^{-5}**	1.18	1.09-1.28

Table 4.9 continued

In this large combined sample, none of the SNPs in the replication panel showed genome-wide levels of significant association with DD. However, 3 SNPs showed a higher level of significance in the whole sample than they had in the initial GWAS sample. These are rs10512712 (GWAS sample minimum P = 1.42 x 10^{-5}; replication sample minimum P = 0.0066; whole sample minimum P = 3.78 x 10^{-6}), rs6498274 (GWAS sample minimum P = 4.47 x 10^{-5}; replication sample minimum P = 0.0169; whole sample minimum P = 2.65 x 10^{-5}) and rs905950 (GWAS sample minimum P = 9.86 x 10^{-5}; replication sample minimum P = 0.0134; whole sample minimum P = 4.52 x 10^{-5}). These 3 SNPs showed the same direction of effect in both the GWAS and replication sample, but a number of other SNPs in this replication panel showed a different direction of effect in the two samples. The SNP rs10512712 is in an intergenic region on chromosome 5, 260kb upstream of the nearest gene which is disabled homolog 2, mitogen-responsive phosphoprotein (*DAB2*). The other two SNPs are in perfect LD with each other in the CEU HapMap population ($r^2 = 1$, $D' = 1$) and are both within intron 7 of the gene sorting nexin 29 (*SNX29*) on chromosome 16, as shown in Figure 4.7. 25 other SNPs in this gene had P-values < 0.05 in the GWAS study but were not in the replication panel of SNPs. These are shown in Table 4.10. Many of these SNPs are not in LD with the most significant SNP in this gene, but are not independent of each other, as shown by the LD blocks in Figure 4.7.

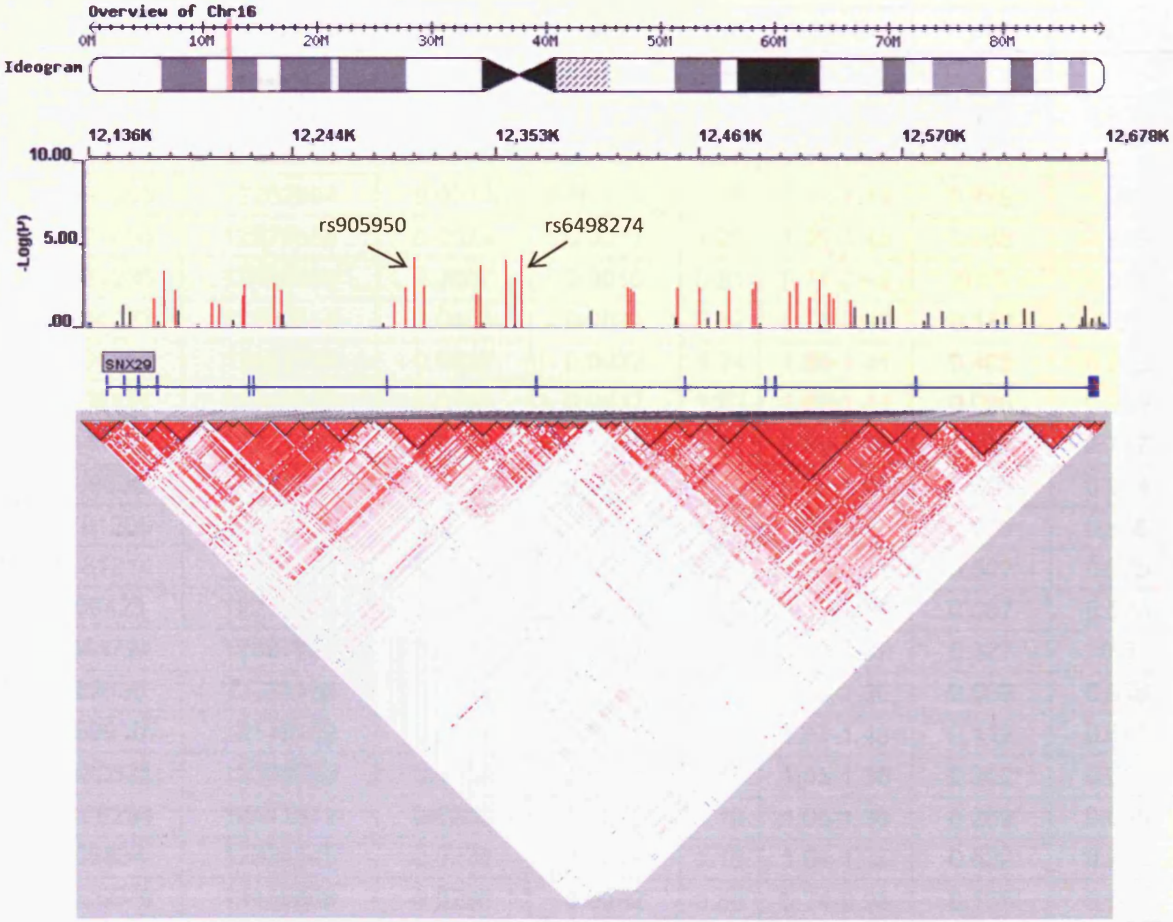


Figure 4.7: LD plot and results of SNPs in *SNX29* that were genotyped in the GWAS. Red bars denote P-values <0.05. The two most significant SNPs are LD with each other.

						LD with rs6498274	
SNP	**Position (bp)**	**P-gen**	**P-add**	**OR**	**95% CI**	r^2	*D'*
rs6498274*	12273876	2.46 x 10^{-4}	4.47 x 10^{-5}	1.31	1.15-1.50	1	1
rs830727	12217285	3.50 x 10^{-4}	6.77 x 10^{-5}	1.30	1.14-1.48	0.506	0.905
rs905950*	12265706	5.19 x 10^{-4}	9.86 x 10^{-5}	1.30	1.14-1.48	1	1
rs7195058	12252684	0.0011	2.16 x 10^{-4}	1.28	1.12-1.46	0.759	0.88
rs7191435	12270535	0.0029	0.0012	1.25	1.09-1.42	0.756	0.896
rs4781236	12432914	0.0067	0.0016	0.81	0.71-0.92	0.15	0.654
rs4781223	12399254	0.0073	0.0021	0.82	0.72-0.93	0.172	0.67
rs7201310	12421297	0.0029	0.0022	1.24	1.08-1.41	0.405	0.663
rs1704147	12142377	0.0106	0.0027	1.23	1.07-1.41	0.066	0.259
rs2941081	12221951	0.0139	0.0038	1.21	1.06-1.38	0.442	0.717
rs209836	12331057	0.0172	0.0044	1.21	1.06-1.37	0.604	0.954
rs4781209	12357278	0.0160	0.0047	1.21	1.06-1.37	0.675	0.956
rs4781214	12369368	0.0172	0.0049	1.20	1.06-1.37	0.502	0.975
rs709423	12212118	0.0114	0.0050	1.40	1.11-1.76	0.057	0.573
rs8043724	12397838	0.0111	0.0053	1.21	1.06-1.37	0.322	0.6
rs209835	12332132	0.0213	0.0059	1.20	1.05-1.36	0.569	0.976
rs12599107	12146102	0.0227	0.0067	1.26	1.07-1.49	0.112	0.651
rs2432625	12384723	0.0269	0.0071	1.19	1.05-1.36	0.262	0.697
rs6498294	12417371	0.0202	0.0078	1.19	1.05-1.36	0.269	0.665
rs209834	12334145	0.0278	0.0079	1.19	1.04-1.35	0.522	0.951
rs1035579	12438659	0.0296	0.0092	0.84	0.74-0.96	0.155	0.621

Table 4.10: SNPs within *SNX29* that had a minimum P-value < 0.05 in the GWAS and their LD with the most significant SNP. * These SNPs were genotyped in the replication sample. P-gen – genotypic P-value, P-add – additive P-value, OR – Odds ratio, CI – confidence interval.

						LD with rs6498274	
SNP	Position (bp)	P-gen	P-add	OR	95% CI	r^2	*D'*
rs830733	12250472	0.0149	0.0132	1.19	1.04-1.37	0.409	0.734
rs7205270	12427699	0.0184	0.0178	1.18	1.03-1.36	0.403	0.675
rs8048462	12441121	0.0202	0.0434	1.14	1.00-1.30	0.201	0.517
rs7188465	12450174	0.0576	0.0240	0.86	0.75-0.98	0.098	0.599
rs4353467	12629947	0.0325	0.0658	0.88	0.78-1.01	0.013	0.155
rs1472979	12206734	0.0419	0.0643	0.85	0.72-1.01	0.054	0.673

Table 4.10 continued

4.3.3 Results of Candidate Genes in GWAS

The results of the SNPs within previous candidate genes for DD and for those genes that were investigated in Chapter 3 of this thesis were extracted from the GWAS data for comparison. Table B.5 of the Appendix show the results of these SNPs in the whole GWAS sample, in the UK sample alone and in the Cardiff case-control sample if they were genotyped as part the candidate gene study. Only one SNP within *KIAA0319* gave a P-value < 0.05 (rs2817200, P-add in UK sample = 0.0092, OR = 0.80). This SNP is in a high level of LD with the SNP rs4504469 ($r^2 = 0.90$, $D' = 0.96$) which has shown association with DD in a number of previous studies (Francks et al. 2004; Cope et al. 2005a; Paracchini et al. 2008). This SNP was not significant in the whole GWAS sample (P-add = 0.057, OR = 0.88). Three other SNPs in *KIAA0319* that have shown significant association with DD previously were all in LD with the SNP rs3756819 (rs1061925, $r^2 = 0.72$, $D' = 1$; rs9461045, $r^2 = 0.76$, $D' = 1$; rs2143340, $r^2 = 1$, $D' = 1$). This SNP was not significant in the GWAS (P-gen = 0.549, P-add = 0.955, OR = 1.01) (see Table B.6 of Appendix). The SNPs on the Illumina HumanHap300 array provide 48% coverage of this gene at an $r^2 > 0.8$ based on phase II of HapMap.

A number of SNPs in *DCDC2* showed significant association (see Table B.5 of the Appendix). The array provides 77% coverage of the SNPs within this gene. The most significant of these SNPs was rs4712804 with an additive P-value of 3.66 x 10^{-4} in the whole sample. This SNP is not in a high level of LD with any previously associated SNPs in *DCDC2* ($r^2 < 0.8$). However, 2 SNPs show high levels of LD with the SNP rs807701, which was previously shown to be significantly associated with DD in the German sample as part of a haplotype with rs793862 (Schumacher et al. 2006a). The SNPs in high LD with rs807701 are rs870601 ($r^2 = 0.89$, $D' = 1$) and rs2274305 ($r^2 = 0.93$, $D' = 0.96$) and in the GWAS these SNPs had additive P-values of 0.0051 (OR = 1.21) and 0.0041 (OR = 1.22), respectively. These two SNPs are also in high LD with each other ($r^2 = 0.82$, $D' = 0.96$). Another significant SNP in *DCDC2* (rs9295619: P-add = 0.0051, OR = 1.21) is also in LD with rs807701 ($r^2 = 0.72$, $D' = 1$). This SNP is also in LD with the previous two SNPs ($D' > 0.96$) but the correlation between these SNPs is low ($r^2 < 0.7$). The SNP rs807724 which has also shown previous association with DD in *DCDC2* (Meng et al. 2005b) was in complete LD with rs2792682 ($r^2 = 1$, $D' = 1$), but this SNP was not significant in the GWAS (P-gen = 0.32, P-add = 0.15, OR = 1.12) (see Table B.6 of Appendix).

One SNP within *DYX1C1* (of which 68% is covered by the array) showed marginally significant association with DD in the UK sample (rs3759864, P-add = 0.0456). This SNP was not in high LD with SNPs in these genes that have shown association with DD in these genes previously ($r^2 < 0.8$). Three SNPs in *DYX1C1* that have shown association with DD recently (Dahdouh et al. 2009; Bates et al. 2009) were genotyped directly in this GWAS, but none of these were significantly associated (rs3743204, P-add = 0.48, OR = 1.06; rs600753, P-add = 0.65, OR = 0.97; rs685935 P-add = 0.92, OR = 1.01). Another two SNPs in *DYX1C1* that have shown association with DD (Taipale et al. 2003; Wigg et al. 2004; Marino et al. 2007; Dahdouh et al. 2009) are in LD with SNPs in this GWAS (see Table B.6 of Appendix). These were rs3743205, which is in LD with rs7181226, ($r^2 = 0.614$, $D' = 1$) and rs17819126, which is in LD with rs11857829 ($r^2 = 1$, $D' = 1$). Neither of these SNPs were significantly associated in this GWAS (rs7181226: P-add = 0.68, OR = 1.05; rs11857829: P-add = 0.61, OR = 1.07).

No other SNPs in the other DD candidate genes showed a significant association in the GWAS. Those SNPs that were typed in the candidate gene study and were not found to be associated in the case control sample (see Chapter 3) were not significantly associated with DD in the GWAS sample either.

4.3.4 Results of Main *DYX* Linkage Regions in GWAS

The results of those SNPs within the main *DYX* linkage regions were extracted from the GWAS data and Manhattan plots were produced for each region as shown in Figure 4.8 to Figure 4.15 (see Chapter 1 for more information on these regions). All of these regions have a large number of SNPs with P-values < 0.05 as would be expected by chance. For example, the *DYX1* region had 68 SNPs with P-values < 0.05. 1767 SNPs were tested in this region, so 88 SNPs would be expected to have P-values < 0.05 by chance. *DYX5* is the only locus that harbours a SNP with an additive P-value $< 5 \times 10^{-5}$ (rs6796074: P-add = 4.8×10^{-6}, OR = 1.45). After correcting for the number of independent SNPs in this region using the SET based analysis, this SNP had a P-value of 0.008 and the product of the P-values in this region was 0.007 (see Table B.7 of the Appendix). The SNP rs6796074 lies within an intergenic region towards the edge of *DYX5*, 117kb downstream from the nearest gene LOC152225. No other SNPs in the linkage regions were significant after correcting for the number of SNPs in the regions.

However, *DYX3* and *DYX4* both had significant product of P-values (P = 0.016 and 0.041, respectively). This indicates that these regions are likely to contain multiple independent signals of modest effect, rather than a single marker of strong effect.

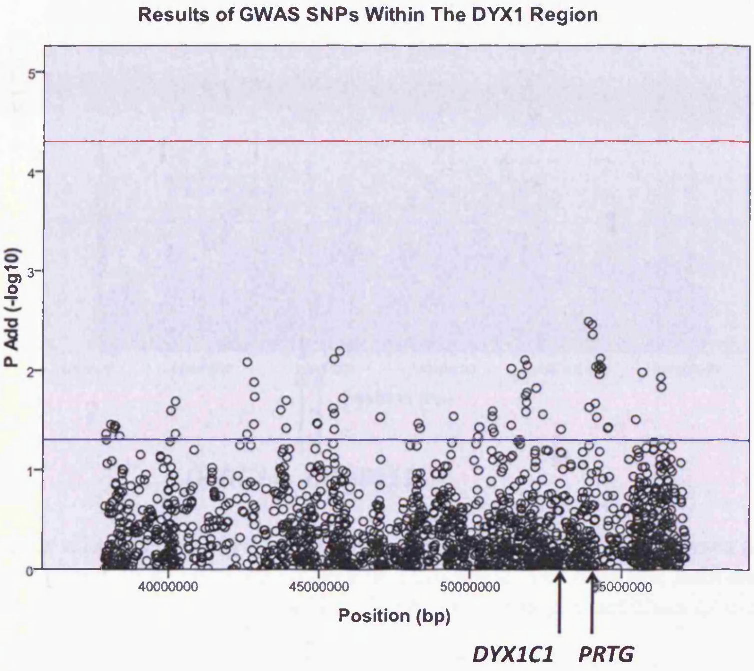


Figure 4.8: Manhattan plots of the additive (P-add) P-values from the logistic regression analyses of SNPs in the GWAS within the *DYX1* region when using country as a covariate. The blue line indicates a P-value of 0.05 and the red line indicates a P-value of 5 x 10^{-5}. The arrows show the positions of the DD candidate genes in this region, *DYX1C1* and *PRTG*.

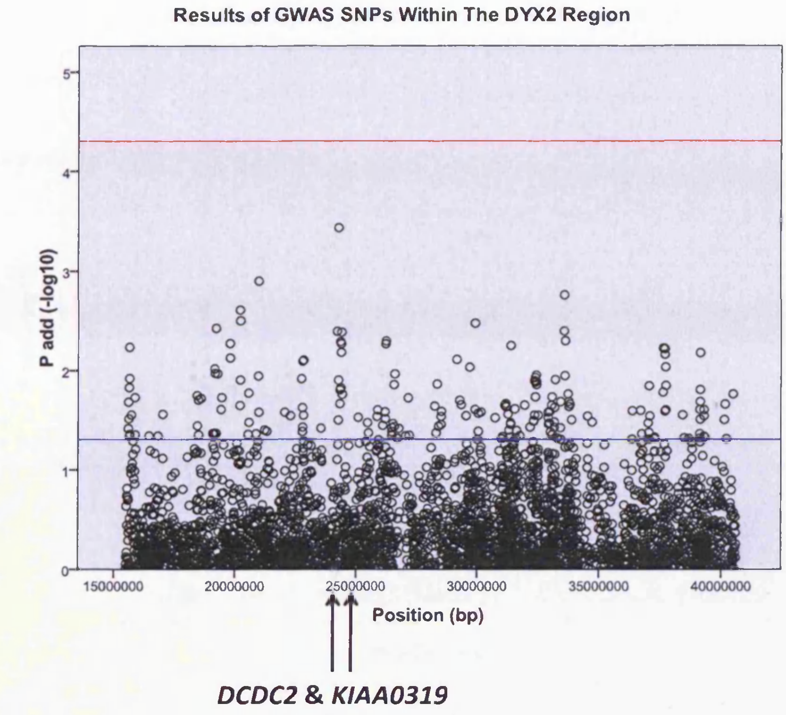


Figure 4.9: Plots of the additive (P-add) P-values from the logistic regression analyses of SNPs in the GWAS within the *DYX2* region when using country as a covariate. The blue line indicates a P-value of 0.05 and the red line indicates a P-value of 5 x 10^{-5}. The arrows show the positions of the DD candidate genes in this region, *DCDC2* and *KIAA0319*.

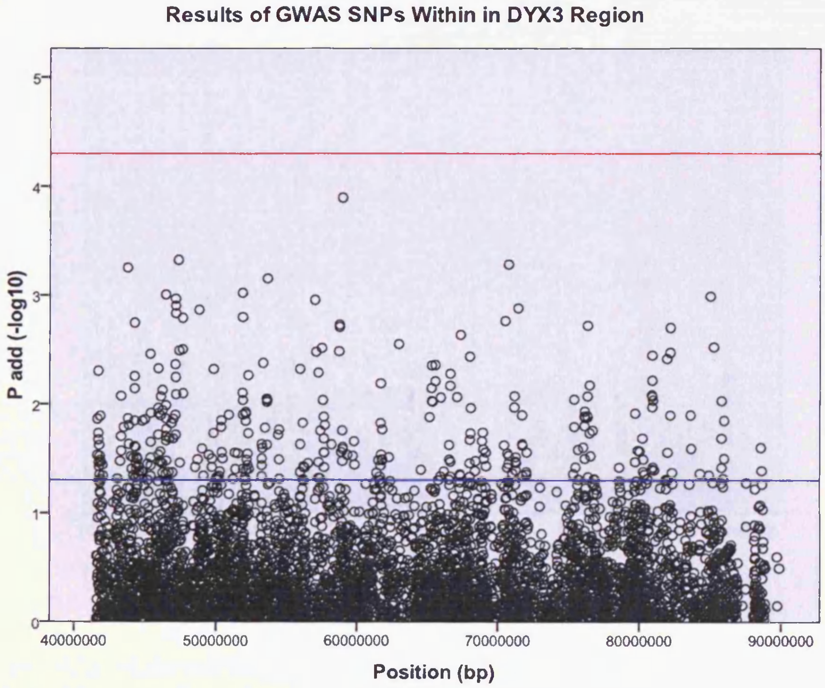


Figure 4.10: Plots of the additive (P-add) P-values from the logistic regression analyses of SNPs in the GWAS within the *DYX3* region when using country as a covariate. The blue line indicates a P-value of 0.05 and the red line indicates a P-value of 5 x 10^{-5}.

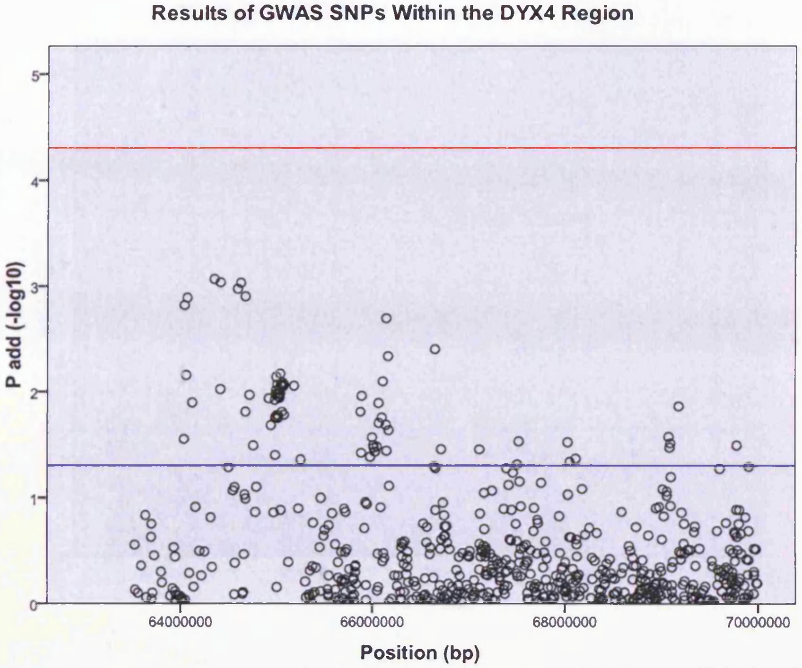


Figure 4.11: Plots of the additive (P-add) P-values from the logistic regression analyses of SNPs in the GWAS within the *DYX4* region when using country as a covariate. The blue line indicates a P-value of 0.05 and the red line indicates a P-value of 5 x 10^{-5}.

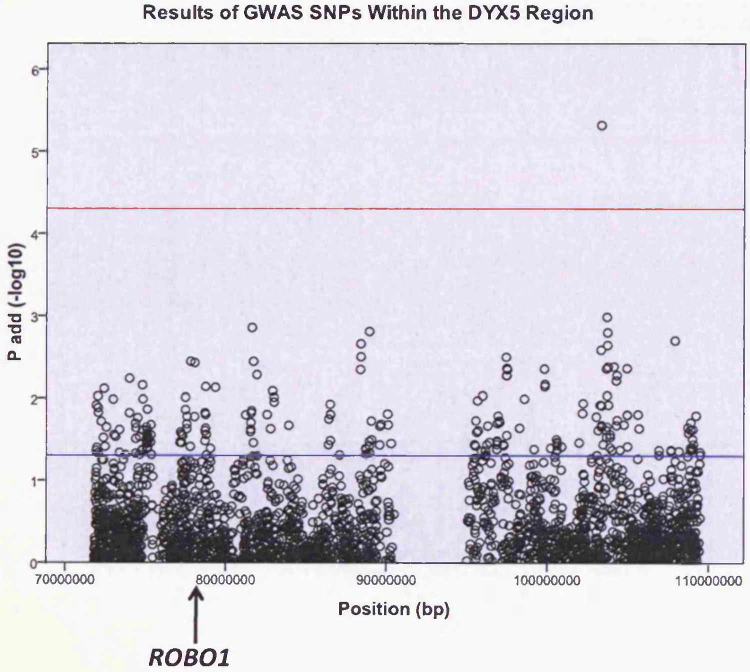


Figure 4.12: Plots of the additive (P-add) P-values from the logistic regression analyses of SNPs in the GWAS within the *DYX5* region when using country as a covariate. The blue line indicates a P-value of 0.05 and the red line indicates a P-value of 5 x 10^{-5}. The arrow shows the position of the DD candidate gene in this region, *ROBO1*.

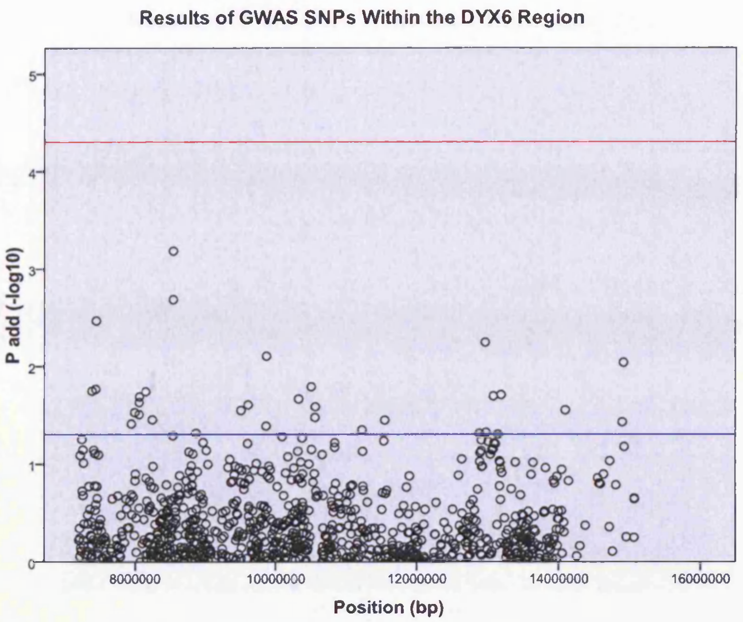


Figure 4.13: Plots of the additive (P-add) P-values from the logistic regression analyses of SNPs in the GWAS within the *DYX6* region when using country as a covariate. The blue line indicates a P-value of 0.05 and the red line indicates a P-value of 5 x 10^{-5}.

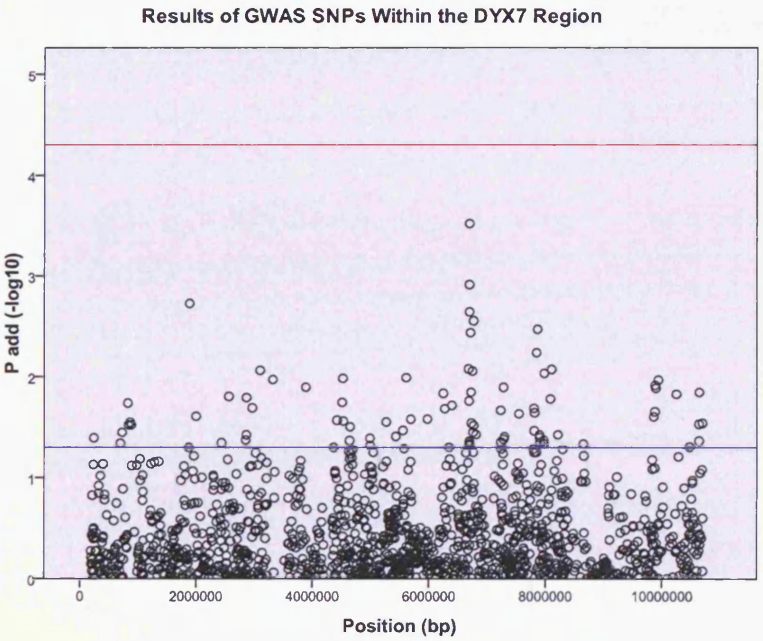


Figure 4.14: Plots of the additive (P-add) P-values from the logistic regression analyses of SNPs in the GWAS within the *DYX7* region when using country as a covariate. The blue line indicates a P-value of 0.05 and the red line indicates a P-value of 5 x 10^{-5}.

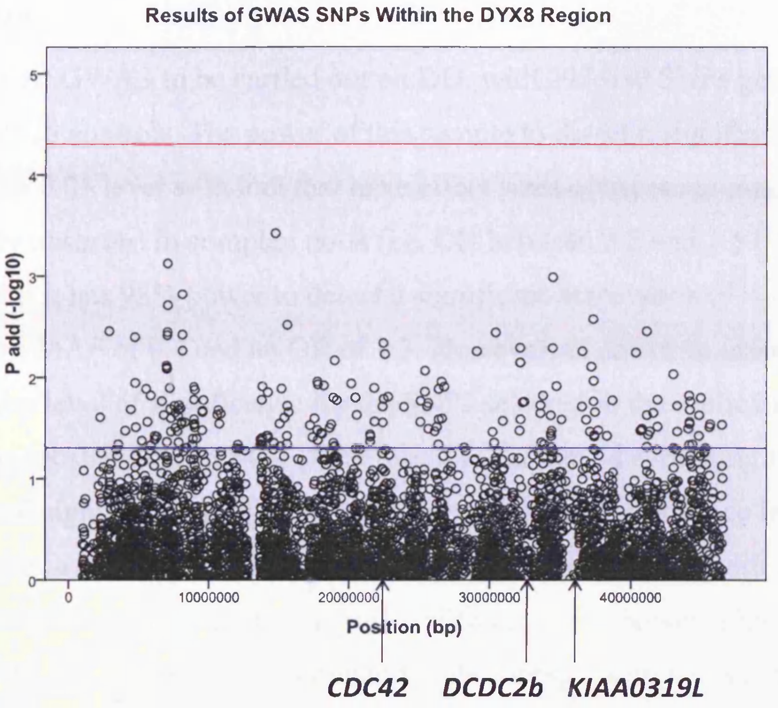


Figure 4.15: Plots of the additive (P-add) P-values from the logistic regression analyses of SNPs in the GWAS within the *DYX8* region when using country as a covariate. The blue line indicates a P-value of 0.05 and the red line indicates a P-value of 5 x 10^{-5}. The arrows show the positions of the DD candidate gene in this region, *CDC42, DCDC2b* and *KIAA0319L.*

4.4 Discussion

This is the first GWAS to be carried out on DD, with 297,650 SNPs genotyped in 585 cases and 2326 controls. The power of this sample to detect a significant association at the 0.05 level with loci that have effect sizes of the same magnitude to those commonly observed in complex traits (i.e. OR between 1.2 and 1.5 (WTCCC 2007)) is modest. It has 98% power to detect a significant association ($P < 0.05$) with a variant that has a MAF of 0.4 and an OR of 1.3. However, to detect an association at $P < 1 \times 10^{-4}$ (i.e. the level of significance for the SNPs selected in the replication panel), the power drops substantially to 54%. Therefore it is perhaps unsurprising that none of the SNPs showed significant association at the genome-wide significance level. Despite this lack of power, a number of SNPs showed suggestive significance with P-values less than 5×10^{-5} in the whole sample. The majority of the top hits showed a higher level of significance in the UK sample alone compared to the German sample. As the effect sizes of the top hits are generally similar in each of the two samples, the higher level of significance observed in the UK subset is likely to have been due to the larger size of this sample. This sample alone has 90% power to detect a significant association ($P < 0.05$) with a variant that has a MAF of 0.4 and an OR of 1.3, while the German sample has 65% power.

Three genes had multiple hits within the SNPs that had P-values $< 5 \times 10^{-5}$. *LPP* had two SNPs in the top 25 hits, which were significantly associated in the whole sample, as well as in the UK and German samples individually. This gene is on chromosome 3 and is thought to play a structural role at sites of cell adhesion in maintaining cell adhesion and motility (Petit et al. 2000), but is only expressed at very low levels in the brain. Another two SNPs in the top hits lie just outside of *IRF8*; this gene is on chromosome 16 and is a transcription factor of the interferon (IFN) regulatory factor (IRF) family (Weisz et al. 1992) . The IRF family of proteins control expression of IFN-α and IFN-β-regulated genes that are induced by viral infection and so play regulatory role in cells of the immune system. This gene is predominantly expressed in lymphoid tissues. The other gene with multiple hits is *C9orf5* on chromosome 9. The function of this gene is unknown, but it is highly expressed in the brain, especially in fetal brain tissue and the hypothalamus. Based on what is known about the functions of these genes, they do not appear to be obvious candidates for DD, however they cannot be discounted and may be involved in biological processes that have not yet been linked to DD.

In an attempt to replicate the findings of the initial GWAS, 27 SNPs were selected from the top hits (P-min < 5 x 10^{-4}) and were genotyped in a large independent sample consisting of 1244 cases and 1955 controls. This sample had 100% power to detect a significant association ($P < 0.05$) with a variant that has a minor allele frequency of 0.4 and an odds ratio of 1.3, and 88% power to detect association with the same variant at P < 1 x 10^{-4}. When genotyped in the replication sample, none of the SNPs approached the same level of significant association seen in the initial GWAS. 19 of these SNPs were not significant at the 0.05 level. This is often seen in GWAS and is thought to be caused by an overestimation of effect sizes of the variants in the original study in comparison to their effect sizes observed in the follow up study. This is referred to as the 'winner's curse' and can result in a failure to replicate the initial results (Ioannidis 2008). 11 SNPs had smaller effect sizes and the remaining 16 showed a different direction of effect in the replication sample. This suggests that many of the top hits in the initial GWAS were actually false positives arising out of chance due to the large number of SNPs tested rather than due to the presence of a true association. This finding highlights the need to adopt stringent P-value cut offs when selecting SNPs for follow up in order to reduce the chance that they are false-positives. This in turn demonstrates the requirement for large, well-powered studies in order to achieve these highly significant associations in the initial GWAS.

Combining the GWAS and replication samples produced a total of 1828 cases and 4274 controls, providing 100% power to detect an association with a variant that has a MAF of 0.4 and an OR of 1.3 at the 0.05 significance level and 99% power at the 1 x 10^{-4} level of significance. Combining both samples resulted in an increase in significance for 3 SNPs compared to the initial GWAS study. The most significant of these was rs10512712 which is in an intergenic region on chromosome 5, with the nearest gene being *DAB2* over 260kb away. The next two SNPs to show an increased significance in the whole sample were rs6498274 and rs905950 which are in complete LD with each other and are both within the gene *SNX29*. Not much is known about the function of this particular gene, but it is ubiquitously expressed and is a member of the sorting nexin family. Members of this family contain phosopholipd-binding motifs (PX domains) and are thought to facilitate membrane trafficking and protein sorting (Worby

& Dixon 2002). PX domains have also been found in yeast proteins that are involved in cell polarity (Ago et al. 2001), which could suggest a possible link for this gene in neuronal migration. However, the PX domains in the nexin family of proteins evolved independently from other PX domains (Teasdale et al. 2001) and are therefore likely to have unique cellular functions.

The results of this GWAS were compared with previous findings in DD. Only one SNP (rs2817200) within the DD candidate gene *KIAA0319* showed marginal significance in the UK sample. Interestingly, this SNP is in high LD with rs4504469 (r^2 = 0.90), which was previously found to be associated with DD in 3 independent UK samples (Francks et al. 2004; Cope et al. 2005a; Paracchini et al. 2008), as presented in Chapter 1. Two of these samples overlap with the Oxford (Francks et al. 2004) and Cardiff (Cope et al. 2005a) samples used in this GWAS and so the significant association of rs2817200 in the UK sample of the GWAS cannot be regarded as an independent replication. This SNP was not significant in the German subset. The sample used in the study by Paracchini and colleagues (2006) was independent of this GWAS and was unselected for reading ability. Three other SNPs in this gene that have shown association with DD previously (Francks et al. 2004; Cope et al. 2005a; Luciano et al. 2007; Paracchini et al. 2008; Dennis et al. 2009; Couto et al. 2010) were in LD with the SNP rs3756819 which was genotyped in this GWAS but was not significant. Two of the previously associated SNPs, rs1061925 and rs2143340 have shown previous association in UK samples that share some overlap with the UK case sample in this GWAS. Both of these SNPs were significantly associated in the UK sample used by Francks and colleagues, particularly with measures of orthographic choice. Cope and colleagues (2005a) also found significant association with rs2143340 in a subset of this UK case sample as part of a haplotype that spans *KIAA0319* and *TTRAP*. However, these studies used screened controls, whereas this GWAS used population controls and this may explain why rs3756819 did not show significant association with DD despite being in LD with these previously associated SNPs. Just 48% of *KIAA0319* was covered by the SNPs on the array and so even though none of the SNPs in this gene achieved genome-wide significance in this study, it still remains a plausible candidate gene for DD.

A number of SNPs in *DCDC2* showed significant association in both the UK and the whole GWAS sample. The most significant SNP in *DCDC2* was rs4712804, which has not shown significant association with DD in any previous studies, nor is it in LD with any previously associated SNPs in *DCDC2*. However, three of the SNPs that showed a significant association (namely rs2274305, rs870601 and rs9295619) are in LD with rs807701 which was previously shown to be significantly associated with DD in the German sample as part of a haplotype with rs793862 (Schumacher et al. 2006a). Interestingly, this SNP was also tested for association in the Cardiff cases and a severe subset of the Oxford cases in a previous study (Harold et al. 2006) but was not significantly associated with DD. All three SNPs in LD with rs807701 showed nominal levels of significant association with DD in the UK sample alone of this GWAS ($P > 0.02$), which may have been down to chance due to the large number of SNPs tested or may be because of the increased power of this GWAS as a result of the large population control sample. Of all the replicated candidate genes for DD, *DCDC2* showed the most association with DD in this GWAS. This may have been because the SNPs on the Illumina HumanHap300 array cover this gene more densely than either of the genes *KIAA0319* or *DYX1C1*.

None of the SNPs in the genes that were tested for association in Chapter 3 of this thesis were significant in the GWAS. The Cardiff case sample in the GWAS overlaps with the cases that were genotyped in the candidate gene studies, and even though these genes were not as well covered in the GWAS, this backs up the previous findings that none of these genes appear to be significantly associated with DD.

A large number of SNPs across all of the *DYX* susceptibility loci that were genotyped in this GWAS were significant at $P < 0.05$ level, as would be expected by chance. However, one SNP showed a P-add value $< 5 \times 10^{-5}$, which was still significant after correcting for the number of SNPs tested in this region (corrected $P = 0.0080$). The SNP rs6796074 lies within an intergenic region towards the edge of *DYX5,* 129kb away from D3S3665 which was found to be linked with DD in one of the first studies that showed evidence for linkage in *DYX5* (Nopola-Hemmi et al. 2001). None of the other regions had significant P-values after correcting for the number of SNPs tested, however *DYX3* and *DYX4* both had significant product of P-values (*DYX3* $P = 0.016$; *DYX4* $P = 0.041$) suggesting that these regions may contain multiple susceptibility variants of weak effect rather than a single variant of strong effect.

The only similar GWAS that has been carried out previously was that by Meaburn and colleagues (2008) using the TEDS cohort. None of the 65 most significant SNPs in this particular GWAS were in their 300 most significant SNPs which may have been due to the difference in the genotyping platforms used in each study.

This GWAS was limited by the size of the SNP array that was used. The Illumina HumanHap300 array has been estimated to cover 76% of the genome in the CEU population (based on Phase II of the HapMap project) (Mägi et al. 2007), and therefore it is possible that some susceptibility variants for DD were missed. The coverage of some of the previously associated DD candidate genes such as *KIAA0319* was even lower than this. In addition, this study was under-powered to detect susceptibility variants due to the size of the sample. Despite employing a two-stage design and forming collaborations with other groups as well as using population controls to substantially increase the sample, this GWAS was still under-powered to detect SNPs that have small effect sizes (which are commonly observed in complex traits such as DD) with a genome-wide level of significance.

Future work for this GWAS could include carrying out imputation to highlight additional susceptibility variants for follow up, and pathway analysis (e.g. testing for over-representation of gene ontology (GO) categories amongst the most significantly associated SNPs) could be used to identify interesting gene networks which should be prioritised for investigation. The increase in significance of the SNPs in the chromosome 5 intergenic region and within the *SNX29* gene in the whole sample also warrants further investigation. The latter in particular could be fine mapped in order to identify any functional variants, or it could be sequenced in a number of DD cases to identify novel common variants.

In conclusion, this GWAS has not found any new convincing susceptibility variants for DD. While some SNPs showed a high level of significance and some SNPs within previously associated regions were significant, none achieved genome-wide significance. Larger sample sizes and larger arrays need to employed in order to improve the chances of identifying convincing susceptibility variants for this complex disease.

Chapter 5: NeuroDys Genome-wide Pooling Study

5.1 Introduction

As discussed in Chapter 4, GWAS of complex diseases require large sample sizes in order to be powerful enough to detect genome-wide significant levels of association. One approach to reducing the cost, time and labour involved in performing a GWAS is to pool DNA samples into case and control pools (Kirov et al. 2006; Docherty et al. 2007; Macgregor et al. 2008). This commonly involves combining equal amounts of DNA from each sample to form pools containing cases and controls. Alternatively, individuals with trait values at the two extremes of a quantitative trait can also be pooled into two samples (Sham et al. 2002). These pools are then genotyped to estimate the difference in allele frequency for each variant. The results from these pooling studies are then used to select a smaller number of SNPs to be genotyped in the samples individually and tested for association with a particular disease at a fraction of the cost of a typical GWAS (Sham et al. 2002; Norton et al. 2002). For example, the allele frequencies in a sample of 500 cases and 500 controls can be measured from two pooled samples, rather than by genotyping 1000 samples, which represents an increase in efficiency of 500-fold (Sham et al. 2002). However, in order to achieve accurate estimates of allele frequencies, most studies run replicates of the pooled samples and so this increase in efficiency would be closer to 100-fold (based on 5 replicates of each pool being genotyped).

Unfortunately, this increase in efficiency is achieved at the cost of detailed information that could have been obtained through individual genotyping (Sham et al. 2002) and a loss in statistical power due to imprecise estimates of allele frequencies (Barratt et al. 2002). Pooling studies are subject to experimental errors that do not apply to individual genotyping and which result in an overall variance in the allele frequency estimation. Quantitative errors can be introduced during the formation of the pools, due to inaccuracies in PCR reactions and also during allele frequency estimation (Barratt et al. 2002). However, these errors can be taken into account during association analysis by using an appropriate test, such as the combined Z-test (Sham et al. 2002). Further information is lost because haplotypes cannot be constructed and epistasis and heterosis cannot be studied (Norton et al. 2002). In addition, when using pooled samples to study quantitative traits, even more information is lost because within-pool differences cannot be explored (Sham et al. 2002). However, these issues may be offset in part by the ability to test many more SNPs than would be feasible if genotyping the same number of samples individually (Norton et al. 2002). In order to reach a compromise between the cost savings of a pooling study and the full information that is provided by individual genotyping, pooling studies are often conducted using a two-stage design in which SNPs that show a significant association in a pooling study are then followed up with individual genotyping.

Several genome-wide pooling studies have been published which have been successful in either identifying new susceptibility loci for diseases or replicating previously known loci (Melquist et al. 2007; Steer et al. 2007; Stokowski et al. 2007; Abraham et al. 2008; Kirov et al. 2008; Shifman et al. 2008), providing proof of principle that DNA pooling can provide an effective alternative to a large and expensive GWAS. For example, in an Alzheimer's disease (AD) case-control pooling study using the Illumina HumanHap 300 and Illumina Sentrix HumanHap240S arrays carried out by Abraham and colleagues (2008), *APOE* showed the most significant association with AD. The association of this locus with the late-onset form of AD has been well replicated in studies previously (Saunders et al. 1993; Farrer et al. 1995; Coon et al. 2007) and has also shown the most significant association with AD in more recent genome-wide studies of Alzheimer's disease (Harold et al. 2009; Lambert et al. 2009; Seshadri et al. 2010). A genome-wide pooling study of schizophrenia by Kirov and colleagues (2009) also illustrated that Illumina arrays can provide highly reproducible results. They used a parent-offspring trios design to screen SNPs on the Illumina HumanHap550 array and identified a significant SNP within *RBP1.* This gene inhibits PI3K/Akt signalling (Farias et al. 2005) and genes in this pathway have been implicated in schizophrenia pathogenesis (Kalkman 2006).

5.1.2 Aims

As significant SNPs in the initial NeuroDys GWAS did not show strong evidence of replication in the follow-up sample (see Chapter 4), the aim of this section of the thesis was to undertake another GWAS which would test a much larger number of SNPs for an association with DD. Since the initial GWAS was undertaken, Illumina released the Human1M-Duo array which allows nearly 1.2 million markers to be genotyped, capturing 95% of the common variation in the HapMap CEU population at an $r^2 > 0.8$ (InfiniumHD Data Sheet, www.illumina.com), whereas the Illumina HumanHap 300 array has been estimated to cover just 76% of the genome in the CEU population (based on Phase II of the HapMap project) (Mägi et al. 2007). Therefore, the Illumina Human1M-Duo array was selected for this GWAS and as Illumina arrays have been shown to produce reliable results when using pooled DNA (Abraham et al. 2008; Kirov et al. 2009), this study was undertaken using pooled DNA samples to reduce the cost. As discussed above, pooling studies can only provide estimates of allele frequencies so selected SNPs were followed up with individual genotyping in an attempt to confirm the findings. In addition, the results of this pooling study were compared with the initial GWAS in order to ascertain the concordance between the two.

5.2 Methods

5.2.1 NeuroDys Pooling Study

5.2.1.1 Pooled Samples

Samples from the NeuroDys replication sample (see Chapter 4) were combined to form 3 case pools and 3 control pools corresponding to different geographical locations as shown in Table 5.1. This sample is referred to as the NeuroDys pooling sample from now on. Only those samples that had a call rate > 98% in the replication panel of SNPs and which had sufficient DNA available were included.

Country/Centre	Case Pool	Control Pool
UK NeuroDys Pool:	**461**	**219**
Cardiff	187	219
Oxford	274	0
Central European NeuroDys Pool:	**532**	**912**
Bonn	196	400
Munich	104	188
Switzerland	25	40
Netherlands	100	103
Austria	107	181
Finnish NeuroDys Pool	**286**	**321**
Total	**1279**	**1452**

Table 5.1: Samples from the replication sample that were pooled to form 3 pooled sample sets; UK, Central European and Finnish.

The UK NeuroDys pool was constructed at the University of Cardiff, the Central European NeuroDys pool was constructed at the University of Bonn and the Finnish pool was constructed at the University of Jyvăskylă (see Chapter 2 for method). After construction, the concentration of the UK pools was 10ng/μl so they were concentrated using Microcon tubes (see Chapter 2 for method) and subsequently diluted to a concentration of 50ng/μl. To test the accuracy of the pool construction, primers for the SNPs rs11648084 and rs1892577 from the GWAS replication panel of SNPs were designed using primer 3 software (http://frodo.wi.mit.edu/primer3/) and the extension primers were designed using FP PRIMER 1.0.1b (http://m034.pc.uwcm.ac.uk/FP_Primer.html) (see Table C.1 of the Appendix for primer sequences). These were then used to genotype the case and control pools using the SNaPshot method in order to estimate the allele frequencies for these SNPs (see Chapter 2 for method). The estimated differences in allele frequencies between the case and control pools for these SNPs were then compared to the actual differences in allele frequencies when the pooled samples were genotyped individually.

5.2.1.2 Genotyping

Genotyping was performed using the Illumina Human1M-Duo chip at the University of Bonn according to the manufacturer's protocol (see Chapter 2 for description). Each case pool and each control pool were run in replicates of 6. Chips were scanned using the Illumina iScan system and the raw intensities were normalised using BeadStudio v3.2 software. The normalised intensities were then extracted for statistical analysis.

To exclude poorly performing replicates, replicate arrays were excluded if the estimated allele frequencies produced a Pearson correlation of $r \leq 0.991$ with at least two other replicate arrays for that pool as these appeared to be outliers (see Table C.2 in Appendix for Pearson correlations between all arrays).

An approximation of allele A frequencies for each replicate were calculated by Valentina Moskvina using the normalised intensities and the following equation, where Xnorm is the normalised intensity of allele A and Ynorm is the normalised intensity of allele B:

$$\text{Frequency of allele A} = \text{Xnorm} / (\text{Xnorm} + \text{Ynorm}).$$

These frequencies were then averaged over the number of replicates in each pool.

SNP QC was carried out by Valentina Moskvina and SNPs were excluded from the analysis if they had a $\text{MAF} < 0.05$ in either cases or controls. In order to exclude those SNPs whose allele frequencies appeared to be poorly predicted by the pooling analysis, the MAFs of SNPs in the control pools were also compared with their frequencies in the CEPH population of the HapMap project. This is a filter that has also been used by Kirov et al.(2006), which assumes that the frequency of these SNPs in the CEPH population approximates to their true frequency in the control sample and that differences between the two are the result of a bias in estimating the frequency of the alleles in the pooling experiment. The correction coefficient, *k*, was calculated using the following formula:

$$k = \frac{H_A}{H_B} \cdot \frac{f_B}{f_A}$$

In this formula, H_A is the frequency of allele A in the controls, H_B is the frequency of allele B in the controls f_B is the frequency of allele B in the HapMap CEPH population and f_A is the frequency of allele A in the HapMap CEPH population. If $k = 1$ then the frequency estimated in the control pool is identical to the frequency in the CEPH population. It has previously been shown that the use of SNPs with extreme values of k results in high error rates (Moskvina et al. 2005). Therefore, the SNPs with the worst 10% of k values (5% in each direction) were filtered out in order to remove SNPs with extreme values of k without excluding a large proportion of true positive results. The coefficient of variation for each SNP across replicates was calculated by dividing the standard deviation of the allele frequencies by the mean allele frequency. Those SNPs that had a coefficient of variation > 0.5 across the replicate arrays were excluded in order to remove the outlying SNPs that showed the most variance across the arrays, indicating poorly performing assays (see Figure C.1 in Appendix for histograms of coefficient of variation). Finally, SNPs which did not show the same direction of effect across all 3 pooled sets were also excluded.

5.2.1.3 Association Analyses

In a pooling study, a standard Pearson χ^2-test should not be used to test the magnitude of the difference in allele frequencies between cases and controls because the assumption that any variance is determined entirely by sampling variation is unrealistic (Sham et al. 2002). The variance in allele frequencies can be inflated by experimental errors that are specific to pooling studies. Therefore, association analyses were carried out with the help of Valentina Moskvina using the Combined Z-test (Sham et al. 2002; Macgregor 2007; Abraham et al. 2008; Kirov et al. 2008). This test combines a chi-square statistic T for testing differences between two proportions (in this situation, allele frequencies) in cases and in controls accounting for sampling variance, with Z-statistics for testing the differences in mean allele frequencies between cases and controls accounting for standard error due to experimental error, as shown in the formula:

$$T_{comb} = \frac{(\bar{f}^{(1)} - \bar{f}^{(2)})^2}{v_1 + v_2 + \varepsilon_1^2 + \varepsilon_2^2}$$

where $f^{(1)}$ and $f^{(2)}$ are the allele frequencies in cases and controls, v_1 and v_2 are the sampling variances in cases and controls and ε_1 and ε_2 are the standard errors due to experimental error in cases and controls.

P-values across all pools were then combined using Fisher's combined probability test shown below, where p is the probability of the i^{th} hypothesis test:

$$\chi^2 = -2\sum_{i=1}^{k} \log_e(p_i).$$

When all the null hypotheses are true, and the pi (or their corresponding test statistics) are independent, χ^2 has a chi-square distribution with $2k$ degrees of freedom, where k is the number of tests being combined (3 in this case, so there are 6 degrees of freedom).

5.2.2 Individual Genotyping of NeuroDys Pools

5.2.2.1 Sample

Individual genotyping was carried out on those samples that were used to construct the pools (the 'NeuroDys individual genotyping sample', as well as additional cases and controls that fulfilled the replication sample ascertainment criteria (the 'whole NeuroDys individual genotyping sample'), as shown in Table 5.2. The additional samples from France were recruited by the University of Toulouse and the Centre National de la Recherche Scientifique (CNRS) in Paris. Additional samples from Hungary were recruited by the University of Budapest. No individual genotyping was carried out on the Finnish pooled samples.

	NeuroDys Individual Genotyping Sample		Additional NeuroDys Samples		
Country/Centre	**Case Pool**	**Control Pool**	**Cases**	**Controls**	**Centre**
UK:	**461**	**219**			
Cardiff	187	219			1
Oxford	274	0	53	359	2
Central Europe:	**527**	**902**			
Bonn	196	400	4	285	3
Munich	104	188	103	26	4
Switzerland	25	39	2	5	5
Netherlands	96	95	57	79	6
Austria	106	180	80	28	7
Finland	**0**	**0**			
France	-	-	161	204	8
Hungary	-	-	78	154	9
Total	**988**	**1121**	**538**	**1140**	

Table 5.2: Samples from the pools that were individually genotyped using the follow up panel of SNPs. Not all samples that were in the pools were genotyped individually due to lack of available DNA. Some additional samples that were not in the pools were also genotyped using this panel.

5.2.2.2 Genotyping of Replication Panel

SNPs were not solely chosen for follow up with individual genotyping based on their top ranking significance in the pooling study; SNPs were also chosen if they were significant in more than one pool, in genes or regions with multiple hits, of functional interest or in a pathway of interest and if they were significant in both the pooling study and the initial NeuroDys GWAS, as shown in Table 5.5.

The SNPs selected for follow-up were entered into the Sequenom MassARRAY Assay Design 3.1 software in order to design a multiplex panel of 40 SNPs (see Table C.3 in Appendix for primer sequences). This panel of SNPs was genotyped in the replication sample using the Sequenom MassARRAY iPlex GOLD system as described in Chapter 2. Genotype calling was carried out using the Typer 3.4 software. All SNP assays were initially optimised by genotyping DNA from 30 CEPH parent-offspring trios. Cluster plots for all SNPs were inspected manually, and SNP assays that did not produce distinct clusters were excluded. 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 rates. 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.

After genotyping, SNPs were tested for Hardy-Weinberg equilibrium in controls and their MAFs were calculated using PLINK v1.05 (Purcell et al. 2007). Samples with a call rate < 70% were not included in the analysis.

SNPs were tested for an association with DD using logistic regression carried out in PLINK v1.05 (Purcell et al. 2007) using both the additive and genotypic models. To correct for possible population stratification, a covariate corresponding to centre was applied as shown in Table 5.2. This analysis was performed for those samples that were in the DNA pools alone and also with the additional samples indicated in Table 5.2.

In order to ascertain the concordance across studies for those SNPs that were individually genotyped, their association results in this study were compared with those in the initial NeuroDys GWAS (in both the UK GWAS sample and the whole NeuroDys GWAS sample). As the initial NeuroDys GWAS was carried out on the smaller Illumina HumanHap300 array, not all of the SNPs that were individually genotyped in this pooling study had been genotyped in the initial GWAS as well. For those SNPs, imputation was carried out on the NeuroDys dataset using PLINK v1.05 (Purcell et al. 2007), as described in Chapter 2. Association analysis was then carried out in the same way as in the GWAS (see Chapter 4) using imputed genotypes for SNPs that had an information score greater than 0.8 as recommended in the PLINK documentation.

5.3 Results

5.3.1 NeuroDys Pooling Study

Due to the lack of convincing evidence for associations with DD from the initial GWAS, a larger GWAS was carried out using pooled DNA from the NeuroDys replication sample to form case and control pools from the UK, Central Europe and Finland. Table 5.3 shows the results of validating these pooled samples using SNaPshot. For the SNP rs11648084, the UK pool had an error rate of 0.97% and for rs1892577 there was an error rate of 1.07%. This gave an overall average error rate of 1.02%. The Central European and Finnish pools were validated in a similar way by Kerstin Ludwig and colleagues at the University of Bonn and Myriam Peyrard-Janvid and colleagues at the University of Jyvăskylă.

Genome-wide pooled genotyping was carried out on the Illumina 1M-Duo chip. Predicted frequencies for each SNP were averaged over the replicate case and replicate control assays.

SNP	Difference in allele frequencies from individual genotyping (%)	Difference in allele frequencies estimated from pools (%)	% Error rate
rs11648084	4.14	3.17	0.97
rs1892577	6.62	5.55	1.07

Table 5.3: Comparison of difference in allele frequencies when sample were genotyped individually and in the UK pools.

For two of the case replicates and two of the control replicates of the Central European pools, the chips had sections that could not be imaged by the iScan system (either due to hybridisation or staining issues) so were not included in subsequent analysis. The remaining replicates all passed QC with a Pearson's correlation $r > 0.993$, as shown in Figure C.2 of the Appendix.

Following stringent QC filters, 501,409 SNPs were analysed. Figure 5.1 shows the results of these SNPs when their P-values were combined across all pools using the Fisher's combined probability test. Two of these SNPs achieved a genome-wide level of significance, rs11686995 on chromosome 2 (P-Fisher = 1.74×10^{-10}) and rs12743401 on chromosome 1 (P-Fisher = 3.37×10^{-9}). A further 109 SNPs had P-values < 1×10^{-4} (see Table C.4 of the Appendix for a list of the 200 most significant SNPs).

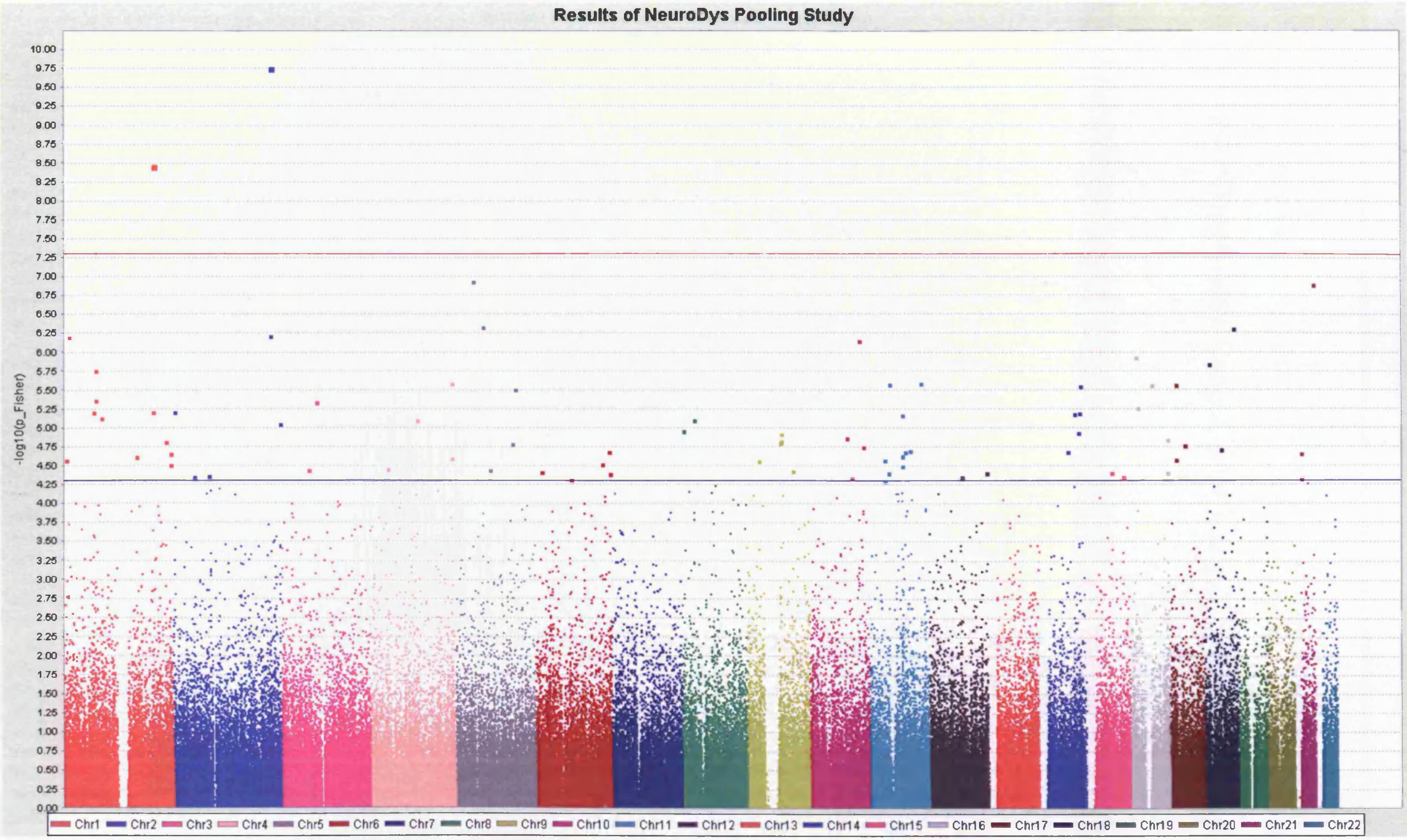


Figure 5.1: Manhattan plot of Fisher P-values (-log10) from all the pools in the NeuroDys Pooling study. The red line indicates genome-wide significance with a P-value of 5 x 10^{-8}, the blue line indicates a P-value of 5 x 10^{-5}.

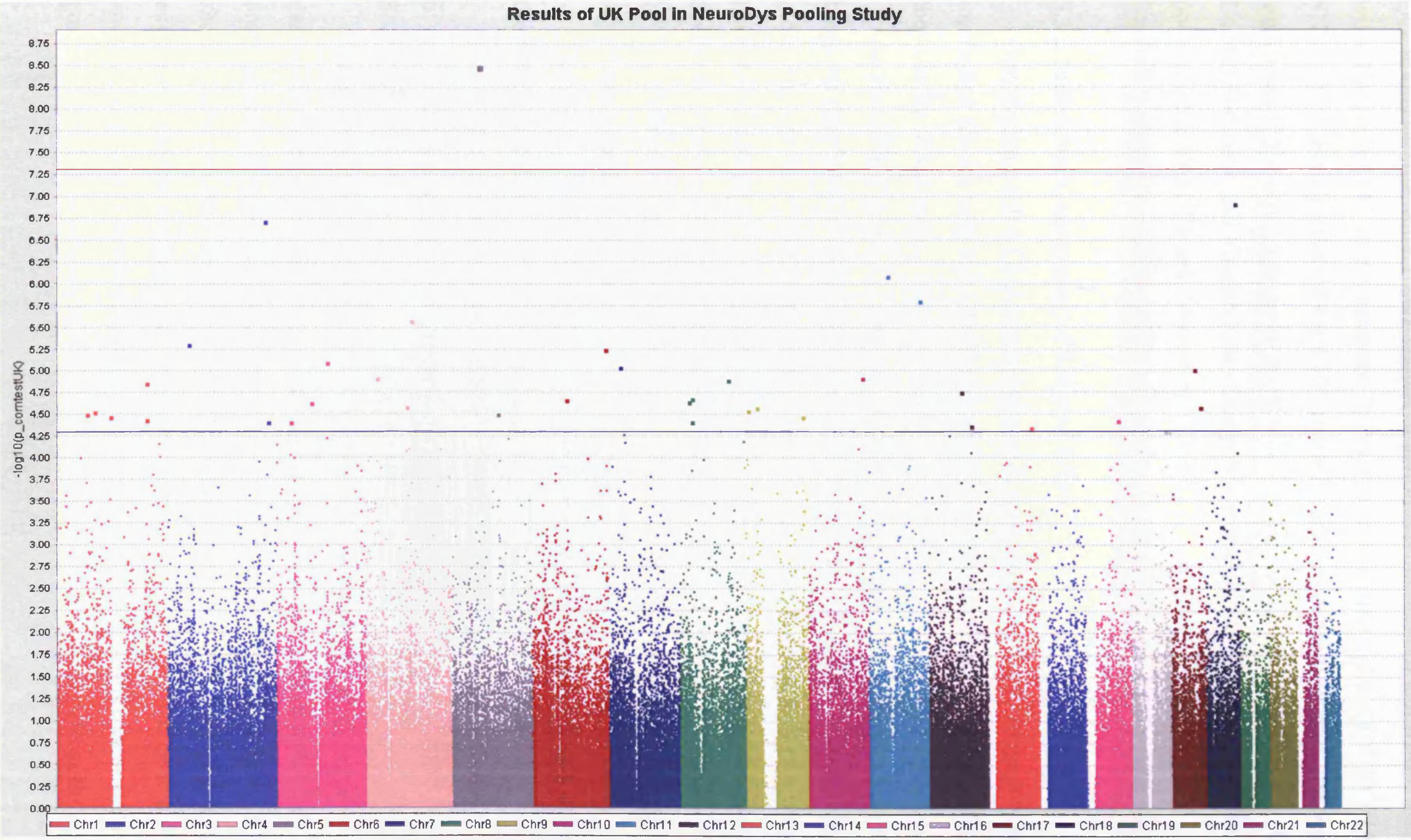


Figure 5.2: Manhattan plot of combined P-values (-log10) from the UK pool in the NeuroDys Pooling Study. The red line indicates genome-wide significance with a P-value of 5 x 10^{-8}, the blue line indicates a P-value of 5 x 10^{-5}

Figure 5.2 shows the P-values in the UK NeuroDys pools for those SNPs that passed QC. In this sample, only 1 SNP showed a genome-wide level of significant association with DD: rs6865447 on chromosome 9, P-value from combined test = 3.30 x 10^{-9}. This SNP had a P-Fisher value of 4.6 x 10^{-7} in the NeuroDys pooling sample, but was not significant in the other individual pools with P-values of just 0.87 and 0.73 in the Central European and Finnish NeuroDys pools, respectively. In the UK NeuroDys pool, another 59 SNPs had P-values < 1 x 10^{-4}.

5.3.2 Individual Genotyping of Selected SNPs

Genotyping pooled DNA can only provide an estimate of allele frequency and any interesting results need to be confirmed by genotyping the samples individually before they can be relied upon.

As shown in Table 5.5, a total of 40 SNPs were selected for individual genotyping based on their level of significance in the pooling study and if they were significant in more than one pool (29 SNPs), if they were functional or were located in a gene in a pathway of interest (3 SNPs), if they were in a gene that had multiple significant hits (4 SNPs) or if they were significant in both the pooling study and the initial GWAS (4 SNPs). Unfortunately, Sequenom assays for the two SNPs that achieved genome-wide significance could not be designed.

The SNPs rs5063, rs945386 were chosen because they are functionally interesting as they cause an amino acid change in the genes they are in. As shown in

Table 5.4, the non-synonymous SNP rs5063 (32Val →Leu) is within the first exon of the gene natriuretic peptide precursor A (*NPPA*) and rs945386 is a non-synonymous SNP (38Met→Thr) within exon 2 of *KIAA1984*.

The SNP rs420121 was selected for follow-up because it is within intron 1 of the gene glutamate receptor ionotropic kainate 1 (*GRIK1*). Glutamate receptors are the predominant excitatory neurotransmitters in the mammalian brain and play a role in short- and long-term synaptic plasticity (Headley & Grillner 1990). *GRIK1* is mainly expressed in the cerebellum and the suprachiasmatic nuclei (SCN) of the hypothalamus and its expression is developmentally regulated with a peak of expression during early postnatal development at a stage of intense synaptogenesis (Bettler & Mulle 1995). This suggests that this gene may be functionally interesting within DD and is worthy of follow-up.

The gene protein phosphatase 1, regulatory (inhibitor) subunit 12B (*PPP1R12B*) had multiple significant hits in the pooling study. The top hit, rs12743401 (P-Fisher = 3.37 x 10^{-9}) in this gene was selected for follow up with individual genotyping. Two more SNPs in this gene were in the top 200 hits: rs3817222 (P-Fisher = 6.23 x 10^{-6}, rank = 24) and rs12734338 (P-Fisher = 2 x 10^{-4}, rank = 199). It is important to note that in the HapMap CEU population, rs12734338 is in perfect LD ($r^2 = 1$, $D' = 1$) with rs12743401 and so these SNPs are likely to be picking up the same association. The top hit within the gene prostaglandin E receptor 3 (subtype EP3) (*PTGER3*) was also selected for follow up (rs6687859, P-Fisher = 1.78 x 10^{-4}, rank = 12). Another SNP in this gene, rs17131481, was in the top 20 hits (P-Fisher = 4.4 x 10^{-6}, rank = 20) and is in high LD with rs6687859 ($r^2 = 0.98$). The gene isoleucyl-tRNA synthetase 2 (*IARS2*) contains 2 SNPs within the top 270 SNPs. The top hit, rs2289191 (P-Fisher = 1.79 x 10^{-4}, rank = 171) was selected for follow up and is in high LD ($D' = 0.88$) with the other significant hit in this gene, rs17007135 (P-Fisher = 4 x 10^{-4}, rank = 269), although these SNPs are not highly correlated ($r^2 = 0.56$). The gene WD-repeat domain 78 (*WDR78*) has two SNPs within the top 350 hits: rs2454320 (P-Fisher = 6.21 x 10^{-6}, rank = 23) and rs4655653 (P-Fisher = 5.11 x 10^{-4}, rank 341). These SNPs are in perfect LD ($r^2 = 1$, D' = 1) with each other and rs4655653 was put in the panel for follow up because the assay for the more significant SNP, rs2454320, could not be designed alongside the other SNPs in the panel.

SNP	Rank	Reason for inclusion	Chr	Position (bp)	Nearest RefSeq Gene	Position Relative to Gene
rs10932727	7	Most significant in pooling study and significant in more than one pool	2	218313957	*DIRC3*	Intronic
rs10509910	9		10	111991750	*MXI1*	Intronic
rs12290752	13		11	115939639	*BUD13*	Intergenic
rs6812487	14		4	183697713	*ODZ3*	Intronic
rs2189167	31		4	104953292	*TACR3*	Intergenic
rs12344734	44		9	74383363	*TMC1*	Intronic
rs11232875	48		11	81191362	*FAM181B*	Intergenic
rs7934218	51		11	72884377	*FAM168A*	Intronic
rs17615558	65		6	12463160	*EDN1*	Intergenic
rs705790	70		6	166286499	*C6orf176*	Intronic
rs7904542	75		10	95245571	*CEP55*	Upstream
rs16932422	87		8	67103552	*DNAJC5B*	Intronic
rs12352208	100		9	14563137	*ZDHHC21*	Intergenic
rs10821663	103		10	61480286	*ANK3*	Intronic
rs4436151	105		8	114902482	*CSMD3*	Intergenic
rs4510693	109		6	156270620	*NOX3*	Intergenic
rs2311445	135		16	17456460	*XYLT1*	Intronic
rs9916926	138		18	12908318	*SEH1L*	Intergenic
rs16900429	140		8	90845430	*RIPK2*	Intronic
rs2817764	143		6	111087345	*CDK19*	Intronic
rs7381	159		22	44375446	*FBLN1*	3' UTR
rs9535442	172		13	49821641	*FAM10A4*	Intergenic
rs1546929	184		6	81104278	*BCKDHB*	Intronic
rs9397276	185		6	156298468	*NOX3*	Intergenic
rs4330611	220		7	94109934	*SGCE*	Intronic
rs9324005	246		14	98565680	*BCL11B*	Intergenic
rs3736403	302		2	219613491	*CCDC108*	Exonic
rs7686728	372		4	184541581	*CDKN2AIP*	Upstream
rs34871518	401		19	63046077	*ZNF587*	Upstream
rs5063	178	Functional / In pathway of interest	1	11830235	*NPPA*	Exonic
rs945386	316		9	138813417	*KIAA1984*	Exonic
rs420121	79		21	30068479	*GRIK1*	Intronic
rs12743401	2	In gene with multiple hits	1	200743271	*PPP1R12B*	Intronic
rs6687859	12		1	71135175	*PTGER3*	Intronic
rs2289191	171		1	218366658	*IARS2*	Intronic
rs4655653	341		1	67104024	*WDR78*	Intronic
rs1569012	27	Significant in pooling and GWAS studies	14	80923160	*STON2*	Intronic
rs1350317	53		4	183649628	*ODZ3*	Intronic
rs268598	173		8	71677425	*TRAM1*	Intronic
rs1581413	217		3	158532859	*VEPH1*	Intronic

Table 5.4: Table showing the position (based on NCBI b36) of the SNPs chosen for individual genotyping, their neareast gene (in RefSeq) and their location relative to this gene. Chr – chromosome; UTR – un-translated region.

					UK Pool		Finnish Pool		Central Europe Pool		All Samples	GWAS	
SNP	Rank	Reason for inclusion	Chr	Pos (bp)	P	OR	P	OR	P	OR	P-Fisher	P-add	OR
rs10932727	7	Most significant in pooling study and significant in more than one pool	2	218313957	5.89 x 10^{-4}	1.97	0.0749	1.42	6.51 x 10^{-5}	1.62	6.15 x 10^{-7}	0.290	1.13
rs10509910	9		10	111991750	0.0024	1.89	0.0033	1.73	4.44 x 10^{-4}	1.56	7.13 x 10^{-7}	0.930	1.01
rs12290752	13		11	115939639	1.60 x 10^{-6}	2.62	0.0567	1.53	0.156	1.27	2.58 x 10^{-6}		
rs6812487	14		4	183697713	0.216	1.29	3.30 x 10^{-6}	2.23	0.0200	1.36	2.61 x 10^{-6}		
rs2189167	31		4	104953292	0.0216	0.59	7.45 x 10^{-5}	0.54	0.0313	0.82	8.01 x 10^{-6}		
rs12344734	44		9	74383363	0.0012	1.65	0.0637	1.33	0.0015	1.36	1.61 x 10^{-5}		
rs11232875	48		11	81191362	0.882	0.97	7.17 x 10^{-4}	0.45	2.38 x 10^{-4}	0.60	2.11 x 10^{-5}		
rs7934218	51		11	72884377	0.664	0.93	5.32 x 10^{-6}	0.52	0.0485	0.80	2.36 x 10^{-5}		
rs17615558	65		6	12463160	0.0655	1.37	7.05 x 10^{-5}	1.97	0.0629	1.32	3.76 x 10^{-5}		
rs705790	70		6	166286499	0.0798	1.51	0.0403	1.38	9.80 x 10^{-5}	1.73	4.04 x 10^{-5}		
rs7904542	75		10	95245571	9.19 x 10^{-4}	0.62	0.00145	0.67	0.271	0.91	4.55 x 10^{-5}		
rs16932422	87		8	67103552	0.332	1.22	1.03 x 10^{-4}	2.18	0.0175	1.41	7.09 x 10^{-5}	0.484	1.11
rs12352208	100		9	14563137	0.414	1.15	2.21 x 10^{-4}	1.79	0.0075	1.35	7.89 x 10^{-5}	0.813	1.03
rs10821663	103		10	61480286	2.65 x 10^{-4}	0.53	0.00450	0.65	0.617	0.94	8.46 x 10^{-5}	0.079	0.85
rs4436151	105		8	114902482	0.438	0.90	0.00907	0.72	1.93 x 10^{-4}	0.72	8.76 x 10^{-5}		
rs4510693	109		6	156270620	0.695	0.95	1.69 x 10^{-4}	0.60	0.0072	0.75	9.50 x 10^{-5}	0.961	1.00
rs2311445	135		16	17456460	0.190	0.81	3.43 x 10^{-4}	0.61	0.0177	0.78	1.25 x 10^{-4}	0.974	1.00
rs9916926	138		18	12908318	4.12 x 10^{-4}	1.68	0.0217	1.34	0.131	1.19	1.27 x 10^{-4}	0.073	1.15
rs16900429	140		8	90845430	0.952	0.99	4.29 x 10^{-4}	0.62	0.0030	0.76	1.31 x 10^{-4}	*0.508*	*0.91*
rs2817764	143		6	111087345	0.989	1.00	0.00231	1.55	5.48 x 10^{-4}	1.37	1.34 x 10^{-4}	*0.414*	*1.08*
rs7381	159		22	44375446	0.0303	1.43	0.0563	1.42	9.03 x 10^{-4}	1.44	1.60 x 10^{-4}	0.426	1.11
rs9535442	172		13	49821641	0.0094	1.67	0.248	1.22	7.55 x 10^{-4}	1.49	1.80 x 10^{-4}	0.981	1.00
rs1546929	184		6	81104278	0.953	0.99	7.11 x 10^{-5}	0.52	0.0298	0.74	2.02 x 10^{-4}	0.780	0.98
rs9397276	185		6	156298468	0.801	1.03	2.53 x 10^{-4}	1.58	0.0101	1.25	2.04 x 10^{-4}	*0.734*	*1.03*
rs4330611	220		7	94109934	1.63 x 10^{-4}	0.49	0.294	0.87	0.0591	0.82	2.70 x 10^{-4}	0.790	0.97
rs9324005	246		14	98565680	2.10 x 10^{-4}	2.15	0.246	1.26	0.0675	1.26	3.22 x 10^{-4}		
rs3736403	302		2	219613491	0.0761	0.76	7.38 x 10^{-4}	0.61	0.0851	0.84	4.22 x 10^{-4}	*0.460*	*0.92*

Table 5.5: SNPs in follow-up panel for individual genotyping. Results of imputed SNPs are presented in italics.

					UK Pools		Finnish Pools		German Pools		All Samples	GWAS	
SNP	Rank	Reason for inclusion	Chr	Pos (bp)	P	OR	P	OR	P	OR	P-Fisher	P-add	OR
rs7686728	372	Most significant in pooling study and significant in more than one pool	4	184541581	1.66×10^{-4}	1.79	0.0676	1.30	0.617	1.06	5.77×10^{-4}	*0.189*	*0.84*
rs34871518	401		19	63046077	0.886	1.02	5.16×10^{-4}	1.66	0.0173	1.26	6.46×10^{-4}		
rs5063	178	Functional / In pathway of interest	1	11830235	0.249	1.23	4.11×10^{-4}	2.09	0.0182	1.40	1.89×10^{-4}		
rs945386	316		9	138813417	0.0711	0.78	0.0011	0.63	0.0663	0.82	4.57×10^{-4}	0.963	1.00
rs420121*	79		21	30068479	0.0051	1.42	0.167	1.19	0.0173	1.33	5.20×10^{-4}	0.371	1.06
rs12743401	2	In gene with multiple hits	1	200743271	3.70×10^{-5}	0.57	0.225	0.86	1.16×10^{-6}	0.63	3.37×10^{-9}		
rs6687859	12		1	71135175	0.0498	1.31	4.91×10^{-5}	1.87	0.0038	1.32	1.78×10^{-6}		
rs2289191	171		1	218366658	0.0649	0.71	5.98×10^{-5}	0.41	0.451	0.87	1.79×10^{-4}	0.654	0.95
rs4655653	341		1	67104024	0.0152	1.44	5.54×10^{-4}	1.61	0.711	1.04	5.11×10^{-4}		
rs1569012	27	Significant in pooling and GWAS studies	14	80923160	0.426	1.19	4.90×10^{-4}	1.81	1.98×10^{-4}	1.60	6.71×10^{-6}	0.027	1.27
rs1350317	53		4	183649628	0.0285	0.51	2.09×10^{-4}	0.47	0.0303	0.73	2.48×10^{-5}	0.049	0.78
rs268598	173		8	71677425	0.0264	1.58	0.130	1.36	5.16×10^{-4}	1.73	1.81×10^{-4}	0.028	1.32
rs1581413	217		3	158532859	0.878	0.98	9.40×10^{-5}	0.60	0.0332	0.82	2.63×10^{-4}	0.006	0.83

Table 5.5 Continued. *rs420121 genotyped via proxy SNP rs461119.

5.3.2.1 Individual Genotyping of Samples Included in Pools

The panel of follow-up SNPs were genotyped in all of those samples that had been pooled, except for the Finnish pool (NeuroDys individual genotyping sample). They were also genotyped with some additional samples (whole NeuroDys individual genotyping sample) to ascertain if any SNPs showed a higher level of significance when they were genotyped in a larger sample (see next section of this Chapter).

Out of the NeuroDys individual genotyping sample, a total of 988 cases and 1121 controls passed QC when they were genotyped individually. As shown in Table C.5 in the Appendix, two SNPs in the panel (rs12743401 and rs4510693) failed optimisation. All remaining SNPs had a call rate > 70% in both the UK and Central European subsets and in the entire sample. 2 SNPs had MAFs < 0.05 but were still included in the analysis (rs12290752, MAF = 0.04 in UK subset, 0.02 in Central European subset and 0.03 in entire sample; rs9324005, MAF = 0.04 in UK sample, 0.049 in Central European subset and 0.046 in the whole sample). Nine of the SNPs were out of Hardy-Weinberg equilibrium in the control sample. These were not excluded from the association analyses but any association found with these SNPs should be treated with caution.

The results from the individual genotyping of the NeuroDys individual genotyping sample are shown in Table 5.6, in order of their additive P-value in the complete sample. A total of 14 SNPs gave significant P-values (<0.05) in the complete sample, with the top hit being rs2189167 (P-add = 5.0×10^{-5}, OR = 1.41, P-gen = 2.7×10^{-4}). This SNP was also the most significant hit in the Central European subset (P-add = 6×10^{-4}, OR = 1.37) but was not quite as significant in the UK subset despite having a larger effect size (P-add = 0.0264, OR = 1.60), which may have been due to a lack of power in this smaller sample. However, this SNP showed highly significant departure from Hardy-Weinberg equilibrium and had relatively low call rates, suggesting that this may have been a poor assay, even though it passed optimisation and QC filters.

The next most significant SNP in the NeuroDys individual genotyping sample was rs7381 (P-add = 8.1×10^{-4}, OR = 1.58, P-gen = 0.0118). This SNP was also significant in the Central European subset alone (P-add = 0.0037, OR = 1.58) but was not significantly associated in the UK subset (P-add = 0.0964, OR = 1.59). Apart from the top hit rs2189167 (which, as indicated previously, should be interpreted with caution), two SNPs showed significant association with DD in both the UK subset and the

Central European subset. The most significant of these is rs10932727 (UK subset: P-add = 0.0304, OR = 2.04; Central European subset: P-add = 0.0135, OR = 1.45) and the other SNP is rs461119 (UK subset: P-add = 0.0177, OR = 1.5; Central European subset: P-add = 0.0191, OR = 1.22). Both of these SNPs had relatively high effect sizes in the UK sample, which may be explain why these were significant in both sample subsets whereas other SNPs with lower effect sizes were only significant in the larger Central European subset. Overall, the SNPs selected for individual genotyping did not reach the level of significant association with DD that was found in the pooling study. In addition, the four SNPs that had shown a significant association in both the initial NeuroDys GWAS and in the pooling study were not significantly associated when genotyped in this sample (rs1350317, P-add = 0.2119; rs1569012, P-add = 0.2937, rs268598, P-add = 0.4003, rs1581413, P-add = 0.601).

	Pooling Study	NeuroDys Individual Genotyping Sample											
		UK Subset				Central European Subset				Complete Sample			
SNP	**P-Fisher**	**P-gen**	**P-add**	**OR**	**95% CI**	**P-gen**	**P-add**	**OR**	**95% CI**	**P-gen**	**P-add**	**OR**	**95% CI**
rs2189167	**8.0 x 10^{-6}**	**0.0248**	**0.0264**	1.60	1.06-2.43	**0.0026**	**6.0 x 10^{-4}**	1.37	1.15-1.64	**2.7 x 10^{-4}**	**5.0 x 10^{-5}**	1.41	1.19-1.66
rs7381	**1.6 x 10^{-4}**	0.4832	0.0964	1.59	0.92-2.74	**0.0236**	**0.0037**	1.58	1.16-2.14	**0.0118**	**8.1 x 10^{-4}**	1.58	1.21-2.06
rs10932727	**6.2 x 10^{-7}**	0.0641	**0.0304**	2.04	1.07-3.90	**0.0260**	**0.0135**	1.45	1.08-1.94	**0.0028**	**0.0015**	1.54	1.18-2.01
rs461119	**5.2 x 10^{-4}**	**0.0161**	**0.0177**	1.50	1.07-2.08	0.0621	**0.0191**	1.22	1.03-1.44	**0.0041**	**0.0015**	1.27	1.10-1.47
rs6687859	**1.8 x 10^{-6}**	0.2108	0.0778	1.36	0.97-1.92	0.0621	**0.0224**	1.24	1.03-1.49	**0.0149**	**0.0044**	1.27	1.08-1.49
rs4436151	**8.8 x 10^{-5}**	0.0836	0.1789	0.78	0.53-1.12	**0.0153**	**0.0112**	0.76	0.62-0.94	**0.0021**	**0.0041**	0.77	0.64-0.92
rs12344734	**1.6 x 10^{-5}**	0.1524	0.1765	1.44	0.85-2.46	0.0861	**0.0268**	1.31	1.03-1.66	**0.0292**	**0.0100**	1.33	1.07-1.65
rs10509910	**7.1 x 10^{-7}**	0.2780	0.1229	1.48	0.90-2.43	0.1033	0.0527	1.25	1.00-1.57	**0.0324**	**0.0160**	1.29	1.05-1.58
rs5063	**1.9 x 10^{-4}**	0.9909	0.4463	1.27	0.69-2.32	**0.0465**	**0.0135**	1.58	1.10-2.28	**0.0387**	**0.0122**	1.49	1.09-2.04
rs7686728	**5.8 x 10^{-4}**	**0.0307**	**0.0317**	0.63	0.41-0.96	0.2798	0.1823	0.86	0.68-1.08	0.0868	**0.0270**	0.80	0.65-0.97
rs4330611	**2.7 x 10^{-4}**	**0.0324**	**0.0099**	0.56	0.36-0.87	0.1402	0.2213	0.87	0.70-1.09	**0.0181**	**0.0232**	0.80	0.66-0.97
rs16900429	**1.3 x 10^{-4}**	0.1400	0.6929	1.09	0.71-1.68	0.0634	**0.0267**	1.32	1.03-1.68	**0.0249**	**0.0343**	1.26	1.02-1.56
rs11232875	**2.1 x 10^{-4}**	0.7986	0.5251	1.17	0.72-1.89	0.1349	**0.0453**	1.31	1.01-1.70	0.1199	**0.0396**	1.27	1.01-1.61
rs2311445	**1.3 x 10^{-4}**	0.6025	0.6629	0.90	0.57-1.43	0.1326	**0.0497**	0.78	0.61-1.00	0.1170	0.0518	0.81	0.65-1.00
rs9397276	**2.0 x 10^{-4}**	0.8814	0.6265	0.93	0.69-1.25	**0.0036**	**0.0133**	1.23	1.05-1.46	**0.0249**	0.0539	1.15	1.00-1.33
rs945386	**4.6 x 10^{-4}**	0.1365	0.0815	1.35	0.96-1.90	0.5403	0.2672	1.12	0.92-1.35	0.1529	0.0666	1.17	0.99-1.38
rs9535442	**1.8 x 10^{-4}**	0.9891	0.8825	1.05	0.54-2.06	0.0728	0.0679	1.30	0.98-1.73	0.0760	0.0827	1.26	0.97-1.64
rs9916926	**1.3 x 10^{-4}**	**0.0191**	**0.0048**	0.62	0.44-0.86	0.3259	0.7401	0.97	0.80-1.17	0.1320	0.0893	0.87	0.74-1.02
rs1546929	**2.0 x 10^{-4}**	0.6287	0.4600	1.16	0.78-1.74	0.1670	0.1238	1.17	0.96-1.43	0.1778	0.0878	1.17	0.98-1.40
rs9324005	**3.2 x 10^{-4}**	0.4281	0.1281	0.59	0.29-1.17	0.4765	0.2388	0.80	0.55-1.16	0.2091	0.0768	0.74	0.54-1.03
rs4655653	**5.1 x 10^{-4}**	**0.0192**	**0.0049**	1.72	1.18-2.52	0.5255	0.8614	1.02	0.83-1.24	0.1514	0.1330	1.14	0.96-1.36
rs705790	**4.0 x 10^{-5}**	0.1662	0.3653	1.32	0.72-2.41	0.4225	0.2012	1.22	0.90-1.65	0.1648	0.1209	1.24	0.95-1.62

Table 5.6: Results of individual genotyping of the NeuroDys Individual Genotyping Sample, sorted by the additive P-value in the complete sample. OR – odds ratio; CI – confidence interval; P-add – P-value from additive test; P-gen – P-value from genotypic test. Significant P-values (<0.05) are in bold. N.B the most significant SNP, rs2189167, is not in Hardy Weinberg equilibrium.

	Pooling Study	NeuroDys Individual Genotyping Sample											
		UK Subset				Central European Subset				All Pooled Samples			
SNP	**P-Fisher**	**P-gen**	**P-add**	**OR**	**95% CI**	**P-gen**	**P-add**	**OR**	**95% CI**	**P-gen**	**P-add**	**OR**	**95% CI**
rs34871518	**6.5 x 10^{-4}**	0.3811	0.643	1.10	0.73-1.66	0.1527	0.1014	1.19	0.97-1.46	0.2302	0.0945	1.17	0.97-1.40
rs12290752	**2.6 x 10^{-6}**	0.3239	0.0751	2.27	0.92-5.61	0.7932	0.7181	1.10	0.66-1.83	0.5717	0.2058	1.32	0.86-2.03
rs2817764	**1.3 x 10^{-4}**	0.3428	0.5701	0.90	0.62-1.30	0.1269	0.0614	1.21	0.99-1.48	0.1686	0.1726	1.13	0.95-1.35
rs12352208	**7.9 x 10^{-5}**	0.4306	0.2164	1.42	0.81-2.49	0.3664	0.3316	1.16	0.86-1.57	0.2079	0.1476	1.22	0.93-1.59
rs1350317	**2.5 x 10^{-5}**	0.8076	0.5133	1.25	0.64-2.47	0.3043	0.2789	1.16	0.89-1.52	0.2462	0.2119	1.17	0.91-1.51
rs7934218	**2.4 x 10^{-5}**	0.8850	0.9777	1.01	0.64-1.60	**0.0382**	0.2333	0.86	0.67-1.10	0.1158	0.3000	0.89	0.71-1.11
rs1569012	**6.7 x 10^{-6}**	0.3801	0.1640	1.47	1.86-2.51	0.1359	0.6178	1.07	0.83-1.37	0.1261	0.2937	1.13	0.90-1.41
rs7904542	**4.5 x 10^{-5}**	0.2825	0.1116	0.69	0.44-1.09	0.1792	0.7649	0.97	0.77-1.21	0.1779	0.3168	0.90	0.74-1.10
rs3736403	**4.2 x 10^{-4}**	0.1308	0.1294	1.45	0.90-2.34	0.9061	0.6966	1.05	0.82-1.35	0.4735	0.2913	1.13	0.90-1.40
rs268598	**1.8 x 10^{-4}**	0.7333	0.9787	0.99	0.55-1.78	0.5302	0.3410	0.85	0.61-1.19	0.2864	0.4003	0.88	0.66-1.18
rs10821663	**8.5 x 10^{-5}**	0.4309	0.5564	0.90	0.64-1.27	0.7049	0.6670	0.96	0.78-1.17	0.4172	0.5015	0.94	0.79-1.12
rs1581413	**2.6 x 10^{-4}**	0.4685	0.9188	1.02	0.77-1.35	0.1032	0.5254	0.95	0.82-1.11	0.0506	0.6101	0.97	0.84-1.11
rs2289191	**1.8 x 10^{-4}**	0.5651	0.2854	1.33	0.79-2.24	0.5179	0.9949	1.00	0.72-1.39	0.5164	0.5682	1.08	0.82-1.43
rs16932422	**7.1 x 10^{-5}**	0.7968	0.5002	1.27	0.63-2.55	0.8905	0.8762	1.03	0.73-1.45	0.7907	0.6621	1.07	0.79-1.45
rs17615558	**3.8 x 10^{-5}**	0.5555	0.6520	1.11	0.70-1.77	0.8523	0.5729	0.92	0.69-1.23	0.7776	0.8053	0.97	0.76-1.24
rs6812487	**2.6 x 10^{-6}**	0.5547	0.7267	0.90	0.48-1.66	0.5226	0.8282	1.03	0.77-1.39	0.2052	0.9641	1.01	0.77-1.31

Table 5.6 continued.

5.3.2.2 Genotyping of Whole NeuroDys Individual Genotyping Sample

When genotyping the whole NeuroDys individual genotyping sample, a total of 1518 cases and 2261 controls passed QC. The results of the QC tests on the panel of follow up SNPs when the additional samples are included are shown in Table C.6 in the Appendix. As stated previously, 2 SNPs failed optimisation (rs12743401 and rs4510693). All remaining SNPs had a call rate > 70% in each individual sample group and in the whole sample combined. As with the NeuroDys individual genotyping sample, nine SNPs were out of Hardy-Weinberg equilibrium when the additional samples were included. As before, these were not excluded from the association analyses at this stage but any association found with these SNPs should be treated with caution.

The results from the individual genotyping when including the additional samples are shown in Table 5.7, in order of their additive P-value in the whole sample. With the additional samples, 9 SNPs gave significant P-values (< 0.05), with the top hit now being rs461119 (P-add = 3.0 x 10^{-4}, OR = 1.23). This SNP was included in the panel of follow-up SNPs as a proxy for the SNP rs420121, which had a Fisher P-value of 5.2 x 10^{-4} in the NeuroDys pooling study. Both of these SNPs are within the first intron of *GRIK1*, as shown in Figure 5.3. As mentioned previously, the SNP in this gene was selected for individual genotyping because *GRIK1* encodes a glutamate receptor, and these receptors are the predominant excitatory neurotransmitters in the mammalian brain (Headley & Grillner 1990). The SNP rs461119 showed a higher level of significance when the NeuroDys additional samples were included, compared to the NeuroDys individual genotyping sample alone (P-add = 0.0015, OR = 1.27).

Another 5 SNPs showed a higher level of significance when including the NeuroDys additional samples. Of these, the next most significant SNP was rs12344734 (P-add = 0.0021, OR = 1.25). This SNP had a P-add value of 0.01 when genotyped in the NeuroDys individual genotyping sample alone (OR = 1.33). This SNP is in the transmembrane channel-like 1 gene (*TMC1*) on chromosome 9 as shown in Figure 5.4. The specific function of this gene in unknown but mutations in this gene have been associated with progressive postlingual hearing loss and profound prelingual deafness (Kurima et al. 2002; Meyer et al. 2005; Santos et al. 2005; Kitajiri et al. 2007b; Tlili et al. 2008; Kitajiri et al. 2007a; Kalay et al. 2005; Vreugde et al. 2002).

The next most significant SNP was rs16900429 (NeuroDys individual genotyping sample: P-add = 0.034, OR = 1.26; whole NeuroDys individual genotyping sample: P-add = 0.009, OR = 1.21). This SNP is within the gene receptor-interaction serine-threonine kinase 2 (*RIPK2*) on chromosome 8, as shown in Figure 5.5. *RIPK2* encodes a component of signalling complexes in both the innate and adaptive immunity pathways and overexpression of this gene can lead to cell death (McCarthy et al. 1998).

The fourth SNP to show an increase in significance was rs4655653 (NeuroDys individual genotyping sample: P-add = 0.133, OR = 1.14; whole NeuroDys individual genotyping sample: P-add = 0.015, OR = 1.15) which is an intronic SNP within the gene (*WDR78*), as shown Figure 5.6. The exact function of this gene is unknown, but the family of WD-repeat domain containing proteins play key roles in the formation of protein-protein complexes and are critical for a wide range of biological functions including transduction, transcription regulation, cytoskeletal assembly and apoptosis (Li & Roberts 2001; Smith 2008). As mentioned previously, this SNP was selected for follow up as this gene had two significant SNPs in the top 350 hits (rs4655653 and rs2454320, as shown in Figure 5.6).

The fifth SNP that showed an increase in significance was rs7686728 (NeuroDys individual genotyping sample: P-add = 0.027, OR = 0.80; whole NeuroDys individual genotyping sample: P-add = 0.0201, OR = 0.84). This SNP is in an intergenic region on chromosome 4, 61kb upstream from the *CDKN2*A interacting protein gene (*CDKN2AIP*) which is thought to be involved the negative regulation of cell growth and the regulation of protein stability.

The final SNP was rs1581413 (NeuroDys individual genotyping sample: P-gen = 0.051; whole NeuroDys individual genotyping sample: P-gen = 0.005, OR = 0.84) within intron 9 of the ventricular zone expressed PH domain homolog 1 gene (*VEPH1*) on chromosome 3, as shown in Figure 5.7. *VEPH1* is a homolog of a zebrafish gene but its function is unknown.

The majority of the other SNPs showed a decreased level of significance when the additional samples were combined with those that had been in the NeuroDys individual genotyping sample. For example, the top hit in the NeuroDys individual genotyping sample (rs2189167, P-add = 5 x 10^{-5}, OR = 1.41) was no longer significantly associated with DD when the additional samples were also analysed (P-add = 0.0602, OR = 1.12).

This suggests that the association identified in the in the smaller sample may have been a false positive finding.

Only one SNP showed significant association in more than one sample group (rs10932727, P-gen in UK sample = 0.0022, P-add in Central European sample = 0.0267). This SNP is within intron 1 of the disrupted in the gene renal carcinoma 3 (*DIRC3*) on chromosome 2. As the name suggests, this gene has been associated with renal cell cancer as it spans a breakpoint that was identified in a family with renal cancer (Bodmer et al. 2003). It shows low expression in fetal and adult tissues, with the highest level of expression in the placenta (Bodmer et al. 2003). Very little is known about the function of this gene as yet.

	Pooling Study	Whole NeuroDys Individual Genotyping Sample																	
		UK Subset			Central European Subset			French Subset			Hungarian Subset			All Samples			
SNP	P-Fisher	P-gen	P-add	OR	P-gen	P-add	OR	P-gen	P-add	OR	P-gen	P-add	OR	P-gen	P-add	OR	95% CI
rs461119	**5.2 x 10^{-4}**	**3.7 x 10^{-4}**	**1.2 x 10^{-4}**	1.45	0.2523	0.1104	1.12	0.6884	0.5628	1.10	0.8003	0.5194	1.15	**0.0010**	**3.0 x 10^{-4}**	1.23	1.11-1.36
rs12344734	**1.6 x 10^{-6}**	0.1143	**0.0414**	1.35	**0.0341**	**0.0146**	1.29	0.8169	0.5943	0.87	0.2416	0.1543	1.51	**0.0080**	**0.0021**	1.25	1.08-1.46
rs16900429	**1.3 x 10^{-4}**	0.2548	0.5324	1.09	**0.0314**	**0.0137**	1.30	0.5779	0.4451	1.18	0.1943	0.2653	1.51	**0.0115**	**0.0085**	1.21	1.04-1.40
rs7381	**1.6 x 10^{-4}**	0.0958	0.2995	1.19	**0.0157**	**0.0042**	1.47	0.9985	0.9566	1.02	0.9163	0.6761	0.83	**0.0205**	**0.0138**	1.24	1.03-1.49
rs4655653	**5.1 x 10^{-4}**	**0.0014**	**0.0031**	1.39	0.5486	0.5678	1.05	0.2966	0.2344	1.26	0.8795	0.7928	0.94	**0.0428**	**0.0149**	1.15	1.02-1.29
rs7686728	**5.8 x 10^{-4}**	**0.0365**	**0.0157**	0.73	0.3267	0.2070	0.88	0.3029	0.1433	0.71	0.3805	0.2193	1.39	0.0670	**0.0201**	0.84	0.73-0.96
rs10932727	**6.1 x 10^{-7}**	**0.0022**	0.2451	1.18	0.0528	**0.0267**	1.33	0.9981	0.8013	1.08	0.2902	0.3007	0.66	**0.0041**	**0.0312**	1.20	1.01-1.42
rs4436151	**8.8 x 10^{-5}**	**0.0039**	**0.0115**	0.75	0.2565	0.1554	0.88	0.1749	0.0676	1.44	0.1661	0.1112	0.61	**0.0145**	**0.0353**	0.86	0.76-0.97
rs705790	**4.0 x 10^{-5}**	**0.0398**	0.2042	1.28	0.1751	0.0664	1.27	0.9678	0.9474	0.98	0.9970	0.8694	0.94	0.0732	0.0506	1.17	0.97-1.42
rs9916926	**1.3 x 10^{-4}**	0.0560	**0.0166**	0.79	0.6787	0.5016	0.95	0.7258	0.8751	1.03	0.9708	0.9449	0.98	0.1469	0.0545	0.90	0.81-1.00
rs2189167	**8.0 x 10^{-6}**	0.2732	0.9424	1.01	**0.0109**	**0.0073**	1.24	0.1210	0.8730	1.03	0.0558	0.7589	0.94	0.1534	0.0602	1.12	1.01-1.25
rs1581413	**2.6 x 10^{-4}**	0.3887	0.1806	0.89	**0.0059**	0.4653	0.95	0.1597	0.0566	0.75	0.5633	0.6907	1.08	**0.0054**	0.0707	0.92	0.84-1.01
rs9397276	**2.0 x 10^{-4}**	**0.0204**	0.0925	1.16	0.0770	0.0762	1.13	0.2644	0.2271	0.82	0.5415	0.8388	0.96	0.2071	0.0766	1.10	0.99-1.21
rs9535442	**1.8 x 10^{-4}**	0.3611	0.1722	1.27	0.2808	0.1653	1.19	0.3284	0.3981	0.79	0.7298	0.4692	1.26	0.0719	0.0846	1.19	1.01-1.42
rs1350317	**2.5 x 10^{-5}**	0.8671	0.7541	0.95	0.1515	0.1306	1.19	0.4898	0.3610	1.25	0.3072	0.1682	1.49	0.0748	0.0987	1.15	0.98-1.36
rs11232875	**2.1 x 10^{-5}**	0.6577	0.5954	1.08	0.4106	0.2982	1.13	0.4108	0.2901	1.30	0.5197	0.5507	1.23	0.2476	0.1088	1.09	0.93-1.27
rs2311445	**1.2 x 10^{-4}**	0.3422	0.9482	1.01	0.2405	0.1510	0.86	0.2576	0.1053	0.65	0.8689	0.8857	0.96	0.1328	0.1239	0.87	0.75-1.01
rs34871518	**6.5 x 10^{-4}**	0.6884	0.4505	1.10	0.1667	0.0673	1.18	0.9975	0.9825	1.00	0.8957	0.7001	0.91	0.2009	0.1324	1.13	1.00-1.27
rs945386	**4.6 x 10^{-4}**	0.5440	0.4720	1.08	0.9287	0.7503	1.03	**0.0350**	**0.0106**	1.59	0.9274	0.9936	1.00	0.3236	0.1361	1.08	0.97-1.21
rs3736403	**4.2 x 10^{-4}**	0.1211	0.1213	1.26	0.2032	0.1004	1.19	0.9280	0.9338	0.98	0.0907	0.7013	0.88	0.2939	0.1508	1.18	1.02-1.37
rs5063	**1.9 x 10^{-4}**	0.8405	0.6075	1.10	0.1073	0.0600	1.35	0.8294	0.3849	0.73	0.9516	0.7530	0.84	0.2852	0.1849	1.13	0.92-1.40
rs6687859	**1.8 x 10^{-6}**	0.6255	0.3869	1.09	0.4488	0.2075	1.11	0.6548	0.6243	0.91	0.9737	0.9637	1.01	0.4219	0.1904	1.06	0.95-1.18
rs17615558	**3.8 x 10^{-5}**	0.6674	0.4254	0.89	0.8104	0.5189	0.92	0.8958	0.3830	0.79	0.3124	0.7763	0.91	0.4631	0.2285	0.90	0.76-1.06
rs7934218	**2.4 x 10^{-5}**	0.3898	0.1865	0.83	0.5871	0.7250	0.96	0.4680	0.3505	1.26	0.5030	0.1834	0.60	0.5298	0.2770	0.93	0.80-1.08
rs7904542	**4.5 x 10^{-5}**	0.2934	0.1178	0.80	**0.0233**	0.7772	0.97	0.9895	0.8845	0.96	0.9853	0.8807	1.05	0.0909	0.2932	0.95	0.83-1.10
rs16932422	**7.1 x 10^{-5}**	0.8496	0.2704	1.24	0.8614	0.5905	1.08	0.9914	0.8911	0.96	0.9201	0.9299	1.04	0.6763	0.3782	1.12	0.91-1.37

Table 5.7: Results of individual genotyping in the Whole NeuroDys Individual Genotyping Sample within each sample subset and for all samples combined, sorted by the additive P-value for all samples combined. P-add – P-value from additive test; P-gen – P-value from genotypic test; OR – odds ratio; CI – confidence interval. Significant P-values (<0.05) are in bold.

	Pooling Study	Whole NeuroDys Individual Genotyping Sample															
		UK Subset			Central European Subset			French Subset			Hungarian Subset			All Samples			
SNP	P-Fisher	P-gen	P-add	OR	P-gen	P-add	OR	P-gen	P-add	OR	P-gen	P-add	OR	P-gen	P-add	OR	95% CI
rs1546929	**2.0×10^{-4}**	0.4825	0.4232	1.10	0.7430	0.5339	1.06	0.9014	0.6632	1.09	0.5198	0.2519	0.73	0.4720	0.4174	1.05	0.93-1.19
rs10509910	**7.1×10^{-7}**	0.4993	0.2408	1.18	0.0507	0.2013	1.13	0.3033	0.1222	0.71	0.7566	0.2941	0.72	0.1963	0.4349	1.06	0.92-1.22
rs12352208	**7.9×10^{-5}**	0.2922	0.3692	0.86	0.1809	0.2116	1.18	0.9961	0.3130	1.35	0.9803	0.8418	1.09	0.0615	0.4419	1.07	0.89-1.28
rs1569012	**6.7×10^{-6}**	0.8352	0.8951	0.98	0.1232	0.4153	1.09	0.5657	0.2866	1.33	0.6799	0.6568	0.86	0.0815	0.4527	1.11	0.95-1.30
rs9324005	**3.2×10^{-4}**	0.9998	0.8576	0.96	0.4911	0.2541	0.83	0.8630	0.8125	1.08	0.9923	0.9014	1.06	0.6427	0.4875	0.89	0.71-1.11
rs6812487	**2.6×10^{-5}**	0.2908	0.7739	0.95	0.2995	0.3526	1.12	0.5568	0.2870	1.31	0.7270	0.4260	0.72	0.2932	0.4966	1.09	0.91-1.30
rs4330611	**2.7×10^{-4}**	0.3476	0.1513	0.82	0.1272	0.7132	0.97	0.4831	0.3152	1.25	0.9970	0.9397	0.98	0.1131	0.5223	0.94	0.82-1.08
rs12290752	**2.6×10^{-6}**	0.1621	0.0568	1.55	0.9374	0.5824	0.89	0.9980	0.9499	1.03	0.2706	0.1412	0.21	0.7823	0.5576	1.06	0.80-1.40
rs2817764	**1.3×10^{-4}**	0.6148	0.3396	0.90	0.3151	0.1410	1.14	0.7131	0.4489	1.16	0.2681	0.1051	0.64	0.7994	0.5927	1.00	0.89-1.13
rs10821663	**8.5×10^{-5}**	0.4253	0.3635	0.90	0.7299	0.9301	1.01	0.4183	0.4507	0.86	0.8342	0.5578	1.16	0.6635	0.6289	0.96	0.85-1.08
rs2289191	**1.8×10^{-4}**	0.5135	0.2483	1.19	0.2495	0.8375	1.03	0.9829	0.5281	0.84	0.5019	0.2445	0.59	0.7259	0.7448	1.02	0.85-1.22
rs268598	**1.8×10^{-4}**	0.8682	0.7953	1.05	0.3995	0.5640	0.92	0.7729	0.2377	0.67	0.3160	0.1373	1.73	0.8752	0.8225	0.94	0.77-1.15

Table 5.7 continued

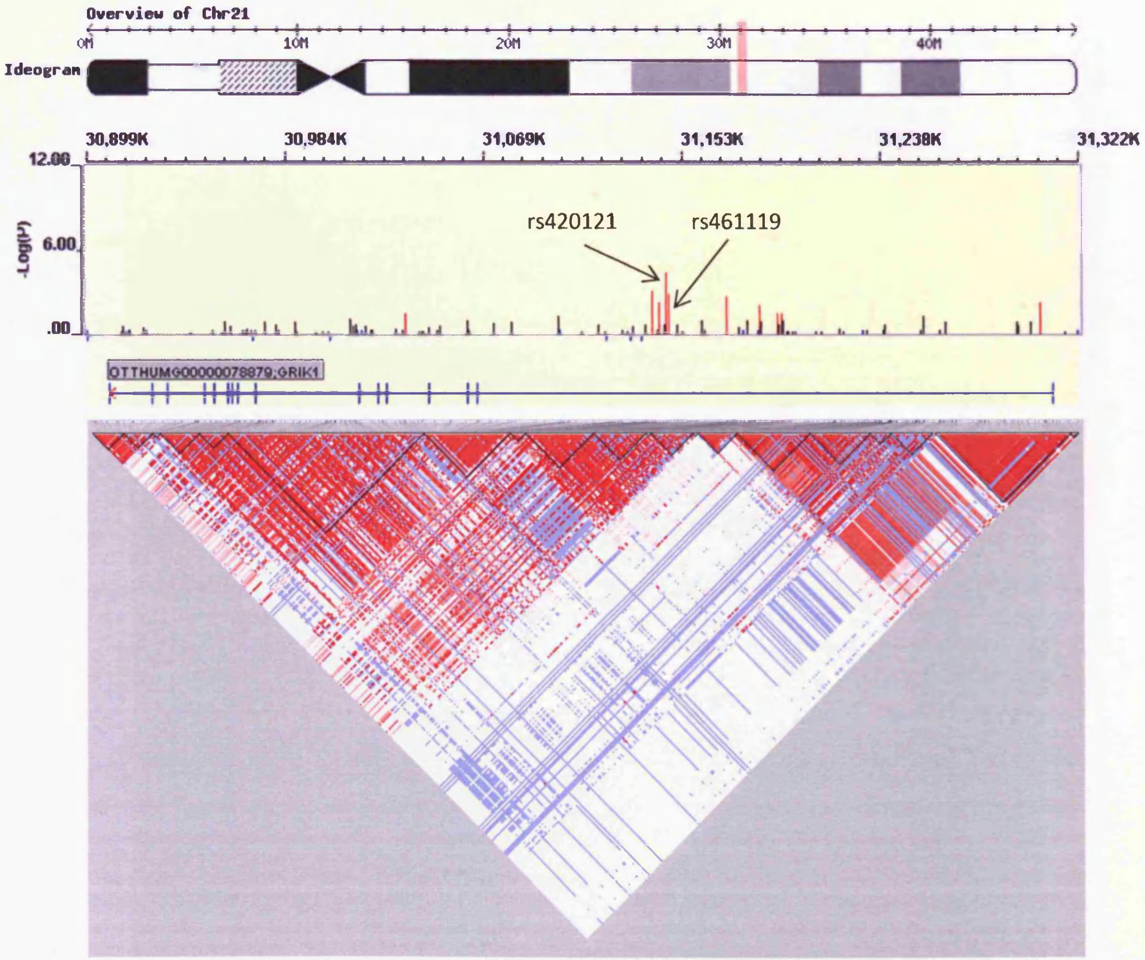


Figure 5.3: LD plot and results of SNPs (P-Fisher values) in *GRIK1* that were genotyped in the NeuroDys pooling study. The SNP rs461119 was genotyping the NeuroDys Individual Genoytping Sample as a proxy for rs420121, as these SNPs are in perfect LD with each other ($r^2 = 1$). Red bars denote P-values <0.05.

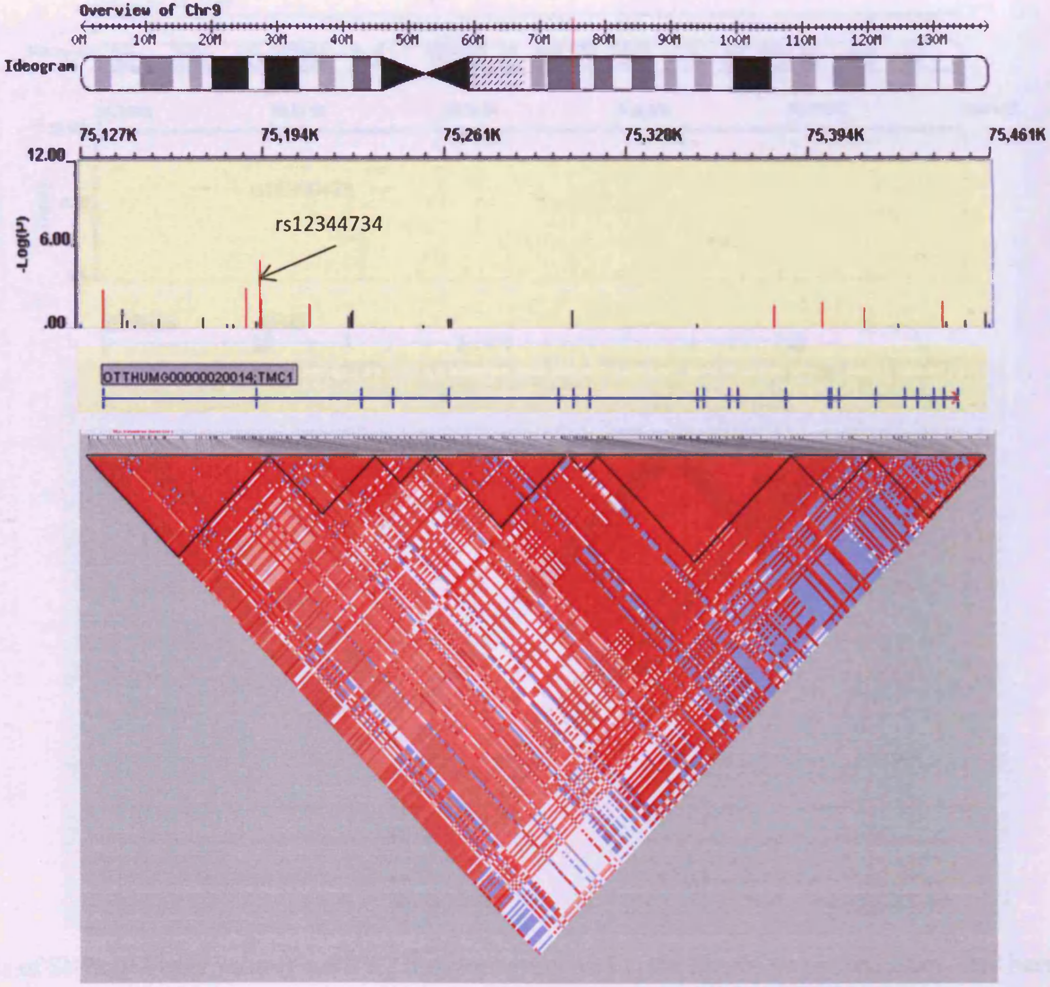


Figure 5.4: LD plot and results of SNPs (P-Fisher values) in *TMC1* that were genotyped in the NeuroDys pooling study. Red bars denote P-values <0.05.

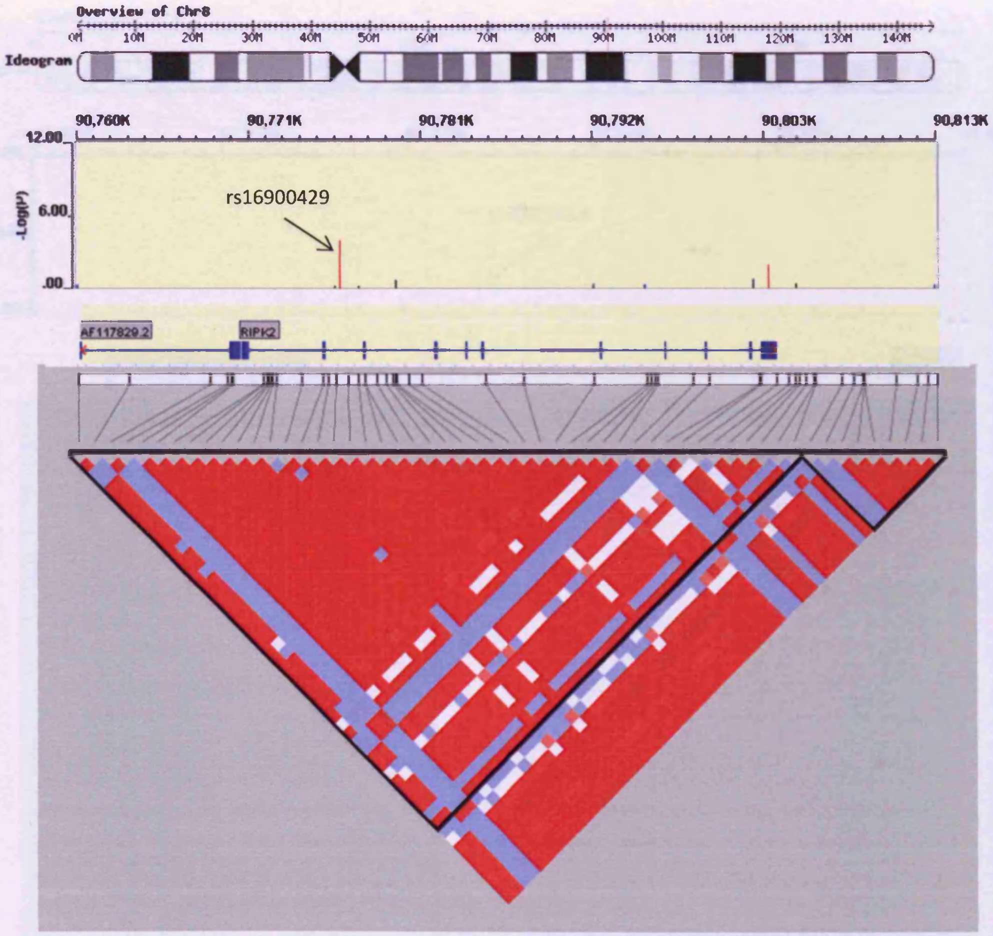


Figure 5.5: LD plot and results of SNPs (P-Fisher values) in *RIPK2* that were genotyped in the NeuroDys pooling study. Red bars denote P-values <0.05.

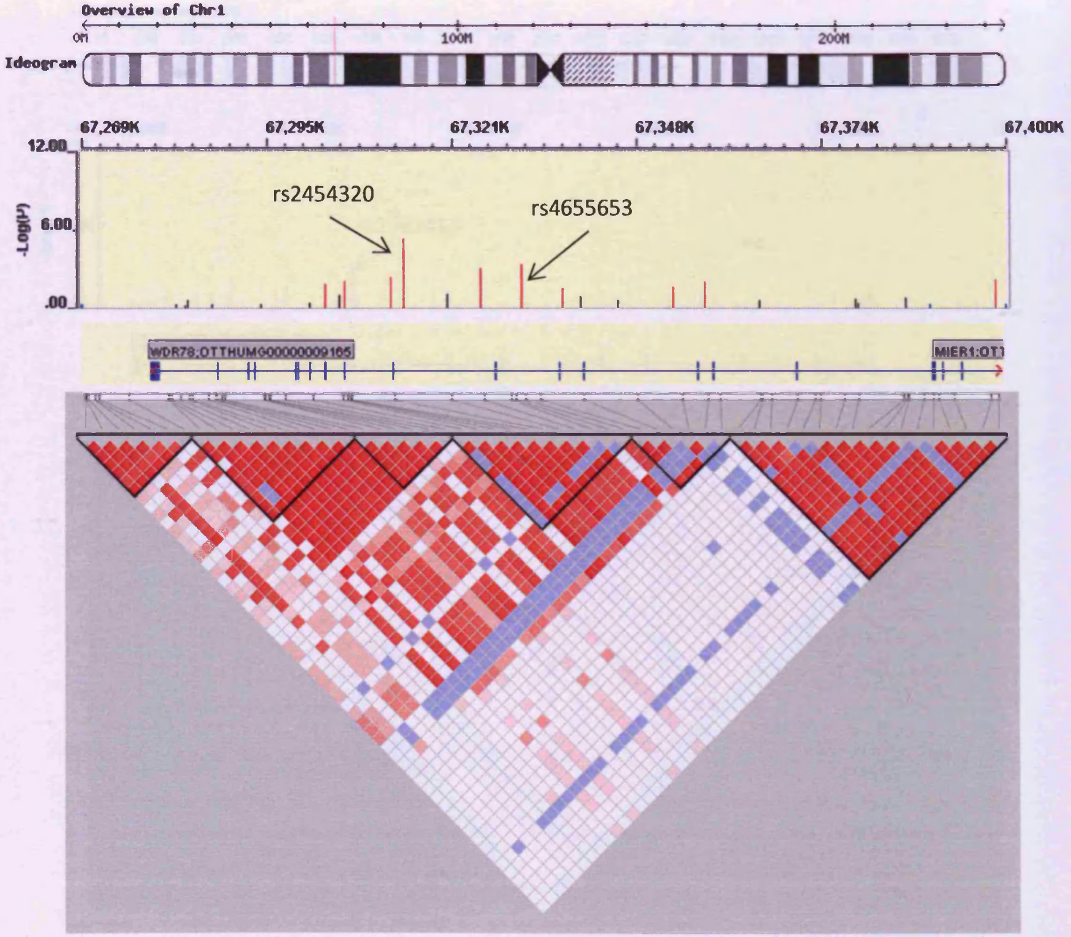


Figure 5.6: LD plot and results of SNPs (P-Fisher values) in *WDR78* that were genotyped in the NeuroDys pooling study. Red bars denote P-values <0.05. The SNP rs4655653 was followed up in the NeuroDys Individual Genotyping Sample. This is in perfect LD with the most significant SNP in this gene, rs2454320 for which an assay could not successfully designed alongside the other SNPs in the follow-up panel.

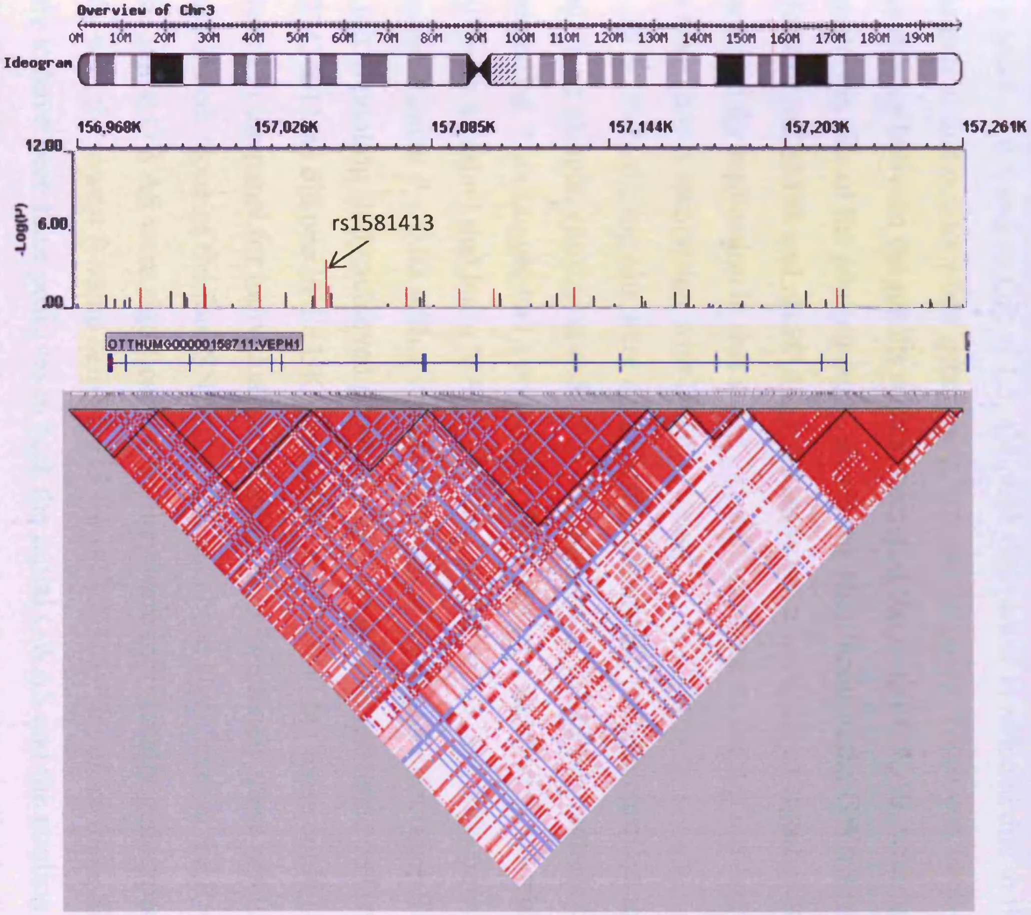


Figure 5.7: LD plot and results of SNPs (P-Fisher values) in *VEPH1* that were genotyped in the NeuroDys pooling study. Red bars denote P-values <0.05

5.3 Discussion

A second GWAS study of DD was carried out using the NeuroDys replication sample from the initial GWAS and the Illumina Human1M-Duo array in the form of a pooling study. After QC filtering, 501,409 SNPs were successfully screened in a total of 1279 cases and 1452 controls from the UK, central Europe and Finland. This sample had a power of 91% to detect a significant association at $P < 5 \times 10^{-4}$ with a variant that has a MAF of 0.4 and an OR of 1.3, although this power is reduced due to the loss of information that occurs when genotyping samples in pools. There was very little concordance between the results of this study and the initial GWAS, with just 4 SNPs from the top hits of the pooling study also being significant in the GWAS (rs1350317, rs1569012, rs268598 and rs1581413) and these were not deemed significant enough to be selected for replication in that study. In addition, none of the top hits were in genes that have shown association with DD previously.

38 SNPs in the top hits were successfully genotyped in the NeuroDys individual genotyping sample, consisting of 988 of the cases and 1121 of the controls that had been pooled. This sample had a power of 77% to detect a significant association at $P < 5 \times 10^{-4}$ with a variant that has a MAF of 0.4 and an OR of 1.3, and 99% power to detect an association at $P < 0.05$ with a variant of the same MAF and effect size. Two SNPs in the initial pooling study achieved genome-wide levels of significance (rs11686995 and rs12743401), as did one in the UK pool alone (rs6865447) but these could not be included in the panel for individual genotyping so unfortunately these results remain unconfirmed. None of the four SNPs that had also been significant at the $P < 0.05$ level in the initial GWAS were significant when they were individually genotyped in this study with the lowest P-value being 0.213 for rs1350317. Therefore these SNPs are likely to have been false positives in both the initial GWAS and the pooling study. Out of the 38 SNPs that were selected, 14 remained significant when individually genotyped in those samples that had been pooled.

When additional samples were included in the individual genotyping, the SNPs were genotyped in a total of 1518 cases and 2261 controls. This sample had a power of 98% to detect a significant association at $P < 5 \times 10^{-4}$ with a variant that has a MAF of 0.4 and an OR of 1.3, and 100% power to detect an association at $P < 0.05$ with a variant that had the same MAF and effect size. Despite the increase in power, the significance of most of the SNPs decreased when including these additional samples. Most of the additional samples were from France and Hungary and so the decrease in significance is may have been due to the lack of association of these SNPs with DD in these samples; just one SNP showed a significant association in the French sample alone and none were significant in the Hungarian sample. This difference between samples may have been due to population differences in LD or allele frequencies, or could have been the result of different ascertainment criteria that was used in each population. However, some SNPs also showed a reduction in significance within the UK and central European samples with the addition of more samples which were recruited using the same ascertainment criteria as those in the pools, suggesting that it is more likely that these SNPs are not truly associated with DD.

Six SNPs that passed QC filters showed an increase in significance in the whole NeuroDys individual genotyping sample compared to the subset that had been pooled. The most significant of these was rs461119 which is within the first intron of (*GRIK1*). This gene encodes a glutamate receptor and as these receptors are the most predominant excitatory neurotransmitters in the mammalian brain, with a role in both short- and long-term synaptic plasticity (Headley & Grillner 1990), making it a good functional candidate for DD. The next most significant SNP was rs12344734 which is in the transmembrane channel-like 1 gene (*TMC1*) on chromosome 9. The specific function of this gene is unknown but mutations in this gene have been associated with progressive postlingual hearing loss and profound prelingual deafness, particularly with deafness autosomal dominant type 36 (DFNA36) (Kurima et al. 2002; Meyer et al. 2005; Santos et al. 2005; Kitajiri et al. 2007b; Tlili et al. 2008; Kitajiri et al. 2007a; Kalay et al. 2005; Vreugde et al. 2002; Hilgert et al. 2008). DFNA36 is a form of sensorineural hearing loss which results from damage to the neural receptors of the inner ear, the nerve pathways to the brain, or the area of the brain that receives sound information. *TMC1* is specifically expressed in the inner and outer hair cells of the cochlea and in neurosensory epithelia of the vestibular end organs (Kurima et al. 2002), and mouse models with a missense mutation in *Tmc1* have shown degeneration of these hair cells (Vreugde et al. 2002) suggesting that mutations in this gene cause deafness through damage to the cochlea hair cells. The role of this gene within hearing and its association with DD in this study suggest that it is possible that *TMC1* may influence susceptibility to DD by affecting an auditory component of the disorder.

Perhaps the most functionally interesting of the other 5 SNPs that showed an increase in significance was rs4655653 within the gene *WDR78*. The exact function of this gene is unknown, but the family of WD-repeat domain containing proteins play key roles in the formation of protein-protein complexes and are critical for a wide range of biological functions including transduction, transcription regulation, cytoskeletal assembly and apoptosis (Li & Roberts 2001; Smith 2008). Interestingly, these proteins have also been associated with a number of human diseases including lissencephaly which is a rare brain formation disorder caused by abnormal neuronal migration (Dobyns & Truwit 1995). The lissencephaly gene 1 (*LIS1*) was the first WD-repeat gene to be identified as responsible for a human disease and is thought to be involved in a signal transduction pathway that is crucial for cerebral development (Reiner et al. 1993). Lissencephaly has also been found to be caused by a mutation in *DCX* – the double cortin gene that shares a common domain with *DCDC2*. If the WD-repeat domain of *WDR78* is found to have a similar function to that of *LIS1*'s domain and is also involved in neuronal migration then it could be a promising candidate gene for DD.

The individual genotyping follow up study was far from ideal because the SNPs were not individually genotyped in the Finnish subset of the sample due to insufficient funding. Of the SNPs that were followed up, 14 (37% of SNPs genotyped) were significant in the NeuroDys individual genotyping sample at the P < 0.05 significance level suggesting that many of these SNPs were false positives in the initial pooling study. These false positives are likely to have been caused by inaccuracies in the estimation of the allele frequencies due to pool-formation or pool-measurement errors. The pool-formation error rate may have been inflated as result of combining samples that had been prepared in different centres. The design of this study may have been improved by forming separate pools for each centre involved, however this would have involved genotyping 8 case pools and 8 control pools in replicate, more than doubling the cost of the experiment. These issues highlight the need to confirm significant findings from pooling studies through individual genotyping due to the inflated false-positive rate that comes with such studies, even when controlling for this inflation as much as is feasibly possible.

Another limitation of this study was that a relatively small number of SNPs were followed up with individual genotyping and more variants with a significant association with DD may have been identified if a large proportion of SNPs further down the list of the top hits from the pooling stage had also been genotyped in the individual sample. However, it was not financially feasible to individually genotype all of the significant SNPs.

Future work for this study should involve the collection of more samples in order to increase the power to detect causal variants for DD. This could include increasing the control sample through the use of more population controls such as those in the Avon Longitudinal Study of Parents and Children (ALSPAC) cohort. These children have been tested for a variety of cognitive measures including reading ability, so a screened subset could be utilised in future DD studies. Those SNPs that were shown to be more significantly associated when including additional samples at the individual genotyping stage need to be replicated in a large independent sample before investigating further. Any interesting genes could then be fine mapped to identify functional variants. The function of the WD-repeat domains within *WDR78* could be of particular interest if they are found to have similar functions to those within the lissencephaly gene, *LIS1*.

In conclusion, there was very little concordance between this pooling study and the initial GWAS. This study has identified a small number of SNPs that could be worthy of follow up, but they did not achieve genome-wide significance and would need to be replicated in a large independent sample before investigating further. This study was also unable to confirm many of the most significant hits through individual genotyping suggesting a larger error rate in the pools than suggested by the pool validation. Pooling studies remain to be an efficient approach to conducting a GWAS but this study suggests that when combining samples from different centres, high error rates may offset the savings made in terms of time and cost.

Chapter 6: Cardiff Genome-wide Pooling Study

6.1 Introduction

Both the initial NeuroDys GWAS (Chapter 4) and the NeuroDys pooling study (Chapter 5) were carried out in collaboration with a number of groups across Europe. Whilst this increased the power of the study by increasing the size of the sample, it also meant that cases were not recruited using the same criteria. This may have resulted in the analysis of cases from different subsamples of the DD spectrum. For example, the German cases were recruited based on a test of spelling, but the UK cases were recruited using reading measures. The use of different ascertainment criteria could partly explain why SNPs that have shown an association in the UK sample have failed to show an association in samples from Germany in previous studies. For example, SNPs that were significantly associated with *KIAA0319* (in the *DYX2* linkage region) in samples from the UK (Francks et al. 2004; Cope et al. 2005a; Harold et al. 2006) were not significantly associated in this German sample (Schumacher et al. 2006a).

As well as genotyping different samples from the collaborative groups, the initial NeuroDys GWAS (Chapter 4) utilised genotypic data from population controls. Again this increased the power of the study due to a larger sample of controls and without an increase in cost, but these samples had not been screened for DD and so individuals with symptoms of DD may have been present in this control group which would reduce the power of this study to detect an association with the disease. In addition, the population control samples were genotyped separately from the cases, in different labs and using different arrays which may have resulted in systematic genotyping error differences existing between the DD cases and the population controls, inflating the false-positive rate even further (Clayton et al. 2005). Whilst these issues were controlled for as much as possible by selecting stringent QC filters, it is acknowledged that there may have still been a higher false-positive rate than if matched, screened controls had been used which had been prepared and genotyped under the same conditions as the cases.

Although the UK cases from Oxford and Cardiff were ascertained using similar tests, Cardiff cases were recruited based on different criteria than those used in the NeuroDys replication and pooled samples. In the NeuroDys studies, these samples were considered to be cases if their reading performance was less than 1.25 standard deviations from the expected age-based norms. However, the Cardiff cases were selected if they had a reading age that was at least 2.5 years behind that expected based on their chronological age. It is thought that reading ability and disability occur along a continuum, with dyslexic individuals forming the lower tail end of the normal distribution (Shaywitz et al. 1992). Selecting individuals with severe phenotypes may increase the ability of a study to find a significant association with DD. A number of studies have found an increased level of significant association when testing a subset of their cases with more severe forms of DD (e.g. Deffenbacher et al. 2004; Francks et al. 2004; Schumacher et al. 2006). Even though this results in smaller sample sizes, it is hoped that this is offset by an increase in the effect size due to the severity of the cases. The *DYX2* locus in particular appears to be associated with the low tail of the distribution of reading ability. For example, Deffenbacher and colleagues (2004) only found linkage to 6p21.3 when they selected a severe subset of their sample and Schumacher and colleagues (2006a) found an increased relative-risk for their associated SNPs in *DCDC2* when increasing the severity of their sample. As the Oxford ascertainment criteria did not select for as severe cases as the Cardiff criteria, this may have affected the ability to detect an association with DD in the NeuroDys UK case sample.

6.1.2 Aims

To address these issues, an additional GWAS was undertaken using the Cardiff case-control sample. The aim of this study was to identify novel susceptibility variants that were associated in the uniform sample of severe DD cases and matched, screened controls. To reduce costs, this GWAS was also carried out as a pooling study using the Illumina Human1M-Duo array in the first stage, with the most significant SNPs being selected for individual genotyping in the second stage. As well as attempting to identify novel variants, these results were compared with the results from the NeuroDys studies (Chapters 4 and 5) in an attempt to ascertain if any variants showed a significant association with DD across all studies.

6.2 Methods

6.2.1 Cardiff Pooling Study

6.2.1.1 Sample

Because the NeuroDys UK control pool from the NeuroDys pooling sample consisted of only Cardiff controls (Chapter 5), this pool was also used as the control pool in the Cardiff pooling sample; as such, only a pool of Cardiff cases needed to be constructed. Samples from the Cardiff cases were quantified twice using PicoGreen. Samples that had a concentration > 5ng/ul and a call rate > 98% in the NeuroDys GWAS replication panel of SNPs were included. Samples also had to meet the ascertainment criteria as outlined in Chapter 2 (IQ ≥ 85 and a reading age ≥ 2.5 years below their chronological age). This resulted in a sample consisting of 302 cases. This sample overlaps both the initial Neurodys GWAS and NeuroDys replication samples to some degree, as shown in Table 6.1.

		Samples in Cardiff Pool Present in Other Sample Sets		
	Cardiff Pool	NeuroDys GWAS Sample	NeuroDys Replication Sample	NeuroDys UK Pool
Cases	302	132	170	157
Controls	219	0	219	219

Table 6.1: Sample used in the Cardiff Pools and overlap with other sample sets used in the NeuroDys studies

6.2.1.2 Pooling

Pools containing only Cardiff cases were constructed as outlined in Chapter 2. After construction, the case pool was quantified using PicoGreen and had a concentration of 10 ng/µl so it was concentrated using Microcon tubes (see Chapter 2 for method) and subsequently diluted to a concentration of 50 ng/µl. As with the NeuroDys UK pools, to test the accuracy of the pool construction, primers for the SNPs rs11648084 and rs1892577 from the NeuroDys GWAS replication panel of SNPs were designed using primer 3 software (http://frodo.wi.mit.edu/primer3/) and the extension primers were designed using FP PRIMER 1.0.1b (http://m034.pc.uwcm.ac.uk/FP_Primer.html) (see Table D.1 in Appendix for primer sequences). These were then used to genotype the Cardiff case and control pools using the SNaPshot method in order to estimate the allele frequencies for these SNPs (see Chapter 2 for method). The estimated differences in allele frequencies between the case and control pools for these SNPs were then compared to the actual differences in allele frequencies when the pooled samples were genotyped individually.

6.2.1.3 Genotyping of Pools

Genotyping was performed using the Illumina 1M-Duo chip in Cardiff according to the manufacturer's protocol (see Chapter 2 for brief description). Each case pool and each control pool were run in replicates of 8. Chips were scanned using the Illumina iScan system and the raw intensities were normalised using BeadStudio v3.2 software. The normalised intensities were then extracted for statistical analysis.

To exclude poorly performing replicates, replicate arrays were excluded if their estimated allele frequencies produced a Pearson correlation of $r \leq 0.995$ with at least one other replicate array for that pool as these appeared to be outliers (see Table D.2 in Appendix for Pearson correlations between all arrays).

As with the NeuroDys pooling study, an approximation of allele A frequencies for each replicate was calculated with the help of Valentina Moskvina using the normalised intensities and the following equation:

$$\text{Frequency of allele A} = \text{Xnorm} / (\text{Xnorm} + \text{Ynorm}).$$

These frequencies were then averaged over the number of replicates in each pool.

SNPs were excluded from the analysis if they had a MAF < 0.05 in either cases or controls. In order to exclude those SNPs whose allele frequencies appeared to be poorly predicted by the pooling analysis, the MAFs of SNPs in the control pools were also compared with their frequencies in the CEPH population of the HapMap project. The correction coefficient, *k*, was calculated using the same method as was used in the NeuroDys pooling study (see section 5.2.1.6 for more information). SNPs with the worst 10% of *k* values (5% in each direction) were excluded in order to remove those SNPs with extreme values of *k* without excluding a large proportion of true positive results. The coefficient of variation for each SNP across replicates was calculated by dividing the standard deviation of the allele frequencies by the mean allele frequency. Those SNPs that had a coefficient of variation > 0.5 across the replicate arrays were excluded in order to remove the outlying SNPs that showed the most variance across the arrays, indicating poorly performing assays (see Figure D.1 in Appendix for histogram of coefficient of variation).

As with the NeuroDys pooling study, association analyses were carried out with the help of Valentina Moskvina using the Combined Z-test (see Chapter 5, section 5.2.1.7 for more information).

6.2.2 Individual Genotyping of Cardiff Pools

6.2.2.1 Sample

Individual genotyping was carried out in those samples that were used to construct the pools, as well as additional cases and controls that fulfilled the ascertainment criteria, as shown in Table 6.2. Genotyping data from 3751 population controls from the 1958 Birth Cohort were also used to compare against the cases, 1437 of which had made up the UK control sample in the initial NeuroDys GWAS sample (Chapter 4).

		Samples in Cardiff Individual Genotyping Sample Present in Other Sample Sets		
	Cardiff Individual Genotyping Sample	NeuroDys GWAS Sample	NeuroDys Replication Sample	NeuroDys UK Pool
Cases	357	147	208	187
Controls	269	0	268	219

Table 6.2: Sample used to individually genotype the Cardiff pooling study follow-up panel and its overlap with other sample sets used in the NeuroDys studies.

6.2.2.2 Individual Genotyping

SNPs were chosen for follow up with individual genotyping based on their significance in the Cardiff pooling study. The top 100 non-redundant SNPs were entered into the Sequenom MassARRAY Assay Design 3.1 software in order to design 2 multiplex panels of 35 SNPs (see Table D.3 in Appendix for primer sequences).

The panels of SNPs were genotyped in the case control sample using the Sequenom MassARRAY iPlex GOLD system as described in Chapter 2. Genotype calling was carried out using the Typer 3.4 software. The UK population controls had previously been genotyped on the Illumina HumanHap550 chip.

All SNP assays were optimised as described in the Chapter 2. Sequenom cluster plots for all SNPs were inspected manually, and SNP assays that did not produce distinct clusters were excluded. Genotyping assays were only considered suitable for analysis if genotypes for CEU individuals were the same as those in the HapMap when available during optimisation. SNPs were tested for Hardy-Weinberg equilibrium in cases and controls and their MAFs were calculated using PLINK v1.05 (Purcell et al. 2007). To exclude poorly performing samples, those with a call rate < 70% were excluded.

SNPs were tested for an association with DD using logistic regression carried out in PLINK v1.05 (Purcell et al. 2007) using both the additive and genotypic models. This analysis was performed with the Cardiff pooling sample and with the additional Cardiff samples that were individually genotyped but were not in the pools (Cardiff individual genotyping sample). Data from the 1958 British Birth Cohort typed on the Illumina HumanHap550 array (n = 3748) were also included at a later stage to increase the size of the control sample (Cardiff sample with population controls).

6.2.2.3 Comparison With Previous GWAS

In order to ascertain the concordance across studies for those SNPs that were individually genotyped, their association results in this study were compared with those in the NeuroDys pooling study (Chapter 5), both when genotyped in the NeuroDys UK pool alone and in all NeuroDys pools. The results were also compared with those in the initial NeuroDys GWAS sample described in Chapter 4 (in both the UK subset and the whole sample). As the initial NeuroDys GWAS was carried out using the smaller Illumina HumanHap300 array, not all of the SNPs that were individually genotyped in this pooling study had been genotyped in the initial GWAS as well. For those SNPs, imputation was carried out on the NeuroDys dataset using PLINK v1.05 (Purcell et al. 2007), as described in Chapter 2. Association analysis was then carried out in the same way as in the GWAS (see Chapter 4) using imputed genotypes for SNPs that had an information score greater than 0.8 as recommended in the PLINK documentation.

6.3 Results

6.3.1 Cardiff Pooling Study

The aim for this section of the thesis was to identify variants that were significantly associated with DD in the Cardiff case-control sample alone, so a GWAS was conducted using pooled DNA. As explained in section 6.2.1.1 of this chapter, the control pool is the same as the UK pool of the NeuroDys study. However, the case pool consists of only Cardiff cases, plus additional cases that were not included in the NeuroDys pooling sample (Chapter 5) but had been genotyped in the initial GWAS (Chapter 4). Table 6.3 shows the results of validating these pooled samples using SNaPshot. For the SNP rs11648084, the pools had an error rate of 0.30% and for rs1892577 there was an error rate of 0.78%. This gave an overall average error rate of 0.54%.

SNP	Difference in allele frequencies from individual genotyping (%)	Difference in allele frequencies estimated from pools (%)	% Error rate
rs11648084	0.57	0.87	0.30
rs1892577	4.71	3.93	0.78

Table 6.3: Comparison of difference in allele frequencies when the Cardiff case control samples were genotyped individually and in pools.

Genome-wide pooled genotyping of the Cardiff cases and controls was also carried out on the Illumina 1M-Duo chip. Predicted allele frequencies for each SNP were averaged over the replicate case and replicate control assays. One of the chips for the case replicates had sections that could not be imaged by the iScan system (either due to hybridisation or staining issues) so it was not included in subsequent analyses. Another case replicate and a control failed QC with Pearson's correlations r = 0.995 and 0.992 respectively. The remaining 4 case replicates and 5 control replicates all passed QC with Pearson's correlations $r > 0.997$ (see Table D.2 in Appendix for all correlations).

Following stringent QC filters, 753,768 SNPs were analysed. Figure 6.1 shows the P-values for these SNPs, calculated using the Combined Z-test. Three of these SNPs achieved a genome-wide level of significance, rs11198878 on chromosome 10 ($P = 1.14 \times 10^{-9}$), rs6865447 on chromosome 5 ($P = 1.51 \times 10^{-9}$) and rs4687806 on chromosome 3 ($P = 4.37 \times 10^{-8}$). A further 181 SNPs had P-values $<1 \times 10^{-4}$ (see Table D.4 in Appendix for a list of the 200 most significant SNPs).

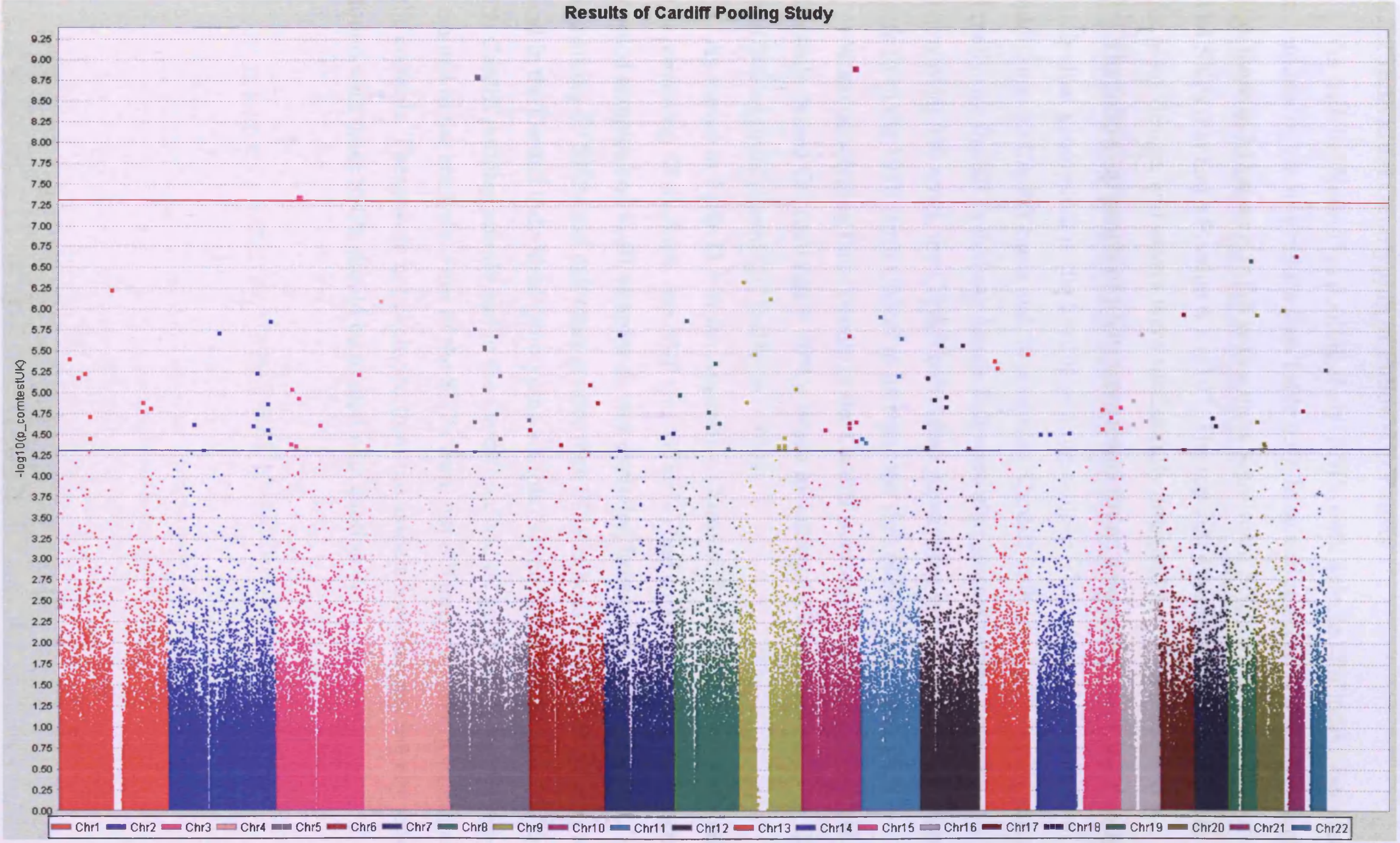


Figure 6.1: Manhattan plot of combined P-values (-log10) from the Cardiff Pooling Study. The red line indicates genome-wide significance with a P-value of 5 x 10^{-8}; the blue line indicates a P-value of 5 x 10^{-5}

6.3.2 Individual Genotyping of Follow-Up Panels

As with the NeuroDys pooling study, SNPs were selected for individual genotyping to validate interesting results (see Table 6.4). In selecting these SNPs, redundant SNPs (i.e. those in LD at an $r^2 > 0.8$ with a more significant SNP) were removed from those 184 SNPs that had a P-value $< 1 \times 10^{-4}$. The top 100 remaining SNPs were then selected for assay design, and assays were successfully designed for 70 of these SNPs.

The follow-up panels of SNPs were tested for an association with DD through individual genotyping in the Cardiff pooling sample. They were also genotyped in an additional 55 Cardiff cases and 50 controls ('Cardiff individual genotyping sample'). Finally, for the SNPs in these panels that were also present on the Illumina HumanHap550 array, the Cardiff individual genotyping sample data was merged with data from the 1958 Birth Cohort to increase the size of the control sample.

When genotyping these panels in the Cardiff pooling sample, 292 cases and 215 controls passed QC (call rate > 70%). When including the additional samples, 331 cases and 262 controls passed QC (call rate > 70%).

As shown in Table D.5 in the Appendix, 12 out of 70 SNPs failed optimisation. Of the remaining 58 markers, one SNP (rs10978074) had a call rate of 66% in both the pooled samples and in all samples so was excluded from association analyses. The remaining 57 SNPs had call rates greater than 72% in both the Cardiff pooling sample and in the Cardiff individual genotyping sample. Seven SNPs had MAFs < 0.05 in both the Cardiff pooling sample and in the Cardiff individual genotyping sample but were retained in the analysis. Four of the SNPs were out of Hardy-Weinberg equilibrium in the controls. These were not excluded from the association analyses but any association found with these SNPs should be treated with caution.

SNP	Rank	Chr	Position (bp)	Nearest RefSeq Gene	Position Relative to Gene	MAF cases	MAF controls	P-Comb	OR
rs4687806	3	3	52119222	*POC1A*	Intronic	0.23	0.11	4.4×10^{-8}	2.53
rs11088038	4	21	27924144	*NCRNA00113*	Intergenic	0.14	0.05	2.1×10^{-7}	3.23
rs11083783	5	19	51050518	*SYMPK*	Intronic	0.07	0.19	2.5×10^{-7}	0.35
rs10978074	6	9	9894481	*PTPRD*	Intronic	0.15	0.05	4.5×10^{-7}	3.13
rs17536837	8	9	72366156	*TRPM3*	Intronic	0.29	0.16	7.1×10^{-7}	2.16
rs6846073	9	4	38634119	*FAM114A1*	Intergenic	0.23	0.11	7.5×10^{-7}	2.35
rs2007343	10	20	59592324	*CDH4*	Intronic	0.40	0.25	9.8×10^{-7}	3.23
rs9890811	11	17	52767460	*MSI2*	Intronic	0.19	0.08	1.1×10^{-6}	2.56
rs6045824	12	20	1909977	*PDYN*	Intronic	0.19	0.09	1.1×10^{-6}	2.50
rs327216	15	8	26547380	*DPYSL2*	Intronic	0.20	0.09	1.3×10^{-6}	2.37
rs488007	16	2	230393214	*TRIP12*	Intronic	0.20	0.09	1.3×10^{-6}	2.44
rs1423363	17	5	58175888	*RAB3C*	Intronic	0.11	0.23	1.6×10^{-6}	0.43
rs7404238	19	16	47320826	*N4BP1*	Intergenic	0.29	0.17	1.9×10^{-6}	2.06
rs10486656	20	7	34767285	*NPSR1*	Intronic	0.22	0.11	1.9×10^{-6}	2.3
rs1352726	21	5	107065959	*EFNA5*	Intergenic	0.14	0.05	2.0×10^{-6}	2.78
rs10745796	25	12	96330306	*RMST*	Intergenic	0.15	0.06	2.6×10^{-6}	2.70
rs3758268	28	9	33394401	*SUGT1P1*	Intronic	0.18	0.08	3.3×10^{-6}	2.39
rs7330054	29	13	109784633	*COL4A2*	Intronic	0.22	0.11	3.4×10^{-6}	2.23
rs7320998	31	13	34345783	*NBEA*	Intergenic	0.08	0.18	4.1×10^{-6}	0.39
rs10504912	32	8	92991692	*RUNX1T1*	Intergenic	0.05	0.15	4.3×10^{-6}	0.33
rs11617247	33	13	41003320	*KIAA0564*	Intergenic	0.20	0.10	5.0×10^{-6}	2.27
rs10493241	36	1	58556559	*DAB1*	Intergenic	0.23	0.12	5.8×10^{-6}	2.10
rs7999	37	3	187853044	*FETUB*	Exonic	0.14	0.06	5.8×10^{-6}	2.63
rs12281150	39	11	84631655	*DLG2*	Intronic	0.08	0.18	6.1×10^{-6}	0.40
rs17440080	40	12	17276384	*LMO3*	Intergenic	0.13	0.05	6.4×10^{-6}	2.95
rs6695238	41	1	43411023	*WDR65*	Exonic	0.14	0.06	6.4×10^{-6}	2.63
rs4431050	43	3	36855688	*TRANK1*	Intronic	0.20	0.10	8.8×10^{-6}	2.19
rs2779708	44	9	129019546	*RALGPS1*	Intronic	0.35	0.23	8.9×10^{-6}	1.84
rs1043180	46	8	11682230	*NEIL2*	3' UTR	0.22	0.11	1.0×10^{-5}	2.13

Table 6.4: SNPs in follow-up panel for individual genotyping. Chr – chromosome; UTR – untranslated region; MAF – minor allele frequency; OR – odds ratio; P-Comb – P-value from combined Z-test.

SNP	Rank	Chr	Position (bp)	Nearest RefSeq Gene	Position Relative to Gene	MAF cases	MAF controls	P-Comb	OR
rs10036598	48	5	5842072	*KIAA0947*	Intergenic	0.19	0.09	1.0×10^{-5}	2.33
rs6581224	49	12	57670904	*LRIG3*	Intergenic	0.18	0.09	1.1×10^{-5}	2.25
rs10492922	52	16	27519162	*KIAA0556*	Intronic	0.19	0.09	1.2×10^{-5}	2.22
rs1885170	53	9	17554267	*SH3GL2*	Intergenic	0.13	0.05	1.3×10^{-5}	2.78
rs11800516	55	1	188170120	*FAM5C*	Intergenic	0.19	0.09	1.3×10^{-5}	2.22
rs2017069	56	2	223630015	*KCNE4*	Downstream	0.06	0.14	1.3×10^{-5}	0.37
rs1600677	58	15	98535787	*ADAMTS17*	Intronic	0.05	0.14	1.5×10^{-5}	0.37
rs11119153	59	1	207142181	*LOC642587*	Intergenic	0.06	0.15	1.5×10^{-5}	0.38
rs10519003	60	15	57498029	*FAM81A*	Intergenic	0.23	0.37	1.6×10^{-5}	0.53
rs2838088	61	21	41965270	*NCRNA00111*	Intergenic	0.29	0.18	1.6×10^{-5}	1.9
rs13250254	62	8	78160148	*PEX2*	Intergenic	0.45	0.31	1.6×10^{-5}	1.78
rs12402777	63	1	188122473	*FAM5C*	Intergenic	0.19	0.09	1.7×10^{-5}	2.22
rs3906517	64	5	107169152	*FBXL17*	Intergenic	0.14	0.05	1.7×10^{-5}	1.85
rs17533238	65	2	201209676	*AOX1*	Intronic	0.32	0.20	1.8×10^{-5}	2.45
rs13375505	66	1	69191724	*DEPDC1*	Intergenic	0.15	0.06	1.9×10^{-5}	2.21
rs1872183	69	5	180178532	*MGAT1*	Upstream	0.18	0.09	2.0×10^{-5}	1.9
rs3785327	71	16	56235199	*GPR56*	Intergenic	0.32	0.20	2.2×10^{-5}	1.84
rs12679969	73	8	102156725	*ZNF706*	Intergenic	0.31	0.20	2.2×10^{-5}	1.85
rs12218153	75	10	109013136	*SORCS1*	Intergenic	0.31	0.19	2.2×10^{-5}	1.85
rs4787965	76	16	27458838	*GTF3C1*	Intronic	0.09	0.18	2.4×10^{-5}	0.44
rs13059624	77	3	100016640	*DCBLD2*	Intronic	0.15	0.07	2.4×10^{-5}	2.37
rs10490093	78	2	58779250	*FANCL*	Intergenic	0.37	0.25	2.4×10^{-5}	1.82
rs10497719	79	2	192085839	*MYO1B*	Intergenic	0.24	0.14	2.4×10^{-5}	1.98
rs10844773	82	12	7401511	*CD163L1*	Intronic	0.10	0.20	2.5×10^{-5}	0.46
rs2745615	85	6	1582259	*GMDS*	Intronic	0.22	0.12	2.6×10^{-5}	2.01
rs11898211	86	2	223658537	*KCNE4*	Intergenic	0.21	0.11	2.7×10^{-5}	2.08
rs17269545	87	15	57269689	*MYO1E*	Intronic	0.06	0.15	2.7×10^{-5}	0.40
rs10267147	89	7	155359690	*SHH*	Intergenic	0.25	0.14	2.9×10^{-5}	2.00

Table 6.4 continued

SNP	Rank	Chr	Position (bp)	Nearest RefSeq Gene	Location relative to gene	MAF cases	MAF controls	P-Comb	OR
rs2748516	91	14	23800302	*TGM1*	Intronic	0.18	0.09	3.1×10^{-5}	2.22
rs1125198	93	7	130580090	MKLN1	Intergenic	0.21	0.11	3.3×10^{-5}	2.08
rs1571581	95	9	102698290	LPPR1	Intergenic	0.34	0.22	3.4×10^{-5}	1.80
rs17180009	96	16	84978318	LOC732275	Intergenic	0.11	0.21	3.4×10^{-5}	0.47
rs804163	97	5	115706855	COMMD10	Intergenic	0.25	0.15	3.4×10^{-5}	1.92
rs968592	98	11	1619112	MOB2	Intronic	0.32	0.2	3.4×10^{-5}	1.83
rs8045270	99	16	85243688	FOXL1	Intergenic	0.17	0.08	3.4×10^{-5}	2.21
rs1529074	104	3	34481950	PDCD6IP	Intergenic	0.16	0.08	4.0×10^{-5}	2.24
rs2050406	106	20	16563735	KIF16B	Intergenic	0.17	0.08	4.1×10^{-5}	2.21
rs10989439	110	9	103021853	LPPR1	Intronic	0.15	0.07	4.3×10^{-5}	2.33
rs17804825	114	20	18183516	ZNF133	Intergenic	0.11	0.21	4.5×10^{-5}	0.47
rs6989022	115	8	121787847	SNTB1	Intronic	0.18	0.09	4.6×10^{-5}	2.12
rs3741781	117	12	107702132	SSH1	3' UTR	0.19	0.09	4.6×10^{-5}	2.19

Table 6.4 continued

The results from the individual genotyping are shown in Table 6.5, in order of their rank in the Cardiff pooling study. Although these SNPs were not as significantly associated as they had been in the pooling study, when genotyped individually in just the Cardiff pooled sample, 54 SNPs gave significant P-values (<0.05) and when genotyped in the whole Cardiff individual genotyping sample, 50 SNPs gave significant P-values. None of these SNPs were significant at a genome-wide level.

The top hit was rs4687806 (P = 1.0 x 10^{-5}). Out of those SNPs selected for individual genotyping, this was also the most significant SNP in the pooling study (P-comb = 4.4 x 10^{-8}). It remained the most significant SNP when analysing the additional Cardiff cases and controls as well (P = 2.0 x 10^{-5}). This SNP was also significantly associated in the UK NeuroDys pool (P-comb = 0.0025) and in the entire NeuroDys pooling sample (P-Fisher = 0.0072). However, it didn't show a high enough level of significance and was not selected for individual genotyping in that study. This SNP was not significantly associated with DD in the NeuroDys UK GWAS sample (P-add = 0.618) or in the complete NeuroDys GWAS sample (P-add = 0.601) (Chapter 4). This SNP is an intronic SNP in the gene POC1 centriolar protein homolog A (*POC1A*) on chromosome 3. This gene is a homolog of the *POC1* gene which is thought to be involved in centriole duplication and length control (Keller et al. 2009). Centrioles are organelles that are involved in the organisation of the mitotic spindle and in the completion of cytokinesis (Salisbury et al. 2002). By homology, *POC1A* may have a similar involvement in centrioles, but it is unclear how this may be linked with a DD phenotype and this gene does not show a high level of expression in the brain.

When the additional samples were combined with the pooled samples, the level of significance was reduced for a number of SNPs. However, 9 SNPs showed an increased level of significant association in all samples. The largest difference in P-values was for the SNP rs2050406 (P = 0.0015 and OR = 2.03 when genotyped in the pooled samples, P = 4.6 x 10^{-4} and OR = 1.99 when genotyped in all samples). This SNP also showed a high level of significance in the NeuroDys UK pool (P = 1 x 10^{-4}) but was not as significantly associated in the complete NeuroDys pooling sample (P-Fisher = 0.0038) so was not selected for individual genotyping in the NeuroDys study (Chapter 5). This SNP is in an intergenic region on chromosome 20, with the nearest gene being kinesin family member 16B (*KIF16B*) which is over 360kb away. However, it is possible that this SNP affects the expression of this gene as the P-value of this SNP with expression levels of *KIF16B* is 0.0332 in lymphoblastoid cell lines in the Genevar database, but is not significantly associated in fibroblasts or T-cells (Dimas et al. 2009).

The most significant SNP to show an increase in significance with the additional samples was rs1125198 (P = 1.8 x 10^{-4} and OR = 2.35 when genotyped in the Cardiff pooled samples, P = 1.3 x 10^{-4} and OR = 2.19 when genotyped in the Cardiff individual genotyping sample). This SNP did not show a high level of significance in the NeuroDys UK pool (P = 0.0012) or in the complete NeuroDys pooling sample (P-Fisher = 0.0031) (Chapter 5). This is an intronic SNP is in the gene muskelin 1, intracellular mediator containing kelch motifs (*MKLN1*) as shown in Figure 6.2. This gene encodes an intracellular protein that acts as a mediator of cell spreading and cytoskeletal responses to the extracellular matrix (ECM) component thrombospondin I (TSP-1) (Adams et al. 1998).

The next most significant SNP to show an increase in significance with the additional samples was rs804163 (P = 7.9 x 10^{-4} and OR = 1.92 when genotyped in the Cardiff pooling sample, P = 6.2 x 10^{-4} and OR = 1.83 when genotyped in the Cardiff individual genotyping sample). This SNP did not show a high level of significance in the NeuroDys UK pool (P = 0.0355) or in all NeuroDys pools combined (P-Fisher = 0.0366) (Chapter 5). This SNP is in an intergenic region of chromosome 5, ~50kb downstream of the nearest gene, COMM domain containing 10 (*COMMD10*). It is possible that this SNP may affect the expression of this gene but this SNP is not in either of the gene expression databases looked at so this was not tested (Dimas et al. 2009; Dixon et al. 2007).

The remainder of the SNPs showing increased significance with the additional samples would not remain significant after correcting for 57 tests (P-values > 0.0010).

Of the SNPs that were selected for individual genotyping from the Cardiff pooling study, 2 were significant in the initial NeuroDys GWAS sample (Chapter 4), both in the UK subset and the whole sample (rs11898211, UK NeuroDys GWAS P-add = 0.0332, NeuroDys GWAS P-add = 0.016; rs17804825, UK NeuroDys GWAS P-add = 0.0086, NeuroDys GWAS P-add = 0.027) and 1 was significant in the UK NeuroDys subset only (rs488007, UK P-add = 0.0159). These SNPs remained significant when genotyped individually in the Cardiff pooled samples (P = 3 x 10^{-4}- 0.04) and also showed significance in the NeuroDys pooling sample in both the UK pool alone (P = 3.5 x 10^{-4} – 0.025) and when all pools were combined (P-Fisher = 1.4 x 10^{-4} - 0.011) (Chapter 5).

However, rs11898211 was no longer significantly associated when the additional Cardiff cases and controls were included in the Cardiff individual genotyping sample (P = 0.0654). Therefore, the only SNP that showed significant association in both the UK and combined samples across all studies is rs17804825. This SNP is in an intergenic region, with the nearest gene being *ZNF133* which is 33.6 Kb away. This SNP is not significantly associated with the expression levels of *ZNF133* in fibroblasts, lymphoblastoid cell lines or in T-cells in the Genevar database so it does not appear to influence the expression of this gene (Dimas et al. 2009).

	Cardiff Pooling Study		Individual Genotyping in Cardiff Pooling sample		Whole Cardiff Individual Genotyping Sample		Cardiff Sample With Population Controls		NeuroDys UK Pool		All NeuroDys Pools	NeuroDys UK Subset of GWAS		NeuroDys GWAS Sample	
SNP	P-comb	OR	P-add	OR	P-add	OR	P-add	OR	P-comb	OR	P-Fisher	P-add	OR	P-add	OR
rs4687806	**4.4 x 10^{-8}**	2.53	**1.0 x 10^{-5}**	3.15	**2.0 x 10^{-5}**	2.79	**0.0012**	1.51	**0.0025**	1.65	**0.0072**	0.6175	1.08	0.6007	1.07
rs11088038	**2.1 x 10^{-7}**	3.23	**0.0020**	2.42	**0.0145**	1.82	0.5605	1.09	**1.0 x 10^{-5}**	2.57	**1.4 x 10^{-6}**	0.1890	0.81	0.1648	0.84
rs17536837	**7.1 x 10^{-7}**	0.35	**0.0024**	2.16	**0.0019**	2.01			**4.2 x 10^{-4}**	1.84	**0.0024**				
rs6846073	**7.5 x 10^{-7}**	2.35	**0.0015**	2.01	**0.0024**	1.83	0.1337	1.20	**0.0170**	1.88	0.1902				
rs2007343	**9.8 x 10^{-7}**	1.93	**1.7 x 10^{-4}**	1.78	**0.0019**	1.51	0.0504	1.20	**1.4 x 10^{-4}**	1.61	**0.0010**	*0.9588*	*1.01*	*0.4991*	*0.94*
rs327216	**1.3 x 10^{-6}**	2.37	**0.0085**	2.06	**0.0198**	1.69	0.7732	1.04	**3.9 x 10^{-5}**	2.06	**1.6 x 10^{-4}**	0.9956	1.00	0.7563	1.05
rs488007	**1.3 x 10^{-6}**	2.44	**3.0 x 10^{-4}**	2.36	**0.0030**	1.82	0.0536	1.28	**0.0143**	1.74	**0.0105**	**0.0159**	1.37	0.0917	1.22
rs1423363	**1.6 x 10^{-6}**	0.43	**3.3 x 10^{-4}**	0.52	**7.1 x 10^{-4}**	0.57			**1.7 x 10^{-4}**	0.55	**0.0032**				
rs7404238	**1.9 x 10^{-6}**	2.06	**0.0329**	1.68	**0.0432**	1.62	0.6620	0.94	0.0572	1.41	0.1094	0.6240	0.94	0.3002	0.91
rs10486656	**1.9 x 10^{-6}**	2.3	**0.0023**	2.03	**0.0120**	1.64	0.0667	1.25	**4.1 x 10^{-5}**	1.96	**5.0 x 10^{-5}**	0.9412	1.01	0.7148	1.03
rs1352726	**2.0 x 10^{-6}**	2.78	**0.0037**	2.10	**0.0021**	2.12	0.4609	1.11	**0.0014**	0.51	**0.0128**	0.9503	1.01	0.4881	1.09
rs10745796	**2.6 x 10^{-6}**	2.70	**0.0048**	3.17	**0.0059**	2.66			**4.3 x 10^{-5}**	2.21	**8.2 x 10^{-4}**				
rs3758268	**3.3 x 10^{-6}**	2.39	**0.0058**	2.02	**0.0274**	1.64	0.6925	1.06	**0.0030**	1.83	**0.0073**	0.9311	1.01	0.3992	0.90
rs7330054	**3.4 x 10^{-6}**	2.23	**5.6 x 10^{-4}**	2.43	**5.7 x 10^{-4}**	2.13	**2.8 x 10^{-5}**	1.71	**0.0401**	1.43	0.0997	0.2930	1.17	0.8549	1.04
rs11617247	**5.0 x 10^{-6}**	2.27	**0.0015**	2.23	**0.0040**	1.91	0.9611	1.01	**0.0084**	1.54	**0.0441**	0.9864	1.00	0.9753	1.01
rs10493241	**5.8 x 10^{-6}**	2.10	**0.0408**	1.51	0.2379	1.21	0.6287	1.06	**0.0380**	1.43	0.0611	0.7169	1.05	0.7970	0.96
rs7999	**5.8 x 10^{-6}**	2.63	**0.0064**	2.36	**0.0328**	1.77			**0.0157**	1.76	**0.0287**				
rs12281150	**6.1 x 10^{-6}**	0.40	**7.2 x 10^{-5}**	0.32	**1.3 x 10^{-4}**	0.37	**0.0434**	0.65	**3.0 x 10^{-4}**	0.50	**0.0034**	0.5872	1.10	0.4356	1.09
rs17440080	**6.4 x 10^{-6}**	2.95	**0.0034**	2.08	**0.0041**	1.86			**0.0029**	2.27	**0.0043**				
rs6695238	**6.4 x 10^{-6}**	2.63	**0.0186**	2.49	**0.0072**	2.64			**2.7 x 10^{-4}**	2.46	**6.7 x 10^{-5}**				
rs4431050	**8.8 x 10^{-6}**	2.19	**0.0035**	1.99	**0.0217**	1.57	0.0875	1.24	**0.0198**	1.53	**3.2 x 10^{-4}**	0.2802	1.15	0.6156	1.06
rs2779708	**8.9 x 10^{-6}**	1.84	**6.4 x 10^{-4}**	1.66	**0.0064**	1.43	0.2437	1.11	**4.4 x 10^{-4}**	1.59	**0.0033**	0.3272	0.91	0.8178	0.98
rs1043180	**1.0 x 10^{-5}**	2.13	**0.0062**	1.81	**0.0025**	1.80	**0.0202**	1.31	**0.0025**	1.58	**0.0014**	0.1689	1.18	0.2353	1.10
rs10036598	**1.0 x 10^{-5}**	2.33	0.0677	1.68	0.1985	1.37	0.7030	0.94	**0.0062**	1.71	**0.0393**	0.7904	1.04	0.8702	1.03
rs6581224	**1.1 x 10^{-5}**	2.25	**0.0036**	1.82	**0.0010**	1.85	0.0930	1.22	**0.0111**	1.60	**0.0282**	0.2610	1.15	0.2207	1.11
rs1885170	**1.3 x 10^{-5}**	2.78	**0.0052**	2.80	**0.0121**	2.29			**3.7 x 10^{-4}**	2.35	**5.3 x 10^{-4}**				
rs11800516	**1.3 x 10^{-5}**	2.22	**0.0012**	2.45	**0.0101**	1.92			0.3725	1.20	0.1740				

Table 6.5: Results of the individual genotyping of the follow up panel for the Cardiff pooling samples, for the whole Cardiff individual genotyping sample, and in the Cardiff sample with population controls (if on the Illumina HumanHap550 array). The results of these SNPs in the NeuroDys GWAS sample (if genotyped) and the NeuroDys pooling sample are also shown. OR – odds ratio; P-comb – P value from combined Z-test; P-Add – P-value from additive test. Significant P-values (<0.05) are in bold. GWAS results in italics are based on imputed SNPs.

	Cardiff Pooling Study		Individual Genotyping in Cardiff Pooling sample		Whole Cardiff Individual Genotyping Sample		Cardiff Sample With Population Controls		NeuroDys UK Pool		All NeuroDys Pools	NeuroDys UK Subset of GWAS		NeuroDys GWAS Sample	
SNP	**P-comb**	**OR**	**P-add**	**OR**	**P-add**	**OR**	**P-add**	**OR**	**P-comb**	**OR**	**P-Fisher**	**P-add**	**OR**	**P-add**	**OR**
rs2017069	**1.3 x 10^{-5}**	0.37	**0.0089**	0.48	**0.0201**	0.54	0.0743	0.69	**0.0092**	0.60	**7.7 x 10^{-5}**	0.7980	0.95	0.9544	0.97
rs11119153	**1.5 x 10^{-5}**	0.38	**0.0027**	0.50	**0.0061**	0.57	**0.0036**	0.62	**9.4 x 10^{-7}**	0.35	**9.6 x 10^{-6}**	0.2423	0.84	0.3528	0.88
rs10519003	**1.6 x 10^{-5}**	0.53	**5.7 x 10^{-4}**	0.59	**0.0014**	0.63	**0.0066**	0.76	**0.0022**	0.58	**7.4 x 10^{-4}**	0.4797	0.93	0.1600	0.89
rs2838088	**1.6 x 10^{-5}**	1.9	**1.8 x 10^{-4}**	1.90	**0.0043**	1.56			**0.0224**	1.42	0.0550				
rs13250254	**1.6 x 10^{-5}**	1.78	**4.9 x 10^{-5}**	1.76	**1.0 x 10^{-4}**	1.64			**0.0088**	1.41	**0.0235**	*0.7196*	*0.97*	*0.8662*	*1.01*
rs12402777	**1.7 x 10^{-5}**	2.22	**6.7 x 10^{-5}**	2.88	**7.4 x 10^{-4}**	2.31			0.1222	1.37	0.3057				
rs17533238	**1.8 x 10^{-5}**	1.85	**0.0043**	1.68	**0.0055**	1.64	0.1422	1.17	**0.0075**	1.52	**0.0145**	*0.4449*	*0.91*	*0.5687*	*0.95*
rs13375505	**1.9 x 10^{-5}**	2.45	**0.0088**	1.89	0.1076	1.40	0.1920	1.19	**0.0136**	1.69	0.0677	0.3853	0.88	0.4528	0.91
rs1872183	**2.0 x 10^{-5}**	2.21	**5.1 x 10^{-4}**	2.29	**0.0010**	2.02	0.1166	1.23	**2.4 x 10^{-4}**	1.86	**6.6 x 10^{-6}**	0.4437	1.11	0.5638	1.07
rs12218153	**2.2 x 10^{-5}**	1.85	**0.0026**	1.78	**0.0221**	1.48	**0.0085**	1.34	**0.0026**	1.55	**0.0013**	0.6034	1.06	0.1568	1.15
rs4787965	**2.4 x 10^{-5}**	0.44	**0.0140**	0.50	**0.0291**	0.55	0.0639	0.68	**0.0012**	0.55	**0.0264**	0.3308	0.83	0.6489	0.94
rs10490093	**2.4 x 10^{-5}**	1.82	**0.0050**	1.73	**0.0099**	1.56			**8.1 x 10^{-5}**	1.73	**0.0021**				
rs10497719	**2.4 x 10^{-5}**	1.98	**0.0113**	1.77	**0.0130**	1.60	**0.0358**	1.28	**9.7 x 10^{-4}**	1.66	**0.0065**	0.6124	1.06	0.5996	0.96
rs10844773	**2.5 x 10^{-5}**	0.46	**2.6 x 10^{-4}**	0.42	**0.0026**	0.51	**7.8 x 10^{-5}**	0.51	**4.2 x 10^{-6}**	0.43	**1.9 x 10^{-4}**	0.1391	1.19	0.2459	1.12
rs2745615	**2.6 x 10^{-5}**	2.01	**0.0164**	1.79	0.1956	1.34	0.3896	1.13	0.0545	1.39	**0.0198**	*0.1927*	*1.20*	*0.1616*	*1.18*
rs11898211	**2.7 x 10^{-5}**	2.08	**0.0398**	2.06	0.0654	1.81	0.3829	0.84	**0.0251**	1.51	**0.0100**	**0.0332**	1.34	**0.0160**	1.31
rs17269545	**2.7 x 10^{-5}**	0.40	0.0538	0.55	0.0930	0.62			0.2820	0.84	**0.0384**				
rs10267147	**2.9 x 10^{-5}**	2.00	**1.7 x 10^{-4}**	2.03	**3.9 x 10^{-4}**	1.93			**0.0122**	1.55	0.0615				
rs2748516	**3.1 x 10^{-5}**	2.22	**0.0306**	1.84	**0.0466**	1.67			**4.6 x 10^{-4}**	1.97	**3.6 x 10^{-4}**				
rs1125198	**3.3 x 10^{-5}**	2.08	**1.8 x 10^{-4}**	2.35	**1.3 x 10^{-4}**	2.19	**2.2 x 10^{-4}**	1.56	**0.0012**	1.86	**0.0031**	0.1293	1.22	0.0612	1.21
rs1571581	**3.4 x 10^{-5}**	1.8	**0.0015**	1.74	**0.0061**	1.52	**0.0090**	1.30	**0.0121**	1.42	**0.0389**	0.5240	1.07	0.4093	1.07
rs804163	**3.4 x 10^{-5}**	1.92	**7.9 x 10^{-4}**	1.92	**6.2 x 10^{-4}**	1.83			**0.0355**	1.37	**0.0366**	*0.6275*	*0.93*	*0.7378*	*0.96*
rs968592	**3.4 x 10^{-5}**	1.83	**0.0011**	1.80	**0.0022**	1.71	**0.0109**	1.30	**0.0246**	1.38	0.1248	*0.3481*	*1.12*	*0.9066*	*1.01*
rs8045270	**3.4 x 10^{-5}**	2.21	**0.0016**	2.87	**0.0078**	2.20			**0.0052**	1.73	**0.0039**				
rs1529074	**4.0 x 10^{-5}**	2.24	0.1210	1.67	0.1257	1.58			**4.2 x 10^{-4}**	1.95	**0.0044**				
rs2050406	**4.1 x 10^{-5}**	2.21	**0.0015**	2.03	**4.6 x 10^{-4}**	1.99	**0.0084**	1.36	**1.0 x 10^{-4}**	2.09	**0.0038**	0.2489	1.16	0.0583	1.20
rs10989439	**4.3 x 10^{-5}**	2.33	**0.0060**	3.34	**0.0129**	2.54			**0.0063**	1.69	**0.0162**				
rs17804825	**4.5 x 10^{-5}**	0.47	**0.0010**	0.47	**0.0020**	0.51	**0.0123**	0.67	**3.5 x 10^{-4}**	0.55	**1.4 x 10^{-4}**	**0.0086**	0.67	**0.0270**	0.79
rs6989022	**4.6 x 10^{-5}**	2.12	**0.0043**	1.97	**0.0013**	2.05	0.1731	1.20	**0.0102**	1.65	**0.0226**	0.0937	1.25	0.2158	1.15
rs3741781	**4.6 x 10^{-5}**	2.19	**0.0036**	2.52	**0.0085**	2.06			**0.0014**	1.83	**0.0080**				

Table 6.5 continued

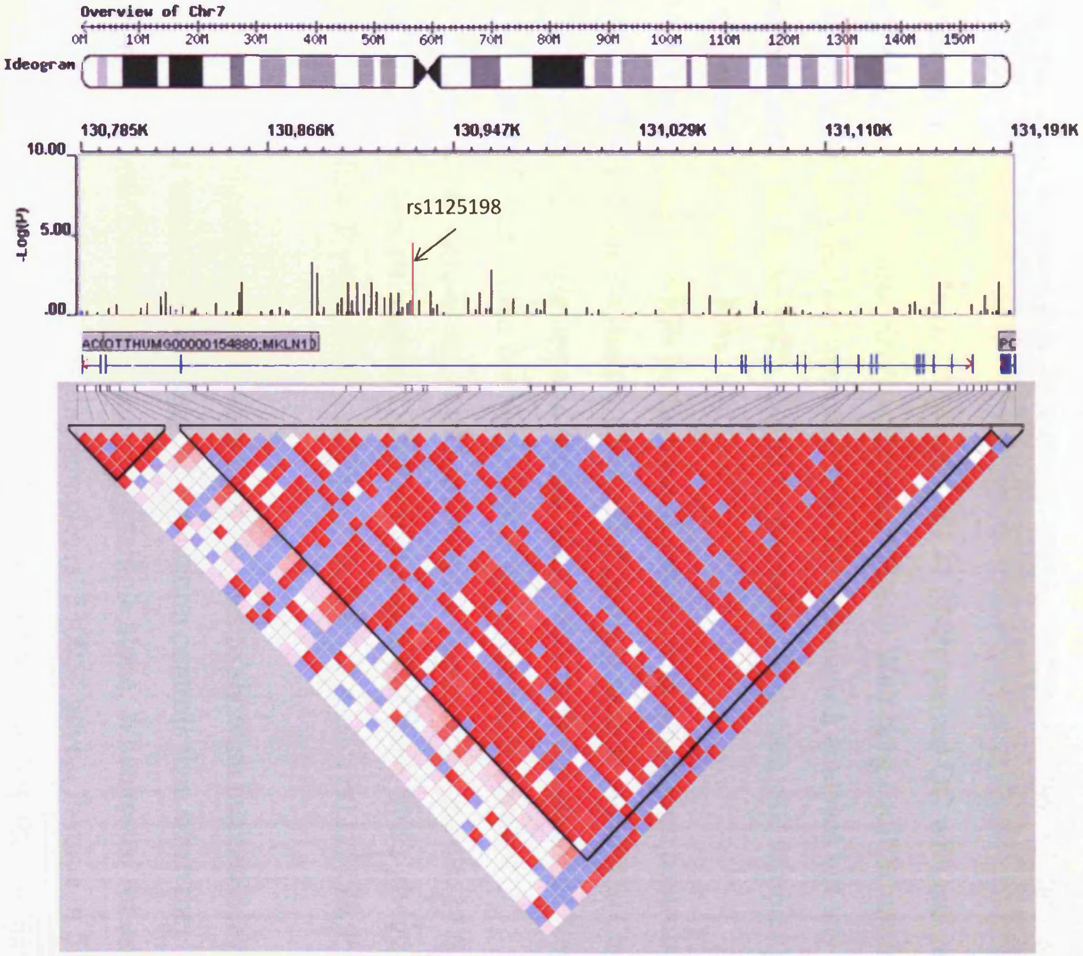


Figure 6.2: LD plot and results of SNPs in *MKLN1* that were genotyped in the Cardiff pooling study. Red bars indicate P-values < 5 x 10^{-5}.

Of the SNPs that were selected for individual genotyping in the Cardiff pooling study, 37 are on the Illumina HumanHap550 array so have been previously genotyped in the 1958 British Birth Cohort. In order to ascertain if any of the SNPs would increase in significance when including more controls, the data from the 262 controls that were individually genotyped were merged with data from 3748 population controls from the 1958 British Birth Cohort to form a combined sample of 4010 controls ('Cardiff sample with population controls'). Association analyses were then performed comparing these controls with all of the cases that were individually genotyped as part of the Cardiff pooling study (n = 331).

As shown in Table D.6 in the Appendix, all 37 SNPs passed QC with call rates > 0.96 and MAFs > 0.05. Three SNPs were out of Hardy Weinberg equilibrium in the control sample. These were not excluded from the association analyses but any association found with these SNPs should be treated with caution. All samples passed QC with genotyping rates > 80%.

The results of these 37 SNPs in the Cardiff sample with population controls are shown in Table 6.5. Their P-values in the Cardiff pooling sample and in the Cardiff individual genotyping sample are also shown, along with their results in the NeuroDys pooling sample (Chapter 5) and the NeuroDys GWAS sample (Chapter 4).

When the Cardiff samples are combined with the population controls, 14 SNPs showed significant association with DD (P < 0.05) with 3 of these SNPs giving P-values < 5 x 10^{-4} (rs7330054, P = 2.8 x 10^{-5}; rs10844773, P = 7.8 x 10^{-5}; rs1125198, P = 2.2 x 10^{-4}).

The two most significant SNPs were the only ones to show an increased level of significance in this larger sample using the population controls than when they were analysed in the Cardiff individual genotyping sample alone. When tested for association using the Cardiff individual genotyping sample, rs7330054 gave a P-value of 5.7 x 10^{-4} (OR = 2.13) and a P-value of 2.8 x 10^{-5} (OR = 1.71) when including the population controls. This SNP lies within intron 3 of the gene collagen type IV alpha 2 (*COL4A2*), as shown in Figure 6.3. No other SNPs in this gene had a P-value < 1 x 10^{-4} in the Cardiff pooling study, but the SNP rs7323190 ~1Kb away, had a P-value of 0.0052. These SNPs are in LD with each other (D' = 1) but they are not highly correlated (r^2 = 0.19).

The next most significant SNP was rs10844773 which gave a P-value of 0.0026 when tested for association using the Cardiff individual genotyping sample and a P-value of 7.8 x 10^{-5} when including the population controls. This SNP lies within intron18 of the gene CD163 molecule-like 1 (*CD163L1*) as shown in Figure 6.4. Again, no other SNPs in this gene had a P-value < 1 x 10^{-4} in the Cardiff pooling study, with the next most significant SNP being rs11053657 (P = 0.008) which is ~40Kb upstream of rs10844773. These SNPS are not in high LD with each other, with D' = 0.238 (r^2 = 0.01).

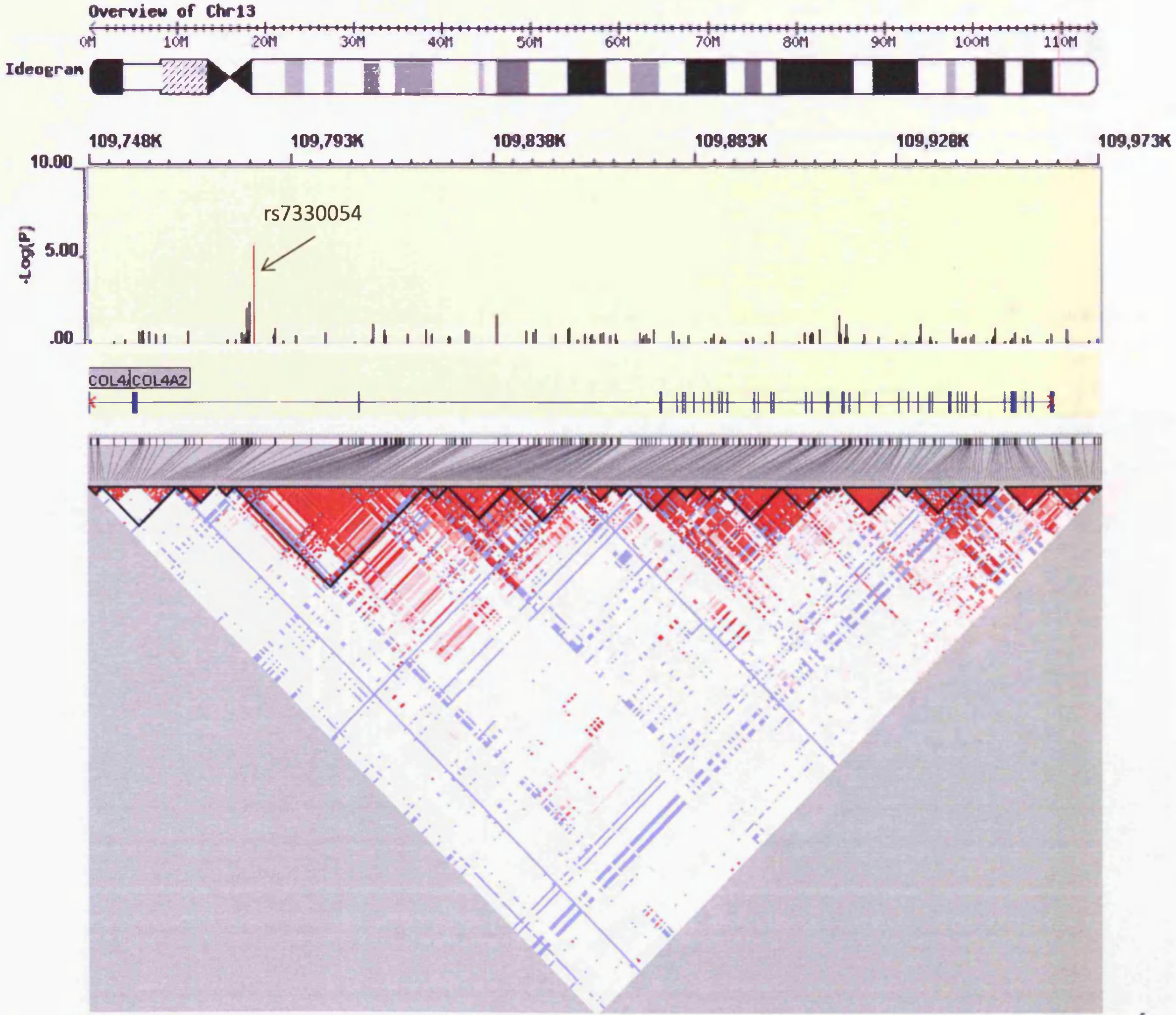


Figure 6.3: LD plot and results of SNPs in *COL4A2* that were genotyped in the Cardiff pooling study. Red bars indicate P-values < 5 x 10^{-5}.

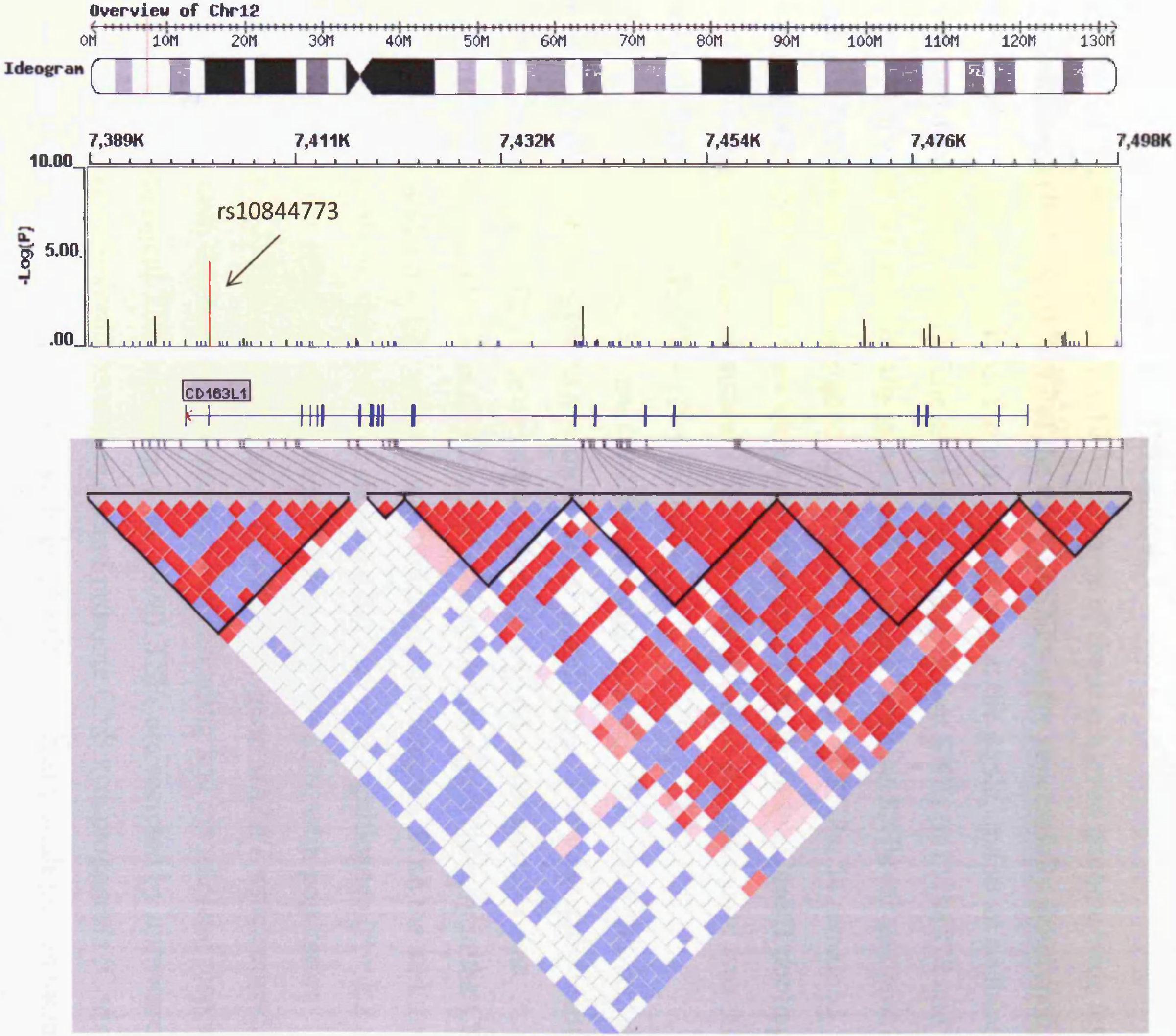


Figure 6.4: LD plot and results of SNPs in *CD163L1* that were genotyped in the Cardiff pooling study. Red bars indicate P-values < 5 x 10^{-5}.

6.3 Discussion

A genome-wide pooling study was conducted using Cardiff cases and controls in order to identify variants that are associated with DD in this more homogeneous sample. As with the NeuroDys pooling study, the Illumina Human1M-Duo array was employed. After QC filtering, 753,768 SNPs were tested for an association in the Cardiff pools which consisted of 302 cases and 219 controls.

184 SNPs had P-values < 1 x 10^{-4}, and three of these achieved genome wide levels of significance (P < 5 x 10^{-8}). 57 of the top 120 SNPs were successfully genotyped individually in 292 cases and 215 controls that were in the pools, and in an additional 39 cases and 47 controls. Two of the genome wide significant SNPs (rs11198878 and rs6865447) were not individually genotyped as assays for these SNPs (or any proxies with $r^2 > 0.8$) could not be designed into the panels. Of the 57 SNPs, 54 remained significant (P < 0.05) when they were genotyped individually in the Cardiff pooling sample. The genome-wide significant SNP that was individually genotyped was the most significant hit (rs4687806, P = 1 x 10^{-5}).

When including the additional Cardiff cases and controls, 50 of the SNPs remained significant, and 9 SNPs showed a higher level of significance. The most significant SNP to show an increase in significance when including the additional samples was rs1125198 which had a P-value of 1.8 x 10^{-4} when genotyped individually in the Cardiff pooling sample and 1.3 x 10^{-4} with the additional Cardiff samples. With the addition of the population controls, this SNP still showed a high level of significance (P = 2.2 x 10^{-4}) but had a smaller effect size (with Cardiff controls OR = 2.19, with population controls OR = 1.56). This is an intronic SNP within the gene *MKLN1* which encodes an intracellular protein that acts as a mediator of cell spreading and cytoskeletal responses to the ECM component TSP-1 (Adams et al. 1998). TSPs are secreted by immature astrocytes during embryonic development and promote CNS synaptogenesis (Christian et al. 2008) and *MKLN1* transcripts have been identified in many adult tissues including the brain (Adams et al. 1998). Tagnaouti and colleagues (2007) have shown that transcripts of *MKLN1* are expressed throughout the CNS, particularly in the hippocampus and cerebellum. They also showed that muskelin localises to the nucleus of neurons as well as to axonal and dendritic projections, including synaptic sites, giving it a possible role within synaptogenesis (Tagnaouti et al. 2007). Therefore this gene appears to have a role within synaptogenesis in the hippocampus and cerebellum and so could have an interesting role within the context of DD.

Data from the 1958 Birth Cohort for the 37 SNPs in the follow up panel that are also present on the Illumina HumanHap550 array were combined with the Cardiff case-control data in order to increase the control sample size. After adding the population controls to the Cardiff sample, many SNPs were no longer significant, with 14 remaining significant. A possible explanation for the reduction in significance of these SNPs is that the initial findings were false positives due to sampling variation. Another possibility is that the use of population controls that had not been screened for DD may have reduced the effect sizes of these variants and therefore the ability to detect a significant association. This could be tested in future studies by comparing the effects of including controls from the 1958 Birth Cohort against using population controls from the ALSPAC cohort. As discussed in Chapter 5, the children in this cohort have been tested for a variety of cognitive measures and could therefore be screened for symptoms of DD.

Of those that remained significant, two SNPs showed an increased level of significance in this larger sample. The most significant of these was rs7330054 which had a P-value of 5.7 x 10^{-4} when genotyped in just the Cardiff case control sample and a P-value of 2.8 x 10^{-5} with the addition of the population controls. This is an intronic SNP within the gene *COL4A2* on chromosome 13. The next most significant SNP in this gene was rs7323190, which lies ~1Kb upstream of rs7330054 and had a P-value of 0.0052. These SNPs are in high LD with each other. The SNP rs7330054 had been genotyped in all the previous studies, but only showed a low level of significance in the UK NeuroDys pool (P = 0.0441), despite the large overlap in samples between the UK NeuroDys pool and the Cardiff pooling sample. This may either suggest that the UK pool was not able to estimate the allele frequencies of this SNP as accurately, resulting in a false negative, or this SNP does not show association in the Oxford sample. *COL4A2* encodes one of the six units of type IV collagen. The C-terminal portion of this protein is an inhibitor of angiogenesis and tumour growth; it inhibits proliferation and migration of endothelial cells and induces apoptosis (Kamphaus et al. 2000). This gene has ubiquitous expression, but shows its highest level of expression in the placenta. However, it has recently been linked with a range of cerebral small-vessel diseases in humans so may also have a role to play in the brain by affecting the blood vessels (Volonghi et al. 2010). Small-vessel disease is more common in older people and it is thought that it may contribute to cognitive decline by affecting information processing speed and executive function (Prins et al. 2005). It is possible that this gene may influence an individual's susceptibility to DD by affecting the blood vessels in their brain, impairing their cognition and therefore their ability to read effectively.

The other SNP which showed a higher level of significance when including the population controls was rs10844773 which had a P-value of 0.0026 in the Cardiff individual genotyping sample and a P-value of 7.8 x 10^{-5} in the Cardiff sample with population controls. This is an intronic SNP within the gene *CD163L1* on chromosome 12. The SNP rs10844773 was also genotyped in all previous studies, and while it wasn't significantly associated in the initial NeuroDys GWAS study, it showed a high level of significance in the UK pool (P = 4.2 x 10^{-5}) and in all NeuroDys pools combined (P-Fisher = 1.9 x 10^{-4}) but it wasn't selected for follow up with individual genotyping in that study because it was not significantly associated in either the Central European pool or the Finnish pool alone. The gene *CD163L1* encodes a member of the scavenger receptor cysteine-rich (SRCR) superfamily (Gronlund et al. 2000). Members of this family are secreted or membrane-anchored proteins, mainly found in cells associated with the immune system (Van Gorp et al. 2010). This may be an interesting function within the context of DD as immune disorders have previously been linked with DD due to a hypothetical common aetiology via prenatal effects of testosterone (Behan & Geschwind 1985), although the evidence for this has been inconclusive and the two types of disorders do not consistently cluster in families (Gilger et al. 1998). In addition, the *DYX2* region overlaps the human histocompatbility antigen (HLA) region, raising the possibility that the association observed between DD and immune disorders in some studies may be genetically linked (Cardon et al. 1994; Grigorenko et al. 1997). This region is highly polymorphic however and contains many genes that influence immune function.

This sample had a power of < 3.3% to detect a significant association (P< 1 x 10^{-4}) with a variant that has a MAF of 0.4 and an OR of 1.3, and a power of 54% to detect a significant association at the P < 0.05 level with a variant with the same MAF and effect size. However, this power is reduced due to the loss of information that occurs when genotyping samples in pools. This shows that this pooling study was greatly underpowered to detect variants at this level of significance in comparison with the

NeuroDys pooling study due to the very small sample size employed. However, after individual genotyping, the Cardiff pooling study identified more significant associations than the NeuroDys pooling study. This may be due to the ability of this study to identify association with variants of large effect sizes due to the use of a homogeneous sample that was collected using consistent ascertainment criteria. After individual genotyping in the Cardiff pooling sample, the effect sizes of the follow-up SNPs ranged from 1.66 to 3.34, which are remarkably higher than the effect sizes of the follow up SNPs in the NeuroDys pooling study (OR = 1.01 – 1.58).

Whilst the level of significance of the SNPs after individual genotyping was lower than it had been in the Cardiff pooling study, the proportion of SNPs that remained significant after individual genotyping in the Cardiff pooling sample (95% of SNPs genotyped) was higher than in the NeuroDys pooling study (37% of SNPs genotyped), which suggests that the Cardiff pooling study was able to estimate allele frequencies more accurately. This indicates that there was a higher error rate in the NeuroDys study, and as discussed in Chapter 5, this may have been a result of combining samples from different countries. In addition, the NeuroDys pools were genotyped in Bonn, whereas the Cardiff pools in this study were genotyped in Cardiff. It may be that there was a higher pool-measurement error rate in the pooled samples genotyped in Bonn, resulting in inaccuracies in the estimation of the allele frequencies. These inaccuracies may have offset the savings made in terms time and money. This suggests that in order for pooling studies to be a valuable method of screening large samples at a lower cost, they ideally need to be conducted using samples that have been ascertained, prepared and genotyped in the same centres. However, this is not often possible in large collaborative GWAS studies involving several research groups.

In general, there was little concordance across all the GWAS studies conducted in this project. 47 of the 50 SNPs that were significant after individual genotyping in this study were also significant in the UK pool of the NeuroDys study and 42 of these were also significant when combining all the NeuroDys pools. It is not surprising that many of these SNPs were also significant in the UK pool as the case pools in each study had a large overlap of samples and the control pools were identical. None of these SNPs were selected for individual genotyping in that study, mainly because they were not significant in more than one pool. This suggests that these SNPs do not show a high level of association with DD in other population samples, possibly due to the difference in ascertainment criteria used. In the initial GWAS study, only one of these SNPs showed significant association in both the UK and combined samples (rs17804825, P = 0.0086 in UK sample, P = 0.0270 in all samples). This SNP is in an intergenic region, 33.6 Kb upstream of the nearest gene, *ZNF133*.

In conclusion, this pooling study has highlighted three variants of potential interest within the genes *MKLN1, COL4A2* and *CD163L1* which may be worthy of further investigation in larger, independent samples. If they show replication in other samples, then these genes could be fine-mapped in order to identify any underlying causal variants. It should be noted that a relatively small proportion of SNPs were followed up and more significantly associated SNPs may have been confirmed using individual genotyping. This study has provided further evidence that pooling studies represent a cost effective approach to identifying variants that are significantly associated with a complex trait.

Chapter 7: Copy Number Variant Analysis Using GWAS Data

7.1 Introduction

7.1.1 Structural Variation in the Human Genome

Single nucleotide polymorphisms were once thought to be the main source of genetic and phenotypic variation, but the advent of genome scanning technologies has enabled the identification of an unexpectedly large amount of structural variation in the human genome (Feuk et al. 2006; Stankiewicz & Lupski 2010). Structural variants are defined as 'a change of genomic DNA greater than 1 kb in size that distinguishes two genomes in one species' (Stankiewicz & Lupski 2010) and can be microscopic (i.e. those large enough to view with a microscope) or submicroscopic (Feuk et al. 2006). The microscopic variants were observed decades before the availability of sequencing technology and include aneuplodies, rearrangements, heteromorphisms and fragile sites. The development of new technologies (e.g. genome-scanning arrays and next-generation sequencing) has allowed the researcher to investigate the genome at a much higher resolution and this has led to the identification of smaller, submicroscopic variants including copy number variants (CNVs), segmental duplications, inversions and translocations (Feuk et al. 2006).

CNVs are segments of DNA ranging from 1 kb to several Mbs in size that are variable in copy number compared to a reference genome of the same species (Feuk et al. 2006; Stankiewicz & Lupski 2010). CNVs can be insertions, translocations, deletions, duplications and triplications. They can be simple in structure or may involve complex gains or losses of homologous sites at multiple sites in the genome (Redon et al. 2006). They can be inherited or sporadic and large *de novo* CNVs are thought to be the most pathogenic (Stankiewicz & Lupski 2010).

7.1.2 Mechanisms of Copy Number Change

A change in copy number requires a change in chromosome structure, joining two formerly separated DNA sequences (Hastings et al. 2009). Segmental duplications are defined as sequences in the reference genome assembly sharing >90% sequence similarity over >1 kb with another genomic location (Bailey et al. 2002) and CNVs are often reported to occur in regions that are flanked by segmental duplications (Freeman et al. 2006; Redon et al. 2006). When segmental duplications have DNA sequence identity greater than ~97% and are located less than ~10Mb away from each other, they can lead to misalignment of chromosomes or chromatids and can mediate non-allelic homologous recombination (NAHR) - one of the three major mechanisms proposed for genomic rearrangements (Stankiewicz & Lupski 2002; Gu et al. 2008). As shown in Figure 7.1, non-allelic segmental duplications can sometimes be misaligned during meiosis or mitosis and this misalignment and the subsequent crossover between them can result in genomic rearrangements in progeny cells (Gu et al. 2008). When this occurs between segmental duplications that are on the same chromosome and in direct orientation with each other, this can result in duplications and/or deletions, as shown in Figure 7.1A. When it occurs between segmental duplications on the same chromosome in reverse orientation it can result in inversions (shown in Figure 7.1B) and when it occurs between segmental duplications on different chromosomes it can result in translocations (shown in Figure 7.1C).

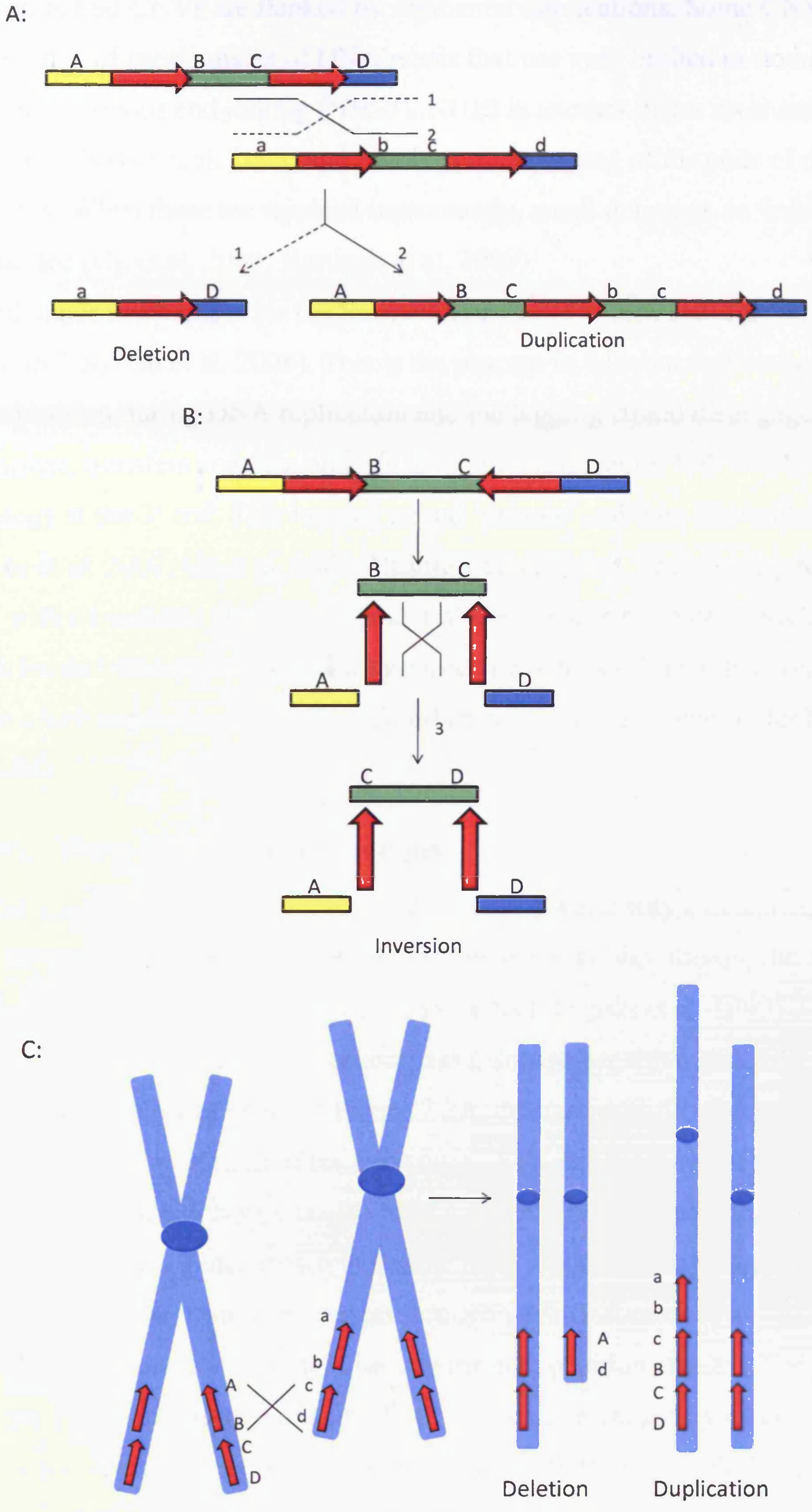


Figure 7.1: Diagrams illustrating how NAHR via segmental duplications (indicated by the red arrows) can result in A: deletions (via mechanism 1) and/or duplications (via mechanism 2), B: inversions and C: translocations. Adapted from Gu et al. (2008).

However, not all CNVs are flanked by segmental duplications. Some CNVs may be formed by a range of mechanisms of DNA repair that use very limited or no homology, such as non-homologous end joining (NHEJ). NHEJ is another major mechanism in the formation of CNVs (Gu et al. 2008) and involves the rejoining of the ends of double-stranded breaks. When these are rejoined inaccurately, small deletions or 'information scars' are formed (Gu et al. 2008; Hastings et al. 2009).

The third major mechanism for the formation of CNVs is fork stalling and template switching (FoSTeS) (Gu et al. 2008). This is the process in which a replication fork stalls at one position during DNA replication and the lagging strand disengages from the original template, transfers and then anneals to another replication fork nearby due to microhomology at the 3' end. This lagging strand 'primes' and then DNA synthesis is restarted (Lee et al. 2007; Gu et al. 2008; Hastings et al. 2009). The priming results in a 'join point' with a transition from one segment of the genome to another. Switching to another fork located downstream (forward invasion) would result in a deletion whilst switching to a fork located upstream (backward invasion) would result in duplication (Gu et al. 2008).

7.1.3 Possible Effects of CNVs on Phenotypes

CNVs can cause phenotypic variation or disease in several ways, as shown in Figure 7.2. One of the most commonly recognised mechanisms involves altering the copy number of a gene (or genes) sensitive to a dosage effect (Lupski et al. 1992). CNVs can affect gene dosage directly when they encompass a dosage-sensitive gene (Figure 7.2A). As shown in the lower panel of Figure 7.2A, dosage-insensitive genes can also be affected by CNVs if a deletion of the gene unmasks a recessive mutation on the homologous chromosome. CNVs can also have a direct effect on gene-dosage when they only partly overlap a gene, shown in Figure 7.2B. CNVs overlapping a gene can also lead to the formation of new transcripts through gene fusions or exon shuffling. CNVs may also alter gene expression indirectly through position effects (Kleinjan & van Heyningen 2005), as shown in Figure 7.2C. This can be caused by deletion or duplication of an important regulatory element (upper panel) or through the unmasking of a functional polymorphism within an effector (lower panel).

CNVs can be benign, can have subtle influences on phenotypes (such as modifying an individual's response to a particular drug), can cause disease or susceptibility to a disease in the current generation, or can predispose to disease in the next generation

(Inoue & Lupski 2002). Although some CNVs might appear to be benign and are prevalent in certain populations, they may contribute to a complex disease phenotype when present in combination with other genetic (e.g. SNPs and other CNVs) and environmental factors (Feuk et al. 2006).

Some evidence has also suggested that CNVs may contribute to phenotypic variation that has a role in determining fitness. A gene ontology (GO) analysis of genes that were known to be affected by CNVs that had been identified in the literature to date by Feuk et al (2006) showed an enrichment of genes that are involved in immune responses and responses to biotic stimuli. These enrichments indicated that structural variation may have a role in the adaptability and fitness of an organism in response to external pressures (Feuk et al. 2006).

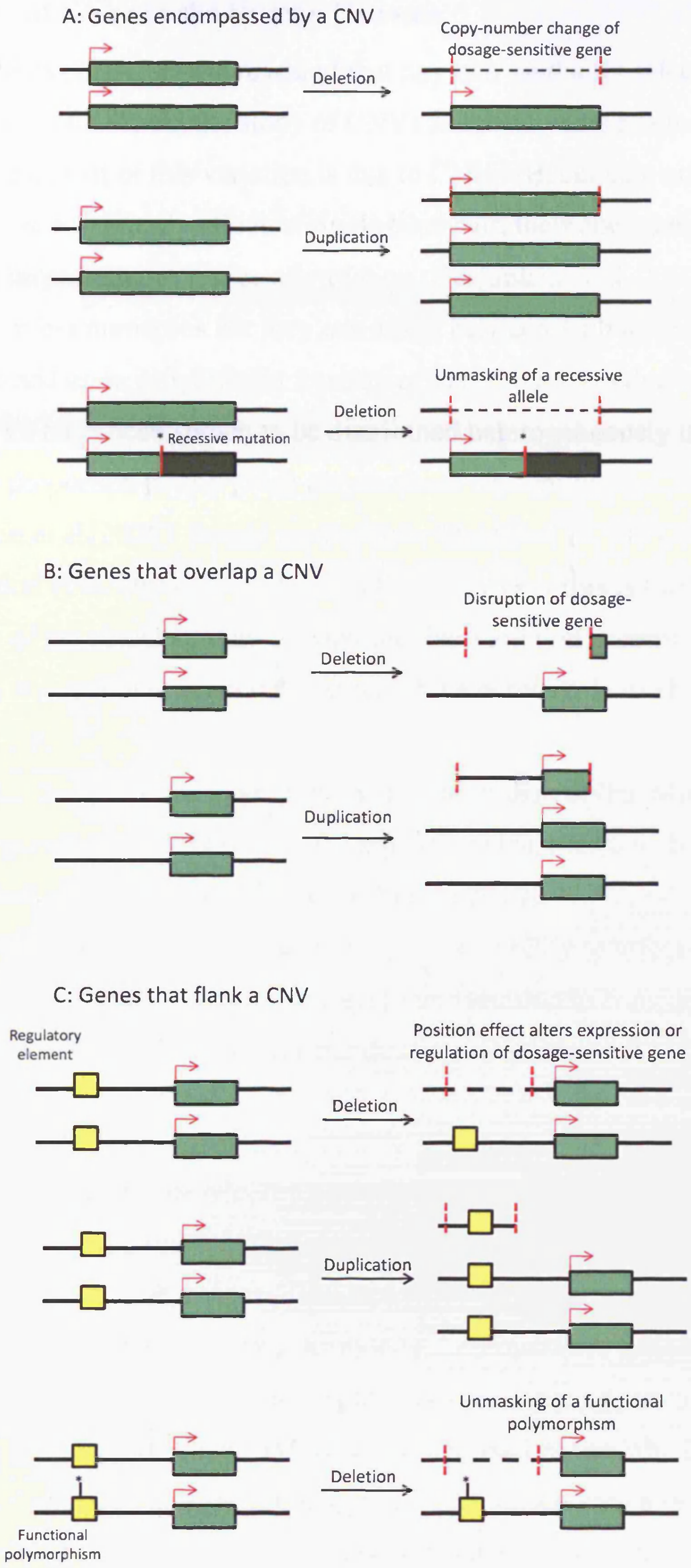

A: Genes encompassed by a CNV
Copy-number change of dosage-sensitive gene
Deletion
Duplication
Unmasking of a recessive allele
Deletion
Recessive mutation
B: Genes that overlap a CNV
Disruption of dosage-sensitive gene
Deletion
Duplication
C: Genes that flank a CNV
Regulatory element
Position effect alters expression or regulation of dosage-sensitive gene
Deletion
Duplication
Unmasking of a functional polymorphsm
Deletion
Functional polymorphism

Figure 7.2: Diagrams illustrating how CNVs can influence phenotypes when they A: encompass a whole gene, B: overlap a gene and C: flank a gene but overlap a functional polymorhphism. The red dotted lines indicate the endpoints of the CNVs. Adapted from Feuk et al. (2006).

7.1.4 Frequency of CNVs in the Human Genome

Previous studies of SNPs have revealed that any two randomly selected human genomes differed by 0.1%, but the study of CNVs in recent years has increased this estimate to 1% and most of this variation is due to CNVs (Beckmann et al. 2007; Redon et al. 2006). Whilst a SNP only affects a single base pair, their abundance in the genome makes them the most frequent source of variation (Beckmann et al. 2007). CNVs on the other hand are far less numerous but they can affect between 1 kb to several Mb of sequence and so add up to a significant fraction of the genome (Redon et al. 2006). Large-scale CNVs have been shown to be distributed heterogeneously throughout the genome and the proportion of any given chromosome susceptible to CNV varies from 6% to 9% (Redon et al. 2006). Recent studies have identified 11,700 CNVs overlapping 1000 genes (Redon et al. 2006; Conrad et al. 2010), however this is likely to be an underestimation of the actual amount of copy number variation present in the genome as current methods are unable to accurately detect CNVs between 1-50 kb in size (Redon et al. 2006).

CNVs are also thought to be more prone to mutation than SNPs. Mutation rates for genomic rearrangements range between 10^{-4} and 10^{-5}, at least 1000 to 10,000 fold more frequent than point mutations (Stankiewicz & Lupski 2010).

Altogether, their frequency, high mutation rates and ability to affect phenotypic variation make CNVs important sources of genomic variation to consider when identifying susceptibility variants for genetic diseases (Beckmann et al. 2007).

7.1.5 Identifying CNVs

As mentioned earlier, the development of new technologies has allowed us to investigate the genome at a much higher resolution, either in a genome-wide or a targeted manner. The main approaches for identifying structural variants have been array-based analyses (Feuk et al. 2006). Array-based comparative genome hybridisation (aCGH) approaches provide the most robust methods for carrying out genome-wide scans to identify CNVs (Pinkel et al. 1998). These approaches use labelled fragments from a genome of interest, which are competitively hybridised with a second differentially labelled reference genome to arrays that are spotted with cloned DNA fragments (e.g.bacterial artificial chromosome (BAC) clones). Differences in the fluorescent intensities identify regions of relative loss and gain in the test sample (Feuk et al. 2006; Carter 2007).

Another array-based approach to identifying CNVs is to use commercial genome-wide SNP-arrays. The combined intensity signals from both alleles at a particular SNP are measured and expressed as a $\log_2$ R ratio (LRR) between the recorded intensity and the expected intensity. The expected intensity is derived from the average intensity of the genotype clusters (Wain et al. 2009). Additionally, some studies use the ratio of fluorescence signals between allelic probes, termed the B allele frequency (BAF), since these ratios would be expected to be 0, 0.5 and 1.0 in the absence of CNVs. A large number of calling algorithms have been developed to turn these intensity signals into CNV calls and many of these, such as QuantiSNP (Colella et al. 2007) and PennCNV (Wang et al. 2007), use a hidden Markov model (HMM). This model segments the contiguous data into several predefined and biologically meaningful discrete states (see section 7.2.2 for more information on the PennCNV algorithm). However, the probes on these SNP arrays are not uniformly distributed across the genome and are particularly sparse in regions of segmental duplication, which creates problems for the design of robust genotyping SNP assays in these regions (Carter 2007). For example, a deletion might cause contiguous SNPs to show a loss of heterozygosity because hemizygous genotypes will be judged to be homozygous. This can cause deviation from the Hardy-Weinberg equilibrium or they may appear to violate Mendelian inheritance and so many SNPs in such regions were not included in early genome-wide genotyping arrays (Wain et al. 2009). As a result, the resolutions of these arrays vary across the genome and they particularly lack probe coverage in and around duplicated sequences (Cooper et al. 2008).

Assays for screening targeted regions of the genome are mainly PCR-based, the most commonly used of which has been real-time quantitative PCR (qPCR). This method enables the comparison of the relative quantification of a gene of interest with a reference gene known to be a single copy. Alternative PCR-based methods exist which allow simultaneous screening of multiple regions, such as quantitative multiplex of short fluorescent fragments (QMPSF), multiplex amplifiable probe hybridisation (MAPH) and multiplex ligation-dependent probe amplification (MLPA) (Feuk et al. 2006).

Sequencing-based approaches can map CNVs far more accurately than the above methods, providing resolution of CNV boundaries at the single nucleotide level and enabling the detection of inversions or translocations as well as deletions and duplications (Fanciulli et al. 2010). The DNA sequence of an individual of interest can be compared against a reference genome to identify structural variants and Khaja et al (2006) used this method to identify 13,066 previously undescribed structural variations in the human genome. With the development of next-generation sequencing in recent years, it has become possible to generate new assemblies of complete sequences from single individuals, enabling more robust and reliable genome comparisons and CNV identification (Fanciulli et al. 2010).

7.1.7 CNVs in Complex Neuropsychiatric Diseases

Large duplications and deletions have been associated with a number of specific genetic disorders for many years and this had led to the establishment of genetic diagnostic tests for certain microdeletion and microduplication syndromes such as DiGeorge syndrome and Charcot-Marie-Tooth disease (Freeman et al. 2006).

However, recent studies have also found CNVs that are associated with more complex diseases such as autism spectrum disorders (ASDs), schizophrenia and ADHD. ASDs comprise a heterogeneous group of neurodevelopmental abnormalities characterised by impairment of social interactions, problems in communication and a restricted range of behaviours and interests. The genetic cause of ASDs is only recognised in ~10% - 20% of cases (Stankiewicz & Lupski 2010). A number of CNV studies have shown the importance of CNVs in the aetiology of ASDs, particularly with *de novo* CNVs (Autism Genome Project Consortium 2007; Sebat et al. 2007; Marshall et al. 2008; Christian et al. 2008; Morrow et al. 2008; Weiss et al. 2008; Kumar et al. 2008; Glessner et al. 2009; Pinto et al. 2010). The most recent of these was conducted by Pinto and colleagues (2010) who compared the rates of rare (<1%) CNVs in 996 individuals with ASD against 1287 matched controls using the Illumina Human1M array. They found a significant burden of CNVs in the ASD cases and identified the rate of *de novo* CNVs to be 5.3%. These CNVs have implicated many novel genes for ASD (Pinto et al. 2010).

Schizophrenia is a severe psychiatric disorder characterised by hallucinations, delusions, cognitive deficits and apathy. Epidemiologic studies on twins indicate that schizophrenia has a complex genetic background with heritability estimated at 73% -

90% (Stankiewicz & Lupski 2010). The two largest CNV studies in schizophrenia were carried out by the International Schizophrenia Consortium (ISC 2008) and Stefansson et al. (2008). The ISC carried out their CNV analysis by screening 3391 patients with schizophrenia and 3181 controls from six European populations using the Affymetrix Human SNP 5.0 and 6.0 arrays and found an increased burden of CNVs in their cases (ISC 2008). Stefansson et al. (2008) identified 66 *de novo* CNVs in an Icelandic population using the Illumina HumanHap300 array and then examined their frequencies in a total of 4718 schizophrenia cases and 41,199 controls from nine European countries and China using a combination of the Illumina HumanHap300 and HumanHap550 arrays and the Affymetrix Human SNP 6.0 array. They found deletions associated with schizophrenia in the regions 1q21.1, 15q11.2 and 15q13.3. Interestingly, the ISC study (2008) also found association of deletions with schizophrenia in the 1q21.1 and 15q13.3 loci.

Two recent studies have also identified CNVs in individuals with ADHD. Elia and colleagues (2010) screened 335 individuals with ADHD using the Illumina HumanHap550 array and found 222 CNVs (158 deletions and 64 duplications) that had not been identified in the sample of 2026 healthy controls. They found that their ADHD CNV gene set was significantly enriched for genes that had been reported as candidates in studies of autism, schizophrenia and Tourette syndrome as well as for genes known to be important for psychological and neurological functions, including learning and central nervous system development. Lesch and colleagues (2010) screened 99 individuals with ADHD using aCGH, using a pool of 100 unscreened blood donors as reference DNA. They found 17 CNVs (4 deletions and 13 duplications) in cases that were not present in a total of 2726 screened controls. These ADHD CNVs could be of particular interest within the context of DD as a high rate of ADHD is observed in children with DD (Gilger et al. 1992; Shaywitz et al. 1995; Willcutt & Pennington 2000) and twin and family studies have suggested that there are shared genetic links between these disorders (Willcutt et al. 2000; Willcutt et al. 2007) (as discussed in Chapter 1). Therefore, any regions of the genome in which CNVs are found in both disorders could provide some information about the genetic overlap of these disorders.

From large CNV studies such as these, it has been noticed that some regions of the genome appear to be 'hotspots' for CNVs associated with a range of diseases. Two studies have identified a number of CNV hotspots across the genome. Mefford and

Eichler (2009) defined hotspots as 'regions predicted to be susceptible to recurrent rearrangement based on the flanking genomic architecture'. The criteria they used to identify these hotspots was a unique sequence (50kb-10Mb) flanked by large (>10kb), highly homologous (>95%) segmental duplications that provide the substrate for NAHR. They identified a total of 12 regions across the genome in which CNVs had been identified for a range of diseases, including schizophrenia, MR and autism (see Table E.3 in Appendix). Itsara and colleagues (2009) combined their CNV data from ~2500 healthy individuals with published CNVs from more than 12,000 individuals from other control and neurological disease collections (including schizophrenia, MR and autism) and identified 27 candidate neurological disease loci across the genome using the same definition of hotspots adopted by Mefford and Eichler (2009) (see Table E.3 in Appendix). As expected, many of the hotspot loci in these two studies share some degree of overlap.

7.1.8 Aims

CNVs could provide an additional source of genetic variation that has not yet been investigated within DD. The aim for this section of the thesis was to carry out CNV analysis using the Cardiff DD cases and the 1958 Birth Cohort that were genotyped in the first stage of the NeuroDys GWAS (see Chapter 4). The overall burden of CNVs in cases and controls were compared, and regions of the genome that harboured significantly more CNVs in the cases than the controls were identified. The findings were also compared with previous CNV studies to identify if any CNVs in the DD cases lay in the CNV disease 'hotspot' regions or overlapped associated CNVs in other diseases such as ADHD.

7.2 Methods

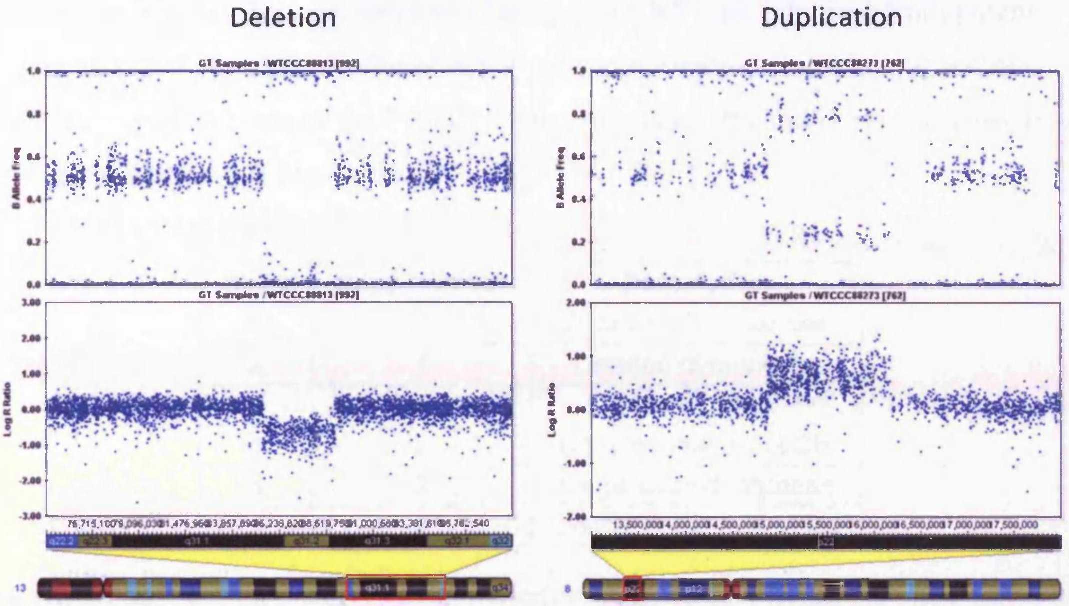
7.2.1 Sample

The sample used in this study consisted of the Cardiff subset of 178 cases that were genotyped on the Illumina HumanHap300 array and 1135 samples from the 1958 Birth Cohort (1958 BC) that were used in the initial NeuroDys GWAS sample, as discussed in Chapter 4.

As previously described in Chapter 2, the Cardiff case criteria is an IQ ≥ 85 and a reading age that is ≥ 2.5 years below their chronological age. DNA for these case samples were extracted from either blood or saliva samples using phenol/chloroform methodology as previously described (see Chapter 2). DNA quantification and dilution was also as described (Chapter 2), with a final sample dilution of 50ng/μl. As part of the NeuroDys GWAS, 102 of these cases were genotyped on the Illumina HumanHap300 array according to manufacturer's instructions at the University of Bonn while the rest of the cases (n = 76) were genotyped using the same array at Oxford University. The 1958 Birth Cohort had previously been genotyped on the Illumina HumanHap550 array. These samples all had SNP call rates > 97% in the GWAS. For this CNV study, a more relaxed call rate filter was used than had been previously implemented in the GWAS because SNPs within regions of CNVs are likely to have lower call rates. Using such a stringent call rate filter could therefore result in some true CNVs being missed.

7.2.2 Calling CNVs

In order to allow for intensity differences occurring between SNPs due to factors other than the presence of CNVs, it is important to carry out CNV analysis using normalised intensity signals. Intensity signals for these samples were normalised using Illumina BeadStudio v3.2. Because the performance of external controls can vary from sample to sample, Illumina have developed a self-normalisation algorithm that uses information contained within the array itself (www.illumina.com). This algorithm is designed to adjust for channel-dependant background and global intensity differences. These normalised intensities were then used to calculate the LRR and BAF of each marker in each sample using the BeadStudio software. The LRR is a measure of the total fluorescent intensity from both sets of alleles at each SNP and the BAF is a measure of the relative ratio of the fluorescent signals between two alleles at each SNP figures. Examples of the plots of these measures are shown in Figure 7.3.


Deletion

Duplication

Figure 7.3: Examples of BAF (top plots) and LRR plots (bottom plots) for a deleted region (left side) and a duplication region (right side)

The LRRs and BAFs were then used to generate CNV calls using the PennCNV software (2009 Aug 27 version), applying the GC-model wave adjustment (Wang et al. 2007). As the cases and controls were genotyped using different arrays, the CNV calling was carried out for each sample group separately. The PennCNV software detects CNVs using a HMM based approach and uses a six-state definition to model CNV states, as shown in Table 7.1. This algorithm attempts to exploit all the available information for each SNP by incorporating the LRR and BAF together with the distance between neighbouring SNPs and the population frequency of the B allele into the HMM. Incorporating the distance between neighbouring SNPs enables the probability of having a copy number state change between them to be determined. The population B allele frequency for each SNP had been calculated using a large set of individuals with mixed ethnic backgrounds and of normal phenotypes and enables the determination of the likelihood of the copy number genotypes for each copy number state. PennCNV also allows the user to adjust the data for ‘genomic waves’. These waves refer to variations in hybridisation intensity which show high correlation with DNA quantity and GC content (Diskin et al. 2008). The GC adjustment procedure implemented in PennCNV is a regression model which corrects and adjusts for genomic waves (Diskin et al. 2008).

By assuming that the vast majority of offspring CNVs are inherited from parents, Wang et al. (2007) used family-based CNV calls as a reference to indirectly estimate that the false-positive rate of the PennCNV algorithm is 1.0% and it has a sensitivity of 82.2% in the absence of family data.

State	Copy Number	Description
1	0	Deletion of two copies
2	1	Deletion of one copy
3	2	Normal state
4	2	Copy-neutral with LOH
5	3	Single copy duplication
6	4	Double copy duplication

Table 7.1: Six-state definition of CNVs used by PennCNV (adapted from Wang et al. 2007). LOH – Loss of heterozygosity.

7.2.3 Sample QC Filtering

Intensity data from poor quality DNA samples are likely to produce more false positive CNV calls. A high LRR standard deviation (SD) can indicate large variability between the signal intensities from the SNPs and so can be used to filter out poorly performing samples. Removing those samples that have high LRR SDs and high CNV call rates improves the quality of the data and reduces the number of false positives. In order to identify outlying samples, histograms of the raw CNVs called by PennCNV were produced showing the LRR SDs in each sample (shown in Figure 7.4) and the number of CNVs identified in each sample (shown in Figure 7.5, zoomed in to those samples with fewer than 200 CNVs). Based on these histograms, samples with LRR SDs > 0.3 and those that harboured more than 30 CNVs were removed.

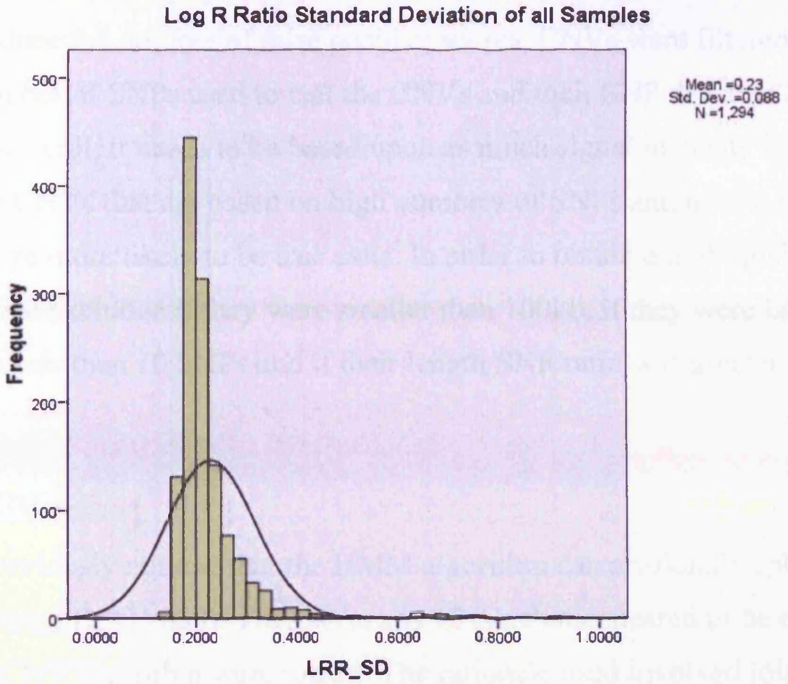


Figure 7.4: Histogram of the Log R Ratio standard deviations of all samples before QC filtering

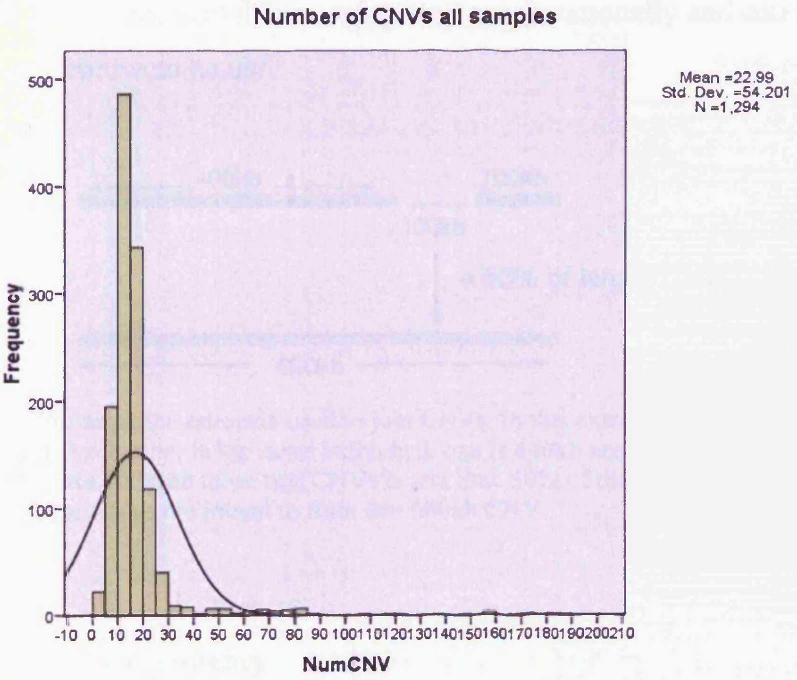


Figure 7.5: Histogram of the number of CNVs in all samples before QC filtering. The scale has been adjusted to show only those samples that had fewer than 200 CNVs in order to identify the cut off more clearly.

7.2.4 CNV QC Filtering

In order to reduce the number of false positive scores, CNVs were filtered based on their size, the number of SNPs used to call the CNVs and their SNP density. To be confident in a CNV call, it needs to be based upon as much signal intensity information as possible; large CNVs that are based on high numbers of SNPs and have a high density of SNPs are more likely to be true calls. In order to obtain a high-quality dataset, CNVs were excluded if they were smaller than 100kb, if they were based on information from less than 10 SNPs and if their length/SNP ratio was greater than 30kb per SNP.

7.2.5 Merging CNVs

It has been previously noticed that the HMM algorithm can artificially split CNVs into smaller segments (ISC 2008). Therefore, any CNVs that appeared to be artificially split by the PennCNV algorithm were joined. The rationale used involved joining two CNVs if the length of the sequence between them was less than 50% of the length of the larger CNV, as shown in Figure 7.6. A programme has been created by Dobril Ivanov which enables the user to perform this merging step computationally and can be found here: http://x001.psycm.uwcm.ac.uk/.

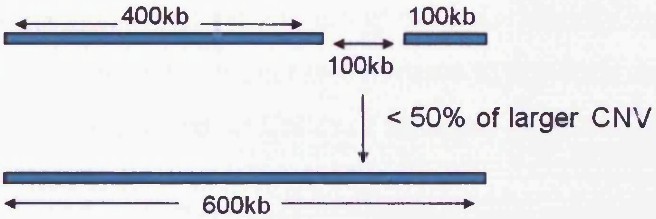


Figure 7.6: Diagram illustrating the rationale used to join CNVs. In this example, two CNVs of the same type are identified next to each other in the same individual, one is 400kb and the other is 100kb and lies 100kb away. As the distance between these two CNVs is less than 50% of the length of the larger CNV (i.e. less than 200kb), these CNVs are joined to form one 600kb CNV.

7.2.6 Filtering CNVs on Frequency

For this study, analysis was restricted to rare CNVs that occurred in less than 1% of the total sample. Common CNV regions are not likely to be adequately covered by the SNP arrays as markers that lie within common CNV regions are likely to show departure from Mendelian inheritance or Hardy Weinberg equilibrium, causing them to be excluded from the SNPs arrays (Beckmann et al. 2007). As such, common CNVs are not thought to be reliably called and past CNV studies have focused on rare (i.e. < 1% frequency) CNVs. In addition, it is likely that many of the common CNVs are tagged by the SNPs on this array and so any association of common CNVs with DD should have been picked up in the initial GWAS (Conrad et al. 2010; WTCCC 2010). It has been estimated that even on first generation SNP arrays, 40-50% of common copy number variants (MAF >5%) were tagged (r^2 >0.8) and this proportion has increased with the newest arrays (~ 65% for the Illumina 1M array) (McCarroll et al. 2008). Rare CNVs are also considered to be more likely to be pathogenic (WTCCC 2010).

7.2.7 Statistical Analysis

The CNV association analyses were carried out using PLINK v1.06 (Purcell et al. 2007). P-values given are 2-tailed, based on comparing the rates of CNVs in cases and controls with the use of 10,000 permutations. The genomic coordinates used in this study are based on the March 2006 human genome sequence assembly (UCSC hg18, National Centre for Biotechnology Information build 36).

7.2.7.1 Burden Analysis

Burden analysis involves performing a global association test of CNV burden in cases and controls to identify if either sample group has a statistically higher rate of CNVs. As larger CNVs have shown a higher rate in cases in previous studies (e.g ISC 2008), burden analysis was performed on CNVs of all sizes, as well as those greater than 500kb and greater than 1Mb.

7.2.7.2 Regional Analysis

In order to identify specific loci that showed a significant excess of CNVs in cases compared with controls, overlapping CNVs were grouped using PLINK. These groups were then tested for a significant excess of CNVs in the cases compared with the controls.

To investigate whether or not any of the genes that were overlapped by CNVs within the significant regions identified had shown association in any of the previous studies, association results for SNPs within these genes (plus 10kb upstream and downstream) were extracted from the previous NeuroDys GWAS, NeuroDys pooling study and the Cardiff pooling study.

7.2.8 Proof of Principle Test Using HapMap Samples

The controls in this dataset were genotyped on the Illumina HumanHap550 array whereas the cases were genotyped using the HumanHap300 array. As these arrays differ in the number of SNPs, this could affect the number of CNVs that are identified in each sample. As the controls were genotyped on the larger chip, this is likely to result in the data being biased to having more CNVs in the controls than in the cases. In order to test if this difference in genotyping platform is likely to cause a significant difference in the rate of CNVs identified, 120 samples from the HapMap CEU population which had been genotyped on both array were compared.

BeadStudio projects for these samples when genotyped on each platform were downloaded from the Illumina ftp website (http://www.illumina.com/forms/ftp.ilmn). The LRRs and BAFs were extracted and were run through PennCNV. As with the case control samples, data from each array type were called separately from each other. The same QC filters were applied to the HapMap samples as were applied to the case control dataset, but samples were only included in further analysis if they passed QC in both the 300 array dataset and the 550 array dataset. CNVs at a frequency greater than 1% were excluded, as before.

7.2.9 Comparison with Other CNV Studies

7.2.9.1 ADHD CNV Studies

The ADHD CNV regions identified in the studies by Elia et al (2010) and Lesch et al (2010) were investigated in this DD dataset and CNVs were reported if they overlapped these regions at all in either the cases or controls (see Table E.2 in Appendix for a list of these regions). Association analysis was then performed comparing the rates of these CNVs in the DD cases with the 1958 Birth Cohort controls.

7.2.9.2 Disease Hotspots

Any CNVs that overlapped the regions identified as CNV 'hotspots' by Mefford and Eichler (2009) and Itsara and colleagues (2009) (see table E.3 in Appendix for a list of these regions) in this DD dataset were reported and association analyses were performed comparing their rates in the DD cases vs. the 1958 Birth Cohort controls.

7.3 Results

7.3.1 QC Filtering

After CNV calling using PennCNV, a total of 29,745 CNVs were identified in 159 cases and 1135 controls. After applying QC filters to the samples, 120 cases (of which 74 had been genotyped in Bonn and 46 had been genotyped in Oxford) and 1030 control samples remained, with 16,063 CNVs. Figure 7.7 shows the LRR SDs for those samples passing QC, separated by sample group. This plot shows very similar mean values between the population controls (mean LRR SD = 0.21) and those cases that were genotyped in Bonn (mean LRR SD = 0.20), but those samples genotyped in Oxford tended to have higher SDs (mean LRR SD = 0.30). This could indicate some form of inter-centre experimental variability existing between the intensity scores from those samples that were genotyped in Bonn and those that were genotyped in Oxford.

After QC filtering and merging, 1148 CNVs remained, 163 of which were in cases. The intensity plots of the 9 CNVs that were larger than 2 Mb were manually inspected. The largest of these was a 33 Mb duplication in one of the controls that covered the majority of chromosome 21, which could indicate a that this individual had Down's syndrome, and so this sample was removed from the analysis, leaving 1147 CNVs and 1030 controls. The number of CNVs that passed QC in each sample is shown in Figure 7.8. On average, more CNVs were identified in those samples which were genotyped in Oxford (average CNV count = 3.24) than those genotyped in Bonn (average CNV count = 1.84) and those within the 1958 Birth Cohort (average CNV count = 2.06). This could be a reflection of the higher LRR SDs that were observed in those samples genotyped in Oxford.

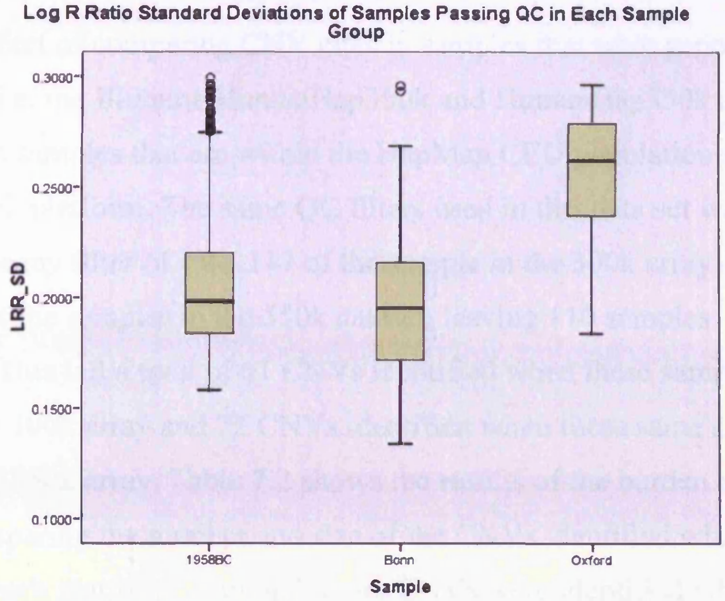


Figure 7.7: Box plots of the Log R Ratio standard deviations of samples that passed QC within each sample group.

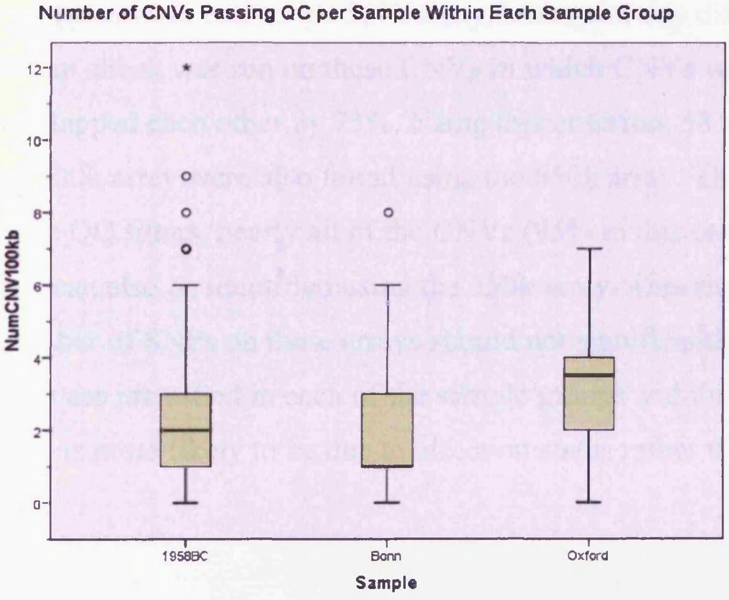


Figure 7.8: Box plot showing the number of CNVs identified in samples that passed QC within each sample group.

CNVs that were present in more than 1% of the whole sample were excluded, leaving 1147 CNVs, of which 163 were in cases (see Table E.1 in Appendix for a list of all CNVs passing QC). The average length of these CNVs was 293 kb in cases and 277 kb in controls.

7.3.2 Proof of Principle Test Using HapMap Samples

To test the effect of comparing CNV rates in samples that were genotyped using different arrays (i.e. the Illumina HumanHap300k and HumanHap550k arrays), CNVs were identified in samples that are within the HapMap CEU population and had been genotyped on each platform. The same QC filters used in this data set were applied, including a frequency filter of 1%. 117 of the sample in the 300k array dataset passed QC, as did 110 of the samples in the 550k dataset, leaving 110 samples that passed QC in both datasets. This left a total of 61 CNVs identified when these samples were genotyped on the 300k array and 72 CNVs identified when these same samples were genotyped on the 550k array. Table 7.2 shows the results of the burden analyses on these CNVs, comparing the number and size of the CNVs identified when using intensities from each platform. Although more CNVs were identified when genotyping these samples on the 550k array, there was no significant difference in the number of CNVs called for any of the size ranges or CNV types (i.e. deletions or duplications). As would be expected due to the larger number of SNPs, CNVs identified were on average larger when these samples were run on the 550k array, but again, this difference was not significant. An overlap check was run on these CNVs in which CNVs were classed as the same if they overlapped each other by 75%. Using this criterion, 58 of the CNVs identified using the 300k array were also found using the 550k array. This suggests that when using the above QC filters, nearly all of the CNVs (95% in this case) identified using the 300k array can also be identified using the 550k array. This shows that the difference in the number of SNPs on these arrays should not significantly affect the number of CNVs that are identified in each of the sample groups and that any significant difference is more likely to be due to affection status rather than the genotyping platform.

	Deletions					Duplications					Deletions and Duplications				
	Number	Rate	Rate P	Av. Size (kb)	Size P	Number	Rate	Rate P	Av. Size (kb)	Size P	Number	Rate	Rate P	Av. Size (kb)	Size P
> 100kb															
300k chip	22	0.20	0.545	199.2	0.741	36	0.33	1	346.3	0.514	58	0.5273	0.418	291	0.760
550k chip	28	0.25		213		41	0.37		425.9		69	0.6273		319.9	
> 500kb															
300k chip	0	0	1	0	1	7	0.06	1	965.8	0.401	7	0.06364	1	965.8	0.514
550k chip	1	0.01		579.8		7	0.06		1345		8	0.07273		1249	
> 1Mb															
300k chip	0	0	1	0	1	3	0.03	1	1480	0.572	3	0.02727	1	1480	0.571
550k chip	0	0		0		4	0.04		1857		4	0.03636		1857	

Table 7.2: Results of burden analysis between CNVs in the HapMap CEPH samples when genotyped on the 300k and 550k arrays. Av – average; Rate P – P-value when the rates of CNVs are compared between the two arrays; Size P – P-value when the average sizes of the CNVs are compared between the two arrays.

7.3.3 Burden Analyses

7.3.3.1 All Cases Compared With 1958 Birth Cohort

When comparing the rates of CNVs in DD cases with those in the 1958 Birth Cohort controls, there was a significant excess of all CNVs in the cases ($P = 3 \times 10^{-4}$, case/control ratio = 1.42), as shown in Table 7.3. This excess appeared to be largely driven by deletions, which were present in cases at double the rate found in the controls ($P = 1 \times 10^{-4}$, case/control ratio = 2.00). When looking at duplications alone, these were present at a higher rate in the controls but the difference was not significant (P = 0.207, case/control ratio = 0.81). When focusing the burden analysis on larger CNVs, this excess of deletions in cases was no longer significant (P = 0.286, case/control ratio = 1.83). However, the case/control ratio is still similar to the ratio for smaller CNVs which suggests that the lack of significance may be due to the small number of large CNVs identified. There was no significant difference in the sizes of CNVs found in cases compared with controls (P = 0.646), including when looking at deletions (P = 0.594) and duplications (P = 0.325) separately.

7.3.3.2 Cases Genotyped in Bonn Compared with Those Genotyped in Oxford

As the two Cardiff case subsets were genotyped on the arrays in different centres, burden analysis was carried out between the subsets to investigate whether or not the excess of CNVs in cases could be attributable to experimental variation between centres rather than a true excess of CNVs in DD cases. The results are shown in Table 7.4. When comparing the rate of CNVs found in Cardiff cases genotyped in Oxford with those identified in Cardiff cases genotyped in Bonn, there was a greater than 2-fold excess of CNVs in those cases genotyped in Oxford ($P = 1 \times 10^{-4}$, Oxford/Bonn ratio = 2.25). Again this overall excess appeared to be down to deletions in particular ($P = 1 \times 10^{-4}$, Oxford/Bonn ratio = 2.73) and there was no significant difference in the rate of duplications (P = 0.332, Oxford/Bonn ratio = 1.44). Each of the case sample groups were then compared separately with the controls, and a significant excess of deletions was found in the Oxford-typed cases ($P = 1 \times 10^{-4}$, Oxford cases/controls ratio = 3.28), but no significant excess was observed when comparing the rate of deletions identified in the Bonn-typed cases with those identified in the controls (P = 0.324, Bonn cases/controls ratio = 1.20). The only significant result when comparing the Bonn-typed cases with the controls was for the average sizes of the duplications, which showed a trend towards being larger in the Bonn-typed cases (P = 0.047).

Therefore technical issues appear to be driving the significant excess of CNVs seen in the complete group of cases compared with the controls.

	Deletions					Duplications					Deletions and Duplications				
	Number	Rate	Rate P	Av. Size (kb)	Size P	Number	Rate	Rate P	Av. Size (kb)	Size P	Number	Rate	Rate P	Av. Size (kb)	Size P
> 100kb															
Cases	118	0.983	**1×10^{-4}**	238.1	0.594	45	0.375	0.207	380.9	0.325	163	1.358	**3×10^{-4}**	293	0.646
Controls	507	0.492		215.9		477	0.463		332.2		984	0.955		277	
> 500kb															
Cases	4	0.033	0.286	1253	0.968	10	0.083	1.000	1033	0.314	14	0.117	0.675	1096	0.477
Controls	19	0.018		1285		86	0.084		876.2		105	0.102		949	
> 1Mb															
Cases	1	0.008	1.000	2622	0.949	4	0.033	0.503	1555	0.726	5	0.042	0.385	1769	0.846
Controls	7	0.007		2393		20	0.019		1413		27	0.026		1660	

Table 7.3: Results of burden analysis between CNVs in cases and controls. Significant P-values are in bold. Av – average; Rate P – P-value when the rates of CNVs are compared between cases and controls; Size P – P-value when the average sizes of the CNVs are compared between cases and controls.

	Deletions					Duplications					Deletions and Duplications				
Sample Groups	**Number**	**Rate**	**Rate P**	**Av. Size (kb)**	**Size P**	**Number**	**Rate**	**Rate P**	**Av. Size (kb)**	**Size P**	**Number**	**Rate**	**Rate P**	**Av. Size (kb)**	**Size P**
Oxford Cases	74	1.61	**1×10^{-4}**	228.7	0.752	21	0.46	0.332	291.9	0.167	95	2.07	**1×10^{-4}**	263	0.299
Bonn Cases	44	0.59		249.3		24	0.32		455.1		68	0.92		320	
Oxford Cases	74	1.61	**1×10^{-4}**	228.7	0.853	21	0.46	1.000	291.9	0.593	95	2.07	**1×10^{-4}**	263	0.803
1958 BC Controls	507	0.49		215.9		477	0.46		332.2		984	0.96		277	
Bonn Cases	44	0.59	0.324	249.3	0.498	24	0.32	0.089	455.1	**0.047**	68	0.92	0.783	320	0.317
1958 BC Controls	507	0.49		215.9		477	0.46		332.2		984	0.96		277	

Table 7.4: Results of burden analysis of CNVs between cases typed in Bonn and Oxford, cases typed in Oxford vs. 1958 Birth Cohort controls and between cases typed in Bonn vs. 1958 Birth Cohort Controls. Significant P-values are in bold. Av – average; Rate P – P-value when the rates of CNVs are compared between cases and controls; Size P – P-value when the average sizes of the CNVs are compared between cases and controls.

7.3.3.3 Effect of Sample Origin on CNV Burden

To test if the type of sample that the DNA was extracted from (i.e. blood or saliva, referred to hereafter as the 'sample origin') affected the number of CNVs identified, burden analysis was carried out on the small subset of samples for which their origin had been recorded (n = 61, 36 from blood and 25 from saliva). The results are shown in Table 7.5. When comparing the rates of CNVs between each sample origin, although more CNVs were identified in those samples that were extracted from saliva, this difference was not significant (P = 0.167, saliva/blood ratio = 1.78), nor was there any significant difference when looking at duplications (P = 0.808, saliva/blood ratio = 0.79). However, there was a trend towards a significant excess of deletions in those samples extracted from saliva, which had a deletion rate which was over double that found in those extracted from blood (P = 0.058, saliva/blood ratio = 2.41). This suggests that DNA extracted from saliva samples may have a higher rate of deletions than those extracted from blood due to technical artefacts. There was no significant difference between the average sizes of the CNVs identified in these samples (P = 0.238).

As all of the samples known to be extracted from blood were genotyped in Bonn and those samples genotyped in Oxford showed a significant excess of CNVs, sample origin burden analysis was also carried out using just those cases that were genotyped in Bonn in order to allow for any experimental variation that may have existed between the centres. There was no overall significant excess of CNVs in samples from either origin (P = 0.786, saliva/blood ratio = 1.14). There was a higher rate of deletions in those samples extracted from saliva (saliva/blood ratio = 1.60) and a higher rate of duplications in those samples extracted from blood (saliva/blood ratio = 0.39), but these differences were not significant (P = 0.401 and 0.219, respectively). There was also no significant difference between the average sizes of the CNVs in each sample origin group (P = 0.534).

	Deletions					Duplications					Deletions and Duplications				
Sample Type	**Number**	**Rate**	**Rate P**	**Av. Size (kb)**	**Size P**	**Number**	**Rate**	**Rate P**	**Av. Size (kb)**	**Size P**	**Number**	**Rate**	**Rate P**	**Av. Size (kb)**	**Size P**
Oxford and Bonn															
Saliva (n = 36)	38	1.06	0.058	219.2	0.688	8	0.22	0.808	230.2	0.181	46	1.28	0.167	229.1	0.238
Blood (n = 25)	11	0.44		204.5		7	0.28		549.1		18	0.72		306.9	
Bonn Only															
Saliva (n = 27)	19	0.70	0.401	469	0.302	3	0.11	0.219	280.8	0.464	22	0.82	0.786	257.7	0.534
Blood (n = 25)	11	0.44		218.1		7	0.28		549.1		18	0.72		306.9	

Table 7.5: Comparison of CNVs in DNA samples extracted from saliva and blood for both all the cases and just those genotyped in Bonn. Av – average; Rate P – P-value when the rates of CNVs are compared between cases and controls; Size P – P-value when the average sizes of the CNVs are compared between cases and controls.

7.3.4 Regional Analysis

No CNVs were found to be overlapping the putative DD susceptibility genes, *KIAA0319, DCDC2,* and *DYX1C1*. Regional analysis was carried out to identify any regions across the autosomal chromosomes that showed a significant excess of CNVs in cases. The results are shown in Table 7.6. 18 regions showed a significant excess of CNVs in the cases, the most significant of these was on chromosome 15q23 shown in Figure 7.9, with 5 deletions identified in the cases (rate = 0.042) and 2 deletions identified in the controls (rate = 0.002).

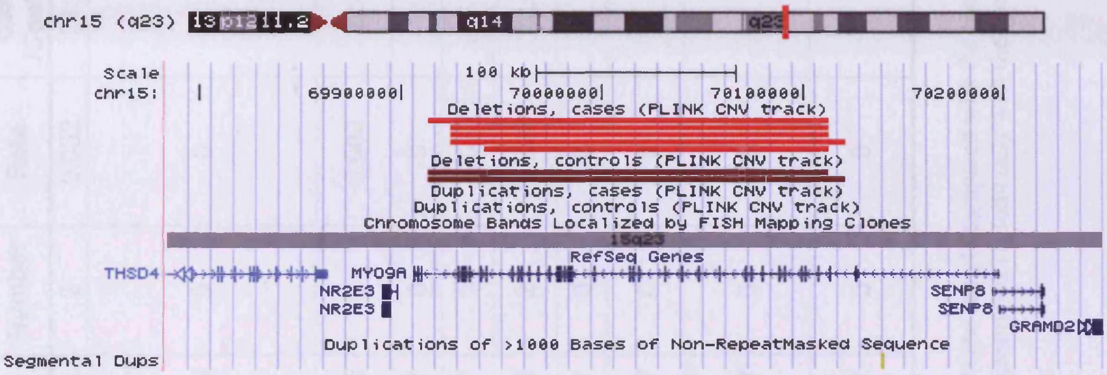

chr15 (q23)
13 p12 11.2
q14
q23
Scale
chr15:
100 kb
69900000|
70000000|
70100000|
70200000|
Deletions, cases (PLINK CNV track)
Deletions, controls (PLINK CNV track)
Duplications, cases (PLINK CNV track)
Duplications, controls (PLINK CNV track)
Chromosome Bands Localized by FISH Mapping Clones
15q23
RefSeq Genes
THSD4
MYO9A
NR2E3
NR2E3
SENP8
SENP8
GRAMD2
Duplications of >1000 Bases of Non-RepeatMasked Sequence
Segmental Dups

Figure 7.9: UCSC track showing the deletions identified within the 15q23 locus with the deletions in the cases depicted as the lighter red bars and those in the controls as the darker red bars. RefSeq genes within this region are also shown.

All of the cases that had CNVs within this locus had been genotyped in Oxford. As this sample group showed higher LRR SDs and significantly higher CNV rates than those that were genotyped in Bonn, the regional analysis was repeated comparing only those cases that had been genotyped in Bonn with the 1958 Birth Cohort to investigate whether or not any of these significant loci would remain significant. The results are shown in Table 7.7.

	Significant Region				Cases (n = 120)		Controls (n = 1030)				
Locus	**Start bp**	**End bp**	**Length (bp)**	**Type**	**Number**	**Rate**	**Number**	**Rate**	**Cases /Controls**	**P**	**Genes Intersected by CNVs**
15q23	69941054	70112649	171595	Del	5	0.042	2	0.002	21.46	0.0002	***MYO9A***
16p13.3*	651906	686885	34979	Del	3	0.025	0	0		0.0016	*RAB11FIP3, C16orf11,* ***SOLH****, NHLRC4, PIGQ, RAB40C, WFIKKN1, C16orf13, FAM195A, WDR90, RHOT2,* ***STUB1****, WDR24, JMJD8, FBXL16, RHBDL1,* ***METRN****, HAGHL, NARFL, FAM173A, MSLN, MSLNL, MIR662, RPUSD1, CHTF18*
6q14.3	86602842	86718986	116144	Del	3	0.025	2	0.002	12.88	0.0095	*NT5E, SNX14, SYNCRIP, SNORD50A, SNORD50B, SNHG5*
7q36.1	151667867	151701091	33224	Del	2	0.017	0	0		0.0099	*MLL3, FABP5L3, LOC100128822*
15q22.31	63587909	63776413	188504	Del	2	0.017	0	0		0.0102	*DPP8, PTPLAD1, C15orf44, SLC24A1, DENND4A*
14q21.1-q21.3*	42981353	43306266	324913	Dup	3	0.025	2	0.002	12.88	0.0104	-
4q35.1	186489132	186618822	129690	Del	2	0.017	0	0		0.0108	*SNX25, LRP2BP, UFSP2, ANKRD37, C4orf47, CCDC110*
19p13.3*	1317545	1364574	47029	Del	2	0.017	0	0		0.0111	*CDC34, GZMM,* ***BSG****, HCN2, POLRMT, FGF22, RNF126, FSTL3, PRSSL1, C19orf21*
16p13.3*	1325651	1442858	117207	Del	2	0.017	0	0		0.0112	***CACNA1H****, TPSG1, TPSB2, TPSAB1, TPSD1, UBE2I,* ***BAIAP3****, C16orf42, GNPTG, UNKL, CCDC154, C16orf91, CLCN7*
19p13.3	556985	602852	45867	Del	2	0.017	0	0		0.0115	*MUM1,* ***NDUFS7****,* ***GAMT****, DAZAP1, RPS15, APC2, C19orf25, PCSK4, REEP6, ADAMTSL5, PLK5P*

Table 7.6 Regions of the genome showing a significant excess of CNVs in the whole DD case sample compared with the controls and the genes that are intersected by these CNVs. * Indicates those regions that remain significant when comparing CNVs in cases genotyped in Bonn only with the controls (see Table 7.7). Functionally interesting genes are in bold.

	Significant Region				Cases (n = 120)		Controls (n = 1030)				
Locus	Start bp	End bp	Length (bp)	Type	Number	Rate	Number	Rate	Cases /Controls	P	Genes Intersected by CNVs
18q22.3	70700538	70779005	78467	Del	2	0.017	0	0		0.0117	*ZNF407*
2p16.3	48351619	48481460	129841	Del	2	0.017	0	0		0.0119	*FOXN2, KLRAQ1*
12q24.31	121565488	121682631	117143	Del	2	0.017	0	0		0.0119	*ZCCHC8, RSRC2, KNTC1*
17q25.1*	69345596	70164827	819231	Dup	2	0.017	0	0		0.0127	*RPL38, MGC16275, TTYH2, DNAI2, KIF19, BTBD17, GPR142, GPRC5C, CD300A, CD300LB, CD300C, C17orf77, CD300LD, CD300E*
6q26	162644237	162769931	125694	Dup	3	0.025	3	0.003	8.58	0.019	***PARK2***
11p11.2	46350333	46480589	130256	Del	2	0.017	1	0.001	17.17	0.0291	*CREB3L1, DGKZ, MDK, CHRM4,* ***AMBRA1****, HARBI1, KIAA0652*
9p21.1*	28288064	28332179	44115	Del	2	0.017	1	0.001	17.17	0.0297	*LINGO2*
1p21.1*	102439976	102571645	131669	Del	2	0.017	1	0.001	17.17	0.0305	-

Table 7.6 continued.

	Significant Region				**Cases** (n =74)		**Controls** (n = 1030)			
Locus	**Start bp**	**End bp**	**Length**	**Type**	**Number**	**Rate**	**Number**	**Rate**	**Cases/Controls**	**P**
15q23	69941054	70112649	171595	Del	0	0	2	0.002		1.0
16p13.3	651906	686885	34979	Del	2	0.027	0	0		**0.004**
6q14.3	86602842	86718986	116144	Del	0	0	2	0.002		1.0
7q36.1	151667867	151701091	33224	Del	1	0.014	0	0		0.069
15q22.31	63587909	63776413	188504	Del	0	0	0	0.000		1.0
14q21.1-q21.3	42981353	43306266	324913	Dup	2	0.027	2	0.002	13.92	**0.025**
4q35.1	186489132	186618822	129690	Del	0	0	0	0		1.0
19p13.3	556985	602852	45867	Del	1	0.014	0	0		0.067
16p13.3	1325651	1442858	117207	Del	2	0.027	0	0		**0.004**
18q22.3	70700538	70779005	78467	Del	1	0.014	0	0		0.066
19p13.3	1317545	1364574	47029	Del	2	0.027	0	0		**0.004**
2p16.3	48351619	48481460	129841	Del	0	0	0	0		1.0
12q24.31	121565488	121682631	117143	Del	1	0.014	0	0		0.070
17q25.1	69345596	70164827	819231	Dup	2	0.027	0	0		**0.004**
6q26	162644237	162769931	125694	Dup	1	0.014	3	0.003	4.64	0.252
11p11.2	46350333	46480589	130256	Del	0	0	1	0.001		1.0
9p21.1	28288064	28332179	44115	Del	2	0.027	1	0.001	27.84	**0.012**
1p21.1	102439976	102571645	131669	Del	2	0.027	1	0.001	27.84	**0.014**

Table 7.7: Comparison of CNVs in cases genotyped in Bonn only compared with controls for those regions that showed a significant excess of CNVs in all cases when compared with controls (see Table 7.6). Regions of the genome a significant excess of CNVs in the whole DD case sample compared with the controls. Seven regions remain significant as indicated by the P-values < 0.05 that are in bold.

Seven of the loci remained significant when only including CNVs from those samples genotyped in Bonn, and these did not include the most significant region from the original analysis. Of the seven loci that remained significant, the four most significant loci consisted of 2 regions on 16p13.3, one on 19p13.3 and the fourth on 17q25.1, all of which had P-values of 0.004. Within both of the 16p13.3 and the 19p13.3 loci, 2 deletions were identified in the cases and none were found in the controls (as shown in Figures 7.10 and 7.11) and within the 17q25.1 locus 2 duplications were identified with no CNVs in the controls (Figure 7.12). The two significant loci within 16p13.3 are nearly 630kb away from each other and three deletions that flank these regions were identified in the controls.

Table 7.8 shows the genes that are overlapped by CNVs in the first 16p13.3 CNV locus, together with their coverage within the NeuroDys GWAS (Chapter 4), NeuroDys pooling study (Chapter 5) and within the Cardiff pooling study (Chapter 6) as well as the SNPs that showed the most significant association within each gene (plus 10kb of sequence in either direction) in each study. Three genes within this locus had significant results in the initial NeuroDys GWAS, and 2 of these also showed significance in the Cardiff pooling study. The SNP rs2038227 within intron 5 of the gene RAB11 family interacting protein 3 (class II) (*RAB11FIP3*) showed significant association in the initial GWAS (P = 0.029) and another SNP just 1.5kb upstream of this gene (rs3760048) gene showed significant association in the Cardiff pooling study (0.008). Both of these SNPs are within a deletion that was identified in an individual diagnosed with DD. A SNP within both the chromosome 16 open reading frame 11 (*C16orf11*) and small optic lobes homolog (*SOLH*) genes was significantly associated with DD in the initial NeuroDys GWAS (rs7763, P = 0.012). Another SNP within *SOLH* was also significantly associated in the Cardiff pooling study (rs9934705, P = 0.030) and both of these SNPs were overlapped by a single deletion identified in an individual with DD. None of the significant SNPs in this region are in high LD with each other ($r^2 < 0.8$). Within this region, *SOLH* could be of particular interest as it is thought to be involved in protein-protein interactions during visual development and shows high expression in the brain (Kamei et al. 1998). Other genes within this region which could be of functional interest but were not significantly associated in any of the GWAS or pooling studies include the gene STIP1 homology and U-box containing protein 1 (*STUB1*). This gene was overlapped by two deletions which were identified in

DD cases (see Figure 7.10) and it encodes the U-box family ubiquitin protein ligase CHIP (carboxy terminus of Hsc70-Interacting Protein) which has been shown to interact with the DD susceptibility gene *DYX1C1* (Hatakeyama et al. 2004). CHIP participates in the degradation of estrogen receptors alpha (ERα) and beta (ERβ) which are important in brain development and cognition functions and this led Massinen and colleagues (2009) to propose that *DYX1C1* may affect brain development through the regulation of ERα and ERβ. Another functionally interesting gene within this locus is Meteroin (*METRN*) which is thought to be involved in both glial cell differentiation and axonal network formation during neurogenesis (Nishino et al. 2004) and so may have a role within neuronal migration. This gene was overlapped by one deletion identified in an individual with DD (Figure 7.10).

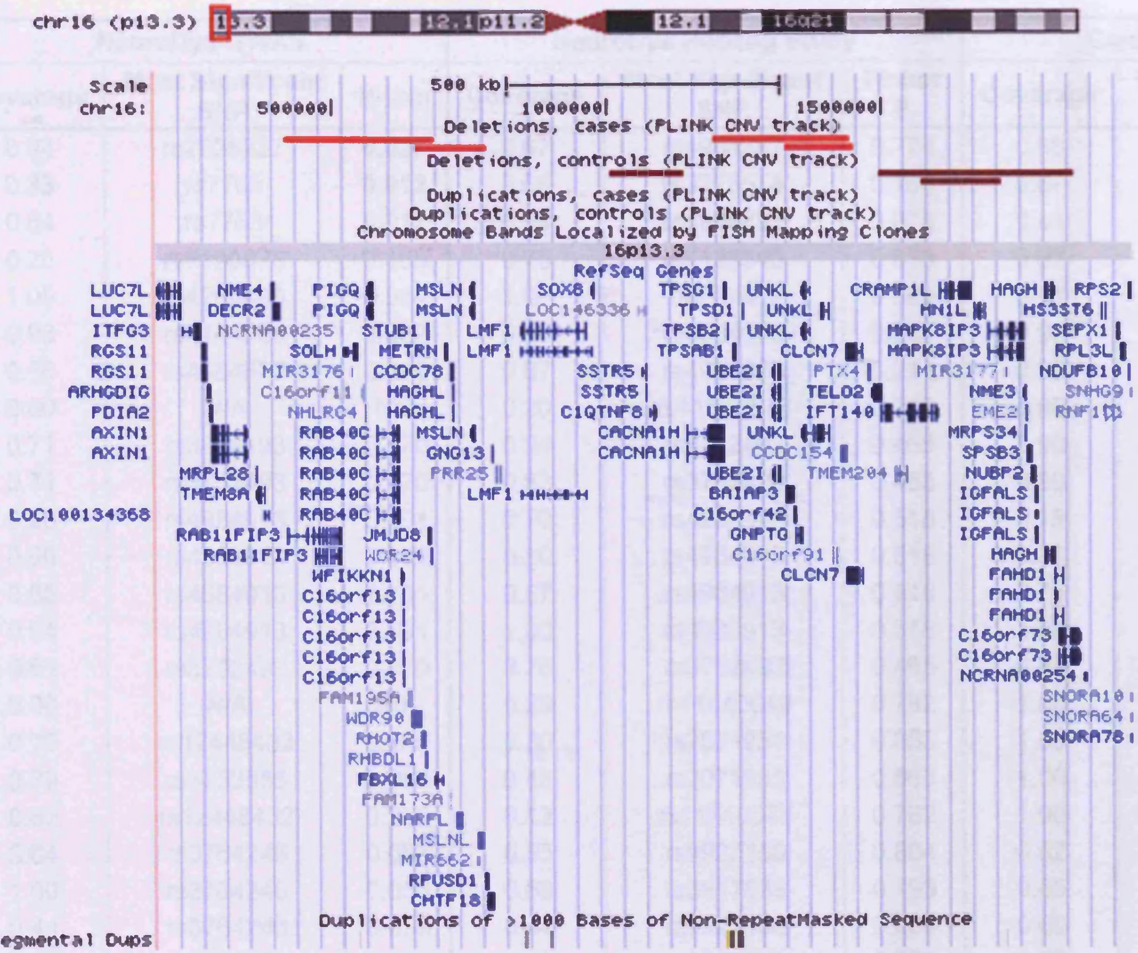


Figure 7.10: UCSC track showing the deletions identified within the two significant loci on 16p13.3 with the deletions in the cases depicted as the lighter red bars and those in the controls as the dark red bars. RefSeq genes within this region are also shown.

		NeuroDys GWAS			NeuroDys Pooling Study			Cardiff Pooling Study		
Locus	Gene	Coverage	Most Significant SNP	P-min	Coverage	Most Significant SNP	Fisher P	Coverage	Most Significant SNP	P-comb
16p13.3a	*RAB11FIP3*	0.94	rs2038227	**0.029**	0.67	rs4984887	0.778	0.88	rs3760048	**0.008**
	C16orf11	0.33	rs7763	**0.012**	0.66	rs3213574	0.168	0.67	rs3213574	0.080
	SOLH	0.54	rs7763	**0.012**	0.54	rs4984887	0.078	0.41	rs9934705	**0.030**
	NHLRC4	0.25	rs7190878	0.169	0.75	rs7189540	0.833	0.67	rs3213574	0.080
	PIGQ	1.00	rs4984890	0.314	1.00	rs710925	0.361	0.91	rs7190878	0.700
	RAB40C	0.96	rs4144003	0.322	0.96	rs4984678	0.211	0.96	rs1045277	0.820
	WFIKKN1	0.86	rs4984677	0.396	0.87	rs4984678	0.211	0.83	rs2269556	0.510
	FAM195A	0.00	N/A	N/A	0.20	rs11642546	0.712	0.80	rs2269561	0.710
	WDR90	0.71	rs3752493	0.170	0.94	rs3752493	0.465	0.90	rs12930932	0.830
	RHOT2	0.71	rs3752493	0.170	0.83	rs3752493	0.465	0.89	rs3752496	0.810
	STUB1	0.90	rs4984913	0.101	0.70	rs4984913	0.518	0.78	rs1139897	0.900
	WDR24	0.90	rs4984913	0.101	0.80	rs4984913	0.518	0.80	rs3830141	0.750
	JMJD8	0.85	rs4984913	0.101	0.57	rs4984913	0.518	0.75	rs1128550	0.910
	FBXL16	0.64	rs4984913	0.101	1.00	rs4984913	0.518	1.00	rs11640115	0.750
	RHBDL1	0.61	rs3752493	0.170	0.76	rs3752493	0.465	0.84	rs1045763	0.348
	METRN	0.00	N/A	N/A	0.29	rs11540048	0.782	1.00	rs12599342	0.940
	HAGHL	0.70	rs12448432	0.250	0.30	rs2071951	0.663	1.00	rs4589552	0.090
	NARFL	0.79	rs3752556	0.061	0.43	rs2071951	0.663	1.00	rs2071950	0.080
	FAM173A	0.62	rs12448432	0.250	0.12	rs11540048	0.782	1.00	rs3809663	0.360
	MSLN	0.64	rs3764246	0.053	0.55	rs9927150	0.804	0.82	rs13336445	0.600
	MSLNL	1.00	rs3764246	0.053	0.68	rs3817833	0.195	0.65	rs3764247	0.760
	MIR662	0.44	rs3764246	0.053	0.44	rs9927150	0.804	0.60	rs2235505	0.340
	RPUSD1	0.56	rs1052629	0.607	0.59	rs1052629	0.263	0.82	rs3765334	0.380
	CHTF18	0.53	rs1052629	0.607	0.55	rs1052629	0.263	0.58	rs3765334	0.380

Table 7.8: This table shows the genes in the first 16p13.3 locus with their coverage (if any) in the NeuroDys GWAS and pooling study and in the Cardiff Pooling study, along with the most significant SNP within the gene (+/- 10kb) in each study and the P-value for this SNP. N.B. Coverage may differ between the pooling studies due to different QC filters. P-values <0.05 are in bold, as are functionally interesting genes.

Table 7.9 shows the genes that are overlapped by CNVs in the second 16p13.3 CNV locus. SNPs near the gene calcium channel, voltage-dependent, T type, alpha 1H subunit (*CACNA1H*) (rs11865234, P = 0.007) and tryptase delta 1 (*TPSD1*) (rs3765436, P = 0.039) showed significant association in the initial Neurodys GWAS. *CACNA1H* is only partially overlapped by the case CNVs in this region, as shown in Figure 7.10 and rs11865234 lies 6.5kb upstream of this gene so is actually overlapped by a deletion identified in a control individual rather than in a DD case. The SNP rs3765436 lies 2.8kb downstream of *TPSD1* and is overlapped by a CNV identified in a DD case individual. Three other SNPs within this region showed significant association in the NeuroDys pooling study. The most significant of these was rs1132356 (P = 0.005) which is an exonic SNP within BAI1-associated protein 3 (*BAIAP3*), and also lies 4.7 kb downstream of chromosome 16 open reading frame 42 (*C16orf42*) and 7.3 kb upstream of N-acetylglucosamine-1-phosphate transferase, gamma subunit (*GNPTG*). The next most significant SNP within this region in the NeuroDys pooling study was rs2369696 (P = 0.008) which is an intronic SNP within the gene unkempt homolog (Drosophila)-like (*UNKL*) and the third was rs3751894 (P = 0.022) which is an intronic SNP within the gene coiled-coil domain containing 154 (*CCDC154*), and also lies 7.5 kb upstream of chromosome 16 open reading frame 91 (*C16orf91*) and 8 kb downstream of chloride channel 7 (*CLCN7*). All three of these SNPs are overlapped by 2 deletions in the cases and none of the SNPs within this significant CNV region are in high LD with each other. Two genes within this region could be of functional interest. *CACNA1H* encodes a T-type member of the alpha-1 subunit family which is a protein in the voltage-dependent calcium channel complex. These T-type channels may be involved in the modulation of firing patterns of neurons which is important for information processing as well as in cell growth processes (Perez-Reyes 2006) and studies have suggested that certain mutations in the *CACNA1H* gene may lead to childhood absence epilepsy (Tan et al. 2006). *BAIAP3* encodes a brain-specific angiogenesis inhibitor which is a member of the secretin receptor family and is predominantly expressed in the brain (Shiratsuchi et al. 1998). The expression profile of the protein encoded by this gene and its similarity to other proteins suggest that it may be involved in synaptic functions (Shiratsuchi et al. 1998).

		NeuroDys GWAS			NeuroDys Pooling Study			Cardiff Pooling Study		
Locus	**Gene**	**Coverage**	**Most Significant SNP**	**P-min**	**Coverage**	**Most Significant SNP**	**Fisher P**	**Coverage**	**Most Significant SNP**	**P-comb**
16p13.3b	***CACNA1H***	0.50	rs11865234	**0.007**	0.53	rs4984639	0.279	0.72	rs909910	0.470
	TPSG1	0.43	rs1054645	0.079	0.89	rs4984639	0.279	0.89	rs3897150	0.800
	TPSB2	0.42	rs1054645	0.079	0.50	rs9937881	0.130	0.80	rs11866966	0.150
	TPSAB1	0.00	N/A	N/A	0.50	rs9937881	0.130	1.00	rs35262813	0.900
	TPSD1	0.57	rs3765436	**0.039**	0.79	rs9937881	0.130	0.79	rs9937881	0.080
	UBE2I	0.75	rs4984803	0.221	0.81	rs7187167	0.112	0.89	rs7187167	0.940
	BAIAP3	0.35	rs742460	0.190	0.61	rs1132356	**0.005**	0.77	rs8063	0.930
	C16orf42	0.22	rs742460	0.190	0.67	rs1132356	**0.005**	0.89	rs2235632	0.400
	GNPTG	0.25	rs1061497	0.089	0.63	rs1132356	**0.005**	0.71	rs2235632	0.400
	UNKL	0.14	rs1061497	0.089	0.68	rs2369696	**0.008**	0.72	rs8058617	0.920
	CCDC154	0.43	rs7194275	0.114	0.17	rs3751894	**0.022**	0.68	rs4984834	0.200
	C16orf91	0.25	rs7194275	0.114	0.12	rs3751894	**0.022**	0.83	rs4984834	0.200
	CLCN7	0.50	rs1040497	0.088	0.43	rs3751894	**0.022**	0.71	rs3751894	0.630

Table 7.9: This table shows the genes in the second 16p13.3 locus with their coverage (if any) in the NeuroDys GWAS and pooling studies and in the Cardiff pooling study, along with the most significant SNP within the gene (+/- 10kb) in each study and the P-value for this SNP. N.B. Coverage may differ between the pooling studies due to different QC filters. P-values <0.05 are in bold, as are functionally interesting genes.

Figure 7.11 shows the genes that are overlapped by CNVs in the significant 19p13.3 CNV locus and Table 7.10 shows the most significant SNPs in or near each of these genes in the previous studies. In this region, the only gene that harbours a significant SNP in any of the previous studies is polo-like kinase 5, pseudogene (*PLK5P*). The SNP rs1040499 lies in the 3' UTR of this gene and showed marginal significant association in the initial NeuroDys GWAS (P = 0.048). However, this particular gene is only partially overlapped by the CNVs in this region, and this SNP is not in the same region of the gene that is overlapped by the deletions identified in the cases.

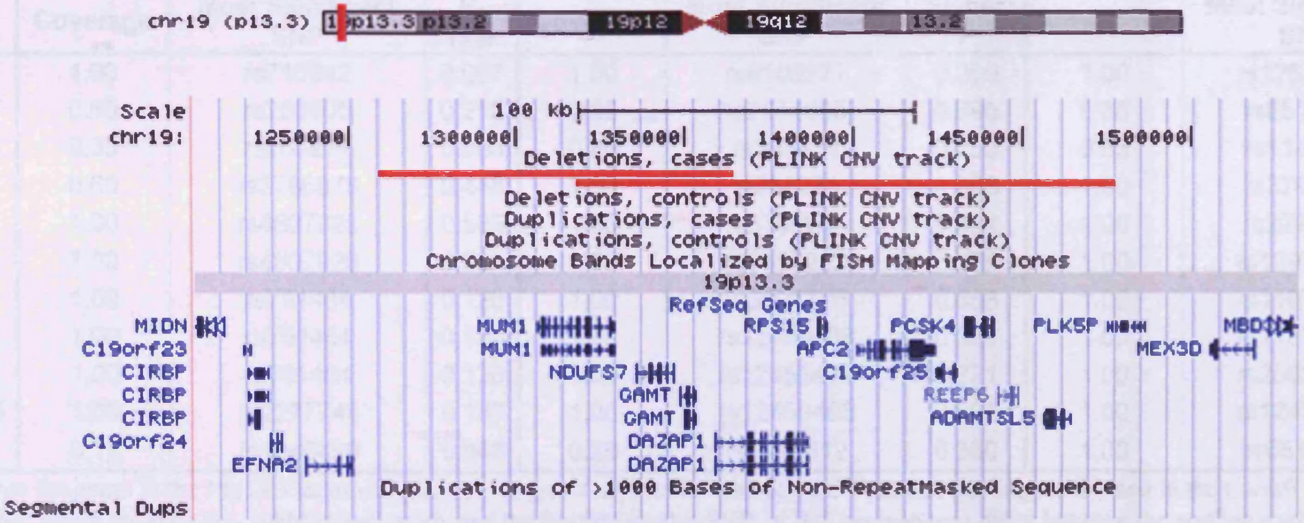


Figure 7.11: UCSC track showing the deletions identified within the 19p13.3 locus in two cases. RefSeq genes within this region are also shown.

		NeuroDys GWAS			**NeuroDys Pooling Study**			**Cardiff Pooling Study**		
Locus	**Gene**	**Coverage**	**Most Significant SNP**	**P-min**	**Coverage**	**Most Significant SNP**	**Fisher P**	**Coverage**	**Most Significant SNP**	**P-comb**
19p13.3	*MUM1*	1.00	rs713042	0.097	1.00	rs8109377	0.269	1.00	rs12609120	0.460
	NDUFS7	0.60	rs266805	0.210	0.86	rs2074895	0.695	0.88	rs6510605	0.200
	GAMT	0.33	rs265271	0.953	0.80	rs265271	0.283	0.83	rs1142530	0.180
	DAZAP1	0.60	rs3786978	0.448	1.00	rs265271	0.283	1.00	rs3786974	0.760
	RPS15	1.00	rs4807928	0.585	1.00	rs3760994	0.864	1.00	rs2292457	0.930
	APC2	1.00	rs4807928	0.585	1.00	rs8100242	0.629	1.00	rs12977033	0.640
	C19orf25	1.00	rs791464	0.120	1.00	rs3894776	0.653	1.00	rs11878689	0.410
	PCSK4	1.00	rs791464	0.120	1.00	rs12459408	0.321	1.00	rs11878689	0.410
	REEP6	1.00	rs791464	0.120	1.00	rs12459408	0.321	1.00	rs28658577	0.720
	ADAMTSL5	1.00	rs2277748	0.133	1.00	rs12459408	0.321	1.00	rs12459408	0.470
	PLK5P	0.12	rs1040499	**0.048**	0.20	rs6510612	0.360	1.00	rs6510612	0.980

Table 7.10: This table shows the genes in the 19p13.3 locus with their coverage (if any) in the NeuroDys GWAS and pooling studies and in the Cardiff pooling study, along with the most significant SNP within the gene (+/- 10kb) in each study and the P-value for this SNP. N.B. Coverage may differ between the pooling studies due to different QC filters. P-values <0.05 are in bold, as are functionally interesting genes.

Figure 7.12 shows the genes that are overlapped by CNVs in the significant 17q25.1 CNV locus and Table 7.11 shows the most significant SNPs in or near each of these genes in the previous studies. Two SNPs within these genes showed significant association in the initial NeuroDys GWAS. These are rs1171196 which is 4.4kb downstream of CD300 molecule-like family member b (*CD300LB*) (P = 0.011) and rs783239 which is an intronic SNP within the chromosome 17 open reading frame 77 (*C17orf77*) and an exonic SNP within CD300 molecule-like family member d (*CD300LD*) (P = 0.009). Another SNP in this region (rs2706506) showed significant association in the NeuroDys pooling study (P = 0.011). This SNP lies 5.6 kb upstream of CD300a molecule (*CD300A*). These three SNPs are all overlapped by the two duplications that were identified within this region and they are not in high LD with each other ($r^2 < 0.14$, $D' < 0.68$). As yet, none of the genes in this region are known to be of functional interest within DD.

The significantly associated SNPs in these loci would not have survived correction for multiple testing in these previous studies and as such they were far down the list of significant hits and were not followed up in any of the previous studies. However, the finding of a significant excess of CNVs in DD cases in these regions could warrant further investigation of the affected genes.

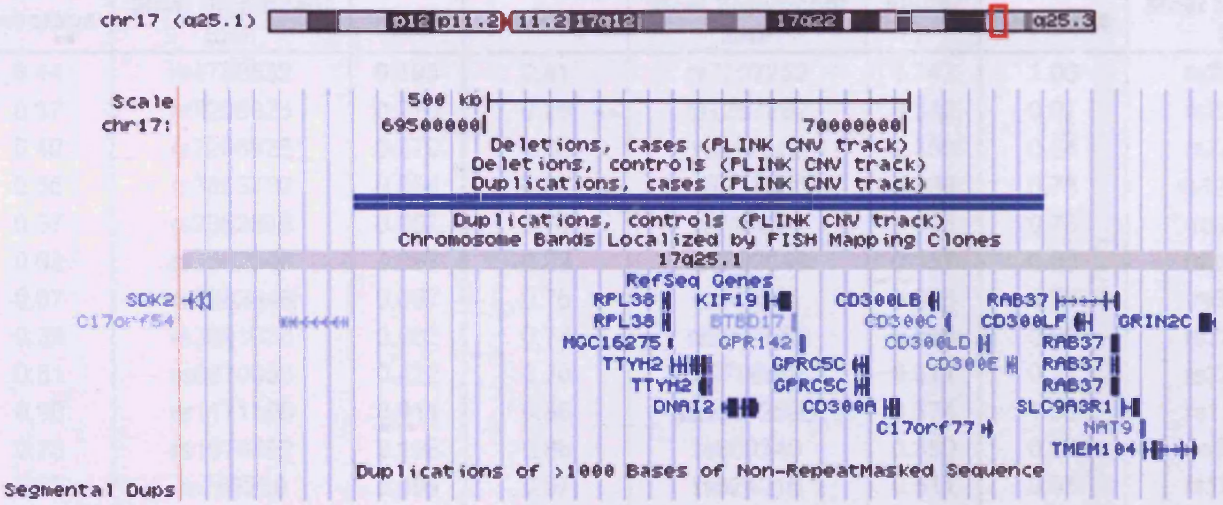


Figure 7.12: UCSC track showing the duplications identified within the 17q25.1 locus in two cases. RefSeq genes within this region are also shown.

		NeuroDys GWAS			NeuroDys Pooling Study			Cardiff Pooling Study		
Locus	***Gene***	**Coverage**	**Most Significant SNP**	**P-min**	**Coverage**	**Most Significant SNP**	**Fisher P**	**Coverage**	**Most Significant SNP**	**P--comb**
17q25.1	*RPL38*	0.44	rs4789632	0.193	0.41	rs7207252	0.342	1.00	rs2054834	0.880
	MGC16275	0.37	rs7206926	0.270	0.25	rs7207252	0.342	0.87	rs7207252	0.160
	TTYH2	0.49	rs7206926	0.270	0.82	rs747742	0.353	0.84	rs2382838	0.860
	DNAI2	0.55	rs3803792	0.064	0.66	rs7219585	0.266	0.73	rs11650353	0.580
	KIF19	0.67	rs2382646	0.087	0.76	rs747321	0.075	0.78	rs938287	0.640
	BTBD17	0.92	rs2382646	0.087	0.92	rs2382646	0.157	0.80	rs3869467	0.650
	GPR142	0.67	rs2382646	0.087	0.76	rs747321	0.075	0.78	rs938287	0.640
	GPRC5C	0.39	rs2891033	0.082	0.77	rs2251065	0.065	0.69	rs7215253	0.670
	CD300A	0.51	rs8070953	0.422	0.70	rs2706506	**0.011**	0.71	rs2706506	0.340
	CD300LB	0.90	rs1171196	**0.011**	0.86	rs10512596	0.674	1.00	rs1107704	0.350
	CD300C	0.75	rs1976492	0.198	0.88	rs809740	0.350	0.87	rs965229	0.850
	C17orf77	0.83	rs783239	**0.009**	0.97	rs524216	0.511	0.95	rs1522875	0.210
	CD300LD	0.65	rs783239	**0.009**	0.86	rs11650378	0.293	0.89	rs1699597	0.440
	CD300E	0.74	rs581157	0.095	0.35	rs16978145	0.407	0.86	rs6501728	0.480

Table 7.11: This table shows the genes in the 17q25.1 locus with their coverage (if any) in the NeuroDys GWAS and pooling studies and in the Cardiff pooling study, along with the most significant SNP within the gene (+/- 10kb) in each study and the P-value for this SNP. N.B. Coverage may differ between the pooling studies due to different QC filters. P-values <0.05 are in bold, as are functionally interesting genes.

7.3.5 Comparison with Other CNV Studies

Whilst the burden analysis appeared to suggest that there may be some inter-centre experimental variation that resulted in a significant excess of deletions being identified in those cases that had been genotyped in Oxford in comparison with those genotyped in Bonn, this needs to be confirmed with further validation. Therefore, in order to retain as much power as possible and to reduce the possibility of false-negative results, all cases were used in the comparisons with other CNV studies. It is important to note however that due to the nature of the data, interesting results will need to be interpreted with caution until they can be reliably validated.

7.3.5.1 ADHD CNV studies

The CNVs found in this study were compared with two CNV studies of ADHD. These studies identified a number of regions in which CNVs were found in ADHD cases but not in healthy controls (Elia et al. 2010; Lesch et al. 2010). The results of any CNVs identified in DD cases that overlapped these ADHD CNV regions are shown in Table 7.12. This table also shows if any CNVs in these regions were identified within the 1958 Birth Cohort and gives the P-value from the association test between the rates of CNVs in DD cases with those in controls for each of these regions.

CNVs were identified in DD cases for 12 of the ADHD CNV regions. Interestingly, CNVs were also found in the 1958 Birth Cohort controls in six of these regions. Figure 7.13 shows the ADHD CNV region on chromosome 17 identified by Lesch et al. (2010) which had significantly more duplications in the DD cases than were identified in the 1958 Birth Cohort controls (P = 0.012), as was found during the regional analysis in section 7.3.3 of this chapter. This was one of the 5 regions that remained significant when excluding the cases that had been genotyped in Oxford.

Elia et al (2010) performed Gene Ontology (GO) analysis on the genes that were overlapped by CNVs, and interestingly, 'learning' was among the six most highly enriched GO Biological Process categories. Two of the ADHD CNV genes that were associated with learning were protein tyrosine phosphatase, receptor type, D (*PTPRD*) and Parkinson disease (autosomal recessive, juvenile) 2, parkin (*PARK2*), both of which were also overlapped by CNVs in DD. Figure 7.14 depicts those CNVs that overlap *PTPRD* with one deletion identified in an individual with DD (rate = 0.008), and 3 deletions identified in the controls (rate = 0.0029). This gives a CNV case/control ratio of 2.76 and a P-value of 0.36, suggesting that whilst there was an excess of CNVs in DD cases within this gene, this excess was not significant. The deletion identified in the DD case overlaps one of the ADHD deletions entirely, disrupting the first two introns and the second exon of this gene. The three other ADHD deletions are overlapped by deletions that were identified in the controls.

As shown in Figure 7.15, one deletion and 3 duplications identified in DD cases (rate of deletions = 0.008, rate of duplications = 0.025, total rate = 0.033) and 7 deletions and 3 duplications identified in the controls (rate of deletions = 0.0068, rate of duplications = 0.0029, total rate = 0.0097) overlap *PARK 2*. This gives a CNV case/control ratio of 3.40 and a P-value of 0.048 for this gene, suggesting that there is a significantly higher rate of CNVs in the DD cases than in the controls within this gene. This region was highlighted as one of the significant regions in section 7.3.3 of this chapter, giving a P-value of 0.019 when looking at duplications alone. When looking at deletions alone the P-value is 1, suggesting that the excess of CNVs in this gene is due to the presence of the duplications. The duplications identified in the DD cases all overlap the ADHD duplication region defined by Elia et al (2010) entirely, whereas the deletion identified in one DD case is downstream from the ADHD deletion region. Two of the DD duplications appear to disrupt the first two introns and second exon of *PARK2* while the third is larger and also disrupts the third intron and the third exon. Interestingly, CNVs within *PARK2* have also been reported in schizophrenia by Xu et al (2008).

Another ADHD CNV region of potential interest is 3p26.3, shown in Figure 7.16. Although there is not a significant excess of CNVs in DD cases in this region, the genes close homolog of L1 (*CHL1*) and contactin 6 (*CNTN6*) are of functional interest. *CHL1* encodes for an extracellular matrix and cell adhesion protein that is thought to play a role in nervous system development and in synaptic plasticity by regulating cell migration in nerve regeneration and cortical development and is expressed at high levels in the adult and fetal brain (Holm et al. 1996; Hillenbrand et al. 1999). *CNTN6* (previously known as *NB*-3) encodes a contactin which mediate cell surface interactions during development and is involved in oligodendrocyte generation through the activation of *NOTCH1* (Takeda et al. 2003). It shows high expression in the fetal cerebellum and this expression increases until adulthood (Lee et al. 2000).

The region 9q24.3 is also of potential interest as this is considered to be a 'hotspot' region for CNVs. This region is discussed in the next section.

Study	Region			Type in Study	DD Cases			1958 Birth Cohort controls			Case /Control	P	Genes Overlapped by CNVs
	Chr	Start bp	End bp		Dels	Dups	Rate	Dels	Dups	Rate			
Elia *et al.* 2010 (n = 335 trios)	3	38411	1118424	dup	1	0	0.008	0	1	0.001	8.58	0.202	***CHL1, CNTN6***
	3	176011208	176181941	del	0	1	0.008	0	0	0		0.106	*NAALADL2*
	3	176307146	176434458	del	0	1	0.008	0	0	0		0.106	*NAALADL2*
	4	190175635	190482064	dup	0	1	0.008	1	1	0.002	4.29	0.288	-
	6	162672945	162801747	del /dup	0	3	0.025	4	3	0.007	3.68	0.077	***PARK2***
	9	36587	415228	del	0	2	0.017	0	5	0.005	3.43	0.159	*FOXD4, CBWD1, LOC642313, C9orf66,* ***DOCK8***
	9	9084805	10423023	del	1	0	0.008	3	0	0.003		0.360	***PTPRD***
	9	11685785	11847464	del	2	0	0.017	3	0	0.003	5.72	0.085	-
	9	12032535	12665264	del	1	0	0.008	7	0	0.007	1.23	0.646	-
	13	62104068	62161921	del	1	0	0.008	0	0	0		0.103	-
	17	31889664	33323543	del	0	1	0.008	0	0	0		0.101	*PIGW, ZNHIT3, MYO19, GGNBP2, MRM1, DHRS11, LHX1, AATF, MIR2909, ACACA, C17orf78, TADA2A, DUSP14, SYNRG, DDX52, HNF1B, LOC284100*
Lesch *et al.*2010 (n = 99)	17	69250000	70180000	dup	0	2	0.017	0	0	0		**0.012**	*RPL38, MGC16275, TTYH2, DNAI2, KIF19, BTBD17, GPR142, GPRC5C, CD300A, CD300LB, CD300C, C17orf77, CD300LD, CD300E, RAB37*

Table 7.12: Presence of CNVs in DD cases and 1958BC controls within the CNVs regions identified in ADHD cases in studies by Elia *et al.* (2010) and Lesch *et al.* (2010). Only those regions in which CNVs were found within DD cases are shown. P-values < 0.05 are in bold. Chr – chromosome; del – deletion; dup – duplication. Genes of functional interest are highlighted in bold.

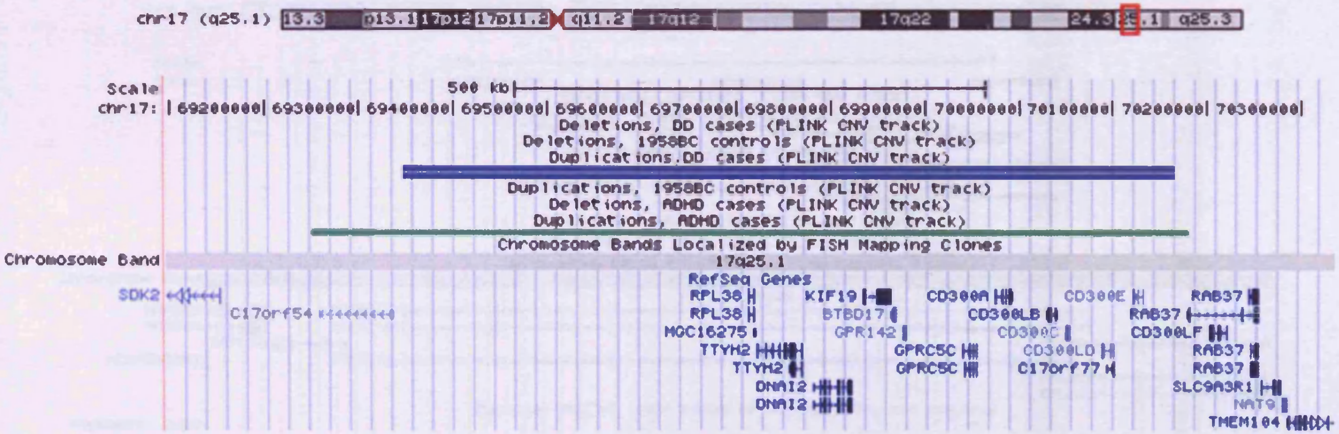

chr17 (q25.1)
13.3
p13.1
17p12
17p11.2
q11.2
17q12
17q22
24.3
25.1
q25.3
Scale
500 kb
chr17: 69200000 69300000 69400000 69500000 69600000 69700000 69800000 69900000 70000000 70100000 70200000 70300000
Deletions, DD cases (PLINK CNV track)
Deletions, 1958BC controls (PLINK CNV track)
Duplications,DD cases (PLINK CNV track)
Duplications, 1958BC controls (PLINK CNV track)
Deletions, ADHD cases (PLINK CNV track)
Duplications, ADHD cases (PLINK CNV track)
Chromosome Bands Localized by FISH Mapping Clones
Chromosome Band
17q25.1
RefSeq Genes
SDK2
C17orf54
RPL38
RPL38
MGC16275
TTYH2
TTYH2
DNAI2
DNAI2
KIF19
BTBD17
GPR142
GPRC5C
GPRC5C
CD300A
CD300LB
CD300C
CD300LD
C17orf77
CD300E
RAB37
RAB37
CD300LF
RAB37
RAB37
SLC9A3R1
NAT9
TMEM104

Figure 7.13: UCSC track showing the duplications identified within the 17q25.1 locus in two DD cases (shown in blue) which overlap a duplication region identified in ADHD cases by Lesch et al (2010) (shown in green). RefSeq genes within this region are also shown.

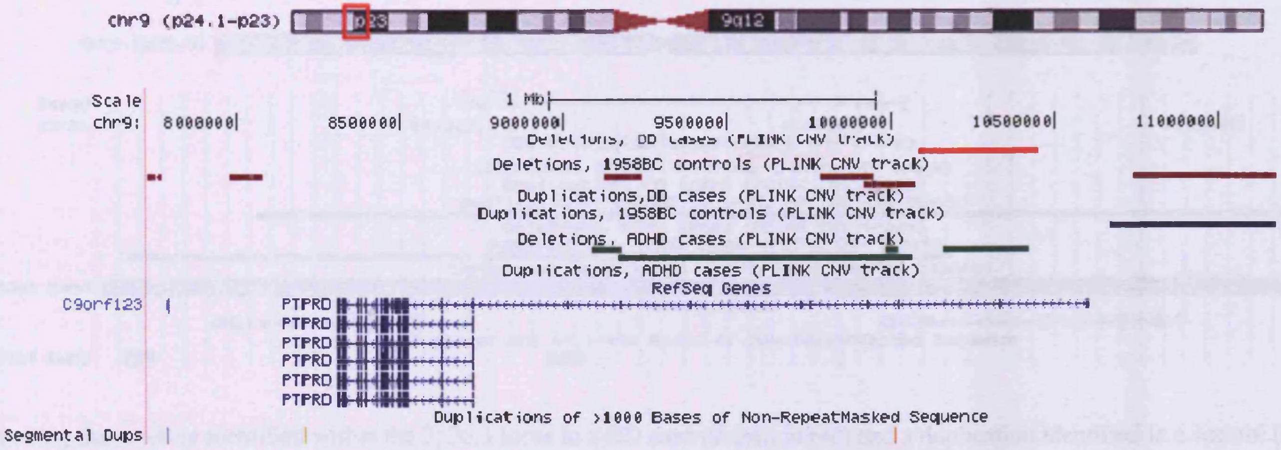

chr9 (p24.1-p23)
p23
9q12
Scale
1 Mb
chr9: 8000000 8500000 9000000 9500000 10000000 10500000 11000000
Deletions, DD cases (PLINK CNV track)
Deletions, 1958BC controls (PLINK CNV track)
Duplications,DD cases (PLINK CNV track)
Duplications, 1958BC controls (PLINK CNV track)
Deletions, ADHD cases (PLINK CNV track)
Duplications, ADHD cases (PLINK CNV track)
RefSeq Genes
C9orf123
PTPRD
PTPRD
PTPRD
PTPRD
PTPRD
PTPRD
Duplications of >1000 Bases of Non-RepeatMasked Sequence
Segmental Dups

Figure 7.14 UCSC track showing the CNVs overlapping *PTPRD* in DD cases (deletions shown in red) and the 1958BC controls (deletions shown in dark red, duplications shown in dark blue). The deletions identified in ADHD cases by Elia et al (2010) are also shown (in dark green).

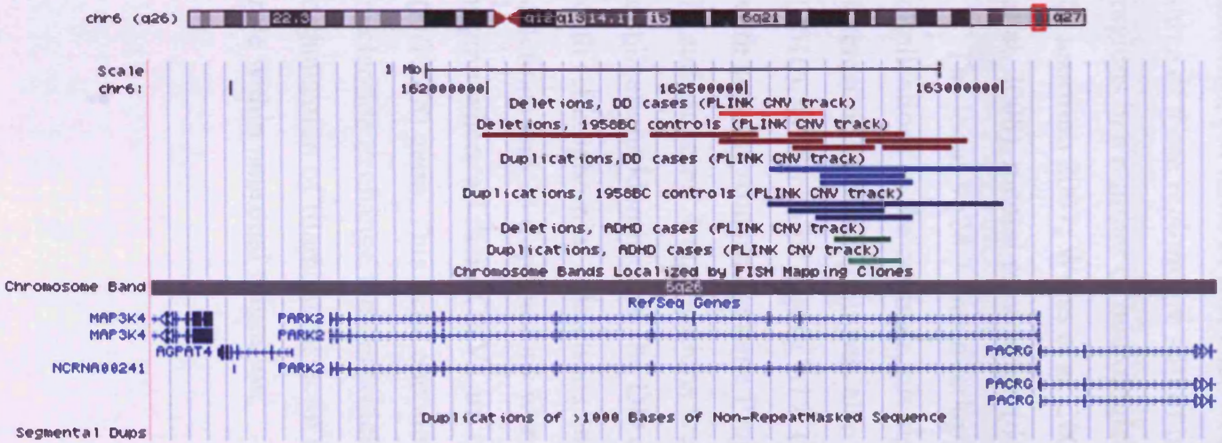

chr6 (q26)
Scale
chr6:
1 Mb
162000000
162500000
163000000
Deletions, DD cases (PLINK CNV track)
Deletions, 1958BC controls (PLINK CNV track)
Duplications,DD cases (PLINK CNV track)
Duplications, 1958BC controls (PLINK CNV track)
Deletions, ADHD cases (PLINK CNV track)
Duplications, ADHD cases (PLINK CNV track)
Chromosome Bands Localized by FISH Mapping Clones
Chromosome Band
6q26
RefSeq Genes
MAP3K4
MAP3K4
AGPAT4
NCRNA00241
PARK2
PARK2
PARK2
PACRG
PACRG
PACRG
Duplications of >1000 Bases of Non-RepeatMasked Sequence
Segmental Dups

Figure 7.15: UCSC track showing the CNVs overlapping *PARK2* in DD cases (deletions shown in red and duplications shown in blue) and the 1958BC controls (deletions shown in dark red, duplications shown in dark blue). The CNV regions identified in ADHD cases by Elia et al (2010) are also shown (deleted region shown in dark green, duplicated region shown in light green).

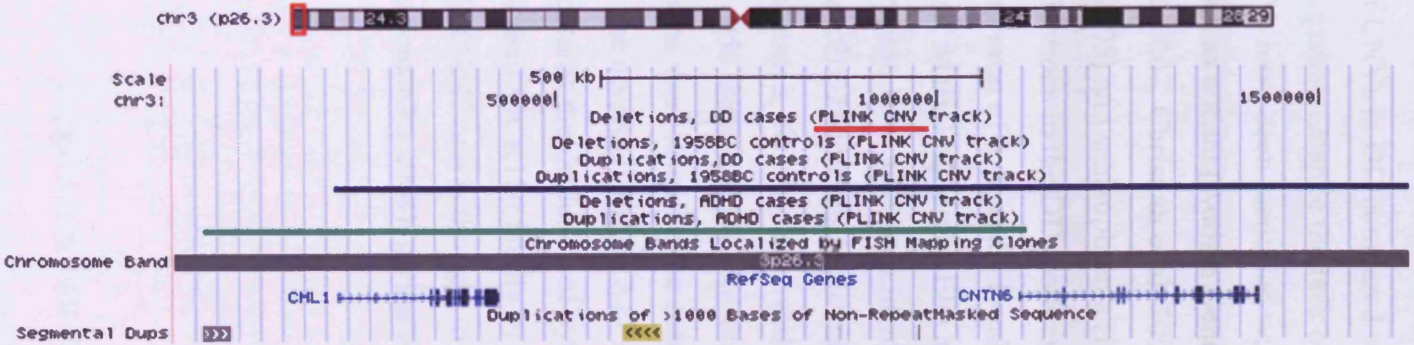

chr3 (p26.3)
Scale
chr3:
500 kb
500000
1000000
1500000
Deletions, DD cases (PLINK CNV track)
Deletions, 1958BC controls (PLINK CNV track)
Duplications,DD cases (PLINK CNV track)
Duplications, 1958BC controls (PLINK CNV track)
Deletions, ADHD cases (PLINK CNV track)
Duplications, ADHD cases (PLINK CNV track)
Chromosome Bands Localized by FISH Mapping Clones
Chromosome Band
3p26.3
RefSeq Genes
CHL1
CNTN6
Duplications of >1000 Bases of Non-RepeatMasked Sequence
Segmental Dups

Figure 7.16 UCSC track showing the deletion identified within the 3p26.3 locus in a DD case (shown in red) and a duplication identified in a control (shown in dark blue) which overlap a duplication region identified in ADHD cases by Elia et al (2010) (shown in green). RefSeq genes within this region are also shown.

7.3.5.2 Disease Hotspots for CNVs

Of the 26 regions identified as being human genome hotspots for CNVs in the recent reviews by Itsara et al. (2009) and Mefford and Eichler (2009), 5 were overlapped by CNVs identified in this dataset, the results of which are shown in Table 7.13. Four of these regions have a higher rate of CNVs in DD compared with the controls, but none of them showed a significant excess. Despite the lack of significance, three of these regions are still of interest as CNVs have been identified within these regions in a number of neurological diseases such as autism (Autism Genome Project Consortium 2007; Weiss et al. 2008; Sebat et al. 2007; Christian et al. 2008; Marshall et al. 2008), mental retardation (de Vries et al. 2005) and schizophrenia (ISC 2008).

The locus 9q24.3 was overlapped by 2 duplications in the DD cases and 5 duplications in the controls, giving a case/control ratio of 3.43 and a P-value of 0.158. Interestingly, this region was also identified as an ADHD region by Elia and colleagues (2010), as shown in Figure 7.17. CNVs in this region have also been identified in cases with autism (Autism Genome Project Consortium 2007) and schizophrenia (ISC 2008). In autism cases, 4 duplications were found in the cases, all of which overlap the duplications identified in the DD cases and the 1958 Birth Cohort controls. In cases with schizophrenia, 3 deletions and 18 duplications were found in the cases and 2 deletions and 23 duplications were identified in the controls. Figure 7.17 shows that the region where all these CNVs are overlapping contains the dedicator of cytokinesis 8 (*DOCK8*) gene. This gene encodes a member of the DOCK180 family of guanine nucleotide exchange factors which interacts with Cdc42 and is thought to affect the organisation of filamentous actin (Ruusala & Aspenström 2004), giving it a possible role within neuronal migration.

						DD Cases				1958 BC controls					
Locus	**Start bp**	**End bp**	**Type**	**Diseases**	**Type in this study**	**All CNVs**	**Dup**	**Del**	**Rate**	**All CNVs**	**Dup**	**Del**	**Rate**	**Cases /Controls**	**P**
9p24.3	140575	1599250	del /dup	Aut[a], Schiz[b]	dup	2	2	0	0.017	5	5	0	0.005	3.43	0.158
15q11.1-q13.3	18376200	30756771	dup	Aut[a,c-f], Schiz[b], MR[g], Cons[b]	del/dup	0	0	0	0	11	10	1	0.011	0	0.399
16p12.2-16p12.1	21441805	22688093	dup	Schiz[b], Cons[b]	del/dup	1	1	0	0.008	3	2	1	0.003	2.86	0.358
17q12	31800000	33300000	del	Renal Abnormalities[h]	dup	1	1	0	0.008	0	0	0	0		0.097
22q11.21-q11.23	17200000	22000000	del /dup	Aut[a,e], MR[g], Schiz[b,i], Prematurity &/or Dev Delay[j,k], Cons[b]	del/dup	2	2	0	0.017	12	8	4	0.012	1.43	0.65

Table 7.13: Table showing the number of CNVs in DD cases and the population controls within regions that have been highlighted as human disease hotspots in reviews by Itsara *et al.* (2009) and Mefford and Eichler (2009). Only those 'hotspot' regions which were overlapped by CNVs identified in this study are shown. Aut – Autism; Schiz – Schizophrenia; MR – Mental Retardation; Cons – controls in ISC 2008 study; Dev delay – Developmental delay. References: [a] Autism Genome Project Consortium (2007) [b] International Schizophrenia Consortium (2008), [c] Weiss *et al.* (2008), [d] Sebat *et al.* (2007), [e] Christian *et al.*(2008), [f] Marshall *et al.*(2008), [g] de Vries *et al.* (2005), [h] Mefford *et al.* (2007), [i] Xu *et al.* (2008), [j] Ben-Shachar *et al.* (2008), [k] Ou *et al.* (2008).

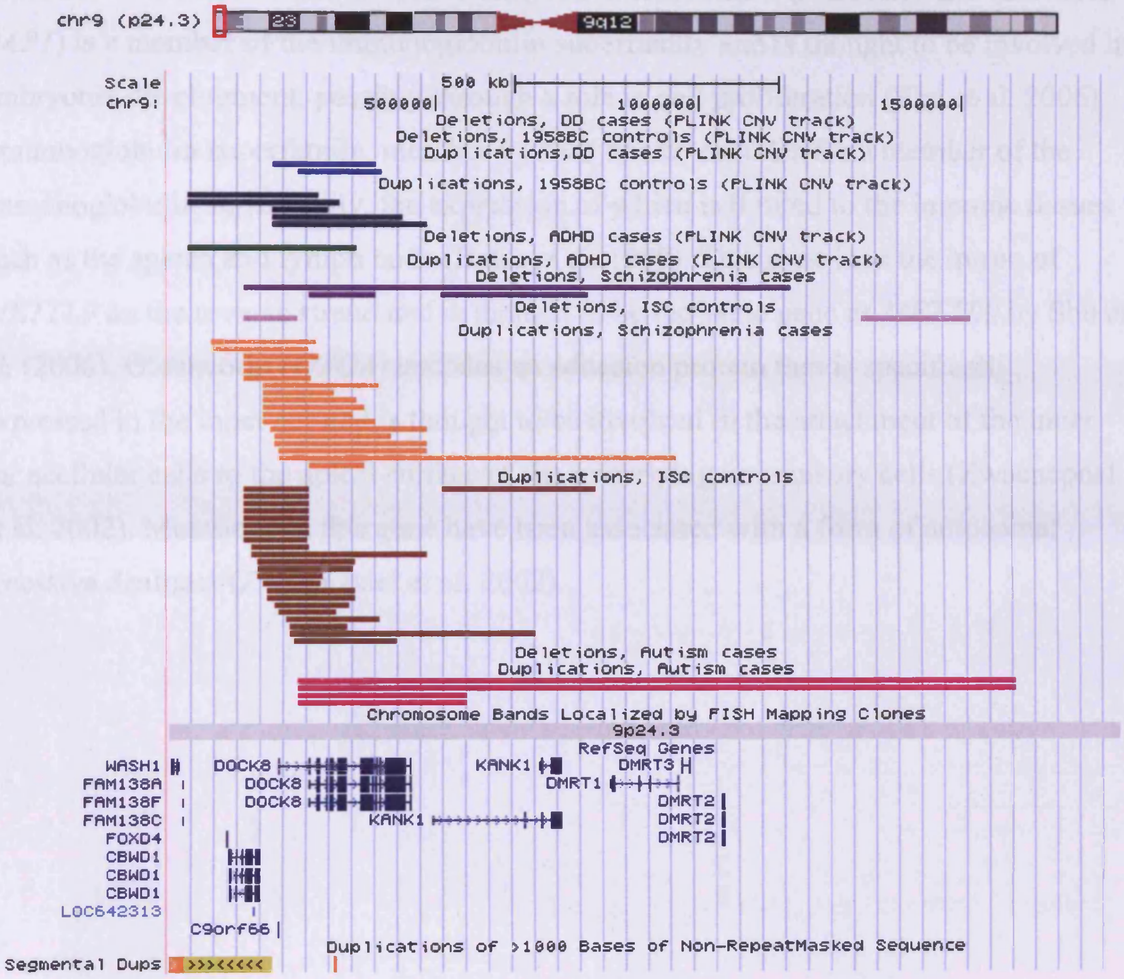


Figure 7.17: UCSC track showing the duplications in DD cases (light blue) and 1958 Birth Cohort controls (dark blue) which are within the 9p24.3 hotspot region. Also shown are the ADHD deletion region (shown in green), deletions and duplications found in schizophrenia cases (purple and orange), deletions and duplications found in the ISC controls (pale pink and brown) and the duplications found in autism cases (dark pink). RefSeq genes within this region are also shown.

Another 'hotspot' region in which a higher rate of CNVs was identified in the DD cases compared with the controls is the locus 16p12.2-16p12.1, shown in Figure 7.18. This locus was overlapped by one duplication identified in a DD case, and two duplications and one deletion identified in the 1958 BC controls, giving a case/control ratio of 2.86 and a P-value of 0.358. CNVs in this region have also been identified in cases with autism (Autism Genome Project Consortium 2007) and schizophrenia (ISC 2008) , as well as in controls in the ISC study (2008). The duplications identified in the DD case and the 1958 BC controls lie in between two regions of segmental duplications, as do 4 deletions and 8 duplications identified in schizophrenia cases, and 8 deletions and 6 duplications identified in the ISC controls. Within this region are the genes *METTL9, IGSF6* and *OTOA*. Methyltransferase-like 9 (*METTL9,* also known as *PAP1*) is a member of the immunoglobulin superfamily and is thought to be involved in embryonic development, possibly through a role in cell proliferation (Shu et al. 2006). Immunoglobulin superfamily, member 6 (*IGSF6*) also encodes for a member of the immunoglobulin superfamily, the expression of which is limited to the immune tissues such as the spleen and lymph node (Bates et al. 1998). This gene is in the intron of *METTL9* on the reverse strand and is thought to be the same gene as *METTL9* by Shu *et al.* (2006). Otoancorin (*OTOA*) encodes an adhesion protein that is specifically expressed in the inner ear and is thought to be involved in the attachment of the inner ear acellular cells to the apical surface of the underlying nonsensory cells (Zwaenepoel et al. 2002). Mutations in this gene have been associated with a form of autosomal recessive deafness (Zwaenepoel et al. 2002).

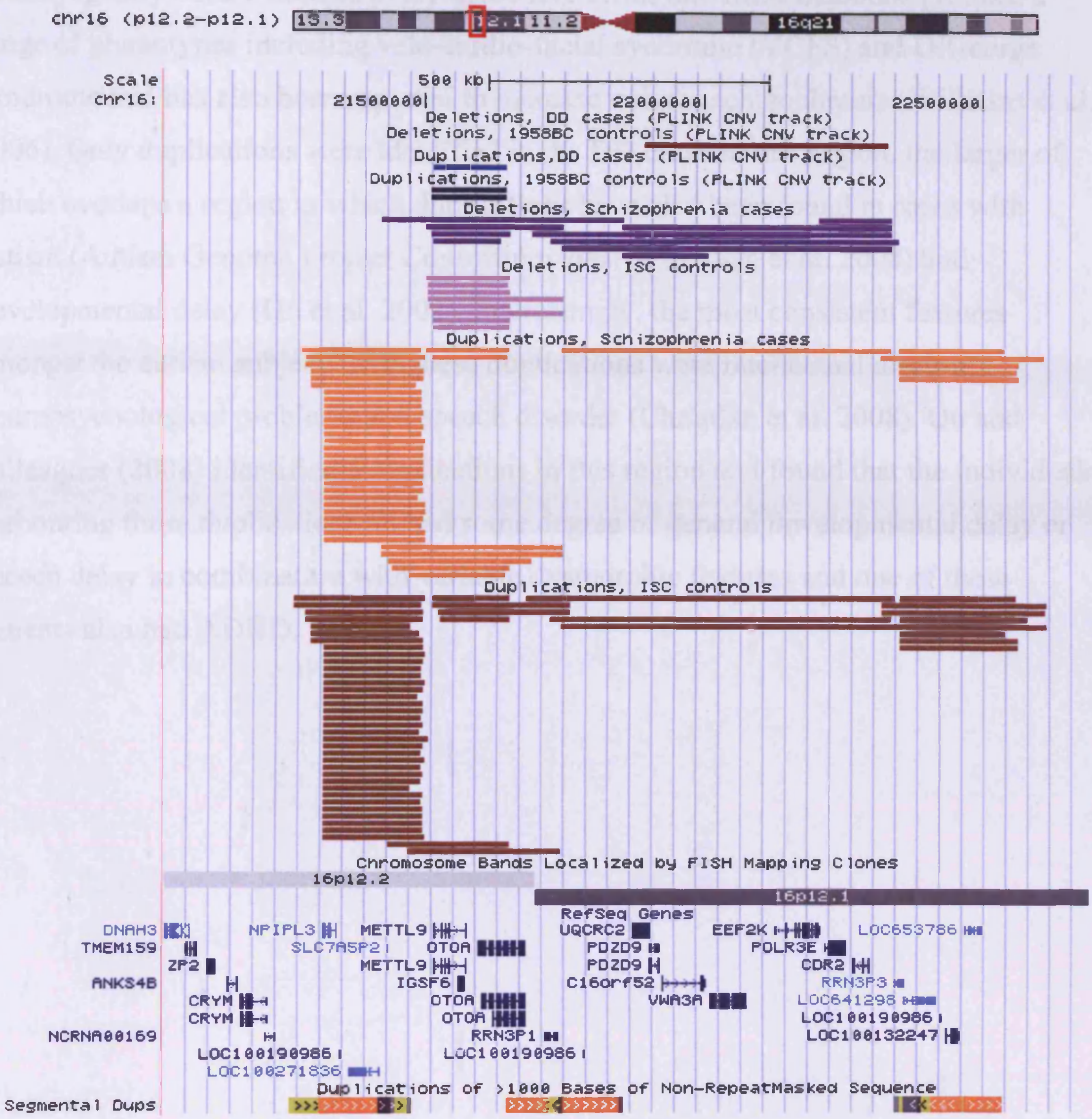


Figure 7.18: UCSC track showing the duplication in a DD case (light blue), and the deletion (dark red) and two duplications (dark blue) in 3 the 1958 BC controls which are within the 16p12.2-p12.1 hotspot region. Deletions and duplications found in schizophrenia cases (purple and orange) and deletions and duplications found in the ISC controls (pale pink and brown) are also shown as are the RefSeq genes in this region.

Figure 7.19 shows the hotspot locus 22q11.21-q11.23 which was overlapped by two duplications in the DD cases and eight duplications and four deletions in the controls, giving a case/control ratio of 1.43 and a P-value of 0.65. CNVs in this region have also been identified in a number of cases with autism (Autism Genome Project Consortium 2007; Sebat et al. 2007; Christian et al. 2008), schizophrenia (ISC 2008; Xu et al. 2008) mental retardation (de Vries et al. 2005), developmental delay (Ben-Shachar et al. 2008; Ou et al. 2008) and in the controls in the ISC study (2008). In this region, hemizyogosity occurs once in every 4,000 live births and these deletions produce a range of phenotypes including velo-cardio-facial syndrome (VCFS) and DiGeorge syndrome and has also been reported to increase risk for schizophrenia (Williams et al. 2006). Only duplications were identified in the DD cases in this region, the larger of which overlaps a region in which duplications have also been found in cases with autism (Autism Genome Project Consortium 2007; Christian et al. 2008) and developmental delay (Ou et al. 2008). Interestingly, the most consistent features amongst the autism subjects with these duplications were intellectual disability, neuropsychological problems and speech disorder (Christian et al. 2008). Ou and colleagues (2008) identified 5 duplications in this region and found that the individuals harbouring these duplications all had some degree of general developmental delay or speech delay in combination with variable dysmorphic features and one of these patients also had ADHD.

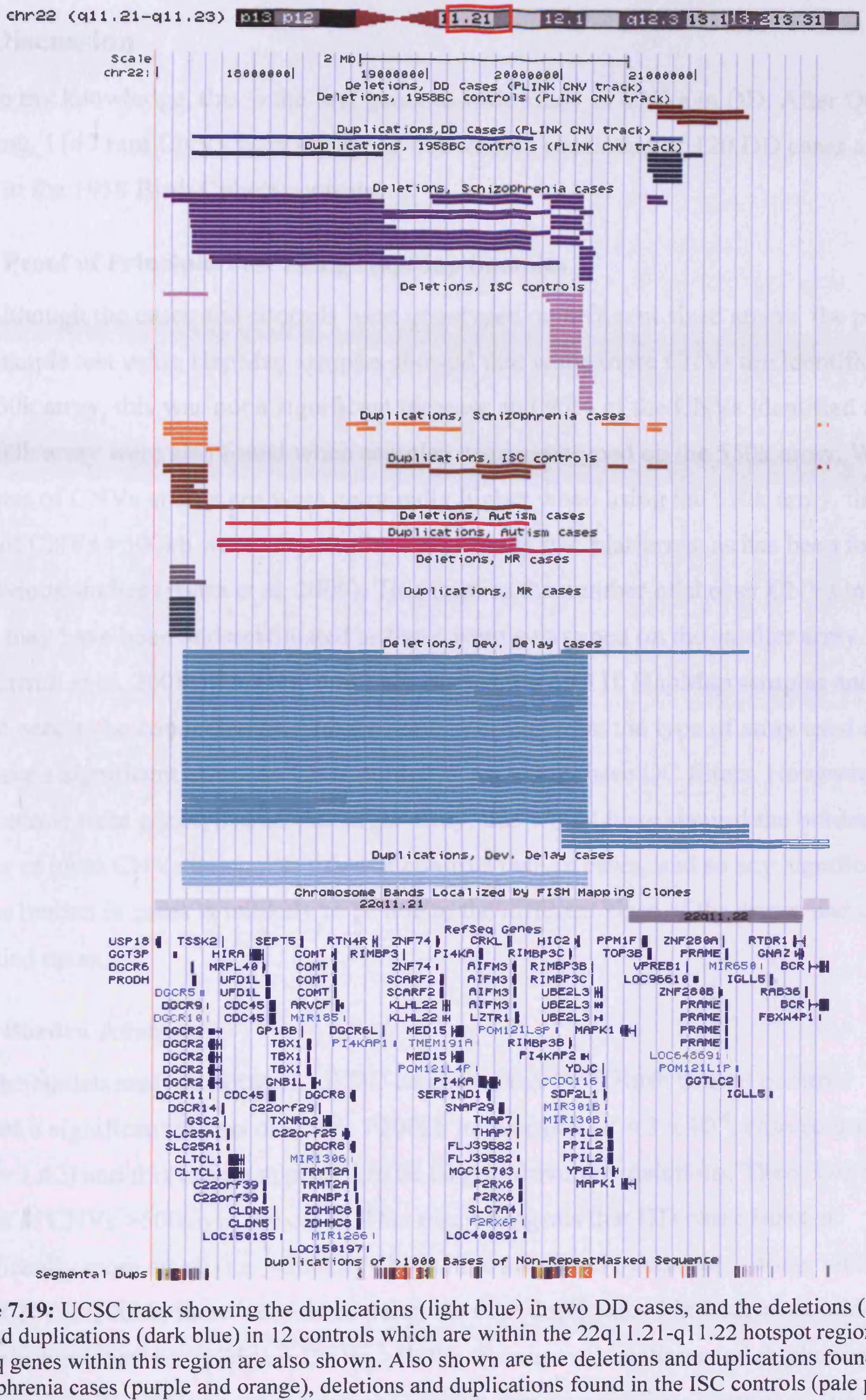


Figure 7.19: UCSC track showing the duplications (light blue) in two DD cases, and the deletions (dark red) and duplications (dark blue) in 12 controls which are within the 22q11.21-q11.22 hotspot region. RefSeq genes within this region are also shown. Also shown are the deletions and duplications found in schizophrenia cases (purple and orange), deletions and duplications found in the ISC controls (pale pink and brown), deletions and duplications found in autism cases (pink and dark pink), deletions and duplications found in cases with mental retardations (grey and dark green), and deletions and duplications found in individuals with developmental delay (green and pale blue).

7.4 Discussion

To my knowledge, this is the first genome-wide study of CNVs in DD. After QC filtering, 1147 rare CNVs were identified in a sample consisting of 120 DD cases and 1030 in the 1958 Birth Cohort controls.

7.4.1 Proof of Principle Test Using HapMap Samples

Although the cases and controls were genotyped on different sized arrays, the proof of principle test using HapMap samples showed that while more CNVs are identified on the 550k array, this was not a significant increase and 95% of the CNVs identified using the 300k array were also found when samples were genotyped on the 550k array. Whilst the rates of CNVs at all sizes were marginally higher when using the 550k array, the rates of CNVs >500kb were very similar between the two platforms, as has been found in previous studies (Itsara et al. 2009). This implies the number of shorter CNVs in the cases may have been underestimated as these were genotyped on the smaller array (McCarroll et al. 2008). This test was only carried out on 110 HapMap samples and would need to be conducted on a larger scale to be sure that the type of array used does not have a significant effect on CNV burden when using these QC filters. However, as the controls were genotyped on the larger array, this would have skewed the burden in favour of more CNVs being identified in controls than in cases, and so any significant excess burden in cases is unlikely to be due to the different sizes of the arrays and can be relied upon.

7.4.2 Burden Analyses

The burden analysis between all DD cases and the 1958 Birth Cohort controls showed a significant excess of CNVs >100kb in the cases ($P = 3 \times 10^{-4}$, case/control ratio = 1.42) and this excess appeared to be largely driven by deletions. There was no excess of CNVs >500kb in the cases. This result suggests that DD cases harbour significantly more small (i.e. <500kb) CNVs than are found in controls. When looking at all CNVs >100kb, there were more deletions than duplications in both cases and controls, but when focussing on CNVs >500kb the rates of deletions and duplications were very similar. This has also been observed in previous studies (Itsara et al. 2009) and it has been suggested that the relative enrichment of deletions at smaller sizes may reflect higher *de novo* rates of occurrence of deletions, whereas their lower rates at larger sizes is consistent with large deletions being more detrimental than duplications (Turner et al. 2008).

However, the cases had been genotyped in two different centres in Bonn (n = 74) and Oxford (n = 46) and the excess of deletions in cases seemed to be coming from those cases that had been genotyped in Oxford. Bonn cases alone did not show a significant excess of CNVs when compared with the controls, whereas the Oxford cases still showed a significant excess of deletions, even though it was the smaller case sample ($P = 1 \times 10^{-4}$, Oxford cases/controls ratio = 3.28). This suggests that there is some inter-centre variability between the cases genotyped in these two centres. This excess of deletions in cases genotyped in Oxford may well be real, but it is perhaps more likely that these CNVs represent false positives. A high rate of CNVs can indicate poor quality samples, but these samples were all extracted and prepared in the same centre (Cardiff). Therefore it is likely to be a genotyping quality issue and the higher LRR SDs (indicating a higher level of variation in the intensities between SNPs) that were observed with the cases genotyped in Oxford correlates with this. Interestingly, the rate of duplications between the two case samples was not significantly different, which suggests that genotyping quality issues can result in more deletions being identified but doesn't have such a large effect on duplications. As these CNVs in cases genotyped in Oxford passed the QC filters, they cannot be confidently excluded until validation has been attempted using another method, such as aCGH or qPCR, but they are worth bearing in mind when carrying out the regional analyses.

In addition, the case DNAs originated from either blood or saliva samples. DNA extracted from saliva samples is often at a lower concentration and of poorer quality than blood-derived DNA which may result in the identification of more CNVs. Whilst there was no significant difference in the rate of CNVs identified when DNA was extracted from either sample type, information on the origin of the samples was only available for a small subset of the sample. A larger number of samples would need to be tested to be confident that there was no significant difference when extracting DNA from saliva compared to blood samples. The 1958 Birth Cohort would have been extracted from cell lines, but DNA from cell-lines and blood-derived DNA have yielded similar results in previous studies suggesting that cell line artefacts are not a major contributor to estimates of CNV burden (Itsara et al. 2009).

It is possible that even if these cases had all originated from the same sample type and had been genotyped in the same centre, batch effects may still have existed which could affect intensity signals from sample to sample (McCarroll 2008). To address such issues, algorithms need to be designed so that sample-specific and batch-specific influences on signal intensities can be controlled for as much as possible.

7.4.3 Regional Analyses

No CNVs were identified in DD susceptibility genes. However, 18 regions across the genome were identified which harboured significantly more CNVs in the DD cases than the controls. The number of CNVs across the human genome is not yet known, and therefore a correction for multiple testing based on this number is not possible to define (Wain et al. 2009).

The most significant of the regions was 15q23 which had 5 deletions in the cases. All of these cases with CNVs in this region had been genotyped in Oxford. In order to prioritise regions of interest and to reduce the possibility that these significant regions are false positives, regional analysis was also conducted on just those cases that had been genotyped in Bonn to see if any would remain significant. Of the original 18 regions, 7 had significantly more CNVs in the DD cases genotyped in Bonn than were identified in the controls. The four most significant of these were 2 regions on 16p13.3, one on 17q25 and one on 19p13.3. These regions all harboured a number of genes, each of which were investigated to see if they had shown significance in a previous NeuroDys study. Within the first region on 16p13.3, 2 deletions were identified in the cases and none were found in the controls. The two CNVs in this region overlap 25 genes. Of these, perhaps the most interesting is the gene STIP1 homology and U-box containing protein 1 (*STUB1*). *STUB1* lies on the section of this region which is overlapped by both of the deletions identified in the cases and did not show significant association with DD in the previous GWAS. What makes this gene particularly interesting is that it encodes the protein CHIP which has been shown to interact with the DD susceptibility gene, *DYX1C1* (Hatakeyama et al. 2004). CHIP promotes the degradation of a variety of proteins, including the estrogen receptors ERα and ERβ. The CHIP-mediated degradation of ERα has been shown to be ligand-dependent and is blocked when estrogen is added, whilst degradation of ERβ is estrogen-dependent (Fan et al. 2005; Tateishi et al. 2004; Tateishi et al. 2006). Estrogen receptors are important in brain development (Wang et al. 2001; Wang et al. 2003; McCarthy 2008) and are also thought to be involved in cognitive processes and memory (Liu et al. 2008; Luine et al. 1998; Fugger et al. 2000; Rissman et al. 2002). ERβ in particular has a role in neuronal migration and neuronal survival in the developing cortex and *Erβ* knock-out mice show very similar phenotypes as the post-mortem brains of DD individuals (Wang et al. 2001; Wang et al. 2003; Massinen et al. 2009). Interestingly, Massinen et al. (2009) showed that over-expression of *DYX1C1* reduces the protein levels of ERα and ERβ, which doesn't quite fit in with the findings that knockdown via *in utero* RNAi of *Dyx1c1* prevents correct neuronal migration and causes malformations similar to those seen in post-mortem studies of individuals with DD (Threlkeld et al. 2007; Wang et al. 2006; Rosen et al. 2007). If *STUB1* had a reduced copy number (as in these DD cases), one might expect the expression of this gene to be reduced, resulting in a reduction of CHIP and therefore a reduction in the degradation of the estrogen receptors, which doesn't fit in with the model that a reduction of ERβ affects neuronal migration and may result in a DD phenotype. However, as explained in section 7.1.3 of this chapter, CNVs can influence phenotypes in a variety of ways and the effects of deleting *STUB1* would need to be investigated in functional studies in order to fully understand how CNVs in this region may result in a DD phenotype.

There were two other genes in this region which are also of potential functional interest. *SOLH*, which is thought to be involved in protein-protein interactions during visual development (Kamei et al. 1998) and two different SNPs in this gene had shown significant association with DD in the previous DD GWAS studies; one in the initial NeuroDys GWAS (rs7763, P = 0.012) and the other in the Cardiff pooling study (rs9934705, P = 0.030), however neither of these SNPs would have passed multiple testing in these studies. The other functionally interesting gene in this region is *METRN* which is thought to be involved in both glial cell differentiation and axonal network formation during neurogenesis (Nishino et al. 2004), giving it a potential role within neuronal migration. However, these genes were only overlapped by one of the deletions in this region as opposed to *STUB1* which was overlapped by both.

The other significantly associated region of 16p13.3 lies ~640kb away from the previous one and also harbours two deletions in cases with DD and no CNVs in the controls. These CNVs overlap 13 genes, of which two are particularly interesting. *BAIAP3* (or *BAP3*) lies in the region which is overlapped by both of the case deletions and this gene encodes a brain-specific angiogenesis inhibitor which shows predominant expression in the brain (Shiratsuchi et al. 1998). Shiratsuchi et al. (1998) found that the expression profile of this inhibitor protein and its sequence similarity to Munc13 and synaptotagmin suggests that it may be involved in neuronal processes such as regulating the release of neurotransmitters. Synaptotagmin promotes the formation of filopodia in fibroblasts (Feany & Buckley 1993) which are involved in motility and recognition in growth cones and so, by similarity, *BAIAP3* may also have a role in neuronal migration.

The other interesting gene in this region is *CACNA1H* which encodes a T-type member of the alpha-1 subunit family which is a protein in the voltage-dependant calcium channel complex and these types of channels may be involved in modulating the firing patterns of neurons (Perez-Reyes 2006).

Another significant region was 19p13.3 which harboured two deletions in the cases with none identified in the controls. These CNVs overlap 11 genes, none of which are of known functional interest within DD. Perhaps the most interesting gene in this region is *PLK5P* which showed nominal significance in the initial NeuroDys GWAS (rs1040499, P = 0.048). However, this gene is on the very edge of the region and is only partially overlapped by one of the deletions. In addition, *PLK5P* is thought to be a psuedogene and so there is very little information regarding its function.

The fourth most significant region was 17q25.1 in which two duplications were identified in the cases and none were found in the controls. These duplications overlap exactly the same region which contains 14 genes. SNPs in or near three of these genes showed significant association with in the initial NeuroDys GWAS (rs11711196 4.4kb downstream of *CD300LB*, P = 0.11 and rs783239 in *CD300LD* and *CD300E*, P = 0.009) and a SNP in another gene showed significant association in the NeuroDys pooling study (rs2706506 5.6kb upstream of *CD300A*, P = 0.011). These genes are all members of the *CD300* molecule family which are leucocyte surface molecules that regulate dendritic cell and monocyte function and so are thought to trigger or inhibit immune responses (Clark et al. 2009). As mentioned previously, this may be interesting as the *DYX2* locus lies over the human histocompatability antigen (HLA) region which contains many genes that influence immune function. As such, immune disorders have previously been linked with DD, although the evidence for this has been inconclusive (Gilger et al. 1998).

All of these CNVs need to be validated using other methods before they can be relied upon. In addition, the CNVs identified in these regions are very rare and identified in only a couple of cases and so may only confer a susceptibility to DD in a small proportion of individuals. CNV studies on larger samples will need to be conducted to see if the results in these regions can be replicated. Based on this evidence alone, the *STUB1* gene appears to be the most interesting gene from the regional analysis and this gene had not been significantly associated with DD in previous studies.

7.4.4 Overlap with ADHD CNVs

12 regions across the genome were identified in which DD CNVs overlapped CNVs that had been found in cases with ADHD by Elia et al. (2010) and Lesch et al. (2010). CNVs in 6 of these regions were also found in the 1958 Birth Cohort but no CNVs in these regions were identified in the screened controls in each of these ADHD studies. As the 1958 Birth Cohort had not been screened, it is possible that some of the individuals may have symptoms of DD or ADHD and so the pathogenicity of the CNVs in these regions for these disorders cannot be ruled out. Further investigation of these regions may provide information behind the mechanisms underlying the comorbidity of ADHD and DD.

Interestingly, the region 17q25.1 discussed above which showed a significant excess of duplications in the DD cases was also found to contain a CNV in a case with ADHD (Lesch et al. 2010). As genes in this region are associated with the immune system, this region may provide a link between ADHD and DD via disorders of the immune system.

Another interesting result is that the genes *PTPRD* and *PARK2*, which were overlapped by ADHD CNVs, were also overlapped by CNVs in DD cases and these genes are thought to be involved in learning. Deletions were found to overlap *PTPRD* in 4 ADHD cases by Elia et al. (2010) and in one DD case and 3 controls in this study. The protein encoded by *PTPRD* is a member of the protein tyrosine phosphatise (PTP) family and has been shown to have high expression in the hippocampus of mice (Mizuno et al. 1993). It is thought to be involved in spatial learning and axonal guidance of motor neurons (Uetani et al. 2000; Uetani et al. 2006), making it good candidate gene for DD as well as ADHD.

PARK2 was overlapped by a duplication and a deletion in two ADHD cases in the study by Elia et al (2010). The same region of *PARK2* was also overlapped by 3 duplications identified in DD cases and 7 deletions and 3 duplications identified in the controls. The precise function of *PARK2* is still unknown but mutations in this gene are known to cause a familial form of Parkinson's disease (Kitada et al. 1998). It encodes Parkin which is an E3 ubiquitin ligase and loss of this ubiquitin ligase activity appears to be the mechanism underlying autosomal-recessive juvenile Parkinsonism (von Coelln et al. 2004). Interestingly, Parkin has been shown to be positively regulated by CHIP (Imai et al. 2002), the protein that is encoded by *STUB1* which was overlapped by a significant excess of deletions in the DD cases. This suggests that these genes may lie on a pathway which may be involved in DD. CHIP was shown to enhance Parkin-mediated ubiquitination (and subsequent degradation) of the Pael receptor (Pael-R), the accumulation of which in the endoplasmic reticulum of dopaminergic neurons can lead to neurodegeneration (Imai et al. 2002). CNVs overlapping this gene have also been found in cases with schizophrenia (Xu et al. 2008), suggesting that this gene is a hotspot for CNVs and may be involved in the pathogenicity of a variety of neurological disorders.

7.4.5 CNVs in Hotspot Regions

The hotspot regions that have been identified across the genome may indicate that there is a lack of specificity for phenotypes caused by CNVs in the same region. The wide range of phenotypes associated with rearrangements in a number of loci points to a common disease mechanism for a wide range of neurocognitive deficits. It is possible that while these deletions and duplications are primarily responsible for disease, the actual specificity of disease is determined by other genetic, epigenetic and environmental modifiers (Mefford & Eichler 2009).

Of the hotspot regions highlighted by Itsara et al (2009) and Mefford and Eichler (2009), 5 were overlapped by CNVs in this study. Perhaps the most interesting of these regions is 9q24.3 which was overlapped by 2 duplications identified in the DD cases and 5 duplications in the controls. CNVs in this region have also been identified in cases with schizophrenia (ISC 2008) and autism (Autism Genome Project Consortium 2007) as well as in an individual with ADHD (Elia et al. 2010). The CNVs identified in these disorders all overlap the *DOCK8* gene. This gene has been shown to interact with Cdc42 and is thought to affect the organisation of filamentous actin (Ruusala & Aspenström 2004). Cdc42 has a well recognised role in neuronal migration (Kozma et al. 1995; Nobes & Hall 1995; Luo 2000), as discussed in Chapter 3, but it was not shown to be significantly associated with DD in the candidate gene study of this thesis. This suggests that as well as in DD, impaired/altered neuronal migration may also be involved in the pathogenicity of ADHD, autism and schizophrenia. The majority of CNVs found to be overlapping *DOCK8* are duplications, suggesting that having extra copies of this gene may somehow be involved in neurological diseases. This needs to be investigated further through functional studies.

The region 22q11.21-q11.22 seems to be particularly prone to CNVs, with deletions and duplications identified in individuals with DD in this study, schizophrenia (ISC 2008; Xu et al. 2008), autism (Autism Genome Project Consortium 2007; Christian et al. 2008), mental retardation (de Vries et al. 2005) and developmental delay (Ben-Shachar et al. 2008; Ou et al. 2008). This is likely to be due to the segmental duplications flanking this region altering copy number via NAHR. This study identified a duplication in a DD case and in one of the controls. Autistic individuals with duplications in this region were observed to have intellectual disability, neuropsychological problems and speech disorder (Christian et al. 2008). This is particularly interesting as speech disorders are thought to share some comorbidity with DD, with half of the individuals with SLI also having DD (Flax et al. 2003), as discussed in Chapter 1. In addition, Ou et al (2008) found that individuals in this region all had some degree of general developmental delay or speech delay and one individual also had ADHD. This suggests that this region may be important during intellectual and speech development, but the effects of rearrangements in this region may be modified though other genetic or environmental factors to cause a range of diseases with varying severity.

Combing the hotspot regions identified by Itsara and colleagues (2009) and Mefford and Eichler (2009) resulted in large regions of the genome being tested which may have been over-inclusive. As such, these regions may have contained CNVs within DD by chance so it is difficult to be sure that these CNVs are contributing to disease in a similar way to other CNVs that have been identified in these regions. However, the DD CNVs still overlap some interesting regions, and if validated, could be worthy of follow-up.

7.4.6 Summary and Future Work

This CNV study has highlighted a number of novel candidate regions for DD that had not shown association in the previous GWAS studies. Carrying out CNV analysis using GWAS data enables the investigation of genetic variants that may not have been identified in the GWAS at no extra experimental cost. However, larger studies need to be conducted to replicate the presence of CNVs in these regions in individuals with DD and functional studies need to be carried out to understand how copy number changes in these genes may be involved in increasing an individual's susceptibility to DD. It is important to note that the CNVs identified in this study need to be validated using an alternative method such as aCGH or qPCR before they can be relied upon. If these CNVs are indeed real, then the genes *STUB1* and *PARK2* and the interaction between them and *DYX1C1* would be of particular interest.

This study also highlights the issues that can arise when carrying out CNV analysis using GWAS data. Whilst SNP genotype calls may not be as adversely affected by carrying out genotyping in different centres (provided the appropriate QC is carried out), the intensity scores are far more sensitive to experimental variation which may increase the rate of false positive results. Future GWAS should ideally be carried out with cases and controls being genotyped on the same arrays at the same centres in order to reduce the variability in the intensity signals. The case sample size used in this study is very small and underpowered to detect the rare CNVs. Future CNV studies of DD need to be carried out using much larger samples and with age-matched screened controls in order to confidently identify CNVs that may be involved in DD.

The array used to genotype the cases was one of the early SNP arrays that did not contain any probes for CNVs. New hybrid arrays include non-polymorphic probes to detect common CNVs and probes for many more SNPs, such as those purposely excluded from early arrays (McCarroll et al. 2008) and so future CNV studies should be conducted using these arrays. However, in general, the resolution of these platforms still needs to be improved and there is a need for technologies that can confidently resolve genes down to less than 1kb in size, as these are much more abundant in the human genome while still having the capacity to disrupt genes (Kirov 2010). The development of next-generation sequencing has enabled resolution of whole genomes down to the single nucleotide level, and as the cost of these techniques decrease, they may well become the method of choice when conducting CNV analysis.

In conclusion, whilst this is a small study on the role of CNVs in DD, it has highlighted a number of novel regions that may have not been considered interesting based on the initial GWAS results. CNVs appear to affect a wide range of phenotypes and are worthy of further investigation in larger, more homogeneous samples for future studies of DD.

Chapter 8: Final Discussion

The main aim of the research presented in this thesis was to identify susceptibility variants for DD using several approaches. DD is one of the most common neurobehavioral disorders and has a highly heritable component (see Chapter 1). Linkage and association studies have been conducted to identify genes that underlie the genetic predisposition to DD. So far, nine putative susceptibility loci for DD have been identified and designated *DYX1-DYX9* by the Human Gene Nomenclature Committee (www.genenames.org).

A candidate gene study was conducted selecting genes within the *DYX* susceptibility loci that had plausible roles in neuronal migration or showed homology with genes which have shown previous association with DD. The first GWAS of DD was conducted by the NeuroDys collaboration and data from this study was analysed to identify new susceptibility variants for DD. An additional GWAS was also carried out as part of the NeuroDys collaboration in the form of a pooling study, testing a larger number of SNPs across the genome for an association with DD. Pooled DNA was also used to conduct a GWAS using just the Cardiff case control sample in an effort to identify susceptibility variants that are significantly associated with DD in this homogenous case-control sample Finally, CNV analysis was also conducted using data from the first NeuroDys GWAS to investigate this source of variation for an association with DD.

8.1 Research Findings

Initial work focused on identifying new susceptibility variants for DD by testing variants within a small number of candidate genes for an association with DD in the Cardiff case-control sample. The genes *CDC42* and *PRTG* were selected based on their location within DD susceptibility loci (*DYX8* and *DYX1*, respectively) and because of their putative roles within neuronal migration. The genes *KIAA0319L* and *DCDC2b* were also selected for this candidate gene study due to their position within the *DYX8* susceptibility locus and their homology with two replicated susceptibility genes for DD, *KIAA0319* and *DCDC2*, respectively. Finally, *RIOK3* was also tested for an association with DD because a survey of gene expression showed that two SNPs within the putative DD susceptibility gene *KIAA0319* were associated with expression levels of *RIOK3* (Myers et al. 2007). These genes were tested for association using tag SNPs identified from HapMap with the exception of *DCDC2b* in which high resolution DNA melting analysis was used to identify novel SNPs. None of the variants tested within these genes showed a significant association with DD in the Cardiff case-control sample after correction for multiple testing. However, no study of this size can confidently exclude a gene from involvement in disease susceptibility. Furthermore, the lack of significant association of the putative neuronal migration genes *CDC42* and *PRTG* with DD in this study does not discount this pathway from having a role within DD.

The next stage of research focussed on a GWAS of DD which was conducted as part of the NeuroDys collaboration. This was carried out using the Illumina HumanHap300 array and a discovery sample of 585 cases and 2326 population controls from the UK and Germany. Twenty-seven of the most significant SNPs identified were selected from this GWAS and followed up in an independent replication sample of 1244 cases and 1955 screened controls from the UK, Germany, Switzerland, the Netherlands, Austria and Finland. Eight of these SNPs had P-values < 0.05 in the replication sample with none achieving genome-wide significance. Many SNPs showed a different direction of effect in the replication sample compared to the discovery sample, suggesting a high false positive rate in the SNPs selected for follow-up from the intial GWAS. When combining genotyping data from both sample sets, the most interesting result came from the *SNX29* gene in which two SNPs showed an increased significance in the combined sample (rs6498274: P-gen = 2.46×10^{-4}, P-add = 2.65×10^{-5}, OR = 1.19; rs905950: P-gen = 5.19×10^{-4}, P-add = 4.52×10^{-5}, OR = 1.18) compared to their results in the initial discovery sample alone (rs6498274: P-gen = 1.32×10^{-4}, P-add = 4.47×10^{-5}, OR = 1.31; rs905950: P-gen = 5.19×10^{-4}, P-add = 2.04×10^{-5}, OR = 1.30). This gene is a member of the sorting nexin family and this family of proteins contain PX domains which are thought to be involved in cell polarity (Worby & Dixon 2002; Teasdale et al. 2001), which may suggest a possible link with neuronal migration. Of the previously replicated susceptibility genes for DD, only *DCDC2* showed significant association in the GWAS. Eleven SNPs within *DCDC2* were significantly associated with DD in the GWAS and three of these were in LD with each other and the SNP rs807701 which has shown association with DD in previous studies (Schumacher et al. 2006a). A large number of SNPs within the *DYX* susceptibility loci were significant at the 0.05 level. However, after running set-based analysis to correct for the number of independent SNPs in each region, one SNP within the *DYX5* locus remained significant (rs6796074, best corrected P-add = 0.008). This SNP lies in an intergenic region towards the edge of the *DYX5* susceptibility locus, 129kb from the polymorphism D3S3665 which has shown evidence of linkage with DD previously (Nopola-Hemmi et al. 2001).

As the significant SNPs in the initial GWAS did not show strong evidence of association with DD in the replication study, the NeuroDys collaboration conducted a larger GWAS using the replication sample from the first GWAS and the Illumina 1M-Duo array. This was carried out in the form of a pooling study in order to reduce the time and cost of this second GWAS. There was little concordance between the two GWAS, with only 4 SNPs of the most significant SNPs of the pooling study also being significant in the initial GWAS at $P < 0.05$. 38 of the most significant SNPs were selected for individual genotyping in the 988 cases and 1121 controls that had been pooled. After individual genotyping, 14 SNPs were significant at the 0.05 level. This indicates that there were some inaccuracies in the estimations of the allele frequencies from the pooled samples, possibly due to pool-formation or pool-measurement errors. When including an additional 568 cases and 1140 controls, 6 SNPs showed an increased level of significance. The most significant of these was rs461119 (P-add = 3.0×10^{-4}, OR = 1.23) which is within the *GRIK1* gene on chromosome 21. This gene encodes a glutamate receptor, and as these are the predeominant excitatory neurotransmitters in the mammalian brain (Headley & Grillner 1990), this gene could be a good functional candidate for DD. The next most significant SNP was rs12344734 (P-add = 0.0021, OR = 1.2) which is within the *TMC1* gene on chromosome 9. Mutations in this gene have previously been associated with progressive post-lingual hearing loss and profound pre-lingual deafness (Kurima et al. 2002; Vreugde et al. 2002) and so this gene could have a possible role in an auditory component of DD. Although not as significant, another interesting SNP was rs4655653 (P-add = 0.0149, OR = 1.15) which is within the gene *WDR78*. This gene belongs to the WD-repeat family of proteins. *LIS1* (a gene involved in the neuronal migration disorder lissencephaly) also contains WD-repeat domains and is thought to be in a crucial pathway for cerebral development (Reiner et al. 1993). If the WD-repeat domains of *WDR78* share a similar function to those of *LIS1* then *WDR78* could be a good candidate for DD.

An additional GWAS was conducted using the Illumina 1M-Duo array on just the Cardiff cases and controls in order to identify susceptibility variants that show association in this more homogeneous sample of severe DD cases and screened controls. This was also carried out in the form of a pooling study in order to reduce the time and cost. 57 of the most significant SNPs were genotyped in 292 cases and 215 controls which had been included in the pools. Whilst the significance of these SNPs was generally lower when individually genotyped, 54 remained significant at the 0.05 level, which is a higher rate than in the NeuroDys pooling study. This suggests that estimations of allele frequencies were more accurate for these pools, possibly due to the use of a homogeneous sample which had been collected, extracted and prepared in the same centre. When including an additional 39 cases and 47 controls, the most significant SNP was rs1125198 (P-value = 1.3 x 10^{-4}, OR = 2.19) which is within a potentially functionally interesting gene, *MLKN1*. This gene shows highest expression in the cerebellum and hippocampus and may have a role within synaptogenesis (Tagnaouti et al. 2007). When including data for 3748 population controls from the 1958 Birth Cohort, two SNPs showed a higher level of significance. The most significant SNP was rs7330054 (P = 2.8 x 10^{-5}, OR = 1.71) which is within *COL4A2*. This gene has been linked with a number of cerebral small vessel diseases in humans which are thought to affect information processing speed and executive function (Volonghi et al. 2010; Prins et al. 2005). Therefore, this gene may have a role within DD susceptibility by affecting an individual's ability to link graphemes with phonemes and so impairing their ability to read efficiently. The other significant SNP was rs10844773 (P = 7.8 x 10^{-5}, OR = 0.51). This SNP is within *CD163L1* which encodes a member of SRCR superfamily which are mainly found in cells of the immune system (Gronlund et al. 2000; Van Gorp et al. 2010). This could be interesting within the context of DD as DD has previously been linked with immune disorders (Gilger et al. 1998). In general, there was little concordance across all the GWAS. Although many of the most significant SNPs within the Cardiff pool also showed a high level of significance in the UK pool, they were not significant in the Central European or Finnish pools.

Finally, a CNV analysis was conducted using data from the Cardiff cases in the initial GWAS and the 1958 Birth Cohort as controls, identifying 1147 rare CNVs. There was a significant excess of deletions in the cases compared with the controls (P = 1 x

10^{-4}, case/control ratio = 2.00). However, this excess appeared to be driven by technical differences in those cases that had been genotyped in Oxford, suggesting that the CNVs in this dataset need to be treated with caution. This highlights the need for homogenous samples that have been prepared and genotyped in the same centre when conducting CNV studies. When removing those samples that were genotyped in Oxford, there was no longer a global significant excess of CNVs in the cases, but a number of regions showed a higher rate of CNVs in the remaining cases compared with the controls. Perhaps the most interesting of these is the region on 16p13.3 (P = 0.004, 2 deletions in cases, none in controls). The two deletions identified in cases in this region both overlap the gene *STUB1* which encodes the protein CHIP. CHIP has been shown to interact with the putative DD susceptibility gene *DYX1C1* (Hatakeyama et al. 2004) and promotes the degradation of estrogen receptors which are thought to be involved in cognitive processes and memory (Liu et al. 2008; Luine et al. 1998; Fugger et al. 2000; Rissman et al. 2002). In particular, the estrogen receptor ERβ has a role within neuronal migration and the brains of *ERβ* knock-out mice show similar phenotypes to those identified in post-mortem studies of brains of individuals with DD (Wang et al. 2001; Wang et al. 2003; Massinen et al. 2009). This suggests that *STUB1* may affect an individual's susceptibility to DD via the action of estrogen receptors. Another interesting region identified in this CNV study was 6q26. There was a significant excess of CNVs overlapping the *PARK2* gene within this region (P = 0.048, case/control ratio = 3.4), particularly for duplications (P = 0.019, case/control ratio = 8.62). Duplications and deletions in this region have also been identified in individuals with ADHD (Elia et al. 2010), suggesting that this region could provide information about the shared genetic aetiology between these disorders. Mutations in *PARK2* cause a familial form of Parkinson's disease (Kitada et al. 1998). Interestingly, PARK2 has also been shown to be positively regulated by CHIP (Imai et al. 2002), the protein encoded by *STUB1*. This raises the possibility that both of these genes lie on a pathway which could affect an individual's susceptibility to DD. CNVs in this region have also been identified in schizophrenia (Xu et al. 2008), suggesting that this region is a hotspot for CNVs. CNVs in individuals with DD were found to overlap other regions of the genome which appear to be hotspots for CNVs in a number of neurocognitive diseases, suggesting that CNVs may be involved in a wide range of diseases. However, the CNVs identified in this study need to be interpreted with caution and validated with alternative methods before these results can be relied upon.

8.2 Thesis Limitations

The main limitation of the studies conducted in this thesis has been power. Power calculations show that in order to have at least 80% power to detect a significant association with a variant that has a MAF of 0.4 and an OR of 1.3 at the 0.05 level, more than 450 cases and 450 controls are required. At the genome-wide level of significance, this increases to over 2300 cases and 2300 controls. The study designs used in this thesis would have only detected common variants of large to moderate effect sizes and so it is likely that some variants associated with DD were missed due to their small effect size. The recruitment of a larger sample set of cases and screened controls which have been ascertained using the same, stringent criteria could greatly increase the power to detect true associations with DD.

In addition, the methods used in these studies would only be able to detect association with common polymorphisms in the genome. As discussed in Chapter 1, the genetic aetiology of DD appears to be complex and it could be caused by a combination of common polymorphisms of various effect sizes as well as rare polymorphisms, gene-gene interactions, gene-environment interactions and other unsuspected genomic mechanisms which were not explored in this thesis.

Conducting a candidate gene study allows for genes to be prioritised based on previous evidence of association or linkage as well as in terms of their suspected functional relevance within a disorder or trait. The nature of candidate gene studies makes them considerably cheaper than GWAS and enables specific genes of interest to be densely mapped and tested for association. However, the precise biology of DD is unknown. Whilst replicated susceptibility genes for DD so far have highlighted a possible role for neuronal migration within DD, this is a complex process involving a large network of pathways. Therefore, it is possible to find evidence connecting a large number of genes to DD via neuronal migration. Selecting those that were within replicated linkage regions of DD may have improved the selection of these candidate genes, but these regions are quite broad, containing a large number of genes of possible functional interest.

Carrying out the first GWAS of DD enabled the systematic testing of the whole genome for an association with reading disability. Because no prior assumptions are made about the biology of this complex disorder or the location of the variants, this represents an unbiased approach to identifying new susceptibility loci for diseases (Hirschhorn & Daly 2005). In order to increase the sample size, this study was conducted by the NeuroDys consortium, involving samples from 6 European countries. Whilst this increased the power of this study, using heterogeneous sample sets may have introduced other limitations such as population stratification. Population stratification may cause differences in allele frequencies between cases and controls, whether the alleles are causally related to the disease or not. These samples were collected by different centres, using slightly different ascertainment criteria and were genotyped in different centres, all of which may have contributed to the false positive rate. The use of population controls can increase the power of a study without a significant increase in genotyping costs. However, the population controls were genotyped on different arrays to the cases and in different centres, which may have also contributed to the false positive rate. Furthermore, these population controls had not been screened and so it is possible that some of the controls had symptoms of DD, which would reduce the power of this study to identify susceptibility variants. Population stratification was controlled for as much as possible through the use of the MDS component and centre as covariates when performing logistic regression analysis, but the false positive rate may have still been inflated in comparison to a study that used a more homogeneous sample of cases and controls, which may explain the lack of replication in the replication sample. The Illumina HumanHap300 array used in this GWAS is one of the earlier SNP arrays and provides 76% coverage of common genetic variation in the human genome (Mägi et al. 2007). It is therefore possible that some of the variants that confer susceptibility for DD were missed. Furthermore, the distribution of markers across the genome is not even, which means that some loci have good coverage, while others have very poor. For example, one of the well replicated susceptibility genes for DD, *KIAA0319,* only had coverage of 48% in this GWAS.

The second GWAS conducted by the NeuroDys consortium used the Illumina Human1M-Duo array which has an improved coverage of 95%, however a large proportion of the SNPs on this array were excluded during QC filtering. Conducting this second GWAS in the form of a pooling study provided significant savings in terms of both time and money. However, it also resulted in a loss of detailed information that could have been obtained through individual genotyping. For example, the design of this study prevented the analysis of componential phenotypes of DD and does not allow the formation of haplotypes. There was also likely to have been a loss in power as a result of specific errors that can arise in pooling studies, such as pool-formation and pool-measurement errors. This pooling study combined samples which had been extracted and prepared in different centres and this may have resulted in inflation of the pool-formation error rate. These errors affect the ability of a study to estimate allele frequency differences accurately and so can increase the rate of false positives, which may have been why only a small proportion of the top SNPs remained significant after individual genotyping. Due to insufficient funding, individual genotyping was not carried out with those samples which had been included in the Finnish pool, which may have also affected the number of SNPs that were significant after individual genotyping. In contrast, the Cardiff case-control pooling study used a uniform sample of cases and controls and had a much higher proportion of SNPs remaining significant after individual genotyping, indicating a greater level of accuracy when estimating the allele frequencies. This suggests that in order for pooling studies to be a valuable method of screening large samples at a lower cost, they ideally need to be conducted using samples that have been collected and prepared in the same centres. However, this is not always possible and meant that the Cardiff pooling study had a much smaller sample, and therefore a reduced power to detect variants of small effect sizes. This is illustrated by the observation that the SNPs which remained significant after individual genotyping all had relatively large effect sizes (OR between 1.43 and 2.79). In both of these pooling studies, only a relatively small number of significant SNPs could be followed up with individual genotyping due to financial constraints and more significant associations may have been confirmed with individual genotyping if larger panels of SNPs had been selected.

The use of intensity data from GWAS using SNP arrays enables CNVs within the genome to be tested for an association with a particular disease or trait, without an increase in experimental costs on top of the initial GWAS. However, the use of SNP arrays for CNV analysis has a number of limitations. As mentioned earlier, the SNP probes on an array are not uniformly distributed across the genome, particularly with the earlier arrays. They are especially sparse in regions of segmental duplications which create problems when designing robust SNP assays in these regions. This means that common CNVs cannot be reliably detected using such methods. These arrays are also not able to reliably identify smaller CNVs, which is why stringent QC filters in terms of size and probe density needed to be used. However, this may have resulted in the exclusion of some causal CNVs. Further limitations come from the design of the initial GWAS, in which the population controls were genotyped on a larger array to the cases, which may have introduced bias in the rates of CNVs identified between the cases and controls. However, this difference did not appear to have a significant impact in this study. A more significant limitation was that the cases had been genotyped in two different centres, and technical differences in those genotyped in Oxford appeared to be driving a significant excess of deletions in the cases. Whilst SNP genotype calls may not be adversely affected by carrying out genotyping in different centres, the intensity scores are far more sensitive to experimental variation which may increase the rate of false positive results. As such, it is important that these CNVs are validated using other methods before the results of this study can be relied upon. However, even if all the cases and controls had been conducted in the same centre, it is likely that these samples would have still been extracted at different times, and analysed in different experimental plates or batches, all of which would influence the intensity signals from sample to sample (McCarroll 2008). It is therefore important to carefully dissect the sample- and batch-specific influences on signal intensities and develop algorithmic approaches which can control for these factors. This is beyond the scope of this current project however.

8.3 Further Work

There is still a lot more research to be done in order to gain a better understanding of the genetic component of DD. This study has highlighted a number of interesting variants, but these all need to be replicated in large independent samples before they can confidently be considered to have a role within susceptibility to DD. However, the power of these analyses is inadequate and further susceptibility variants for DD could be identified upon further interrogation of the genome.

Future meta-analysis of GWAS of DD may provide greater power to detect susceptibility variants for this disorder. This will require collaboration with more groups that have used appropriate ascertainment criteria. Additional control data from GWAS of other population cohorts could also be included to increase power further. Even if these studies are conducted on different arrays to the one used in this dataset, imputation could be used in order to impute the genotypes of un-typed markers before combining the data.

In terms of the GWAS data already collected, further work may involve sub-phenotype analysis, haplotype analysis and gene ontology analysis. Although heritability estimates of DD have suggested that there may be a shared genetic aetiology between many components of reading, Castles and colleagues (1999) have suggested that there may also be some partial genetic independence between the cognitive processes involved in reading, such as phonological skills and orthographic skills. Analysis of these sub-phenotypes of DD may highlight variants which are strongly associated with particular components of DD. This would require detailed phenotypic information for all the cases and controls and it may be difficult to standardise the component tests used across all groups of the NeuroDys collaboration, particularly when comparing those conducted in different languages.

Haplotype analysis of the GWAS data may identify significant associations that were not seen with single SNPs. Such an association may suggest that the haplotype itself is directly associated with DD or that the haplotype indirectly tags the true associated variant more effectively than any individually genotyped SNPs (McCarthy et al. 2008). Over recent years novel methods of haplotype analysis have been implicated or are under development, which are fast enough and hence practical to use even for marker densities of 500,000 SNPs/genome (Nolte et al. 2007).

Gene ontology (GO) is a method used to describe gene products in terms of their associated biological processes, cellular components and molecular functions in a species-independent manner (Harris et al. 2004). A number of methods have been developed to identify GO categories that are amongst association signals in a GWAS and so prioritise genes/pathways for further investigation. These methods are based on the assumption that susceptibility alleles for a given disorder are likely to be distributed among genes whose functions are, to some extent, related (Holmans et al. 2009). Under this model one would expect to see an overall excess of SNPs with moderate P-values for association on a list of SNPs representing a set of genes from relevant related biological pathways (Holmans et al. 2009). In Cardiff, Prof. Pete Holmans has designed the ALIGATOR program, which converts a list of significant SNPs into a list of significant genes, applying corrections for gene size and non-independent categories and is not influenced by LD (Holmans et al. 2009). These genes are then examined for any enrichment of individual GO categories.

In terms of CNV analysis, future work could involve attempting to validate the interesting CNVs identified in this study using an alternative method such as qPCR. Alternatively, an additional whole genome scan of CNVs could be conducted using aCGH methods. These approaches provide the most robust method for scanning the whole genome (Pinkel et al. 1998) and custom arrays (e.g. Agilent custom arrays) could be produced in order to densely cover particular regions of interest. In addition, a CNV study could be conducted on probands and their parents in the form of trios to determine if any *de novo* or inherited CNVs are associated with DD.

The recent development of 'next-generation' sequencing technologies are set to have a large impact on the field of genetics. What sets next-generation sequencing apart from conventional capillary-based sequencing is the ability to process millions of sequence reads in parallel rather than 96 at a time (Mardis 2008). Being a relatively new technology, next generation sequencing is still expensive. However, sequencing platforms are now available which enable researchers to prioritise genome-wide exon (or 'exome') sequencing in order to sequence coding regions at a workable financial cost. With the appropriate algorithms, analysing the genome on a single nucleotide level will enable the identification of rare variants which cannot be tested using commercial SNP arrays. It will also enable accurate detection of large deletions and duplications, as well as inversions and translocations (Fanciulli et al. 2010).

The development of next-generation sequencing has been of particular importance to the 1000 Genomes Project (Via et al. 2010). This project is an international collaboration and involves sequencing 2000 individuals from at least 20 different populations representing Africa, Europe, East Asia, and the Americas. The main goal of this project is to describe most of the genetic variation that occurs at a population frequency of $> 1\%$ (Via et al. 2010). The first stage of this analysis has already been conducted and of the 9 million novel SNPs that have been identified so far, approximately 8 million are seen in only one HapMap population (Via et al. 2010). It is hoped that the results of this project will allow researchers to identify genetic variation at a greater degree of resolution and also improve current imputation methods. It will enable the development of new population-specific genotyping arrays, maximising genome coverage while minimising the ascertainment bias that affects currently available arrays, especially for non-European populations (Via et al. 2010). It will also provide a valuable resource as a reference genome when attempting to identify novel rare variants or conducting CNV analysis.

8.5 Conclusions

This study has sought to identify susceptibility variants for DD using a number of techniques including a candidate gene study, collaborative GWAS, pooling studies and CNV analysis. Despite the lack of genome-wide significant findings, these approaches have been successful in identifying a number of interesting regions but these regions need to be confirmed in large independent samples. Future studies of DD may benefit from a combination of approaches, with systematic analysis of the genome using SNP arrays or next-generation sequencing in large, well-powered samples to highlight regions of interest, followed by CNV analysis and targeted candidate gene studies.

By improving our understanding of the genetics underlying this complex neurobehavioral disorder, we can hope to gain a better understanding of the biological processes involved in reading in general and how and where particular deficits may occur. It may also enable young children to be screened for their potential risk of developing DD and allow tailored tuition to be provided in order to reduce this risk.

References

Abraham, R. et al., 2008. A genome-wide association study for late-onset Alzheimer's disease using DNA pooling. *BMC Medical Genomics*, 1, 44.

Adams, J.C. et al., 1998. Muskelin, a novel intracellular mediator of cell adhesive and cytoskeletal responses to thrombospondin-1. *The EMBO Journal*, 17(17), 4964-4974.

Adlard, A. & Hazan, V., 1998. Speech Perception in Children With Specific Reading Disabilities (Dyslexia). *The Quarterly Journal of Experimental Psychology*, 51A, 153-177.

Ago, T. et al., 2001. The PX domain as a novel phosphoinositide- binding module. *Biochemical and Biophysical Research Communications*, 287(3), 733-738.

Ahissar, M. et al., 2000. Auditory processing parallels reading abilities in adults. *Proceedings of the National Academy of Sciences of the United States of America*, 97(12), 6832-6837.

Altmüller, J. et al., 2001. Genomewide scans of complex human diseases: true linkage is hard to find. *American Journal of Human Genetics*, 69(5), 936-950.

Amitay, S. et al., 2002. Disabled readers suffer from visual and auditory impairments but not from a specific magnocellular deficit. *Brain: A Journal of Neurology*, 125(Pt 10), 2272-2285.

Anderson, C.A. et al., 2008. Evaluating the effects of imputation on the power, coverage, and cost efficiency of genome-wide SNP platforms. *American Journal of Human Genetics*, 83(1), 112-119.

Anthoni, H. et al., 2007. A locus on 2p12 containing the co-regulated MRPL19 and C2ORF3 genes is associated to dyslexia. *Human Molecular Genetics*, 16(6), 667-677.

Arcos-Burgos, M. & Muenke, M., 2002. Genetics of population isolates. *Clinical Genetics*, 61, 233-247.

Autism Genome Project Consortium, 2007. Mapping autism risk loci using genetic linkage and chromosomal rearrangements. *Nature Genetics*, 39(3), 319-328.

Ayala, R. et al., 2007. Trekking across the Brain: the journey of neuronal migration. *Cell*, 128, 29-43.

Bache, I. et al., 2006. Systematic re-examination of carriers of balanced reciprocal translocations: a strategy to search for candidate regions for common and complex diseases. *European Journal of Human Genetics*, 14, 410-417.

Bailey, J.A. et al., 2002. Recent segmental duplications in the human genome. *Science*

(New York, N.Y.), 297(5583), 1003-1007.

Bakker, S. et al., 2003. A whole-genome scan in 164 Dutch sib pairs with attention-deficit/hyperactivity disorder: suggestive evidence for linkage on chromosomes 7p and 15q. *American Journal of Human Genetics*, 72, 1254-1260.

Bakwin, H., 1973. Reading disability in twins. *Developmental Medicine and Child Neurology*, 15(2), 184-187.

Balding, D.J., 2006. A tutorial on statistical methods for population association studies. *Nature Reviews. Genetics*, 7(10), 781-791.

Balogh, S.A. et al., 1998. Effects of neocortical ectopias upon the acquisition and retention of a non-spatial reference memory task in BXSB mice. *Brain Research. Developmental Brain Research*, 111(2), 291-293.

Barratt, B.J. et al., 2002. Identification of the sources of error in allele frequency estimations from pooled DNA indicates an optimal experimental design. *Annals of Human Genetics*, 66, 393-405.

Barrett, J.C. et al., 2005. Haploview: analysis and visualization of LD and haplotype maps. *Bioinformatics (Oxford, England)*, 21(2), 263-265.

Bates, E.E. et al., 1998. CD40L activation of dendritic cells down-regulates DORA, a novel member of the immunoglobulin superfamily. *Molecular Immunology*, 35(9), 513-524.

Bates, T.C. et al., 2009. Dyslexia and DYX1C1: deficits in reading and spelling associated with a missense mutation. *Molecular Psychiatry*. Available at: internal-pdf://Bates et al 2009-1432061821/Bates et al 2009.pdf.

Bates, T.C. et al., 2007. Replication of reported linkages for dyslexia and spelling and suggestive evidence for novel regions on chromosomes 4 and 17. *European Journal of Human Genetics*, 15, 197-203.

Beckmann, J.S. et al., 2007. Copy number variants and genetic traits: closer to the resolution of phenotypic to genotypic variability. *Nature Reviews Genetics*, 8, 639-646.

Behan, P. & Geschwind, N., 1985. Dyslexia, congenital anomalies, and immune disorders: the role of the fetal environment. *Annals of the New York Academy of Sciences*, 457, 13-18.

Bellini, G. et al., 2005. No evidence for association between dyslexia and DYX1C1 functional variants in a group of children and adolescents from Southern Italy. *Journal of Molecular Neuroscience*, 27, 311-314.

Ben-Shachar, S. et al., 2008. 22q11.2 distal deletion: a recurrent genomic disorder distinct from DiGeorge syndrome and velocardiofacial syndrome. *American Journal of Human Genetics*, 82(1), 214-221.

Berg, M.J. et al., 1998. X-linked female band heterotopia-male lissencephaly syndrome. *Neurology*, 50(4), 1143-1146.

Bettler, B. & Mulle, C., 1995. Review: neurotransmitter receptors. II. AMPA and kainate receptors. *Neuropharmacology*, 34(2), 123-139.

Bielas, S. et al., 2004. Cortical neuronal migration mutants suggest separate but intersecting pathways. *Annual Review of Cell and Developmental Biology*, 20, 593-618.

Bishop, D.V. & Adams, C., 1990. A prospective study of the relationship between specific language impairment, phonological disorders and reading retardation. *Journal of Child Psychology and Psychiatry, and Allied Disciplines*, 31(7), 1027-1050.

Bishop, D.V. et al., 1999. Different origin of auditory and phonological processing problems in children with language impairment: evidence from a twin study. *Journal of Speech, Language, and Hearing Research: JSLHR*, 42(1), 155-168.

Blachman, B.A., 1984. Relationship of Rapid Naming Ability and Language Analysis Skills to Kindergarten and First-Grade Reading Acheivement. *Journal of Educational Pyshology*, (76), 610-622.

Bodmer, D. et al., 2003. Disruption of a novel gene, DIRC3, and expression of DIRC3-HSPBAP1 fusion transcripts in a case of familial renal cell cancer and t(2;3)(q35;q21). *Genes, Chromosomes & Cancer*, 38(2), 107-116.

Bowers, P.G. & Swanson, L.B., 1991. Naming speed deficits in reading disability: multiple measures of a singular process. *Journal of Experimental Child Psychology*, 51(2), 195-219.

Brown, W.E. et al., 2001. Preliminary evidence of widespread morphological variations of the brain in dyslexia. *Neurology*, 56(6), 781-783.

Browning, S.R., 2008. Missing data imputation and haplotype phase inference for genome-wide association studies. *Human Genetics*, 124(5), 439-450.

Brunswick, N. et al., 1999. Explicit and implicit processing of words and pseudowords by adult developmental dyslexics: A search for Wernicke's Wortschatz? *Brain: A Journal of Neurology*, 122 (Pt 10), 1901-1917.

Burbridge, T.J. et al., 2008. Postnatal analysis of the effect of embryonic knockdown and overexpression of candidate dyslexia susceptibility gene homolog Dcdc2 in the rat. *Neuroscience*, 152(3), 723-733.

Bycroft, M. et al., 1999. The structure of a PKD domain from polycystin-1L implications for polycystic kidney disease. *EMBO Journal*, 18, 297-305.

Cardon, L.R. & Bell, J.I., 2001. Association study designs for complex diseases. *Nature*

Reviews Genetics, 2, 91-99.

Cardon, L.R. et al., 1994. Quantitative trait locus for reading disability on chromosome 6. *Science*, 266(5183), 276-279.

Carter, N.P., 2007. Methods and strategies for analysing copy number variation using DNA microarrays. *Nature Genetics*, 39, S16-S21.

Castles, A. et al., 1999. Varieties of developmental reading disorder: genetic and environmental influences. *Journal of Experimental Child Psychology*, 72(2), 73-94.

Castles, A. & Coltheart, M., 2004. Is there a causal link from phonological awareness to success in learning to read? *Cognition*, 91(1), 77-111.

Cestnick, L., 2001. Cross-modality temporal processing deficits in developmental phonological dyslexics. *Brain and Cognition*, 46(3), 319-325.

Chan, D. et al., 2007. Prevalence, gender ratio and gender differences in reading-related cognitive abilities among Chinese children with dyslexia in Hong Kong. *Educational Studies*, 33(2), 249-265.

Chang, B.S. et al., 2005. Reading impairment in the neuronal migration disorder of periventricular nodular heterotopia. *Neurology*, 64(5), 799-803.

Chapman, N.H. et al., 2004. Linkage analyses of four regions previously implicated in dyslexia: confirmation of a locus on chromosome 15q. *American Journal of Medical Genetics Part B (Neuropsychiatric Genetics)*, 131B, 67-75.

Cheng, C. et al., 2002. Cloning, expression and characterisation of a novel human VMP gene. *Molecular Biology Reports*, 29, 281-286.

Chiu, M.M. & McBride-Chang, C., 2006. Gender, Context, and Reading: A Comparison of Students in 43 Countries. *Scientific Studies of Reading*, 10(4), 331-362.

Christian, S.L. et al., 2008. Novel submicroscopic chromosomal abnormalities detected in autism spectrum disorder. *Biological Psychiatry*, 63(12), 1111-1117.

Clark, G.J. et al., 2009. The CD300 molecules regulate monocyte and dendritic cell functions. *Immunobiology*, 214(9-10), 730-736.

Clayton, D.G. et al., 2005. Population structure, differential bias and genomic control in a large-scale, case-control association study. *Nature Genetics*, 37(11), 1243-1246.

von Coelln, R. et al., 2004. Parkin-associated Parkinson's disease. *Cell and Tissue Research*, 318(1), 175-184.

Cohen, M. et al., 1989. Neuropathological abnormalities in developmental dyslexia.

Annals of Neurology, 25, 567-570.

Colella, S. et al., 2007. QuantiSNP: an Objective Bayes Hidden-Markov Model to detect and accurately map copy number variation using SNP genotyping data. *Nucleic Acids Research*, 35(6), 2013-2025.

Conrad, D.F. et al., 2010. Origins and functional impact of copy number variation in the human genome. *Nature*, 464(7289), 704-712.

Coon, K.D. et al., 2007. A high-density whole-genome association study reveals that APOE is the major susceptibility gene for sporadic late-onset Alzheimer's disease. *The Journal of Clinical Psychiatry*, 68(4), 613-618.

Cooper, G.M. et al., 2008. Systematic assessment of copy number variant detection via genome-wide SNP genotyping. *Nature Genetics*. Available at: internal-pdf://Cooper et al 2008-1031725866/Cooper et al 2008.pdf.

Cope, N. et al., 2005a. Strong evidence that KIAA0319 on chromosome 6p is a susceptibility gene for developmental dyslexia. *American Journal of Human Genetics*, 76, 581-591.

Cope, N.A. et al., 2005b. No support for association between Dyslexia Susceptibility 1 Candidate 1 and developmental dyslexia. *Molecular Psychiatry*, 10, 237-238.

Cordell, H.J. & Clayton, D.G., 2005. Genetic association studies. *Lancet*, 366, 1121-1131.

Corina, D.P. et al., 2001. fMRI auditory language differences between dyslexic and able reading children. *Neuroreport*, 12(6), 1195-1201.

Cornelissen, P. et al., 1995. Contrast sensitivity and coherent motion detection measured at photopic luminance levels in dyslexics and controls. *Vision Research*, 35(10), 1483-1494.

Couto, J.M. et al., 2010. Association of reading disabilities with regions marked by acetylated H3 histones in KIAA0319. *American Journal of Medical Genetics. Part B, Neuropsychiatric Genetics: The Official Publication of the International Society of Psychiatric Genetics*, 153B(2), 447-462.

Couto, J.M. et al., 2008. The KIAA0319-Like (KIAA0319L) gene on chromosome 1p34 as a candidate for reading disabilities. *Journal of Neurogenetics*, 22, 295-313.

Couto, J.M. et al., 2006. Evidence for association of the gene KIAA0319-Like (KIAA0319L) on chromosome 1p34 to reading disabilities. *American Journal of Medical Genetics Part B (Neuropsychiatric Genetics)*, 141B(7), 729.

Craddock, N. et al., 2008. Genome-wide association studies in psychiatry: lessons from early studies of non-psychiatric and psychiatric phenotypes. *Molecular Psychiatry*, 13(7), 649-653.

Culotti, J.G. & Merz, D.C., 1998. DCC and netrins. *Current Opinion in Cell Biology*, 10, 609-613.

van Daal, V. & van der Leij, A., 1999. Developmental dyslexia: Related to specific or general deficits? *Annals of Dyslexia*, 49, 71-104.

Dahdouh, F. et al., 2009. Further evidence for DYX1C1 as a susceptibility factor for dyslexia. *Psychiatric Genetics*, 19(2), 59-63.

Dahlman, I. et al., 2002. Parameters for reliable results in genetic association studies in common disease. *Nature Genetics*, 30(2), 149-150.

Defagot, M.C. et al., 1997. Distribution of D4 dopamine receptor in rat brain with sequence-specific antibodies. *Brain Research. Molecular Brain Research*, 45, 1-12.

Deffenbacher, K.E. et al., 2004. Refinement of the 6p21.3 quantitative trait locus influencing dyslexia: linkage and association analyses. *Human Genetics*, 115, 128-138.

DeFries, J.C. et al., 1997. Genetic aetiologies of reading and spelling deficits: developmental differences. In M. Snowling, ed. *Dyslexia: biology, cognition and intervention*. London: Whurr.

DeFries, J.C. et al., 1987. Evidence for a genetic aetiology in reading disability of twins. *Nature*, 329, 537-539.

Démonet, J. et al., 2004. Developmental Dyslexia. *Lancet*, 363, 1451-1460.

Denenberg, V.H. et al., 1991. Spatial learning, discrimination learning, paw preference and neocortical ectopias in two autoimmune strains of mice. *Brain Research*, 562(1), 98-104.

Dennis, M.Y. et al., 2009. A common variant associated with dyslexia reduces expression of the KIAA0319 gene. *Public Library of Science Genetics*, 5(3). Available at: internal-pdf://Dennis et al 2009-0156804654/Dennis et al 2009.pdf.

Devlin, B. & Risch, N., 1995. A comparison of linkage disequilibrium measures for fine-scale mapping. *Genomics*, 29(2), 311-322.

Dimas, A.S. et al., 2009. Common regulatory variation impacts gene expression in a cell type-dependent manner. *Science (New York, N.Y.)*, 325(5945), 1246-1250.

Diskin, S.J. et al., 2008. Adjustment of genomic waves in signal intensities from whole-genome SNP genotyping platforms. *Nucl. Acids Res.*, 36(19), e126.

Dixon, A.L. et al., 2007. A genome-wide association study of global gene expression. *Nature Genetics*, 39(10), 1202-1207.

Dobyns, W.B. & Truwit, C.L., 1995. Lissencephaly and other malformations of cortical development: 1995 update. *Neuropediatrics*, 26(3), 132-147.

Dobyns, W.B. et al., 1999. Differences in the gyral pattern distinguish chromosome 17-linked and X-linked lissencephaly. *Neurology*, 53(2), 270-277.

Docherty, S.J. et al., 2007. Applicability of DNA pools on 500 K SNP microarrays for cost-effective initial screens in genomewide association studies. *BMC Genomics*, 8, 214.

Doerge, R.W. & Churchill, G.A., 1996. Permutation tests for multiple loci affecting a quantitative character. *Genetics*, 142(1), 285-294.

Dudbridge, F., 2003. Pedigree disequilibrium tests for multilocus haplotypes. *Genetic Epidemiology*, 25, 115-121.

Dupont, W.D. & Plummer, W.D., 1990. Power and sample size calculations. A review and computer program. *Controlled Clinical Trials*, 11(2), 116-128.

Dwyer, S. et al., 2010. Mutation screening of the DTNBP1 exonic sequence in 669 schizophrenics and 710 controls using high-resolution melting analysis. *American Journal of Medical Genetics. Part B, Neuropsychiatric Genetics: The Official Publication of the International Society of Psychiatric Genetics*, 153B(3), 766-774.

Eberharter, A. & Becker, P.B., 2002. Histone acetylation: a switch between repressive and permissive chromatin. Second in review series on chromatin dynamics. *EMBO Reports*, 3(3), 224-229.

Eckert, M.A. et al., 2003. Anatomical correlates of dyslexia: frontal and cerebellar findings. *Brain: A Journal of Neurology*, 126(Pt 2), 482-494.

Elia, J. et al., 2010. Rare structural variants found in attention-deficit hyperactivity disorder are preferentially associated with neurodevelopmental genes. *Molecular Psychiatry*, 15(6), 637-646.

Elliot, C.D., 1983. *British Ability Scales*, Windsor: NFER-Nelson.

Facoetti, A. et al., 2000. The spatial distribution of visual attention in developmental dyslexia. *Experimental Brain Research. Experimentelle Hirnforschung. Expérimentation Cérébrale*, 132(4), 531-538.

Fagerheim, T. et al., 1999. A new gene (DYX3) for dyslexia is located on chromosome 2. *Journal of Medical Genetics*, 36, 664-669.

Fan, M. et al., 2005. CHIP (carboxyl terminus of Hsc70-interacting protein) promotes basal and geldanamycin-induced degradation of estrogen receptor-alpha. *Molecular Endocrinology (Baltimore, Md.)*, 19(12), 2901-2914.

Fanciulli, M. et al., 2010. Gene copy number variation and common human disease. *Clinical Genetics*, 77(3), 201-213.

Faraone, S.V. et al., 1999. Meta-analysis of the association between the 7-repeat allele of the dopamine D(4) receptor gene and attention deficit hyperactivity disorder. *American Journal of Psychiatry*, 158, 1052-1057.

Farias, E.F. et al., 2005. Cellular retinol-binding protein-I inhibits PI3K/Akt signaling through a retinoic acid receptor-dependent mechanism that regulates p85-p110 heterodimerization. *Oncogene*, 24(9), 1598-1606.

Farrag, A.F. et al., 1988. Prevalence of specific reading disability in Egypt. *Lancet*, 2(8615), 837-839.

Farrag, A.F. et al., 2002. Impaired parvocellular pathway in dyslexic children. *European Journal of Neurology: The Official Journal of the European Federation of Neurological Societies*, 9(4), 359-363.

Farrer, L.A. et al., 1995. Apolipoprotein E genotype in patients with Alzheimer's disease: implications for the risk of dementia among relatives. *Annals of Neurology*, 38(5), 797-808.

Fawcett, A.J. & Nicolson, R.I., 2001. Dyslexia: The role of the cerebellum. In A. J. Fawcett, ed. *Dyslexia: Theory and Good Practice*. London: Whurr, pp. 89-105.

Fawcett, A.J. et al., 1996. Impaired Performance of Children With Dyslexia on a Range of Cerebellar Tasks. *Annals of Dyslexia*, 46, 259-283.

Feany, M.B. & Buckley, K.M., 1993. The synaptic vesicle protein synaptotagmin promotes formation of filopodia in fibroblasts. *Nature*, 364(6437), 537-540.

Fernández-Ruiz, J. et al., 2000. The endogenous cannabinoid system and brain development. *Trends in Neurosciences*, 23, 14-20.

Feuk, L. et al., 2006. Structural variation in the human genome. *Nature Reviews Genetics*, 7, 85-97.

Field, L.L. & Kaplan, B.J., 1998. Absence of linkage of phonological coding dyslexia to chromosome 6p23-p21.3 in a large family data set. *American Journal of Human Genetics*, 63, 1448-1456.

Finucci, J.M. et al., 1976. The genetics of specific reading disability. *Annals of Human Genetics*, 40(1), 1-23.

Fisher, S.E. et al., 2002. Independent genome-wide scans identify a chromosome 18 quantitative-trait locus influencing dyslexia. *Nature Genetics*, 30, 86-91.

Fisher, S.E. et al., 1999. A quantitative-trait locus on chromosome 6p influences different aspects of developmental dyslexia. *American Journal of Human Genetics*, 64, 146-156.

Flannery, K.A. et al., 2000. Male prevalence for reading disability is found in a large sample of black and white children free from ascertainment bias. *Journal of the International Neuropsychological Society: JINS*, 6(4), 433-442.

Flax, J.F. et al., 2003. Specific language impairment in families: evidence for co-occurrence with reading impairments. *Journal of Speech, Language, and Hearing Research: JSLHR*, 46(3), 530-543.

Fox, J.W. et al., 1998. Mutations in filamin 1 prevent migration of cerebral cortical neurons in human periventricular heterotopia. *Neuron*, 21(6), 1315-1325.

Francks, C. et al., 2003. Familial and genetic effects on motor coordination, laterality, and reading-related cognition. *The American Journal of Psychiatry*, 160(11), 1970-1977.

Francks, C. et al., 2002. Fine mapping of the chromosome 2p12-16 dyslexia susceptibility locus: quantitative association analysis and positional candidate genes SEMA4F and OTX1. *Psychiatric Genetics*, 12(1), 35-41.

Francks, C. et al., 2004. A 77-kilobase region of chromosome 6p22.2 is associated with dyslexia in families from the United Kingdom and from the United States. *American Journal of Human Genetics*, 75, 1046-1058.

Frank, Y. & Pavlakis, S.G., 2001. Brain imaging in neurobehavioural disorders. *Pediatric Neurology*, 25, 278-287.

Franke, B. et al., 2006. Dyslexia susceptibility locus on chromosome 1P confirmed in Dutch sib pair collection. *American Journal of Medical Genetics Part B (Neuropsychiatric Genetics)*, 141B(7), 765.

Frayling, T.M., 2007. Genome-wide association studies provide new insights into type 2 diabetes aetiology. *Nature Reviews. Genetics*, 8(9), 657-662.

Freeman, J.L. et al., 2006. Copy number variation: new insights in genome diversity. *Genome Research*, 16(8), 949-961.

Froster, U. et al., 1993. Cosegregation of balanced translocation (1;2) with retarded speech development and dyslexia. *Lancet*, 342, 178-179.

Fugger, H.N. et al., 2000. Novel effects of estradiol and estrogen receptor alpha and beta on cognitive function. *Brain Research*, 883(2), 258-264.

Gabriel, S.B. et al., 2002. The structure of haplotype blocks in the human genome. *Science (New York, N.Y.)*, 296(5576), 2225-2229.

Galaburda, A.M. et al., 1994. Evidence for aberrant auditory anatomy in developmental dyslexia. *Proceedings of the National Academy of Sciences of the United States of America*, 91(17), 8010-8013.

Galaburda, A.M. & Kemper, T., 1979. Cytoarchitectonic abnormalities in developmental dyslexia: a case study. *Annals of Neurology*, 6, 94-100.

Galaburda, A.M. et al., 1985. Developmental dyslexia: four consecutive patients with cortical anomalies. *Annals of Neurology*, 18, 222-233.

Gayán, J. & Olson, R.K., 1999. Reading disability: evidence for a genetic aetiology. *European Child and Adolescent Psychiatry*, 8, 52-55.

Gayán, J. et al., 1999. Quantitative-trait locus for specific language and reading deficits on chromosome 6p. *American Journal of Human Genetics*, 64, 157-164.

Gayán, J. et al., 2005. Bivariate linkage scan for reading disability and attention-deficit/hyperactivity disorder. *Journal of Child Psychology and Psychiatry*, 46(10), 1045-1056.

Geiger, G. et al., 1994. Dyslexic children learn a new visual strategy for reading: a controlled experiment. *Vision Research*, 34(9), 1223-1233.

Geschwind, N. & Behan, P., 1982. Left-handedness: association with immune disease, migraine, and developmental learning disorder. *Proceedings of the National Academy of Sciences of the United States of America*, 79(16), 5097-5100.

Gilger, J.W. et al., 1994. Commingling and segregation analysis of reading performance in families of normal reading probands. *Behavior Genetics*, 24(4), 345-355.

Gilger, J.W. et al., 1996. Differential risk for developmental reading disorders in the offspring of compensated versus noncompensated parents. *Reading and Writing*, 8(5), 407-417.

Gilger, J.W. et al., 1991. Risk for reading disability as a function of parental history in three family studies. *Reading and Writing: An Interdisciplinary Journal*, 3, 205-217.

Gilger, J.W. et al., 1992. A twin study of the etiology of comorbidity: attention-deficit hyperactivity disorder and dyslexia. *Journal of the American Academy of Child and Adolescent Psychiatry*, 31(2), 343-348.

Gilger, J.W. et al., 1998. A twin and family study of the association between immune system dysfunction and dyslexia using blood serum immunoassay and survey data. *Brain and Cognition*, 36(3), 310-333.

Gleeson, J.G. et al., 1998. Doublecortin, a brain-specific gene mutated in human X-linked lissencephaly and double cortex syndrome, encodes a putative signaling protein. *Cell*, 92(1), 63-72.

Gleeson, J.G. & Walsh, C.A., 1997. New genetic insights into cerebral cortical development. In A. M. Galaburda & E. Christen, eds. *Normal and Abnormal Development of Cortex.* Berlin: Springer-Verlag, pp. 145-163.

Gleeson, J. et al., 1999. Doublecortin is a microtubule-associated protein and is expressed widely by migrating neurons. *Neuron*, 23, 257-271.

Glessner, J.T. et al., 2009. Autism genome-wide copy number variation reveals ubiquitin and neuronal genes. *Nature*, 459(7246), 569-573.

Graham, R. et al., 2005. Distinguishing different DNA heterozygotes by high-resolution melting. *Clinical Chemistry*, 51(7), 1295-1298.

Grant, A.C. et al., 1999. Tactile perception in developmental dyslexia: a psychophysical study using gratings. *Neuropsychologia*, 37(10), 1201-1211.

Grigorenko, E.L. et al., 2003. Continuing the search for dyslexia genes on 6p. *American Journal of Medical Genetics Part B (Neuropsychiatric Genetics)*, 118B, 89-98.

Grigorenko, E.L. et al., 1997. Susceptibility loci for distinct components of developmental dyslexia on chromosomes 6 and 15. *American Journal of Human Genetics*, 60, 27-39.

Grigorenko, E.L. et al., 2000. Chromosome 6p influences on different dyslexia-related cognitive processes: further confirmation. *American Journal of Human Genetics*, 66, 715-723.

Grigorenko, E.L. et al., 2001. Linkage studies suggest a possible locus for developmental dyslexia on chromosome 1p. *American Journal of Medical Genetics Part B (Neuropsychiatric Genetics)*, 105B, 120-129.

Gronlund, J. et al., 2000. Cloning of a novel scavenger receptor cysteine-rich type I transmembrane molecule (M160) expressed by human macrophages. *Journal of Immunology (Baltimore, Md.: 1950)*, 165(11), 6406-6415.

Gross-Tsur, V. et al., 1996. Developmental dyscalculia: prevalence and demographic features. *Developmental Medicine and Child Neurology*, 38(1), 25-33.

Gu, W. et al., 2008. Mechanisms for human genomic rearrangements. *PathoGenetics*, 1(1), 4.

Hallgren, B., 1950. Specific dyslexia (congenital word-blindness); a clinical and genetic study. *Acta Psychiatrica Et Neurologica. Supplementum*, 65, 1-287.

Hannula-Jouppi, K. et al., 2005. The axon guidance receptor gene ROBO1 is a candidate gene for developmental dyslexia. *PLoS Genetics*, 1(4), e50.

Hari, R. et al., 2001. Left minineglect in dyslexic adults. *Brain: A Journal of Neurology*, 124(Pt 7), 1373-1380.

Harold, D. et al., 2009. Genome-wide association study identifies variants at CLU and PICALM associated with Alzheimer's disease. *Nature Genetics*, 41(10), 1088-1093.

Harold, D. et al., 2006. Further evidence that the KIAA0319 gene confers susceptibility to developmental dyslexia. *Molecular Psychiatry*, 11(12), 1085-1091.

Harris, M.A. et al., 2004. The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Research*, 32(Database issue), D258-261.

Haslum, M.N. & Miles, T.R., 2007. Motor performance and dyslexia in a national cohort of 10-year-old children. *Dyslexia (Chichester, England)*, 13(4), 257-275.

Hastings, P.J. et al., 2009. Mechanisms of change in gene copy number. *Nature Reviews. Genetics*, 10(8), 551-564.

Hatakeyama, S. et al., 2004. Interaction of U-box-type ubiquitin-protein ligases (E3s) with molecular chaperones. *Genes To Cells*, 9, 533-548.

Hatten, M.E., 1999. Central nervous system neuronal migration. *Annual Review of Neuroscience*, 22, 511-539.

Hattersley, A.T. & McCarthy, M.I., 2005. What makes a good genetic association study? *Lancet*, 366, 1315-1323.

Headley, P.M. & Grillner, S., 1990. Excitatory amino acids and synaptic transmission: the evidence for a physiological function. *Trends in Pharmacological Sciences*, 11(5), 205-211.

Heath, S.M. et al., 1999. Auditory temporal processing in disabled readers with and without oral language delay. *Journal of Child Psychology and Psychiatry, and Allied Disciplines*, 40(4), 637-647.

Heintzman, N.D. et al., 2007. Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. *Nature Genetics*, 39(3), 311-318.

Hennings, J.M. et al., 2009. Clinical characteristics and treatment outcome in a representative sample of depressed inpatients - findings from the Munich Antidepressant Response Signature (MARS) project. *Journal of Psychiatric Research*, 43(3), 215-229.

Hermann, K., 1956. Congenital word-blindness; poor readers in the light of Gerstmann's syndrome. *Acta Psychiatrica Et Neurologica Scandinavica. Supplementum*, 108, 177-184.

Hilgert, N. et al., 2008. Mutation analysis of TMC1 identifies four new mutations and suggests an additional deafness gene at loci DFNA36 and DFNB7/11. *Clinical Genetics*, 74(3), 223-232.

Hill, N.I. et al., 1999. Frequency acuity and binaural masking release in dyslexic listeners. *The Journal of the Acoustical Society of America*, 106(6), L53-58.

Hillenbrand, R. et al., 1999. The close homologue of the neural adhesion molecule L1

(CHL1): patterns of expression and promotion of neurite outgrowth by heterophilic interactions. *The European Journal of Neuroscience*, 11(3), 813-826.

Hinshelwood, J., 1907. Four cases of congenital word-blindness occuring in the same family. *British Medical Journal*, 1, 608-609.

Hirschhorn, J.N. & Daly, M.J., 2005. Genome-wide association studies for common diseases and complex traits. *Nature Reviews Genetics*, 6, 95-108.

Hohnen, B. & Stevenson, J., 1999. The structure of genetic influences on general cognitive, language, phonological and reading abilities. *Developmental Psychology*, 35(2), 590-603.

Holm, J. et al., 1996. Structural features of a close homologue of L1 (CHL1) in the mouse: a new member of the L1 family of neural recognition molecules. *The European Journal of Neuroscience*, 8(8), 1613-1629.

Holmans, P. et al., 2009. Gene ontology analysis of GWA study data sets provides insights into the biology of bipolar disorder. *American Journal of Human Genetics*, 85(1), 13-24.

Hoogendoorn, B. et al., 2000. Cheap, accurate and rapid allele frequency estimation of single nucleotide polymorphisms by primer extension and DHPLC in DNA pools. *Human Genetics*, 107, 488-493.

Hsiung, G.R. et al., 2004. A dyslexia susceptibility locus (DYX7) linked to Dopamine D4 Receptor (DRD4) region on chromosome 11p15.5. *American Journal of Medical Genetics Part B (Neuropsychiatric Genetics)*, 125B, 112-119.

Humphreys, P. et al., 1990. Developmental dyslexia in women: Neuropathological findings in three cases. *Annals of Neurology*, 28, 727-738.

Hyde, L.A. et al., 2001. Effects of ectopias and their cortical location on several measures of learning in BXSB mice. *Developmental Psychobiology*, 39(4), 286-300.

Ibraghimov-Beskrovnaya, O. et al., 2000. Strong homophilic interactions of the Ig-like domains of polycystin-1, the protein product of an autosomal dominant polycystic kidney disease gene, PKD1. *Human Molecular Genetics*, 9, 1641-1649.

Igo Jr, R. et al., 2006. Genomewide scan for real-word reading subphenotypes of dyslexia: Novel chromosome 13 locus and genetic complexity. *American Journal of Medical Genetics Part B (Neuropsychiatric Genetics)*, 141B, 15-27.

Imai, Y. et al., 2002. CHIP is associated with Parkin, a gene responsible for familial Parkinson's disease, and enhances its ubiquitin ligase activity. *Molecular Cell*, 10(1), 55-67.

Inoue, K. & Lupski, J.R., 2002. Molecular mechanisms for genomic disorders. *Annual Review of Genomics and Human Genetics*, 3, 199-242.

International Schizophrenia Consortium, 2008. Rare chromosomal deletions and duplications increase risk of schizophrenia. *Nature*, 455(7210), 237-241.

Ioannidis, J.P.A., 2008. Why most discovered true associations are inflated. *Epidemiology (Cambridge, Mass.)*, 19(5), 640-648.

Ioannidis, J.P.A. et al., 2003. Genetic associations in large versus small studies: an empirical assessment. *Lancet*, 361(9357), 567-571.

Itsara, A. et al., 2009. Population analysis of large copy number variants and hotspots of human genetic disease. *The American Journal of Human Genetics*, 84, 148-161.

James, W.H., 1992. The sex-ratio of dyslexic children and their sibs. *Developmental Medicine and Child Neurology*, 34, 530-533.

Johannes, S. et al., 1996. Developmental dyslexia: passive visual stimulation provides no evidence for a magnocellular processing defect. *Neuropsychologia*, 34(11), 1123-1127.

Johnson, A.D. et al., 2008. SNAP: a web-based tool for identification and annotation of proxy SNPs using HapMap. *Bioinformatics*, 24(24), 2938 -2939.

Jorgensen, T.J. et al., 2009. Hypothesis-driven candidate gene association studies: practical design and analytical considerations. *American Journal of Epidemiology*, 170(8), 986-993.

Kadesjö, B. & Gillberg, C., 1999. Developmental coordination disorder in Swedish 7-year-old children. *Journal of the American Academy of Child and Adolescent Psychiatry*, 38(7), 820-828.

Kalay, E. et al., 2005. Four novel TMC1 (DFNB7/DFNB11) mutations in Turkish patients with congenital autosomal recessive nonsyndromic hearing loss. *Human Mutation*, 26(6), 591.

Kalkman, H.O., 2006. The role of the phosphatidylinositide 3-kinase-protein kinase B pathway in schizophrenia. *Pharmacology & Therapeutics*, 110(1), 117-134.

Kamei, M. et al., 1998. SOLH, a human homologue of the Drosophila melanogaster small optic lobes gene is a member of the calpain and zinc-finger gene families and maps to human chromosome 16p13.3 near CATM (cataract with microphthalmia). *Genomics*, 51(2), 197-206.

Kaminen, N. et al., 2003. A genome scan for developmental dyslexia confirms linkage to chromosome 2p11 and suggests a new locus on 7q32. *Journal of Medical Genetics*, 40, 340-345.

Kamphaus, G.D. et al., 2000. Canstatin, a novel matrix-derived inhibitor of angiogenesis and tumor growth. *The Journal of Biological Chemistry*, 275(2), 1209-1215.

Kaplan, D.E. et al., 2002. Evidence for linkage and association with reading disability, on 6p21.3-22. *American Journal of Human Genetics*, 70, 1287-1298.

Katusic, S.K. et al., 2001. Incidence of reading disability in a population-based birth cohort, 1976-1982, Rochester, Minn. *Mayo Clinic Proceedings. Mayo Clinic*, 76(11), 1081-1092.

Keller, L.C. et al., 2009. Molecular architecture of the centriole proteome: the conserved WD40 domain protein POC1 is required for centriole duplication and length control. *Molecular Biology of the Cell*, 20(4), 1150-1166.

Keogh, B.K. & Margolis, J.S., 1976. A component analysis of attentional problems of educationally handicapped boys. *Journal of Abnormal Child Psychology*, 4(4), 349-359.

Khaja, R. et al., 2006. Genome assembly comparison identifies structural variants in the human genome. *Nature Genetics*, 38(12), 1413-1418.

Kidd, T. et al., 1998. Roundabout controls axon crossing of the CNS midline and defines a novel subfamily of evolutionarily conserved guidance receptors. *Cell*, 92, 205-215.

Kirov, G. et al., 2009. A genome-wide association study in 574 schizophrenia trios using DNA pooling. *Molecular Psychiatry*, 14(8), 796-803.

Kirov, G., 2010. The role of copy number variation in schizophrenia. *Expert Review of Neurotherapeutics*, 10(1), 25-32.

Kirov, G. et al., 2006. Pooled DNA genotyping on Affymetrix SNP genotyping arrays. *BMC Genomics*, 7(27). Available at: internal-pdf://Kirov et al 2006-2682444544/Kirov et al 2006.pdf.

Kirov, G. et al., 2008. A genome-wide association study in 574 schizophrenia trios using DNA pooling. *Molecular Psychiatry*. Available at: internal-pdf://Kirov et al 2008-2464851968/Kirov et al 2008.pdf.

Kitada, T. et al., 1998. Mutations in the parkin gene cause autosomal recessive juvenile parkinsonism. *Nature*, 392(6676), 605-608.

Kitajiri, S. et al., 2007a. A novel mutation at the DFNA36 hearing loss locus reveals a critical function and potential genotype-phenotype correlation for amino acid-572 of TMC1. *Clinical Genetics*, 71(2), 148-152.

Kitajiri, S. et al., 2007b. Identities, frequencies and origins of TMC1 mutations causing DFNB7/B11 deafness in Pakistan. *Clinical Genetics*, 72(6), 546-550.

Kleinjan, D.A. & van Heyningen, V., 2005. Long-range control of gene expression:

emerging mechanisms and disruption in disease. *American Journal of Human Genetics*, 76(1), 8-32.

Knopik, V.S. et al., 1997. Comorbidity of mathematics and reading deficits: evidence for a genetic etiology. *Behavior Genetics*, 27(5), 447-453.

de Kovel, C. et al., 2004. Genomewide scan identifies susceptibility locus for dyslexia on Xq27 in an extended Dutch family. *Journal of Medical Genetics*, 41, 652-657.

Kozma, R. et al., 1995. The Ras-related protein Cdc42Hs and bradykinin promote formation of peripheral actin microspikes and filapodia in Swiss 3T3 fibroblasts. *Molecular Cell Biology*, 15, 1942-1952.

Kröger, K. et al., 2006. Prevalence of peripheral arterial disease - results of the Heinz Nixdorf recall study. *European Journal of Epidemiology*, 21(4), 279-285.

Kronbichler, M. et al., 2002. Dyslexia: verbal impairments in the absence of magnocellular impairments. *Neuroreport*, 13(5), 617-620.

Kumar, R.A. et al., 2008. Recurrent 16p11.2 microdeletions in autism. *Human Molecular Genetics*, 17(4), 628-638.

Kurdistani, S.K. et al., 2004. Mapping global histone acetylation patterns to gene expression. *Cell*, 117(6), 721-733.

Kurima, K. et al., 2002. Dominant and recessive deafness caused by mutations of a novel gene, TMC1, required for cochlear hair-cell function. *Nature Genetics*, 30(3), 277-284.

Lai, C. et al., 2001. A forkhead-domain gene is mutated in a severe speech and language disorder. *Nature*, 413, 519-523.

Lambert, J. et al., 2009. Genome-wide association study identifies variants at CLU and CR1 associated with Alzheimer's disease. *Nature Genetics*, 41(10), 1094-1099.

Lee, J.A. et al., 2007. A DNA replication mechanism for generating nonrecurrent rearrangements associated with genomic disorders. *Cell*, 131(7), 1235-1247.

Lee, S. et al., 2000. Expression and regulation of a gene encoding neural recognition molecule NB-3 of the contactin/F3 subgroup in mouse brain. *Gene*, 245(2), 253-266.

Leonard, C.M. et al., 2001. Anatomical risk factors for phonological dyslexia. *Cerebral Cortex (New York, N.Y.: 1991)*, 11(2), 148-157.

Lesch, K. et al., 2010. Genome-wide copy number variation analysis in attention-deficit/hyperactivity disorder: association with neuropeptide Y gene dosage in an extended pedigree. *Molecular Psychiatry*. Available at: http://www.ncbi.nlm.nih.gov/pubmed/20308990 [Accessed July 12, 2010].

Levitt, P. et al., 1997. New evidence for neurotransmitter influences on brain development. *Trends in Neurosciences*, 20, 269-274.

Lewis, C. et al., 1994. The prevalence of specific arithmetic difficulties and specific reading difficulties in 9- to 10-year-old boys and girls. *Journal of Child Psychology and Psychiatry, and Allied Disciplines*, 35(2), 283-292.

Li, D. & Roberts, R., 2001. WD-repeat proteins: structure characteristics, biological function, and their involvement in human diseases. *Cellular and Molecular Life Sciences: CMLS*, 58(14), 2085-2097.

Liu, F. et al., 2008. Activation of estrogen receptor-beta regulates hippocampal synaptic plasticity and improves memory. *Nature Neuroscience*, 11(3), 334-343.

Livingstone, M.S. et al., 1991. Physiological and anatomical evidence for a magnocellular defect in developmental dyslexia. *Proceedings of the National Academy of Sciences of the USA*, 88, 7943-7947.

Lohmueller, K.E. et al., 2003. Meta-analysis of genetic association studies supports a contribution of common variants to susceptibility to common disease. *Nature Genetics*, 33(2), 177-182.

Loo, S.K. et al., 2004. Genome-wide scan of reading ability in affected sibling pairs with attention-deficit/hyperactivity disorder: unique and shared genetic effects. *Molecular Psychiatry*, 9, 485-493.

Lorenzi, C. et al., 2000. Use of temporal envelope cues by children with developmental dyslexia. *Journal of Speech, Language, and Hearing Research: JSLHR*, 43(6), 1367-1379.

Lovegrove, W.J. et al., 1980. Specific reading disability: differences in contrast sensitivity as a function of spatial frequency. *Science (New York, N.Y.)*, 210(4468), 439-440.

Luciano, M. et al., 2007. A haplotype spanning KIAA0319 and TTRAP is associated with normal variation in reading and spelling ability. *Biological Psychiatry*, 62, 844-817.

Luine, V.N. et al., 1998. Estradiol enhances learning and memory in a spatial memory task and effects levels of monoaminergic neurotransmitters. *Hormones and Behavior*, 34(2), 149-162.

Luo, L., 2000. Rho GTPases in neuronal morphogenesis. *Nature Reviews Neuroscience*, 1, 173-180.

Lupski, J.R. et al., 1992. Gene dosage is a mechanism for Charcot-Marie-Tooth disease type 1A. *Nature Genetics*, 1(1), 29-33.

Macgregor, S., 2007. Most pooling variation in array-based DNA pooling is attributable to array error rather than pool construction error. *European Journal of Human Genetics: EJHG*, 15(4), 501-504.

Macgregor, S. et al., 2008. Highly cost-efficient genome-wide association studies using DNA pools and dense SNP arrays. *Nucleic Acids Research*, 36(6), e35.

Mägi, R. et al., 2007. Evaluating the performance of commercial whole-genome marker sets for capturing common genetic variation. *BMC Genomics*, 8, 159.

Malmquist, E., 1958. *Factors related to reading disabilities in the first grade of elementary school* Malmquiest & Wiksell, eds., Stockholm.

Manis, F.R. et al., 1997. Are speech perception deficits associated with developmental dyslexia? *Journal of Experimental Child Psychology*, 66(2), 211-235.

Manolio, T.A. et al., 2009. Finding the missing heritability of complex diseases. *Nature*, 461(7265), 747-753.

Manolio, T.A. et al., 2008. A HapMap harvest of insights into the genetics of common disease. *The Journal of Clinical Investigation*, 118(5), 1590-1605.

Mardis, E.R., 2008. Next-generation DNA sequencing methods. *Annual Review of Genomics and Human Genetics*, 9, 387-402.

Marín, O. & Rubenstein, J.L., 2001. A long, remarkable journey: tangential migration in the telencephalon. *Nature Reviews. Neuroscience*, 2(11), 780-790.

Marino, C. et al., 2007. Association of short-term memory with a variant within DYX1C1 in developmental dyslexia. *Genes, Brain, and Behavior*, 6(7), 640-646.

Marino, C. et al., 2005. A family-based association study does not support DYX1C1 on 15q21.3 as a candidate gene in developmental dyslexia. *European Journal of Human Genetics*, 13, 491-499.

Marino, C. et al., 2004. A locus on 15q15-15qter influences dyslexia: further support from a transmission/disequilibrium study in an Italian speaking population. *Journal of Medical Genetics*, 41, 42-46.

Marlow, A.J. et al., 2001. Investigation of quantitative measures related to reading disability in a large sample of sib-pairs from the UK. *Behavior Genetics*, 31(2), 219-230.

Marlow, A.J. et al., 2003. Use of multivariate linkage analysis for dissection of a complex cognitive trait. *American Journal of Human Genetics*, 72, 561-570.

Marshall, C.M. et al., 2001. Rapid auditory processing and phonological ability in normal readers and readers with dyslexia. *Journal of Speech, Language, and Hearing Research: JSLHR*, 44(4), 925-940.

Marshall, C.R. et al., 2008. Structural variation of chromosomes in autism spectrum disorder. *American Journal of Human Genetics*, 82(2), 477-488.

Massinen, S. et al., 2009. Functional interaction of DYX1C1 with estrogen receptors suggests involvement of hormonal pathways in dyslexia. *Human Molecular Genetics*, 18(15), 2802-2812.

McAnally, K.I. & Stein, J.F., 1996. Auditory temporal coding in dyslexia. *Proceedings. Biological Sciences / The Royal Society*, 263(1373), 961-965.

McArthur, G.M. & Hogben, J.H., 2001. Auditory backward recognition masking in children with a specific language impairment and children with a specific reading disability. *The Journal of the Acoustical Society of America*, 109(3), 1092-1100.

McArthur, G.M. et al., 2000. On the "specifics" of specific reading disability and specific language impairment. *Journal of Child Psychology and Psychiatry, and Allied Disciplines*, 41(7), 869-874.

McCarroll, S.A., 2008. Extending genome-wide association studies to copy-number variation. *Human Molecular Genetics*, 17, R135-R142.

McCarroll, S.A. et al., 2008. Integrated detection and population-genetic analysis of SNPs and copy number variation. *Nature Genetics*, 40(10), 1166-1174.

McCarthy, J.V. et al., 1998. RIP2 is a novel NF-kappaB-activating and cell death-inducing kinase. *The Journal of Biological Chemistry*, 273(27), 16968-16975.

McCarthy, M.M., 2008. Estradiol and the developing brain. *Physiological Reviews*, 88(1), 91-124.

McCarthy, M.I. et al., 2008. Genome-wide association studies for complex traits: consensus, uncertainty and challenges. *Nature Reviews. Genetics*, 9(5), 356-369.

McCrory, E. et al., 2000. Abnormal functional activation during a simple word repetition task: A PET study of adult dyslexics. *Journal of Cognitive Neuroscience*, 12(5), 753-762.

McManus, M.F. & Golden, J.A., 2005. Neuronal migration in developmental disorders. *Journal of Child Neurology*, 20(4), 280-286.

Meaburn, E.L. et al., 2008. Quantitative trait locus association scan of early reading disability and ability using pooled DNA and 100K SNP microarrays in a sample of 5760 children. *Molecular Psychiatry*, 13, 729-740.

Mefford, H.C. & Eichler, E.E., 2009. Duplication hotspots, rare genomic disorders, and common disease. *Current Opinion in Genetics and Development*, 19, 196-204.

Melquist, S. et al., 2007. Identification of a novel risk locus for progressive supranuclear palsy by a pooled genomewide scan of 500,288 single-nucleotide polymorphisms. *American Journal of Human Genetics*, 80(4), 769-778.

Meng, H. et al., 2005a. TDT-association analysis of EKN1 and dyslexia in a Colorado twin cohort. *Human Genetics*, 118, 87-90.

Meng, H. et al., 2005b. DCDC2 is associated with reading disability and modulates neuronal development in the brain. *Proceedings of the National Academy of Sciences of the USA*, 102(47), 17053-17058.

Meyer, C.G. et al., 2005. Novel TMC1 structural and splice variants associated with congenital nonsyndromic deafness in a Sudanese pedigree. *Human Mutation*, 25(1), 100.

Miscimarra, L.E. et al., 2007. Further evidence of pleiotropy influencing speech and language: analysis of the DYX8 region. *Human Heredity*, 63, 47-58.

Mizuno, K. et al., 1993. MPTP delta, a putative murine homolog of HPTP delta, is expressed in specialized regions of the brain and in the B-cell lineage. *Molecular and Cellular Biology*, 13(9), 5513-5523.

Mody, M. et al., 1997. Speech perception deficits in poor readers: auditory processing or phonological coding? *Journal of Experimental Child Psychology*, 64(2), 199-231.

Morris, D.W. et al., 2004. Association analysis of two candidate phospholipase genes that map to the chromosome 15q15.1-15.3 region associated with reading disability. *American Journal of Medical Genetics Part B (Neuropsychiatric Genetics)*, 129B, 97-103.

Morris, D.W. et al., 2000. Family-based association mapping provides evidence for a gene for reading disability on chromosome 15q. *Human Molecular Genetics*, 9(5), 843-848.

Morrow, E.M. et al., 2008. Identifying autism loci and genes by tracing recent shared ancestry. *Science (New York, N.Y.)*, 321(5886), 218-223.

Morton, N.E., 1955. Sequential tests for the detection of linkage. *American Journal of Human Genetics*, 7(3), 277-318.

Moskvina, V. et al., 2009. Gene-wide analyses of genome-wide association data sets: evidence for multiple common risk alleles for schizophrenia and bipolar disorder and for overlap in genetic risk. *Molecular Psychiatry*, 14(3), 252-260.

Moskvina, V. et al., 2005. Streamlined analysis of pooled genotype data in SNP-based association studies. *Genetic Epidemiology*, 28, 273-282.

Mueller, J.C., 2004. Linkage disequilibrium for different scales and applications. *Briefings in Bioinformatics*, 5(4), 355 -364.

Myers, A.J. et al., 2007. A survey of genetic human cortical gene expression. *Nature*

Genetics, 39(12), 1494-1499.

Nagarajan, S. et al., 1999. Cortical auditory signal processing in poor readers. *Proceedings of the National Academy of Sciences of the United States of America*, 96(11), 6483-6488.

Neale, B.M. & Sham, P.C., 2004. The future of association studies: gene-based analysis and replication. *American Journal of Human Genetics*, 75(3), 353-362.

Neale, M.D., 1989. *Analysis of reading ability*, Windsor: NFER-Nelson.

Nicolson, R.I. et al., 1999. Association of abnormal cerebellar activation with motor learning difficulties in dyslexic adults. *Lancet*, 353(9165), 1662-1667.

Nicolson, R.I. et al., 2001. Developmental dyslexia: the cerebellar deficit hypothesis. *Trends in Neurosciences*, 24(9), 508-511.

Nicolson, R.I. et al., 1995. Time estimation deficits in developmental dyslexia: evidence of cerebellar involvement. *Proceedings. Biological Sciences / The Royal Society*, 259(1354), 43-47.

Nicolson, R.I. & Fawcett, A.J., 1990. Automaticity: a new framework for dyslexia research? *Cognition*, 35, 159-82.

Nishino, J. et al., 2004. Meteorin: a secreted protein that regulates glial cell differentiation and promotes axonal extension. *The EMBO Journal*, 23(9), 1998-2008.

Nittrouer, S., 1999. Do temporal processing deficits cause phonological processing problems? *Journal of Speech, Language, and Hearing Research: JSLHR*, 42(4), 925-942.

Nobes, C.D. & Hall, A., 1995. Rho, Rac, and Cdc42 GTPases regulate the assembly of multimolecular focal complexes associated with actin stress fibers, lamellipoida, and filapodia. *Cell*, 81, 53-62.

Nolte, I.M. et al., 2007. Association testing by haplotype-sharing methods applicable to whole-genome analysis. *BMC Proceedings*, 1 Suppl 1, S129.

Nopola-Hemmi, J. et al., 2001. A dominant gene for developmental dyslexia on chromosome 3. *Journal of Medical Genetics*, 38, 658-664.

Nopola-Hemmi, J. et al., 2000. Two translocations of chromosome 15q associated with dyslexia. *Journal of Medical Genetics*, 37, 771-775.

Norton, N. et al., 2000. Suggestive evidence of linkage with reading disability in a large Norwegian family. *American Journal Of Medical Genetics Part B (Neuropsychiatric Genetics)*, 96(4), 556.

Norton, N. et al., 2002. Universal, robust, highly quantitative SNP allele frequency measurement in DNA pools. *Human Genetics*, 110, 471-478.

Nöthen, M.M. et al., 1999. Genetic linkage analysis with dyslexia: Evidence for linkage of spelling disability to chromosome 15. *European Child and Adolescent Psychiatry*, 8(Suppl.3), III/56-III/59.

Olson, R.K. et al., 1994. Genes, Environment, and the Development of Orthographic Skills. In V. W. Berninger, ed. *The Varieties of Orthographic Knowledge. 1: Theoretical and Developmental Issues*. Dordrecht, The Netherlands: Kluwer Academic Publishers, pp. 1-31.

Ou, Z. et al., 2008. Microduplications of 22q11.2 are frequently inherited and are associated with variable phenotypes. *Genetics in Medicine: Official Journal of the American College of Medical Genetics*, 10(4), 267-277.

Paracchini, S. et al., 2007. The Genetic Lexicon of Dyslexia. *Annual Review of Genomics and Human Genetics*, 8, 57-79.

Paracchini, S. et al., 2008. Association of the KIAA0319 dyslexia susceptibility gene with reading skills in the general population. *American Journal of Psychiatry*, 165, 1576-1584.

Paracchini, S. et al., 2006. The chromosome 6p22 haplotype associated with dyslexia reduces the expression of KIAA0319, a novel gene involved in neuronal migration. *Human Molecular Genetics*, 15(10), 1659-1666.

Paulesu, E. et al., 2001. Dyslexia: cultural diversity and biological unity. *Science (New York, N.Y.)*, 291(5511), 2165-2167.

Paulesu, E. et al., 1996. Is developmental dyslexia a disconnection syndrome? Evidence from PET scanning. *Brain: A Journal of Neurology*, 119 (Pt 1), 143-157.

Pe'er, I. et al., 2008. Estimation of the multiple testing burden for genomewide association studies of nearly all common variants. *Genetic Epidemiology*, 32(4), 381-385.

Pennington, B.F., 1990. Annotation: the genetics of dyslexia. *Journal of Child Psychology and Psychiatry*, 31, 193-201.

Pennington, B.F., 1989. Using genetics to understand dyslexia. *Annals of Dyslexia*, 39, 81-93.

Pennington, B.F. et al., 1991. Evidence for major gene transmission of developmental dyslexia. *JAMA*, 266, 1527-34.

Perez-Reyes, E., 2006. Molecular characterization of T-type calcium channels. *Cell Calcium*, 40(2), 89-96.

Peschansky, V.J. et al., 2009. The effect of variation in expression of the candidate Dyslexia susceptibility gene homolog Kiaa0319 on neuronal migration and dendritic morphology in the rat. *Cerebral Cortex*. Available at: internal-pdf://Peschansky et al 2009-3322868229/Peschansky et al 2009.pdf.

Petit, M.M. et al., 2000. LPP, an actin cytoskeleton protein related to zyxin, harbors a nuclear export signal and transcriptional activation capacity. *Molecular Biology of the Cell*, 11(1), 117-129.

Petryshen, T.L. et al., 2002. Supportive evidence for the DYX3 dyslexia susceptibility gene in Canadian families. *Journal of Medical Genetics*, 39, 125-126.

Petryshen, T.L. et al., 2000. Absence of significant linkage between phonological coding dyslexia and chromosome 6p23-21.3, as determined by use of quantitative-trait methods: confirmation of qualitative analyses. *American Journal of Human Genetics*, 66, 708-714.

Petryshen, T.L. et al., 2001. Evidence for a susceptibility locus on chromosome 6q influencing phonological coding dyslexia. *American Journal of Medical Genetics Part B (Neuropsychiatric Genetics)*, 105B, 507-517.

Peyrard-Janvid, M. et al., 2004. Fine mapping of the 2p11 dyslexia locus and exclusion of TACR1 as a candidate gene. *Human Genetics*, 114, 510-516.

Pinkel, D. et al., 1998. High resolution analysis of DNA copy number variation using comparative genomic hybridization to microarrays. *Nature Genetics*, 20(2), 207-211.

Pinto, D. et al., 2010. Functional impact of global rare copy number variation in autism spectrum disorders. *Nature*, 466(7304), 368-372.

des Portes, V. et al., 1998. A novel CNS gene required for neuronal migration and involved in X-linked subcortical laminar heterotopia and lissencephaly syndrome. *Cell*, 92(1), 51-61.

Power, C. & Elliott, J., 2006. Cohort profile: 1958 British birth cohort (National Child Development Study). *International Journal of Epidemiology*, 35(1), 34-41.

Primus, R.J. et al., 1997. Localisation and characterisation of dopamine D4 binding sites in rat and human brain by use of the novel, D4 receptor-selective ligand [3H]NGD 94-1. *Journal of Pharmacology and Experimental Therapeutics*, 282, 1020-1027.

Prins, N.D. et al., 2005. Cerebral small-vessel disease and decline in information processing speed, executive function and memory. *Brain: A Journal of Neurology*, 128(Pt 9), 2034-2041.

Pugh, K.R. et al., 2000. The angular gyrus in developmental dyslexia: task-specific differences in functional connectivity within posterior cortex. *Psychological Science: A Journal of the American Psychological Society / APS*, 11(1), 51-56.

Purcell, S. et al., 2007. PLINK: A Tool Set for Whole-Genome Association and

Population-Based Linkage Analyses. *American Journal of Human Genetics*, 81(3), 559-575.

Pype, S. et al., 2000. TTRAP, a novel protein that associates with CD40, tumor necrosis factor (TNF) receptor-75 and TNF receptor-associated factors (TRAFs), and that inhibits nuclear factor kappa B activation. *Journal of Biological Chemistry*, 275, 18586-18593.

Rabin, M. et al., 1993. Suggestive linkage of developmental dyslexia to chromosome 1p34-p36. *Lancet*, 342, 178.

Rae, C. et al., 1998. Metabolic abnormalities in developmental dyslexia detected by 1H magnetic resonance spectroscopy. *Lancet*, 351(9119), 1849-1852.

Rakic, P., 1982. Early developmental events: cell lineages, acquisition of neuronal positions, and areal and laminar development. *Neurosciences Research Program Bulletin*, 20(4), 439-451.

Ramus, F. et al., 2003a. The relationship between motor control and phonology in dyslexic children. *Journal of Child Psychology and Psychiatry, and Allied Disciplines*, 44(5), 712-722.

Ramus, F. et al., 2003b. Theories of developmental dyslexia: insights from a multiple case study of dyslexic adults. *Brain*, 126, 841-865.

Raskind, W.H. et al., 2005. A genome scan in multigenerational families with dyslexia: identification of a novel locus on chromosome 2q that contributes to phonological decoding efficiency. *Molecular Psychiatry*, 10, 699-711.

Redon, R. et al., 2006. Global variation in copy number in the human genome. *Nature*, 444(7118), 444-454.

Reed, G.H. & Wittwer, C.T., 2004. Sensitivity and specificity of single-nucleotide polymorphism scanning by high-resolution melting analysis. *Clinical Chemistry*, 50(10), 1748-1754.

Reed, M.A., 1989. Speech perception and the discrimination of brief auditory cues in reading disabled children. *Journal of Experimental Child Psychology*, 48(2), 270-292.

Reich, D.E. & Lander, E.S., 2001. On the allelic spectrum of human disease. *Trends in Genetics: TIG*, 17(9), 502-510.

Reiner, O. et al., 1993. Isolation of a Miller-Dieker lissencephaly gene containing G protein beta-subunit-like repeats. *Nature*, 364(6439), 717-721.

Richardson, A.J. & Ross, M.A., 2000. Fatty acid metabolism in neurodevelopment disorder: a new perspective on associations between attention-deficit/hyperactivity disorder, dyslexia, dyspraxia and the autistic spectrum. *Prostoglandins, Leukotrienes and Essential Fatty Acids*, 63, 1-9.

Richlan, F. et al., 2009. Functional abnormalities in the dyslexic brain: a quantitative meta-analysis of neuroimaging studies. *Human Brain Mapping*, 30(10), 3299-3308.

Ririe, K.M. et al., 1997. Product differentiation by analysis of DNA melting curves during the polymerase chain reaction. *Analytical Biochemistry*, 245(2), 154-160.

Risch, N. & Merikangas, K., 1996. The future of genetic studies of complex human diseases. *Science (New York, N.Y.)*, 273(5281), 1516-1517.

Risch, N.J., 2000. Searching for genetic determinants in the new millennium. *Nature*, 405, 847-856.

Rissman, E.F. et al., 2002. Disruption of estrogen receptor beta gene impairs spatial learning in female mice. *Proceedings of the National Academy of Sciences of the United States of America*, 99(6), 3996-4001.

Roh, T. et al., 2005. Active chromatin domains are defined by acetylation islands revealed by genome-wide mapping. *Genes & Development*, 19(5), 542-552.

Roh, T. et al., 2007. Genome-wide prediction of conserved and nonconserved enhancers by histone acetylation patterns. *Genome Research*, 17(1), 74-81.

Rosen, G.D. et al., 1995. Behavioral consequences of neonatal injury of the neocortex. *Brain Research*, 681(1-2), 177-189.

Rosen, G.D. et al., 2007. Disruption of neuronal migration by RNAi of Dyx1c1 results in neocortical and hippocampal malformations. *Cerebral Cortex*, 17(11), 2562-2572.

Rosen, S. & Manganari, E., 2001. Is there a relationship between speech and nonspeech auditory processing in children with dyslexia? *Journal of Speech, Language, and Hearing Research: JSLHR*, 44(4), 720-736.

Rozen, S. & Skaletsky, H., 2000. Primer3 on the WWW for general users and for biologist programmers. In S. Krawetz & S. Misener, eds. *Bioinformatics Methods and Protocols: Methods in Molecular Biology.* Totowa, NJ: Humana Press, pp. 365-386.

Rutter, M. & Yule, W., 1975. The concept of specific reading retardation. *Journal of Child Psychology and Psychiatry, and Allied Disciplines*, 16(3), 181-197.

Ruusala, A. & Aspenström, P., 2004. Isolation and characterisation of DOCK8, a member of the DOCK180-related regulators of cell morphology. *FEBS Letters*, 572(1-3), 159-166.

Salisbury, J.L. et al., 2002. Centrin-2 is required for centriole duplication in mammalian cells. *Current Biology: CB*, 12(15), 1287-1292.

Salyakina, D. et al., 2005. Evaluation of Nyholt's procedure for multiple testing correction. *Human Heredity*, 60(1), 19-25; discussion 61-62.

Santos, R.L.P. et al., 2005. Novel sequence variants in the TMC1 gene in Pakistani families with autosomal recessive hearing impairment. *Human Mutation*, 26(4), 396.

Saunders, A.M. et al., 1993. Apolipoprotein E epsilon 4 allele distributions in late-onset Alzheimer's disease and in other amyloid-forming diseases. *Lancet*, 342(8873), 710-711.

Scerri, T.S. et al., 2004. Putative functional alleles of DYX1C1 are not associated with dyslexia susceptibility in a large sample of sibling pairs from the UK. *Journal of Medical Genetics*, 41, 853-857.

Schlaug, G. et al., 1995. In vivo evidence of structural brain asymmetry in musicians. *Science (New York, N.Y.)*, 267(5198), 699-701.

Schrott, L.M. et al., 1992. Environmental enrichment, neocortical ectopias, and behavior in the autoimmune NZB mouse. *Brain Research. Developmental Brain Research*, 67(1), 85-93.

Schulte-Körne, G. et al., 1998a. Auditory processing and dyslexia: evidence for a specific speech processing deficit. *Neuroreport*, 9(2), 337-340.

Schulte-Körne, G. et al., 1998b. Role of auditory temporal processing for reading and spelling disability. *Perceptual and Motor Skills*, 86(3 Pt 1), 1043-1047.

Schulte-Körne, G. et al., 2001. [Diagnosis of reading and spelling disorder]. *Zeitschrift Für Kinder- Und Jugendpsychiatrie Und Psychotherapie*, 29(2), 113-116.

Schulte-Körne, G. et al., 1998. Evidence for linkage of spelling disability to chromosome 15. *American Journal of Human Genetics*, 63, 279-282.

Schumacher, J. et al., 2006a. Strong genetic evidence of DCDC2 as a susceptibility gene for dyslexia. *American Journal of Human Genetics*, 78, 52-62.

Schumacher, J. et al., 2006b. Linkage analyses of chromosomal region 18p11-q12 in dyslexia. *Journal of Neural Transmission*, 113, 417-423.

Schumacher, J. et al., 2008. Further evidence for a susceptibility locus contributing to reading disability on chromosome 15q15-q21. *Psychiatric Genetics*, 18, 137-142.

Sebat, J. et al., 2007. Strong associations of de novo copy number mutations in Autism. *Science*, 316, 445-449.

Seeger, M. et al., 1993. Mutations affecting growth cone guidance in Drosophila: Genes necessary for guidance toward or away from the midline. *Neuron*, 10, 409-426.

Semiclaes, W. et al., 2001. Perceptual discrimination of speech sounds in developmental dyslexia. *Journal of Speech, Language, and Hearing Research: JSLHR*, 44(2), 384-399.

Seshadri, S. et al., 2010. Genome-wide analysis of genetic loci associated with Alzheimer disease. *JAMA: The Journal of the American Medical Association*, 303(18), 1832-1840.

Sham, P.C. & Curtis, D., 1995. Monte Carlo tests for associations between disease and alleles at highly polymorphic loci. *Annals of Human Genetics*, 59, 97-105.

Sham, P. et al., 2002. DNA pooling: a tool for large-scale association studies. *Nature Reviews Genetics*, 3, 862-871.

Sham, P. & McGuffin, P., 2002. Linkage and association. In P. McGuffin et al., eds. *Psychiatric Genetics & Genomics*. New York: Oxford University Press.

Share, D.L. et al., 2002. Temporal processing and reading disability. *Reading and Writing: An Interdisciplinary Journal*, 15, 151-178.

Shaywitz, B.A. et al., 1995. Defining and classifying learning disabilities and attention-deficit/hyperactivity disorder. *Journal of Child Neurology*, 10 Suppl 1, S50-57.

Shaywitz, B.A. et al., 2002. Disruption of posterior brain systems for reading in children with developmental dyslexia. *Biological Psychiatry*, 52(2), 101-110.

Shaywitz, S.E. et al., 1992. Evidence that dyslexia may represent the lower tail of a normal distribution of reading ability. *The New England Journal of Medicine*, 326(3), 145-150.

Shaywitz, S.E. et al., 1990. Prevalence of reading disability in boys and girls. Results of the Connecticut Longitudinal Study. *JAMA: The Journal of the American Medical Association*, 264(8), 998-1002.

Shaywitz, S.E. et al., 1998. Functional disruption in the organization of the brain for reading in dyslexia. *Proceedings of the National Academy of Sciences of the United States of America*, 95(5), 2636-2641.

Shifman, S. et al., 2008. A whole genome association study of neuroticism using DNA pooling. *Molecular Psychiatry*, 13(3), 302-312.

Shiratsuchi, T. et al., 1998. Cloning and characterization of BAP3 (BAI-associated protein 3), a C2 domain-containing protein that interacts with BAI1. *Biochemical and Biophysical Research Communications*, 251(1), 158-165.

Shu, K. et al., 2006. Characterization of the human PAP1 gene and its homologue possible involvement in mouse embryonic development. *Colloids and Surfaces. B, Biointerfaces*, 52(1), 22-30.

Simos, P.G. et al., 2000a. Cerebral mechanisms involved in word reading in dyslexic children: a magnetic source imaging approach. *Cerebral Cortex (New York, N.Y.: 1991)*, 10(8), 809-816.

Simos, P.G. et al., 2000b. Brain mechanisms for reading: the role of the superior temporal gyrus in word and pseudoword naming. *Neuroreport*, 11(11), 2443-2447.

Skottun, B.C., 2000. The magnocellular deficit theory of dyslexia: the evidence from contrast sensitivity. *Vision Research*, 40(1), 111-127.

Smith, S.D. et al., 1991. Screening for multiple genes influencing dyslexia. *Reading and Writing: An Interdisciplinary Journal*, 3, 285-298.

Smith, S.D. et al., 1983. Specific reading disability: Identification of an inherited form through linkage analysis. *Science*, 219, 1345-1347.

Smith, S.D. et al., 2005. Linkage of speech sound disorder to reading disability loci. *Journal of Child Psychology and Psychiatry*, 46(10), 1057-1066.

Smith, T.F., 2008. Diversity of WD-repeat proteins. *Sub-Cellular Biochemistry*, 48, 20-30.

Snowling, M.J., 2000. *Dyslexia* 2nd ed., Oxford: Blackwell.

Snowling, M.J., 1981. Phonemic deficits in developmental dyslexia. *Psychological Research*, 43(2), 219-234.

Snowling, M.J., 1995. Phonological processing and developmental dyslexia. *Journal of Research in Reading*, 18, 132-138.

Stang, A. et al., 2005. Baseline recruitment and analyses of nonresponse of the Heinz Nixdorf Recall Study: identifiability of phone numbers as the major determinant of response. *European Journal of Epidemiology*, 20(6), 489-496.

Stankiewicz, P. & Lupski, J.R., 2002. Genome architecture, rearrangements and genomic disorders. *Trends in Genetics: TIG*, 18(2), 74-82.

Stankiewicz, P. & Lupski, J.R., 2010. Structural variation in the human genome and its role in disease. *Annual Review of Medicine*, 61, 437-455.

Steer, S. et al., 2007. Genomic DNA pooling for whole-genome association scans in complex disease: empirical demonstration of efficacy in rheumatoid arthritis. *Genes and Immunity*, 8(1), 57-68.

Stefansson, H. et al., 2008. Large recurrent microdeletions associated with schizophrenia. *Nature*, 455(7210), 232-236.

Stein, C.M. et al., 2006. Speech sound disorder influenced by a locus in a 15q14 region. *Behavior Genetics*, 36, 858-868.

Stein, C.M. et al., 2004. Pleiotropic effects of a chromosome 3 locus on speech-sound disorder and reading. *American Journal of Human Genetics*, 74, 283-297.

Stein, J. & Walsh, V., 1997. To see but not to read; the magnocellular theory of dyslexia. *Trends in Neurosciences*, 20(4), 147-152.

Steinman, S.B. et al., 1998. Vision and attention. II: Is visual attention a mechanism through which a deficient magnocellular pathway might cause reading disability? *Optometry and Vision Science: Official Publication of the American Academy of Optometry*, 75(9), 674-681.

Stephenson, S., 1907. Six cases of congenital word-blindness affecting three generations of one family. *Opthalmoscope*, 5, 482-484.

Stevenson, J., 1991. Which aspects of processing texts mediate genetic effects? *Reading and Writing*, 3, 249-269.

Stevenson, J. et al., 1987. A twin study of genetic influences on reading and spelling ability and disability. *Journal of Child Psychology and Psychiatry*, 28, 229-247.

Stokowski, R.P. et al., 2007. A genomewide association study of skin pigmentation in a South Asian population. *American Journal of Human Genetics*, 81(6), 1119-1132.

Stoodley, C.J. et al., 2000. Selective deficits of vibrotactile sensitivity in dyslexic readers. *Neuroscience Letters*, 295(1-2), 13-16.

Stoodley, C.J. et al., 2005. Impaired balancing ability in dyslexic children. *Experimental Brain Research*, 167, 370-380.

Tagnaouti, N. et al., 2007. Neuronal expression of muskelin in the rodent central nervous system. *BMC Neuroscience*, 8, 28.

Taipale, M. et al., 2003. A candidate gene for developmental dyslexia encodes a nuclear tetratricopeptide repeat domain protein dynamically regulated in brain. *PNAS*, 100(20), 11553-11558.

Takeda, Y. et al., 2003. Impaired motor coordination in mice lacking neural recognition molecule NB-3 of the contactin/F3 subgroup. *Journal of Neurobiology*, 56(3), 252-265.

Tallal, P., 1980. Auditory temporal perception, phonics, and reading disabilities in children. *Brain and Language*, 9(2), 182-198.

Tallal, P. et al., 1993. Neurobiological basis of speech: a case for the preeminence of temporal processing. *Annals of the New York Academy of Sciences*, 682, 27-47.

Tan, N.C.K. et al., 2006. Genetic dissection of the common epilepsies. *Current Opinion in Neurology*, 19(2), 157-163.

Tateishi, Y. et al., 2004. Ligand-dependent switching of ubiquitin-proteasome pathways for estrogen receptor. *The EMBO Journal*, 23(24), 4813-4823.

Tateishi, Y. et al., 2006. Turning off estrogen receptor beta-mediated transcription requires estrogen-dependent receptor proteolysis. *Molecular and Cellular Biology*, 26(21), 7966-7976.

Teare, M.D. & Barrett, J.H., 2005. Genetic linkage studies. *Lancet*, 366, 1036-1044.

Teasdale, R.D. et al., 2001. A large family of endosome-localized proteins related to sorting nexin 1. *The Biochemical Journal*, 358(Pt 1), 7-16.

Temple, E. et al., 2001. Disrupted neural responses to phonological and orthographic processing in dyslexic children: an fMRI study. *Neuroreport*, 12(2), 299-307.

The International HapMap Consortium, 2007. A second generation human haplotype map of over 3.1 million SNPs. *Nature*, 449(7164), 851-861.

Thomas, C., 1905. Congenital ''word-blindness'' and its treatment. *Opthalmoscope*, 3, 380-385.

Threlkeld, S.W. et al., 2007. Developmental disruptions and behavioural impairments in rats following in utero RNAi of Dyx1c1. *Brain Research Bulletin*, 71, 508-514.

Tlili, A. et al., 2008. TMC1 but not TMC2 is responsible for autosomal recessive nonsyndromic hearing impairment in Tunisian families. *Audiology & Neuro-Otology*, 13(4), 213-218.

Toyoda, R. et al., 2005. Identification of Protogenin, a novel immunoglobulin superfamily gene expressed during early chick embryogenesis. *Gene Expression Patterns*, 5, 778-785.

Turic, D. et al., 2003. Linkage disequilibrium mapping provides evidence for a gene for reading disability on chromosome 6p21.3-22. *Molecular Psychiatry*, 8, 176-185.

Turner, D.J. et al., 2008. Germline rates of de novo meiotic deletions and duplications causing several genomic disorders. *Nature Genetics*, 40(1), 90-95.

Tzenova, J. et al., 2004. Confirmation of a dyslexia susceptibility locus on chromosome 1p24-p36 in a set of 100 Canadian families. *American Journal Of Medical Genetics Part B (Neuropsychiatric Genetics)*, 127B, 117-124.

Uetani, N. et al., 2000. Impaired learning with enhanced hippocampal long-term potentiation in PTPdelta-deficient mice. *The EMBO Journal*, 19(12), 2775-2785.

Uetani, N. et al., 2006. Mammalian motoneuron axon targeting requires receptor protein tyrosine phosphatases sigma and delta. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience*, 26(22), 5872-5880.

Van Gorp, H. et al., 2010. Scavenger receptor CD163, a Jack-of-all-trades and potential target for cell-directed therapy. *Molecular Immunology*, 47(7-8), 1650-1660.

Van Ingelghem, M. et al., 2001. Psychophysical evidence for a general temporal processing deficit in children with dyslexia. *Neuroreport*, 12(16), 3603-3607.

Vargha-Khadem, F. et al., 1995. Praxic and nonverbal cognitive deficits in a large family with a genetically transmitted speech and language disorder. *Proceedings of the National Academy of Sciences of the United States of America*, 92, 930-933.

Vesque, C. et al., 2006. Cloning of vertebrate Protogenin (Prtg) and comparative expression analysis during axis elongation. *Developmental Dynamics*, 235, 2836-2844.

Via, M. et al., 2010. The 1000 Genomes Project: new opportunities for research and social challenges. *Genome Medicine*, 2(1), 3.

Victor, J.D. et al., 1993. Visual evoked potentials in dyslexics and normals: failure to find a difference in transient or steady-state responses. *Visual Neuroscience*, 10(5), 939-946.

Visscher, P.M. & Montgomery, G.W., 2009. Genome-wide association studies and human disease: from trickle to flood. *JAMA: The Journal of the American Medical Association*, 302(18), 2028-2029.

Vogler, G.P. et al., 1985. Family history as an indicator of risk for reading disability. *Journal of Learning Disabilities*, 18(7), 419-421.

Volonghi, I. et al., 2010. Role of COL4A1 in basement-membrane integrity and cerebral small-vessel disease. The COL4A1 stroke syndrome. *Current Medicinal Chemistry*, 17(13), 1317-1324.

Vreugde, S. et al., 2002. Beethoven, a mouse model for dominant, progressive hearing loss DFNA36. *Nature Genetics*, 30(3), 257-258.

de Vries, B.B.A. et al., 2005. Diagnostic genome profiling in mental retardation. *American Journal of Human Genetics*, 77(4), 606-616.

Wacholder, S. et al., 2004. Assessing the probability that a positive report is false: an approach for molecular epidemiology studies. *Journal of the National Cancer Institute*, 96(6), 434-442.

Wadsworth, S.J. et al., 1992. Gender ratios among reading-disabled children and their siblings as a function of parental impairment. *Journal of Child Psychology and Psychiatry, and Allied Disciplines*, 33(7), 1229-1239.

Wadsworth, S.J. et al., 2000. Differential genetic etiology of reading disability as a function of IQ. *Journal of Learning Disabilities*, 33(2), 192-199.

Wain, L.V. et al., 2009. Genomic copy number variation, human health, and disease.

Lancet, 374(9686), 340-350.

Wang, K. et al., 2007. PennCNV: An integrated Markov model designed for high-resolution copy number variation detection in whole-genome SNP genotyping data. *Genome Research*, 17, 1665-1674.

Wang, L. et al., 2001. Morphological abnormalities in the brains of estrogen receptor beta knockout mice. *Proceedings of the National Academy of Sciences of the United States of America*, 98(5), 2792-2796.

Wang, L. et al., 2003. Estrogen receptor (ER)beta knockout mice reveal a role for ERbeta in migration of cortical neurons in the developing brain. *Proceedings of the National Academy of Sciences of the United States of America*, 100(2), 703-708.

Wang, W.Y.S. et al., 2005. Genome-wide association studies: theoretical and practical concerns. *Nature Reviews. Genetics*, 6(2), 109-118.

Wang, Y. et al., 2006. DYX1C1 functions in neuronal migration in developing neocortex. *Neuroscience*, 143, 515-522.

Weiss, L.A. et al., 2008. Association between microdeletion and microduplication at 16p11.2 and autism. *The New England Journal of Medicine*, 358(7), 667-675.

Weisz, A. et al., 1992. Human interferon consensus sequence binding protein is a negative regulator of enhancer elements common to interferon-inducible genes. *The Journal of Biological Chemistry*, 267(35), 25589-25596.

Weshler, D., 1992. WISC III UK. In *The Psychological Corporation Ltd.* London: Harcourt Brace and Co.

Wigg, K. et al., 2004. Support for EKN1 as the susceptibility locus for dyslexia on 15q21. *Molecular Psychiatry*, 9(12), 1111-1121.

Wigg, K. et al., 2008. Association of ADHD and the Protogenin gene in the chromosome 15q21.3 reading disabilities region. *Genes, Brain and Behavior*, 7, 877-886.

Wigginton, J.E. et al., 2005. A note on exact tests of Hardy-Weinberg equilibrium. *American Journal of Human Genetics*, 76(5), 887-893.

Wijsman, E.M. et al., 2000. Segregation analysis of phenotypic components of learning disabilities. I. Nonword memory and digit span. *American Journal of Human Genetics*, 67(3), 631-646.

Wilcke, A. et al., 2009. The role of gene DCDC2 in German dyslexics. *Annals of Dyslexia*, 59, 1-11.

Willcutt, E.G. & Pennington, B.F., 2000. Comorbidity of reading disability and attention-deficit/hyperactivity disorder: differences by gender and subtype.

Journal of Learning Disabilities, 33(2), 179-191.

Willcutt, E.G. et al., 2000. Twin study of the etiology of comorbidity between reading disability and attention-deficit/hyperactivity disorder. *American Journal of Medical Genetics*, 96(3), 293-301.

Willcutt, E. et al., 2007. Understanding comorbidity: a twin study of reading disability and attention-deficit/hyperactivity disorder. *American Journal of Medical Genetics Part B (Neuropsychiatric Genetics)*, 144B, 709-714.

Willcutt, E.G. et al., 2002. Quantitive trait locus for reading disability on chromosome 6p is pleiotropic for attention-deficit/hyperactivity disorder. *American Journal of Medical Genetics Part B (Neuropsychiatric Genetics)*, 114B, 260-268.

Williams, J. & O'Donovan, M.C., 2006. The genetics of developmental dyslexia. *European Journal of Human Genetics*, 14, 681-689.

Williams, N.M. et al., 2006. Chromosome 22 deletion syndrome and schizophrenia. *International Review of Neurobiology*, 73, 1-27.

Wimmer, H. et al., 1998. Poor reading: a deficit in skill-automatization or a phonological deficit? *Scientific Studies of Reading*, 2, 321-340.

Witton, C. et al., 1998. Sensitivity to dynamic auditory and visual stimuli predicts nonword reading ability in both dyslexic and normal readers. *Current Biology: CB*, 8(14), 791-797.

Wittwer, C.T. et al., 2003. High-resolution genotyping by amplicon melting analysis using LCGreen. *Clinical Chemistry*, 49(6 Pt 1), 853-860.

Wolf, M. & Bowers, P.G., 1999. The double-deficit hypothesis for the developmental dyslexias. *Journal of Educational Pyshology*, 91(3), 415-438.

Worby, C.A. & Dixon, J.E., 2002. Sorting out the cellular functions of sorting nexins. *Nature Reviews. Molecular Cell Biology*, 3(12), 919-931.

World Health Organisation, 2003. *The International Classification of Diseases, Vol. 10: Classification of Mental and Behavioural Disorders*, Geneva, Switzerland.

WTCCC, 2007. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature*, 447, 661-684.

WTCCC, 2010. Genome-wide association study of CNVs in 16,000 cases of eight common diseases and 3,000 shared controls. *Nature*, 464(7289), 713-720.

Wu, C. et al., 2009. BioGPS: an extensible and customizable portal for querying and organizing gene annotation resources. *Genome Biology*, 10(11), R130.

Xu, B. et al., 2008. Strong association of de novo copy number mutations with sporadic schizophrenia. *Nature Genetics*, 40(7), 880-885.

Yap, R.L. & van der Leij, A., 1994. Testing the automatization deficit hypothesis of dyslexia via a dual-task paradigm. *Journal of Learning Disabilities*, 27(10), 660-665.

Ziegler, A. et al., 2005. Developmental dyslexia--recurrence risk estimates from a german bi-center study using the single proband sib pair design. *Human Heredity*, 59(3), 136-143.

Zondervan, K.T. & Cardon, L.R., 2004. The complex interplay among factors that influence allelic association. *Nat Rev Genet*, 5(2), 89-100.

Zwaenepoel, I. et al., 2002. Otoancorin, an inner ear protein restricted to the interface between the apical surface of sensory epithelia and their overlying acellular gels, is defective in autosomal recessive deafness DFNB22. *Proceedings of the National Academy of Sciences of the United States of America*, 99(9), 6240-6245.