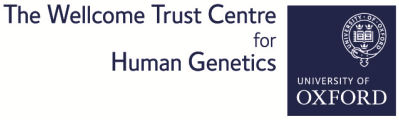
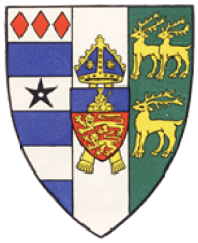




Lincoln College
Oxford

Molecular Genetics of Language Impairment

Ron Nudel

Supervisors: Prof. Anthony P. Monaco and Dr. Dianne F. Newbury

Molecular Genetics of Language Impairment

Ron Nudel

Thesis submitted for the degree of

Doctor of Philosophy

Supervisors: Prof. Anthony P. Monaco and Dr. Dianne F. Newbury

Ron Nudel

Lincoln College, University of Oxford

Molecular Genetics of Language Impairment

Abstract

Developmental language impairments are neurodevelopmental disorders in which the acquisition of language, a task which children typically perform with ease, is hindered or fraught with difficulty. This work focuses on specific language impairment (SLI), a common and highly heritable language impairment in which language development is abnormal while other developmental domains are normal. Additionally, a case-study of a child with a broader linguistic and behavioural phenotype is also presented. The work described in this thesis includes both genetic and functional investigations which were aimed at identifying candidate genes for language impairment and provide insight into the genetic mechanisms that underlie language development.

I performed a genome-wide association study of SLI which included child genotype effects, maternal genotype effects, parent-of-origin effects, and maternal-foetal interaction effects. This study found significant paternal parent-of-origin effects with the gene *NOP9* on chromosome 14, and suggestive maternal parent-of-origin effects with a region on chromosome 5 which had previously been implicated in autism and ADHD. Case-control and quantitative association analyses of HLA genes and SLI identified several risk alleles and protective alleles. A case-control association analysis for related individuals which used an isolated population affected by SLI identified a non-synonymous coding variant in the gene *NFXL1* which was significantly more frequent in affected individuals than in unaffected individuals. High-throughput sequencing of the coding regions of *NFXL1* and LD blocks surrounding associated variants in *ATP2C2*, *CMIP* and *CNTNAP2* (as reported in previous studies) identified novel or rare non-synonymous coding variants in *NFXL1* and *ATP2C2* in SLI families as well as intronic variants in all four genes that were significantly more frequent in SLI probands than in population controls.

I describe a functional study of *NFXL1* examining its expression in various brain regions, the presence of different splice variants across several tissues, its effect on genes it potentially interacts with, and the subcellular localisation of the protein.

Finally, I present the case-study of a child with language impairment who had chromosomal rearrangements which spanned the location of *FOXP2*. I examine the potential influence the chromosomal rearrangements had on *FOXP2* expression and describe a lincRNA gene which was disrupted by the chromosomal inversion.

In conclusion, this work identified new candidate genes for language impairment, provided further support for the involvement of previously-identified candidate genes in SLI and contributed to the understanding of the molecular function of a newly-identified candidate gene for SLI.

Submitted for the degree of Doctor of Philosophy, Trinity 2014

Acknowledgements

Firstly I would like to thank my supervisors Anthony (Tony) Monaco and Dianne Newbury. I thank Tony for giving me the opportunity to join his research group, for his invaluable advice and guidance and for always being approachable. I thank Dianne for guiding my first steps in research, for teaching me how to work in a lab and for encouraging me to explore my own research directions.

I would also like to thank the present and past members of the Monaco and Newbury groups; it was a great pleasure working with and getting to know such a diverse group of people. In particular, I would like to thank Antonio Velayos-Baeza for instructing me in some of the methods I used in the functional study described in this thesis.

Lastly I would like to thank my parents and sister, to whom I dedicate this thesis, for their continued and unconditional love and support. I remember visiting my dad's lab often when I was younger, and I believe that his work sparked my interest in biology. My mum, who studied English and French at university, often talked to me about languages, a subject which I found very interesting. This work is a testament to both of those influences.

Table of Contents

List of Tables

List of Figures

Abbreviations

A – adenine
ADHD – attention deficit hyperactivity disorder
ARHGEF6 – Rac/Cdc42 guanine nucleotide exchange factor 6
ARHGEF19 – Rho guanine nucleotide exchange factor 19
ASD – autism spectrum disorder
ATP2C2 – ATPase, Ca++ transporting, type 2C, member 2
BAM – binary SAM
BCA – bicinchoninic acid
BCF – binary variant call format
bp – base-pair
C – cytosine
CDNA – complementary DNA
CELF – Clinical Evaluation of Language Fundamentals
ChIP-Seq – chromatin immunoprecipitation sequencing
cM – centimorgan
CMIP – c-Maf inducing protein
CNTNAP2 – contactin associated protein-like 2
CNV – copy-number variation
CT – cycle threshold
CTOPP – Comprehensive Test of Phonological Processing subtest
DAB2 – disabled homolog 2, mitogen-responsive phosphoprotein
DAPI – 4',6-diamidino-2-phenylindole
DNA – deoxyribonucleic acid
dNTP – deoxyribonucleotide
DVD – developmental verbal dyspraxia
DZ – dizygotic
EDTA - ethylenediaminetetraacetic acid
ELS – expressive language score
EQTL – expression QTL
EVS – Exome Variant Server
FDR – false discovery rate
FISH – fluorescence *in situ* hybridization
fMRI – functional magnetic resonance imaging
FOXO3 – forkhead box O3
FOXP2 – forkhead box P2
G – guanine
GAPDH – glyceraldehyde-3-phosphate dehydrogenase
GDNA – genomic DNA
GFTA – Goldman Fristoe Test of Articulation
GORT – Gray Oral Reading Test
GUSB – glucuronidase, beta

GWAS – genome-wide association study
HE – Haseman-Elston (linkage analysis method)
HEK – human embryonic kidney (cells)
HKG – housekeeping gene
HLA – human leukocyte antigen
HLA-A – major histocompatibility complex, class I, A
HLA-B – major histocompatibility complex, class I, B
HLA-C – major histocompatibility complex, class I, C
HLA-DQA1 – major histocompatibility complex, class II, DQ alpha 1
HLA-DQB1 – major histocompatibility complex, class II, DQ beta 1
HLA-DRA – major histocompatibility complex, class II, DR alpha
HLA-DRB1 – major histocompatibility complex, class II, DR beta 1
HLOD – heterogeneity LOD
HWE – Hardy-Weinberg equilibrium
IBD – identity/identical by descent
IFN-γ – interferon gamma
IPO8 – importin 8
IQ – intelligence quotient
IVT – in-vitro transcription
kbp – kilo base-pair
kDa – kilo dalton
LB – lysogeny broth
LD – linkage disequilibrium
LINC01162 – long intergenic non-coding RNA 1162
lincRNA – long intergenic non-coding RNA
LOD – logarithm of odds
LRRC16A – leucine rich repeat containing 16A
MACROD2 – MACRO domain containing 2
MAF – minor-allele frequency
Mb – mega base-pair
MDFIC – MyoD family inhibitor domain containing
MFG – maternal-foetal genotype
MLU – mean length of utterance
MQLS – more powerful/modified quasi-likelihood score
MRI – magnetic resonance imaging
MRPS28 – mitochondrial ribosomal protein S28
MZ – monozygotic
NFX1 – nuclear transcription factor, X-box binding 1
NFXL1 – nuclear transcription factor, X-box binding-like 1
NF-κB – nuclear factor-kappaB
NOP9 – nucleolar protein 9
NWR – non-word repetition (in the text, may also refer to non-word repetition score)
OMIM – online Mendelian inheritance in man
PACSIN1 – protein kinase C and casein kinase substrate in neurons 1
PAX6 – paired box 6
PBS – phosphate buffered saline

PCR – polymerase chain reaction
PDE1C – phosphodiesterase 1C
PET – positron emission tomography
PIQ – performance IQ
POBI – People of the British Isles study
PPIA – peptidylprolyl isomerase A (cyclophilin A)
PPL – posterior probability of linkage
PPVT – Peabody Picture Vocabulary Test
PT – past tense marking task
PTGER4 – prostaglandin E receptor 4
PVDF – polyvinylidene difluoride
Q-Q – quantile-quantile
qPCR – quantitative PCR
QTDT – quantitative transmission/disequilibrium test
QTL – quantitative trait locus
RBFOX3 – RNA Binding Protein, Fox-1 Homolog (C. Elegans) 3
RLS – receptive language score
RNA – ribonucleic acid
RPM – rounds per minute
RR – relative risk
RT-PCR – reverse-transcription polymerase chain reaction
SAM – sequence alignment/map
SD – standard deviation
SDS – sodium dodecyl sulfate
SLI – specific language impairment
SLIC – SLI Consortium
SNP – single-nucleotide polymorphism
SOC – super optimal broth with catabolite repression
SP4 – Sp4 transcription factor
SP8 – Sp8 transcription factor
STAT – signal transducer and activator of transcription
T – thymine
TBE – Tris-borate EDTA
TBS – Tris-buffered saline
TDT – transmission/disequilibrium test
TEGI – Test of Early Grammatical Impairment
TM4SF20 – transmembrane 4 L six family member 20
TNS – tense marking
TOLD – Test of Language Development
TRAF7 – TNF receptor-associated factor 7
UTR – untranslated region
VC – variance-components (linkage analysis method)
VCF – variant call format
WB – western blot
WGEF – weakly similar to Rho GEF
WISC – Wechsler Scales of Intelligence

WRMT – Woodcock Reading Mastery Tests
WTCHG – Wellcome Trust Centre for Human Genetics
ZNF277 – zinc finger protein 277
ZNF423 – zinc finger protein 423

Chapter 1: Introduction

Language is one of the most unique human (and uniquely human) traits. Despite the complexity of human language, children are able to acquire it with ease. Some individuals, however, manifest problems in their linguistic capabilities, leading to language impairment. Such impairments may be developmental, i.e. presumably when one or more of the cognitive mechanisms involved in language acquisition do not develop normally, or acquired, e.g. following brain injury. Developmental language impairments could arise in the context of hearing impairment, intellectual disability, and/or other neurological or cognitive deficits. If language is taken to include speech as well, then problems in motor development may also cause and/or be involved in such disorders. Specific language impairment (SLI) is the term used to denote a developmental language impairment in which the child has language skills which are below age expectations while other cognitive and sensory-motor skills are within age expectations. Most of the work described in this thesis is focused on the genetics of SLI, with one chapter (Chapter 8) discussing a case with a broader phenotype. The current chapter introduces some of the methodology used for identifying genes involved in disease and discusses studies of the genetics of a particular speech and language disorder (in which a gene involved in speech and language was identified for the first time) and the genetics of SLI and related disorders.

Studying the genetics of language impairment may provide answers to several questions about the nature of both typical and atypical language development. Firstly, in the context of the nature of the impairment itself, elucidating the genetic underpinnings of the impairment may provide insight into the biological mechanisms that, when impaired, result in atypical language development. In other words, it could implicate biological pathways that might not have previously been thought to be involved in language (or provide further support to biological pathways that had been implicated in language development through other means).

Secondly, when genetic variants conferring susceptibility to language impairment are identified, they can be studied in the context of language abilities in the general population.

Such studies would be useful for comparing language development trajectories between individuals with typical language development and those with atypical language development.

Lastly, genetic studies of language impairment are important for unravelling the aetiology of the impairment in that they provide a framework for the identification of causal genetic variants, which could be followed up on with functional work, and, in turn, those could be examined in the context of other neurodevelopmental disorders.

1.1: Identifying Genes Involved in Disease

Two of the main methods of identifying genes involved in diseases, disorders, or various traits are linkage analysis and association analysis. It is usually the case that evidence for the genetic component of the disease in question should be obtained before one starts searching for the genes involved, either through the examination of large pedigrees (where the disease appears to be monogenic and has a clear pattern of inheritance) or by performing studies such as familial aggregation studies and/or twin studies (discussed later in this chapter in the context of SLI). Both linkage analysis and association analysis methods use genetic markers in the identification of genes involved in disease. Linkage analysis methods aim to identify regions in the genome which contain genes involved in disease by testing for co-segregation of marker alleles and disease phenotypes within pedigrees. The power to detect linkage typically depends upon a cohort of families which have sufficient recombination events e.g. one large multigenerational family or several smaller families. Parametric linkage methods are powerful methods that may be used when the disease model is known, which includes: the pattern of inheritance, the penetrance of the disease (the probability of expressing the disease phenotype given the risk genotype), and the disease and marker allele frequencies in the population (Lander and Schork, 1994, Dawn Teare and Barrett, 2005). When one or more of these parameters are unknown, or when the disease is complex, i.e. when several genes and possibly environmental factors, are involved, and the pattern of inheritance is not Mendelian, non-parametric methods may be used (Lander and Schork, 1994, Dawn Teare and Barrett, 2005). Non-parametric linkage methods test for identity by descent (IBD, the property of two alleles being identical copies of the same ancestral allele) sharing at marker loci in affected relatives. It is expected that markers

physically close to genes involved in the disease in question would also show increased IBD sharing (Lander and Schork, 1994, Dawn Teare and Barrett, 2005). Results of linkage analyses typically consist of logarithm of odds (LOD) scores. The LOD score is a function of the recombination frequency (which depends, among other things, on the distance of the investigated locus from the marker) and is defined as a logarithm (base 10) of the ratio between the probability of obtaining the observed data assuming linkage and the probability of obtaining the observed data assuming no linkage (Morton, 1955). The LOD function can be maximised over the recombination frequency to produce a maximum LOD score for each tested marker. A LOD score of 3 used to be considered significant evidence for linkage (Chotai, 1984), but a higher score (3.3) is needed to show significant linkage at a genome-wide significance level of 0.05, based on calculations for the human genome, and, depending on the study design, higher LOD scores may be needed as evidence for genome-wide linkage (Lander and Kruglyak, 1995). Association analysis methods test for a statistical correlation between a genetic variant (which can be a marker allele) and a disease phenotype in a sample which contains affected and unaffected individuals (in the context of an association study they are usually called cases and controls, respectively). The cohort of a case-control association study typically consists of a large number of unrelated cases and controls (Lander and Schork, 1994). Nonetheless, some association analysis methods are based on models that inherently use family-based data, and, in the instance of case-control association methods that use family-based data, the test statistics can be corrected for relatedness, and, depending on the test, “internal controls” may be used e.g. alleles that were not transmitted to a case (compared with alleles that were transmitted to a case). The genetic variants implicated in association studies may themselves be involved in the disease or may be in linkage disequilibrium (LD) with the disease-causing variant (Lander and Schork, 1994). Genome-wide association studies (GWAS) may be suitable for the study of complex phenotypes (McCarthy *et al.*, 2008). Results of association studies typically consist of p-values, and, when plotted, are given in the form of $-\log_{10}(P)$. The accepted significance threshold in GWAS is $P \leq 5 \times 10^{-8}$ (Risch and Merikangas, 1996), but this threshold was calculated assuming 10^6 independent tests, which is not necessarily adequate for every study design (Hoggart *et al.*, 2008). An alternative system for classifying gene-disease associations

has been proposed, in which a first-class association is one with $P<2\times10^{-7}$ (Manly, 2005). This system is intended for case-control as well as family-based association studies and takes into account the reproducibility of the association (Manly, 2005). Others have adopted a Bayesian approach and suggested a threshold of $P\leq5\times10^{-5}$, which is based on a ratio of 20:1 between true positives and false positives (Colhoun *et al.*, 2003). The threshold used by the Wellcome Trust Case Control Consortium is $P\leq5\times10^{-7}$ (The Wellcome Trust Case Control Consortium, 2007). In summary, it is important to take into account the study design, including the purpose of the study, the number of markers used, and the type of analyses performed, when considering significance thresholds.

There exist linkage analysis and association analysis methods for quantitative as well as dichotomous traits, some of which will be discussed later in this chapter. Association studies are powerful in detecting genetic risk variants in complex diseases when the variants are low-risk variants (provided adequate sample sizes are used), as they rely on LD, as measured in the population, between the variant and the disease (whereas in linkage analyses, LD between a marker allele and a risk allele is not always present across families), but they typically require a large number of markers (Risch, 2000), which, in turn, aggravates the problem of multiple testing (Cardon and Bell, 2001). Linkage analysis and association analysis methods may be used in a complementary fashion, which helps reduce the number of tests performed. An association analysis can be performed using only the markers which are located in regions implicated in a linkage analysis (Hirschhorn and Daly, 2005). This approach is called a targeted association study. In my work I used case-control and family-based association methods (quantitative or categorical) applied to genome-wide and/or smaller sets of markers.

Recent technological advances have made it possible to sequence a person's entire genome or exome (the coding parts of the genome) relatively quickly. This results in an abundance of genetic data which may be very useful in the identification of disease-causing genes. In contrast to the use of chip-genotyped markers, which allow testing for an association between a disease and common genetic variants, genome and exome sequencing are better suited for the identification of rare variants or coding changes, but they

can also be used to produce data on genetic variants in sequenced regions, which can be used as markers. These approaches have recently been applied to the study of neurodevelopmental disorders. Part of the work described in this thesis utilised these approaches.

1.2: The First Gene Implicated in a Speech and Language Disorder

The forkhead box P2 gene (*FOXP2*) is the first gene shown to be directly involved in a speech and language disorder. The gene was identified in a study of a large multigenerational British family, the KE family. An examination of the pedigree suggested that the speech and language disorder from which the affected KE family members suffered was autosomal dominant (Hurst *et al.*, 1990). Even before the genetic screening began, there was an ongoing debate regarding the nature of the disorder in terms of the phenotypic spectra thereof. Whereas Hurst *et al.* classified the disorder as developmental verbal dyspraxia (DVD) (Hurst *et al.*, 1990), Gopnik and others argued that the disorder was a manifestation of a grammatical impairment involving the morphological component of grammar, i.e. an impairment in the linguistic component that deals with the form and structure of words (Gopnik, 1990a, Gopnik, 1990b, Gopnik and Crago, 1991). Further studies showed that the latter view is less likely; affected KE family members had, in addition to the morphological impairment, problems with processing linguistic information and severe orofacial dyspraxia (Vargha-Khadem *et al.*, 1995). Additionally, the affected KE family members, as a group, had lower scores on verbal and performance (non-verbal) intelligence quotient (IQ) tests than those of the unaffected members (Vargha-Khadem *et al.*, 1995).

In the years following those early studies, several behavioural and neuroimaging studies of the KE family were performed. A study employing positron emission tomography (PET) and magnetic resonance imaging (MRI) techniques showed differences in activation between affected KE family members and control subjects in various brain regions including cortical regions while performing linguistic tasks (Vargha-Khadem *et al.*, 1998). MRI scans showed differences in the volume of grey matter between affected and unaffected KE family members in some brain regions, including Broca's area, which has been implicated in many cases of impaired language (Hillis, 2007). One region that was implicated in both the PET and

MRI analyses was the caudate nucleus, in which affected family members had significantly smaller left and right volumes than the unaffected family members (Vargha-Khadem *et al.*, 1998). Another study identified more regions in which differences in the amount of grey matter between affected KE family members and unaffected KE family members or controls were found. These included the putamen, the frontal operculum, the medial and lateral motor cortex, the cerebellum and the planum temporale (Watkins *et al.*, 2002b). Significant differences in brain activation in Broca's area and the putamen during linguistic tasks between affected and unaffected KE family members were reported in a functional MRI (fMRI) study (Liegeois *et al.*, 2003). These studies demonstrate that the gene involved in the disorder affects cortical and motor circuits in the brain.

In terms of the behavioural (both linguistic and extra-linguistic) aspects of the disorder, affected KE family members had lower mean performance IQ (PIQ) than unaffected members (Watkins *et al.*, 2002a). The affected KE family members performed poorly on receptive grammar tests, non-word repetition (NWR) tests, tense production tests for regular and irregular verbs, and tests for orofacial praxis, compared to the unaffected KE family members. Performance on the NWR test was shown to be the best marker in discriminating affected KE family members from unaffected KE family members (Watkins *et al.*, 2002a). Performance on NWR tests was found to be a good marker for heritable language impairment in other families (Bishop *et al.*, 1996). This will be discussed later in this chapter.

The first genome-wide linkage analysis done using 27 KE family members identified a region of strong linkage on chromosomal band 7q31 (Fisher *et al.*, 1998). The 5.6 cM wide region of linkage was designated SPCH1 (OMIM#602081). A subsequent study used new polymorphic markers in a linkage analysis to better characterise the region, which contained 20 known genes (Lai *et al.*, 2000). Additionally, another subject included in this study, who had verbal dyspraxia similar to that of the affected KE family members, CS, was found to be carrying a translocation that was mapped to a clone inside SPCH1 (Lai *et al.*, 2000). Further investigation revealed that the translocation breakpoint mapped to a gene inside SPCH1, which was designated *FOXP2* (OMIM#605317) (Lai *et al.*, 2001). A mutation screen of the gene using KE family members revealed a non-synonymous coding change, a G to A

transition in exon 14, which co-segregated perfectly with the disorder in the KE family, so that every affected family member was heterozygous for the mutation, and none of the unaffected members had it (Lai *et al.*, 2001). This missense mutation results in an arginine-to-histidine substitution in the forkhead DNA-binding domain of FOXP2, at position 553 of the protein (R553H), which is important for the DNA-binding function of the protein (Lai *et al.*, 2001). Thus, the KE family mutation is predicted to disrupt the function of the protein (Lai *et al.*, 2001, Vernes *et al.*, 2006), namely, its role as a transcription factor (Lai *et al.*, 2001, Shu *et al.*, 2001). A nonsense mutation (R328X) in *FOXP2* found in a proband who had verbal dyspraxia (MacDermot *et al.*, 2005) which impaired the protein's DNA-binding and transactivation capacities (Vernes *et al.*, 2006) provided further evidence for the connection between the loss of function of the FOXP2 protein and DVD. Other cases of disruptions in *FOXP2* and their effect on speech and language have since been reported, for example, by MacDermot *et al.* (2005), Zeesman *et al.* (2006), Feuk *et al.* (2006), Tomblin *et al.* (2009), and Rice *et al.* (2012). The cases reported in those studies are not SLI cases.

FOXP2 is expressed in several brain regions, including the cerebellum, thalamus, caudate nucleus and the putamen (Lai *et al.*, 2003), and its expression was sometimes limited to specific structures in those regions, e.g. Purkinje cells in the cerebellum (Lai *et al.*, 2003). *FOXP2*, as a transcription factor, has many neuronal targets (Spiteri *et al.*, 2007, Vernes *et al.*, 2007, Vernes *et al.*, 2011). Functional studies of *FOXP2* in other species have demonstrated that it is involved in a variety of processes that are related to vocal communication and motor skills in those species, including: song learning in song-learning birds (Haesler *et al.*, 2007); mouse vocalizations (Shu *et al.*, 2005, Fujita *et al.*, 2008, Gaub *et al.*, 2010); motor-skill learning in mice (Groszer *et al.*, 2008) and auditory-motor association learning in mice (Kurt *et al.*, 2012).

1.3: Specific Language Impairment

SLI is a language impairment which manifests itself primarily in the domain of language; children with SLI have problems with the acquisition of language but show normal development in all other areas (Bishop, 2006). The estimated prevalence of SLI among kindergarten children is ~7% (Tomblin *et al.*, 1997). While there are no standard diagnosis

criteria for SLI, the common diagnostic criteria are as follows (Bishop, 2006): language abilities are significantly below the level expected from age and IQ; non-verbal IQ and non-linguistic aspects of development are within broadly normal limits; language difficulties cannot be accounted for by hearing loss, physical abnormality of the speech apparatus, or environmental deprivation, nor are they caused by brain damage. SLI is a heterogeneous and complex disorder believed to be caused by a combination of genetic factors with possible environmental effects (Newbury *et al.*, 2005).

1.3.1: A Clinical Description of SLI

While SLI is a heterogeneous disorder, broadly speaking, two main language domains may be said to be affected in children with SLI: language form and structure, which would include syntax (the component of language that deals with the construction and the processing of sentences), morphology (the component of language that deals with the structure of words) and phonology (the component of language that deals with the sound system of a language), and language function and use, which would include semantics (the component of language that deals with the meaning of words and sentences) and, mostly, pragmatics (the component of language that deals with the relation between meaning and context). Rapin (1996) divides developmental language disorders into three types, two of which could be classified as SLI in which one of the two domains mentioned above may be impaired, whereas the third type included language disorders in which phonological production was the primary deficit e.g. verbal dyspraxia. Bishop (2000) further distinguishes between pragmatic SLI and typical SLI, but argues that pragmatic SLI is not in itself a diagnostic category, and that children with pragmatic difficulties may additionally have a diagnosis of autism or SLI. This view is supported by the results of a study which examined inferential processing performance in children with typical SLI, children with pragmatic SLI and children with autism. The study found that children with pragmatic SLI did not form a coherent group in terms of their performance (Norbury and Bishop, 2002). This would suggest that the primary deficit in SLI lies with expressive and receptive language, but children with SLI may still show pragmatic difficulties. Moreover, some children may have a language impairment which lies between SLI and the language deficit found in autism (Bishop, 2000).

This may be supported by recent genetic studies which found some overlaps between genes implicated in SLI and genes implicated in autism (discussed in Section 1.8.2), as well as two of the studies described in this thesis.

Co-morbidity between SLI and other developmental disorders has been reported. Substantial co-morbidity between SLI and dyslexia was reported by McArthur *et al.* (2000) and by Pennington and Bishop (2009). Cohen *et al.* (1998) found that a high proportion of children referred to mental health centres examined in their study (who often met criteria for a diagnosis of attention deficit hyperactivity disorder) also had language impairment. Language deficits in children with autism which are sometimes similar to those observed in children with SLI have also been reported (Bishop, 2003). Speech sound disorder, a disorder involving an impairment in the child's phonology and/or articulation, which may include omissions and substitutions of speech sounds as well as speech production errors which effect the intelligibility of speech, was also shown to be commonly co-morbid with SLI (Pennington and Bishop, 2009). The similarities and differences between SLI and those disorders are discussed in Section 1.8. Interestingly, substantial co-morbidity between SLI and poor motor skill has also been reported (Hill, 2001), and language deficits were observed in children with developmental coordination disorder (Archibald and Alloway, 2008).

1.3.2: Models Explaining the Deficit in SLI

There are several hypothesised models which try to explain the deficit that results in SLI. They are based on various theoretical frameworks and often use conceptually-different tools from diverse disciplines e.g. linguistics and psychology, to explain the language impairment. I will very briefly discuss three linguistically-motivated theories of SLI and two other theories which are grounded in a more general cognitive framework.

1.3.2.1: SLI is Caused by Deficits in Grammar

Gopnik and her colleagues studied several language-impaired populations which included native English speakers, native French speakers, native Greek speakers, and native Japanese speakers. The results of these studies (reviewed in (Gopnik, 1997)) indicate that the language-impaired subjects had problems with grammatical rules and representations.

For example, many speakers across different languages had problems with tense-marking, and, since tense-marking is not done in the same way across those languages studied, the above finding supports the notion that the deficit involves the rules that govern tense-marking, rather than the inability to process the verbal endings which mark the tense (which may be short in English, but may be otherwise long or stressed in the other languages studied). Furthermore, Gopnik argues that language-impaired individuals may store inflected words e.g. walk-walked, as separate entities in their lexicon, in contrast to typical-language individuals who would store the stem e.g. walk, and apply a rule to form the past tense form.

Rice *et al.* (1995) argue that children with SLI have an extended optional infinitive stage. The optional infinitive stage describes the stage in language acquisition in which children do not always mark the tense in main clauses, despite there being evidence that shows that they can distinguish between finite verbs e.g. [(I) walked] and non-finite verbs e.g. [to walk] (Wexler, 1994). Rice *et al.* propose that children with SLI remain in the optional infinitive stage for a longer period of time or never leave it. In this account, therefore, children with SLI do intrinsically know the rules that govern tense-marking, but they do not know that they are obligatory.

First proposed by Van der Lely (1994), the Representational Deficit for Dependent Relations model states that some difficulties observed in SLI may arise from a deficit in forming or understanding syntactic elements which involve dependent structural relations in the sentence. In this model, which operates under a theoretical framework that postulates the psychological and neurological autonomies of certain grammatical domains (Van der Lely, 2003), the core deficit is with syntactic operations. It should be noted that this model is meant to explain the language impairment observed only in a specific category of language-impaired children in whom the impairment is shown to be mostly grammatical in nature. An illustration of this model can be seen in a study of reference assignment of pronouns and anaphors in the context of children with the aforementioned type of SLI, called grammatical SLI (Van der Lely and Stollwerck, 1997). The reference is the object in the world that a constituent in the sentence may point to, e.g. [the brown chair] in the sentence [the brown chair is old] points to a specific object in the world, and that is its reference. The assignment of a reference when

analysing a sentence may be dependent on syntactic principles, for example: in the sentence [John asked Dan to call him] we understand that [John] and [him] may refer to the same person, but [Dan] may not, and in the sentence [John asked Dan to call himself] we understand that [Dan] and [himself] refer to the same person, but [John] does not. This is arguably due to some grammatical principles which I will not discuss here. I use names that are associated with specific genders in these examples to make them simple, since English does not have a productive grammatical gender system for nouns. In some cases there may be non-syntactic constraints, such as the context or the lexical-semantic properties of a word, which may guide the assignment of the reference. For example, in the sentence [Mary asked John to call her] we understand that [Mary] and [her] may refer to the same person, but [John] may not, possibly due to the aforementioned syntactic principles, but, in this case, even if the syntactic component were impaired, the agreement in gender between [Mary] and [her] but not between [John] and [her] would enable the correct assignment of the reference. In the study, children were shown pictures and asked yes/no questions that contained a pronoun (e.g. him) or a reflexive (e.g. himself). Based on the child's answers to questions describing the pictures, the researchers could determine whether the child assigned the correct reference to the pronoun or the reflexive. They found that when the reference assignment could be guided by both syntactic and lexical-semantic knowledge both children with grammatical SLI and normal-language children were able to determine the co-reference successfully. However, when the reference assignment relied solely upon syntactic knowledge, only the children with typical language were consistently successful in determining the co-reference.

All of the above models present the case that some aspects of SLI may indeed be caused by or correlated with deficits in grammar.

1.3.2.2: SLI is Caused by a Deficit in Phonological Short-term Memory

In the model of working memory proposed by Baddeley and Hitch (1974), the phonological, or articulatory loop is responsible for storing and rehearsing phonological information. Gathercole and Baddeley (1990) propose that children with deficits in their phonological short-term memory (the ability to retain verbal, or speech-based information for

a short period of time), which involves the phonological loop, should have problems with language acquisition, mainly in the area of vocabulary learning, and provide experimental data that support this hypothesis. Baddeley and Wilson (1993) present the case of a child with impaired phonological short-term memory who also has deficits in language areas other than vocabulary development, such as language comprehension, as determined through an assessment of syntactic processing. The authors, however, do not claim that it can be shown that the deficit in phonological short-term memory is responsible for this. Further support to this hypothesis is given by studies that examined the heritability of non-word repetition abilities in children with language impairment (Bishop *et al.*, 1996, Bishop *et al.*, 1999). This is further discussed later in this chapter.

Adams and Gathercole (2000) showed that children with better non-word repetition skills produced speech characterised by a wider repertoire of words, longer utterances (on average), and more syntactic constructions, compared to children with low non-word repetition skills who were matched to the former group in terms of non-verbal ability. As those measures are indices of language development, the authors argue that phonological short-term memory may be an important role in language development. Importantly, while children were classified as having good or poor non-word repetition skills, they were also assessed with a test which used non-spoken recall, meaning that they did not have to make an utterance, but rather point to pictures. Here, too, the performance on the test was correlated with language development, suggesting that it is not the phonological output ability that is impaired. Furthermore, Archibald and Gathercole (2006) observed impairments in phonological short-term memory as well as working memory in children with SLI, which, they argued, likely represented parallel deficits.

1.3.2.3: SLI is Caused by a Deficit in Auditory Processing

Tallal and Piercy propose that the language impairment in children with developmental aphasia (which, based on their description of the phenotype, is akin to SLI) may arise from a deficit in the processing of rapid auditory stimuli. They show that language-impaired children are impaired (compared to controls) in perceiving the order of rapidly-presented non-verbal auditory stimuli as well as in discriminating between the sounds

presented to them (Tallal and Piercy, 1973a). A further study showed that the decreased performance of the language-impaired children is in direct correlation with the rate at which the stimuli were presented to them; when the tone duration or the intervals between tones were increased, their performance improved (Tallal and Piercy, 1973b). Since then, further experiments confirmed deficits in auditory processing in children with SLI (Neville *et al.*, 1993, Wright *et al.*, 1997). However, Neville *et al.* (1993) report a high degree of within-group heterogeneity in the processing deficits they observed in children with SLI. Moreover, as discussed by (Van der Lely, 2003), other studies have provided results that are not in line with this hypothesis, for example, it was shown in a twin study that environmental factors were more likely to influence performance on a test of auditory processing than genetic factors, in contrast to performance on a non-word repetition test, which showed high heritability (Bishop *et al.*, 1999). McArthur and Bishop (2004) examined auditory perception in individuals with SLI and in control subjects. They found that, in terms of rapid auditory processing, individuals with SLI did not perform more poorly than controls, although a subgroup of the SLI individuals were less able to discriminate between frequencies of sounds compared to controls. A second experiment, which examined auditory event-related potentials (the brain's electrical responses to stimuli), found that in general the individuals with SLI had age-inappropriate event-related potentials. The authors suggest that that could be the result of a delayed maturation of the auditory cortex.

1.3.2.4: SLI is Caused by Deficits in the Procedural Memory System in the Brain

According to the Procedural Deficit Hypothesis (Ullman and Pierpont, 2005), SLI is the result of the abnormal development of brain regions involved in procedural memory. This theory attempts to explain SLI in terms of a neurological deficit in the memory system which is responsible for the learning of motor and cognitive skills, habits, and procedures, as well as the control of those said abilities when they are already long-established. The learning procedure is gradual, but the application of learned procedures is quick and "automatic". The learning process and the memories themselves are generally not accessible consciously. The basal ganglia are of particular importance to this system. The cerebellum is also thought to play a role in the procedural memory system. Apart from the general cognitive and motor

skills, the brain regions that make up the procedural memory system are also involved in grammar and lexical retrieval. In contrast to the procedural memory system, a different system, the declarative memory system, underlies the learning, representation, and use of semantic knowledge and episodic knowledge, which might be consciously recollected (Ullman, 2001). The hippocampus plays a central role in this system.

Ullman and Pierpont argue that many, if not most, individuals with SLI have brain abnormalities affecting the regions involved in the procedural memory system. This, they argue further, could explain both linguistic and non-linguistic deficits observed in individuals with SLI. While the brain abnormalities should be strongly associated with the linguistic and non-linguistic deficits, the dysfunction of different structures should lead to different impairments, which could explain the heterogeneity observed in SLI (Ullman and Pierpont, 2005). The hypothesis is generally agnostic with regards to the declarative memory system, but the authors suggest that in some cases it may compensate for the dysfunction of the procedural memory system, if it is itself functional.

It is important to note that when exclusionary criteria are employed in the diagnosis of SLI, known non-linguistic developmental deficits may preclude a diagnosis of SLI. Therefore, the Procedural Deficit Hypothesis cannot account for all cases of SLI, which is acknowledged by the authors.

It seems that currently no one theory can account for the full range of heterogeneity of SLI. Nonetheless, it is clear that some aspects of the disorder are common across all SLI children, namely the abnormal language acquisition. As will be shown shortly, SLI clusters in families, and, therefore, studying the genetics of SLI, in addition to having its own merit (as a medical study), may also shed some light on the cognitive mechanisms involved in SLI, as it provides an opportunity to study something which is, at least at first sight, more tangible.

1.3.3: The Genetics of SLI

The heritability of SLI was determined before any molecular analyses were performed. Unlike in the case of the KE family, SLI does not have a clear pattern of inheritance, despite the fact that it runs in families (Newbury *et al.*, 2005). The evidence for a

genetic component in SLI comes from familial aggregation studies and twin studies. Familial aggregation studies compare the incidence of a disorder in relatives of probands with that in relatives of control subjects. Stromswold examined 18 familial aggregation studies for language impairment and found that the rate of affectedness among probands' siblings ranged from 15% to 49%, with a mean rate of 30.3%, and the rate of affectedness among probands' parents ranged from 19% to 71%, with a mean rate of 31.6% (Stromswold, 1998). In both cases, the rate is higher than the prevalence of the disorder in the general population. Such studies, however, do not distinguish between genetic factors and environmental factors, particularly those which are shared within families. Twin studies, in contrast, are better suited for that purpose, as they include both monozygotic, or identical, twins, and dizygotic, or non-identical twins, and, in both cases, the twins should have a shared environment. A twin study typically involves comparing the incidence of a disorder in identical co-twins of probands with that in non-identical co-twins of probands, or, in other words, comparing the monozygotic (MZ) twin concordance with the dizygotic (DZ) twin concordance. Since MZ twins are assumed to be 100% genetically similar, and DZ twins are assumed to be 50% genetically similar, if in every case both twins had a shared environment, then a significantly higher concordance for MZ twins (compared to the concordance for DZ twins) would suggest that the disorder has a genetic component. When investigating quantitative traits, twin studies may also allow quantifying the genetic contribution to the phenotypic variance of the trait being investigated (its heritability). One such method is the DeFries-Fulker analysis, which is based on the idea that if a trait is heritable, then the phenotypic scores of MZ and DZ co-twins of selected probands will regress to the population mean, to different extents (DeFries and Fulker, 1985). This method was suitable for cases in which the proband was selected because of a deviant score and was later extended to handle cases in which MZ and DZ twins had normal scores (DeFries and Fulker, 1988). Stromswold performed a meta-analysis of five twin concordance studies of spoken language disorders which yielded an overall concordance of 83.6% for MZ twins and an overall concordance of 50.2% for DZ twins (Stromswold, 2001). These studies showed that SLI had a genetic component, and genetic studies were soon carried out to try to elucidate what that might be. It should be noted that some of the twin studies included in the meta-analysis used small numbers of twin pairs, and

that studies using large twin cohorts might be limited with regards to the type and/or amount of data that can be collected. As discussed earlier, the heritability can be estimated in a twin study which uses quantitative traits. A twin study of a cohort of SLI twins which used quantitative language traits found heritability estimates which were close to 1 for several language traits, although the results were very different when the model controlled for IQ scores (Bishop *et al.*, 1995). It should be noted, however, that different studies reported different heritability estimates, and that that discrepancy may be the result of the differences in the selection of cases; studies which included children who had been referred to speech and language pathology services more frequently reported high heritability estimates (Bishop and Hayiou-Thomas, 2008). Additionally, it was shown that, in preschool children, only when taking into account speech impairment, was the language impairment significantly heritable; it is plausible that children with speech impairment are more likely to be referred to speech and language pathology services (Bishop and Hayiou-Thomas, 2008). In my analyses, the proband in each family, i.e. the child who was referred to the speech and language clinicians, was considered "affected" irrespective of their language test scores.

It is worth mentioning that while the studies in Stromswold's meta-analysis showed that SLI in general had a genetic component, NWR skills in particular were shown to be heritable and could be used as a good behavioural marker for language impairment (Bishop *et al.*, 1996). The accepted criteria for a trait to be considered an endophenotype stipulate that the proposed trait should be state-dependent (i.e. present even when the condition itself is no longer present), heritable, co-segregating in families, more frequent in unaffected relatives of affected individuals than in the general population, and associated with the disorder in the general population (Gottesman and Gould, 2003). NWR meets the above criteria and could therefore be considered an endophenotype of SLI (Bishop *et al.*, 1996, Barry *et al.*, 2007, Levy and Ebstein, 2009). NWR tests are used to measure phonological short-term memory, and it has been proposed that children with language disorders have deficits in their phonological short-term memory (Gathercole and Baddeley, 1990). Sentence repetition, which also relies on short-term memory, has also been shown to be a good marker for SLI (Conti-Ramsden *et al.*, 2001).

I will now present and discuss specific investigations into the genetics of SLI. It is important to bear in mind that the number of genome-wide analyses of SLI is very low, compared to those of other neurodevelopmental disorders.

1.4: The SLI Consortium Studies

The following section discusses several studies of SLI which used a cohort of British probands and their families, the SLI Consortium studies.

1.4.1: The Original SLI Consortium Linkage Studies

The first linkage analysis study by the SLI Consortium (SLIC) used 98 nuclear families, each containing at least one child with SLI. The tests administered to the children included the revised version of the Clinical Evaluation of Language Fundamentals (CELF-R) (Semel *et al.*, 1992), the Wechsler Scales of Intelligence (WISC) (Wechsler, 1992) and a non-word repetition test (Gathercole *et al.*, 1994). The CELF-R is a battery of tests designed to diagnose and evaluate language disorders in school-aged children. The battery consists of an expressive language scale and a receptive language scale, each having three subtests. The combination of tests used depended on the age of the subject. All probands had expressive and receptive language scores (ELS and RLS, obtained from CELF-R) of more than 1.5 standard deviations (SD) below the normative mean for their chronological age and a Perceptual Organization Index of more than 77.5 (a composite score of the non-verbal subtests of WISC) (The SLI Consortium, 2002). NWR scores (NWR will hereafter refer to performance on a non-word repetition test) were used in the genetic analysis, but not for the ascertainment of probands. The two linkage analyses employed in this study were both non-parametric quantitative linkage analysis methods: Haseman-Elston (HE) method (Haseman and Elston, 1972) and variance-components (VC) method (Amos, 1994, Pratt *et al.*, 2000). The HE method implements a regression of the squared phenotype differences on the estimated IBD sharing at marker loci for a given sib-pair. The premise for this method is that siblings who are similar with regards to the trait being studied are likely to show genetic similarity (estimated by the IBD sharing) in genomic regions in which the genetic factors influencing the trait are located. If a given marker is close to a genetic factor influencing the

phenotype, then there should be an inverse association between the sib pair phenotype difference and the IBD sharing at that marker locus. The VC method separates the trait variability between siblings into a major gene variance component, a polygenic variance component, and an environmental variance component. Family data are used in order to estimate the variance components. This method compares the likelihoods of two models: one in which there is a major gene effect and another in which there is no major gene effect. The ratio between these two likelihoods can then be used in to obtain a χ^2 test statistic in order to assess the evidence for linkage. If there is a major gene effect close to a marker locus, the former of the two models will be more likely.

Genome-wide evidence for linkage on chromosomes 16 and 19 was found: a region of ~40 cM on chromosome 16q showed linkage to NWR with a maximum LOD score of 3.55 (using the HE method), and a region of ~30 cM on chromosome 19q showed linkage to ELS with a maximum LOD score of 3.55 (HE) (The SLI Consortium, 2002). The former was designated SLI1 (OMIM#606711), and the latter was designated SLI2 (OMIM#606712). A subsequent study by the SLIC (which had a larger sample size of 184 families) replicated the linkage to NWR on chromosome 16, with a combined maximum LOD score of 7.46; the linkage on chromosome 19 was also replicated, but it was linked to different traits: in the first wave (the 2002 cohort) the trait was ELS, in the second wave (the new samples) it was NWR. The maximum LOD score on chromosome 19, taking into account both cohorts, was 1.4 with NWR (HE) and 1.02 with ELS (HE) (The SLI Consortium, 2004).

1.4.2: A Multivariate Linkage Analysis Using the SLIC Cohort

Monaco and SLIC performed a multivariate genome-wide linkage analysis for SLI using the cohort from the first two SLIC studies (Monaco and The SLI Consortium, 2007). As stated previously, ELS and RLS were obtained from the CELF-R. However, in this study, the scores in the subtests used to derive ELS and RLS were also used separately. These included: formulating sentences, recalling sentences, sentence assembly, oral directions, word classes, and semantic relationships. Literacy was tested using the Wechsler Objective Reading Dimension assignment (Rust, 1993), from which three measures were derived: single-word reading, single-word spelling, and reading comprehension, but only a subset of

the probands were tested for it. These measures, together with NWR and PIQ, were used in the multivariate analysis. The statistical analysis followed that proposed by Marlow *et al.* (Marlow *et al.*, 2003). Preliminary screening of chromosome 16 revealed that only a subset of traits showed significant linkage. When the traits were restricted, further analyses using the NWR and the oral directions (in a subset of the cohort) measures, as well as the single-word reading and the single-word spelling measures, found evidence for linkage (P=0.0089), and determined that the effect of NWR was stronger than that of the oral directions measure, but equally as strong as those of the single-word reading and the single-word spelling measures (Monaco and The SLI Consortium, 2007). Linkage to chromosome 19 was found using the following measures: NWR, formulating sentences, recalling sentences, sentence assembly, oral directions, word classes, and semantic relationships (P=0.0008). The linkage could be contributed to semantic relationships and recalling sentences (which represent both expressive and receptive language measures), and, to a lesser extent, NWR. A further analysis showed that the literacy traits did not contribute to the linkage to chromosome 19 (Monaco and The SLI Consortium, 2007). The results of this study further supported the previous findings, and, furthermore, they showed that the linkage region on chromosome 16 contributed not only to NWR, but also to reading and spelling measures. This study also identified a putative locus on chromosome 10 influenced by NWR and RLS.

1.4.3: Replication of the SLI Consortium Findings

In an attempt to replicate the SLIC findings, Falcaro *et al.* used an independent cohort of 93 nuclear families in a linkage analysis examining the previous regions of linkage on chromosomes 16 and 19 only (Falcaro *et al.*, 2008). The traits used in the study included NWR, ELS, and scores on a past tense marking task (PT) (Marchman *et al.*, 1999). The test requires children to produce the past tense forms of verbs (both regular and irregular) as responses to sentences produced by the assessor which lack the past tense form of the verb e.g. “The boy is walking. He walks every day. Yesterday, he...”. They used the HE as well as the DeFries–Fulker approach (DF) for quantitative trait loci (QTL) analysis (Fulker *et al.*, 1991). This is an extended version of the DeFries-Fulker methods described earlier, which tests for linkage using quantitative traits in a cohort of sibling pairs. This method is based on

the assumption that a sibling's phenotypic score will regress back to the population mean as a function of environmental and genetic effects (estimated by IBD sharing) shared with the proband. Weak linkage to NWR was found on chromosome 16 (maximum LOD score of 1.69, DF). Significant linkage to ELS was found on chromosome 19 (maximum LOD score of 5.8, DF). When PT was measured as a continuous trait, some linkage was found on both chromosome 16 (maximum LOD score of 1.8, DF) and chromosome 19 (maximum LOD score of 2.2, DF). When PT was represented as a categorical trait (affected or unaffected), linkage was found only on chromosome 19 (maximum LOD score of 1.66, HE) (Falcaro *et al.*, 2008). Familiar aggregation for PT was significant only when using PT as a categorical trait, and the authors suggest that PT should not be regarded as a trait measurable as a continuous dimension throughout development, but, rather, that tense-marking is a skill which is either acquired or not acquired by early school age (Falcaro *et al.*, 2008).

1.4.4: A Targeted Association Study of the Linkage Region on Chromosome 16q

The next SLIC study (Newbury *et al.*, 2009) used 211 SLIC families (the families from the previous analyses and a new subset of families) and 1906 single-nucleotide polymorphisms (SNPs) to perform a high density association screen for SLI on the linkage region on chromosome 16 that had been found in the previous study. A follow-up association analysis used 138 SNPs in regions of association from the previous association screen. Two approaches were used: a case-control association analysis and a family-based association test, the quantitative transmission/disequilibrium test (QTDT) (Abecasis *et al.*, 2000). The original transmission/disequilibrium test (TDT) was designed to measure the over-transmission of an allele from heterozygous parents to affected offspring (Spielman *et al.*, 1993). Quantitative TDTs, in general, test for changes in the phenotype value as a function of the number of copies of a given marker allele carried by a child (taking into account the expected number given the mating type, i.e. based on the parents' genotypes) (Ewens *et al.*, 2008). In the case-control association analysis, cases were defined as children with low NWR (> 2 SD below the SLIC cohort mean, n=79), and controls were defined as children with above-average NWR (> 0.5 SD above population mean, n=71), since the region showed

linkage primarily to NWR. Only one case or one control was selected from each family so that relatedness did not influence the association analysis.

The most significantly associated SNPs were located in two genes: c-Maf inducing protein (*CMIP*) (OMIM#610112) and ATPase, Ca++ transporting, type 2C, member 2 (*ATP2C2*) (OMIM#613082). The most significantly associated SNP in *CMIP* had a p-value of 5×10^{-7} (obtained in the case-control association analysis), and the one in *ATP2C2* had a p-value of 2×10^{-5} (obtained in the quantitative analysis) (Newbury *et al.*, 2009). Both of these genes are expressed in the brain. *ATP2C2* is involved in calcium ion transport to the Golgi apparatus (Missiaen *et al.*, 2007), a cell organelle involved in various macromolecule processing steps. Calcium ion homeostasis is very important in various nerve cell functions and processes (Zheng and Poo, 2007). CMIP interacts with RelA (Kamal *et al.*, 2009), a member of the nuclear factor-kappaB (NF-κB) family of molecules. The NF-κB family is involved in synaptic activity and plasticity, and alterations in the signalling pathway of NF-κB have been reported in several diseases of the nervous system, including Alzheimer's disease and Parkinson's disease (Memet, 2006). Each of the genes had an independent effect on SLI susceptibility (Newbury *et al.*, 2009).

The authors then tried to replicate their findings using samples from the Avon longitudinal study of pregnancy and childhood (ALSPAC), which is a cohort study which follows children born in the former county of Avon, England in the early 1990s. 490 cases were selected on the basis of low non-word repetition test scores, for a quantitative association analysis. For the case-control analysis, children who had ≥1 SD NWR below the general-population mean (n=112) were defined as cases, and children who had ≥1 SD NWR above the general-population mean (n=72) were defined as controls. These cut-offs are not the same as the ones used in the SLIC case-control association analysis due to differences in the distribution of the non-word repetition scores between the two cohorts. The association with *ATP2C2* was replicated with the most significantly associated SNP having a p-value of 0.0058; the association with *CMIP* was replicated with the most significantly associated SNP having a p-value of 0.0182, but the association was in the opposite direction, i.e. the SNP alleles associated with high NWR in the SLIC sample were associated with low NWR in the

replication cohort (Newbury *et al.*, 2009). This could reflect differences in the relationship between the SNPs and the causal variant in the two populations (Lin *et al.*, 2007), or it could mean that the association was a false positive. The ALSPAC cohort consisted of families from England, whereas the SLIC cohort consisted of families from England and Scotland. In the linkage analysis described in (The SLI Consortium, 2004), the locus on chromosome 19 was linked to different traits when using only English samples, only Scottish samples, or a combined English and Scottish subset of samples. This suggests that the English and Scottish samples may have different contributions to the language phenotype as assessed in the SLIC studies, but it is not clear whether those differences could account for the opposite association trends with *CMIP* SNPs, as observed in the two cohorts. The authors found no association between NWR and SNPs in *ATP2C2* or *CMIP* in the general population (n=3612), again using the ALSPAC cohort, which suggests that these genes affect NWR (and therefore, phonological short-term memory) only in language-impaired individuals (Newbury *et al.*, 2009).

1.4.5: *FOXP2* and SLI

It is interesting to note that *FOXP2* was not implicated in any of the linkage analyses described above. In fact, a quantitative association and mutation screen of *FOXP2* in 43 SLIC families did not find any evidence for a direct involvement of *FOXP2* in SLI (Newbury *et al.*, 2002). Another study reported that the KE family mutation had not been found in children with SLI (Meaburn *et al.*, 2002), and a study that used a mutation screening of the exon in which that mutation was found, in children with SLI, reported that none of the children had a mutation in that exon, and that neither linkage or association to *FOXP2* were observed (O'Brien *et al.*, 2003).

These results are not surprising, considering the phenotypic differences between SLI and the disorder from which the affected KE family members suffered (which included both linguistic and extra-linguistic features). That said, since FOXP2 is a transcription factor, it is possible that the genes involved in SLI and are FOXP2 targets, or are otherwise involved in genetic pathways shared with FOXP2 targets.

An example of a FOXP2 target that is involved in SLI is contactin associated protein-like 2 (*CNTNAP2*) (OMIM#604569), which codes for a transmembrane protein found at the nodes of Ranvier (gaps in the myelin sheaths insulating the axons of neurons) (Poliak and Peles, 2003). This protein is also involved in cortical development and neuroblast migration and laminar organization in humans (Strauss *et al.*, 2006). Enriched *CNTNAP2* expression was detected in brain regions that are important for language, and its expression patterns in the human and rat brains are remarkably different (Abrahams *et al.*, 2007). Vernes *et al.* show that FOXP2 down-regulates *CNTNAP2* (Vernes *et al.*, 2008). In the same study, an association screen found nine intronic SNPs located between exons 13 and 15 of *CNTNAP2* that were significantly associated with NWR in SLIC families (Vernes *et al.*, 2008). The genomic region in which *CNTNAP2* is located was designated SLI4 (OMIM#612514). Some of the associated SNPs in *CNTNAP2* were also found to be associated with early language development in the general population (Whitehouse *et al.*, 2011). *CNTNAP2* has been implicated in studies of other developmental disorders as well, which will be discussed later in this chapter.

1.5: The Linkage Studies by Bartlett *et al.*

The following section describes the SLI linkage studies performed by another group (Bartlett *et al*). using cohorts from Canada and the United States.

1.5.1: The First Study by Bartlett *et al.*

The first SLI linkage study by Bartlett *et al.* (Bartlett *et al.*, 2002) used extended pedigrees which consisted of 86 individuals from branches of five Canadian families of Celtic ancestry. The families were originally identified in a schizophrenia study, and seven of the individuals in the linkage study had schizophrenia and therefore were coded as having an unknown phenotype. The largest family in this study (n=34) was related to the schizophrenia families only by marriage (Bartlett *et al.*, 2002). The tests administered to the subjects included the following language tests: the age-appropriate version of the Test of Language Development (TOLD) (Adolescent Language, Language Development – Primary, or Language Development – Intermediate), which measures various components of spoken

language (Hammill *et al.*, 1987, Newcomer and Hammill, 1988, Hammill and Newcomer, 1988); the following IQ tests: the age-appropriate version of the Wechsler Intelligence Test (Wechsler Intelligence Scale For Children, Wechsler Intelligence Scale for Adults, Pre-school and Primary Scale of Intelligence, or Wechsler Abbreviated Scale of Intelligence) (Wechsler, 1974, Wechsler, 1981, Wechsler, 1989, Wechsler, 1999); a word identification test: Woodcock Reading Mastery Test (Woodcock, 1987); the age-appropriate version of the Token Test, which measures the subject's ability to perform directions increasing in complexity (DiSimoni, 1978); Oral Speech Mechanism Screening Examination (St Louis and Ruscello, 1987), and self report or parental report questionnaire to assess history of hearing difficulties. The criteria for SLI were: spoken language quotient standard score ≤85 on the age-appropriate version of TOLD; PIQ ≥80 on the age-appropriate intelligence test; PIQ ≥ spoken language quotient standard score; hearing within normal limits; no motor impairments or oral structural deviations affecting speech or non-speech movement of the articulators; no co-morbid diagnosis of autism, schizophrenia, psychoses, or neurological disorders; families were included in the study if at least two members met the criteria for SLI (Bartlett *et al.*, 2002). Subjects were classified as "language impaired" (spoken language quotient standard score ≤85), "reading impaired" (single non-word reading score obtained from the Woodcock Reading Mastery Test 1 SD below their PIQ), and/or "clinically impaired" (language impaired, reading impaired or having a history of language problems or dyslexia). The statistical analysis included parametric linkage methods under two models (dominant and recessive) for each classification of impairment. For the dominant models, penetrance for individuals with one or two copies of the susceptibility allele was set to 0.5. For the recessive models, penetrance for individuals with two copies of the susceptibility allele was set to 0.8, and penetrance for individuals with one copy of the susceptibility allele was set to 0.01. For both dominant and recessive models, the penetrance of individuals with no susceptibility alleles was set to 0.001. The disease-allele frequency was set to 0.08 for the dominant model and 0.3 for the recessive model. (Bartlett *et al.*, 2002). The authors also employed a method called posterior probability of linkage (PPL) (Vieland, 1998), which measures the probability that the genetic distance between the marker and a putative disease gene is smaller than 50 cM.

The results were as follows: significant linkage to reading impairment under the recessive model on chromosomal band 13q21 (maximum LOD score of 3.92, PPL of 0.53); suggestive linkage to language impairment under the recessive model on chromosomal band 2p22 (maximum LOD score of 2.79, PPL of 0.058); suggestive linkage to reading impairment under the dominant model on chromosomal band 17q23 (maximum LOD score of 2.19, PPL of 0.036) (Bartlett *et al.*, 2002). The linkage region on chromosome 13 was designated SLI3 (OMIM#607134).

1.5.2: The Second Study by Bartlett *et al.*

In their next study, Bartlett *et al.* included a second cohort which consisted of 22 nuclear and extended families from the United States (n=279). The criteria for SLI were as described above, for the first study. The statistical analyses included the PPL (combined over the data from both cohorts) and heterogeneity LOD (HLOD) linkage analyses (Bartlett *et al.*, 2004). The HLOD is a function of both the recombination frequency and an admixture parameter which is defined as the proportion of families in which the linkage to the trait is with the marker being tested (Smith, 1963). Thus, the HLOD analysis takes into account possible involvement of more than one locus. Bartlett *et al.* used two approaches for the calculation of the HLOD scores: HLOD-P, which is the HLOD score obtained when the two data sets had been pooled, and HLOD-S, which is the HLOD score obtained when the HLOD was calculated separately for each data set and then summed across both data sets. The same battery of tests as used in (Bartlett *et al.*, 2002) was used in this study to assess various linguistic and non-linguistic skills, and the same three affection categories were used under both a dominant model and a recessive model. This study used markers from specific regions on chromosomes 2, 7 and 13, based on their own previous results and studies of autism.

The linkage to reading impairment under the recessive model on chromosomal band 13q21 had a maximum HLOD-S score of 6.181, a maximum HLOD-P score of 6.031 and a PPL of 0.923 (Bartlett *et al.*, 2004), which further supports the evidence for linkage found in the previous study. The loci on chromosomes 2 and 7 showed weak linkage (HLOD-S scores between 1 and 2).

1.5.3: The Lack of Overlapping Linkage Regions Implicated in the SLIC and Bartlett *et al.* Studies

There is a clear lack of overlap between the linkage regions implicated in the SLIC genome-wide linkage studies (SLI1 and SLI2) and the region implicated in the Bartlett *et al.* studies (SLI3). Several factors might account for that. Firstly, whereas SLI1 and SLI2 were linked to NWR and ELS, respectively, SLI3 was linked to reading impairment. Secondly, there are several crucial differences between the SLIC and Bartlett *et al.* studies in terms of the study design, the statistical analyses used and the linguistic tests administered. The linguistic tests used in the original SLIC linkage studies were used to derive three quantitative language scores (NWR, ELS and RLS). Bartlett *et al.* used different tests to derive three categorical traits. As seen in the case of the PT trait used in (Falcaro *et al.*, 2008), even when using the same trait in two linkage analyses, the only difference being that in one analysis it is used as a categorical trait and in the other analysis it is used as a continuous trait, one may still obtain very different results, because of the way the data are used in the analyses and the difficulty in deciding on a threshold for affectedness. Since the Bartlett *et al.* studies and the SLIC studies used different categorical and quantitative methods, respectively, and different cohorts, it is not surprising that different results had been obtained in the Bartlett *et al.* and SLIC studies. Another important point is that SLIC performed non-parametric linkage analyses, whereas Bartlett *et al.* performed parametric linkage analyses. As previously discussed, parametric linkage methods can be used when the disease model is known. As SLI is a complex disorder, some assumptions need to be made before using parametric linkage approaches to identify genes involved in SLI. Bartlett *et al.* address this issue and say that "parametric analysis using a single-locus model has been shown to be an effective method for detection of linkage to oligogenic disorders" (Bartlett *et al.*, 2002). They do not, however, provide support for the underlying assumption that SLI is oligogenic (influenced by a small number of genes). Furthermore, the SLIC 2002 study shows that when using two different linkage analysis methods, even if they are both non-parametric (HE and VC), one may still obtain very different results, for example: the maximum LOD score obtained for the linkage to NWR found in a region on chromosome 16 using the HE method was 3.55, but when the VC method was used, the maximum LOD score was 2.57 (The SLI Consortium,

2002). Finally, while SLIC used nuclear families in their analyses, Bartlett *et al.* used extended pedigrees. The above differences between the studies may explain the discrepancies between the results obtained in those studies.

1.6: The Robinson Crusoe Island Study

The following section describes a study of an isolated Chilean population in which the prevalence of SLI is very high. The results of this study will then be discussed in the context of the previous studies.

1.6.1: The First Linkage Analysis Study of an Isolated Chilean Population Affected by SLI

The Robinson Crusoe Island is located 677 km west of Chile, South America and forms part of the Juan Fernández archipelago. The island has a population of around 633 (based on a Chilean census from 2002) (Villanueva *et al.*, 2008). The island was most recently colonised in the late 19th century by eight families (these eight families will be hereafter referred to as founder families), and 77% of the population have a colonising surname (Villanueva *et al.*, 2008). The prevalence of SLI among children who are related to a founder family is 35% (Villanueva *et al.*, 2008) while the prevalence of SLI in children who are not related to a founder family is 3.8% (Villanueva *et al.*, 2011), close to that reported for children in mainland Chile, which is ~4% (De Barbieri *et al.*, 1999). Additionally, 27.5% of the children who are related to a founder family have language deficits but do not meet diagnosis criteria for SLI, and 37.5% have normal language development (Villanueva *et al.*, 2008). The diagnostic criteria for SLI were as follows: all children between the ages of 3 and 8 years and 11 months of age (n=66) were subjected to a linguistic battery, which included tests of phonology (Test para Evaluar Procesos de Simplificación Fonológica (Maggiolo and Pavez, 2000) and expressive and receptive morphosyntax (Test Exploratorio de Gramática Española de A. Toronto) (Pavez, 2003). Children who performed >2 SD below that expected for their age on one or more of the tests were classified as having SLI. Exclusion criteria for SLI diagnosis included the following: non-verbal IQ (Columbia Mental Maturity Scale) below the 80th percentile, hearing disability, motor or structural abnormalities, as assessed by the Oral

Motor and Speech Examination (Villanueva, 2000). Children who had a diagnosis of autism, emotional difficulties, or neurological disorders (as assessed by medical history) were excluded. Adults and children older than 8 years and 11 months were assessed by medical history, an auditory screen, and the following tests: Barcelona test (Benito-Cuadrado *et al.*, 2002), to assess verbal fluency; The Token Test (De Renzi and Vignolo, 1962), to assess verbal comprehension; Raven's Progressive Matrices (Raven *et al.*, 2003) , to assess non-verbal intelligence. DNA samples were obtained from 125 individuals from 34 families. The majority of those 34 families themselves were related to each other and formed part of a large pedigree in which most of the affected individuals were related to a pair of founder brothers (Villanueva *et al.*, 2011). For the purpose of performing the linkage analyses, the large pedigree had to be broken into seven smaller pedigrees. The analyses used 6009 SNPs.

As most of the affected individuals were related to a pair of founder brothers, the working hypothesis was that the language impairment might be caused by a mutation in a single gene that was co-segregating with the language impairment in the pedigree, and therefore, both parametric (under dominant and recessive models) and non-parametric linkage methods were used. However, no genomic region reached genome-wide significance in the parametric analyses. Linkage regions on chromosomes 6, 7, 12, 13 and 17 were identified through non-parametric linkage analyses, with HLOD scores of 7.56, 6.73, 6.14, 4.78 and 4.49, respectively (the highest score in several analyses using different expected allele frequencies is shown for each region); the linkage region on chromosome 7 was the most consistently implicated linkage region in all non-parametric analyses with an HLOD score higher than 6 in each analysis (Villanueva *et al.*, 2011). The region spanned across 48 Mb on chromosomal band 7q. The genes *FOXP2* and *CNTNAP2* are located in that region.

1.6.2: Comparison with Previous SLI Linkage Studies

While the linkage regions identified in the SLIC and Bartlett *et al.* studies showed little overlap, the linkage regions identified in the Robinson Crusoe Island study identified regions which contained a gene identified in one of the SLIC studies (*CNTNAP2*). No linkage to SLI1 or SLI2 was found in this study. No linkage to SLI3 was found, although linkage to other regions of chromosome 13 was reported. This discrepancy could be the result of the

differences in the study design between the original SLIC studies and this study, which include: the type of families used (nuclear families or extended families); whether the analysis was quantitative or used affection statuses instead; different linguistic tests. It is also possible that different causes exist for the phenotypes in those two distinct populations. Part of the work in this thesis was aimed at finding genetic variants which increase the risk of having SLI in both the SLIC and the Robinson Crusoe Island cohorts.

1.7: The Study by Rice *et al.*:

Rice *et al.* performed linkage and association analyses for various language, speech and reading measures using a cohort of SLI families (Rice *et al.*, 2009). They were interested in regions previously implicated in dyslexia (on chromosomes 1, 3, 6 and 15) and a speech and language disorder, namely the disorder in the KE family (on chromosome 7). The authors wished to see whether those regions would show linkage in a cohort of SLI families. The cohort consisted of 322 participants, including 86 probands, 134 siblings, and 102 parents. Probands met the following criteria: PIQ >85 as assessed with either The Columbia Mental Maturity Scale (Burgemeister *et al.*, 1972), the Wechsler Intelligence Test for Children (Wechsler, 1991), or the Wechsler Intelligence Test for Adults (Wechsler, 1997) – the above exclusionary criterion applied only to probands and not to parents or siblings in this study; normal hearing acuity; no history of neurological disorders or diagnosis of autism; intelligible speech sufficient for language transcription and production of target phonemes used in word-final morphology. Probands were ascertained based on language performance >1 SD below the mean on an age-appropriate language test. Goldman Fristoe Test of Articulation (GFTA) was used to assess speech (Goldman and Fristoe, 2000). The language measures used included the following: general language skills, which were assessed with an age-appropriate test: Preschool Language Scale (Zimmerman *et al.*, 1992), the Test of Early Language Development (Hresko *et al.*, 1999), the Test of Language Development-Primary (Newcomer and Hammill, 1999); or the CELF (Semel *et al.*, 1995); vocabulary, which was assessed with the Peabody Picture Vocabulary Test (PPVT) (Dunn and Dunn, 1981, Dunn *et al.*, 1997); mean length of utterance (MLU), which was assessed according to (Leadholm and Miller, 1992); tense marking (TNS), which was assessed with the Test of Early Grammatical

Impairment (TEGI) (Rice and Wexler, 2001). The reading measures used included the following: word-level reading, which was assessed with the Woodcock Reading Mastery Tests score (WRMT) (Woodcock, 1998) and text-level reading, which was assessed with the Gray Oral Reading Test (GORT) (Wiederholt and Bryant, 1992). Non-word repetition was assessed with the Comprehensive Test of Phonological Processing subtest (CTOPP) (Wagner *et al.*, 1999). The authors also used a dichotomous affection status in some analyses, and individuals were classified as cases if they had a score ≥1 SD below the mean, for the relevant trait (Rice *et al.*, 2009).

The linkage analyses included regression methods for quantitative traits and non-parametric linkage analysis for the dichotomous traits. When regions were implicated in the linkage analyses, they were also tested using the extended DeFries-Fulker method discussed above. The authors also performed quantitative association analyses using 53 SNPs in two dyslexia genes on chromosome 6 and the *FOXP2* region on chromosome 7. The linkage results obtained (with LOD≥1) were as follows: chromosome 1: linkage to GORT (dichotomous phenotype, LOD 1.25), linkage to general language skills (quantitative phenotype, LOD 1.165); chromosome 3: linkage to PPVT (dichotomous phenotype, LOD 1.03), which was telomeric to a dyslexia candidate gene; chromosome 6: linkage to TEGI (quantitative and dichotomous phenotypes, LOD 2.145 and 1, respectively); chromosome 7: no LOD score above 1; chromosome 15: linkage to WRMT (dichotomous phenotype, LOD 1.29), linkage to GORT (quantitative phenotype, LOD 1.712). The association analyses detected weak associations ($0.0106<P<0.047$) between SNPs in a candidate gene for dyslexia on chromosome 6 and GORT, GFTA and general language skills. Two SNPs within *FOXP2* showed weak association to general language skills ($P=0.0408$) and GFTA ($P=0.0295$) (Rice *et al.*, 2009).

This study, together with the SLIC multivariate linkage analysis and the Bartlett *et al.* studies, shows that some neurodevelopmental disorders affecting language and reading abilities may involve overlapping sets of genes. This will be discussed in more detail in the following section. A recent study reported an association between a deletion in transmembrane 4 L six family member 20 (*TM4SF20*) and communication disorders

(Wiszniewski *et al.*, 2013). The region on chromosome 2q36 containing this gene was designated SLI5 (OMIM#615432). However, the deletion was more often than not associated with cerebral structural abnormalities, namely, white matter hyperintensities. Moreover, several of the individuals included in the study who had the deletion were diagnosed with epilepsy, autism and/or developmental delay, which would normally preclude a diagnosis of SLI. Therefore, I did not consider this locus an SLI locus *per se*.

1.8: SLI and Other Developmental Disorders in Which Language Is Affected

While SLI is defined as a disorder that affects primarily language, it is by no means the only disorder that affects linguistic skills. There are other disorders in which language development is delayed or communication skills are impaired. Moreover, such disorders may also have a strong genetic component, as is the case of SLI. It is therefore interesting to consider whether the phenotypic similarities could be the result of genetic similarities. The rationale for this is made even more apparent considering the high degree of co-morbidity between SLI and several other disorders in which language is affected. The following section will discuss several cases of such disorders.

1.8.1: SLI and Developmental Dyslexia

Developmental dyslexia (sometimes referred to as specific reading disability) is defined as a specific and significant impairment in reading abilities which cannot be accounted for by any kind of deficit in general intelligence, learning opportunity, general motivation or sensory acuity. Its prevalence among school-aged children is 5-10% (Habib, 2000). Several studies found a high degree of co-morbidity between SLI and developmental dyslexia: McArthur *et al.*, using a cohort of 110 children with developmental dyslexia and 102 children with SLI, found that 55% of the children with dyslexia have impaired oral language, and 51% of the children with SLI have a reading disability (McArthur *et al.*, 2000), and in a review of several studies Pennington and Bishop show that children with SLI are in a greater risk of having a reading impairment, compared to controls or the general population (Pennington and Bishop, 2009). Some researchers have suggested that SLI and

developmental dyslexia were, in fact, two developmental disorders that are on one continuum (Kamhi and Catts, 1986). However, Bishop and Snowling argue that, despite the similarities between SLI and developmental dyslexia, a distinction between the two disorders is still favourable, as there are some difficulties which are mostly present in one but not in the other; for example, while phonological processing may be impaired in both disorders, syntax is characteristically impaired in SLI, but not in dyslexia, and children with dyslexia have normal listening comprehension, unlike children with SLI (Bishop and Snowling, 2004). They argue that differential diagnosis should be based on both phonological and non-phonological skills (Bishop and Snowling, 2004). A recent study further determined that the phonological deficits in SLI and dyslexia are distinct (Nithart *et al.*, 2009).

It is clear that, while researchers may disagree on where to draw the line between SLI and developmental dyslexia (or whether such a line needs to be drawn), some similarities do exist. That is where genetic studies may contribute to understanding the hypothetically shared mechanisms that underlie these disorders. There have been several genetic studies aimed at identifying the genes involved in dyslexia. Similar to SLI, dyslexia has been shown to be heritable, and several linkage and association studies have identified loci and linkage regions that might be involved in dyslexia, namely on chromosomes 1, 2, 3, 6, 11, 15 and 18, and the X chromosome (Williams and O'Donovan, 2006, Paracchini *et al.*, 2007).

While these regions do not overlap with the major linkage regions identified by SLIC or Bartlett *et al.*, studies that looked into the involvement in dyslexia of genes implicated in SLI did show some evidence supporting the hypothesis of shared genetic pathways between the two disorders. It is also worth re-mentioning at this point that the region on chromosome 13 identified in the Bartlett *et al.* studies showed linkage to reading impairment in children with SLI.

A study of SLI and dyslexia families found an association between a SNP in *KIAA0319*, a gene involved in dyslexia (Francks *et al.*, 2004, Cope *et al.*, 2005) and RLS in the SLI families (P=0.004), association trends between a SNP in *CNTNAP2* (in the opposite direction than previously described for SLI) (P=0.0425) and a SNP in *ATP2C2* (0.0251) and reading-related measures in the dyslexia families, and associations between SNPs in

CNTNAP2 and *CMIP* and reading-related measures in SLI families (0.0002≤P≤0.0354) (Newbury *et al.*, 2011). In another study an association between one of the SNPs in *CNTNAP2* and NWR in SLI reported in (Vernes *et al.*, 2008) was replicated using a cohort of dyslexia families (Peter *et al.*, 2011). Interestingly, a recent study found an association between a SNP in *CMIP* and reading ability in a population cohort (Scerri *et al.*, 2011).

1.8.2: SLI and Autism Spectrum Disorders

Autism spectrum disorders (ASDs) are a group of developmental disorders affecting social interaction, communication, and behaviour (which tends to include repetitive activities and restricted interests) (Dover and Le Couteur, 2007). Autism is the prototypical disorder in the group (Dover and Le Couteur, 2007). Language deficits are central in ASDs, but the nature of the deficits themselves can be quite different from those in SLI; autistic children may show abnormal language use. An inappropriate use of language in context, i.e. a pragmatic deficit, is the most striking feature in autism, but not in SLI (Bishop, 2003). However, some children with autism do have problems which are usually associated with SLI, and *vice versa* (Bishop, 2003). Interestingly, autistic children who have low CELF scores, as is the case with children with SLI, also perform poorly on a non-word repetition task, again, similar to the case of children with SLI (Kjelgaard and Tager-Flusberg, 2001). In another study, however, no evidence for shared aetiology between autism and SLI was found, and the authors argue that NWR is not an endophenotype of autism, and that the pattern of errors in the non-word repetition test the children make is not the same in SLI and autism (Whitehouse *et al.*, 2007). Bishop proposed a model that accounts for both the similarities and dissimilarities between autism and SLI, which incorporates interactions between genes and can explain an overlap in the genetic aetiologies of autism and SLI without assuming that the genetic effects are additive (Bishop, 2010). In a recent study, Bartlett *et al.* (2012) examined the relationship between autism and SLI using a unique cohort of families, in which every family included at least one person affected by autism and one (other) person affected by SLI. The heritability of several language measures was estimated once using the individuals with autism and individuals with SLI, and again while excluding the individuals with autism. They found that when removing the individuals with autism, the heritability of five language measures

increased significantly: the core language score from the CELF, and scores in four subtests of the Comprehensive Assessment of Spoken Language (Carrow-Woolfolk, 1999): ambiguous sentences, meaning from context, non-literal language, and pragmatic judgments. Since the statistical model used in this study assumed only additive genetic effects, the fact that the exclusion of the individuals with autism increased the heritability of some language traits (while making sure that potential problems with the data e.g. outliers in the autism group did not contribute to this) suggests that an additive model does not fit the data for those language measures in autism and SLI; the authors argue that their results support a model in which interaction between genes may account for genetic overlaps between SLI and autism, while still allowing for phenotypic differences (Bartlett *et al.*, 2012). A subsequent genome-wide linkage study which used the same SLI/autism families identified a region on chromosomal bands 15q23-26 linked with a language impairment phenotype (defined as oral language impairment or autism); the linkage signal was greatly reduced after removing autism probands or SLI probands from the analysis, suggesting that the locus is related to both SLI and autism (Bartlett *et al.*, 2014).

Many studies of various designs have shown that autism has a complex genetic component (Abrahams and Geschwind, 2008). Numerous studies have identified linkage regions involved in autism. They include regions on chromosomes 1, 2, 3, 5, 7, 9, 11, 17; only the loci on chromosomes 7 and 17 have been replicated with genome-wide significance (Abrahams and Geschwind, 2008). Genome-wide association studies for autism have also been performed. For example, the study by Anney *et al.* (2010) found a genome-wide significant association with a SNP in the MACRO domain containing 2 gene (*MACROD2*), but it was not replicated in a later study (Curran *et al.*, 2011). Copy-number variation (CNV) is also thought to play a role in ASDs (Cook and Scherer, 2008). Many candidate genes for autism have been proposed; I will limit my discussion to several studies which identified genes and/or genomic regions that pertain to both autism and SLI, as they pertain to the potential genetic overlap between these disorders.

Several autism and SLI studies have found linkage regions which overlap. A study which examined a language deficit (based on a substantial delay in “phrase speech” i.e. the

developmental stage at which children begin to speak in phrases) in autistic children found a linkage peak on chromosomal band 13q22 (maximum HLOD score of 2.54) (Bradford *et al.*, 2001). This peak overlaps with the one found by Bartlett *et al.* (in fact, the same marker was found to have the maximum LOD or HLOD score in both studies). A genome-wide screen for autism susceptibility loci found evidence for suggestive loci on chromosome 19 in two regions: one on chromosomal band 19p13 (maximum LOD score of 2.53), and one with a peak which falls within SLI2 (maximum LOD score of 1.7) on chromosomal band 19q13 (Liu *et al.*, 2001).

In terms of candidate genes, several studies have identified the involvement of the same genes in both autism and SLI. A deletion in *CMIP*, which has been shown to be involved in SLI, was found in an autistic child (Van der Aa *et al.*, 2012). *CNTNAP2*, which has been found to be associated with NWR in SLI, was implicated in several autism studies. A linkage and association study of autism found a linkage peak on chromosomal band 7q35 (maximum LOD score of 3.4) and a significant association with a SNP in *CNTNAP2* (within the linkage region) ($P=2.4\times10^{-5}$) (Arking *et al.*, 2008). *CNTNAP2* has also been associated with an endophenotype of autism, namely, age at first word, in two independent cohorts (with empirical p-values of 0.002 and 0.028) (Alarcon *et al.*, 2008). An inversion disrupting *CNTNAP2* was detected in a child who had cognitive and social delay (but did not meet the full criteria for ASD) (Bakkaloglu *et al.*, 2008). Recently, an exome sequencing study of autism found a rare missense mutation in *CNTNAP2* which was predicted to be deleterious (O'Roak *et al.*, 2011). More recently, a homozygous microdeletion in a gene located in the autism locus AUTS1 (OMIM#209850), zinc finger protein 277 (*ZNF277*), was reported in an individual with severe language impairment (Ceroni *et al.*, 2014).

1.8.3: SLI and Attention Deficit Hyperactivity Disorder

Attention deficit hyperactivity disorder (ADHD) is a common neurodevelopmental disorder with a worldwide-pooled prevalence of 5.29% (Polanczyk *et al.*, 2007). The core deficits in ADHD include inattention, impulsivity and motor restlessness, and people who suffer from ADHD may have impaired executive functions and working memory, and learning disabilities (Seidman, 2006). Language is also often impaired in children with ADHD (Baird *et*

al., 2000). Interestingly, in terms of their language, children with ADHD were found to be pragmatically-impaired, and their impairments were comparable to those in children with Asperger syndrome, which is an ASD (Bishop and Baird, 2001). The similarities between the pragmatic components of the language impairment in ASD and ADHD have been observed in more than one study; however, the language impairment in children with ASD appears to be more severe in general (Geurts and Embrechts, 2008). Cohen *et al.* (1998) reported that 63.6% of children referred to mental health centres in their study (who often met criteria for a diagnosis of ADHD) also had language impairment. They excluded children with autism, hearing loss, neuromotor impairments, neurosensory impairments, or aphasia.

ADHD has a very strong genetic component, which is evident from several family and twin studies; one review of such studies found a mean heritability estimate of 76% (Faraone *et al.*, 2005). Linkage regions have been identified in several studies. They include regions on chromosomes 4, 5, 7, 8, 9, 10, 11, 12, 15, 16, and 17 (Mick and Faraone, 2008). Most of the linkage regions were identified using one of two available cohorts, and there was little overlap between the results of the various studies in terms of significant linkage (Mick and Faraone, 2008). Nonetheless, a pooled study that used the two cohorts mentioned above did find one linkage region which indicated a common risk locus on chromosomal band 5p13 (maximum LOD score of 3.67) (Ogdie *et al.*, 2006). This region, which had previously been identified as a significant linkage region using one of the cohorts (Ogdie *et al.*, 2003, Ogdie *et al.*, 2004), was designated ADHD4 (OMIM#608906).

Association studies for ADHD have also been performed, and several candidate genes have been identified. I will now discuss results which pertain to both ADHD and SLI. One genome-wide association study identified a SNP in *ATP2C2*, a gene implicated in SLI, which was associated with ADHD ($P=9.06\times10^{-7}$) (Lesch *et al.*, 2008). A CNV analysis of ADHD identified a hemizygous deletion in *CNTNAP2* in an individual with ADHD (Elia *et al.*, 2009). It is possible that, given the similarities in the nature of the language impairment between ADHD and ASD, the model proposed by Bishop to explain the genetic overlap between SLI and ASD, namely, the implication of *CNTNAP2* in both disorders, might account for the same overlap between SLI and ADHD.

1.8.4: SLI and Speech Sound Disorder

Speech sound disorder is defined as a significant delay in the acquisition of articulate speech sounds (Lewis *et al.*, 2006). It may involve omissions and substitutions of speech sounds as well as speech production errors. Its prevalence among 6-year-old children is ~3.8% (Shriberg *et al.*, 1999). Co-morbidity between SLI and speech sound disorder has been reported (Shriberg *et al.*, 1999, Pennington and Bishop, 2009).

The results of several familial aggregation studies and twin studies indicate that the disorder has a genetic component (Lewis *et al.*, 2006). However, genetic studies of speech sound disorder focused on regions that were implicated in reading disability, due to theoretical considerations that assumed an aetiological overlap between the two conditions. Similarly, a locus implicated in autism was also investigated in a study of speech sound disorder. Thus, loci on chromosomes 3, 6 and 15 were implicated in speech sound disorder (Stein *et al.*, 2004, Smith *et al.*, 2005, Stein *et al.*, 2006). Those loci do no overlap with known SLI loci.

1.9: Summary

SLI is a complex and heterogeneous disorder with a strong genetic component. Several genome-wide linkage screens of various designs have identified linkage regions of genome-wide significance. Targeted association studies have identified candidate genes involved in phonological short-term memory. A summary of the main results from genetic studies of SLI is given in Table 1.1. Various neurodevelopmental disorders may be co-morbid with SLI. Although the nature of the language impairment in those disorders may be different from that in SLI, there exist some genetic overlaps between SLI and those disorders, which may be accounted for by the existence of genetic pathways which are shared, to some extent, between SLI and said disorders. With this in mind, I took into account whether a gene had previously been implicated in a neurodevelopmental or neurological disorder when examining the results of the various tests used in this work.

Table 1.1: Major SLI loci.

Locus/Gene	Trait	Method	Reference
SLI1 (16q)	NWR	Linkage (quantitative)	The SLI Consortium (2002)
SLI2 (19q)	ELS	Linkage (quantitative)	The SLI Consortium (2002)
SLI3 (13q)	Reading impairment	Linkage (categorical)	Bartlett *et al.* (2002)
SLI4 (7q); *CNTNAP2*	NWR	Association (quantitative)	Vernes *et al.* (2008)
CMIP	NWR	Association (quantitative + categorical)	Newbury *et al.* (2009)
ATP2C2	NWR	Association (quantitative + categorical)	Newbury *et al.* (2009)
7q	Language impairment	Linkage (categorical)	Villanueva *et al.* (2011)

1.10: Aims of the Work Described in this Thesis

My work aimed to apply methodologies that have not been applied to the study of language impairment in the past, to provide more insight into the molecular genetics of language impairment. I performed several genome-wide association analyses which test for child, maternal, parent-of-origin and interaction effects. I also performed several association analyses using imputed human leukocyte antigen types in the SLIC families and controls. I performed a directed association analysis using the Robinson Crusoe Island cohort. Working with two separate cohorts enabled me to examine the genes identified using one population, in the other one. In addition to association methods, I also looked for rare variants in candidate genes for SLI using high-throughput and exome sequencing data for SLI probands and performed gene expression analyses in a case study of a language disorder. Finally, I performed a functional study of a candidate gene that I had identified.

Chapter 2: Materials and Methods

The following chapter describes the cohorts and samples, techniques, protocols, software, statistical tools and the databases used in this work.

2.1: Subjects

2.1.1: The SLI Consortium Cohort

The analyses in Chapters 3, 4 and 6 used the SLIC cohort. The linguistic tests and inclusion and exclusion criteria were described in Chapter 1; detailed documentation of the cohort may be found in (The SLI Consortium, 2002, The SLI Consortium, 2004, Monaco and The SLI Consortium, 2007, Falcaro *et al.*, 2008, Newbury *et al.*, 2009). This study also used additional DNA samples (from 49 families) from Guy's Hospital, London which were not used in previous SLI Consortium studies. Following the genotyping of the SLIC DNA samples on SNP arrays, which will be discussed shortly, individuals were removed from the experimental cohort, if they had a low genotype rate; if they were outliers in a principal component analysis, or if they had a heterozygosity rate which was ±2 SD from the mean. Individuals who may have been mislabelled were removed, as were those who had aneuploidy of the X or Y chromosomes or discordant sex information. Further details are given in Section 2.6. Affection statuses in categorical association analyses were different from the ones used in the published SLIC studies, and they will be described in Section 2.6. Quantitative analyses used the NWR, ELS and RLS traits, as described previously.

In two of the human leukocyte antigen (HLA) analyses, described in Chapter 4, 566 population DNA samples taken from the cohort of the People of the British Isles (POBI) (Winney *et al.*, 2012) and an Oxfordshire study of gene expression in primary immune cells (Fairfax *et al.*, 2012) were used as controls. Since the POBI samples come from a variety of places in the British Isles, some minor differences in population sub-structure might still be present, even though the POBI samples and Oxfordshire samples were analysed in the principal component analysis, as described above.

The long-range sequencing of candidate genes for SLI used samples from 140 probands taken from the SLIC cohort and the AUTPREV cohort, the latter of which came from Guy's Hospital collection of SLI singleton cases, in a pooled-DNA high-throughput sequencing experiment. High-quality DNA samples from 140 probands were available. Fourteen pools of ten individuals each was the optimal design considering limitations on the number of tags that could be used in one sequencing lane as well as the number of samples and the cost of the sequencing. One control sample was mistakenly included. In some cases, samples from affected siblings were used when samples from the probands could not be used. This was communicated to me after the completion of the analyses; I was not responsible for choosing the samples. These samples alone could not be excluded afterwards, as the reactions and analyses used pooled DNA samples. However, the inclusion of a control sample is not likely to have affected the results much, as it involved only 1 out of 140 individuals, and none of the rare or novel variants described in Chapter 6 were found in that individual. There was one case in which a variant was found in an affected sibling, but all the family members related to the sibling who had language problems were individually sequenced for the variant which had been found in that individual. Six probands who were excluded otherwise (as described in Section 2.1.1) were present in the pools.

2.1.2: The Robinson Crusoe Island Cohort

The analyses described in Chapter 5 used the Robinson Crusoe Island cohort. This cohort was described in Chapter 1, and further details can be found in (Villanueva *et al.*, 2011). The affection statuses used in the case-control analyses are the same as the ones described in Chapter 1. Additionally, several other samples were used in analyses which pertained to the Robinson Crusoe Island study. These include: samples from 320 Colombian unselected controls (collected for an unrelated study); UK Caucasian control samples from the Human Random Control DNA panels; 46 new DNA samples (from 16 families) from the Robinson Crusoe Island collected from individuals who had language impairment and other deficits (excluded for SLI), which were supplied by P. Villanueva (University of Chile, Santiago); samples from 121 Chilean controls, which included samples from 30 male students (supplied by P. Villanueva, University of Chile, Santiago), and 91 adult female controls from a

breast cancer study (supplied by L. Jara, University of Chile, Santiago). It is important to note that affection statuses for the parents and children were determined in different ways (as outlined in Chapter 1). In particular, due to the fact that parents were tested for their IQ and language skills when they were at a later stage of their lives, it was not possible to tell whether their IQ was influenced by a language problem (in case of parents with language problems), or whether their language problem was a result of a general developmental problem. For that reason, parents with a history of language problems who also had children that were diagnosed with SLI in the study were considered affected themselves. Additionally, in the linkage study, some clinical diagnoses overrode affection statuses as determined by the test scores at the clinicians' discretion. Further details are given in Chapter 5.

2.1.3: A Case-study of a Spanish Girl with a Severe Speech and Language Impairment

Chapter 8 describes the investigation of a speech and language-impaired Spanish girl. The case is described in (García-Bellido *et al.*, 2009). Further details are given in Chapter 8.

2.2: Protocols for Experimental Work

2.2.1: Polymerase Chain Reaction

The polymerase chain reaction (PCR) mix consisted of: 5 µl genomic DNA (8 ng/µl); 1 µl of the forward primer (10 µm); 1 µl of the reverse primer (10 µm); 5 µl of buffer (160 mM (NH4)2SO4, 670 mM Tris-HCL, 0.1% stabiliser); 2.5 µl of $MgCl_2$ (50 mM); 1.25 µl of deoxyribonucleotides (dNTPs) (10 mM); 0.2 µl of 9:1 BIOTAQ (Bioline):Pfu (Thermo Scientific) (5 u/µl); 34.05 µl of MilliQ H_2O to a final volume of 50 µl.
The PCR included the following steps:

1. 95°C for 10 minutes
2. 95°C for 30 seconds
3. Ta (annealing temperature) for 30 seconds
4. 72°C for 30 seconds

5. Steps 2-4×29
6. 72°C for 5 minutes

When the above conditions were not suitable, amplification was performed using touchdown PCR, meaning that the first steps included high annealing temperatures which were subsequently reduced in later steps.
The touchdown PCR included the following steps:

1. 95°C for 10 minutes
2. 95°C for 30 seconds
3. Maximum Ta for 30 seconds, -0.5°C per cycle
4. 72°C for 30 seconds
5. Steps 2-4×13
6. Minimum Ta for 30 seconds
7. 72°C for 30 seconds
8. Steps 6-7×29
9. 72°C for 7 minutes

Annealing temperatures differed based on the primer pair. Water-controls (no DNA) were included for each reaction type (i.e. per each fragment) and examined after electrophoresis on agarose gels.

For the purpose of mapping of the inversion breakpoints in the case of the Spanish proband with language impairment (Chapter 8), different PCR conditions and reagents were used, since the expected PCR product was very large. Phire Hot Start II DNA polymerase (Thermo Scientific) was used in the PCR (1 µl), together with 5x Phire buffer (10 µl), 2.5 µl of each primer (10 µM), 1 µl of dNTPs (10 mM), 5 µl of genomic DNA (15 ng/µl), and 28 µl of MilliQ H_2O.

The reaction conditions were as follows:

1. 98°C for 30 seconds
2. 98°C for 5 seconds
3. 68°C for 5 seconds
4. 72°C for 10 minutes
5. Steps 2-4×29
6. 72°C for 1 minute

2.2.2: EXO-SAP Reaction

The EXO-SAP reaction is used to clean the PCR product if it is to be used for sequencing. The following was added to the product of the PCR: 1.6 µl of SAP buffer (25 mM Tris-HCL, 1 mM $MgCl_2$, 0.1 mM $ZnCl_2$, 50% glycerol); 1.6 µl of SAP (1 u/µl); 1.5 µl of MilliQ H_2O and 0.16 µl of Exonuclease 1 (20,000 u/µl). The samples were heated to 37°C for 30 minutes and then to 80°C for 15 minutes.

2.2.3: Sequencing Reaction

Sequencing reactions were performed with 2 µl of the cleaned PCR product together with: 0.5 µl of the forward or the reverse primer (10 µm); 0.5 µl (or 2 µl) of Big Dye; 1.75 µl (or 1 µl) of Big Dye buffer and 5.25 µl (or 4 µl) of MilliQ H_2O.
The reaction included the following steps:

1. 95°C for 1 minute
2. 95°C for 10 seconds
3. 50°C for 10 seconds
4. 60°C for 4 minutes
5. Steps 2-4×35

When less Big Dye was used, more Big Dye buffer was used, and the volume was kept at 10 µl per reaction by correcting the volume of MilliQ H_2O.

2.2.4: Precipitation

The following was added to the product of the sequencing reaction (already in a plate): 2 µl 125 mM (ethylenediaminetetraacetic acid (EDTA), 2 µl 3 M NaAc and 50 µl 100% ethanol. Then, the plates were spun for 30 minutes at 4°C, 3000×g. The plates were spun inverted for 10 seconds at 190 g. Then 70 µl of 70% ethanol were added to each well and the plates were spun again at 4°C, 1650×g for 15 minutes. The plates were then spun inverted for 10 seconds at 190 g and the wells were left to dry and sent to the Zoology department, University of Oxford, for Sanger sequencing.

2.2.5: Reverse-transcription PCR

The QuantiTect Reverse Transcription kit (Qiagen) was used to perform reverse-transcription PCR (RT-PCR). The protocol included two procedures: a genomic DNA (gDNA) elimination reaction and the reverse transcription reaction. In the first step, 0.5 µg of RNA, 2 µl of gDNA Wipeout Buffer 7x, and RNase-free water (up to a final volume of 14 µl) were incubated for 2 minutes at 42°C. In the second step, 1 µl of Quantiscript Reverse Transcriptase, 4 µl of Quantiscript RT Buffer 5x and 1 µl of RT Primer Mix were added to the sample from step 1. The sample was incubated for 15 minutes at 42°C, and then for 3 minutes at 95°C. Two types of controls were added: a reverse-transcriptase negative control for each RNA sample, and one RNA negative control. The complimentary DNA (cDNA) was then diluted with MilliQ H_2O 1:9 to be used in a PCR.

In the case of the RNA I extracted for the functional study described in Chapter 8, DNase treatment was performed using the rigorous protocol of the TURBO DNA-free kit (Life Technologies). The cDNA was treated with RNase-H (Ambion) (1 µl of enzyme; incubation at 37°C for 30 minutes). Those steps were taken in order to eliminate as much of the plasmid (with which the cells were transfected) as possible, and to avoid high levels of the plasmid cDNA from interfering with the PCR.

2.2.6: Agarose Gel Electrophoresis

Agarose was dissolved in Tris-Borate EDTA (TBE) solution to make a gel. A 10x Tris-Borate EDTA (TBE) solution was used which included: 1 M Tris Base, 0.01 M EDTA, and 0.9 M boric acid. Gels were made with 3 g (or 1.2 g, when the size of the expected fragment was a few kb) of agarose in 150 ml of 1x TBE and 15 µl of SYBR Safe DNA gel stain (added to the heated agarose and TBE solution). The ladder used was the 1 kb Plus ladder from Invitrogen (see Figure 2.1). Electrophoresis was performed at a constant voltage of 100-140 V, depending on the expected size of the fragments.

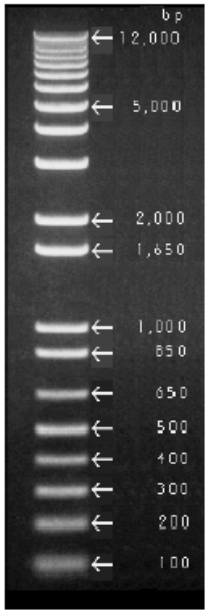


Figure 2.1: 1 kb Plus DNA ladder (Invitrogen).

2.2.7: RNAs and cDNAs

Apart from the RNA extracted from cells grown by me (or others, where stated otherwise), RNA samples from various brain regions were obtained from Clontech (extracted from adult brains). The cDNA panels used were Amsbio cDNA human adult normal tissue major panels 1 and 2, which together included cDNA from the following tissues: heart, brain, kidney, liver, lung, pancreas, spleen, skeletal muscle, and placenta.

2.2.8: DNA Extraction from the Gel

The QIAquick Gel Extraction kit (Qiagen) was used to extract DNA from the gel for analysing bands of varying sizes. The protocol included the following steps:

- The DNA bands were excised from the gel and placed in a tube.
- 750 µl of Buffer QG were added to each tube.
- The tubes were incubated at 50°C until the agarose completely dissolved.
- 250 µl of isopropanol were added to each tube.
- The contents of each tube were transferred to QIAquick spin columns, and the columns were then spun at 13,000 rpm for 1 minute using a table-top microcentrifuge, and the flow-through was discarded. This step was repeated since the contents of each tube were greater in volume than 800 µl (the volume of the column).
- 500 µl of Buffer QG were then added to each column. The columns were spun again, and the flow-through was discarded.
- 750 µl of Buffer PE were added to each column, and the columns were allowed to stand for 5 minutes. The columns were then spun, and the flow-through was discarded.
- The columns were spun again, and any residual Buffer PE (wash buffer) was discarded.
- DNA was eluted in the following way: 30 µl of MilliQ H_2O were added to the centre of each column, and the columns were allowed to stand for 4 minutes. Each column was then placed in a clean tube, and the tubes containing the columns were spun.

2.2.9: Cells

Human embryonic kidney HEK-293T cells were grown in Dulbecco's Modified Eagle's Medium (Sigma) supplemented with 2 mM L-glutamine, 10% foetal bovine serum, and 1% penicillin/streptomycin (from a stock of 10,000 units penicillin and 10 mg streptomycin/ml). SH-SY5Y cells were grown in medium consisting of 1:1 Minimum Essential Medium Eagle (Sigma) and Nutrient Mixture F-12 Ham (Sigma). The medium was supplemented with

2 mM L-glutamine, 1% non-essential amino acids, 15% foetal bovine serum, and 1% penicillin/streptomycin (from a stock of 10,000 units penicillin and 10 mg streptomycin/ml).

The cells were grown at 37°C and 5% CO_2. The medium was changed every 1-3 days, and, when the cells were at least 80% confluent, the cell culture was split using trypsin/EDTA.

2.2.10: Transient Transfection

HEK cells were transfected with either a plasmid containing the cDNA of *NFXL1* with a C-terminal Myc-DDK tag (RC216989) by Origene, or the empty vector (pCMV6-Entry Vector), also by Origene. Transfections were performed in a six-well plate using 3 µg of plasmid, 4 µl of jetPRIME (Polyplus transfections), and 200 µl of jetPRIME buffer per well. Transfections were performed 24 hours after seeding 2.5×10^5 cells per well. 3 wells were transfected with the plasmid containing the cDNA, and three wells were transfected with the empty vector. After 4 hours, the transfection medium was changed. After 24 hours, the cells were processed for either RNA extraction or western blot analysis (WB); ~10^6 cells were used for RNA extraction, and the rest were used to produce lysate for the WB.

2.2.11: Western Blot Analysis

Cells grown on a 6-well plate were washed with phosphate buffered saline (PBS) and detached through pipetting PBS over the cell culture. Some of the cells were then used for RNA extraction (see below), and the rest were spun at 250×g for 5 minutes in order to pellet them.

The cells were lysed using 150 µl of the following lysis buffer: 50 mM Tris-HCl pH 7.8, 150 mM NaCl, 1 mM EDTA, 1% Triton X-100, 1x Protease Inhibitor Cocktail (Roche). The cell lysates were kept on ice for 30 minutes. After spinning the lysates at 1000×g for 5 minutes at 4°C, the post-nuclear supernatant was collected to be used in the WB analysis.

Lysate protein concentration was measured using the Pierce BCA (bicinchoninic acid) Assay (ThermoFisher). Bovine serum albumin standards were prepared according to Table 2.1.

Table 2.1: Preparation of standards for measuring protein concentration in cell lysate.

Standard	mg/ml	µl	From	µl buffer	Initial vol.	Final vol.
A	2	35	stock	0	35	35
B	1.5	30	stock	10	40	40
C	1	60	stock	60	120	40
D	0.8	56	C	14	70	35
E	0.6	24	C	16	40	40
F	0.4	35	D	35	70	40
G	0.2	30	F	30	60	40
H	0.1	20	G	20	40	40

20 µl from each sample (including the standards) were added to a flat-bottom 96-well plate. BCA working reagent was prepared by mixing 50 parts reagent A and 1 part reagent B. 200 µl of working reagent were added to each well, and the plate was kept at 37°C for 30 minutes. Protein concentrations were determined by measuring the absorbance at 550 nm using a plate reader.

Electrophoresis was performed using the NuPAGE system (Invitrogen). Samples to be loaded consisted of up to 65% cell lysate, 25% LDS sample buffer 4x, and 10% reducing agent 10x. The samples were heated at 70°C for 10 minutes. The protein ladder used was the Spectra Multicolor Broad Range Protein Ladder (Fermentas).

Samples were loaded onto a 7% Tris-Acetate mini gel in 1x Tris-Acetate sodium dodecyl sulfate (SDS) running buffer (prepared using 20x Tris-Acetate SDS (Invitrogen) and deionised water). The buffer for the upper chamber of the Mini-Cell system was prepared using 200 ml of the running buffer and 500 µl of antioxidant (Invitrogen).

The gel was placed in a Mini-Cell system so that the wells faced the inner part of the system. The upper chamber was filled with buffer containing the antioxidant, and the lower chamber was filled with running buffer. The gel was run at 150 V for 75 minutes.

For the transfer stage of the WB, proteins were transferred using 1x transfer buffer (prepared using 20x NuPAGE transfer buffer (Invitrogen), 10% methanol, and deionised water). 250 µl of antioxidant were added to 250 ml of transfer buffer to be used per one transfer. A retentive polyvinylidene difluoride (PVDF) membrane was soaked in methanol for

30 seconds, rinsed with deionised water, and left in 1x transfer buffer. After removing the gel from its container, the two ends of the gel were cut. A filter paper-gel-PVDF membrane-filter paper "sandwich" was placed between blotting pads soaked with transfer buffer inside the transfer module of the Mini-Cell system (the gel was places towards the bottom of the cathode). The module was then filled with the remaining transfer buffer, and the tank was filled with deionised water. The transfer was run at 30 V for 75 minutes. The membrane can be kept at 4°C in a solution of Tris-buffered saline (TBS) with 0.1% Tween-20.

For the blotting stage, the PVDF membrane was blocked in 5% skimmed milk dissolved in TBS Tween-20 0.1% for at least 30 minutes. Incubation with the primary antibody was done using 1% skimmed milk in TBS Tween-20 0.1%, with the primary antibody added at a dilution of 1:2000. The membrane was placed on a rocker in a cold room overnight. The antibody used was a rabbit polyclonal anti-*NFXL1* antibody (A303-197A) from Bethyl Laboratories. The membrane was then washed 3 times with TBS Tween-20 0.1%, each wash lasting 10 minutes. The membrane was then placed in 25 ml 1% skimmed milk in TBS Tween-20 with the secondary antibody, goat anti-rabbit IgG (H+L)-HRP conjugate (Bio-Rad), at a dilution of 1:5000, and this was left on a rocker for one hour and then washed 3 times, as described above.

Following the blotting, the membrane was placed on a cut transparent plastic bag on the bench. 0.5 ml of ECL Plus Reagent A (Amersham Biosciences) was mixed with 0.5 ml of Reagent B (Amersham Biosciences), and the mix was pipetted over the membrane. The top side of the cut plastic bag was used to cover the membrane. The bag was covered and left on the bench for 5 minutes. Pictures of the membrane were taken using the Bio-Rad ChemiDoc MP imaging system using two protocols, one to capture the bands, and another to capture the markers.

2.2.12: Immunofluorescence

HEK and SH-SY5Y cells were seeded onto coverslips coated with poly-L-lysine. Some of the cells were transfected with the NFXL1 plasmid (as described previously), with 0.75/0.5 µg of the plasmid for HEK/SH-SY5Y cells and 1 µl of jetPRIME, and others were grown as controls. The day following the transfection, cells were washed with PBS, fixed with either 3% paraformaldehyde (30 minutes at room temperature) or cold methanol (10 minutes at room temperature) and then washed five times with PBS.

A blocking and permeabilisation solution (0.1% Triton X-100, 10 mg/ml (1%) fish gelatine, 10% normal goat serum in TBS) was then applied to the cells for 10 minutes (300 µl per well in a 24-well plate). Primary antibodies diluted in 250 µl of blocking solution were added to each well (the combinations of antibodies are given in the relevant chapter). The antibodies used were: Anti-NFXL1 A303-196A (Bethyl), Anti-NFXL1 A303-197A (Bethyl), Anti-DDK TA50011-100 (Origene), and Anti-Myc obtained from the Imperial Cancer Research Fund (likely to be from mouse clone 9E10). The plates were incubated at room temperature for one hour with slow rocking. Following five washes with PBS, Alexa Fluor-conjugated secondary antibodies diluted in 250 µl of blocking solution were added to each well. The plates were covered and incubated at room temperature for one hour with slow rocking. Cells were then washed five times with PBS and stained with a 5 µg/ml 4',6-diamidino-2-phenylindole (DAPI) solution. Cells were then washed three times with PBS. The coverslips were mounted onto glass slides with Fluoromount-G (Southern Biotech) and left at 4°C until use. Samples were analysed using a Zeiss LSM 510 Meta Confocal Microscope.

2.2.13: Quantitative PCR

The quantitative PCR (qPCR) used the standard TaqMan protocol, TaqMan Gene Expression Assay probes and primers, and the TaqMan Gene Expression Assay Master Mix from Applied Biosystems (Life Technologies). In the gene expression analysis described in Chapter 7, singleplex PCRs were performed. In the gene expression analysis described in Chapter 8, multiplex PCRs were performed (in order to include three RNA batches in the same run). Each sample included: 4 µl of a 1:10 dilution of the RT-PCR product, 10 µl of

TaqMan Gene Expression Master Mix 2x, 1 µl of TaqMan Gene Expression Assay FAM dye-labeled 20x, 1 µl of TaqMan Gene Expression Assay VIC dye-labeled 20x (primer-limited) for the endogenous control gene, and 4 µl of RNAse-free water. Reactions that used only one assay had 5 µl of RNAse-free water. The reaction conditions were as follows:

1. 10 minutes at 95°C
2. 15 seconds at 95°C
3. one minute at 60°C
4. Steps 2-4×40

Reactions were run on a Bio-Rad iQ5. Samples were run in triplicates, and each experiment included reverse-transcriptase negative controls for each sample and an RNA negative control. When one sample in a triplicate was an outlier, it was taken out. Mean cycle thresholds (C_T) were calculated per triplicate (or duplicate, if an outlier was removed).

Quantification of gene expression was performed using the $\Delta\Delta C_T$ method (Livak and Schmittgen, 2001). The C_T values for a given gene were first normalised per sample using the C_T values of a housekeeping gene (HKG). The difference between those C_T values was then normalised using the difference between C_T values belonging to a reference sample. $\Delta\Delta C_T=(C_{T,g,s} - C_{T,h,s})-(C_{T,g,r} - C_{T,h,r})$, where s is a given sample, r is the reference sample, f is a given gene, and h is the HKG. The relative expressions levels were calculated using the following formula: $2^{-\Delta\Delta C_T}$.

2.2.14: Transformation (Cloning Plasmids Using Bacteria)

Before the plasmids were used in the transfections, they were cloned using *E.Coli* bacteria to produce greater amounts of plasmid. The protocol involved the following steps: 1 µl of plasmid at 0.5 µg/µl was added to 50 µl of XL1Blue competent E.Coli cells. The mix was left on ice for 20 minutes. It was incubated at 42°C for 90 seconds, then removed and put on ice for 2 minutes. 200 µl of SOC (super optimal broth with catabolite repression) were added to the mix, and the SOC mix was incubated at 37°C for 30 minutes. The bacteria were then seeded on LB (lysogeny broth) plates with 15 µg/ml kanamycin (as the plasmids used had a kanamycin resistance gene). The plates were left overnight at 37°C. The inoculation step was

performed the following day: One colony was transferred to a flask containing liquid LB medium with 30 µg/ml kanamycin (twice as much as the concentration used for the solid medium). The flask was incubated overnight in a shaking incubator at 37°C, 225 rpm. The following day, the plasmids were extracted from the bacteria with the GeneElute HP Plasmid Midiprep kit (Sigma): The cell culture was pelleted through centrifugation, resuspended in resuspension solution and mixed on a vortex machine. Lysis solution was added, and the sample was left at room temperature for 5 minutes. Four ml of the neutralisation solution were added, and then the sample was mixed by inverting the tube 8 times. Three ml of binding solution were added to the sample, which was then inverted twice. The sample was then added to the barrel of a filter syringe and left there for 5 minutes. The lysate was pressed through the filter syringe and into a column pre-prepared with the column preparation solution, which was put inside a collection tube. The column was spun at 3000×g for two minutes and the flow-through was discarded. This was done twice per column, as the volume of the lysate was bigger than the volume the column could hold. The columns were washed with the wash solution (4 ml) through centrifugation at 3000×g for 5 minutes. The plasmids were eluted with 1 ml of elute solution through centrifugation at 3000×g for 5 minutes. The extracted plasmids were compared with the original plasmids supplied by Origene through running them on a gel following digestions with several restriction enzymes, and the cDNA of the plasmid containing the *NFXL1* cDNA was sequenced using Sanger sequencing.

2.2.15: RNA Extraction

RNA was extracted using the RNeasy Mini kit (Qiagen). About 10^6 cells (as counted with a TC10 automated cell counter, Bio-Rad) were lysed with 350 µl Buffer RLT. 350 µl of 70% ethanol were then added to the sample. The sample was then transferred to an RNeasy Mini spin columns. The sample was spun at 8000×g for 20 seconds. The sample was then treated with DNase while inside the column following the standard protocol of the RNase-free DNase set (Qiagen). The column was washed with Buffer RW1 (350 µl buffer, centrifugation at 8000×g for 20 seconds), and then with buffer RPE (500 µl, centrifugation at 8000×g for 2 minutes). The membrane inside the column was then dried through the centrifugation of the

column at 8000×g for 1 minute. RNA was eluted with water through centrifugation at 8000×g for 1 minute.

2.3: SNP Arrays

The SLIC samples were genotyped using the Illumina Human OmniExpress array 12v1. The Robinson Crusoe Island DNA samples were originally genotyped using the Illumina HumanLinkage-12 array for the linkage analysis and were later genotyped using the Affymetrix Axiom GW LAT 1 array and a custom array for the association analyses described in this thesis. The physical positions of SNPs in the analyses that used the Illumina arrays are in build hg18, and the ones in the analyses that used the Affymetrix array are in build hg19. Further details regarding the quality control for the genotype data used in the various association analyses are given in the Statistical Analyses section in this chapter.

2.4: Databases

The UCSC Genome Browser (Kent *et al.*, 2002) and the ENSEMBL Genome Browser (Flicek *et al.*, 2012) were used to obtain reference sequences of genes, and to obtain general information on genomic regions. Data from the 1000 Genomes project (The 1000 Genomes Project Consortium, 2012) were used in order to obtain variant frequencies in different populations. Further details regarding how the data were used will be given in the relevant chapters. SNP data for SLI candidate genes were obtained from the 22/7/2014 release of the UK10K project (http://www.uk10k.org). Allele frequencies and counts from 2000 ALSPAC samples included in the release, which were listed as controls on the UK10K website, were used in a case-control analysis described in Chapter 6. The SCAN database (Gamazon *et al.*, 2010) was used in the annotation of the SLIC GWAS results which included expression quantitative trait locus (eQTL) information. Allele frequencies in the general population for variants identified in the long-range sequencing of SLI candidate genes were obtained using the Exome Variant Server (EVS) (http://evs.gs.washington.edu/EVS). The protein database used was UniProt (The UniProt Consortium, 2012).

2.5: Software

2.5.1: Software Used for Data Management and Pedigree Analysis

PLINK (Purcell *et al.*, 2007) was used for editing pedigree and map files and for converting files into other formats to be used with other tools, as well as to perform several quality control procedures and analyses, which will be discussed in the relevant sections. PEDSTATS (Wigginton and Abecasis, 2005) was used to check pedigrees for Mendelian errors and markers for Hardy-Weinberg equilibrium (HWE). Pedigrees were drawn with HaploPainter (Thiele and Nurnberg, 2005).

2.5.2: Primer Design

Primers were designed with Primer3 (Rozen and Skaletsky, 2000).

2.5.3: Imputation and Grouping of HLA Types

HLA types for *HLA-A*, *HLA-B*, *HLA-C*, *HLA-DQA1*, *HLA-DQB1* and *HLA-DRB1* were imputed using HLA*IMP version 1 (Dilthey *et al.*, 2011). HLA*IMP selects informative SNPs that are available in both the pre-selected reference set as well as the sample set. SNPs were aligned and the genotypes were phased within HLA*IMP with the HapMap3 b36 reference panel. HLA types were imputed on the HLA*IMP server. Imputations were used in further analyses if they had a Q (maximum of the posterior probability distribution) value of at least 0.9. For SLIC samples, ~90% of the genotypes were retained using this threshold, and, for the POBI and Oxfordshire samples, ~91% of the genotypes were retained. HLA types were grouped based on serological equivalence according to the HLA Dictionary (Schreuder *et al.*, 2001) (HLA-DQA1 types were not grouped, since *HLA-DQA1* types were not included in the HLA Dictionary). The codes used for grouped alleles can be found in the HLA Dictionary tables in the “WHO assigned” column (alleles that were in the HLA Dictionary but did not have a WHO assignment were not examined). For example, HLA-A*0101 and HLA-A*0102, among other HLA-A*01 alleles, would be considered A1. The serological groups will hereafter be

referred to as alleles for simplicity. Alleles that were not present in the HLA Dictionary were left ungrouped and have their four-digit designations e.g. HLA-DQA1*0501.

2.5.4: Variant-calling and Annotation in Candidate Genes for SLI

Three variant-calling tools were tested with data obtained from the sequencing of candidate genes for SLI: SAMtools (Li *et al.*, 2009), Syzygy (Rivas *et al.*, 2011), and Platypus (Rimmer *et al.*, 2012). These programs generated variant call format (VCF) files, which were then annotated using SnpEff (Cingolani *et al.*, 2012). Further details are given in the relevant chapter.

2.5.5: Tools for Predicting the Effects of Amino Acid Substitutions and Nuclear Localisation

Predictions of the effects of amino acid substitutions were obtained with MutationTaster (Schwarz *et al.*, 2010) or PolyPhen (Adzhubei *et al.*, 2010). NucPred was used to predict whether a protein was likely to be localised in the cell nucleus (Brameier *et al.*, 2007).

2.6: Statistical Analyses

The following section briefly describes the statistical tools and tests employed in the work described in this thesis. Further information may be found in the relevant chapters.

2.6.1: Estimation of Child, Maternal, Parent-of-origin and Interaction Effects

Estimation of child, maternal, parent-of-origin and interaction effects in the SLIC cohort was performed with EMIM (Howey and Cordell, 2012). EMIM employs a multinomial model in which various parameters may be estimated from the data (genotypes for each marker for each family and affection statuses). These parameters represent the increase in the risk which may result from the child's genotype, the mother's genotype, paternal or maternal parent-of-origin effects, or interactions between the mother and child genotypes. Several parameters may be estimated together. EMIM estimates the likelihoods of two models: a null model, in which there is no increase in risk, and an alternative model in which

there is an increase in risk (based on the parameter(s) one wishes to estimate, e.g. maternal genetic effects). In order to calculate p-values, one needs to use the value of twice the difference between the maximised log-likelihoods of the two models as a χ^2 test statistic. One degree of freedom was used for tests in which one parameter was tested (trend, parent-of-origin and MFG analyses), and two degrees of freedom were used for tests in which two parameters were tested (genetic effects). In general, the number of degrees of freedom equals the difference between the number of estimated parameters in the alternative model and the number of estimated parameters in the null model, e.g. if the alternative model estimated one parental parent-of-origin effects parameter, and in the null model that parameter was set to 1, then one degree of freedom should be used to calculate the p-value. Further information about the tests and parameters follows. One-sided p-values were calculated in R with the "1-pchisq", with the appropriate number of degrees of freedom.

Input files for EMIM were generated with PREMIM using the -a parameter, which includes the estimation of minor-allele frequencies (MAFs). By default, minor alleles are treated as the risk alleles in the EMIM analyses, but that does not necessarily reflect the actual contribution to the risk of the minor alleles. EMIM can estimate various parameters which may increase (or decrease, if the major allele turns out to be the risk allele) the risk of having the disease in question (Ainsworth *et al.*, 2011). Several tests were run with EMIM:

- Child genetic effects (two parameters estimated: the increase in the risk due to the child's having one risk allele; the increase in the risk due to the child's having two risk alleles).
- Child trend (one parameter estimated: the increase in the risk due to the child's having one risk allele; the increase in the risk due to the child's having two risk alleles is assumed to equal the former, squared).
- Maternal genetic effects (two parameters estimated: the increase in the risk due to the mother's having one risk allele; the increase in the risk due to the mother's having two risk alleles).

- Maternal trend (one parameter estimated: the increase in risk due to the mother's having one risk allele; the increase in the risk due to the mother's having two risk alleles is assumed to equal the former, squared).
- Paternal parent-of-origin effects (one parameter estimated: the increase in the risk due to the child's inheriting the risk alleles from the father).
- Maternal parent-of-origin effects (one parameter estimated: the increase in the risk due to the child's inheriting the risk alleles from the mother).
- Maternal-foetal genotype incompatibility test – null allele and antigen-coding allele (one parameter estimated: the increase in the risk due to genotype mismatching between the mother and the child under the null allele and antigen-coding allele scenario).
- Maternal-foetal genotype incompatibility test – two antigen-coding alleles (one parameter estimated: the increase in the risk due to genotype mismatching between the mother and the child under the two antigen-coding alleles scenario).
- Maternal-foetal genotype matching test (one parameter estimated: the increase in the risk due to genotype matching between the mother and the child).

The "child" analyses test for the increase in the child's risk of having the disease when the child carries the risk allele. The "increase in risk" refers to the factor by which the baseline probability of a disease (when the child carries only the low-risk allele) is multiplied. The "maternal" analyses test for the increase in the child's risk of having the disease when the mother carries the risk allele. "Genetic effects" analyses estimate the increase in risk when having one risk allele and two risk alleles separately, while the "trend" analyses estimate the increase in risk when having one risk allele and assume that the increase in risk when having two risk alleles equals the former, squared. Analyses of parent-of-origin effects test for the increase in risk when the child receives a copy of the risk allele only from the mother or only from the father; the parameterisation used by EMIM is based on (Weinberg *et al.*, 1998). Several maternal-foetal genotype (MFG) analyses were run. The maternal-foetal genotype incompatibility test was run using two parameters which correspond to scenarios 1A and 1B described in (Sinsheimer *et al.*, 2003). The first of the two tests assumes that a locus has two alleles and only one of them codes for an antigen. Under such a scenario, a mismatch

between a maternal genotype which only includes null alleles and a foetal genotype which includes the antigen-coding allele may result in an increased risk of having the disease. The second of the two tests is designed for cases in which both alleles may code for an antigen. Under this scenario, a mother who is homozygous for either of the two alleles may develop an immune response when the child is heterozygous, and this may increase the risk of having the disease. In sum, the two MFG incompatibility analyses test whether the MFG mismatch increases the risk of having the disease. The MFG-matching analysis tests whether a match between the maternal and foetal genotypes, rather than a mismatch, i.e. the child's not having any allele that the mother does not have, increases the risk of having the disease (Palmer *et al.*, 2006).

Affection status was based on the expressive and receptive language scores from the CELF: probands and any siblings with ELS or RLS of 1.5 SD below the general-population mean for their age were defined as cases (mean=100, and SD=15); siblings with ELS and RLS of more than 0.5 SD below the mean were defined as controls; all other individuals were given an "unknown" affection status. EMIM uses cases, case duos (case and one parent), and case trios (case and two parents) in its analyses. EMIM uses every individual only once, and, therefore, the subsets of the families used in the analyses are independent. Individuals who had "unknown" affection statuses, controls, and their parents (unless they were in a subset with a case) were not used in the analyses due to the fact that EMIM assumes that individuals with a missing phenotype may be considered controls, which holds true only for rare diseases.

The raw genotype data were processed with the Genome Studio software (Illumina). The quality control measures for the SNPs used in the EMIM analyses are as follows: 47 samples were duplicated across plates (concordance rate of 0.98968), and all the samples were randomised across plates with cases and controls being spread evenly across plates, as were different sample types (saliva, blood, or mouth swab samples) and samples of different origins (the various UK centres which supplied the samples). SNPs with a Gentrain score below 0.5 were removed. Before performing the association analyses, SNPs and samples were filtered based on several quality control measures, as described in (Anderson *et al.*,

2010): the success thresholds for genotyping for both SNPs and samples were 95%, and SNPs with a minor allele frequency of less than 1% were removed. SNPs and samples with an error rate of 1% or higher, as estimated by bad inheritances, were removed (this was confirmed using PEDSTATS). SNPs were filtered with PLINK based on a Hardy-Weinberg equilibrium p-value threshold of 0.001. Genotypes were compared with existing SNP genotype data for 700 individuals and 734 SNPs, with an error rate of 0.00967. The total number of SNPs used in the EMIM analyses was 614,937. Following the generation of p-values, SNPs that showed trends of association were manually checked for good clustering with Genome Studio i.e. SNP clustering was checked to make sure there were three distinct clusters (for genotypes AA, AB, and BB). SNPs with association p-values ≤0.0001 were checked for HWE using PEDSTATS, which selects unrelated samples in a different way from that used by PLINK, resulting in a more powerful test. SNPs with HWE P≤0.001 were then removed.

2.6.2: Quantitative Association for HLA Data and SNPs in the HLA Region

SNPs from the HLA region on chromosome 6 (7625 SNPs in total) were used in these analyses. Three traits were used: ELS, RLS and NWR.

MERLIN (Abecasis *et al.*, 2002) was used to generate p-values for associations with SNPs in the HLA region, because it is optimised for analysing SNP data and quantitative traits in family-based cohorts. MERLIN tests for association using one SNP at a time, and employs a likelihood-ratio test to generate p-values. MERLIN was also used to estimate IBD values for QTDT.

Quantitative association analysis for grouped HLA types in the SLIC cohort was performed with QTDT (Abecasis *et al.*, 2000), which accepts more than 2 alleles per marker (and HLA genes are multi-allelic). As discussed in Chapter 1, quantitative transmission/disequilibrium tests test for changes in the phenotype value as a function of the number of copies of a given marker allele carried by a child, given the expected number, which is, in turn, based on the parents' genotypes. QTDT, specifically, can use variance

components to calculate the trait means and variances used to construct the association test. Several tests were run with QTDT:

- The orthogonal "within family" model (-ao option).
- Paternal parent-of-origin effects.
- Maternal parent-of-origin effects.
- A test for determining whether maternal and paternal alleles are significantly different.

In the orthogonal model, the likelihoods of two models are estimated and compared using a maximum-likelihood test, a null model, and an alternative model, in which a within-family component of association is estimated. In the paternal and maternal effects tests only paternally-derived or maternally-derived alleles are tested, respectively. The last test was used to test whether the differences between the maternal and paternal effects were significant. If the differences were not significant for any grouped HLA type tested, then the p-value for that HLA type was not considered significant, even if it was smaller than 0.05 in the original parent-of-origin analysis.

All tests were performed with the -wega parameters (to allow for environmental, polygenic, and additive major gene effect variance components). The tests were also run with 1000 permutations to generate empirical p-values, a procedure which does not assume that the quantitative traits are distributed normally (Abecasis *et al.*, 2000). However, where results are reported for the parent-of-origin effects analyses in Chapter 4, alleles with $P \leq 0.05$ on the test of significant differences between paternal and maternal alleles were reported, even if the empirical p-value for that specific test is greater than 0.05. The false discovery rate (FDR) method (Benjamini and Hochberg, 1995) was used to control for the proportion of false positives among the significant results, thus, the proportion of false positives among the significant associations is limited, rather than the proportion of false positives among all association tests. The QVALUE tool, which employs the methodology described in (Storey, 2002), was used to generate q-values (the minimum FDR for the study at which a given association may be called significant) for the alleles of each locus separately (with $\lambda=0$). Results could then be filtered by q-value: if a threshold of 0.05 is

chosen, then the interpretation would be that 5% of the associations with q≤0.05 are expected to be false positives. Alternatively, a p-value threshold could be used, in which case, the corresponding q-value would be the expected proportion of false positives among associations with that p-value or a lower p-value.

2.6.3: Case-control Association Analyses for HLA Data and SNPs in the HLA Region

For the SNP data, PLINK was used to generate p-values in a case-control study design which was based on a chi-squared test.

Two-sided Fisher's exact test p-values for differences in the allele counts of grouped HLA types between SLIC probands and POBI controls were calculated with the False Discovery Rate Calculator, which can handle multiple Fisher's exact test 2×2 contingency tables (Carlson *et al.*, 2009). The contingency tables for each HLA type included the allele/type count in probands, the allele/type count in controls, the allele count of the other alleles/types in probands, and the allele count of the other alleles/types in controls. This tool was also used to calculate Fisher's exact test p-values for differences between variant allele counts in SLI probands and the general population. QVALUE was used to generate q-values for the associations.

2.6.4: Estimation of the Strength of Association for Associations with HLA Alleles

Originally, relative risks (RR) (the increase in the risk of having a disease as a result of having a risk allele) were calculated for alleles that had been shown to be significantly associated with increased or decreased risk in the case-control analysis. Individuals who only had one allele specified for a given HLA locus (due to the other's being removed for not having a quality score of at least 0.9) were taken out of the analysis altogether (unless they carried the allele for which the RR was being calculated), for the loci for which they did not have two known alleles. However, odds ratios were more appropriate for this kind of analysis, and, therefore, odds ratios were calculated as well (see Addendum to Chapter 4). The calculations of the relative risks, odds ratios and confidence intervals followed the method described in (Morris and Gardner, 1988).

2.6.5: Quantitative Association Analysis for Gene Expression

MERLIN (Abecasis *et al.*, 2002) was used for performing a quantitative association analysis with gene expression data in the MEX HapMap3 population. The analysis involved testing for association between a risk allele and levels of expression of various transcripts. The gene expression data were reported in (Stranger *et al.*, 2012), and they were also made available online (http://www.ebi.ac.uk/arrayexpress/experiments/E-MTAB-264).

2.6.6: Case-control Association Analyses for Related Individuals

A case-control association analysis for related individuals specifically within the Robinson Crusoe Island cohort was performed with MQLS (more powerful/modified quasi-likelihood score) and MQLS-XM (for markers which are on the X chromosome) (Thornton and McPeek, 2007, Thornton *et al.*, 2012). The MQLS method takes into account kinship and inbreeding coefficients and corrects the association test statistic accordingly (which allows using related individuals in the analysis). Moreover, it can handle very complex pedigrees; in the case of the Robinson Crusoe Island cohort, the pedigree consisted of over a few hundred people, most of whom were connected to each other. In contrast, currently available linkage analysis software would require such a pedigree to be cut into several smaller pedigrees. Another advantage to using MQLS with this cohort is that its power is quite robust to misspecification of the disease prevalence. Prevalence values of 25% for males, and 27% for females were used, as estimated from the cohort of SLI children or normal-language children aged 3 to 8. The prevalence of SLI in our entire sample (that included mostly islanders but also outsiders) was close to that, namely, 34%. No previous registry data was available, which meant that the prevalence estimates from our sample had to be used. However, the robustness of MQLS to this parameter, as shown through a simulation analysis (Thornton and McPeek, 2007) ensures that the p-values would not change dramatically even if the prevalence were slightly different.

The raw genotype data were processed with the Affymetrix Genotyping Console software. As recommended by Affymetrix, samples with a dish quality score of less than 0.82 were removed. The dish quality score is a measure of the extent of which signal values are

separated from background values across pre-selected sites with no polymorphisms, with 0 indicating no separation, and 1 indicating a perfect separation. Markers and samples with call rates below 95% (per either the LAT array or the custom array) were removed. The genotype data were examined with the Affymetrix SNPolisher script, which provides several quality scores for the clustering, including: Fisher's linear discriminant. heterozygous cluster strength offset, and homozygote ratio offset. As recommended in the Affymetrix guidelines, markers were removed if their scores did not reach the following threshold: 3.6 (Fisher's linear discriminant), -0.1 (heterozygous cluster strength offset), 0.3 (homozygote ratio offset, if the marker had fewer than three clusters), or -0.9 (homozygote ratio offset, if the marker had three clusters). In the custom array, several markers had more than one probe designed for them. In such cases, one probe was selected based on information supplied by Affymetrix. Markers that were SNPs were used in further quality control procedures with PLINK. Samples and SNPs with a Mendelian error rate of more than 1% were removed. Samples were checked for IBD and discordant sex information, and the pedigree was amended where necessary.

In the targeted association analysis of the linkage region on chromosome 7, SNPs were tested in PEDSTATS for their HWE p-values, and SNPs with p-values of less than 0.001 (based on 19 unrelated individuals from the large pedigree) were removed. QVALUE was used to generate q-values for the p-values.

2.6.7: Association Analysis for SNPs in SLI Candidate Genes Following Long-range Sequencing

The Fisher's exact test calculator was used to generate p-values (as described above), and QVALUE was used to generate q-values.

2.6.8: Plots

Manhattan plots were generated with an R script written by Stephen Turner, and Q-Q plots with confidence intervals were generated with an updated version of that script written by Stephen Turner and Daniel Capurso. Plots of regions surrounding association peaks were

generated with LocusZoom (Pruim *et al.*, 2010), using the 1000 Genomes EUR population data (hg19).

2.7: Work Done in Collaboration with Others

The work described in this thesis is my own original work; in cases in which I collaborated with others, their contributions are indicated below:

- The exome sequencing described in this thesis was done in Nijmegen, the Netherlands by A. Hoischen, C. Gilissen, and J. Veltman (Radboud University). Funding was provided by S. E. Fisher (Max Planck Institute for Psycholinguistics). S. Chen, C. Francks, and S. E. Fisher (Max Planck Institute for Psycholinguistics) performed the variant calling and data analysis for the exome sequencing of SLI probands.
- Quality control for the genotype array data was performed by D. F. Newbury (Wellcome Trust Centre for Human Genetics, University of Oxford), and me. Quality control for samples was performed by D. F. Newbury.
- Some of the primers for the validation of the exome sequencing variants in the Robinson Crusoe Island cohort were designed by D. F. Newbury, and some were designed by me.
- Some of the primers for the long-range PCRs for candidate genes for SLI were designed by N. H. Simpson (Wellcome Trust Centre for Human Genetics, University of Oxford), and some were designed by me. PCRs were performed by N. H. Simpson. The sequencing was performed at the High-Throughput Sequencing facility at the Wellcome Trust Centre for Human Genetics, University of Oxford. Analysis of raw data, variant-calling, annotation, statistical analyses and validation of the results and all subsequent work which pertains to this experiment and is described in this thesis were done by me.
- Validation of the *NFXL1* variant in Chilean controls was done by D. F. Newbury.
- Validation of *ZNF423* variants (in SLI probands) identified through exome sequencing was carried out by A. Tchaconas, an undergraduate student who was

doing a project under my supervision, and me. The validation of said variants in probands' families and all subsequent work which pertains to this experiment and is described in this thesis were done by me.

- In the case-study of the Spanish-speaking child who has an inversion and translocation, the cytogenetic work discussed in this thesis (including RNA extraction) was carried out by D. Moralli (Wellcome Trust Centre for Human Genetics, University of Oxford). The mapping of the breakpoint near *FOXP2* was carried out by M. T. M. Chan, who was doing a project at the lab under my supervision, and me. I conceptualised the project, M. T. Chan designed the primers, and the PCRs, sequencing reactions, and analysis of the sequencing results were carried out by both of us.

Sample collection and phenotyping for the SLIC cohort were done by the SLI Consortium: G. Baird, V. Slonims, K. Dworzynski (Newcomen Centre, Guy's Hospital, London); P. F. Bolton (Department of Child and Adolescent, Institute of Psychiatry, King's College London); E. Simonoff (Department of Child and Adolescent Psychiatry, Institute of Psychiatry, King's College London); A. O'Hare (Department of Reproductive and Developmental Sciences, University of Edinburgh); J. Seckl (Molecular Medicine Centre, University of Edinburgh); H. Cowie (Department of Speech and Language Therapy, Royal Hospital for Sick Children, Edinburgh): A. Clark, J. Watson (Speech and Hearing Sciences, Queen Margaret University); W. Cohen (Department of Educational and Professional Studies, University of Strathclyde); A. Everitt, E. R. Hennessy, D. Shaw, P. J. Helms (Department of Child Health, University of Aberdeen); Z. Simkin, G. Conti-Ramsden (School of Psychological Sciences, University of Manchester; D. V. M. Bishop (Department of Experimental Psychology, University of Oxford); A. Pickles (Department of Biostatistics, Institute of Psychiatry, King's College London).

Sample collection and phenotyping for the Robinson Crusoe Island cohort were done by: P. Villanueva (Faculty of Medicine and Faculty of Dentistry, University of Chile); M. A.

Fernández, L. Jara, Z. De Barbieri (Faculty of Medicine, University of Chile); H. Palomino (Faculty of Dentistry, University of Chile).

The Colombian samples were supplied by: L. Carvajal-Carmona (WTCHG, University of Oxford); M. M. Echeverry (Department of Biology, University of Tolima).

Chapter 3: A Genome-wide Association Study of SLI

As discussed in Chapter 1, previous genome-wide studies of SLI only utilised linkage methods, and, furthermore, they identified different loci depending on the method used, the type cohort and the selection criteria for the language impairment. As has been shown, association methods may be used to analyse complex traits (see Chapter 1 for discussion thereof). In this chapter I will describe several genome-wide association analyses of SLI that I performed. Before I performed the analyses described in this chapter, Dianne F. Newbury performed preliminary association analyses with the SLIC genotype data using traditional methods, which included qualitative and quantitative analyses. Those did not identify genome-wide significant associations, but they highlighted a variant found in the HLA region on chromosome 6. Since a previous study highlighted genetic interaction effects between HLA alleles and Schizophrenia, I used a statistical analysis tool called EMIM, which could model interaction effects, with the SLIC genotype data to see if such an effect could be detected with SNPs in the HLA region. As EMIM could model other types of genetic effects, including child genetic effects, maternal genetic effects, and parent-of-origin effects, I used it to do a genome-wide screen of SLI. Importantly, the advantage of this approach is that it provides greater power to detect such effects, when they are present, than the power that traditional case-control analyses have in those cases (Ainsworth *et al.*, 2011). Furthermore, parent-of-origin effects have been shown to be relevant to the issue of the heritability of complex traits, and when parent-of-origin effects are incorporated into the model, it can increase the power to detect genetic risk variants that exhibit such effects (Kong *et al.*, 2009, Mott *et al.*, 2014).

Affection status was determined based on the child's performance on two components of the CELF. Details of the methodology and the samples used in this study can be found in Chapter 2. Included in my analyses were 153 case-parents trios, 54 case-mother duos, 12 case-father duos, and 18 cases from the SLIC cohort, based on the average numbers of these subsets per SNP. I also used a replication set for the maternal parent-of-origin effects analysis which consisted of samples from language-impaired children from the

ALSPAC cohort (Fraser *et al.*, 2013, Boyd *et al.*, 2013). The samples were filtered based on SLI exclusion criteria (hearing problems, PIQ <80, and diagnosis of autism), where data were available. Additionally, where data for the ethnic background of the child or parents were available, samples from children with any ethnic background (for themselves or their parents) other than "white" were excluded, as to be compatible with our sample. Cases were defined as children who had a pragmatic composite score of at least 132 from the Children's Communication Checklist (Bishop, 1998), which was found to be typical of children with SLI (Bishop, 1998), and, additionally, either a low score on a receptive language test: 1.5 SD or more below the sample mean (after preliminary filtering, 7.58) on the Wechsler Objective Language Dimensions (Rust, 1996), or a low score on two indices of language structure (Speech and Syntax) representing expressive language, from the Children's Communication Checklist: 1.5 SD or more below the sample means (after preliminary filtering, 35.56 and 31.9, respectively). In total, there were 313 cases in the replication set (excluding those without genotypes). Genotype data from the children and their mothers for 6 SNPs were used (paternal genotypes were not available), and, on average, there were 193 case-mother duos, 75 cases, and 45 mothers of cases per SNP.

3.1: Results of the Association Analyses

3.1.1: Results of the EMIM Analyses

Nine association analyses were performed using 614,937 SNPs: child and maternal genetic effects, child and maternal trends, paternal and maternal parent-of-origin effects, maternal-foetal genotype incompatibility (under two models), and maternal-foetal genotype matching. Descriptions of the tests can be found in Chapter 2. Manhattan plots and Q-Q plots with confidence intervals are presented in Figures 3.1 to 3.9. Manhattan plots mark association on chromosomes with alternating colours for clarity. The Q-Q plots show the $-\log_{10}(P)$ of all associations (y-axis) plotted against the distribution expected under the null hypothesis of no association (x-axis). The diagonal line represents the null hypothesis (i.e. observed values correspond to expected values).

The results of the various tests were examined based on the following criteria:

- The p-value of a given SNP (a genome-wide significance threshold of $P \leq 5 \times 10^{-8}$)
- Whether the p-value was lower than expected (as can be inferred from a Q-Q Plot)
- The association trends of adjacent SNPs (or lack thereof)

Two analyses identified significant and suggestive associations: the paternal and maternal parent-of-origin effects analyses. The p-values of the SNPs in those clusters can be found in Table 3.1, and zoomed-in plots of the peaks can be found in Figure 3.10 and Figure 3.11. The other analyses did not find suggestive or significant associations. The top associations in those analyses are given in Table 3.2.

Table 3.1: Association clusters identified through analyses of parent-of-origin effects.

SNP ID	Chromosome	Position (hg18)	Test	Minor allele/major allele	I_m/I_p (relative to risk allele)	p-value
rs1353835	5	39784227	Maternal parent-of-origin effects	C*/A	2.14	2.55×10^{-5}
rs1994882	5	39841921		C*/A	2.441	1.56×10^{-6}
rs12658486	5	39841974		G*/A	2.367	1.43×10^{-5}
rs980306	5	39852592		A/G*	3.008	3.91×10^{-7}
rs10447141	5	39852924		A/G*	3.08	$\mathbf{1.16\times10^{-7}}$
rs17194068	5	39857074		G/A*	2.86	1.04×10^{-6}
rs6895329	5	39861497		G/A*	2.926	1.02×10^{-6}
rs1816088	5	39897583		A/C*	2.486	4.72×10^{-5}
rs618051	5	39902670		C/A*	2.429	4.62×10^{-5}
rs542708	5	39968025		A/G*	2.453	8.42×10^{-5}
rs17218399	5	40061719		G/A*	2.4	7.54×10^{-5}
rs2939378	5	40078002		G/A*	2.29	3.43×10^{-5}
rs1567010	5	40086058		G/A*	2.628	9.93×10^{-6}
rs11158632	14	23839502	Paternal parent-of-origin effects	C/A*	3.842	4.62×10^{-8}
rs4280164	14	23841124		A/G*	3.872	$\mathbf{3.74\times10^{-8}}$
rs2144494	14	23843226		G/A*	3.832	4.94×10^{-8}
rs2281472	14	23845685		G/A*	3.421	1.12×10^{-7}
rs3181384	14	23856815		A/G*	3.016	1.58×10^{-6}

The top association in each analysis is in boldface. * Risk Allele.

Table 3.2: Top associations in analyses which did not test for parent-of-origin effects.

Test	SNP	Chromosome	Position (hg18)	p-value	Gene
Child genetic effects	rs10026666	4	99770930	2.79×10^{-6}	*TSPAN5*
Child trend	rs4790018	17	74876673	1.17×10^{-6}	*RBFOX3*
Maternal genetic effects	rs2100346	8	81004527	3.18×10^{-6}	*MRPS28*
Maternal trend	rs9309457	2	71587449	1.02×10^{-6}	*DYSF*
Maternal-foetal genotype incompatibility test – null allele and antigen-coding allele	rs3800471	6	34589118	4.90×10^{-6}	*PACSIN1*
Maternal-foetal genotype incompatibility test – two antigen-coding alleles	rs2813894	7	25602904	4.75×10^{-6}	*LOC646588*
Maternal-foetal genotype matching test	rs11758326	6	34590476	6.94×10^{-7}	*PCASIN1*

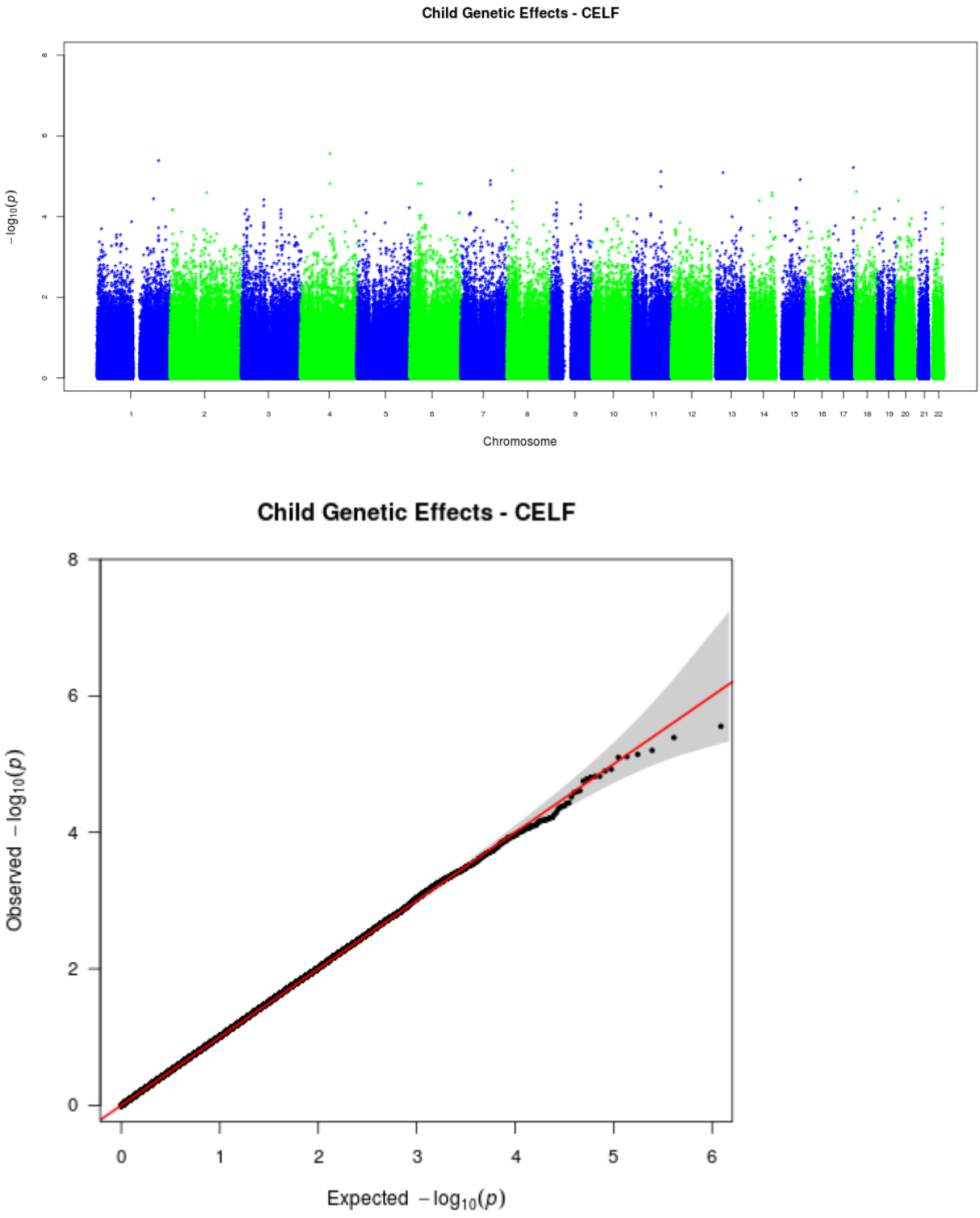
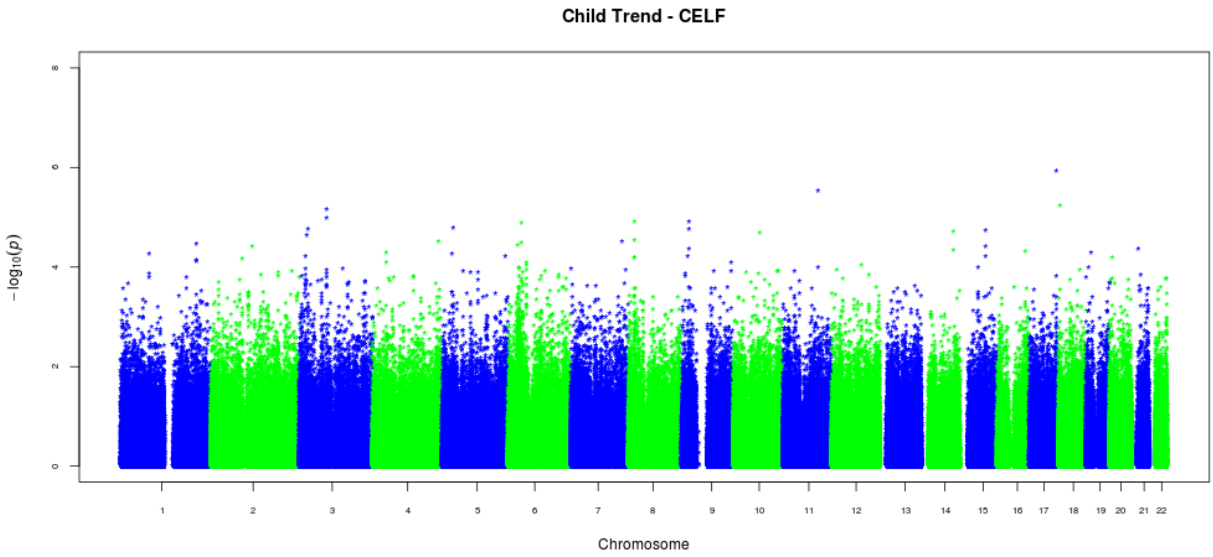


Figure 3.1: Child genetic effects, Manhattan and Q-Q plots (with 95% confidence intervals in grey).



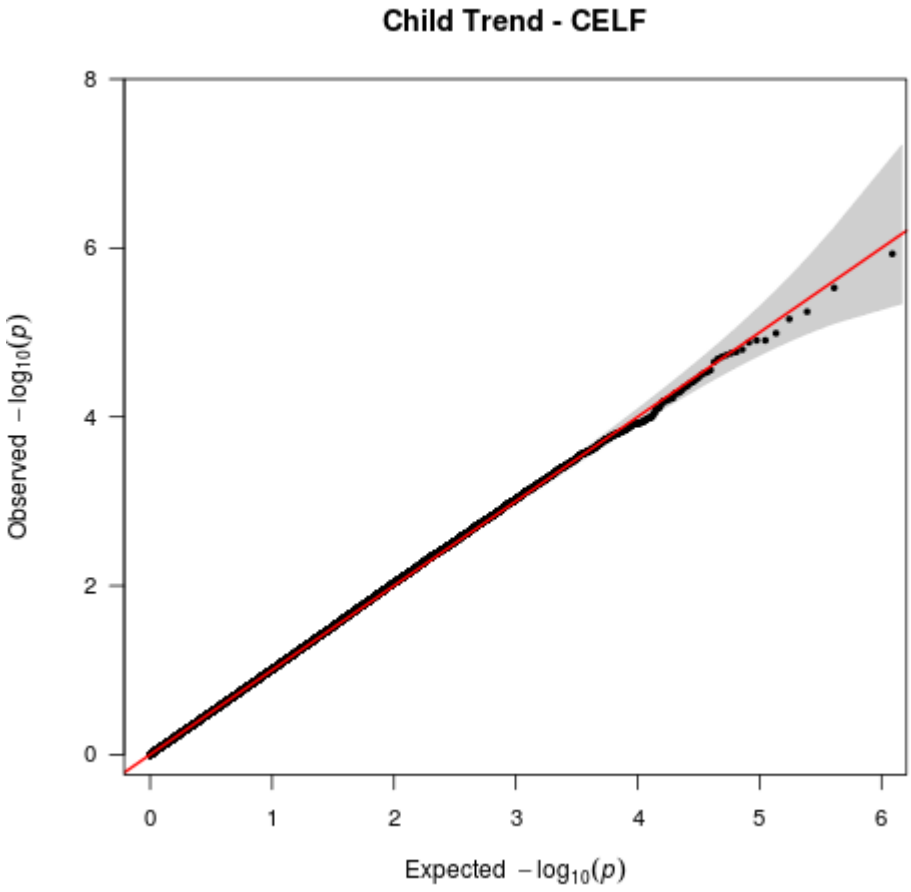


Figure 3.2: Child trend, Manhattan and Q-Q plots (with 95% confidence intervals in grey).

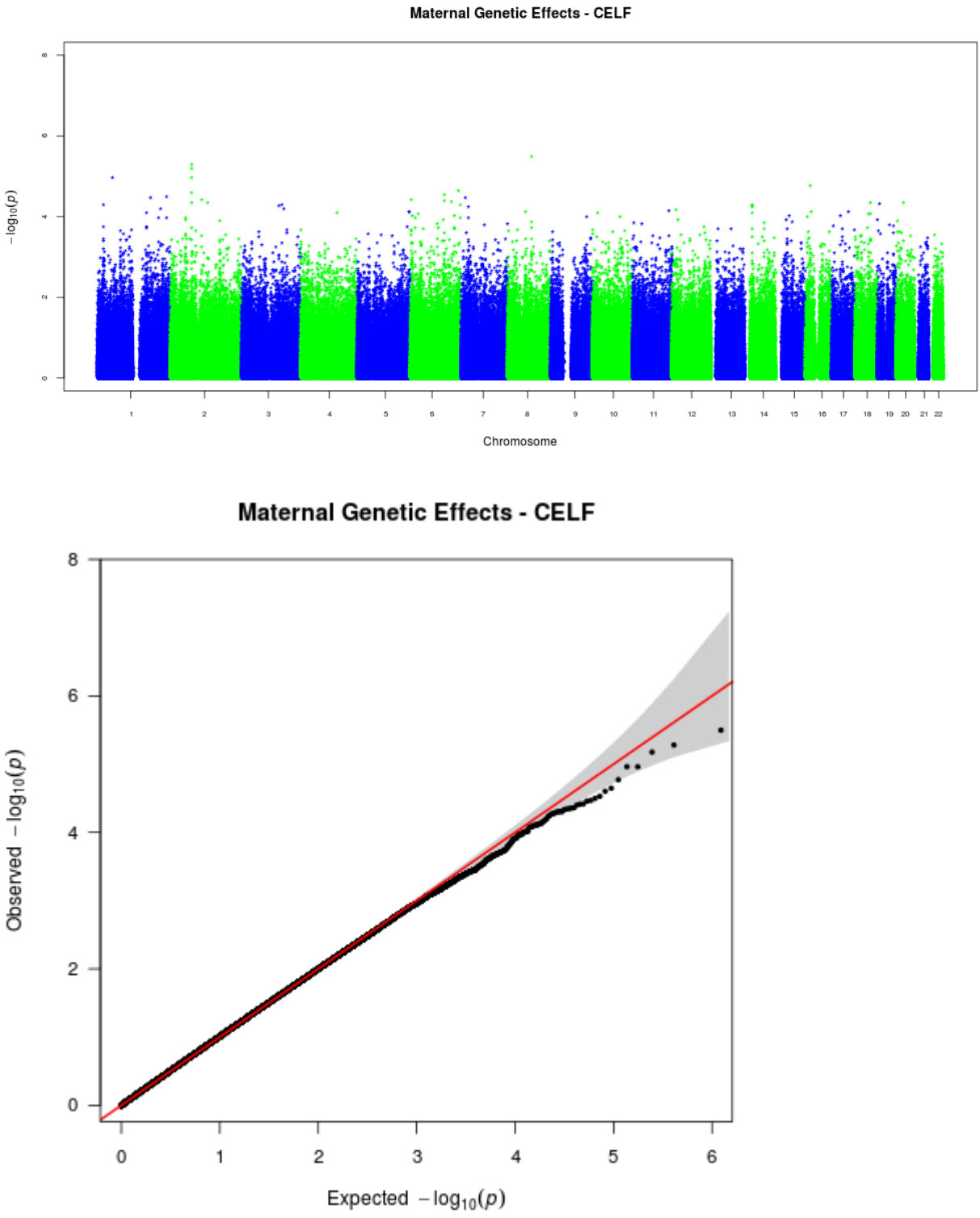


Figure 3.3: Maternal genetic effects, Manhattan and Q-Q plots (with 95% confidence intervals in grey).

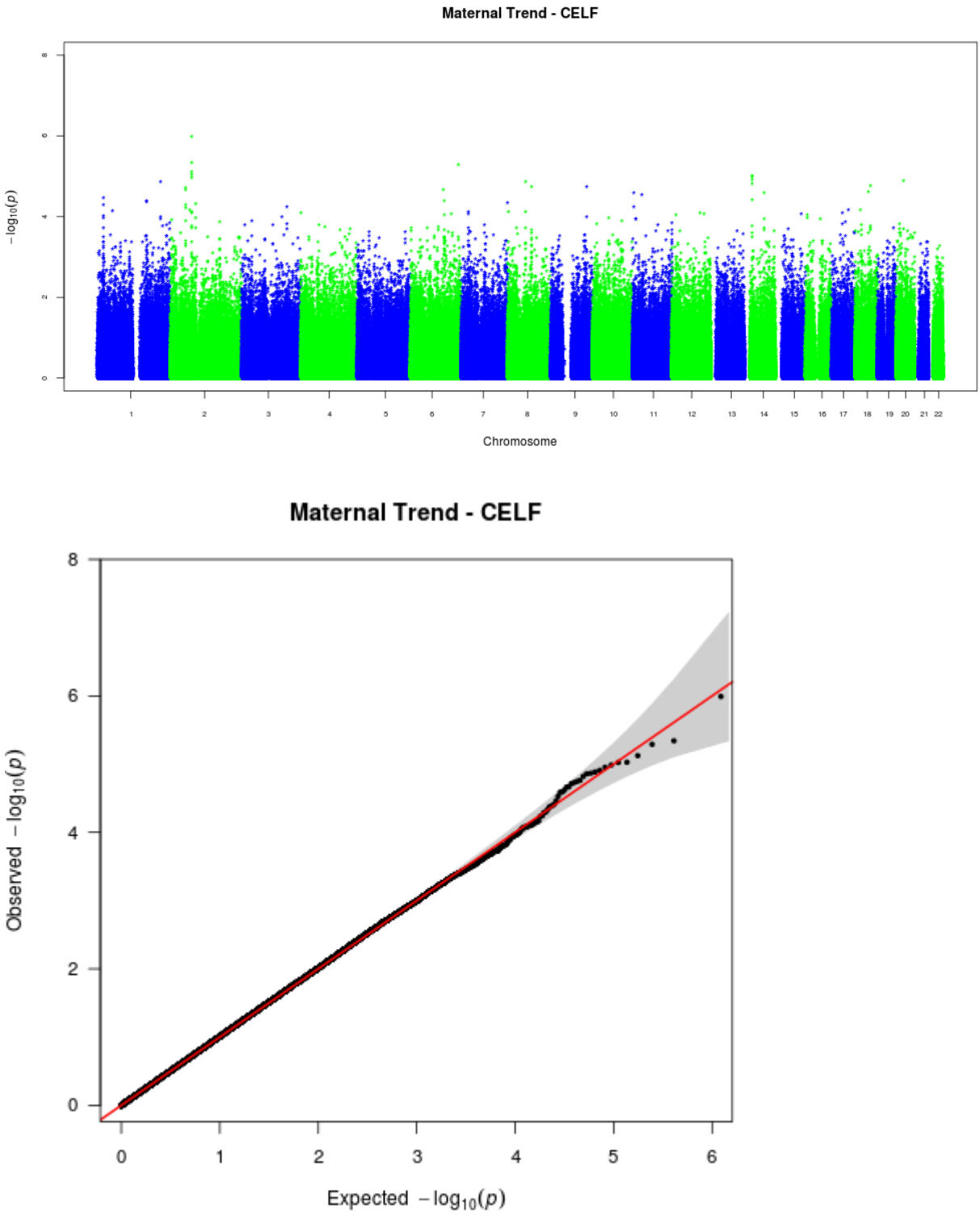


Figure 3.4: Maternal trend, Manhattan and Q-Q plots (with 95% confidence intervals in grey).

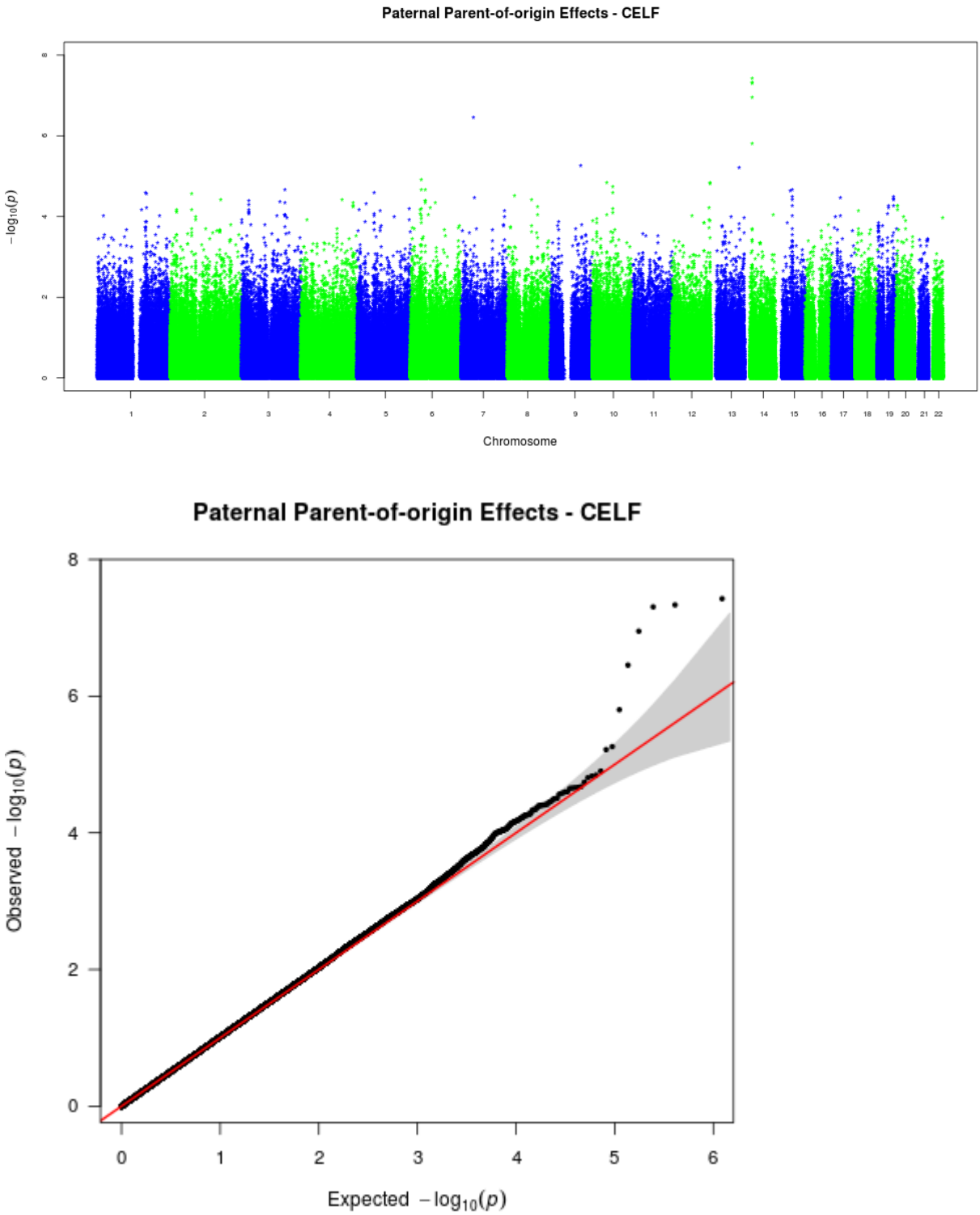


Figure 3.5: Paternal parent-of-origin effects, Manhattan and Q-Q plots (with 95% confidence intervals in grey).

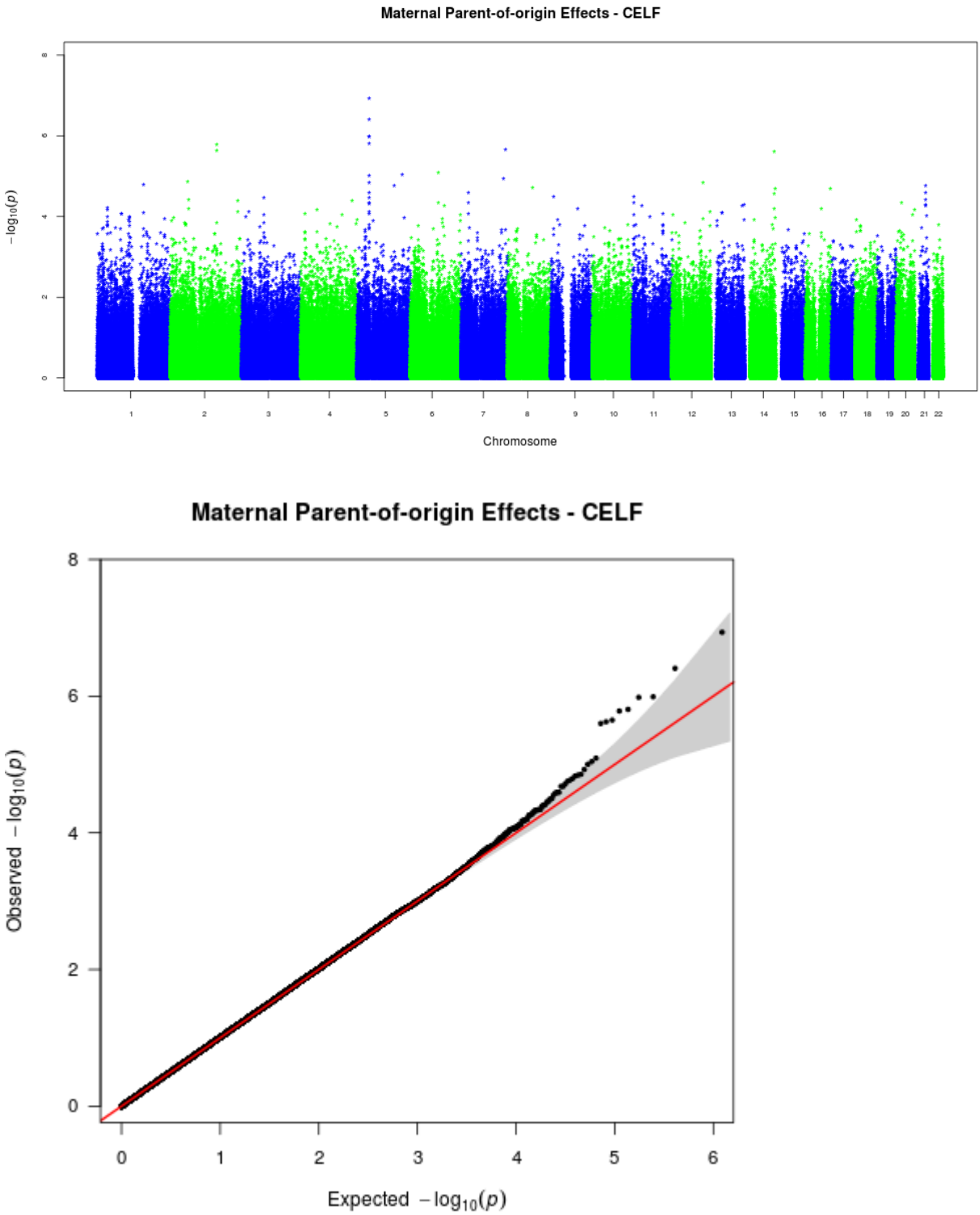


Figure 3.6: Maternal parent-of-origin effects, Manhattan and Q-Q plots (with 95% confidence intervals in grey).

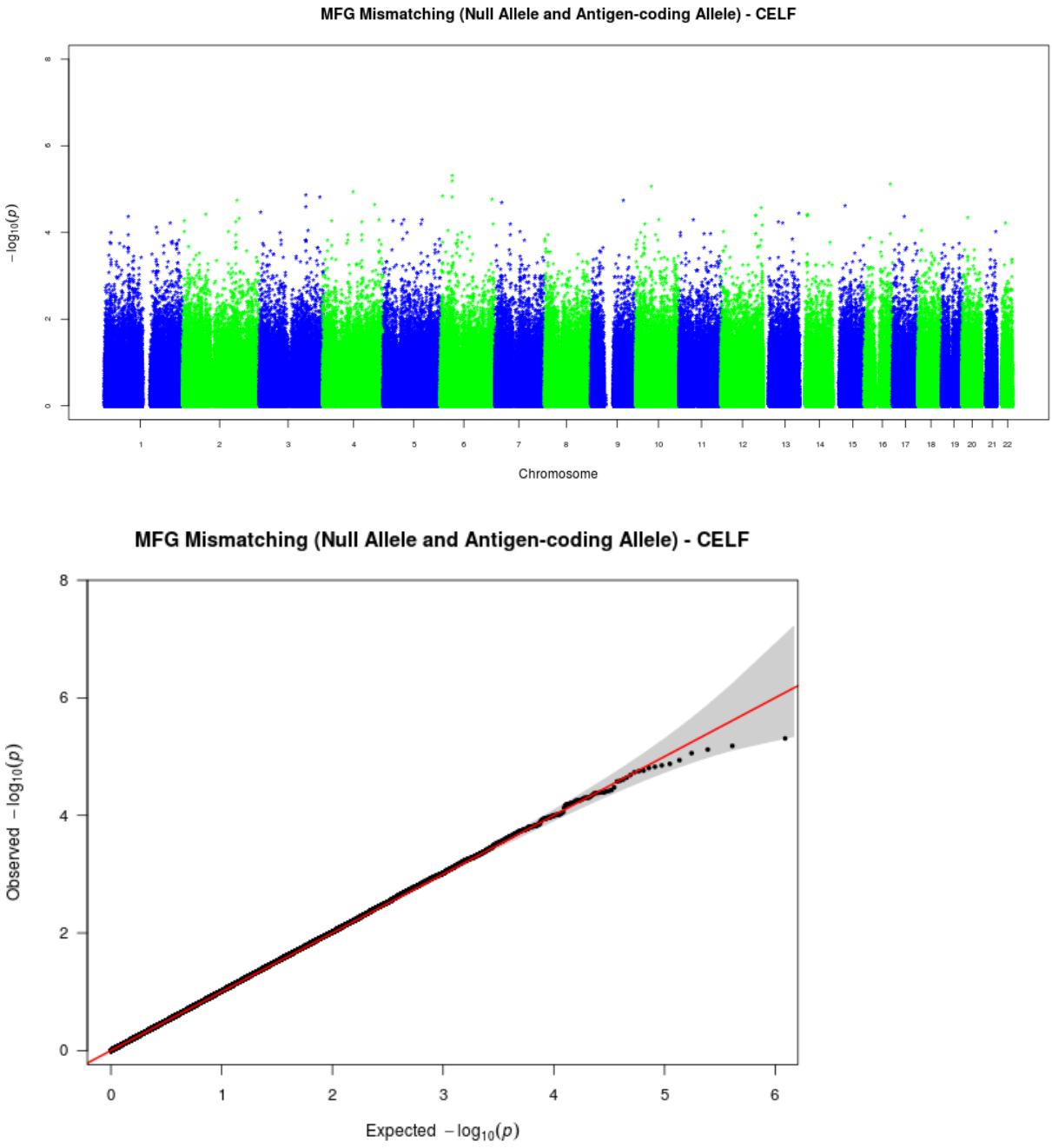


Figure 3.7: Maternal-foetal genotype incompatibility test (null allele and antigen-coding allele), Manhattan and Q-Q plots (with 95% confidence intervals in grey).

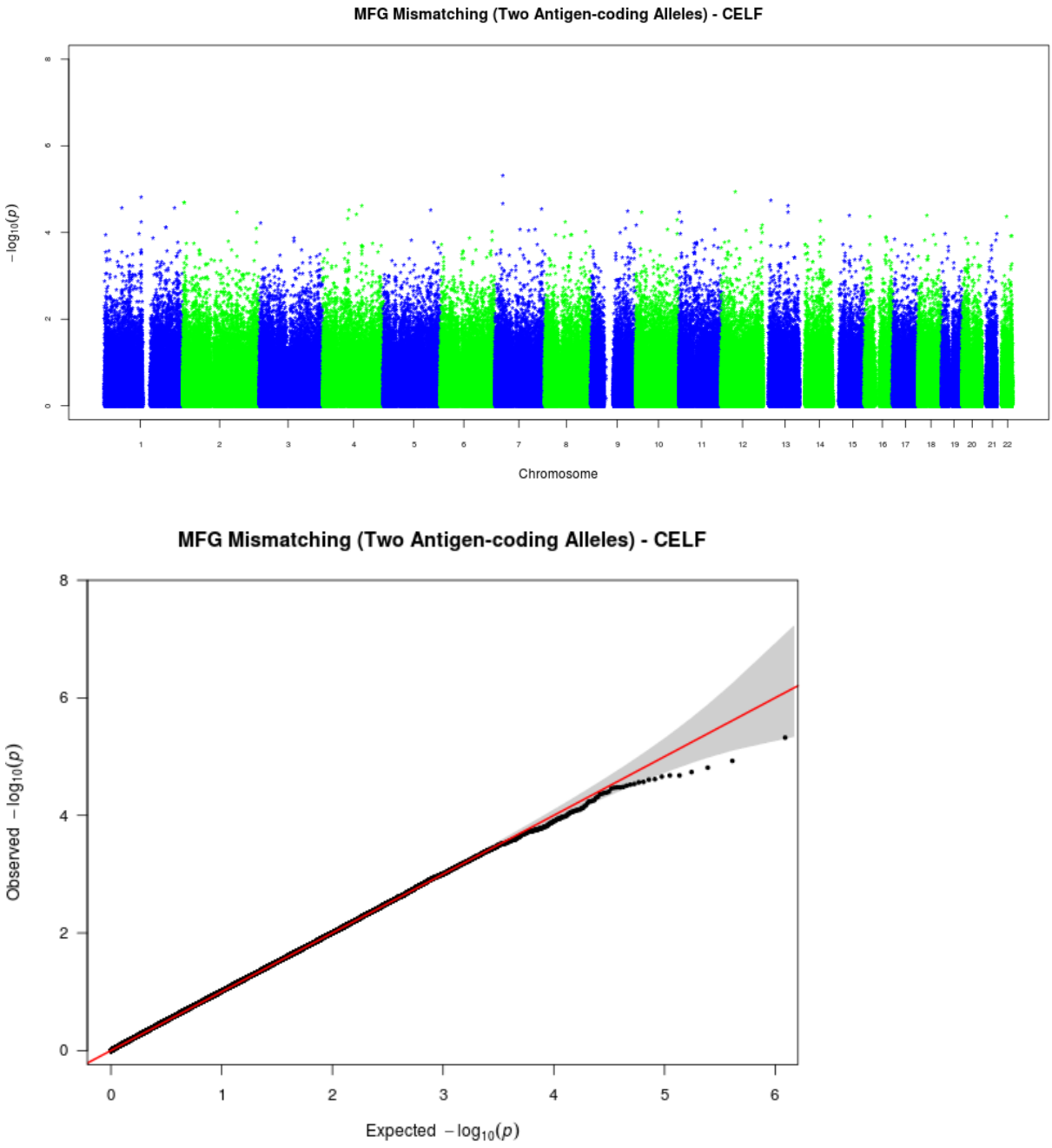


Figure 3.8: Maternal-foetal genotype incompatibility test (two antigen-coding alleles), Manhattan and Q-Q plots (with 95% confidence intervals in grey).

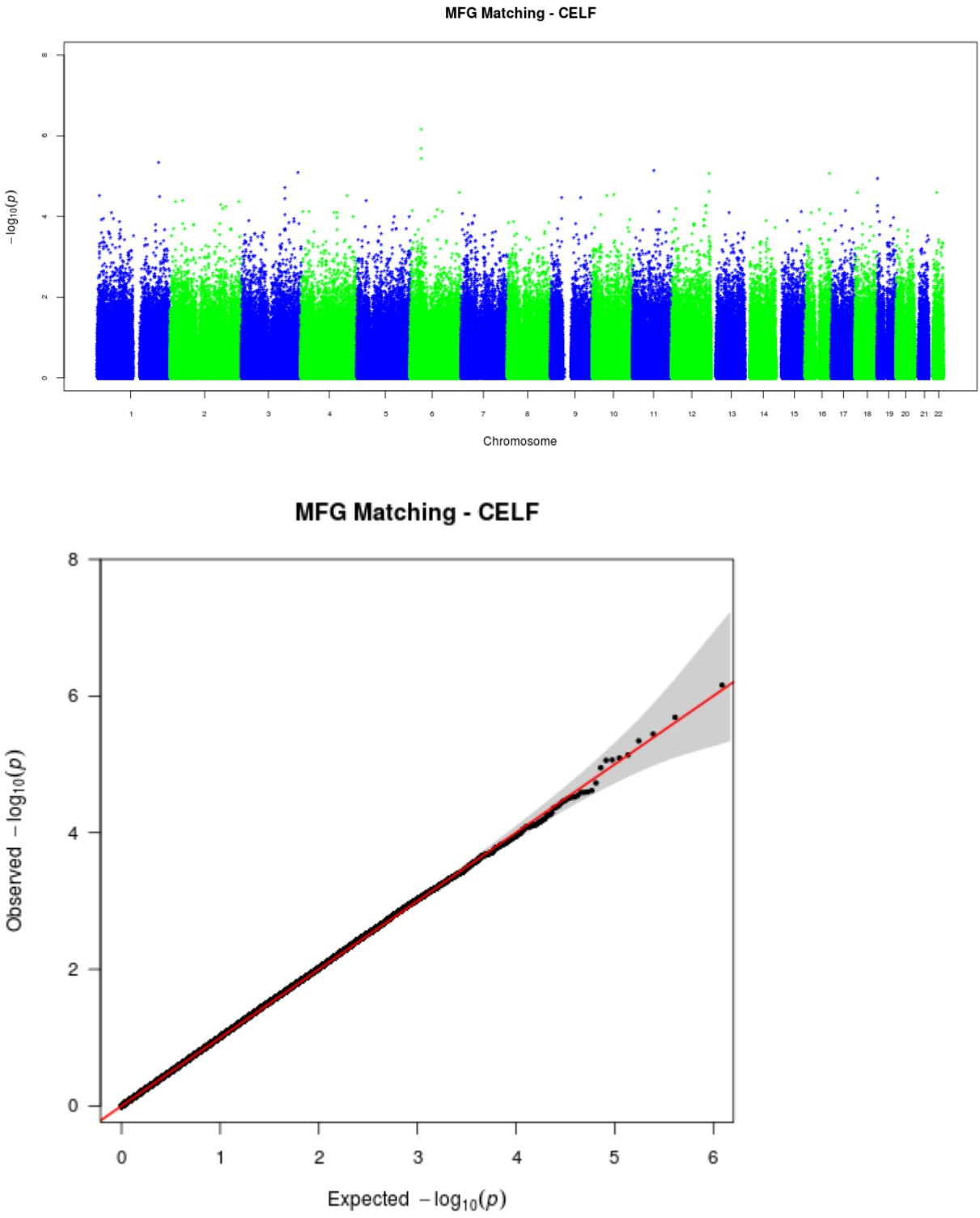


Figure 3.9: Maternal-foetal genotype matching test, Manhattan and Q-Q plots (with 95% confidence intervals in grey).

As can be learned from the plots, two of the nine analyses identified regions that showed association with the affection status. The clusters of SNPs around the top SNPs in 5p13 (in the maternal parent-of-origin effects analysis) and 14q12 (in the paternal parent-of-origin effects analysis) also show association trends.

The most highly-associated SNP on chromosome 14 (rs4280164, MAF of 0.222) reached genome-wide significance with a p-value of 3.74×10^{-8}. This SNP falls within a gene called nucleolar protein 9 (*NOP9*). Furthermore, it is a missense variant. The SNP was associated with paternal parent-of-origin effects. The I_p parameter estimated for the top SNP on chromosome 14 was 3.872 (with 95% confidence intervals of 2.235 and 6.707). The most highly-associated SNP on chromosome 5 (rs10447141, MAF of 0.326) had a p-value of 1.16×10^{-7}; it does not fall within any known gene. The SNP was associated with maternal parent-of-origin effects. The I_m parameter estimated for the top SNP on chromosome 5 was 3.08 (with 95% confidence intervals of 1.975 and 4.801). The most highly-associated SNPs in the other analyses did not meet the aforementioned criteria and so were not followed up any further.

3.1.2: Results of the Replication Analysis for the Maternal Parent-of-origin Effects Analysis

The analysis included the six SNPs that fell within the region of association on chromosome 5 in maternal parent-of-origin effects analysis. The most highly-associated SNP in my analysis was not included in the SNP array on which the ALSPAC samples had been genotyped and was therefore not available for the replication analysis. The SNPs used were: rs1994882, rs12658486, rs17194068, rs618051, rs2939378 and rs1567010, and the p-values were: 0.001, 0.002, 0.027, 0.152, 0.561 and 0.329, respectively. All associations except for the one with rs2939378 were in the opposite direction compared with SLIC (i.e. the risk alleles were different).

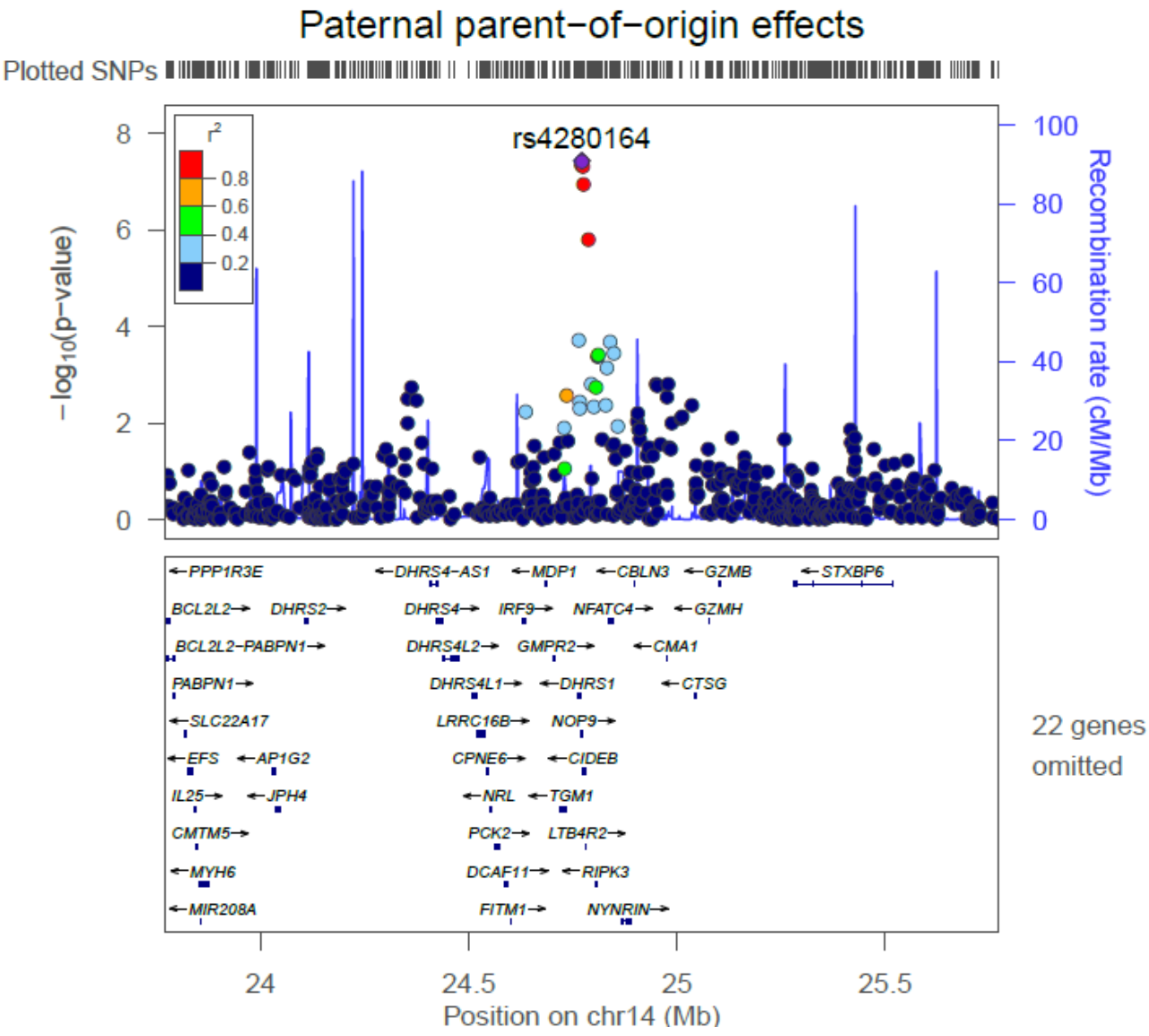


Figure 3.10: The association peak on chromosome 14 (plotted with LocusZoom). The top association (the index SNP) is marked by a purple square overlaid with a circle. LD r^2 values are measured between all SNPs and the index SNP. The direction of transcription is indicated by an arrow next to the name of the gene.

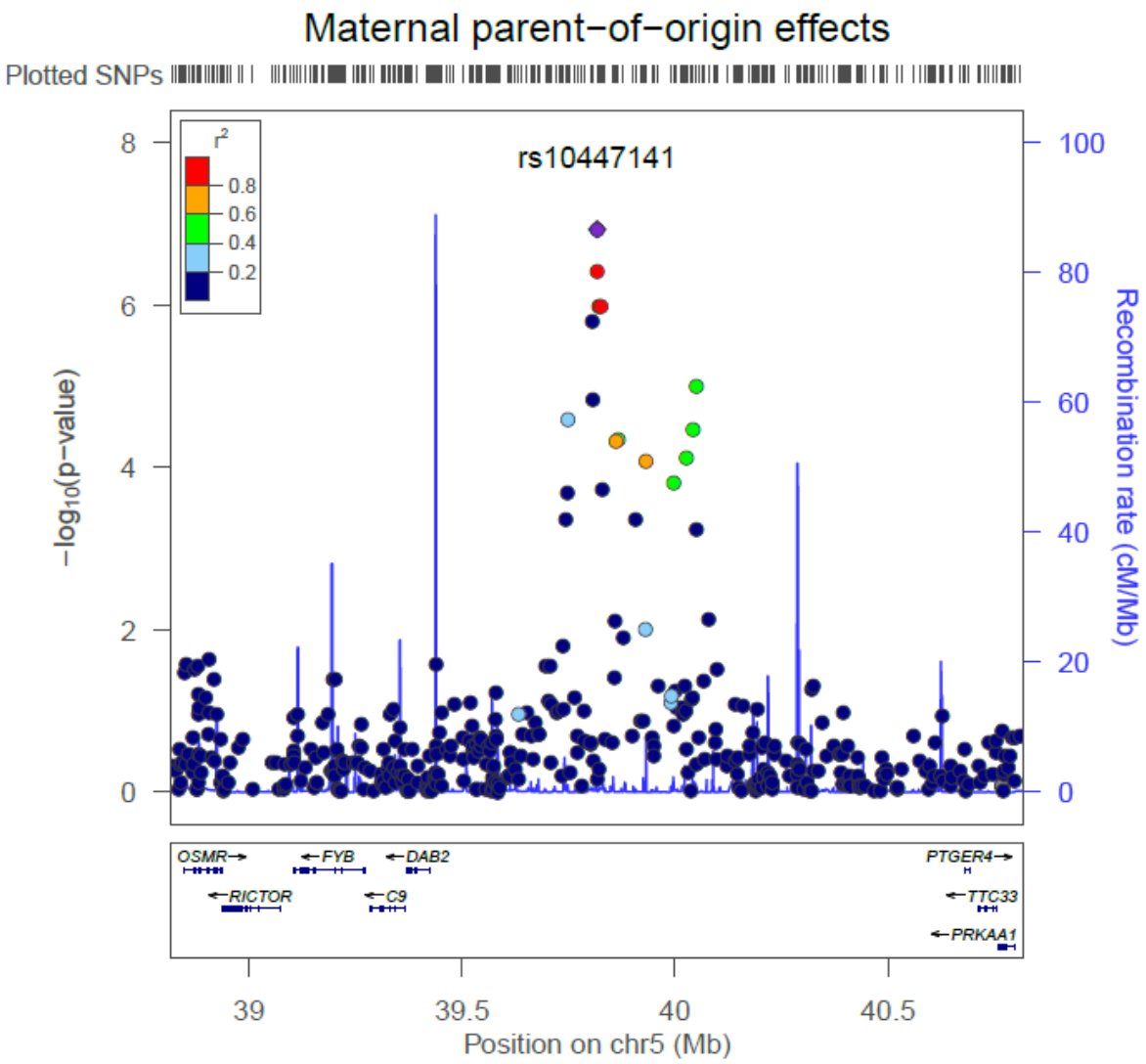


Figure 3.11: The association peak on chromosome 5 (plotted with LocusZoom). The top association (the index SNP) is marked by a purple square overlaid with a circle. LD r^2 values are measured between all SNPs and the index SNP. The direction of transcription is indicated by an arrow next to the name of the gene.

3.2: Discussion of the GWAS Results

In terms of the significance of the results described above, three of the SNPs on chromosome 14 reached genome-wide significance ($P \leq 5 \times 10^{-8}$). However, as discussed in Chapter 1, this threshold was calculated for a case-control association study design assuming 10^6 independent tests, which is not the case here. Under the three other systems discussed in Chapter 1, which have thresholds of $P < 2 \times 10^{-7}$ (for the first category of association), $P \leq 5 \times 10^{-5}$, and $P \leq 5 \times 10^{-7}$ (in the Wellcome Trust Case Control Consortium study), the associations with the top SNPs on chromosome 5 and chromosome 14 would both be considered significant. While these thresholds are corrected for the number of SNPs, they are not corrected for the number of analyses (rather than the number of tests per analysis). The p-values reported here are not corrected for the number of analyses performed (nine in total). This is due to the fact that we tested for different hypotheses and did not have a global hypothesis, as discussed in (Krawczak, 2006). Most of the analyses, including the ones which identified suggestive and significant associations, used different parameters. However, the child genetic effects and child trend analyses had one parameter in common, as were the maternal genetic effects and maternal trend analyses. Therefore, the associations in those analyses may have to survive a more stringent correction.

Interestingly, both the analysis which found the association on chromosome 5 and the one which found the association on chromosome 14 tested for parent-of-origin effects. Such effects have been described for other neurodevelopmental disorders, including: autism (Schanen, 2006), Prader-Willi syndrome and Angelman syndrome (Chamberlain and Lalande, 2010). Imprinted genes expressed in the brain are involved in neuronal differentiation and gene regulation, among other processes (Wilkinson *et al.*, 2007). Each of the top associations in both paternal and maternal parent-of-origin analyses were tested through comparing the effects of one type to those of the other type; two analyses were run in each case: one testing for both types of parent-of-origin effects, and one testing only for the effects of the other type. Thus, when validating the paternal parent-of-origin effects, for example, I used a null model in which maternal parent-of-origin effects operate and calculated twice the difference between the log-likelihood of this model and of a model in which both paternal and maternal parent-of-

origin effects are estimated. In other words, the test allowed for maternal parent-of-origin effects when testing for paternal parents of origin effects. In those analyses, the top association in the paternal parent-of-origin effects analysis had a p-value of 3.87×10^{-7}, and the top association in the maternal parent-of-origin effects analysis had a p-value of 1.29×10^{-7}, indicating that neither association is driven by parental parent-of-origin effects of the other type, even when assuming those were operating. In a recent study of the genetic architecture of complex traits in the mouse, it was shown that the expression of non-imprinted genes nonetheless displayed parent-of-origin effects, possibly through interaction with imprinted loci (Mott *et al.*, 2014). This implies that even if the regions identified in my analyses are not imprinted themselves, the parent-of-origin effects detected here may be originating through interaction with other loci. That said, an analysis of cases of uniparental disomy (the state of having inherited both copies of the same type of chromosome from only one parent) identified a region on chromosome14q11-14q12 (which includes the position of the highly-associated SNP on 14q12 from my analyses) which is likely to be imprinted (Kotzot, 2001). The analysis compared cases of isodisomy (in which two copies of the same parental homologue were inherited) and heterodisomy (in which both homologues for the same chromosome were inherited) on chromosome 14: homozygosity of autosomal recessive alleles is likely to be the result of isodisomy. If it is nonetheless heterodisomy (in which homozygosity of such alleles is less likely) that is present in the same region across cases, then this might suggest that the regions are imprinted. The cases analysed in that study were all maternal uniparental disomy. Therefore, it is not clear whether the phenotypes associated with that condition, which included dysmorphic features, malformations and developmental delay, were the result of an over-expression of the maternal allele or the lack of expression of the paternal allele. In other words, it is not clear whether the region is paternally imprinted (in which case only the maternal allele is expressed) or maternally imprinted (in which case only the paternal allele is expressed). If it were the former, then the maternal uniparental disomy would result in an over-expression of the allele, and, if it were the latter, then lack of expression of the paternal allele would be the outcome. However, the heterodisomy region on 14q12 identified in that study corresponds to cases of pathological paternal deletions of 14q11-14q13 associated with small size at birth (Sutton and Shaffer, 2000), which suggests

that the paternal allele is normally expressed. That is in line with my results in that paternal parent-of-origin effects were observed on 14q12.

3.2.1: The Association with SNPs on Chromosome 14

The paternal parent-of-origin effects association peak (i.e. the cluster of SNPs containing the most highly-associated SNP and the adjacent SNPs showing association trends) lies in chromosomal band 14q12. This region (OMIM#611095) has been implicated in a study of intellectual disability which used homozygosity mapping as a means of identifying regions involved in that disorder (Abou Jamra *et al.*, 2011). A close region on 14q12 was implicated in a linkage study of ADHD, but the top SNP in my analyses was ~432 kbp away from the linkage peak in that study (Romanos *et al.*, 2008). Similarly, it was ~2738 kbp away from a linkage region identified in a linkage study of intellectual disability (Najmabadi *et al.*, 2007).

The top SNP in the peak falls within the *NOP9* gene. As previously stated, the SNP is a missense variant (S308N), and it is predicted by PolyPhen to be "possibly damaging". The amino acid substitution is found five amino acids away from one of the protein's pumilio domains. In my analysis, the minor allele, which is the missense variant, was associated with decreased risk. It may be that the prediction is not correct, or that the mutation is a gain-of-function mutation. The Nop9 protein (the yeast orthologue of NOP9) is RNA-binding and is involved in the nuclear maturation of the ribosomal subunits in the yeast (Thomson *et al.*, 2007). A study of peripheral blood gene expression in the context of schizophrenia found that the *NOP9* gene was significantly dysregulated in schizophrenia patients and their unaffected siblings, compared to controls (Glatt *et al.*, 2011).

3.2.2: The Association with SNPs on Chromosome 5

The maternal parent-of-origin effects association peak lies in chromosome 5p13. Chromosome 5p13 has been implicated in several linkage studies of autism and ADHD, and a locus designated ADHD4 (OMIM#608906) falls therein. As discussed in Section 1.8.3, a genome-wide linkage study of ADHD found strong linkage to 5p13, with the most strongly linked marker being D5S418, which is ~200 kbp proximal to rs10447141 (Ogdie *et al.*, 2003).

Further evidence for this was obtained through a fine-mapping of the region (Ogdie *et al.*, 2004). This region was the only significant common risk locus in a pooled analysis of two separate ADHD cohorts (Ogdie *et al.*, 2006). Autism studies have also identified 5p13 as a risk locus. The highest linkage peak in a genome-wide linkage study of autism fell within 5p13 (Liu *et al.*, 2001). This result was obtained when using a broad phenotype for autism, which included autism, Asperger syndrome and other pervasive developmental disorders. The marker for which the linkage peak was obtained was D5S2494, which is ~400 kbp proximal to rs10447141. The same study also examined the parental contribution to the IBD sharing at marker loci, and at the above marker it was concluded that IBD sharing was mostly maternally contributed to (Liu *et al.*, 2001), which is in line with the maternal parent-of-origin effects on 5p13 identified in my analysis. A recent genome-wide screen of parent-of-origin effects in autism found a maternally-linked region on 5p13 (Fradin *et al.*, 2010). Again, this is in line with the results of my analysis.

The relation between SLI, autism and ADHD is not completely clear neither in terms of the phenotypes, nor in terms of the genetic basis for those disorders (see Chapter 1 for a detailed discussion). However, since there is evidence for some genetic overlap between SLI, autism and ADHD, it is plausible that whatever is driving the association observed here is on a genetic pathway relevant to all three disorders. It is interesting to note that for SLI and autism the association and linkage, respectively, pertain to the maternal side. While the above ADHD studies cited do not report such an effect, other studies looked into parent-of-origin effects in ADHD, with varying results. While some studies reported paternal parent-of-origin effects involved in ADHD (Hawi *et al.*, 2005, Hawi *et al.*, 2010), others reported no overall parent-of-origin effects in ADHD (Kim *et al.*, 2007, Anney *et al.*, 2008), which, in turn, does not preclude such effects acting through a gene-specific mechanism (Anney *et al.*, 2008). Indeed, a recent GWAS of parent-of-origin effects in the context of ADHD detected such effects, but not on 5p13 (Wang *et al.*, 2012).

In contrast to the top SNP on 14q12, the top SNP on 5p13 does not fall within any known gene. Instead, it lies between *PTGER4* (prostaglandin E receptor 4) and *DAB2* (disabled homolog 2, mitogen-responsive phosphoprotein). *PTGER4* has been implicated in a

study of schizophrenia in that it was significantly down-regulated in brains of schizophrenia patients (Schmitt *et al.*, 2011). It is possible that the association peak on 5p13 is driven by causal variant in LD with the top SNP and/or that the SNP has a regulatory function. The top SNP, rs10447141, has an eQTL p-value of 4×10^{-5} for the gene weakly similar to Rho GEF (*WGEF*), also called Rho guanine nucleotide exchange factor 19 (*ARHGEF19*) (OMIM#612496), in the SCAN database (obtained using HapMap CEU cell lines (Duan *et al.*, 2008)). The WGEF protein can activate proteins of the Rho-GTPases family (Wang *et al.*, 2004), which are involved in cellular signalling pathways (Van Aelst and D'Souza-Schorey, 1997). It can induce the rearrangement of cytoskeleton in the cell (Wang *et al.*, 2004) and is involved in a crucial step in embryonic development called convergent extension (Tanegashima *et al.*, 2008). Another gene from the same family, Rac/Cdc42 guanine nucleotide exchange factor 6 (*ARHGEF6*) (OMIM#300267), has been implicated in intellectual disability (Kutsche *et al.*, 2000), and it is believed that the roles of these genes in the development of dendritic structures may explain their involvement in intellectual disability (Newey *et al.*, 2005). The *WGEF* gene itself is regulated through methylation, a process which plays a major role in genomic imprinting (Horii *et al.*, 2009). Interestingly, the gene lies in a dyslexia susceptibility locus, DYX8 (OMIM#608995), found on chromosome 1p34-1p36 (Rabin *et al.*, 1993, Grigorenko *et al.*, 2001, Tzenova *et al.*, 2004).

In the replication analysis, three out of six SNPs on chromosome 5 tested had a p-value below 0.05, with the most associated one being rs1994882, P=0.001. The direction of association for rs1994882 in the ALSPAC analysis was opposite to that observed in the SLIC analysis. However, such a phenomenon may be the result of differences in the interaction between the observed variant and the causal variant in different populations (Lin *et al.*, 2007). Furthermore, the same effect was observed in a previous association study which used samples from the SLIC and ALSPAC cohorts (Newbury *et al.*, 2009). That said, as discussed in Section 1.4.4, it is not clear whether the differences between the SLIC cohort and the ALSPAC cohort are substantial enough to explain the opposite association trends. It remains possible that the associations observed in ALSPAC do not represent true replications. Another important point to make is that the available ALSPAC SNPs were not in high LD with the top SNP on chromosome 5p13 (Figure 3.11). Therefore, the aim of the replication

analysis was to see whether some of the associations from the association peak on chromosome 5p13 could be replicated directly.

3.2.3: The Top Associations in the Other Analyses

Although the top associations in the other analyses did not reach genome-wide significance, some of them were with SNPs that fell within interesting candidate genes. As can be seen in Table 3.2, the protein kinase C and casein kinase substrate in neurons 1 gene (*PACSIN1*) (OMIM#606512) was highlighted by analyses testing for genotype matching as well as genotype mismatching, which makes it harder to explain the biological basis for those associations. Two genes from Table 3.2 have been directly implicated in neurodevelopmental or psychiatric disorders. RNA Binding Protein, Fox-1 Homolog (C. Elegans) 3 (*RBFOX3*, also called *FOX-3*), might be involved in autism. The protein belongs to a family of RNA-binding proteins that act as splice factors (Underwood *et al.*, 2005, Kim *et al.*, 2013). Differences in expression levels between embryonic and adult mouse cortex were reported (increase in adult) (Weyn-Vanhentenryck *et al.*, 2014), which suggests a role during development. Furthermore, autistic brains with reduced levels of *RBFOX1*, *RBFOX2* and *RBFOX3* show alteration in splicing (Weyn-Vanhentenryck *et al.*, 2014). The mitochondrial ribosomal protein gene (S28*MRPS28*) (OMIM#611990) encodes a mithochondrial ribosomal protein and has been implicated in a schizophrenia gene expression study, although the direction of the changes in expression was not replicated in a second cohort used in the same study (Maycox *et al.*, 2009). In a power analysis (through simulation) of models similar to those used by EMIM which examined the parent-of-origin effects parameters and the child genetic effects parameter, Weinberg *et al.* (1998) showed that when controlling for the sample size, the risk allele frequencies, and the increase in the risk, parent-of-origin effects analyses were more powerful than the child genetic effects analysis. This could potentially explain why most tests did not detect significant associations.

3.3: Summary of Chapter 3

This chapter described the first (published) GWAS for SLI. The genome-wide analyses identified two loci of interest which were significantly or suggestively associated with

parent-of-origin effects in SLI. One locus in the gene *NOP9* on chromosomal band 14q12 was associated with paternal parent-of-origin effects, and another locus, on chromosomal band 5p13, was associated with maternal parent-of-origin effects. Both of these loci overlap with loci identified through studies of other neurodevelopmental disorders including autism, ADHD and intellectual disability. These results provide further evidence for a genetic overlap between these neurodevelopmental disorders. The analyses of child genotype effects, maternal genotype effects and interaction effects did not identify significant associations, but that could be the result of insufficient power to detect them. The analyses of parent-of-origin effects highlighted a potential new line of investigation in the field of the genetics of neurodevelopmental disorders. They provide, within some models and when such effects are operating, greater power to detect associations than that which other study designs may achieve, given sample size limitations.

3.4: Addendum to Chapter 3

The affection status of one individual who had previously been excluded from the EMIM analyses was resolved after the analyses had been completed. Additionally, two parents from two nuclear families were found to be related to each other, and the pedigree file was updated accordingly. I chose to keep the results of this study as they were when they were published. However, I decided to check how adding this individual in as a case and correcting the pedigree structure of those two families affected the top associations detected in the parent-of-origin analyses. The p-value for rs4280164 (the most highly-associated SNP in the paternal parent-of-origin analysis) changed to 2.29×10^{-8} (I_p=3.941), and the p-value for rs10447141 (the most highly-associated SNP in the maternal parent-of-origin analysis) changed to 1.29×10^{-7} (I_m=3.067). In other words, the suggestive and significant associations remained so following the corrections. Additionally, two minor errors were found in Table 3.1: the value for the I_m parameter for rs12658486 should be 2.37 and not 2.367 (this was a typographical error); the value for the I_m parameter for rs17194068 should be 2.859 and not 2.86 (this is the result of an error in the rounding, which should be up to the third decimal place for all SNPs). The results from Table 3.1 were published, so I chose to add the above corrections as an addendum.

Chapter 4: An Association Study of HLA Loci and SLI

The HLA (human leukocyte antigen) region is a super locus found on chromosome 6. The various HLA loci are involved primarily in immune function. This region on chromosome 6 is highly polymorphic (Thomson, 1981), and genetic variants of HLA genes have been associated with more than 100 diseases (Shiina *et al.*, 2004), including autoimmune diseases (Fernando *et al.*, 2008). Investigations into the possible involvement of HLA genes in neuropsychiatric and neurodevelopmental disorders were carried out following the discovery of potential connections between immune reaction and such disorders. For example, abnormal immune responses in subjects with autism (Ashwood and Van de Water, 2004). Each HLA gene examined in my analyses belongs to one of two categories of HLA genes: HLA class I genes (*HLA-A, HLA-B* and *HLA-C*, major histocompatibility complex, class I, A, B, and C, respectively), and HLA class II genes (*HLA-DQA1*, *HLA-DQB1* and *HLA-DRB1*, major histocompatibility complex, class II, DQ alpha 1, DQ beta 1, and DR beta 1, respectively) (OMIM#142800, OMIM#142830, OMIM#142840, OMIM#146880, OMIM#604305, and OMIM#142857, respectively). The function of HLA class I proteins is to present peptide antigens to T_C cells (cytotoxic T lymphocytes that kill cells targeted by the immune system for various reasons), whereas the function of HLA class II proteins is to present processed antigenic peptides to T_H cells (helper T lymphocytes that help and/or activate other immune cells).

HLA-DRB1 alleles have been associated with several disorders including autism, ADHD and schizophrenia, but the associated alleles were not always the same ones (Wright *et al.*, 1996, Torres *et al.*, 2002, Lee *et al.*, 2006, Wang *et al.*, 2008b). *HLA-A* has also been implicated in autism (Torres *et al.*, 2006). These findings will be discussed in more detail later in this chapter.

Autoimmunity is thought to be involved in some disorders which affect language. In a study of Landau-Kleffner syndrome, a disorder which includes aphasia and, at times, epileptic seizures, a positive autoimmune reaction to central and peripheral myelin was detected in children during episodes of language disturbance (Nevsimalova *et al.*, 1992). This is

interesting in that *CNTNAP2*, a gene that has been implicated in SLI and other neurodevelopmental disorders (see Chapter 1), codes for a protein that forms part of the structure of the Nodes of Ranvier, gaps in the myelin sheath. In another study, autoantibodies to brain were very common in children with Landau-Kleffner syndrome variant and autism compared to controls, in whom they were uncommon, but they were also found in children who had other neurological disorders (Connolly *et al.*, 1999). Serum obtained from the mother of three children: one with autism, one with language impairment and one who was normal, was found to contain antibodies binding to rodent Purkinje cells (neuronal cells located in the cerebellum, mentioned earlier in the context of *FOXP2* expression) and other neuronal cells in the adult large brainstem and cerebellum and to Purkinje cells in the developing cerebellum; when this serum was injected into pregnant mice, the pups which were later born to those mice showed reduced exploratory behaviour and reduced performance on tests of motor coordination (Dalton *et al.*, 2003).

Quantitative and case-control analyses were performed using both SNP data across the HLA region and the imputed HLA alleles. The quantitative analyses used individuals from 280 SLIC language-impaired families, of whom 914 had been genotyped. Table 4.1 includes the descriptive statistics of the cohort (only probands and siblings had phenotypic data), and Figure 4.1 shows the distributions of the traits. QTDT assumes that the traits are normally distributed, but the permutation procedure used here to determine significance (as described in Chapter 2) does not. Further phenotypic information about the SLIC samples can be found in (The SLI Consortium, 2002, The SLI Consortium, 2004, Falcaro *et al.*, 2008) The case-control analyses used 241 independent probands and 566 unselected British population controls (no phenotypic data for those individuals were available). Further details on the methodology can be found in Chapter 2. The analyses for the SNP data included a quantitative association analysis using MERLIN and a case-control analysis using PLINK. The SNPs included in those analyses numbered 6725, from chromosome 6 25-35 MB (hg18). I performed two types of analyses with the HLA data, a family-based quantitative analysis using QTDT, which included tests of parent-of-origin effects, and a case-control analysis with SLI probands and population controls. The same quality control measures used in the GWAS (with respect to exclusion criteria) were used in these analyses. The same applied to the

population control samples, except that a minority of samples, which had discordant sex information, were not removed, in contrast to SLIC samples. In the SLIC cohort, discordant sex information may suggest a sample mix-up which could affect the affection status of the individual, but in the case of the population controls, all individuals were independent and were assumed to not have SLI, and, therefore, an incorrect ID would affect the results.

Table 4.1: Descriptive statistics of language test scores in SLIC children (as used in the quantitative analyses).

Trait	NWR	ELS	RLS
Mean	90.858	77.812	88.005
Standard deviation	19.406	16.742	17.354
Median	94	76	87
Minimum	55	45	50
Maximum	136	131	138

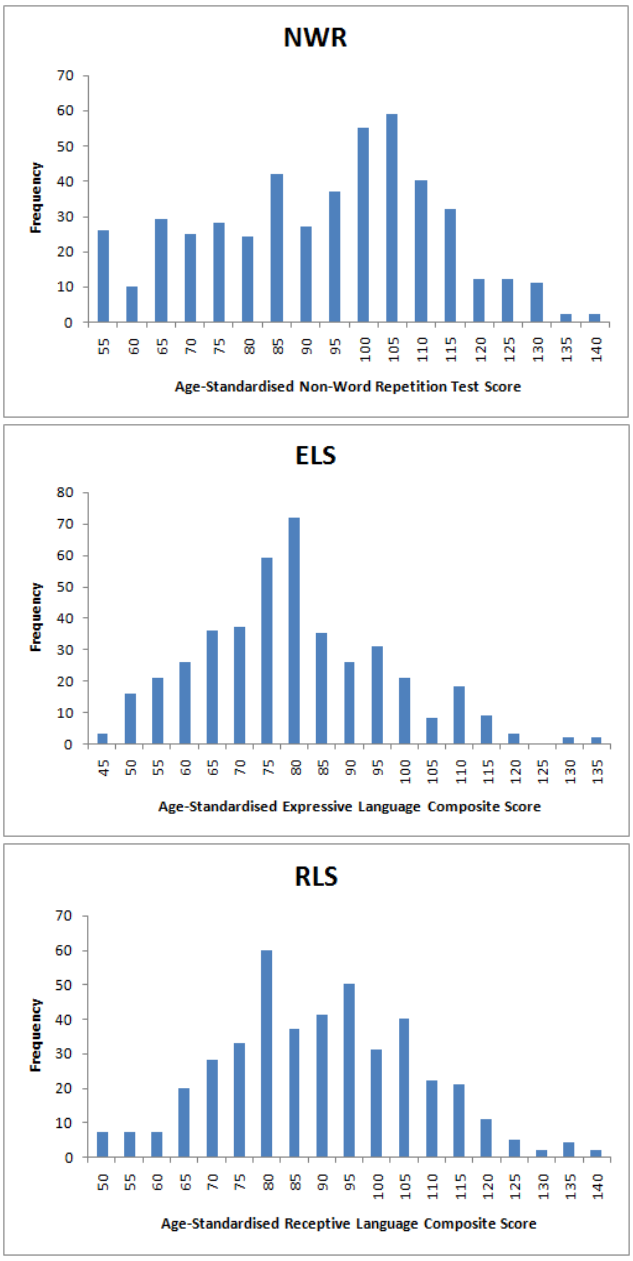


Figure 4.1: Distributions of language test scores in SLIC children (as used in the quantitative analyses).

4.1: Results of the Various HLA Association Analyses

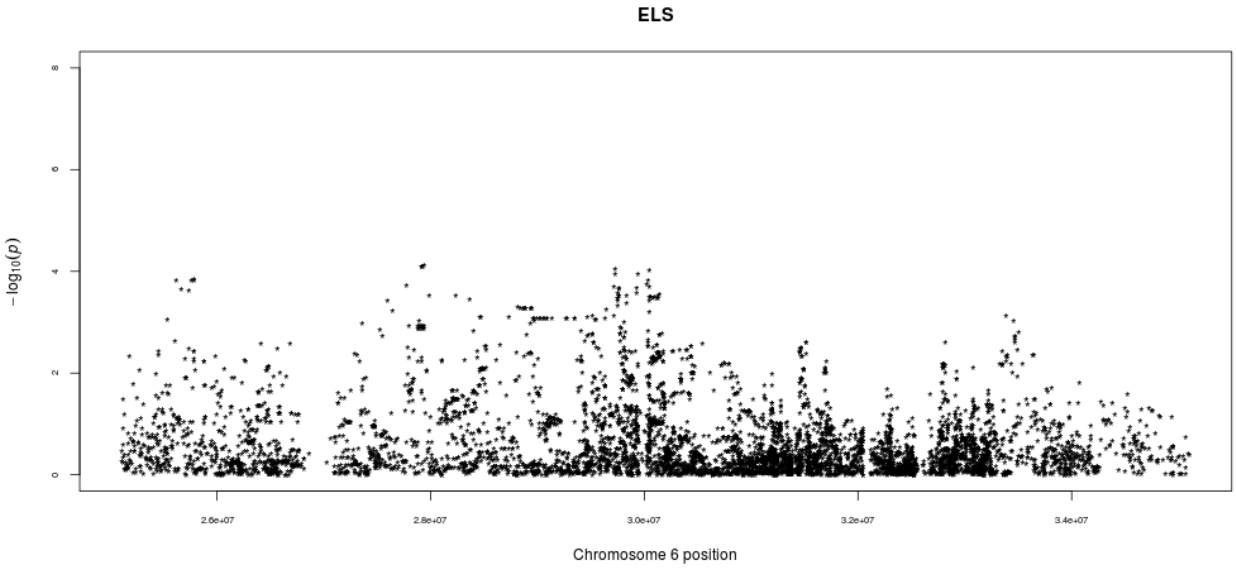
The analyses which used the SNP data did not produce any significant results. In the analyses which used the imputed HLA alleles, several associations were detected.

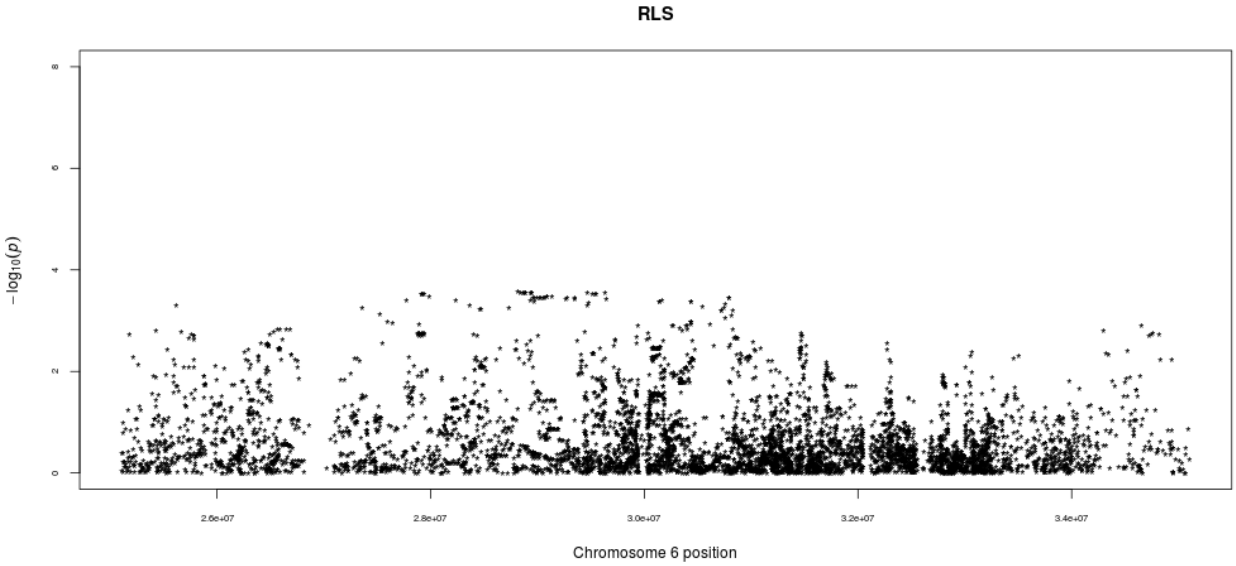
4.1.1: Results of the Quantitative and Case-control Analyses of SNP Data

None of these analyses found significant associations (in terms of having q≤0.05). Regional association plots for the quantitative analyses for ELS, RLS and NWR, and the case-control analysis can be found in Figure 4.2. In the context of HLA genes, only the quantitative analysis for ELS identified an association with an HLA locus (a SNP in HLA-A, rs3823342, P=0.01173). This association had a q-value of 0.205 i.e. 20.5% of the associations with p-values of 0.01173 or below are expected to be false positives. The lowest q-value for the ELS associations was 0.057, and the most highly-associated SNP was rs1319972 ($P=7.72\times10^{-5}$). The lowest q-values for the RLS, NWR and case-control associations were 0.064, 0.155 and 0.671, respectively. The top SNP in the RLS analysis was rs1233579 (P=0.00026). The top SNPs in the ELS and RLS analyses did not fall within any currently known genes. The top SNP in the NWR analysis (rs17266491) fell within the gene *LRRC16A* (leucine rich repeat containing 16A) ($P=6.86\times10^{-5}$). The top SNP in the case-control analysis (rs9467476) fell within the same gene (P=0.00027).

4.1.2: Results of the Quantitative and Case-control Analyses of HLA Data

All nominal p-values of at most 0.05 in the quantitative and case-control analyses can be found in Tables 4.2 and 4.3, respectively (the p-values have been rounded up to the third decimal place). Differences in allele frequencies for all HLA genes between SLIC probands and population controls are shown in Figure 4.3 (this figure includes all HLA types/alleles). Appendix B presents the allele counts and allele frequencies (and confidence intervals) in probands and in controls. A permutation procedure implemented in QTDT was used to account for multiple testing in the quantitative analysis. In the case-control analysis, q-values were calculated for that purpose. The quantitative analysis suggested possible associations of *HLA-A* alleles with NWR and ELS. *HLA-B* alleles were implicated in the parent-of-origin analyses.


ELS
$-\log_{10}(p)$
0
2
4
6
8
2.6e+07
2.8e+07
3.0e+07
3.2e+07
3.4e+07
Chromosome 6 position


RLS
$-\log_{10}(p)$
0
2
4
6
8
2.6e+07
2.8e+07
3.0e+07
3.2e+07
3.4e+07
Chromosome 6 position

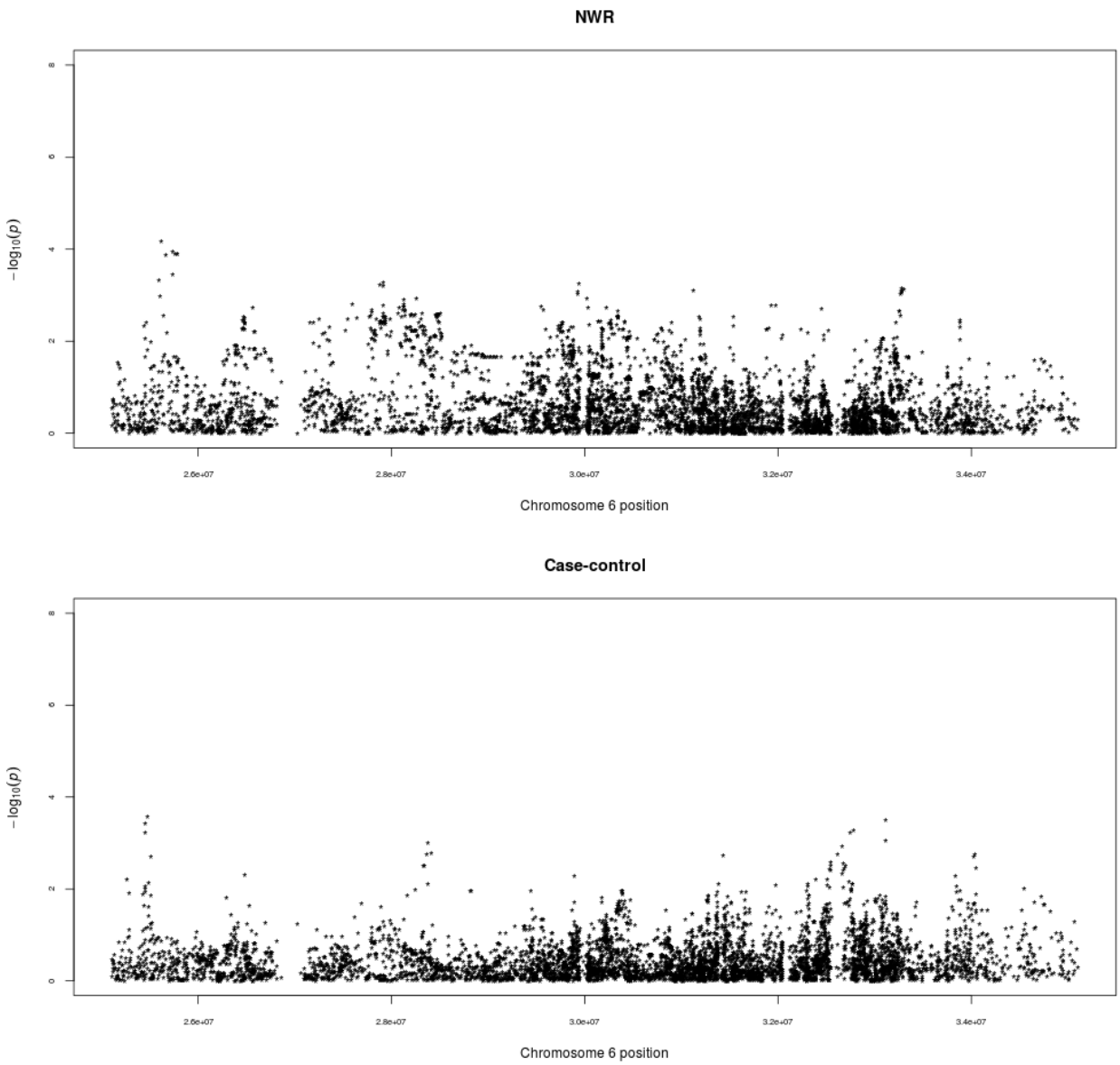


Figure 4.2: Regional association plots for the association analyses of SNPs in the HLA region (previous page as well).

Table 4.2: Results of the HLA QTDT analysis (P≤0.05).

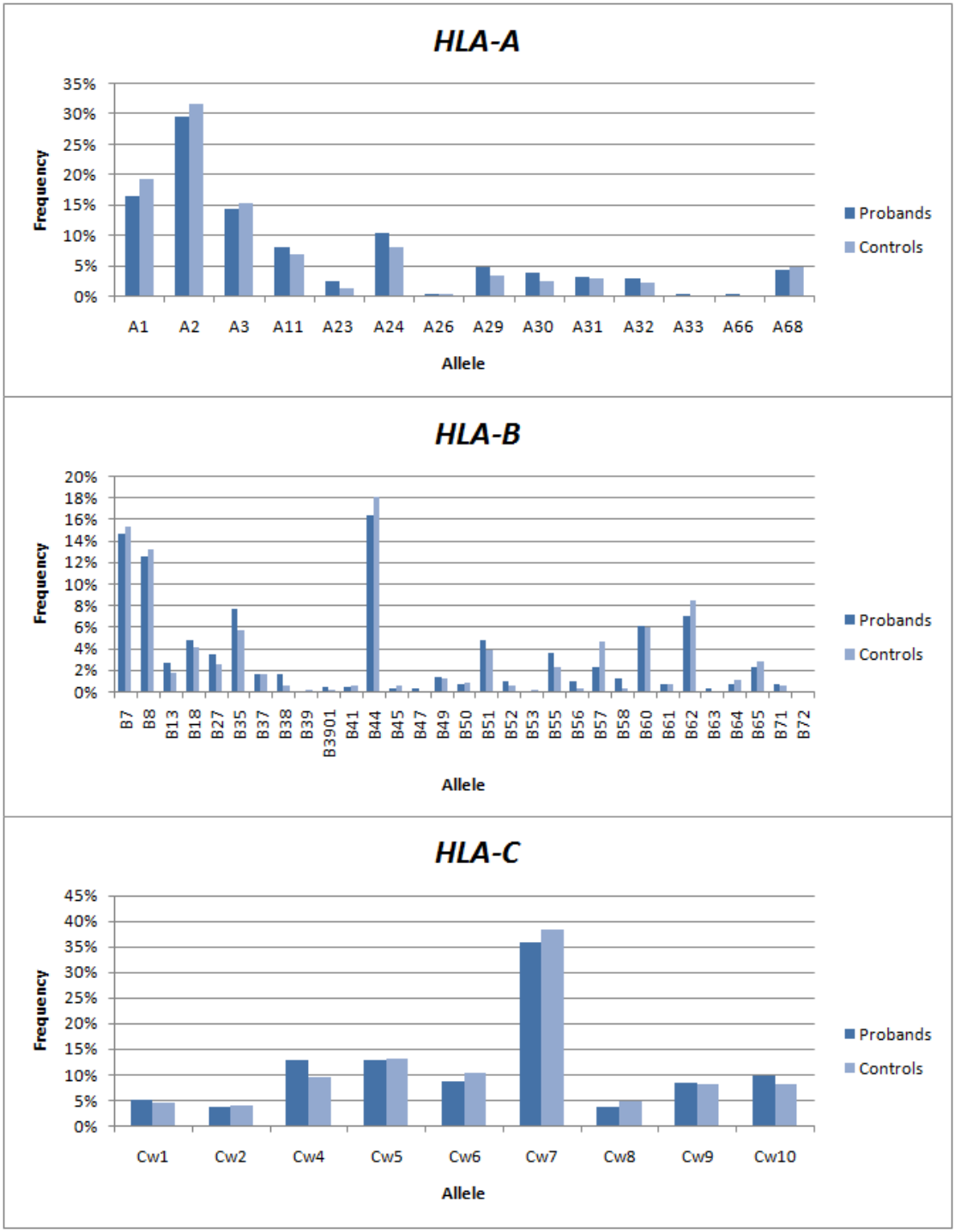
HLA gene/allele	General QTDT	Maternal parent-of-origin analysis	Paternal parent-of-origin analysis
HLA-A A1	0.004 (NWR)+* 0.026 (ELS)+		
HLA-A A3	0.006 (ELS)-*		
HLA-B B8		0.021 (RLS)-*	0.013 (RLS)+*
HLA-C Cw6			0.019 (ELS)-*
HLA-C Cw8	0.039 (ELS)+*		
*DQA1*0301			0.03 (ELS)- 0.011 (RLS)-*
*DQA1*0501		0.034 (RLS)-*	0.029 (RLS)+
DRB1 DR7	0.019 (RLS)+		
DRB1 DR17		0.028 (RLS)-	

+/- refers to the direction of the association; associations marked by * remained significant following the permutation procedure.

Table 4.3: Results of the HLA case-control analysis (P≤0.05).

HLA gene/ allele	AC probands	AC controls	F probands	F controls	Fisher's exact test P	RR	95% CI lower boundary	95% CI upper boundary
HLA-DRB1 DR10	395	912	0.018	0.002	0.004	2.575	1.773	3.737
*HLA-DQA1** 0301	453	1078	0.174	0.231	0.012	0.753	0.59	0.961
HLA-B B57	443	1055	0.023	0.047	0.029	0.541	0.303	0.963
HLA-DRB1 DR4	395	912	0.108	0.158	0.02	0.692	0.51	0.939
HLA-DRB1 DR12	395	912	0.005	0.02	0.05	0.333	0.09	1.243

AC – allele count; F – frequency; RR – relative risk; CI – confidence interval.


HLA-A
35%
30%
25%
20%
15%
10%
5%
0%
Frequency
A1
A2
A3
A11
A23
A24
A26
A29
A30
A31
A32
A33
A66
A68
Allele
Probands
Controls
HLA-B
20%
18%
16%
14%
12%
10%
8%
6%
4%
2%
0%
Frequency
B7
B8
B13
B18
B27
B35
B37
B38
B39
B3901
B41
B44
B45
B47
B49
B50
B51
B52
B53
B55
B56
B57
B58
B60
B61
B62
B63
B64
B65
B71
B72
Allele
Probands
Controls
HLA-C
45%
40%
35%
30%
25%
20%
15%
10%
5%
0%
Frequency
Cw1
Cw2
Cw4
Cw5
Cw6
Cw7
Cw8
Cw9
Cw10
Allele
Probands
Controls

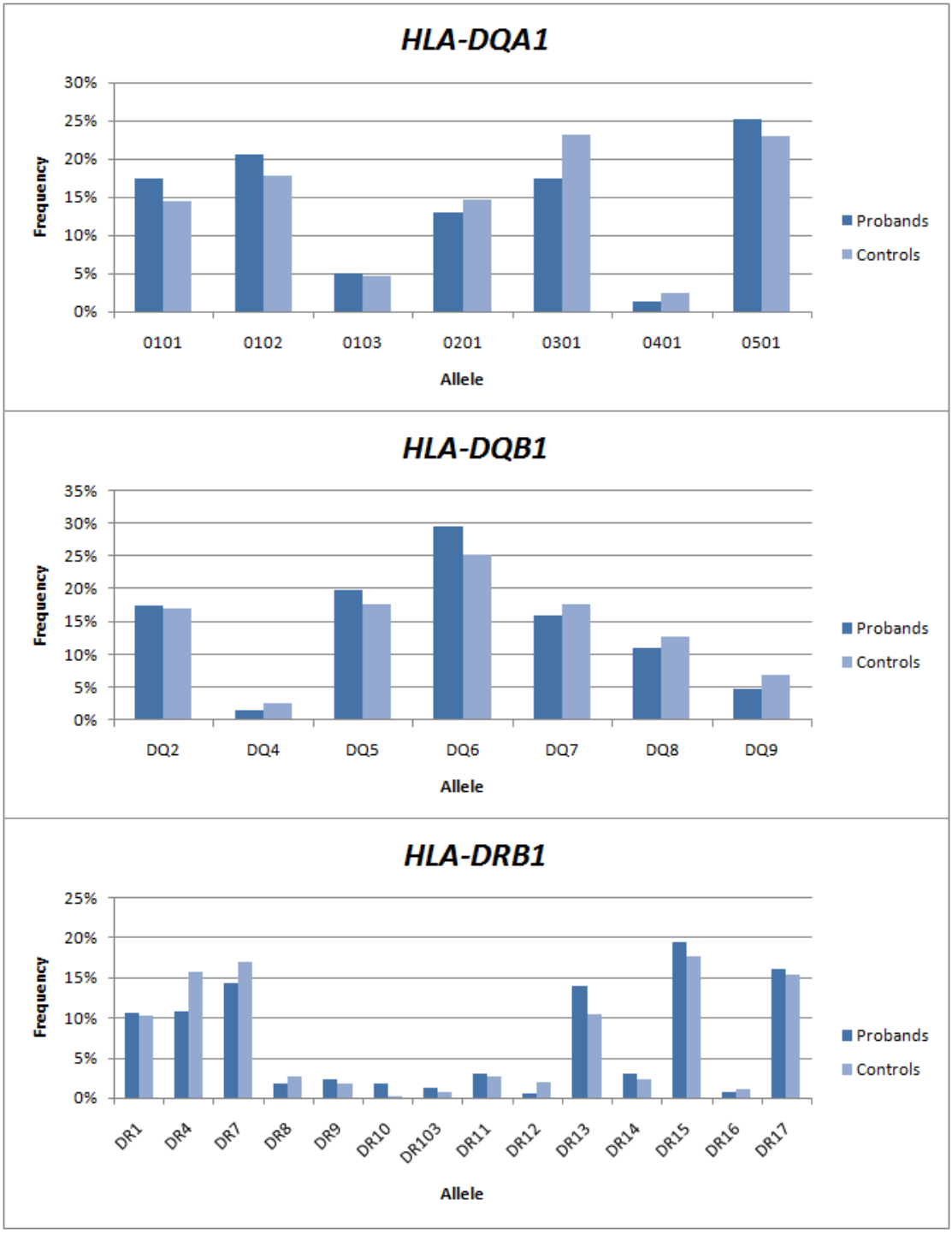


Figure 4.3: HLA allele frequencies in SLI probands and population controls (previous page as well).

In total, nine alleles showed association with P≤0.05 in the quantitative association analysis. There were 13 associations (some alleles had nominally significant p-values in more than one test), 8 of which remained significant following the permutation procedure.

The most highly-associated allele was the A1 allele of *HLA-A*. Allele A1 was positively associated with NWR (P=0.0004, empirical P=0.017). The same allele was also associated with increased ELS (P=0.026), but this association did not remain significant following the permutation procedure (empirical P>0.05). The A3 allele was associated with decreased ELS (P=0.006, empirical P=0.017). Other, less significant associations were also identified, but only one of them, between *HLA-C* CW8 and increased ELS (P=0.039, empirical P=0.048) had empirical P≤0.05.

Parent-of-origin effects were detected as well. The B8 allele of *HLA-B* was associated with increased RLS when inherited from the father (P=0.013, empirical P=0.029) and decreased RLS when inherited from the mother (P=0.021, empirical P=0.013). DQA1*0501 showed a similar pattern of association with RLS; however, the paternal association did not remain significant following the permutation procedure.

The case-control association analysis found five alleles with nominal P≤0.05. The DR10 allele of *HLA-DRB1* was significantly more frequent in cases than in controls (P=0.004). This allele had a relative risk of 2.575 (95% confidence interval: lower boundary: 1.773, upper boundary: 3.737). The remaining four associations pertain to alleles that have a protective effect, i.e. they were more frequent in controls than in cases. These included: HLA-DQA1*0301 (nominal P=0.014), the B57 allele of *HLA-B* (nominal P=0.029), and the DR4 and DR12 alleles of *HLA-DRB1* (nominal P=0.02 and nominal P=0.05, respectively). Both the lower and the upper boundaries of the 95% confidence intervals for the relative risks of all of those alleles save DR12 were below 1, which further supports the associations as being true positives. The q-values for the least highly-associated allele of each gene were as follows: 0.235 for *HLA-DRB1*, 0.911 for *HLA-B*, and 0.098 for *HLA-DQA1*.

4.2: Discussion of the HLA Association Study Results

The quantitative analysis and the case-control analysis identified HLA alleles that were associated with an increase or a decrease in the linguistic test scores, or had a protective effect or increased the risk of SLI, respectively.

4.2.1: Discussion of the Results of the Analyses which Used the SNP Data

As previously stated, no significant associations were found in the SNP-based analyses, considering that none had a q-value of 0.05 or below. In contrast to the case-control analysis which used HLA types, the SNPs themselves do not necessarily represent the causal variants, and with such large numbers of SNPs, it is not advisable to draw conclusions based on a p-value cut-off while considering the q-value (as described below for case-control association analysis with HLA types). To overcome the issue of the high genetic variability of the region, HLA alleles were imputed from the SNP data, and alleles were then grouped into types, where possible.

Considering the following: with a disease allele frequency of 0.01, a disease prevalence of 7% (similar to SLI), and the numbers of probands and controls that were used in this study, the power to detect an association at $\alpha=0.05$ is ~80%, if the relative risk is 3. If the relative risk is set to 2.5, the power decreases to ~60%. If the disease allele frequency is then raised to 0.1, the power increases to 99.9%. These calculations were done with the Genetic Power Calculator (Purcell *et al.*, 2003). Most of the SNPs used in this study had higher minor allele frequencies than 0.01 and even 0.1. However, it is likely that their contribution to the risk was not very high. The analysis which used HLA data allowed the identification of very rare HLA types which had high relative risks.

4.2.2: Discussion of the Results of the QTDT Analysis

The most significant result in the general QTDT analysis was the association of the A1 allele of *HLA-A* with P=0.004. This allele was associated with increased NWR. The A3 allele of the same gene was the most significant negative association (P=0.006), with ELS. Both of these associations remained significant (empirical $P \leq 0.05$) after 1000 permutations.

The *HLA-A* gene has previously been implicated in autism, but the associated allele in that case was A2 (Torres *et al.*, 2006). Both A1 and A3 were examined in that study, but they were not found to be significantly associated with autism.

In the parent-of-origin analyses, allele B8 of the *HLA-B* gene was positively correlated with RLS when inherited from the father (P=0.013) and negatively correlated with RLS when inherited from the mother (P=0.021). Both of these associations remained significant following 1000 permutations. The same trend was observed with DQA1*0501. However, the positive association (with paternal parent-of-origin effects) in that case did not remain significant after the permutation procedure. Interestingly, *HLA-B* was implicated in schizophrenia (Palmer *et al.*, 2006), but the analysis which identified the involvement of this gene in schizophrenia was the MFG-matching test (described in Chapter 2 in the EMIM GWAS section). Thus, it was not any specific *HLA-B* allele that increased the risk of having schizophrenia, but, rather, the matching between the mother's and child's alleles. In my analysis, a specific allele was associated with decreased risk, but only when inherited from the mother. That is, there is some correlation between the MFG-matching test and the parent-of-origin analysis but only when it comes to allele B8 (since the allele's being inherited from the mother indicates at least a partial matching between the mother's and child's genotypes for that allele). An MFG-matching test was not performed with the HLA data, because available software packages which employ the MFG-matching test accept only di-allelic markers.

4.2.3: Discussion of the Results of the Case-control Analysis with HLA Types

The case-control analysis identified five alleles with nominal P≤0.05. The DR10 allele of *HLA-DRB1* was the most highly-associated allele and was more frequent in cases than in controls (P=0.004). Moreover, it had a relative risk of 2.575 with a lower boundary of the 95% confidence interval of 1.773 and an upper boundary of 3.737. The same allele was found to be significantly more frequent in cases than in controls in a study of ADHD of a Chinese cohort (Wang *et al.*, 2008b). Similarly, DR12, which had a protective effect in my analysis (P=0.05, RR=0.333, 95% confidence interval lower boundary of 0.09, upper boundary of 1.243), was more frequent in controls than in children with ADHD. However, it was only marginally significant in my analysis (the 95% confidence interval for the RR encompassed 1).

DR4 also had a protective effect in my analysis (P=0.02, RR=0.692, 95% confidence interval lower boundary of 0.51, upper boundary of 0.939), but the ADHD study did not find a significant association between decreased risk and DR4. DR4 was also found to have a protective effect in the context of schizophrenia (Wright *et al.*, 1996). There is conflicting evidence for the role of DR4 in ADHD: one study found that DR4 was significantly associated with ADHD (Odell *et al.*, 1997), but no such association was found in another study (Wang *et al.*, 2008b). It is interesting to note that association trends of *HLA-DRB1* observed in autism seem to be opposite to those observed in SLI: DR4 was shown to be significantly associated with autism (Torres *et al.*, 2002, Lee *et al.*, 2006), but in the context of SLI it has a protective effect. DR13 and DR14 were more frequent in SLI cases than in controls, as is shown in Figure 4.3 (frequency in cases: 0.14 and 0.03; frequency in controls: 0.11 and 0.02, for DR13 and DR14, respectively). These differences were not significant in SLI, but an opposite association trend in autism was reported, in which case they were significantly less frequent in autism cases (DR13 and DR14 were grouped together) (Torres *et al.*, 2002). These results suggest that while HLA genes may play a role in SLI, autism and ADHD, different alleles may be contributing to the susceptibility to these disorders in different ways.

In order to correct for multiple testing, the FDR method was used, and q-values were generated, as a large number of alleles were tested, and this method is well-suited for such cases (further details can be found in Chapter 2). Of the *HLA-B* alleles tested, only one, B57, had P≤0.05. The q-value for B57 was 0.911, so it is likely to be a false positive. Of the *HLA-DQA1* alleles tested, only DQA1*0301 had P≤0.05, and its q-value was 0.098. Three *HLA-DRB1* alleles had P≤0.05: DR10, DR4 and DR12 (in order of significance). The q-value for DR12 was 0.235, which can be interpreted as the proportion of false positives among the three significantly-associated alleles of *HLA-DRB1*. In other words, it is possible that one of these alleles is a false positive, but it is not possible to know which one is the false positive.

4.2.4: The Lack of Overlap between the Results of the Quantitative Analyses and the Results of the Case-control Analysis with the HLA Data

The quantitative association analyses identified significant associations particularly with alleles of *HLA-A* and *HLA-B*, whereas the case-control analyses identified significant

associations with *HLA-DRB1* alleles (particularly DR10). Several factors may account for this discrepancy. Firstly, NWR was not used in the ascertainment of probands. In fact, as discussed in Chapter 1, while most probands did indeed have low ELS and/or RLS, they were all considered "affected", even though some had scores which were higher than 1.5 SD below the mean for their age. Therefore, associations with NWR are not expected to be replicated in a case-control analysis, and, to a lesser degree, associations with ELS or RLS might not be replicated either. Secondly, as discussed in Chapter 1, different models may detect different results, even when the same cohort is used, let alone when different cohorts are used. In this case, the quantitative analyses used family-based models, while the case-control analyses used population-based models. The only thing both types of analyses had in common was the proband cohort, which was included in both, whereas the other samples came from parents and siblings, in the case of the quantitative analyses, and unselected population controls, in the case of the case-control analysis. Therefore, I would view those two approaches as complementary.

4.3: Summary of Chapter 4

This chapter described the first association study of HLA genes and SLI. The results of the quantitative and case-control association analyses identified several alleles that might influence susceptibility to SLI. Several of these alleles, notably ones in *HLA-DRB1*, *HLA-B* and *HLA-A*, have been implicated in other neurodevelopmental or neuropsychiatric disorders, although the nature of the association is not always the same as in SLI. Furthermore, the associations in *HLA-B* displayed parent-of-origin effects, in which the association trend depended on whether the allele was inherited from the father or from the mother. The influence of the immune system on neural development is an important area of study which is being examined from various perspectives. The results of this genetic study highlight the possible involvement of HLA genes in language impairment, and add to the literature on the possible links between various neurodevelopmental disorders.

4.4: Addendum to Chapter 4

Minor errors were found in Table 4.3: the rounded frequency of DR4 in probands should be 0.109, rather than 0.108, and the rounded frequency of *HLA-DQA1**0301 in controls should be 0.232, rather than 0.231. It should be noted that a more appropriate way to quantify the associations in the case-control analysis would have been to use odds ratios and not relative risks. Therefore, I present the odds ratios for the top 5 associations from Table 4.3, with their 95% confidence intervals: DR10 8.085 (1.661, 39.354); *DQA1**0301 0.674 (0.483, 0.94); B57 0.449 (0.223, 0.903); DR4 0.599 (0.397, 0.904); DR12 0.255 (0.058, 1.115). I recently learned that some samples in the case-control analysis had inaccurate information regarding their proband status and/or should have been removed during the quality control step but were not removed due to a sample ID issue. In total, 15 samples should have been removed from the proband cohort, and 5 samples should have been added to it. Additionally, 3 pairs of related probands should have been removed from the case-control analysis. The quantitative analyses were less affected by this, as they relied on quantitative traits rather than affection status, employed the within-family model, and used a pedigree file from which all samples which had failed the quality control thresholds had been removed already, but further examination revealed that a few ungrouped alleles of *HLA-C* should have not been used, so they were set to have missing values. As this work has been published, and the above corrections did not affect the results drastically, I decided to leave the information in this chapter as it was when the results were published. Nonetheless, I will contrast the results of the old and new analyses here with regards to the significant associations found previously. In the general QTDT, the p-value for A1 and NWR was 0.005, and the p-value for A3 and ELS was 0.007. In the paternal parent-of-origin effects analysis, the p-value for Cw6 and ELS was 0.003. The association trends remained the same for all alleles. Allele Cw8 was not tested in the general QTDT due to lack of sufficient informative families. The rounded p-values for the other significant associations (as described in the text) remained the same. The empirical p-values for the associations that were significant previously were still below 0.05. In the case-control analysis of HLA data, DR10, *DQA1**0301, and DR4 still had p-values below 0.05 (0.003, 0.015, and 0.035, respectively). The q-values

for *DQA1**0301 and DR4 were 0.103 and 0.244, respectively. The q-value for DR10 was 0.044, suggesting that it would survive the FDR correction if results were filtered by q-value. Lastly, with the corrections, the odds ratios for DR10, *DQA1**0301 and DR4 were 8.632, 0.652 and 0.602, respectively. Their confidence intervals did not encompass 1, but it should be noted that the approximation for the confidence intervals does not work well for rare alleles, which is why the lower and upper boundaries of the confidence interval for DR10 are quite far from each other. In summary, the quantitative association analyses still highlight *HLA-A* and *HLA-B* alleles, and the results of the case-control association analysis still highlight the potential involvement of *HLA-DRB1* in SLI: DR10 obtained a lower p-value, a lower q-value and a higher odds ratio following the corrections.

Chapter 5: A Candidate Gene Analysis Study of an Isolated Population Affected by SLI

As discussed in Chapter 1, the population living on the Robinson Crusoe Island is an isolated population with an increased inbreeding coefficient in which the prevalence of SLI is very high (35%). A previous genetic study of this population employed linkage analysis methods and found a region on chromosome 7 that was consistently linked to SLI under several models (Villanueva *et al.*, 2011). Given the relatedness between the families on the island, and the fact that many of the affected individuals could trace their lineage back to two founder brothers, it was possible that a coding mutation which was related to the language impairment was passed on through the generations. To test this hypothesis, exome sequencing was performed on samples from five affected children from different families living on the island. The exome sequencing was done in Nijmegen, the Netherlands. This chapter describes the work done following the exome sequencing, which included the validation of the identified variants in the entire island cohort, direct association analyses to test for a correlation between the variants and SLI, the identification of a candidate gene for SLI, and the study of the genetic variation in that gene using other cohorts. Additionally, using genotypes obtained from a new array designed for South-American populations, I performed a targeted association analysis of the region on chromosome 7 which was consistently implicated across several linkage analyses in the previous study of the language impairment in this population (Villanueva *et al.*, 2011).

The affection statuses used in the targeted association analysis described in this chapter were the same as the ones used in the linkage study, as was the sample. In the association analysis for the exome sequencing variants, also described in this chapter, affection statuses were determined in the following manner: probands were diagnosed with SLI strictly on the basis of their language test scores and exclusionary criteria (as described in Chapter 1); children aged 6 or younger with scores of >2 SD below those expected on the phonology test and children above the age of 6 with performance of two years below that expected for their chronological age on the grammar test, as well as children who performed below the 10th percentile on either the expressive or the receptive scales of the grammar test

were classified as having SLI, if their non-verbal IQ was not below the 10th percentile, and they had normal hearing, oral-motor skills and neurological development. Children whose performance on the phonology test was not 2 SD below expected or two years below expected for their chronological age, and whose performance on the grammar test was above the 10th percentile on both the expressive and receptive scales were considered unaffected (normal language). Other children e.g. children having language problems and low IQ or other developmental problems, were excluded from this study. Parents and siblings of probands were diagnosed with SLI if they performed poorly (below the 10th percentile) on one of the language tests administered to them (as outlined in Chapter 1) and/or if they reported a clinical history of language problems. They were considered unaffected if they performed above the 10th percentile on both language tests and had no history of language problems.

5.1: Results of the Exome Sequencing

Exome enrichment was performed using Agilent SureSelect (v.1; 18,000 genes), and the sequencing platform was the SOLiD 4 system. Variants were called using algorithms developed by our collaborators in Nijmegen. The exome sequencing done on samples from five affected individuals identified, on average, 58,495 variants (medium stringency) out of which 24,683 were untranslated region (UTR) variants, coding variants or splice site variants. In order to narrow down the list of variants to be examined further, the following filters were applied:

- The variant had to be present in three out of the five affected individuals.
- The variant must not have been observed in the 1000 Genomes populations, in dbSNP 132, or in the Nijmegen in-house exome database.
- The variant had to be a non-synonymous coding or splice site variant.

This narrowed down the list to include nine variants. Details on the variants can be found in Table 5.1.

Table 5.1: Variants identified through the exome sequencing that met the filtering criteria.

Transcript	Gene symbol	Chromosome	Position (hg19)	Reference/variant (+ strand)
NM_001004690	*OR2M5*	1	248308783	T/A
NM_001042678	*RHOC*	1	113245326	T/C
NM_152995	*NFXL1*	4	47907320	T/A
NM_001143766	*ZNF438*	10	31134425	G/A
NM_139160	*DEPDC7*	11	33054503	T/G
NM_000418	*IL4R*	16	27363901	G/A
NM_001130141	*PCBP3*	21	47359924	C/T
NM_145174	*DNAJB7*	22	41257834	0/T
NM_006044	*HDAC6*	X	48682972	A/G

5.2: Validation of Variants in the Island Cohort and Case-control Association Analysis

Islander samples were checked for the presence of the above nine variants. This study included 111 Island samples (49 cases and 62 controls, following quality control which identified sample mix-ups and discordant gender information), although some samples were excluded for a given variant, if Mendelian errors were found (two families were removed from the analyses of two variants), or if the sequencing reaction consistently failed for them (one person was removed from the analysis of one variant). The pedigree structure for the association analysis reflected the relatedness of the individuals going back five generations and contained 288 individuals divided into two families (283 in one large family, and 5 in a smaller family). Sanger sequencing was used to determine the genotype each individual had at the positions of the variants (primer sequences can be found in Appendix A). Case-control analyses were then performed using MQLS and MQLS-XM, which control for the relatedness of the individuals (further details can be found in Chapter 2). One variant, in NM_001004690, was found to be a false positive (i.e. a miscalling following the exome sequencing; the variant was not found in any individual). The results of the case-control association analysis for the remaining eight variants can be found in Table 5.2.

Table 5.2: Results of the case-control analyses for the exome sequencing variants.

Transcript	Gene symbol	p-value
NM_001042678	*RHOC*	0.6253
NM_152995	*NFXL1*	0.0002
NM_001143766	*ZNF438*	0.4658
NM_139160	*DEPDC7*	0.3988
NM_000418	*IL4R*	0.0535
NM_001130141	*PCBP3*	0.2279
NM_145174	*DNAJB7*	0.5539
NM_006044	*HDAC6*	0.4565

As can be seen from Table 5.2, only one association was significant, and its p-value was considerably lower than those of the other associations. That association was with a non-synonymous coding change in the nuclear transcription factor, X-box binding-like 1 gene (*NFXL1*). The coding change is of the form N150K, i.e. an asparagine to lysine change at amino acid position 150 of the protein. The position of the change is 10 amino acids away from a zinc finger RING-type domain. This change is predicted to be "disease-causing" by MutationTaster; the nucleotide at this position is highly conserved (phastCons score of 1), and potential splice site changes were identified. This coding variant was designated rs144169475 (after this analysis had been carried out). The *NFXL1* gene has not been studied, but it is predicted to code for a transcription factor, based on structural similarities to nuclear transcription factor, X-box binding 1 (*NFX1*) (OMIM#603255); both proteins have NF-X1 type zinc finger domains, and there is a 22%-27% similarity in their amino acid sequences (NFX1 is slightly larger). Despite the fact that *NFXL1* has not been the focus of any published functional studies (note that its function has been studied as part of this thesis), it has nonetheless been implicated in several studies. NFXL1 was shown to be in abundance in embryonic stem cells prior to their differentiation into oligodendrocytes, the cells which produce myelin (Chaerkady *et al.*, 2011). The chromosomal region in which *NFXL1* is found, 4p12, has been implicated in two studies of neurodevelopmental disorders: a family suffering from a syndrome involving severe mental retardation and other developmental problems (Kahrizi *et al.*, 2009), and a family in which members suffered from a variety of disorders including autism, ADHD, developmental delay and learning disabilities (Polan *et al.*, 2013). In the former case, *NFXL1* fell within the significant linkage region identified in that study, and, in

the latter case, *NFXL1* fell within a 2.42 Mb duplication reported for the family studied. *NFXL1* did not fall within the linkage regions identified by Villanueva *et al.* (2011).

The frequency of the rs144169475 allele A (- strand, reference T) was 4.8% in 62 unaffected individuals, compared to 19.4% in 49 affected individuals. Figure 5.1 includes a sequence containing the variant and a sequence which contains only the reference allele. No individual was homozygous for the variant. The variant was not found in 346 European samples, which included 195 SLIC samples (mostly probands) and 151 samples from the Human Random Control DNA panels, but it was found in 320 Colombian unselected controls, with a frequency of 4.2%, and in 121 Chilean controls, with a frequency of 7.4%. The presence of the variant in samples from those cohorts was determined using Sanger sequencing. Again, no individual was homozygous for this variant. The 1000 Genomes project database indicated that only three populations, all of which were South-American, included individuals who carried the variant. The frequency of the variant in the PUR population (Puerto Rican in Puerto Rico) was 0.9%; the frequency of the variant in the CLM population (Colombian in Medellin, Colombia) was 3.3%; the frequency of the variant in the MXL population (Mexican Ancestry in Los Angeles, CA) was 7.6%. Here, again, no individual was homozygous for the variant. It thus seems to be the case that the variant is only found in South-American populations. Interestingly, the frequency of the variant in unaffected islanders (4.8%) is closer to its frequency in the Colombian unselected controls (4.2%), and in the Chilean controls (7.4%), and to the average frequency of the variant in South-American 1000 Genomes populations (4.5%) than to the frequency of the variant in affected islanders, which is considerably higher (19.4%).

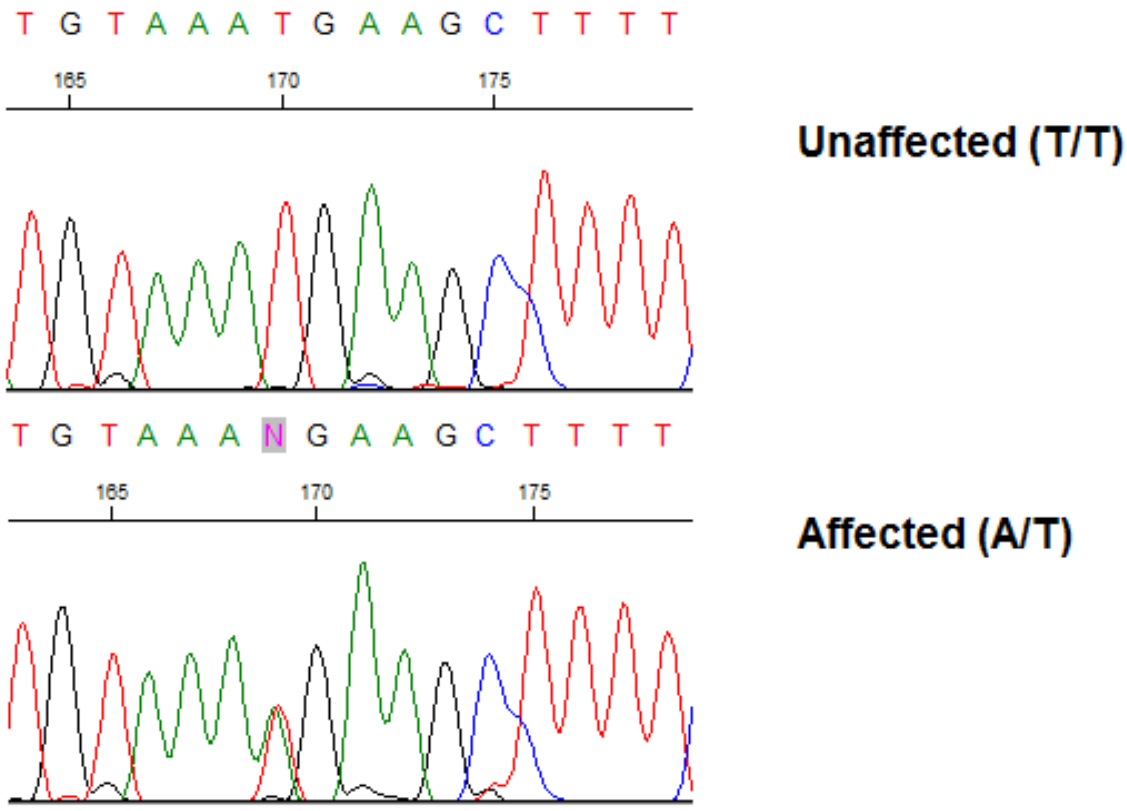


Figure 5.1: The sequence around the NFXL1 variant (- strand). The variant position is marked as "N" in the above sequence.

In the original linkage study of the Robinson Crusoe Island population, some children were identified as having language problems associated with other neurological or developmental deficits; neither these children nor their families were used in the linkage study, since they could not be diagnosed as having SLI (Villanueva *et al.*, 2011). The genomic region containing the variant was sequenced in 16 such families (46 individuals). Case-control analysis was not feasible, as there were no affection status data available for those individuals. The frequency of the variant in those families was 6.5%. This frequency is closer to that in unaffected individuals (4.8%) than to that in affected individuals (19.4%). This suggests that the variant is associated only with SLI and not with language problems in general. Across all Robinson Crusoe Island samples genotyped, the variant was found in 14 nuclear families which were part of the larger pedigree. In those families, the variant co-segregated with SLI in five cases (four families).

5.3: Analyses of Gene Expression Data from the HapMap3 Project in the Context of *NFXL1*

In a study by Stranger *et al.*, genome-wide gene expression data for lymphoblastoid cell lines from 726 individuals from eight HapMap3 populations were generated (Stranger *et al.*, 2012). Data for 618 individuals from seven HapMap3 populations were made available online, and these were used in the analyses described in the following section. The rationale behind these analyses is that, given that the NFXL1 protein is predicted to regulate the expression of genes, it is possible that the amino acid substitution may influence its function, and, therefore, the expression levels of the genes it regulates may differ between individuals who only have the wild-type *NFXL1* and individuals who carry the variant. Similarly, it may be possible to detect correlations between the expression patterns of *NFXL1* and other genes. I performed two analyses: a quantitative gene expression association analysis with gene expression data from the MEX population (Mexican ancestry in Los Angeles, California i.e. the MXL population in 1000 Genomes; the only South-American population among the aforementioned seven populations, and, therefore, the only one which included individuals carrying the identified variant in *NFXL1*), and a correlation analysis of expression patterns between the included *NFXL1* transcript and all other transcripts (21,800 in total, including

NFXL1). The gene expression data used were part of the processed data made available online, which had gone through quality control and quantile normalisation ("within IVTs (in-vitro transcription) and then across samples, then median normalised for each population").

5.3.1: A Quantitative Gene Expression Association Analysis

The quantitative association analysis performed with MERLIN (see Chapter 2 for details) included 42 samples from the MEX population which had genotypes for the position of the *NFXL1* variant. Gene expression data were available for three other samples labelled as MEX, but genotype data for those samples were not available. In total, there were six individuals carrying the variant (all six were heterozygous for the variant), and 36 individuals who were homozygous for the reference allele i.e. the variant allele frequency was 7.14%. The analysis included 21,800 transcripts. No genome-wide significant associations were found; FDR analysis (as performed for the HLA associations) resulted in a lowest q-value of 0.779. There were two transcripts that had that q-value: one was a transcript of phosphodiesterase 1C (*PDE1C*) (OMIM#602987), with $P=6.27\times10^{-5}$, and the other was a transcript of TNF receptor-associated factor 7 (*TRAF7*) (OMIM#606692), with $P=7.15\times10^{-5}$. *TRAF7* induces the activation of NF-κB, a family of transcription factors which includes at least one member that interacts with *CMIP* (see Chapter 1) (Bouwmeester *et al.*, 2004). *PDE1C* is a FOXP2 target (Vernes *et al.*, 2007) which is involved in signal transduction and is expressed in several brain regions (Lakics *et al.*, 2010). An inversion disrupting *PDE1C* was identified in a patient with language and motor delay (Gamage *et al.*, 2013), and the gene has also been implicated in autism, when a missense variant was found in a proband, although an unaffected sibling also had it (Vaags *et al.*, 2012).

In both cases, the wild-type allele was associated with reduced levels of transcripts. If we consider these two associations significant, then, based on the q-value, ~78% of them are expected to be false positives. Also, it should be noted that the number of people in the analysis and the number of people carrying the variant were both very small. The q-values for the remaining transcripts equalled 1.

5.3.2: A Correlation Analysis of Gene Expression Data

Expression levels for 21,799 transcripts were compared with that of the *NFXL1* transcript in lymphoblastoid cell lines across 609 samples from the following populations: MEX (Mexican ancestry in Los Angeles, California), JPT (Japanese in Tokyo, Japan), CHB (Han Chinese in Beijing, China), YRI (Yoruba in Ibadan, Nigeria), GIH (Gujarati Indians in Houston, Texas), LWK (Luhya in Webuye, Kenya), and MKK (Maasai in Kinyawa, Kenya) (six MEX individuals carrying the *NFXL1* variant and three MEX samples with unknown genotypes for the position of the variant were excluded). Pearson correlation coefficient (r) values were calculated for the expression levels of each transcript against the expression levels of the *NFXL1* transcript. The interpretation of the results is difficult, since thousands of correlations had $|r|>0.9$, and 314 had $|r|>0.99$. However, one interesting result is that among the 21,799 transcripts tested against *NFXL1*, the highest correlation (r=0.9946) was between *NFXL1* and *NFX1*. This is, perhaps, not very surprising, given the structural similarities between the proteins the two genes code for, which both have NF-X1 type zinc finger domains. This result suggests that, at least in lymphoblastoid cells, *NFX1* and *NFXL1* have similar expression patterns, as far as this analysis can show. *NFX1* has been shown to regulate HLA class II genes (HLA-DR genes in particular): it has been shown to be a repressor of *HLA-DRA* (Song *et al.*, 1994). This is interesting in light of the potential involvement of HLA genes, including *HLA-DRB1*, in SLI, as suggested from the results of the HLA association analyses discussed in Chapter 4. Since the structure and expression patterns of *NFX1* and *NFXL1* are similar, it is possible that they share some functional properties as well.

Since the gene expression data were generated using lymphoblastoids, which are, in essence, immune system cells, these results can may be very relevant to *NFXL1*, as its paralogue codes for a transcription factor that is known to regulate genes involved in immune response.

5.4: A Targeted Association Analysis of the Linkage Region on Chromosome 7

In a previous report, which used 6009 SNPs in a linkage analysis, a region on chromosome 7q consistently showed linkage across linkage analyses which used different methods, one of which resulted in a LOD score of 6.73 (Villanueva *et al.*, 2011). After that study had been published, the islanders' samples were genotyped on a new array designed for South-American populations, and on a custom array as well (see Chapter 2 for further information on the arrays). Genotypes for SNPs in the linkage region on chromosome 7q were used to perform a targeted association analysis. The association analysis included 15,277 SNPs across 47.5 Mb of chromosome 7q (111,499,581-159,018,204, hg19) and was performed with MQLS. This analysis used the samples, pedigree information and affection statuses used in the linkage study, and family connections between 330 individuals were used to correct for relatedness. The most associated SNP, rs80161634, had a p-value of 1.84×10^{-5} (with q-value of 0.281) and fell within *CNTNAP2*, a gene that has been previously implicated in language impairment and other neurodevelopmental disorders. This G/A SNP is intronic. The frequency of the A allele in affected individuals was 23.2%, which was ~4 times higher than its frequency in unaffected individuals (5.6%). The frequencies of the A allele in the South-American 1000 Genomes populations CLM (5.8%), MXL (2.3%), and PUR (0.9%), are closer to its frequency in unaffected islanders. In these South-American populations, the SNP was not in LD with rs17236239, the most highly-associated SNP in *CNTNAP2* in the study by Vernes *et al.* (2008) (r^2=0.001, as estimated with PLINK). The two associations may represent independent causal variants in those two populations, but it is also possible that both of these SNPs are in LD with a third SNP, which is the true causal variant increasing the risk of SLI in both populations. In such a case, rs80161634 would be in LD with that SNP in South-Americans, and rs17236239 would be in LD with that SNP in Europeans.

5.5: Summary of Chapter 5

A variant in the *NFXL1* gene, which is a non-synonymous coding change, was found to be significantly more frequent in islanders affected by SLI than in unaffected islanders. This

variant was found only in South American populations when examining the 1000 Genomes project data and independent samples from Colombian and European individuals. Where the variant was found in other South American populations, its frequency was always low and close to that of the unaffected islanders. The frequency of the variant in families of children who have language problems in the context of other developmental problems was, again, closer to that in the unaffected islanders than that in affected islanders. This suggests that the variant is only associated with SLI. Moreover, no individual in any population studied was homozygous for the variant, even though the frequency of the variant, even in controls, is at least 5%. This may suggest that the amino acid substitution affects the function of the protein. An analysis of gene expression data of seven HapMap3 populations showed that the expression pattern that was the most-highly correlated with the expression pattern of *NFXL1* was that of *NFX1*. NFX1 is a repressor of HLA class II genes and shares structural similarities with NFXL1. A quantitative analysis of transcript levels in individuals from the MEX population, some of whom carry the *NFXL1* variant, found suggestive association between the variant and increased expression levels of two genes. One of those genes, *PDE1C*, has been implicated in neurodevelopmental disorders and is a FOXP2 target. However, the sample size of the quantitative analysis was very small due to the low frequency of the *NFXL1* variant, and, therefore, the power of the test was low. Additionally, FDR analysis showed that a high proportion of false positives among the top associations is expected. *NFXL1* is an interesting candidate gene for SLI, and it will be the subject of further investigations discussed in this thesis. Additionally, a targeted association analysis of the linkage region on chromosome 7 found that the strongest association was with a SNP in *CNTNAP2*. While the FDR q-value for that association was higher than 0.05, it is possible that that was the result of having a small sample size combined with testing a large number of SNPs (as the linkage region was 47.5 Mb long). It is nonetheless interesting that *CNTNAP2* was implicated in this study of the Robinson Crusoe Island cohort as well as in a previous SLI association study, which used British families (Vernes *et al.*, 2008).

Chapter 6: Sequencing of SLI Candidate Genes in Language-Impaired Individuals

This chapter discusses the sequencing of four SLI candidate genes. Three of those genes, *ATP2C2*, *CMIP*, and *CNTNAP2*, were identified in previous studies (discussed in Chapter 1), and their sequenced regions included LD blocks between associated SNPs (for *ATP2C2*: chr16:84446210-84488988; for *CMIP*: chr16:81639990-81691013; for *CNTNAP2*: chr7:147500441-147652054, hg19). The coding sequences of *NFXL1*, the gene identified in the association study of the exome sequencing variants described in this thesis, were also sequenced. In that respect, the aim of the partial sequencing of the former three genes and that of the sequencing of the coding regions of *NFXL1* are different; the former was to identify the genetic variants that drove the associations identified in previous studies, be they intronic or exonic, and the latter was to identify non-synonymous coding changes in *NFXL1*. The DNA used for the PCR and sequencing reactions was pooled DNA, consisting of 14 pools of 10 SLIC probands each. Information about the samples used in this analysis can be found in Chapter 2. The output of the sequencing was in the form of BAM files, which are the binary versions of sequence alignment/map (SAM) files. These files contain all the reads of short DNA fragments aligned against a reference sequence (hg19). In order to identify positions at which the DNA base differs from the base in the reference sequence, algorithms known as variant callers are used. Three different variant-calling tools were assessed, one of which was designed specifically for use with pooled data. The other two were used with settings that allow for very high sensitivity in terms of read-depth. After choosing the tool which identified the highest number of non-synonymous coding variants, following the annotation of the data, all of said variants were validated using Sanger sequencing, and the results were compared to the predictions of the algorithm based on the quality score generated for each variant. Additionally, DNA samples from all family members of individuals carrying the variants from whom DNA was available were sequenced as well. Some of the non-synonymous coding variants were novel and/or very rare.

One other candidate gene discussed in this chapter was identified in a GWAS of the Robinson Crusoe Island cohort. Non-synonymous coding variants in this gene were later

identified in several SLI probands through exome sequencing. Those variants were validated by Sanger sequencing in those probands and in their families, and the pedigrees were analysed for co-segregation.

6.1: A Comparison of Three Variant-calling Algorithms

Three variant-calling algorithms were used to process the BAM files: Syzygy, SAMtools, and Platypus. Syzygy was designed to be used with sequences derived from pooled samples and was run with the default settings. The input for Syzygy includes the number of chromosomes in the pool, so it can estimate whether a variant with a low number of reads represents a false read or a rare variant. SAMtools and Platypus were designed to work with single-sample sequences, i.e. they expect a frequency of reads of either 100% or 50% for each present allele and call variants according to that expectation. They were therefore run with settings allowing for increased read-depth and/or for bases with low quality scores to be considered.

The following commands were used with SAMtools:

mpileup -BQ0 -d10000000 (disables base alignment quality calculation, counts even low-quality bases and processes all reads).

When creating the VCF file from the BCF (binary variant call format) generated in the previous step:

vcfutils.pl varFilter -D10000000 (increases the maximum read-depth to 10,000,000).

The following commands were used with Platypus:

callVariants --mergeClusteredVariants=1 --maxReads=200000000 (allows calling in divergent samples and increases the maximum allowed coverage in the region being analysed to 200,000,000).

Taking into account only the variants that were found in one of the four aforementioned genes, as to remove reads from regions that were not supposed to be amplified or sequenced, Syzygy called the highest number of variants (4453) compared to

SAMtools (986) and Platypus (2582). In terms of non-synonymous coding changes across all four genes, Syzygy called all the variants called by the other two algorithms, and more (Table 6.1). Even with the parameters set to high sensitivity, SAMtools and Platypus did not call the putative rare variants called by Syzygy (see Table 6.2 for variant frequencies). As will be described in the next section, I checked for the presence of all of the non-synonymous variants in samples from the relevant DNA pools and concluded that Syzygy had provided the most accurate calls. Given this result and the fact that Syzygy was the only variant-caller out of the above three designed to be used with pooled samples, I decided to use the Syzygy output in the further analyses of these data.

Table 6.1: A comparison of the non-synonymous coding changes called by the three algorithms.

Chromosome: Position (hg19)	Gene	Called by Syzygy?	Called by SAMtools?	Called by Platypus?
4:47877165	*NFXL1*	Yes	No	No
4:47887652	*NFXL1*	Yes	No	No
4:47898575	*NFXL1*	Yes	No	No
4:47901468	*NFXL1*	Yes	No	Yes
4:47901476	*NFXL1*	Yes	Yes	Yes
16:84456032	*ATP2C2*	Yes	No	No
16:84474484	*ATP2C2*	Yes	Yes	Yes
16:84476200	*ATP2C2*	Yes	Yes	Yes
16:84482217	*ATP2C2*	Yes	No	No
16:84482218	*ATP2C2*	Yes	No	Yes
16:84485677	*ATP2C2*	Yes	No	Yes
16:84486950	*ATP2C2*	Yes	No	No

6.2: Validation of Non-synonymous Coding Changes in SLI Candidate Genes

Syzygy called 12 non-synonymous coding changes in total: 5 in *NFXL1* and 7 in *ATP2C2*. Eight of those had a “pass” quality score. Table 6.2 includes the non-synonymous coding changes identified with Syzygy.

Table 6.2: Non-synonymous coding changes in *ATP2C2* or *NFXL1* identified through long-range sequencing.

Chromosome: Position (hg19)	Gene	Quality score (Syzygy)	Reference/ variant (+ strand)	Variant frequency (Syzygy)	Variant frequency in EVS (European-Americans)	Amino acid change	Predicted effect (MutationTaster)
4:47877165	*NFXL1*	FAIL	G/T	0.0517	Not found	T742K	Polymorphism
4:47887652	*NFXL1*	PASS	T/C	0.0035*	Found in one African-American individual, but not in European-Americans	T563A	Polymorphism
4:47898575	*NFXL1*	PASS	G/A	0.0071*	0.005	R432C	Disease-causing
4:47901468	*NFXL1*	FAIL	C/A	0.132	Not found	V249L	Disease-causing
4:47901476	*NFXL1*	PASS	G/A	0.3195	0.31	P246L	Polymorphism
16:84456032	*ATP2C2*	PASS	G/T	0.0035*	0.001564	G221W	Disease-causing
16:84474484	*ATP2C2*	PASS	G/A	0.05	0.070163	G411S	Polymorphism
16:84476200	*ATP2C2*	PASS	A/T	0.4309	0.375122	M466L	Polymorphism
16:84482217	*ATP2C2*	FAIL	A/C	0.0438	Not found	I528L	Polymorphism
16:84482218	*ATP2C2*	FAIL	T/C	0.1307	Not found	I528T	Polymorphism
16:84485677	*ATP2C2*	PASS	T/A	0.053	0.072406	L604Q	Disease-causing
16:84486950	*ATP2C2*	PASS	C/G	0.0035*	Not found	S651R	Disease-causing

* The frequency of the novel or rare variants was validated through Sanger sequencing of samples in the DNA pools in which they were found, but the remaining samples were not sequenced.

As can be seen from Table 6.2, several variants were either novel or extremely rare in the general population (NB the number of alleles used by the EVS is higher than 8000, whereas the number of alleles of SLI probands was 280). Primers were designed to sequence the regions containing the variants (see Appendix A). All samples in at least one pool in which each variant was found were sequenced. The results confirmed the predictions made by Syzygy as to the frequencies of the variants (in the pool sequenced) and the quality scores. Hereafter the variants will be referred to by the gene they fall in and their physical position e.g. ATP2C2-84486950 represents the variant in *ATP2C2* on chromosome 16 position 84486950. Of the 12 variants called, all the rare variants were validated: one person was heterozygous for the ATP2C2-84486950 variant; one person was heterozygous for the ATP2C2-84456032 variant; one person was heterozygous for the NFXL1-47887652 variant; two people were heterozygous for the NFXL1-47898575 variant.

6.3: Validation of Rare or Novel Non-synonymous Coding Variants in SLI Families

The four coding variants that were either rare or novel were validated in the families of the probands who carried them, where samples from family members were available (no samples were available from the family members of the person carrying the ATP2C2-84456032 variant). Scores of language tests and clinical descriptions for family members (if available) are given in Table 6.3. Pedigrees which include genotypes and descriptions of language-related disorders (if known) are shown in Figure 6.1, and sequence examples for the rare or novel non-synonymous coding variants in *ATP2C2* and *NFXL1* are given in Figure 6.2.

Table 6.3: Families with rare or novel non-synonymous coding changes in *ATP2C2* or *NFXL1* identified through long-range sequencing.

Variant	Family	Individual	Genotype	ELS	RLS	NWR	Clinical description
ATP2C2-84486950	1	1.1	CG*			94	Speech disorder
		1.2	CC				Family history of speech and language problems
		1.3	CG*	64	63	96	Went to a special school
		1.4	CC	72	85	102	Went to a mainstream school
		1.5	CC	64	76	62	Went to a special school
NFXL1-47898575	2	2.1	Unknown				
		2.2	GG				
		2.3	GA*	72	76	82	
		2.4	GA*	85	86	89	
	3	3.1	GA*			64	Paternal history of speech and language delay
		3.2	GG			88	Dyslexia
		3.3	GA*	99	110	79	Went to a special school; expressive language problem
		3.4	GA*	50	70	72	Went to a special school
		3.5	GG	54	83	85	Speech problem (affected less severely than the proband)
NFXL1-47887652	4	4.1	TC*			91	
		4.2	TT			84	
		4.3	TC*	78	84	95	Referred to a language unit
		4.4	TT	73	83	92	Went to a mainstream school; literacy problem
		4.5	TC*	73	83	99	Went to a mainstream school

* Variant allele

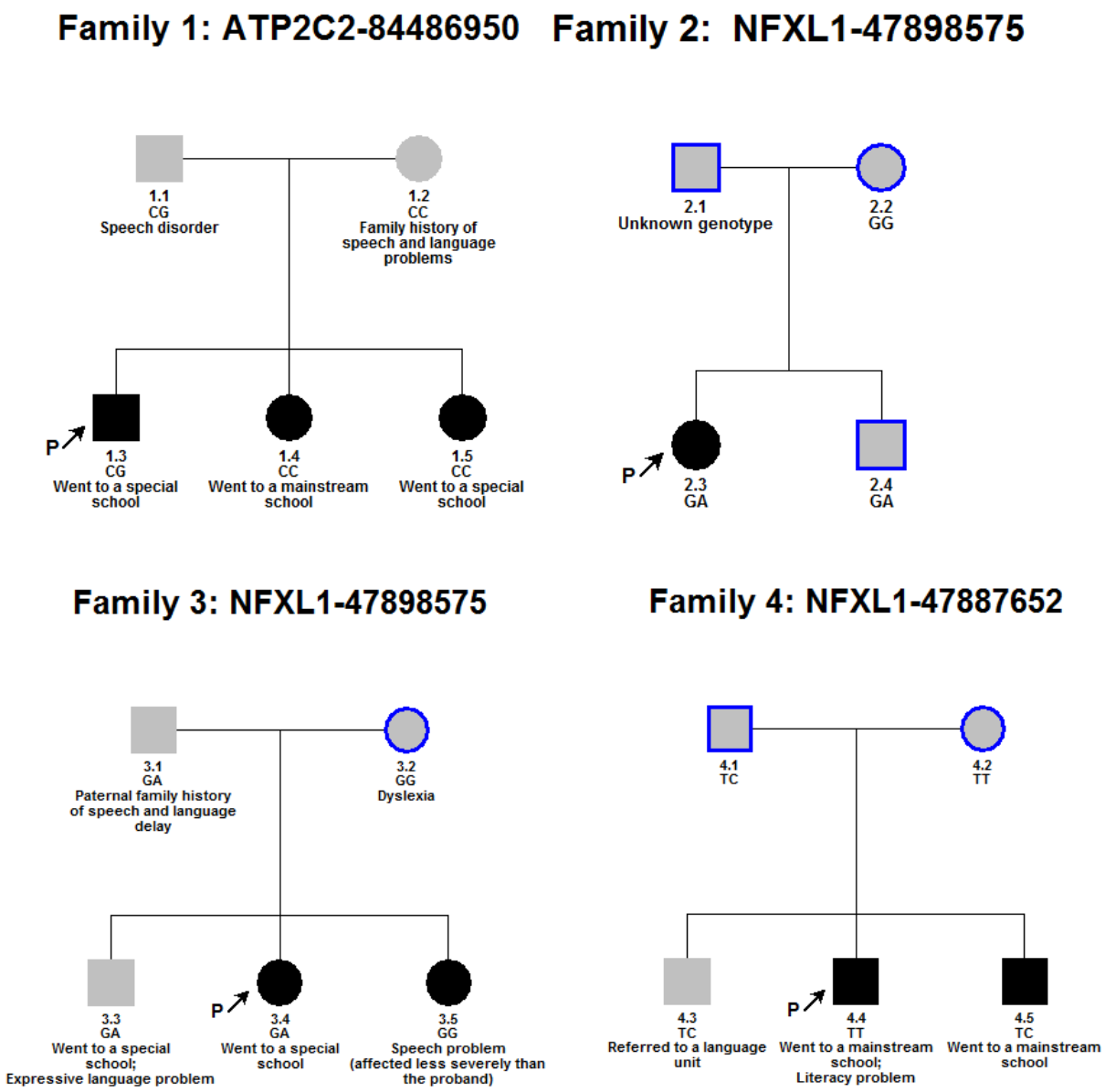


Figure 6.1: Pedigrees of families with non-synonymous coding changes in *ATP2C2* or *NFXL1* identified through long-range sequencing (the proband in each family is marked with an arrow; individuals with SLI, based on the criteria outlined in Section 2.6.1, are marked in black; individuals who are unaffected are circled in blue and are not shaded; individuals with an atypical language phenotype based on the clinical description are marked in grey; individuals with an unkown phenotype are marked in grey and circled in blue).

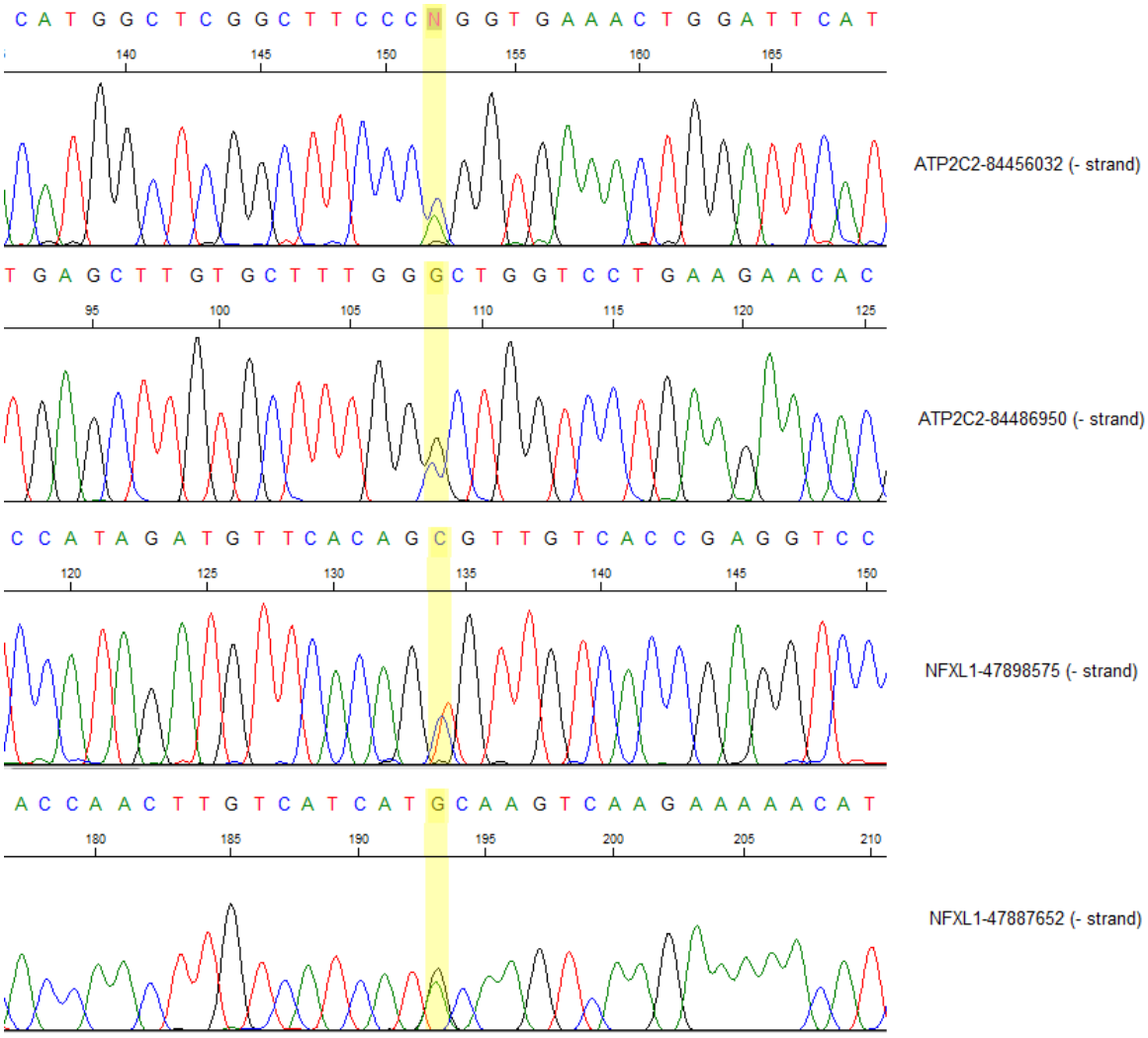


Figure 6.2: Sequences around rare or novel coding variants in *ATP2C2* or *NFXL1* (the positions of the variants are highlighted).

Of the rare variants, the ATP2C2-84486950 variant was inherited by the proband from his father, who suffered from a speech-related problem, in Family 1. While the mother, who did not carry the variant, is indicated as having a family history of speech and language problems, one sibling, who did not have the variant also had low language scores. Notably, compared to that sibling, the father and the proband had high NWR, so that sets them apart from that sibling. However, since *ATP2C2* was associated with NWR, it is unlikely that this variant alone can account for the language impairment in this family, despite the fact that it was predicted to be disease-causing. Moreover, another sibling, who is affected based on her ELS, does not carry the variant. That prediction is based on a high likelihood of a splice site change and the fact that the nucleotide at that position is highly conserved (phastCons score of 1). This variant was not found in EVS.

The NFXL1-47898575 variant was present in two families, Family 2 and Family 3. Family 2 is not informative, since the variant appears to have been inherited by the proband from the father, for whom no phenotypic or genetic data were available (the mother does not have the variant, but both children do, so it is not likely to be a de novo mutation). In Family 3, a proband and a sibling inherited the variant from the father. Both the proband and said sibling went to a special school. Another sibling, though, who was diagnosed with a speech problem, did not carry the variant. While the proband had low language scores across all three language traits, the two siblings scored very differently on different language tests. One consistent finding is that only the individuals carrying the variant had NWR below 1 SD below the mean (the mean was 100 and the SD was 15, for all traits). This is also the case in Family 2, in which the proband had a NWR of 82. This variant co-segregates with low NWR in Family 3. It is predicted to be disease-causing based on potential splice site changes and a phastCons score of 1 for the nucleotide at that position. This variant was found in EVS with a frequency of 0.5%.

The NFXL1-47887652 variant present in Family 4 was inherited by the two siblings from the father. This is one of the cases described in Chapter 2, namely, the one in which a sample from a sibling (who was referred to a language unit) was added to the DNA pool rather than a proband. Another sibling who had the variant went to a mainstream school but

had low language scores nonetheless. The proband in this family did not have the variant. The variant is predicted to be a polymorphism, and there appears to be no co-segregation of the variant with any language phenotype in this family. This variant was found in EVS, but not in European-Americans.

6.4: Intronic Variants in SLI Candidate Genes

In addition to the aforementioned non-synonymous coding changes in *APT2C2* and *NFXL1*, 4441 intronic variants in all four candidate genes were identified (but not validated i.e. Sanger sequencing was not performed). As the focus of sequencing *NFXL1* was on its coding regions, only intronic regions between exons amplified together in one PCR were captured. For *ATP2C2*, *CMIP*, and *CNTNAP2*, only the regions of LD surrounding the reported associations were sequenced, as discussed previously. It is harder to assess the effect of intronic changes, rather than coding changes, on protein function, although they are likely to be relevant in the context of complex disorders, as discussed in Chapter 1. In order to test for association between such changes and SLI, Fisher's exact test was used with data from population controls and the SLI probands. All variants used in this analysis passed the quality test as determined by Syzygy. The final list of variants that were also known SNPs contained mostly intronic changes, but also included several of the aforementioned non-synonymous coding changes, if they were present in the control data. The final list of variants included SNPs for which allele frequencies were available through the UK10K data of 2,000 British controls from the ALSPAC cohort (SNP data were extracted based on the physical positions in the VCF files). Allele designations were consistent across datasets (the UK10K ALSPAC data included the minor and major alleles, the MAF, the total allele count, and the minor allele count). Allele counts were calculated from these data (in the SLI sample, allele counts were calculated by multiplying the allele frequencies by 280, twice the number of samples in the pools, whereas the ALSPAC data had 3854 alleles). A/T and G/C SNPs were excluded from the analysis to avoid strand-flip errors, as well as positions for which the variant was not found in ALSPAC. In total, 1291 variants were included in this analysis.

The non-synonymous coding variants did not show significant association. That included the rare variants in *ATP2C2* and *NFXL1*. It should be noted, however, that they were very rare and may therefore be not suitable for such analyses.

As for the intronic variants, many of them were significantly more frequent either in the probands or in the population controls. Since the regions sequenced were based on LD blocks around previously-associated SNPs, it is perhaps not surprising that many of these variants should show strong association. 93 SNPs had q-values of less than 0.05. An intronic T/C SNP in *CNTNAP2*, rs2707581, had the smallest p-value (4.24×10^{-75}), the variant C being ~8 times more frequent in controls (60.6% in controls compared to 7.14% in probands). In the GBR population (British in England and Scotland) from the 1000 Genomes cohort, this SNP was not in LD with rs17236239, the most-highly associated SNP from the study by Vernes *et al.* (2008) (r^2= 0.041, as estimated with PLINK). Other SNPs (in *ATP2C2*, *CMIP* and *NFXL1*) showing very strong association and similar trends (albeit with higher, but nonetheless significant, p-values) were also identified. Associations with q≤0.05 can be found in Appendix C. These results support the previous findings with regards to the involvement of these genes in SLI. It should be emphasised, however, that the intronic variants were not validated, and the allele counts may be inaccurate due to high-throughput sequencing errors. Moreover, I did not have access to the genotype data for the ALSPAC individuals, which meant that I was not able to employ any quality control measures on those data or on the ALSPAC samples. If there were genotype errors or population stratification in the ALSPAC cohort, then that could have an effect on the associations identified in my study. Therefore, the variants should be validated in SLI cases and in controls in order to support the findings of this study.

6.5: Analysis of Non-synonymous Coding Variants in *ZNF423*

An association with the zinc finger protein 423 (*ZNF423*) (OMIM#604557) was identified in a GWAS I had run with the Robinson Crusoe Island samples using the SNP array that had previously been used in the linkage analysis. The results of that GWAS can be found in Appendix D. *ZNF423* is a transcription factor that has been implicated in several studies of neurodevelopmental and brain disorders. The mouse orthologue is important in the development of neuronal and glial precursors in the developing brain (Alcaraz *et al.*, 2006). In

humans, one study reported a microdeletion that included *ZNF423* in two patients who had developmental delay and intellectual disability (Ballif *et al.*, 2008). Another deletion in *ZNF423* was reported for a boy who had psychomotor delay and microcephaly (Shoukier *et al.*, 2012). A duplication in *ZNF423* has also been reported: one study reported the case of a 4 year old boy with a gain of a region on chromosome 16q which included *ZNF423*; the boy suffered from intellectual disability and verbal dyspraxia, among other things (Zerem *et al.*, 2011). *ZNF423* is therefore a good candidate gene for SLI and was fully sequenced in an exome sequencing study of 49 SLIC probands. Five non-synonymous coding variants in *ZNF423* were identified (Table 6.4), and those were validated in the probands and their families, where DNA samples were available (Table 6.5, Figure 6.3). Sequence examples for the rare or novel non-synonymous coding variants in *ZNF423* are given in Figure 6.4.

Table 6.4: Non-synonymous coding variants in *ZNF423* found in SLI probands.

Chromosome: Position (hg19)	Gene	Reference/ variant (+ strand)	Variant frequency in EVS (European-Americans)	Amino acid change	Predicted effect (MutationTaster)
16:49669923	*ZNF423*	G/A	0.009535	A1047V	Disease-causing
16:49671081	*ZNF423*	T/G	0.000581	K661T	Disease-causing
16:49671177	*ZNF423*	T/C	0.031977	N629S	Disease-causing
16:49671396	*ZNF423*	G/A	Not found	T556M	Disease-causing
16:49672599	*ZNF423*	C/T	Not found	R155K	Disease-causing

Table 6.5: Families with non-synonymous coding changes in *ZNF423*.

Variant	Family	Individual	Genotype	ELS	RLS	NWR	Clinical description
ZNF423-49671396	5	5.1	GA*				
		5.2	GG			61	
		5.3	GA*		72	76	
		5.4	GA*	91	93	83	Borderline language phenotype
ZNF423-49669923	6	6.1	Unknown				
		6.2	GA*			84	Language delay
		6.3	GA*	72	72	73	
		6.4	GA*	73	87	65	Low language test scores; low PIQ
		6.5	GG	110	91	82	Borderline language phenotype
ZNF423-49672599	7	7.1	CC			94	
		7.2	CT*			94	
		7.3	CC	99	103	114	
		7.4	CT*	59	67	69	
ZNF423-49671081	8	8.1	TG*				
		8.2	TT				
		8.3	TT	91	112	90	
		8.4	TG*	50	50		
ZNF423-49671177	9	9.1	TC*			61	
		9.2	TT			74	
		9.3	TC*	62	95	65	
		9.4	TC*	112	101	96	
	10	10.1	TT			104	
		10.2	TC*			91	
		10.3	TC*	72	72	92	
		10.4	TC*	86	93	103	Borderline language phenotype
		10.5	TT			116	
	11	11.1	TC*				
		11.2	TT				
		11.3	TC*			98	
		11.4	TT	82	91	73	Dyslexia
		11.5	TC*	72	103	73	
	12	12.1	TT				
		12.2	TC*			88	Borderline language phenotype
		12.3	Unknown				
		12.4	TT	78	83	94	Borderline language phenotype
		12.5	Unknown				
		12.6	TC*	59	74	62	

* Variant allele

As can be seen from Table 6.4, apart from ZNF423-49671177, the non-synonymous coding changes found are either very rare or novel in a population database, and all variants were predicted to be disease-causing by MutationTaster.

There is a possible co-segregation of the ZNF423-49671396 variant and a language phenotype in Family 5, but no information on the father's phenotype was available. The prediction that this substitution is disease-causing was based only on the amino acid change at that position (i.e. no other elements in the gene or protein seem to be affected).

ZNF423-49669923 also shows a possible co-segregation in Family 6, with the variant being inherited by the proband and a sibling with low language test scores from the mother, who suffers from language delay. It should be noted, however, that another sibling, who was borderline for language problems, did not have the variant, and that the father's genotype and phenotype were unknown. The prediction that this substitution is disease-causing was based only on the amino acid change at that position (i.e. no other elements in the gene or protein seem to be affected).

ZNF423-49672599 did not show any co-segregation with any of the known language traits available for both parents and children in Family 7, and with the lack of other phenotypic data, there is no evidence that the variant is involved in the language impairment in that family. The prediction that this substitution is disease-causing was based only on the conservation of the nucleotide at that position (phastCons score of 1) and splice site changes.

Similarly, due to the lack of phenotypic data for the parents in Family 8, it is not possible to determine if the ZNF423-49671081 variant co-segregates with the language impairment. The prediction that this substitution is disease-causing was based only on the conservation of the nucleotide at that position (phastCons score of 1) and splice site changes.

Four families had the ZNF423-49671177 variant. In Family 9, the variant does not appear to be co-segregating with any language trait. In Family 10, the proband and a sibling who is borderline for language problems both have the variant, and a sibling with an unknown language phenotype does not. However, there were no data available regarding the parents' phenotypes. Based on the information available, the variant may be co-segregating with the

language impairment in that family. In Family 11 there does not seem to be co-segregation of the variant with the language impairment, but information on the parents' and one sibling's language scores was unavailable. In Family 12 there seems to be no co-segregation as both the mother and a sibling were borderline for language problems, but only the mother and the proband have the variant. The prediction that this substitution is disease-causing was based only on the conservation of the nucleotide at that position (phastCons score of 1) and splice site changes.

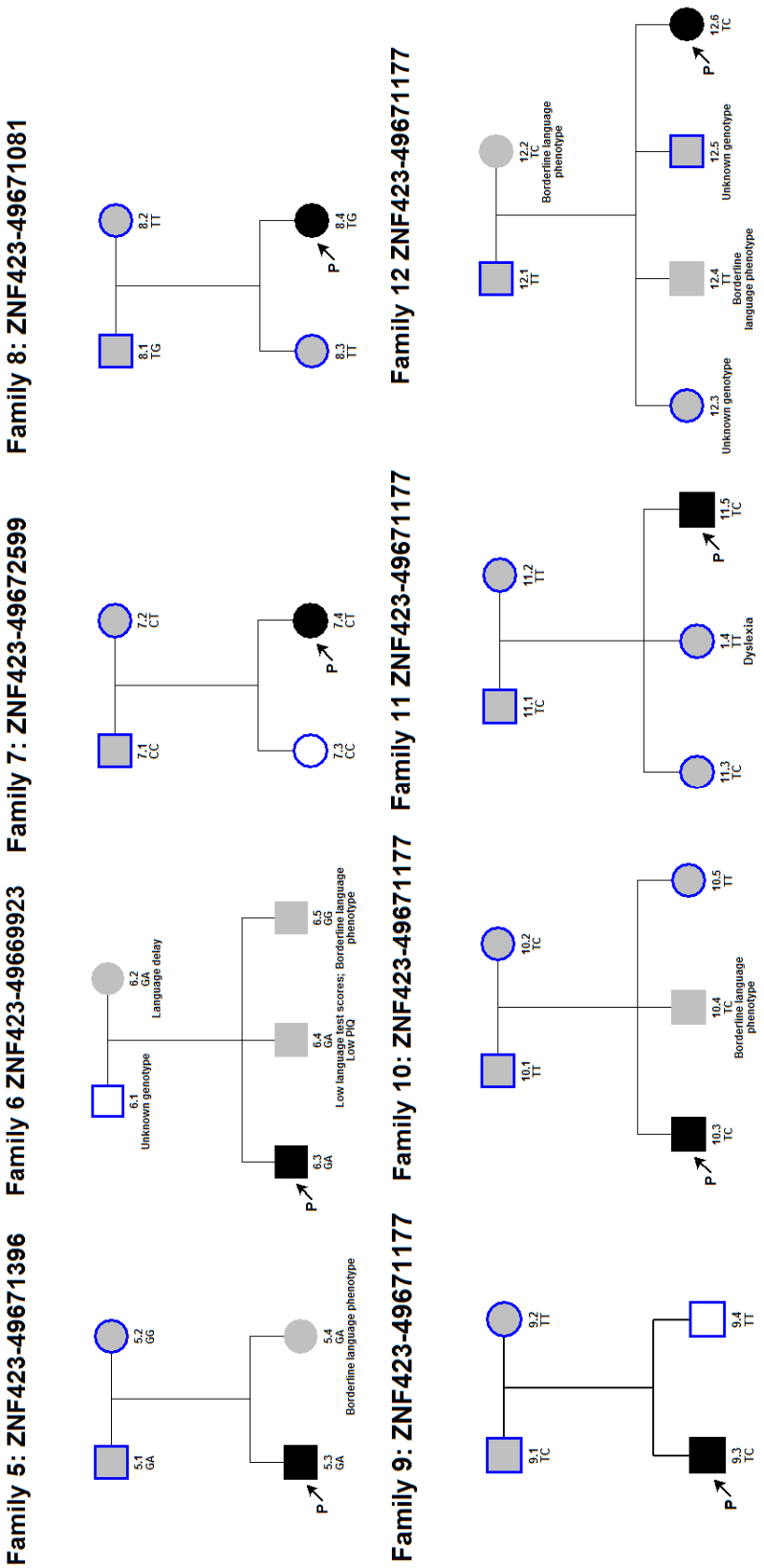


Figure 6.3: Pedigrees of families with non-synonymous coding changes in *ZNF423* (the proband in each family is marked with an arrow; individuals with SLI, based on the criteria outlined in Section 2.6.1, are marked in black; individuals who are unaffected are circled in blue and are not shaded; individuals with an atypical language phenotype based on the clinical description are marked in grey; individuals with an unkown phenotype are marked in grey and circled in blue).

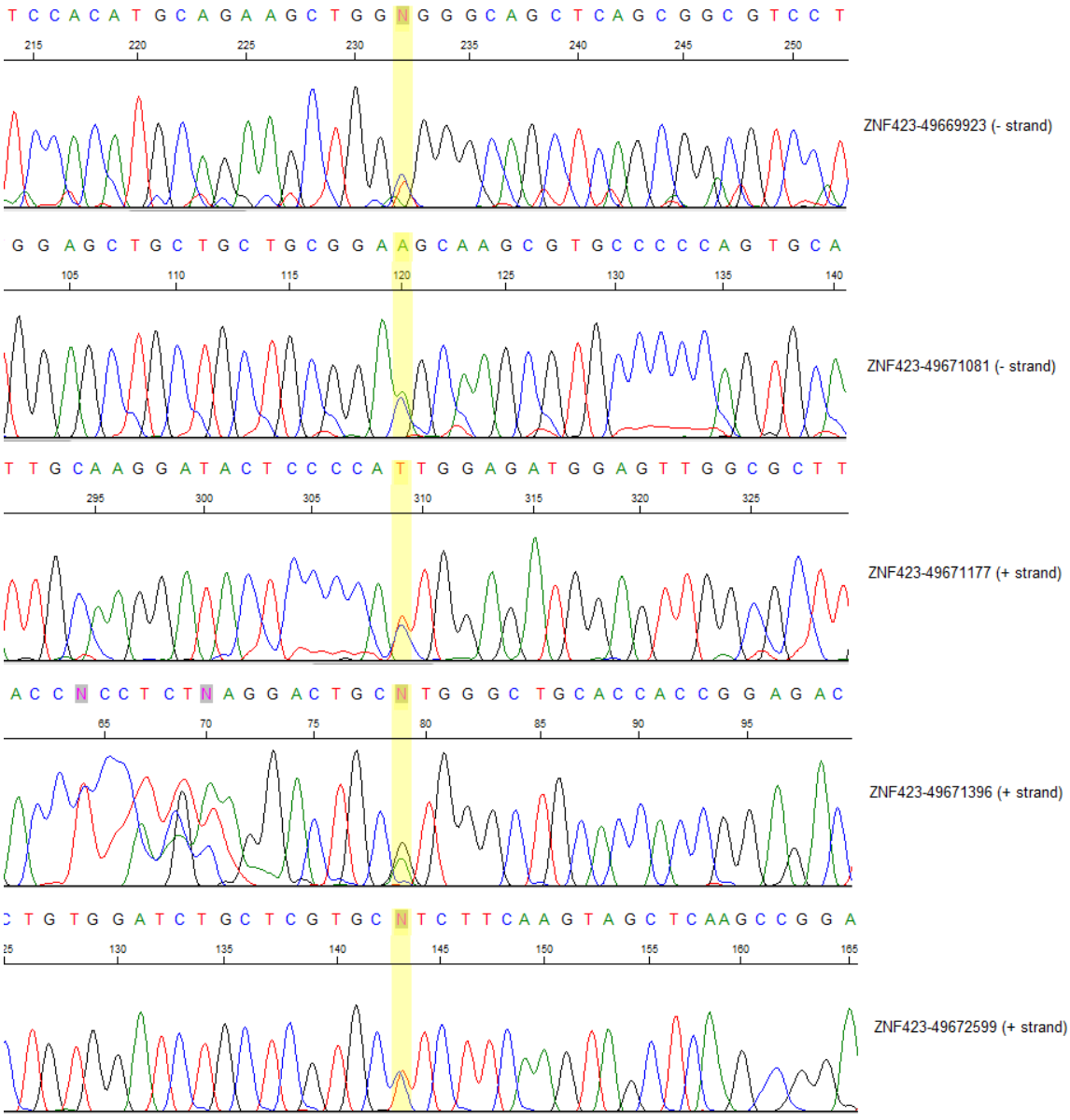


Figure 6.4: Sequences around rare or novel coding variants in *ZNF423* (the positions of the variants are highlighted).

6.6: Summary of Chapter 6

This chapter discussed the study of five SLI candidate genes, namely, *ATP2C2*, *CMIP*, *CNTNAP2*, *NFXL1* and *ZNF423*, which were implicated in SLI through linkage and/or association analyses. Techniques such as high-throughput sequencing and exome sequencing were used together with variant-calling tools and statistical analyses to identify variants in these genes which contributed to the risk of having language impairment.

In *ATP2C2*, one novel non-synonymous coding variant predicted to be disease-causing was identified in one family, but it did not seem to co-segregate with the language impairment in that family.

In *NFXL1*, one rare non-synonymous coding variant predicted to be disease-causing co-segregated with low NWR in one family.

In *ZNF423*, co-segregation could not be ruled out for three variants which were very rare or novel. These variants were all predicted to be disease-causing. However, in many cases the phenotypic data for the families were incomplete.

These variants were so rare that each one individually cannot account for the language impairment in the families in which they were found, given that different families carried different variants. However, one plausible model is that of increased risk caused by a mutational load across genes that are relevant for language, which can account for rare variants that are unique to a given family or otherwise very rare. Evidence for this model has been found in many studies of the genetics of autism (Buxbaum, 2009).

Additionally, intronic variants in regions where associations with SLI had previously been identified were also examined. Several intronic variants showed very high associations with increased or decreased risk of language impairment (minimum p-value of 4.24×10^{-75}). These results support the involvement of those genes in SLI.

Chapter 7: A Functional Study of *NFXL1*

Non-synonymous coding variants in *NFXL1* have been found to be associated with SLI in the Robinson Crusoe population, as described in Chapter 5, and in some British SLI families, as described in Chapter 6. This gene has not been previously studied at the functional level, and, therefore, I decided to perform preliminary functional studies of the gene in order to gain some insight into the mechanisms through which it may be involved in SLI. The NFXL1 protein is predicted to be DNA-binding, based on sequence similarities with NFX1. I examined *NFXL1* expression in brain tissues and potential differences between splice variants expressed in different tissues (including the brain). I carried out a functional study of the effects of over-expressing *NFXL1* on candidate target genes (based on examples of NFX1 target genes) and an immunofluorescence experiment to examine the localisation of the NFXL1 protein in the cell.

7.1: *NFXL1* Expression in the Brain

The presence of *NFXL1* transcripts in brain RNA samples was tested using primers which spanned exons 3 and 4 (counting from the beginning of the cDNA), which had a very long intronic sequence between them (~5.5 kbp), so that the reaction would amplify only the transcript and not any genomic DNA that might still be present despite the steps taken to eliminate it. Primer sequences can be found in Appendix A. RNAs from the total brain, specific brain regions, and the colon (to represent a non-neural tissue) were used to produce cDNA through RT-PCR. The samples consisted of the following: brain, temporal frontal lobe, frontal lobe, parietal lobe, cerebral cortex, cerebellum, and colon. These brain regions were selected for an unrelated experiment, and the RNA samples were made available to me for the purpose of this experiment. As can be seen in Figure 7.1, the transcript was present in all samples, albeit at varying levels (in all cases the same amount of RNA was used in the RT-PCR, the same amount of cDNA was used in the PCR, and the same amount of PCR product was loaded onto the gel). The weakest band on the gel was that of the colon sample. Each of the bands for the specific brain regions was stronger than the band for the total brain

sample, indicating that the gene is not expressed in some brain regions, or that its expression is lower in some brain regions compared to the ones examined here.

In order to assess the differences more accurately, a qPCR was performed using TaqMan assays for *NFXL1* (Hs00541755_m1) and *GADPH* (glyceraldehyde-3-phosphate dehydrogenase) as a housekeeping gene (HKG) (Hs99999905_m1) i.e..it was used as a reference gene for quantifying expression levels. *GAPDH* is a common HKG, and previous studies have confirmed little within-tissue variation in its expression patterns (Barber *et al.*, 2005). This was also the case in my experiment, as the C_T values for *GAPDH* across all brain regions from which I had RNA were between 21.63 and 22.14 (the value for the total brain RNA sample was 24.06). Only one HKG was used due to low amounts of RNA. Expression levels were quantified using the $\Delta\Delta C_T$ method, with *GAPDH* as the reference gene and total brain RNA as the reference sample. Table 7.1 presents the C_T values and expression levels (relative to total brain RNA) of various brain regions. Figure 7.2 shows the expression levels in the various brain tissues compared to total brain RNA.

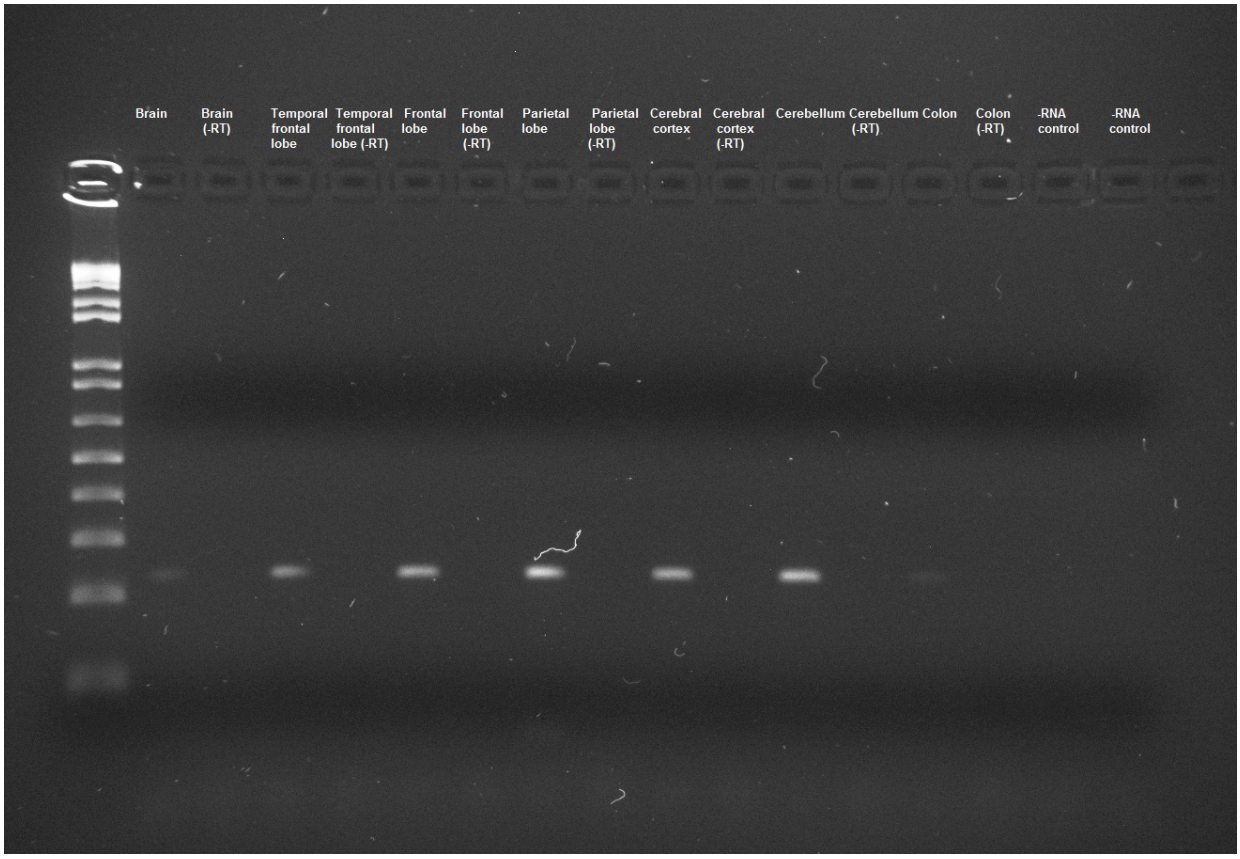


Figure 7.1: *NFXL1* expression in the brain (gel electrophoresis). The bands indicate the presence of the *NFXL1* transcript in the various samples -RT refers to reverse-transcriptase negative controls i.e. RNA samples to which the enzyme which transcribes DNA from RNA was not added before the RT-PCR.

Table 7.1: *NFXL1* expression in the brain.

Tissue	*NFXL1* C_T	*GAPDH* C_T	Expression level compared to total brain RNA
Total brain RNA	32.21	24.06	1
Temporal frontal lobe	30.69	21.84	0.62
Frontal lobe	28.25	21.93	3.56
Parietal lobe	28.35	21.63	2.69
Cerebral cortex	28.47	21.96	3.12
Cerebellum	27.91	22.14	5.21

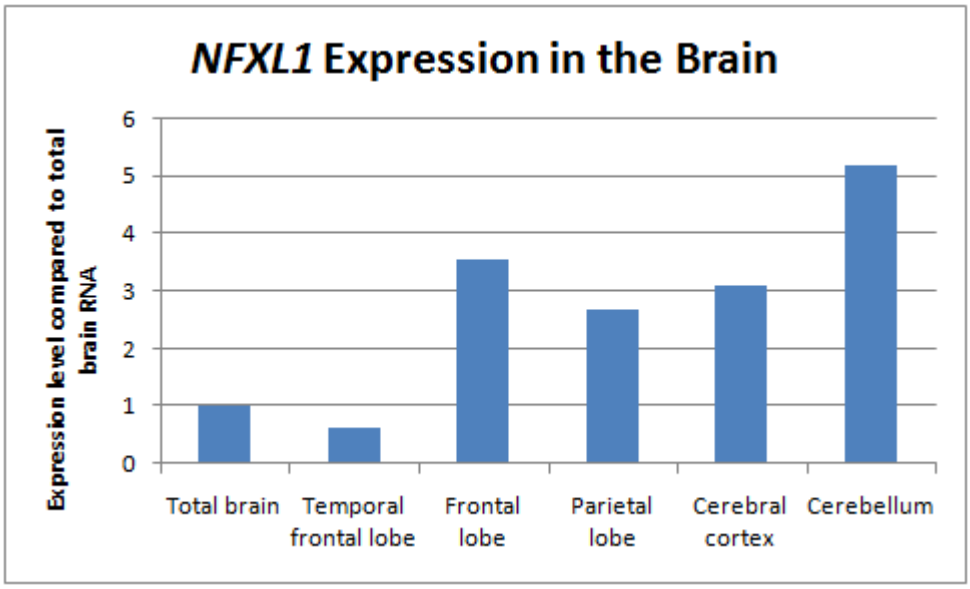


Figure 7.2: *NFXL1* expression in the brain (TaqMan).

The results of the gene expression analysis show that *NFXL1* is highly expressed in the cerebellum, compared to the other brain regions examined in this analysis. The cerebellum is mostly regarded as the part of the brain involved in motor control. In recent decades, however, many studies have implicated the cerebellum in language disorders and cognitive processes (De Smet *et al.*, 2013). A PET study reported cerebellar activation for visually-cued linguistic output, but the same activation pattern was not observed when auditory cues were presented, suggesting a role in a cognitive process associated with the task, rather than only with the motor response when producing the output (Petersen *et al.*, 1989). Another study reported the case of a patient with a right cerebellar infarction who developed problems with the morphological component of language (Silveri *et al.*, 1994). Similarly, Fabbro *et al.* (2000) reported the case of four patients with cerebellar lesions who had problems with morphology and syntax. The lesions were localised to different cerebellar structures in different patients, some of which are not correlated with language disorders (Fabbro *et al.*, 2000), suggesting, according to the authors, that the cerebellum might be involved in language control processes (i.e. the temporal and sequential organisation of systems involved in the production of language). In some cases, the language deficits improved slightly following surgical operations to treat the lesion, but in others it was not the case (Fabbro *et al.*, 2000). Abnormal cerebellar anatomy has also been implicated in neurodevelopmental disorders such as autism. Significantly lower Purkinje cell counts in the cerebellar hemisphere were observed in the brains of subjects with autism compared to controls (Ritvo *et al.*, 1986). According to Bauman and Kemper (2005), this anatomical observation is the most reproducible pathological observation in autopsied brains of subjects with autism. Purkinje cells have been mentioned earlier in the contexts of *FOXP2* expression and the potential involvement of the immune system in language disorders. Lastly, as discussed in Section 1.3.2.4, the cerebellum is involved in the procedural memory system, which some scholars believe is dysfunctional in many cases of SLI.

It should be noted that the results presented here show only a correlation between relatively higher expression levels of *NFXL1* and the potential involvement of a given brain region in language processing. Moreover, this experiment did not use biological replicates, as only one RNA sample from each brain region was available.

7.2: *NFXL1* Splice Variants in Various Tissues

In addition to examining the expression levels of the gene in various brain regions, I also checked whether different splice variants might be present in different brain regions and tissues. In order to cover the entire transcript, primers were designed so that overlapping fragments of the cDNA would be amplified in a PCR (Figure 7.3). Primer sequences can be found in Appendix A. The patterns on agarose gel were similar across all brain regions, indicating that the same transcripts are found in all brain RNA samples investigated. However, one reaction (the one amplifying the sixth fragment in Figure 7.3, which included exon 19) resulted in two bands on the gel (Figure 7.4). The two bands indicate that the samples contained two transcripts, one being shorter than the other, and that the difference in their sequence lay in the boundaries defined by the sixth pair of primers (Figure 7.3). The fragments were excised from the gel and sequenced. Exon 19 was missing in the shorter transcript. When doing a PCR using a primer placed in exon 19, only one band appeared on the gel, confirming the result (Figure 7.5). Figure 7.6 shows the difference in the sequences between the long and short fragments and the proteins the two transcripts code for, and Figure 7.7 shows the point at which the sequences start to differ. The long transcript codes for the NFXL1 protein with the canonical sequence of 911 amino acids. The short transcript codes for an isoform that would diverge from the long isoform in terms of its sequences after amino acid position 749 inclusive, as the reading frame shifts due to the lack of exon 19. Moreover, the frame shift introduces a stop coding at amino acid position 773, thus resulting in a protein 139 amino acids shorter. However, no such isoform is listed on UniProt. There is one other transcript listed on ENSEMBL. That transcript codes for a protein which has 733 amino acids, which makes it shorter than the normal protein. However, the alternative transcript which codes for the short isoform nonetheless contains exon 19 of the canonical sequence, although it appears after the first stop codon.

Figure 7.3: cDNA of the *NFXL1* gene (- strand) showing the overlapping fragments amplified in the PCR (the sequences which the primers bind are highlighted, coding regions are in dark blue, non-coding regions are in red, and exon boundaries are in light blue).

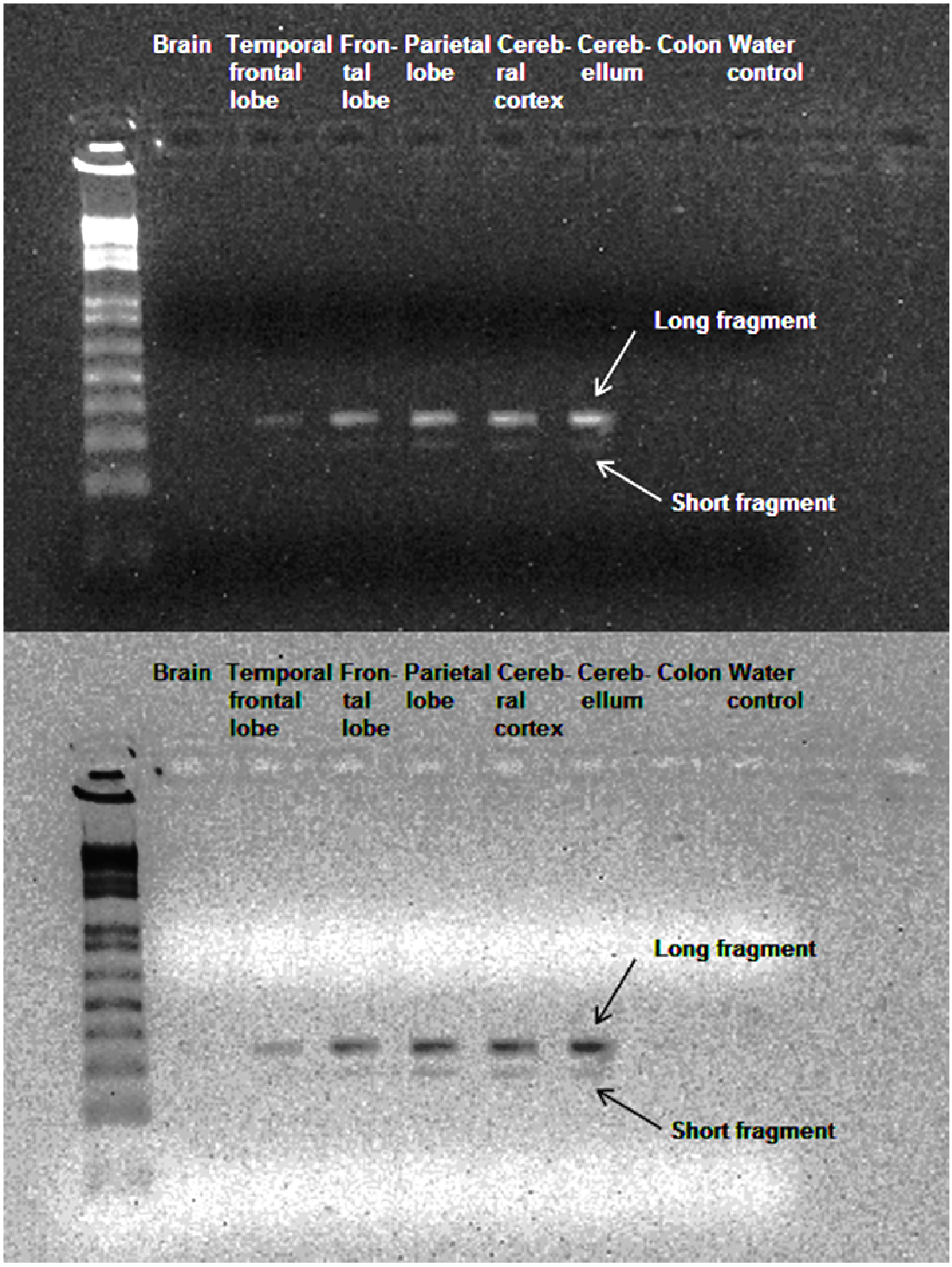


Figure 7.4: Amplification of the sixth fragment of the *NFXL1* cDNA. Two bands are detected for samples from brain regions. The negative of the gel photo is also presented in order to make it easier to spot the bands for the short fragment. The colour, contrast and brightness levels have been adjusted to make the weak bands more visible.

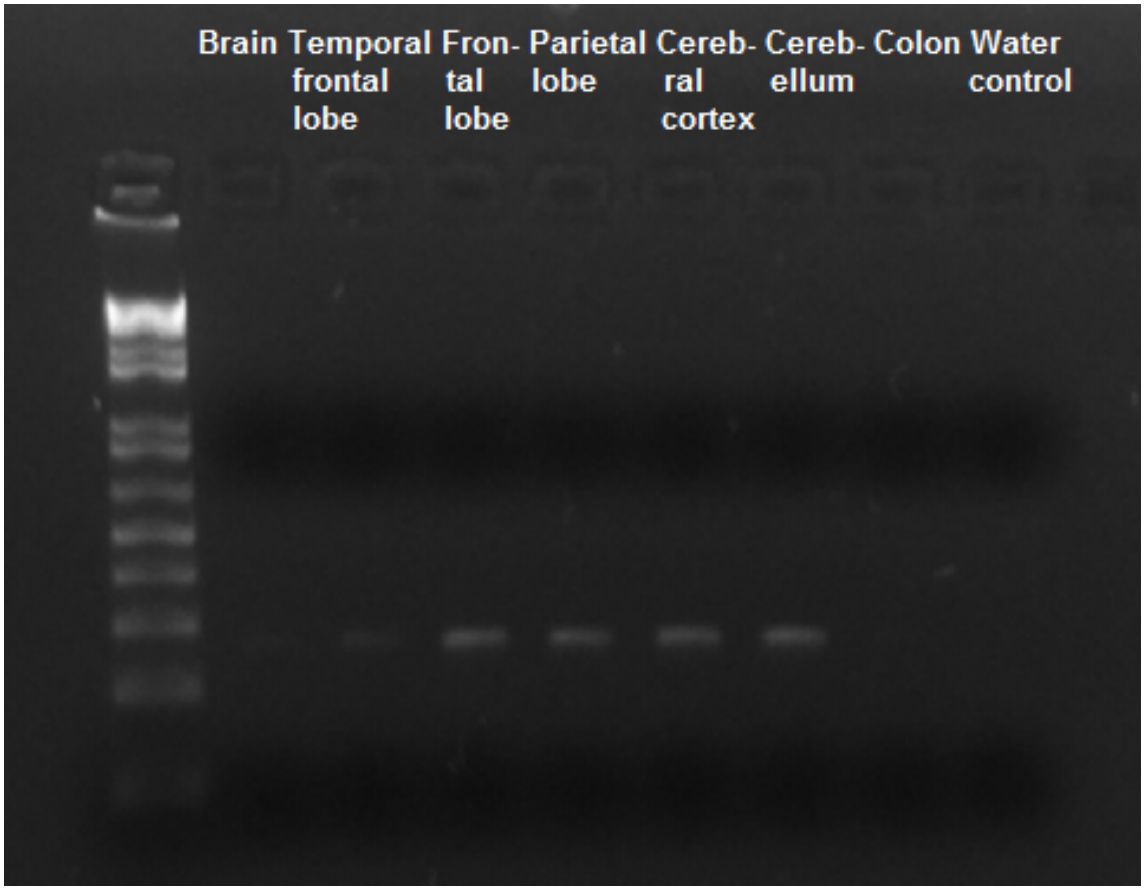


Figure 7.5: Amplification of part of the sixth fragment of the *NFXL1* cDNA using the alternative primer. When one primer in the pair binds to exon 19, only one fragment is amplified in the PCR.

Fragment 6 (- strand, exon boundaries in light blue)

GCCACAAAGTAACCAAAACTGATGGCTGCACTGGAAAAAACAAGGCTGGCCCAGAATGCCTTCATTGTGAGGAAG
GGTGCTCCAAGTCACGGCCACTAGGTTGTCTTCACCCATGTATTTTGCGATGTCACCCTGGAGAATGTCCACCTT
GTGTTCAGATGCTTAGAATAAAATGTCACTGTAAGATCACAAGCCTGTATGTGGAATGTAGAAAAATAACCACAG
CTGATGTAAATGAAAAGAACCTCCTCAGTTGTTGCAAAAATCAGTGCCCTAAAGAGCTTCCTTGTGGTCATAGAT
GCAAAGAGATGTGTCATCCTGGTGAATGTCCCTTTAACTGCAACCAG

Aligned partial sequences of fragment 6 as obtained from the sequencing output files of the long and short fragments (long fragment in green; short fragment in red)

TGCCTTCATTGTGAGGAAGGGTGCTCCAAGTCACGGCCACTAGGTTGTCTTCACCCATGTATTTTGCGATGTCAC
TGCCTTCATTGTGAGGAAGGGTGCTCCAAGTCACGGCCACTAGGTTGTCTTCACCCATGTATTTTGCGATGTCAC
CCTGGAGAATGTCCACCTTGTGTTCAGATGCTTAGAATAAAATGTCACTGTAAGATCACAAGCCTGTATGTGGAA
CCTGGAGAATGTCCACCTTGTGTTCAGATGCTTAGAATAAAATGTCACTGTAAGATCACAAGCCTGTATGTGGAA
TGTAGAAAAATAACCACAGCTGATGTAAATGAAAAGAACCTCCTCAGTTGTTGCAAAAATCAGTGCCCTAAAGAG
TGTAG**
CTTCCTTGTGGTCATAGATGCAAAGAGATGTGTCATCCTGGTGAATGTCCCTTTA
CTTCCTTGTGGTCATAGATGCAAAGAGATGTGTCATCCTGGTGAATGTCCCTTTA

Protein encoded by the long transcript

MEASWRQVAGGRGRSRGRATAAPSGNGVHLRGAGGGREKGSVGAVPSGTS 50
PGGVATTAAAGSRHSPAGSQALQTTAASELMSQKKFEEIKKANQAAARKL 100
VEEQFSSSSEEGDEDFEGKQGKILANTFITYTTQTDGDTRELERTKQYVN 150
EAFQAGAMTCLICIASVKRNQAVWSCSGCFCIFHMPCIQKWAKDSQFLVS 200
SVTDDDFGKKDCPWPCPKCRFEYKRSETPSRYYCYCGKVEDPPLDPWLVP 250
HSCGQVCEREFKPPCGHKCLLLCHPGPCPPCPKMVTTTCYCKKAKPIPRR 300
CSAKEWSCQLPCGQKLLCGQHKCENPCHAGSCQPCPRVSRQKCVCGKKVA 350
ERSCASPLWHCDQVCGKTLPCGNHTCEQVCHVGACGECPRSGKRFCPCQK 400
SKFSLPCTEDVPTCGDSCDKVLECGIHRCSQRCHRGPCETCRQEVEKHCR 450
CGKHTKRMPCHKPYLCETKCVKMRDCQKHQCRRKCCPGNCPPCDQNCGRT 500
LGCRNHKCPSVCHRGSCYPCPETVDVKCNCGNTKVTVPCGRERTTRPPKC 550
KEQCSRPPTCHHTSQEKHRCHFGSCPPCHQPCQKVLEKCGHLCPAPCHDQ 600
ALIKQTGRHQPTGPWEQPSEPAFIQTALPCPPCQVPIPMECLGKHEVSPL 650
PCHAVGPYSCKRVCGRILDCQNHTCMKECHKVTKTDGCTGKNKAGPECLH 700
CEEGCSKSRPLGCLHPCILRCHPGECPPCVQMLRIKCHCKITSLYVECRK 750
ITTADVNEKNLLSCCKNQCPKELPCGHRCKEMCHPGECPFNCNQKVKLRC 800
PCKRIKKELQCNKVRENQVSIECDTTCKEMKRKASEIKEAEAKAALEEEK 850
RRQQAELEAFENRLKGRRKKNRKRDEVAVELSLWQKHKYYLISVCGVVVV 900
VFAWYITHDVN

Protein encoded by the short transcript

MEASWRQVAGGRGRSRGRATAAPSGNGVHLRGAGGGREKGSVGAVPSGTS 50
PGGVATTAAAGSRHSPAGSQALQTTAASELMSQKKFEEIKKANQAAARKL 100
VEEQFSSSSEEGDEDFEGKQGKILANTFITYTTQTDGDTRELERTKQYVN 150
EAFQAGAMTCLICIASVKRNQAVWSCSGCFCIFHMPCIQKWAKDSQFLVS 200
SVTDDDFGKKDCPWPCPKCRFEYKRSETPSRYYCYCGKVEDPPLDPWLVP 250
HSCGQVCEREFKPPCGHKCLLLCHPGPCPPCPKMVTTTCYCKKAKPIPRR 300
CSAKEWSCQLPCGQKLLCGQHKCENPCHAGSCQPCPRVSRQKCVCGKKVA 350
ERSCASPLWHCDQVCGKTLPCGNHTCEQVCHVGACGECPRSGKRFCPCQK 400
SKFSLPCTEDVPTCGDSCDKVLECGIHRCSQRCHRGPCETCRQEVEKHCR 450
CGKHTKRMPCHKPYLCETKCVKMRDCQKHQCRRKCCPGNCPPCDQNCGRT 500
LGCRNHKCPSVCHRGSCYPCPETVDVKCNCGNTKVTVPCGRERTTRPPKC 550
KEQCSRPPTCHHTSQEKHRCHFGSCPPCHQPCQKVLEKCGHLCPAPCHDQ 600
ALIKQTGRHQPTGPWEQPSEPAFIQTALPCPPCQVPIPMECLGKHEVSPL 650
PCHAVGPYSCKRVCGRILDCQNHTCMKECHKVTKTDGCTGKNKAGPECLH 700
CEEGCSKSRPLGCLHPCILRCHPGECPPCVQMLRIKCHCKITSLYVECSF 750
LVVIDAKRCVILVNVPLTATRR

Figure 7.6: The difference in the sequences of the short and long versions of the sixth fragment of the *NFXL1* cDNA and the resulting differences in the amino acid sequences of the proteins.

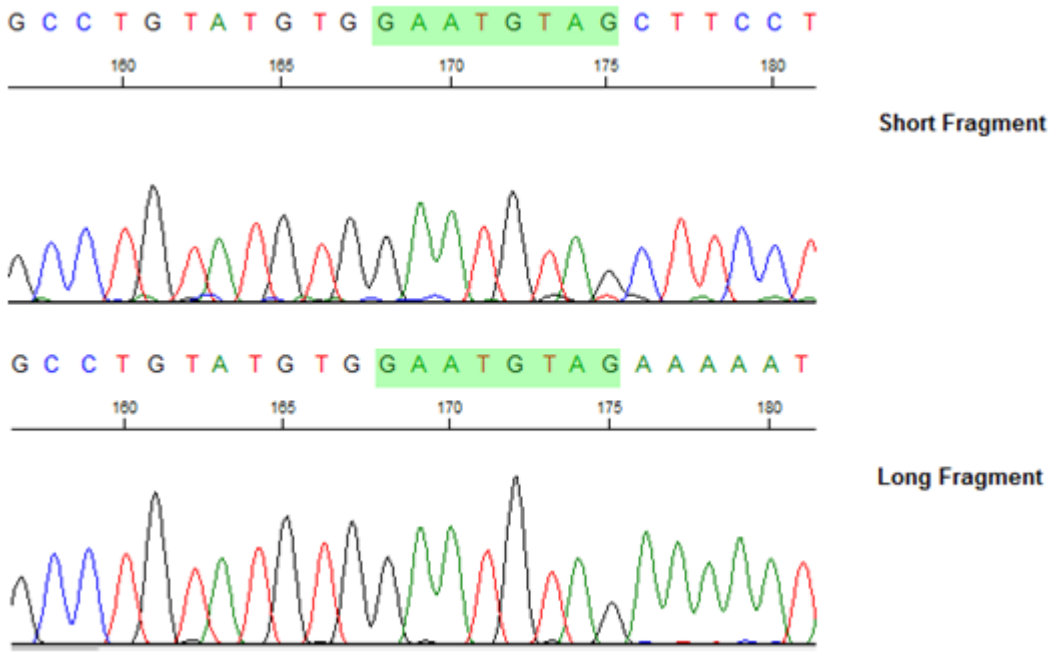


Figure 7.7: Partial sequences of the short and long *NFXL1* transcripts (the sequences do not align after the highlighted sequences).

The presence of the alternative transcript was validated in two human cDNA panels which included cDNA from the following tissues: heart, brain, kidney, liver, lung, pancreas, spleen, skeletal muscle, and placenta. The alternative transcript was observed in all of the above tissues (Figure 7.8). It should also be noted that across all samples tested the long transcript was always more abundant than the short transcript. As will be discussed in the next section, the short transcript was detected in HEK cells, but the presence of a shorter isoform of the protein could not be established in a WB analysis. This may suggest that the short transcript is not translated into protein. Moreover, as will be discussed later in this chapter, the predicted amino acid sequence crucial for the nuclear localisation of NFXL1 is found between positions 866 and 874 (inclusive), suggesting that the short isoform would not be able to enter the nucleus.

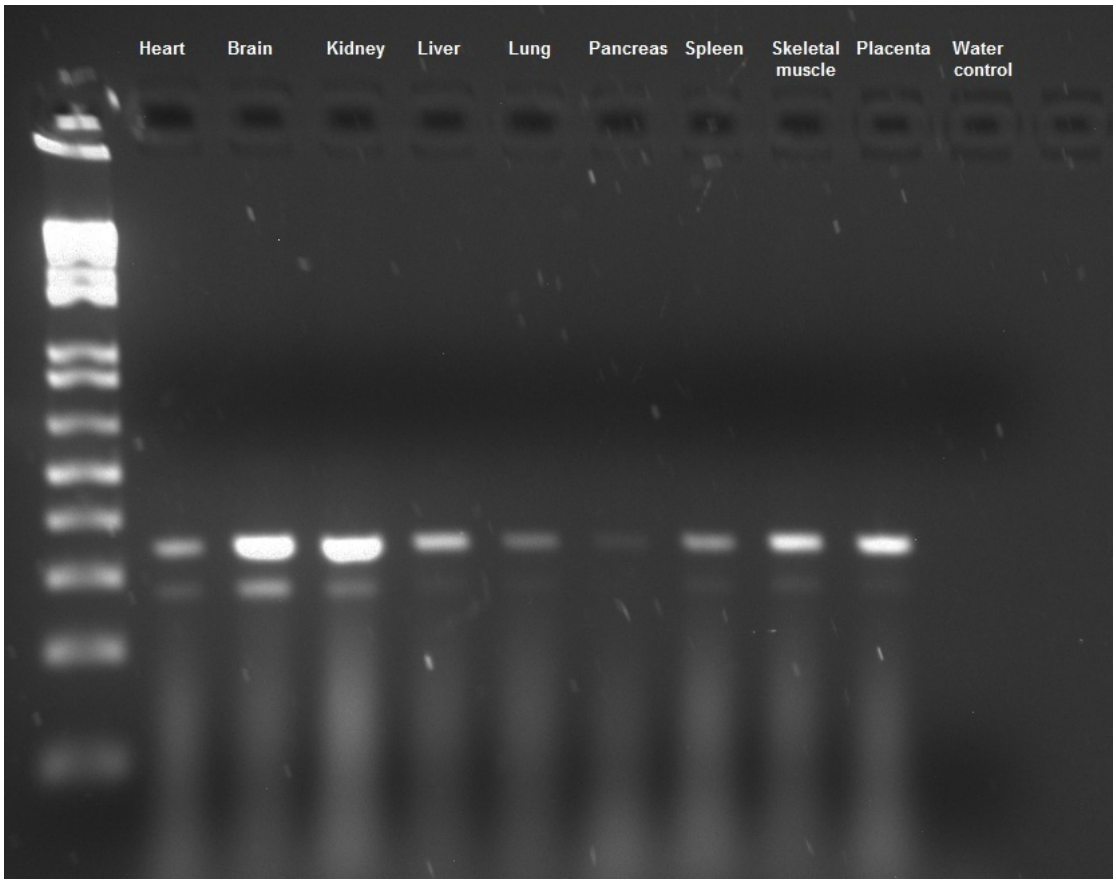


Figure 7.8: Amplification of the sixth fragment of the *NFXL1* cDNA across various tissues.

7.3: The Effects of Over-expressing *NFXL1* on Candidate Genes for SLI

A non-synonymous coding variant in *NFXL1* has been implicated in SLI through exome sequencing and an association analysis, as discussed in Chapter 5. As *NFX1*, a paralogue of *NFXL1*, regulates HLA class II genes (which include HLA-DRB1, HLA-DQA1 and HLA-DQB1), and associations between SLI and genes belonging to that group were found (as described in Chapter 4), I examined whether over-expressing *NFXL1* had an effect on the expression levels of the above genes, as well as major histocompatibility complex, class II, DR alpha (*HLA-DRA*) (OMIM#142860), which had previously been shown to be down-regulated by *NFXL1* (Song *et al.*, 1994). Additionally, two genes showed suggestive association with the N150K variant in *NFXL1* in a gene expression association analysis of the MEX 1000 Genomes population (*PDE1C* and *TRAF7*, as described in Chapter 5), and so expression levels for those genes were also examined. The expression levels were determined through qPCR with TaqMan assays. Three housekeeping genes were chosen for this analysis, based on studies of HEK cells or renal tissue: *GAPDH*, *GUSB* (glucuronidase, beta) and *PPIA* (peptidylprolyl isomerase A (cyclophilin A)) (Lisowski *et al.*, 2008, Caradec *et al.*, 2010). The TaqMan assays used in this analysis were: *NFXL1* (Hs00541755_m1), *GAPDH* (Hs99999905_m1), *GUSB* (Hs99999908_m1), *PPIA* (Hs99999904_m1), *HLA-DRA* (Hs00219575_m1), *HLA-DRB1* (Hs04192464_mH), *HLA-DQA1* (Hs03007426_mH), *HLA-DQB1* (Hs03054971_m1), *PDE1C* (Hs01095682_m1), and *TRAF7* (Hs00260228_m1). Expression levels were quantified using the $\Delta\Delta C_T$ method, with three housekeeping genes, as described above, and using the sample from the cells transfected with the empty vector as the reference.

HEK cells were grown (as described in Chapter 2) for the purpose of this experiment. These cells are easy to transfect and naturally express *NFXL1* and all the other genes examined. The over-expression of *NFXL1* was determined at the transcript level and at the protein level. Figure 7.9 shows a WB for lysates from HEK cells transfected with the *NFXL1* plasmid (left-most lane) and cells transfected with the same construct but without the *NFXL1* cDNA (the empty vector) (right-most lane). The NFXL1 protein is predicted to have a mass of 101 kDa. Since the *NFXL1* plasmid contained the cDNA of the gene and a double tag, it is

predicted to be slightly larger. As can be seen in the figure, the left lane contains two bands that are very close to each other: one for endogenous protein and one for the tagged protein, and the right lane has only one band, for the endogenous protein. The other bands in the left lane may be the result of processed protein which the antibody recognised. Further details on the procedure used to generate this image can be found in Chapter 2. Also shown in Figure 7.9 is the difference in the amounts of PCR product obtained from the cDNA of cells transfected with the NFXL1 plasmid and cells transfected with the empty vector. Primers for the amplification of fragment 6 of the NFXL1 cDNA, as described in the previous section, were used to check whether HEK cells expressed the putative short splice variant. There was a large amount of DNA template and therefore too much PCR product in HEK cells treated with the NFXL1 plasmid, but it can be seen from the right lane on the gel that HEK cells have the two transcripts at the cDNA level. However, as can be seen from the WB, the short protein is not detected in HEK cells. It is possible that some regulatory process such as nonsense-mediated decay prevents the transcript from being translated. It is also possible that the amount of the alternative isoform was not large enough to be detected through WB with those conditions.

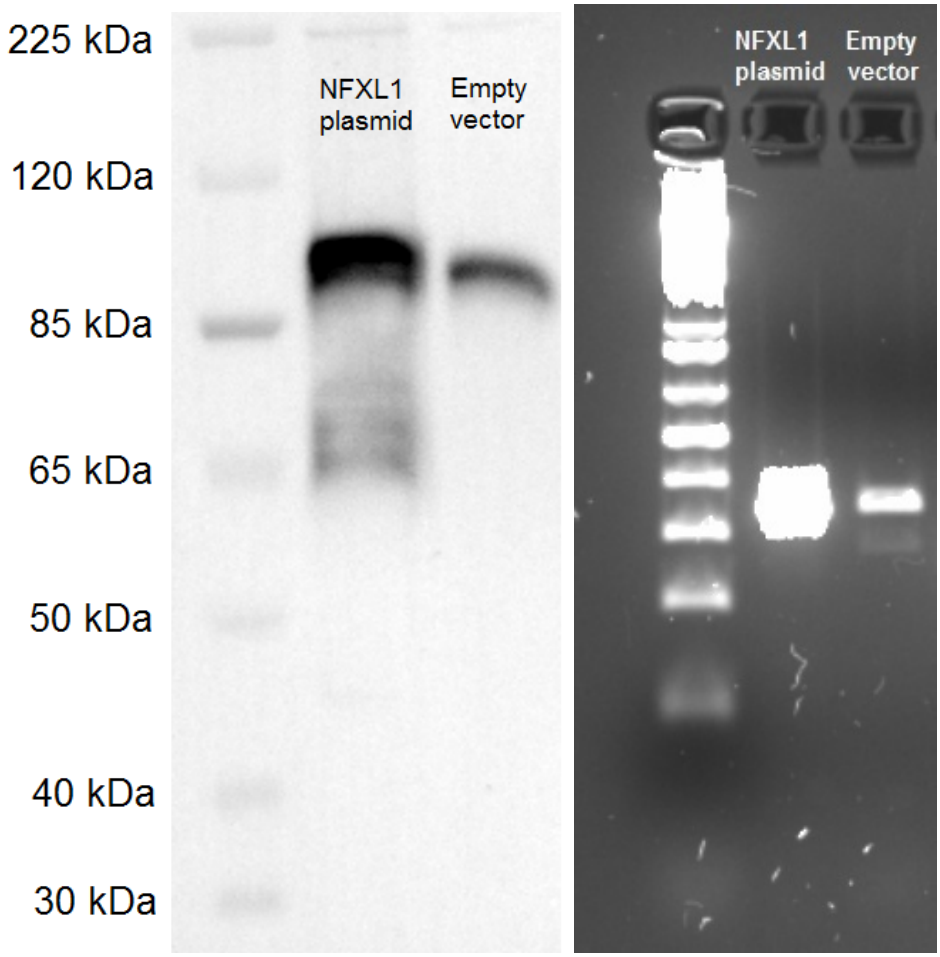


Figure 7.9: Western blot and DNA gel electrophoresis analyses for HEK cells over-expressing *NFXL1* and control HEK cells (the left lane on the gel contained lysate from HEK cells transfected with the NFXL1 plasmid, and the right lane contained lysate from HEK cells transfected with the empty vector). Protein or DNA markers were included in the left-most lane of the gels. For the western blot, samples contained 2 µg protein lysate, and the exposure time for the image was 20 seconds (the picture of the markers was superimposed on the picture of the bands). The PCR used primers for the sixth *NFXL1* cDNA fragment, which show the qualitative difference in expression levels between the samples as well as the presence of the short transcript in (control) HEK cells. The NFLX1 protein is predicted to have a mass of 101 kDa.

In order to study the effects of over-expressing *NFXL1* in HEK cells on potential candidate genes, RNA was extracted from the HEK cells treated with either the NFXL1 plasmid or the empty vector. RT-PCR was performed using the RNA (further details can be found in Chapter 2), and a qPCR was performed with the TaqMan assays to measure gene expression levels for the genes mentioned earlier. The C_T value for *NFXL1* for the cells transfected with the NFXL1 plasmid was 9.21 cycles lower than that obtained for the cells transfected with the empty vector. That indicated that the former had more than 500 times as many *NFLX1* transcripts, compared to the latter. It should be noted that, only in the case of the TaqMan assay for *NFXL1* with the cells transfected with the NFXL1 plasmid, the RT-negative controls were not clean. That is probably due to the presence of the plasmid itself in the extracted RNA, which could serve as a template in the qPCR, as it contained the cDNA of the gene. The plasmid was not eliminated despite DNase treatment. The presence of the plasmid in the RT-negative controls was confirmed through a PCR in which one primer bound to a sequence on the plasmid but outside the *NFXL1* cDNA (supplied by Origene), and the other bound to a sequence inside the cDNA. The PCR amplified the fragment contained by those two primers, indicating that the plasmid was indeed present in the sample. The C_T values of the negative RT controls were slightly higher than those of the positive RT reactions with the cells transfected with the empty vector, suggesting that the low C_T value for the positive RT reaction with cells transfected with the NFXL1 plasmid is the result of the abundance of the transcript and not the contribution of the plasmid. In any case, the aim of measuring *NFXL1* levels was not to quantify the change in expression levels following the transfection, but, rather, only to show that the gene is over-expressed. The results of the expression analysis are presented in Table 7.2.

Table 7.2: Effects of *NFXL1* over-expression on several genes in HEK cells.

Gene	C_T (HEK cells transfected with the NFXL1 plasmid)	C_T (HEK cells transfected with the empty vector)	Expression level compared to the control (*GADPH* as the HKG)	Expression level compared to the control (*GUSB* as the HKG)	Expression level compared to the control (*PPIA* as the HKG)
GAPDH (HKG)	22.48	22.55	1	1.03	1.21
GUSB (HKG)	25.93	25.96	0.97	1	1.17
PPIA (HKG)	22.57	22.37	0.83	0.85	1
*NFXL1**	17.95	27.16	564.18	580.04	680.29
HLA-DRA	37.98	38.08	1.02	1.05	1.23
HLA-DRB1	31.15	31.1	0.92	0.95	1.11
HLA-DQA1	38.22	38.19	0.93	0.96	1.13
HLA-DQB1	30.31	30.38	1	1.03	1.21
PDE1C	35.9	35.94	0.98	1.01	1.18
TRAF7	28.12	28.15	0.97	1	1.17

* *NFXL1* expression levels are not accurate due to the contribution of the plasmid cDNA (see discussion in the text).

As can be seen from Table 7.2, there are no significant differences between the HEK cells transfected with the NFXL1 plasmid and those transfected with the empty vector.

There may be several possible explanations for these results. If the protein does regulate these genes, but this was not captured by the qPCR, it might be due to the fact that while the levels of the *NFXL1* transcript in the cells transfected with the NFXL1 plasmid were very high, the protein levels were not as high compared to the endogenous protein (Figure 7.9). It could be that some of the protein was degraded (which could explain the other bands on the left-most lane in Figure 7.9), or that the tag affects the functionality of the protein. It is also possible that the NFXL1 protein does not affect the expression levels of the candidate genes tested, at least in the conditions used here. This could be restricted to the type of cell used (HEK). All of the genes tested including *NFXL1* itself are naturally expressed in HEK cells. However, transcription factors may operate by interacting with other proteins through protein-protein interaction domains (Wray *et al.*, 2003), and it is possible that another protein, which is necessary for the regulation of these genes, was not present in this cell type. Despite the fact that *NFX1*, a paralogue of *NFXL1*, regulates some of the genes tested, *NFXL1* might have evolved to regulate other genes, or compete with its paralogue, through changes in its DNA-binding domains or changes in its potential protein interaction domains (Singh and Hannenhalli, 2008). Finally, as will be discussed in the next section, the over-expressed

NFXL1 might have not been present in the cell nucleus at all. If none of the transfected cells had nuclear NFXL1, then over-expressing it would not have made any difference with regards to the regulation of target genes, in the context of direct regulation of transcription (but not in the context of interaction with other proteins which regulate gene expression).

7.4: An Investigation into the Localisation of *NFXL1* inside the Cell through Immunofluorescence

An immunofluorescence experiment was performed using control HEK cells and HEK cells transfected with the NFXL1 plasmid in order to examine the subcellular localisation of NFXL1. Table 7.3 shows the antibody combinations I used. Combinations 1 and 2 were applied to control and transfected cells fixed with both paraformaldehyde and methanol (not used together). Combination 3 was applied to control and transfected cells fixed with paraformaldehyde. The same secondary antibody combination was used with all samples. The cells were also treated with a DAPI solution, which stains cell nuclei in blue. Fixation with methanol worked for all antibodies. Fixation with paraformaldehyde did not work well for anti-NFXL1 (A303-196A), but that could be the result of a problem with the transfection (no cells over-expressing *NFXL1* were detected). The over-expressed protein should have a Myc-DDK tag, which should be recognised by anti-Myc and anti-DDK antibodies. The antibodies for the endogenous protein should recognise both the endogenous and the over-expressed protein.

Table 7.3: Antibody combinations for the Immunofluorescence experiment.

Combination	Antibody	Species	Dilution	Colour
Primary 1	Anti-NFXL1 (A303-196A)	Rabbit	1:500	-
	Anti-Myc	Mouse	1:300	-
Primary 2	Anti-NFXL1 (A303-197A)	Rabbit	1:500	-
	Anti-DDK (TA50011-100)	Mouse	1:500	-
Primary 3	Anti-NFXL1 (A303-197A)	Rabbit	1:200	-
	Anti-Myc	Mouse	1:300	-
Secondary	Anti-rabbit (Alexa Fluor 594)	Goat	1:500	Red
	Anti-mouse (Alexa Fluor 488)	Goat	1:500	Green

Figures 7.10-7.13 show the results of the immunofluorescence experiment for cells fixed with methanol. The first observation is that the signal from the endogenous protein is

indiscernible from the background noise (shown in red in Figures 7.10 and 7.12). In the transfected cells, the over-expressed protein is detected by all antibodies (shown in red and green in Figures 7.11 and 7.13). Surprisingly, the protein was not found in cell nuclei in any of the samples. The over-expressed protein appeared to be localised around the nucleus and not evenly spread in the cytoplasm. It is not possible to rule out that the tag or the over-expression itself might have interfered with the localisation of the protein and prevented it from entering the nucleus, based solely on the results of this experiment. Another explanation would be that it is not naturally found in the nucleus (at least, in HEK cells). Whichever scenario is true, it could potentially explain why over-expressing *NFXL1* did not have any effect on the potential candidate genes, as described in the previous section. If the endogenous protein does in fact localise in the cytoplasm, then it means that the NFXL1 protein could have acquired a different function from that of NFX1. The immunofluorescence experiment was repeated, this time with a neuronal cell line (SH-SY5Y, a neuroblastoma cell line), to check whether the localisation of the protein differed based on the type of cell used. Figures 7.14-7.17 show the results of the immunofluorescence experiment, with antibody combinations 1 and 2, as used with the HEK cells. The SH-SY5Y cells were difficult to transfect with the NFXL1 plasmid (and they are generally harder to transfect compared to HEK cells). However, since transfected cells may potentially be easier to detect through immunofluorescence than WB, I decided to use them nonetheless. As can be seen from Figures 7.14-7.17, the transfected cells and control cells are indistinguishable. No signal from the antibodies for the Myc or DDK tags was detected. However, in contrast to the case of the HEK cells, some signal from the antibodies for the endogenous protein was detected, as shown in red in those figures. The signal was weak, and it was only detected in some cells. If it does in fact come from the endogenous NFXL1 protein in those cells, it could mean that some cells had more of the protein than others. It is interesting to note that the signals from both antibodies for the endogenous protein indicate that the protein they recognise is not found in the nucleus. That is in line with the results of the immunofluorescence experiment done with the HEK cells. Furthermore, if it is indeed NFXL1 that the signal comes from, then this shows that even the non-tagged protein is not found in the nucleus. It is not possible to know whether the transfections did not work, or whether the tags were removed through

some cellular processing, and the signal came from the over-expressed protein in some cases, even though a similar signal was obtained when using control cells. However, assuming that the antibodies for the endogenous protein worked properly, these results suggest that the NFXL1 protein is primarily localised in the cytoplasm. In a previous study, Tang *et al.* (2005) cloned part of the *NFXL1* gene (they obtained the cDNA from a human ovary cDNA library). The amino acid sequence of the protein they examined started with the amino acid at position 676 of the NFXL1 protein. Their cDNA mapped to the *NFXL1* transcript on the UCSC Genome Browser, but it was shorter. Tang *et al.* examined the localisation of the protein by inserting the cDNA into a vector containing an enhanced green fluorescent protein tag (which would produce a fusion protein) and transfecting HeLa cells with it. Their results indicate that the fusion protein was found in the cytoplasm of the HeLa cells and not in the nucleus. While the protein they studied did not have the full NFXL1 sequence, their results are in line with mine. One way of confirming whether the tag prevents the protein from entering the nucleus would be to transfect cells with a plasmid containing only the cDNA of the protein without any tags, but, in that case, only antibodies for the endogenous protein would be able to detect it. For that reason, using such antibodies which were produced in different species and shown to work properly – with the same sample (as to allow the capturing of signals of different colours) – would be advantageous. Finally, it is possible that the NFXL1 protein translocates to the nucleus only following interaction with other molecules in the cell, which were not present in the HEK or SH-SY5Y cells in which the protein was detected. I checked the amino acid sequence of NFXL1 with NucPred, a tool for predicting nuclear localisation of proteins. The NucPred score for NFXL1 was 0.87. A threshold of 0.8 means a specificity of 84% and a sensitivity of 32%. In that sense, a score of 0.87 is rather high, indicating that the protein is likely to localise in the nucleus. The sequence "GRRKKNRKR" (amino acid positions 866-874) was predicted to be important for nuclear localisation. If that prediction is true, then the absence of NFXL1 from the nucleus in the examined cell lines could mean that it was being kept out of the nucleus through some cellular process. Another possibility is that NFXL1 must interact with other molecules before it can translocate to the nucleus. There are many examples in the literature of transcription factors that are cytoplasmic and translocate to the nucleus only after interacting with other

molecules in the cell, or transcription factors that move in or out of the nucleus based on external stimuli. The previously mentioned NF-κB transcription factor has two subunits, both of which interact with a cytoplasmic inhibitory protein, which keeps NF-κB in the cytoplasm; when the NF-κB is released from that protein, the former translocates to the nucleus (Beg *et al.*, 1992). A famous example is that of signal transducer and activator of transcription (STAT) proteins. STAT proteins are retained in the cytoplasm and translocate to the nucleus only after they have undergone activation by phosphorylation (Darnell, 1997). Another experiment examined, among other things, the localisation of the forkhead Box O3 (FOXO3) transcription factor in fibroblasts. FOXO3 was naturally localised in the nucleus, but, when the cells were treated with growth factors, FOXO3 was found mainly in the cytoplasm. When the treated cells were exposed to oxidative stress, FOXO3 translocated to the nucleus (Brunet *et al.*, 2004). In the same line of thought, it is possible that the reason NFXL1 was not detected in the nucleus is not that it is not a transcription factor, but, rather, that the necessary conditions for it to be able to translocate to the nucleus were not present. Interestingly, it was shown that NFX1 repressed *HLA-DRA* transcription in HeLa cells following treatment with interferon gamma (IFN-γ) but had no effect on *HLA-DRA* in that cell type without the IFN-γ treatment (Song *et al.*, 1994). If the same applied to the case of NFXL1, and IFN-γ were involved in the activation of NFXL1, then it could explain the cytoplasmic localisation of NFXL1 in HeLa cells, as reported by Tang *et al.* (2005), and the results described here. IFN-γ has indeed been shown to be involved in the nuclear translocation of the transcription factor STAT1-alpha (Subramaniam *et al.*, 2000). To test this hypothesis, I treated HEK cells with human INF-γ recombinant protein (Affymetrix eBioscience). INF-γ was added to the medium following transfections (after the transfection medium had been changed) at the following concentrations: 250 U/ml, 500 U/ml, 1000 U/ml, and 2000 U/ml. The cells were incubated for 48 hours. Song *et al.* (1994) observed an effect 48 hours after treating HeLa cells with 250 U/ml INF-γ. Cells were also treated with 1000 U/ml INF-γ and processed after 24 hours. In all of the above cases, immunofluorescence experiments revealed that the over-expressed NFXL1 protein was still retained in the cytoplasm.

In summary, these results suggest that, at least in HEK and SH-SY5Y cells and with the experimental conditions used here, NFXL1 is retained in the cytoplasm.

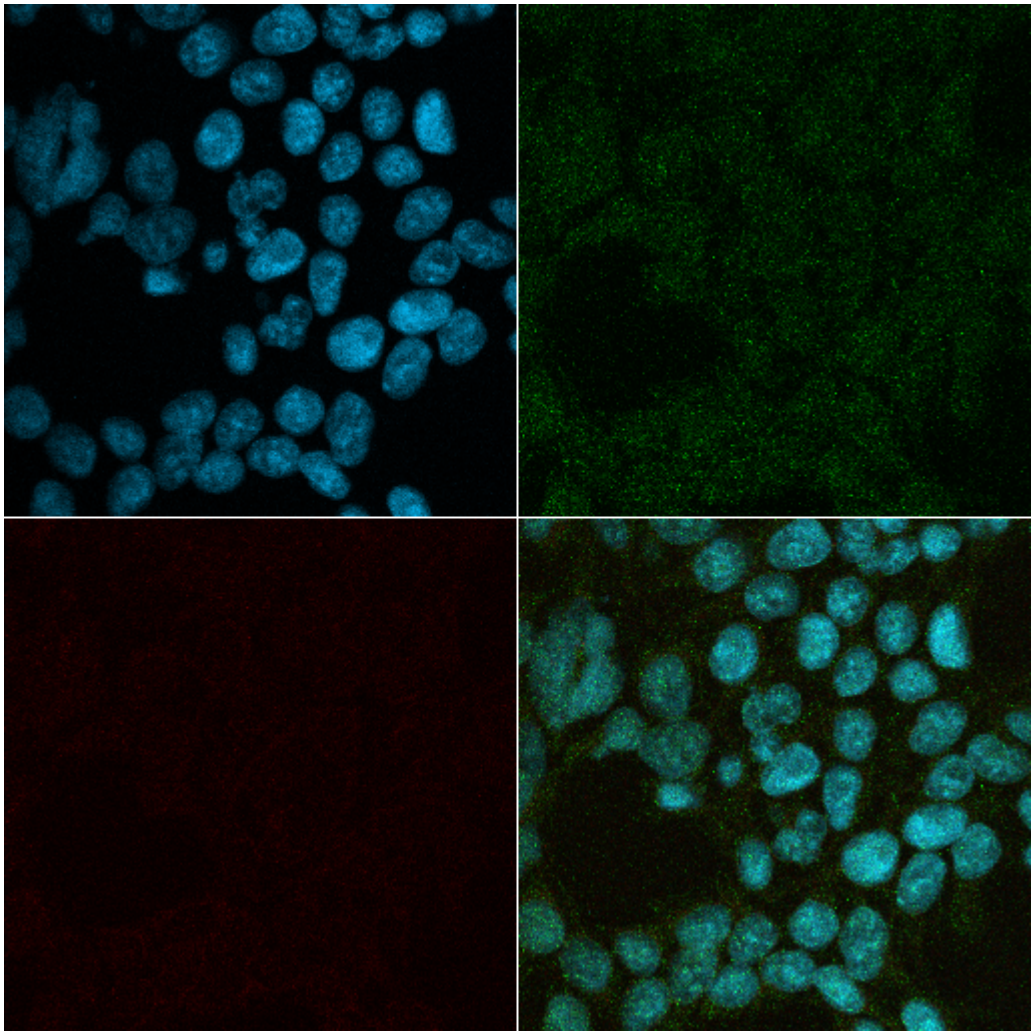


Figure 7.10: Immunofluorescence of methanol-fixed control HEK cells using anti-NFXL1 (A303-196A) and anti-Myc antibodies. From top-right counter-clockwise: only anti-Myc (green), DAPI staining (blue), only anti-NFXL1 (A303-196A) (red), all three images superimposed.

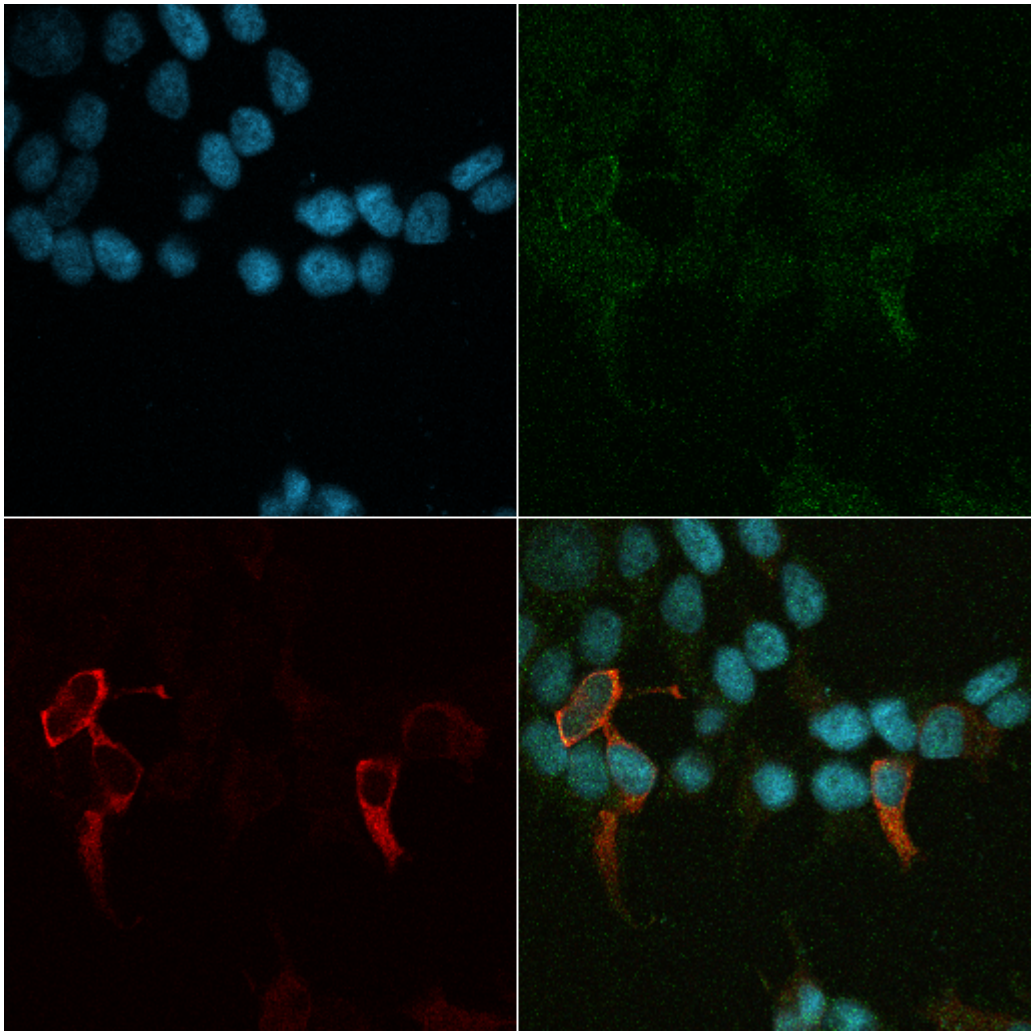


Figure 7.11: Immunofluorescence of methanol-fixed transfected HEK cells using anti-NFXL1 (A303-196A) and anti-Myc antibodies. From top-right counter-clockwise: only anti-Myc (green), DAPI staining (blue), only anti-NFXL1 (A303-196A) (red), all three images superimposed.

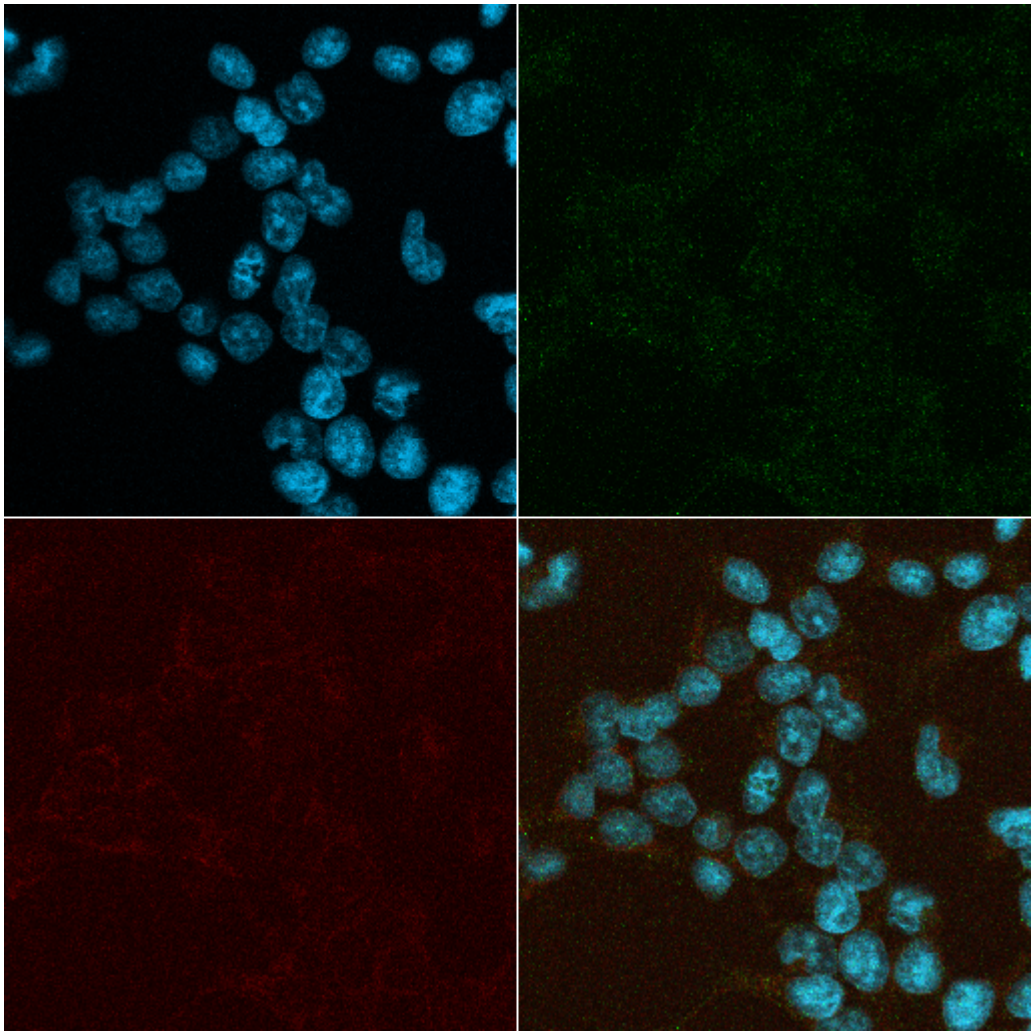


Figure 7.12: Immunofluorescence of methanol-fixed control HEK cells using anti-NFXL1 (A303-197A) and anti-DDK antibodies. From top-right counter-clockwise: only anti-DDK (green), DAPI staining (blue), only anti-NFXL1 (A303-197A) (red), all three images superimposed.

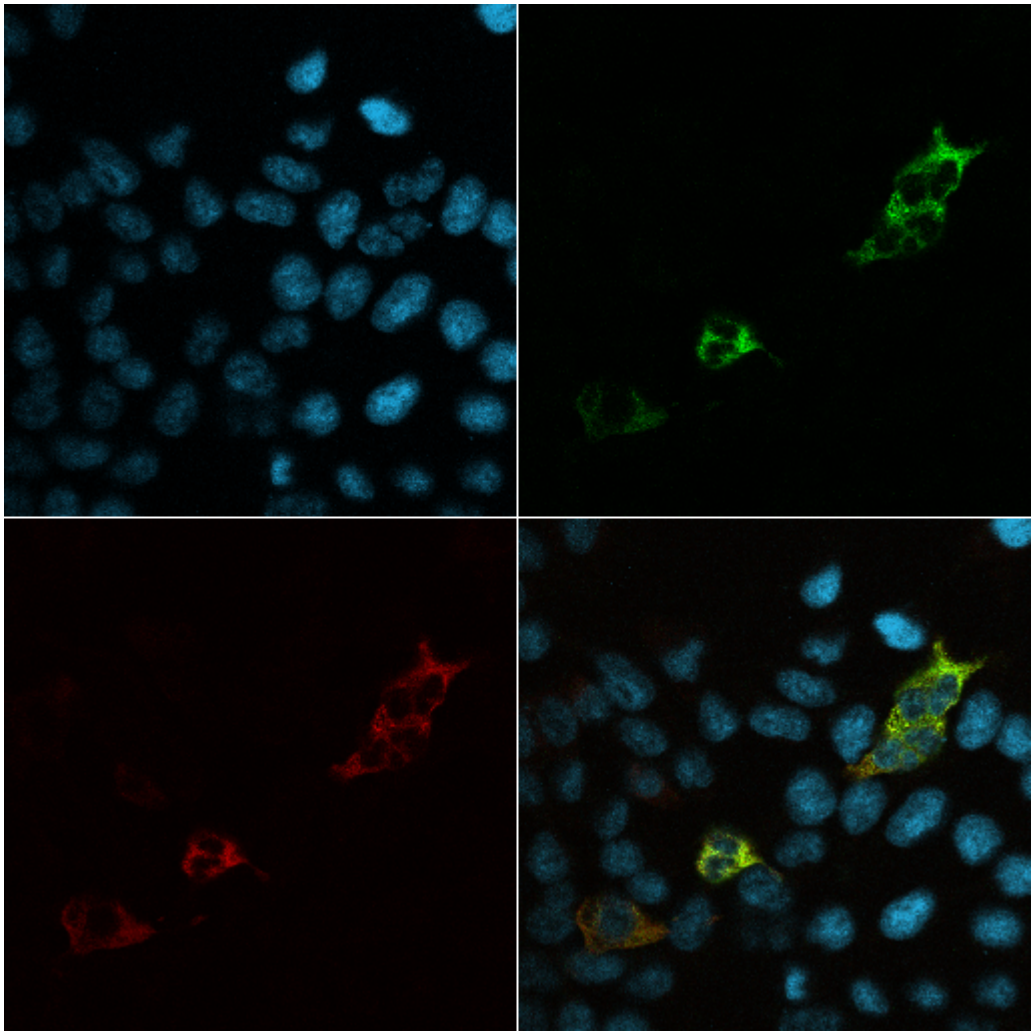


Figure 7.13: Immunofluorescence of methanol-fixed transfected HEK cells using anti-NFXL1 (A303-197A) and anti-DDK antibodies. From top-right counter-clockwise: only anti-DDK (green), DAPI staining (blue), only anti-NFXL1 (A303-197A) (red), all three images superimposed.

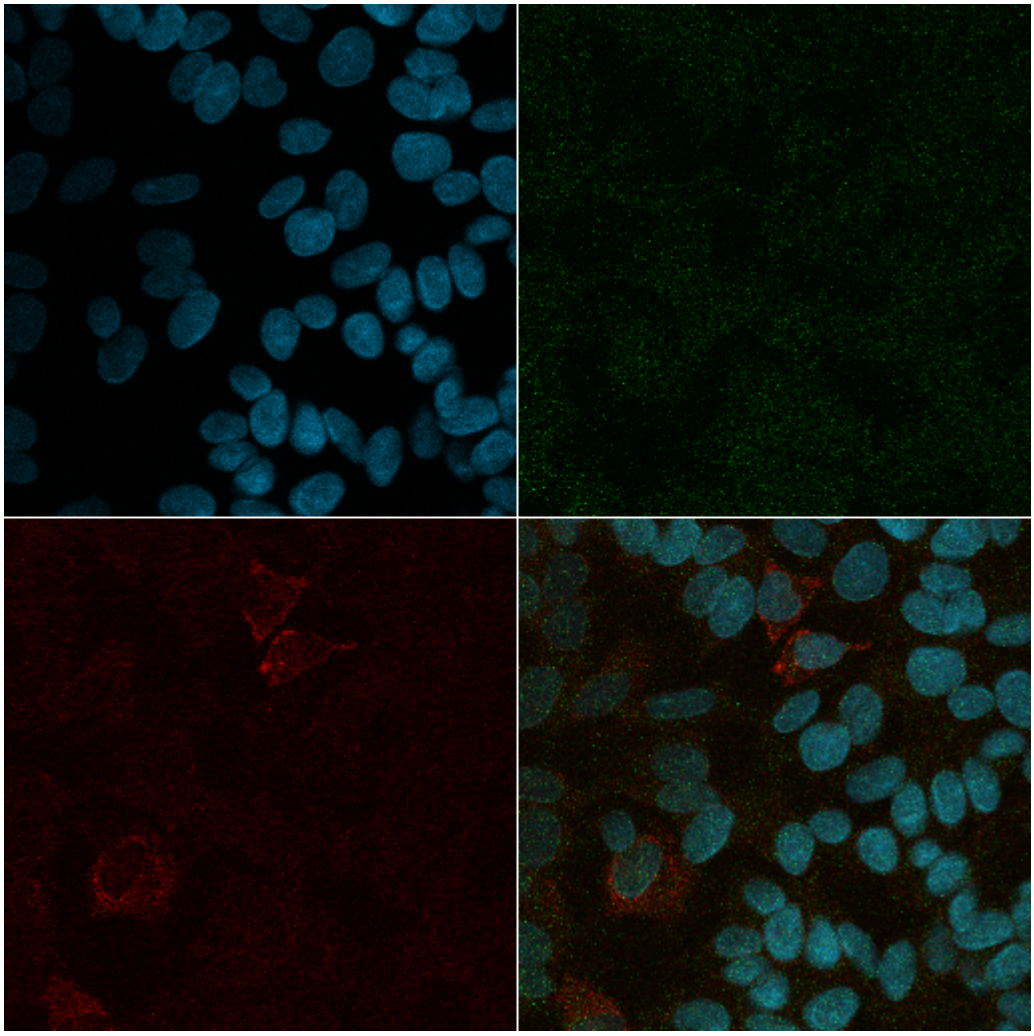


Figure 7.14: Immunofluorescence of methanol-fixed control SH-SY5Y cells using anti-NFXL1 (A303-196A) and anti-Myc antibodies. From top-right counter-clockwise: only anti-Myc (green), DAPI staining (blue), only anti-NFXL1 (A303-196A) (red), all three images superimposed.

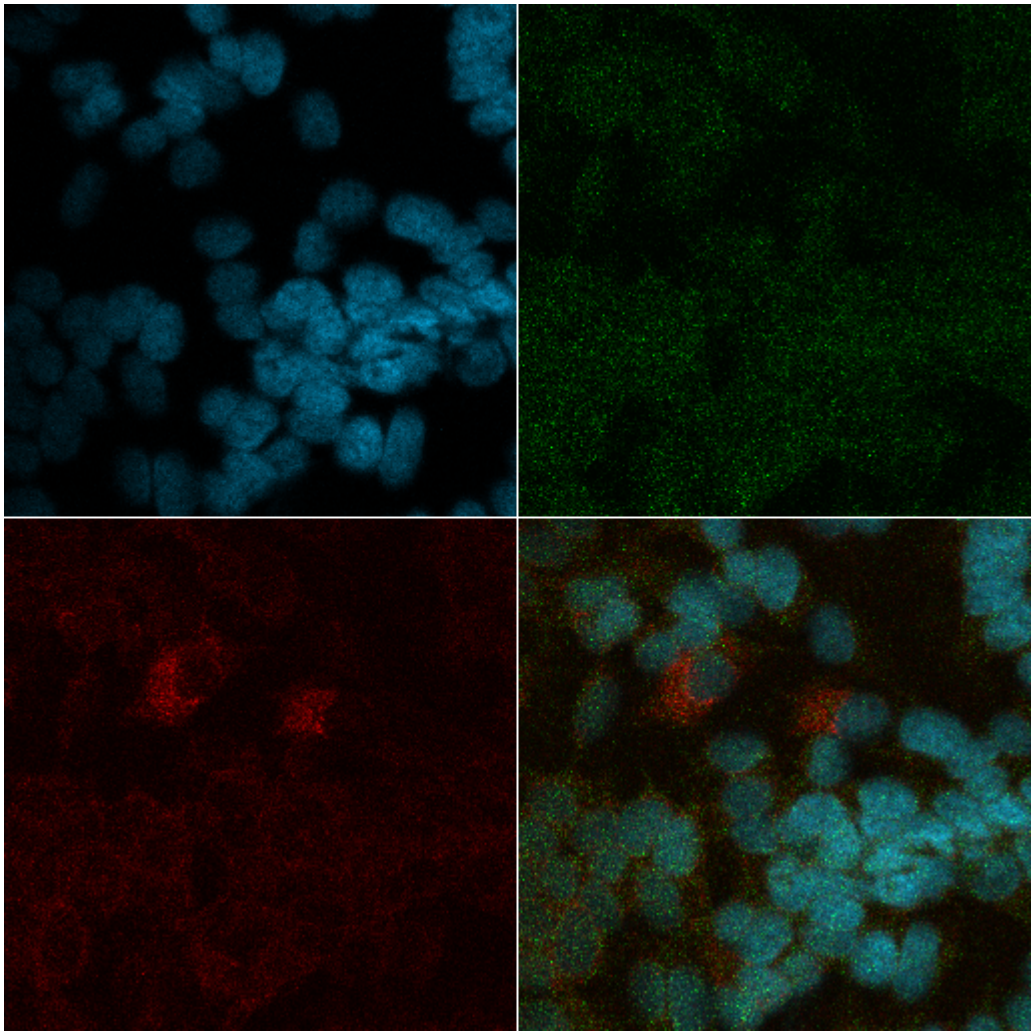


Figure 7.15: Immunofluorescence of methanol-fixed transfected SH-SY5Y cells using anti-NFXL1 (A303-196A) and anti-Myc antibodies. From top-right counter-clockwise: only anti-Myc (green), DAPI staining (blue), only anti-NFXL1 (A303-196A) (red), all three images superimposed.

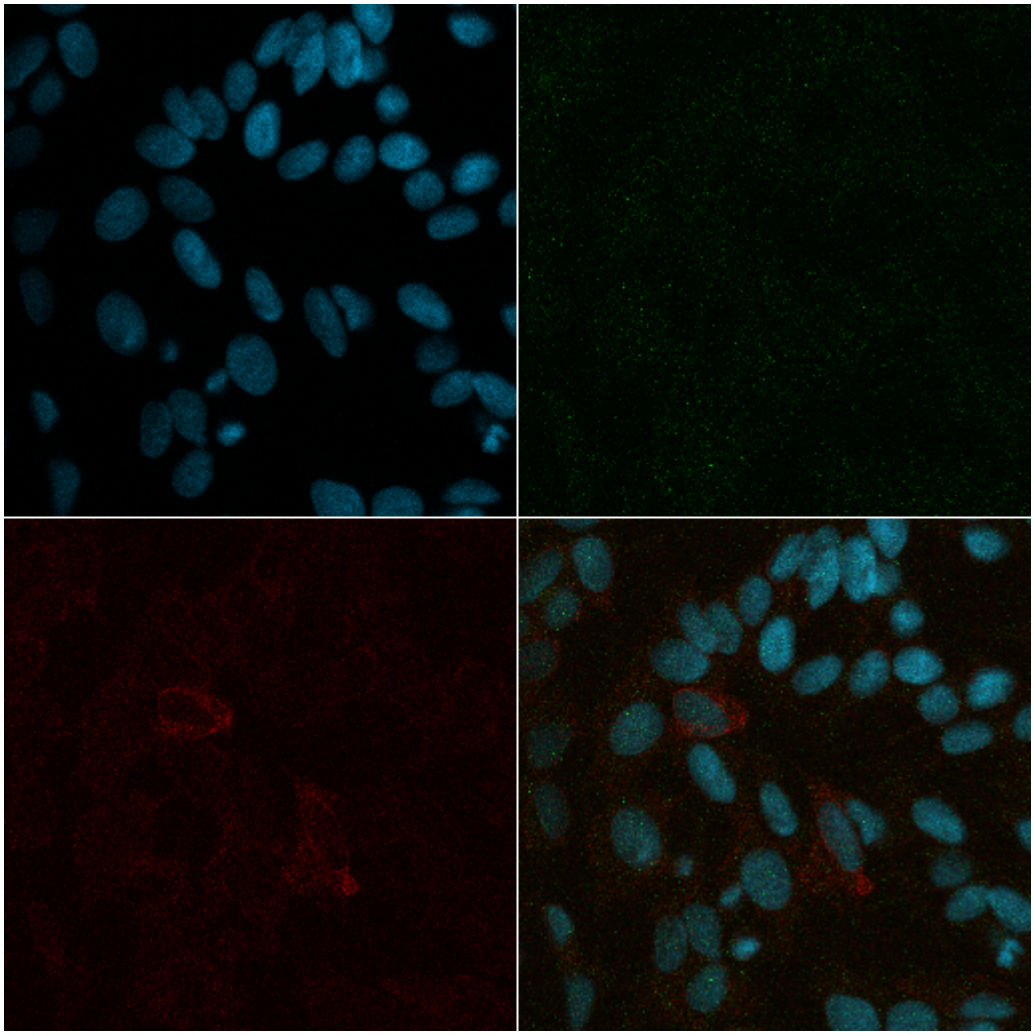


Figure 7.16: Immunofluorescence of methanol-fixed control SH-SY5Y cells using anti-NFXL1 (A303-197A) and anti-DDK antibodies. From top-right counter-clockwise: only anti-DDK (green), DAPI staining (blue), only anti-NFXL1 (A303-197A) (red), all three images superimposed.

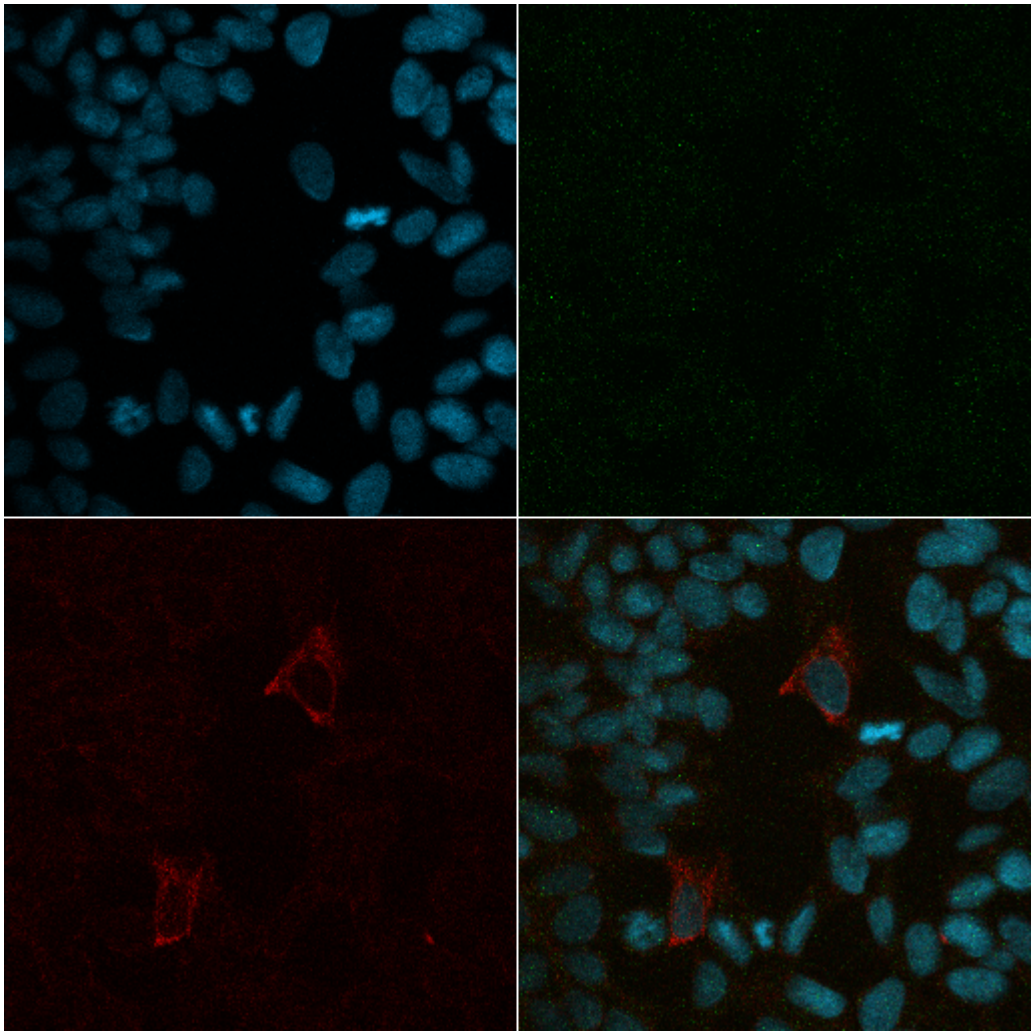


Figure 7.17: Immunofluorescence of methanol-fixed transfected SH-SY5Y cells using anti-NFXL1 (A303-197A) and anti-DDK antibodies. From top-right counter-clockwise: only anti-DDK (green), DAPI staining (blue), only anti-NFXL1 (A303-197A) (red), all three images superimposed.

7.5: Summary of Chapter 7 and Future Directions for the Functional Study of *NFXL1*

A non-synonymous coding variant in *NFXL1* (N150K) has been implicated in language impairment in the Robinson Crusoe Island cohort. Two other, rare (R432C) or novel (in individuals of European ancestry, T563A) coding variants have been found in SLIC families. Little is known about the function of this gene, but it is predicted to code for a transcription factor.

I performed several experiments to further our knowledge regarding the possible role of NFXL1. I investigated gene expression levels of *NFXL1* across various brain regions, to identify splice variants in various tissues, to assess the effect of over-expressing *NFXL1* on potential target genes (based on its predicted function and DNA-binding domains), and to examine the subcellular localisation of the protein.

The highest expression levels in the brain were found in the cerebellum. As outlined earlier in this chapter, the cerebellum is now thought to be important for language processing in the brain. I identified a splice variant lacking one exon which was found in all investigated tissues. It was less abundant than the longer transcript (as can be learned from the intensity of the band on the gel). This could suggest different functions for the different isoforms.

The over-expression and immunofluorescence experiments did not suggest a regulatory role for *NFXL1* in the cell lines used. In the former, no effect was observed when examining genes that were known to be regulated by *NFX1*, a close paralogue of *NFXL1*. In the latter, no nuclear localisation was observed in any sample. The signal from the endogenous protein was too weak to be distinguished from background noise in HEK cells, but the over-expressed protein was observed only in the cytoplasm. In SH-SY5Y cells, the over-expressed protein was not detected with anti-tag antibodies, but the antibodies for the endogenous protein detected a protein in the cytoplasm. However, it is possible that, given the right conditions, NFXL1 could have translocated to the nucleus, and the observable effects would have been very different. It should be noted, with regards to those two experiments, that cells used in the over-expression experiment were grown and transfected

separately from the cells used in the immunofluorescence experiment. For that reason, the lack of observable over-expressed nuclear NFXL1 in the latter sample does not entail that there was no over-expressed nuclear NFXL1 in the former, although it does suggest it. The immunofluorescence experiment was carried out after the over-expression experiment, but, in my opinion, it was worth it doing both, because NFXL1 might have a function which is different from what the predictions say, and that is particularly relevant to the case of studying the effects of over-expressing *NFXL1* on the expression levels of *PDE1C* and *TRAF7*¸ which were also included in the over-expression experiment.

If NFXL1 does act as a transcription factor, a full characterisation of the targets of the *NFXL1* gene would require a ChIP-Seq (chromatin immunoprecipitation sequencing) experiment to identify potential target genes for *NFXL1* within a neuronal cell line (SH-SY5Y). I attempted this approach. However, the most important factor in a ChIP experiment is the antibody, and the commercially-available antibodies for *NFXL1* were not good enough for such an experiment. I tried five different anti-NFXL1 antibodies. Three of them did not detect the protein through WB at all. The other two, which were used in the immunofluorescence experiment, detected the protein through WB, but blotting with them would at times result in more than one band on the WB. They also performed poorly when I tested them for use in immunoprecipitation. I then decided to use a plasmid containing the cDNA and a Myc-DDK tag, which would have enabled me to use antibodies for the tag rather than antibodies for the endogenous protein in the ChIP. However, neither the anti-Myc nor the anti-DDK antibodies detected the tagged protein through WB or immunoprecipitated it successfully. I used a positive control with a Myc tag in the same WB, and it was detected. Furthermore, blotting the same membrane with an anti-NFXL1 antibody produced a band that showed that the protein was indeed over-expressed. The anti-Myc and anti-DDK antibodies worked in the immunofluorescence experiment. Whatever the reason that they did not perform well on WB or immunoprecipitation may be, it meant that they could not be used in ChIP. Furthermore, the neuronal cell line that I was using proved difficult to transfect with the NFXL1 plasmid. I therefore used HEK cells for my experiments, as they were easier to transfect and expressed all the relevant genes. I also tried using the Raji cell line, which is a lymphoblastoid cell line. Raji cells are HLA class II-positive cells, and they also express *NFXL1*. However,

lymphoblastoid cells are hard to transfect, and I could not detect the over-expression of *NFXL1* following transfections with three different transfection reagents, or electroporation. Due to lack of time, I could not test more antibodies, cell lines or experimental conditions, and I therefore decided to not go ahead with the ChIP-seq experiment, as it is very costly and should not be attempted with antibodies that do not perform adequately.

There are two points for future work that require consideration. The first is that a development of a ChIP-grade antibody for NFXL1 would make it possible to perform a ChIP-seq experiment with cells that are hard to transfect, as it would eliminate the need to transfect cells in order for them to produce a tagged version of the protein and thus would provide a more accurate picture of the endogenous protein. The second point is that, given that there is a possibility that NFXL1 may not be present in the nucleus in some types of cells and/or under certain conditions, its presence there should be verified for the cell type and conditions used in the ChIP. Moreover, further support for a regulatory function for *NFXL1* should be obtained (perhaps from structural studies) prior to performing a ChIP-seq experiment. Given that NFXL1 was found in abundance in embryonic stem cells prior to their differentiation into oligodendrocytes, it is possible that it plays a role in differentiation. Therefore, an interesting line of research would be to study it in embryonic stem cells at various stages of differentiation, during which it may translocate to the cell nucleus.

Chapter 8: A Case-study of a Speech and Language Impairment in a Spanish-speaking Child

This chapter describes parts of ongoing research focused on elucidating the genetic mechanisms hypothesised to be involved in a language disorder affecting a Spanish-speaking girl. The following information including the clinical description of the girl is taken from (García-Bellido *et al.*, 2009). The girl, A, was born in 1999. When she was born she suffered from jaundice, which lasted for some months following her birth. She had problems with feeding, which decreased over time but lasted until she was 7. From early on she was unable to produce comprehensible speech, and she has been receiving close paediatric and educational attention, including speech therapy. For most of her schooling years she has been receiving two hours of language therapy per week and additional help at school. A has normal hearing and locomotion, but she suffers from astigmatism. At the age of 10 years and 3 months (Paloma Garcia-Bellido, personal communication, September 1, 2014), A was administered several linguistic tests which included components in morphology, syntax, reading comprehension, conversational skills and narrative task. These tests were not part of a test battery (Paloma Garcia-Bellido, personal communication, November 25, 2013). Her linguistic deficits showed some similarity to those observed in the KE family and in other cases of *FOXP2* disruptions; for example, she had deficits in the area of word morphology (inflectional morphology e.g. plural marking) and an oromotor deficit (mainly in the planning of articulation and in the correct integration of prosodic sequences) (García-Bellido *et al.*, 2009). Furthermore, A had a deficit in the timing of articulations (Paloma Garcia-Bellido, personal communication, December 10, 2014). Results of further tests administered to her indicate lower verbal competence compared to an age-matched control (Paloma Garcia-Bellido, personal communication, November 25, 2013). She had a verbal IQ score below the mean for her age, as determined by her performance on the WISC revised edition, in Spanish (Wechsler, 1993) (Paloma Garcia-Bellido, personal communication, November 28, 2013). Additionally, through cytogenetic analysis it was shown that a region on chromosome 7 close to the location of *FOXP2* was inverted (with breakpoints on 7q31 and 7p21-15), and a part of the inverted chromosome 7 had undergone a reciprocal translocation to chromosome 11 (with

breakpoints on 7q21, which was found on the p arm of the inverted chromosome 7, and on 11p12). Figure 8.1 shows the location of the inversion on chromosome 7 using DAPI banding, which identifies A=T-rich regions on chromosomes, the patterns of which should be the same even in different people. Figure 8.2 shows the reciprocal translocation between chromosomes 7 and 11 through multicolour fluorescence *in situ* hybridization (FISH). This method allows staining each chromosome type with a different colour. The two figures were kindly provided to me by Daniela Moralli. Following these two analyses, a higher-resolution FISH was then performed using bacterial artificial chromosomes as probes. Based on the results of the FISH, the positions of the four breakpoints were expected to fall within the following fragments (hg19): chromosome 7 114,528,369-114,556,605; chromosome 7 20,954,043-21,001,537; chromosome 7 93,884,065-93,933,453; chromosome 11 38,601,145-38,621,572. The first breakpoint occurred close to *FOXP2*, and, for that reason, it was the focus of the work described later in this chapter.

The discovery that a breakpoint existed close to *FOXP2* prompted an investigation of the possible involvement of *FOXP2* in the language disorder presented by this case. This chapter describes the following investigations:

- A mutation screen of the coding regions of *FOXP2*. The coding regions were amplified and then sequenced using primers used by MacDermot *et al.* (2005) and additional primers designed by me.
- An analysis of *FOXP2* expression in the entire family.
- The mapping of the inversion breakpoints on chromosome 7 using a long-range PCR technique.

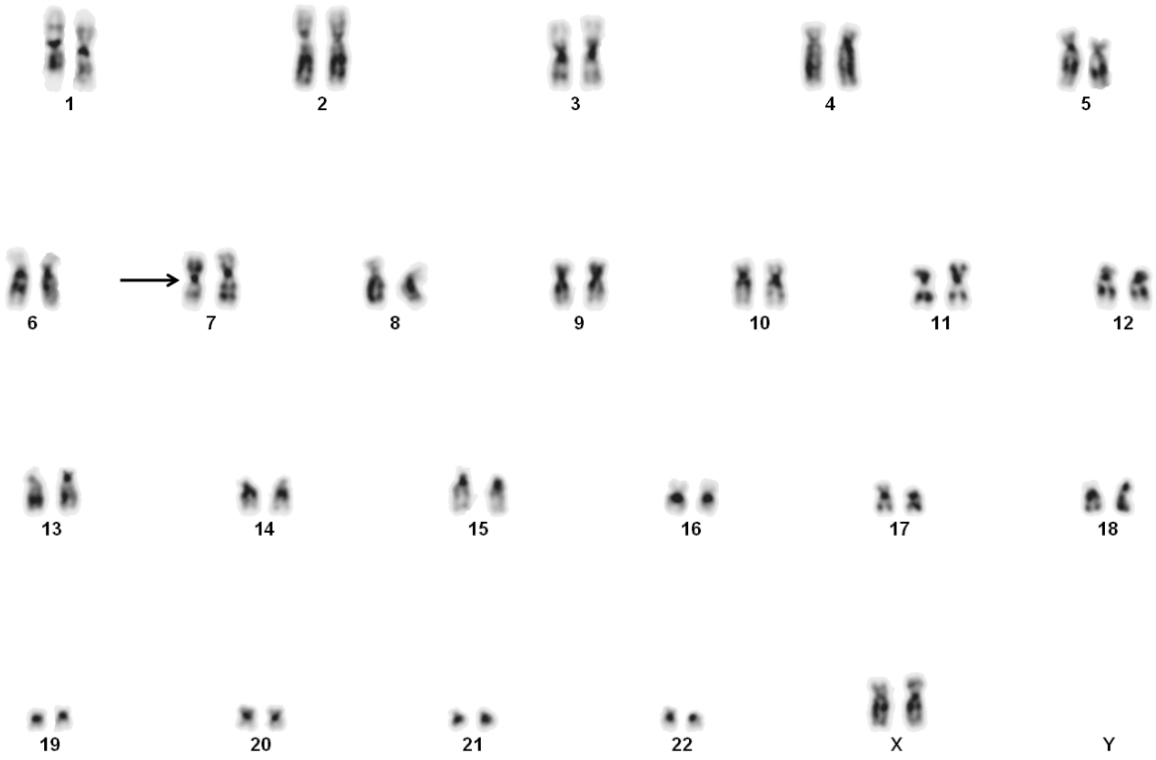


Figure 8.1: DAPI staining of the chromosome of the proband showing the inverted region on chromosome 7 (indicated by an arrow).

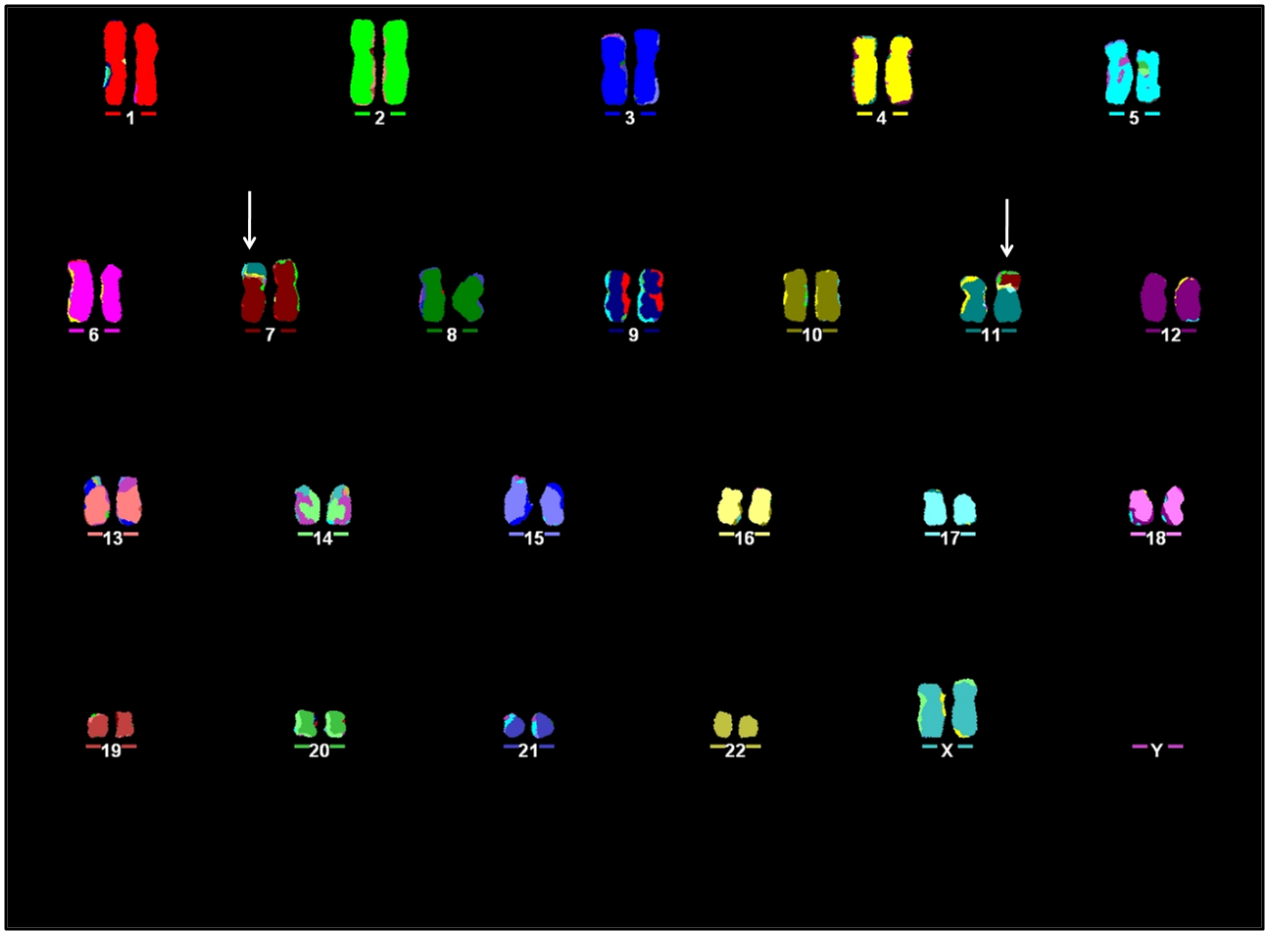


Figure 8.2: Multicolour FISH of the chromosomes of the proband showing the reciprocal translocation between chromosomes 7 and 11 (indicated by arrows).

8.1: Mutation Screen for *FOXP2*

As the behavioural phenotype of A was similar to that of affected KE family members, the coding regions of *FOXP2* were sequenced to see whether a non-synonymous coding change or a splice site mutation were present. Twenty three genomic fragments based on *FOXP2* exons across several isoforms were sequenced. The coding regions and splice sites were identical to those of the reference sequence. Therefore, it is likely that the FOXP2 protein which is translated from transcripts transcribed from the inverted and translocated copy of *FOXP2* is functional. However, it is possible that regulatory elements 3' to *FOXP2* were separated from the gene by the inversion (and, later, by the translocation). Therefore, the next stages of the investigation involved an examination of *FOXP2* expression and the fine-mapping of the breakpoint which separated the hypothetical regulatory elements from *FOXP2*.

8.2: Expression Analysis for *FOXP2*

I performed qPCRs to check whether *FOXP2* expression in the proband differed from that in the other family members. RNA was extracted from fibroblasts obtained from each family member. This cell type was selected due to the unsuitability of lymphoblastoids for *FOXP2* gene expression studies (as *FOXP2* is not expressed in those cells), despite the fact that they are easy to obtain from a blood sample. Furthermore, the epidermis (skin) and the nervous system originate from the same primary germ cell layer in early embryonic development, the ectoderm. Therefore, it was assumed that fibroblasts could be used to study *FOXP2* expression. Given that obtaining cells from the nervous system of the family members was not feasible, fibroblasts were chosen instead. I used two different assays which detected different *FOXP2* transcripts to get a better resolution of the expression levels: Hs01074134_m1 for the detection of NM_148899.3, NM_148900.3, NM_014491.3, and NM_001172766.2, and Hs01081804_m1 for the detection of NM_148898.3 and NM_001172767.2. Figure 8.3 shows the various *FOXP2* transcripts.

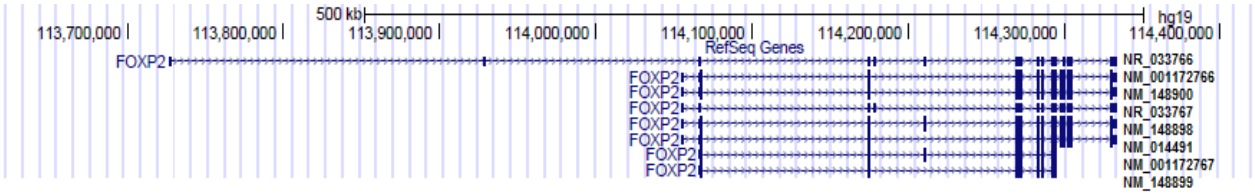


Figure 8.3: *FOXP2* transcripts (source: UCSC Genome Browser). Transcripts with IDs starting with NR are non-coding, and transcripts with IDs starting with NM are coding. The blocks represent exons.

The endogenous control gene that I used was *IPO8* (importin 8) (TaqMan assay Hs00183533_m1). This assay was selected because it performed well in a qPCR which used 32 housekeeping genes from the TaqMan endogenous control plate (the C_T value for *IPO8* was lower than those for the two *FOXP2* assays, but the difference was not as large as with some of the other control genes). Triplicate samples from three RNA batches were used, which differed in the number of cell passages (batch 1 being the one with the lowest number of cell passages). All the negative controls were clean, except for two reverse-transcriptase negative controls (one per triplicate) which had very low C_T values (much lower than the actual sample – a difference of ~10 cycles). Given that the two other samples in the respective triplicates did not have any C_T values, and that the same wells did not have C_T values with the other assay (the qPCR was a multiplex PCR), it is highly likely that the values observed were the result of a measurement error. The C_T values are given in Table 8.1. They are based on the mean value of each triplicate, except for the case of RNA batch 1, for which the C_T values for the endogenous control gene (for Hs01081804_m1) are based on two out of the three samples per triplicate, as each triplicate had an outlier with a very high C_T value. C_T values across the three RNA batches were averaged before calculating the relative expression levels. Table 8.2 includes the expression levels in the proband, relative to the other family members. Expression levels were quantified with the $\Delta\Delta C_T$ method using the proband as a reference. Using the Mann-Whitney U test (Wilcoxon rank-sum test), I compared the proband's ΔC_T values with those of each family member, separately, with the values for each RNA batch used as individual data points. All comparisons had $P>0.05$. The p-values were two-sided.

Table 8.1: Average C_T values (across three RNA batches) obtained from the qPCR.

Assay	Proband	Sibling	Mother	Father
Average C_T values				
FOXP2 (Hs01074134_m1)	29.83	30.39	29.91	29.96
FOXP2 (Hs01081804_m1)	28.97	29.32	28.09	29.59
IPO8 (Hs00183533_m1) for Hs01074134_m1	27.95	28.12	28.24	28.42
IPO8 (Hs00183533_m1) for Hs01081804_m1	30.66	29.29	29.20	29.90
Standard errors for C_T values				
FOXP2 (Hs01074134_m1)	0.18	0.40	0.32	0.29
FOXP2 (Hs01081804_m1)	0.50	0.30	0.36	0.16
IPO8 (Hs00183533_m1) for Hs01074134_m1	0.12	0.35	0.31	0.15
IPO8 (Hs00183533_m1) for Hs01081804_m1	0.72	0.31	0.21	0.21

Table 8.2: *FOXP2* expression levels compared to the proband.

Assay	Proband	Sibling	Mother	Father
FOXP2 (Hs01074134_m1)	1	1.01	1.03	0.94
FOXP2 (Hs01081804_m1)	1	0.86	0.77	0.89

The consistent findings are that the proband has higher expression levels compared with the other family members for assay Hs01081804_m1 (Figure 8.4). This could suggest that there is a problem with *FOXP2* regulation in the proband, and that higher levels of Hs01081804_m1 could be associated with the language phenotype. In this family, the mother and sibling have normal language, but the father has a history of language problems. That said, the differences in the expression levels were not significant, and, even if the differences reflected a true and consistent variation, they could not be associated with the chromosomal rearrangement, since only the proband had it. It is possible that the cell type used (fibroblast, a type of skin cell) was after all not appropriate for studying *FOXP2* expression due to the low expression levels of that gene in that cell type.

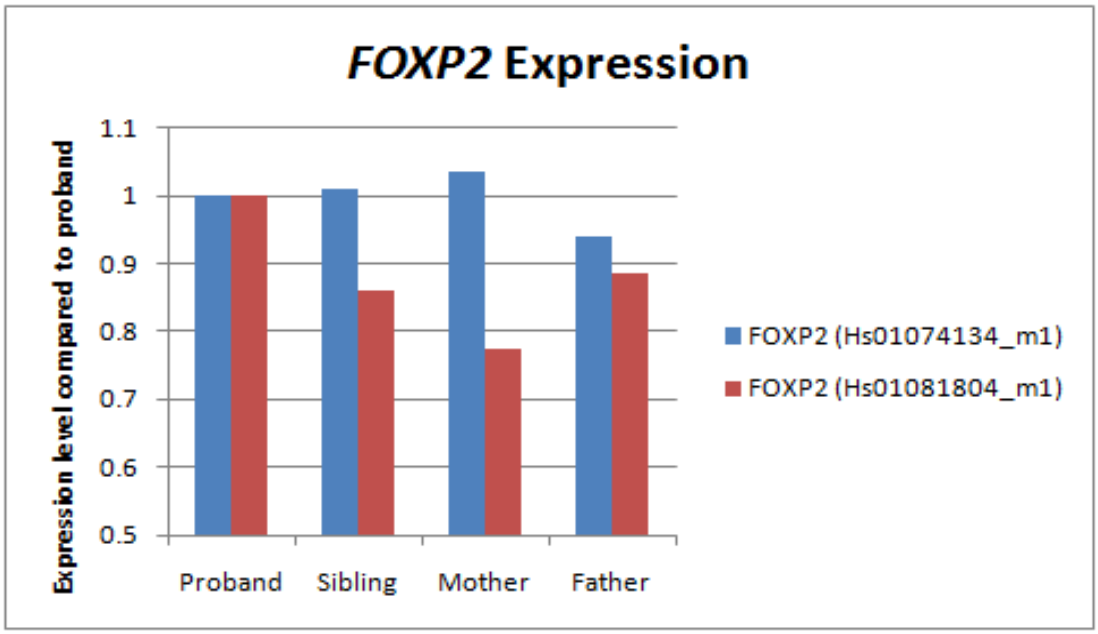


Figure 8.4: *FOXP2* expression in family members (compared to proband).

8.3: Mapping the Inversion Breakpoints on Chromosome 7

As mentioned previously, the inverted region on chromosome 7 included *FOXP2* and did not disrupt it. Despite the fact that *FOXP2* itself was not broken when the inversion or the translocation occurred, it was possible that it was separated from a region important to its function, which remained on the long arm of chromosome 7. However, since the cytogenetic methods did not have the necessary resolution to pinpoint the positions of the breakpoints (due to the nature of the probes used), they had to be found using other methods. The method chosen for this purpose was to try to amplify by PCR the region on the abnormal chromosome which contained the breakpoint. In order to achieve this, primers were designed in such a way that both would bind to the same DNA strand in the reference sequence; following the inversion, the (+) strand of the region before the inversion would be connected to the (-) strand of the region after the breakpoint. As primers were designed using a reference genome, both forward and reverse primers were designed to bind to the same strand. The positions to which the primers bound spanned the hypothesised region of the breakpoint, based on the cytogenetic analysis. Moreover, each primer pair was designed in such a way that one primer would bind to a sequence on 7q, and the other would bind to a sequence on 7p. In that way, the PCR is expected to work only if the region between the positions to which the primers bind actually contains the breakpoint. Several combinations of primers were used, and, eventually, a PCR product was obtained with the following primers: TCATGCAATGTGTCCCCAAA, GATTTGCTTAACTGCCCTGC. Given the size of the product obtained (4-5 kbp, which fit our prediction), it was consequently sequenced using nested Sanger sequencing reactions (i.e. new primers were designed to cover the sequence of the PCR product, based on the primers that worked for the PCR). One of the sequenced fragments (which was sequenced with the primer GACTATTTCCAGCCTCTTTATCCT) contained a sequence which mapped partly to chromosome 7p21 (- strand) and partly to chromosome 7q31 (+ strand) (Figure 8.5). The locations of the inversion breakpoints are 20,969,207 on chromosome 7p21 and 114,539,340 on chromosome 7q31 (hg19). The distances of these breakpoints from the genes nearest to them were as follows: for 7q31: distance to *FOXP2*: 205.5 kbp; distance to MyoD family inhibitor domain containing (*MDFIC*)

(OMIM#614511): 22.9 kbp; for 7p21: distance to Sp8 transcription factor (*SP8*) (OMIM#608306): 142.7 kbp; distance to Sp4 transcription factor (*SP4*) (OMIM#600540): 498.5 kbp. The breakpoint on 7q31 is found 205.5 kbp distal to the 3' end of *FOXP2*. There are reports in the literature of chromosomal rearrangements which were outside of genes in which mutations were shown to cause a certain disorder, and those rearrangements resulted in the same disorder. A good example thereof is that of the paired box 6 (*PAX6*) gene. Mutations in this gene are associated with an eye malformation called aniridia (Glaser *et al.*, 1994). Interestingly, several studies found that chromosomal rearrangements after the 3' end of *PAX6*, even at a distance of more than 100 kbp away from the gene, might still result in aniridia (Fantes *et al.*, 1995, Lauderdale *et al.*, 2000).

Interestingly, the breakpoint on 7p21 disrupts a long intergenic non-coding RNA (lincRNA), *LINC01162* on ENSEMBL, or *AC006481.1*, which is found on chromosome 7:20,875,050-21,062,767 (hg19). There is a growing body of evidence to suggest that long non-coding RNAs are involved in fine-tuning of gene regulation, and that they are important for brain function and development (Barry, 2014). A study which examined the expression of long non-coding RNAs in the brains of subjects with autism found 222 such RNAs that were differentially expressed in the brains of the subjects with autism compared to controls (only the cerebellum and the prefrontal cortex were analysed), and that more than 90% of them were within 500 kbp of a known gene (Ziats and Rennert, 2013). Recently, a study of a subject with developmental delay implicated a lincRNA that had been disrupted by a translocation (Talkowski *et al.*, 2012). A functional study into the involvement of lincRNAs in development in knockout mice for some lincRNAs suggested that they play important roles in development, and that one was notably crucial for the development of the cerebral cortex (Sauvageau *et al.*, 2013). One mechanism through which non-coding RNAs may be regulating gene expression is interaction with RNA-binding proteins. Wang *et al.* (2008a) showed experimentally that non-coding RNA could bind to an RNA-binding protein, inducing an allosteric modification which caused the protein to act as a repressor. Non-coding RNAs can function as both cis- and trans-regulators (Guttman and Rinn, 2012). In the context of cis-regulation, it is interesting to note that the flanking genes of *LINC01162*, *SP8* and *SP4*, both transcription factors, have been shown to be involved in brain development and

neuropsychiatric disorders, respectively. *Sp8*-mutant mice have multiple brain malformations, and it was shown that the gene has an important role during cortical development (Zembrzycki *et al.*, 2007). A SNP in *SP4* showed the second-strongest association in a meta-analysis of three major depressive disorder GWAS (Shyn *et al.*, 2011). Associations with SNPs in *SP4* were also reported for bipolar disorder and schizophrenia (Zhou *et al.*, 2009). Reduced SP4 protein levels in the cerebellum and the prefrontal cortex of subjects with bipolar disorder were also reported, as well as reduced *SP4* transcript and SP4 protein levels in the prefrontal cortex in the same subjects (Pinacho *et al.*, 2011).

Thus, it is possible that the lincRNA disrupted by the inversion breakpoint on 7p21 may be related to the language phenotype in the proband, but further studies will be required in order to substantiate this hypothesis.

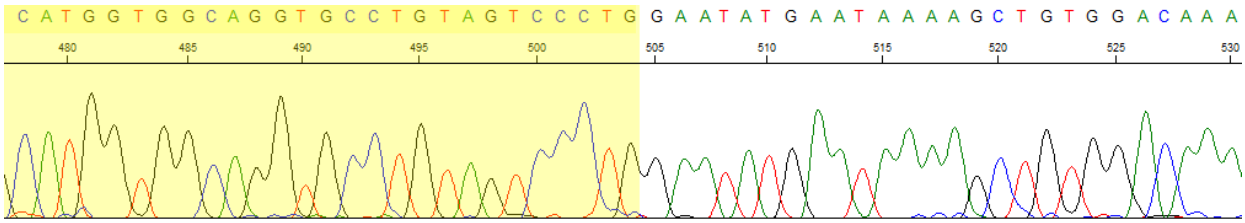


Figure 8.5: The fragment containing the inversion breakpoint on chromosome 7 (the highlighted sequence belongs to chromosome 7p21, and the rest of the sequence belongs to chromosome 7q31).

8.4: Summary of Chapter 8

This chapter describes part of a study of a Spanish girl who has a chromosomal inversion and translocation encompassing the *FOXP2* gene and suffers from a language disorder. No coding changes were found in *FOXP2*, and the cytogenetic analysis showed that the gene was not directly disrupted as a result of the chromosomal rearrangements, suggesting that if *FOXP2* is involved in her language disorder, it is likely to involve dysregulation of gene expression. Some differences in *FOXP2* expression levels between the proband and family members were found, indicating higher levels of some transcripts in the proband, but those were not significant, and it is possible that they are the result of natural variation and/or of using fibroblasts, which may be not suitable for examining *FOXP2* expression. The inversion breakpoint which occurred close to *FOXP2* was mapped to a region 205.5 kbp away from the 3' end of *FOXP2*. The other inversion breakpoint, on 7p21, disrupts a lincRNA gene. Such genes are thought to be involved in gene regulation and have been implicated in neurodevelopmental disorders. Thus, another possibility is that *FOXP2* is not involved in the language disorder, and it is the disruption of *LINC01162* that caused it. Further experimental work is needed in order to determine whether changes in the sequence in that region may affect the function of *FOXP2*, or whether the lincRNA is involved in the language phenotype of the proband. In order to investigate the implication of *FOXP2* in this language phenotype a better-suited cell type should be used, possibly by generating induced pluripotent stem cells from the fibroblasts and differentiating them into neuronal cells. Furthermore, a functional study of the lincRNA which would examine its expression pattern as well as the effects of disrupting it on the expression levels of its flanking genes and/or on genes which are involved in brain development would be helpful in determining whether it plays a role in language abilities.

Chapter 9: General Discussion and Concluding Remarks

Developmental language impairments are among the most common neurodevelopmental disorders. Several of these disorders have been shown to have a strong genetic component, which may manifest itself as a dominant monogenic disorder with full penetrance, as in the KE family members who have a mutation in *FOXP2*, or through complex genetic architectures, as in the case of specific language impairment (SLI). The work described in this thesis focused on the genetics of specific language impairment, a neurodevelopmental disorder in which the development of language is impaired, in the absence of other neurological and behavioural impairments. Additionally, a case study of a language-impaired child in whom an additional atypical behavioural phenotype was also observed is also presented.

Through family and twin studies, it has been shown that SLI is highly heritable. Past studies into the genetics of SLI consisted of linkage analyses or targeted association analyses.

The first part of the thesis describes association analyses of various kinds (including a genome-wide study) and analyses of high-throughput sequencing and exome sequencing data whose purpose was to identify SLI candidate genes, genetic variants which confer risk of SLI, and mutations which potentially drove previously-reported associations in known SLI candidate genes.

The first chapter in this thesis described a GWAS of SLI, which used a family-based design in order to maximise the power of the study and allowed for the testing of parent-of-origin effects, maternal genetic effects, and interaction effects, in addition to the effects of the child's genotype on the risk. Genome-wide significant paternally-associated variants on chromosome 14 were observed. The top association on chromosome 14 (Figure 3.10, $P=3.74\times10^{-8}$) was a non-synonymous coding change in the gene *NOP9*, which codes for an RNA-binding protein and was implicated in schizophrenia. Thus the GWAS identified a new candidate gene for SLI, *NOP9*. Suggestive maternally-associated variants on chromosome 5

were also observed (Figure 3.11, minimum $P=1.16\times10^{-7}$). Those did not fall within any known gene, but the region in which they were located had been implicated in autism and ADHD.

The next chapter described the examination of a potential link between SLI and genes involved in immune reaction. The immune system is known to influence neurodevelopment, and several studies highlighted the over-representation of certain HLA types in subjects with some neurodevelopmental disorders. Both family-based quantitative association analyses and case-control association analyses were performed with SNP data and imputed HLA data. The quantitative analyses highlighted two alleles of *HLA-A* (minimum P=0.004) and one allele of *HLA-B* (with parent-of-origin effects, minimum P=0.013). The case-control analyses highlighted *HLA-DRB1* (P=0.004). All three genes have been implicated in neuropsychiatric or neurodevelopmental disorders.

Next, I described an association study of an isolated population from the Robinson Crusoe Island, for which a very high prevalence of SLI had been reported. This population has previously been studied using linkage methods. Following the exome-sequencing of five affected children and the identification of non-synonymous coding variants that were present in at least three out of the five children by our colleagues in Nijmegen, I checked for the presence of the variants in the entire island cohort and performed an association analysis using a model appropriate for a cohort of related individuals. This analysis identified another candidate gene for SLI, *NFXL1*; a non-synonymous coding variant in this gene was significantly more frequent in affected islanders than in unaffected islanders (P=0.0002). This gene is mentioned only in relatively few publications, and, so far, no studies investigating its function have been reported. It is assumed to code for a transcription factor, based on its DNA-binding domains. Following the association analysis, I examined the frequency of the variant in 1000 Genomes populations and through direct genotyping in Colombian and European cohorts and discovered that it was found in some individuals, but they all belonged to South-American populations. The frequency of the variant in those populations was similar to its frequency in unaffected islanders. Moreover, every individual who had the variant, which was predicted to be deleterious, also had one copy of the wild-type allele. In other words, individuals homozygous for the variant were not found in our cohorts or the 1000 Genomes

cohort. In addition to that, I used gene expression data in an association analysis and a correlation analysis to check which genes were affected by the presence of the variant and which genes were correlated with *NFXL1* expression levels, respectively. However, the former analysis had a small sample size, and the results of the latter analysis were not straightforward to interpret. Additionally, I performed a targeted association analysis of the linkage region on chromosome 7 using new SNP arrays. The top association in this analysis was with a SNP in *CNTNAP2* ($P=1.84\times10^{-5}$), suggesting that it may be involved in the language disorder in the Robinson Crusoe Island.

I then went on to analyse high-throughput sequencing data for four SLI candidate genes: *ATP2C2*, *CMIP*, *CNTNAP2*, and the gene implicated in the Robinson Crusoe Island study, *NFXL1*. LD blocks around previously-identified associations in *ATP2C2*, *CMIP* and *CNTNAP2* were sequenced, together with the coding sequence of *NFXL1*. The sequencing data were based on 14 pools of DNA from SLI probands. Eight non-synonymous coding variants in *ATP2C2* and *NFXL1* were identified, and the rare or novel ones among them were then validated in probands and their family members (DNA was available for three families). One novel variant in *ATP2C2* was predicted to be disease-causing, but it did not co-segregate with the language phenotype or any language trait in the family. The other two non-synonymous coding variants were rare variants in *NFXL1*. One of them, which was found in two families and was predicted to be disease-causing, co-segregated with low NWR in one of the families (the other family was not informative due to missing genotype and phenotype information for the father). The other one was predicted to be a polymorphism and did not co-segregate with the language phenotype in the family. Five non-synonymous variants in another candidate gene for SLI, *ZNF423*, were identified through exome sequencing by our colleagues in Nijmegen. In this case, too, the variants were validated in probands and their families. All of the variants were predicted to be disease-causing; two were novel, two were rare, and one had a frequency of ~0.032. Two *ZNF423* variants appeared to be co-segregating with the language phenotype in two families; however, in both of these cases some parental information was missing. Intronic variants in *ATP2C2*, *CMIP*, *CNTNAP2* and *NFXL1* which were called by Syzygy were not validated, but allele counts for those variants which were also known SNPs were compared against allele counts from general-population

control data. Variants in *CNTNAP2* were at the top of the list (minimum $P=4.24\times10^{-75}$), with very strong associations. The results of that analysis further reaffirmed the role of those genes in SLI.

It is important to note that different variants in the same genes were implicated in different cohorts. Therefore, it is possible that the variants are not causal, but, rather, in LD with the causal variant (and not necessarily with each other) in both populations, as could be the case with the top *CNTNAP2* associations in the SLIC and Robinson Crusoe Island cohorts, or that they independently contribute to the risk, which is likely to be the case when the variants are non-synonymous coding changes, as is the case of the *NFXL1* missense variants.

The second part of the thesis describes two functional studies. The first study investigated *NFXL1* in terms of its gene expression in the brain, splice variants in various tissues, the effects of over-expressing the gene in HEK cells on potential target genes, and subcellular localisation, as shown through immunofluorescence. The cerebellum had high expression levels relative to total brain RNA, and the highest expression level among the brain tissues that were examined (Figure 7.2). This result is interesting in light of recent studies which suggest that the cerebellum may be important for language. I identified a splice variant which results in a transcript that lacks one exon and leads to a frame shift. This shorter transcript was found in all brain regions examined in my analysis, and in various other tissues as well (Figures 7.4 and 7.8). It was always less abundant than the common, longer transcript. In the over-expression experiment, no effects of the over-expression of *NFXL1* in HEK cells on six potential target genes were observed. However, this may be because the cell type might not have been the most appropriate one for testing the effect of *NFXL1* on those genes. In the immunofluorescence experiment, only the over-expressed tagged protein was detected in HEK cells, and it was not present in the cell nucleus, as was expected, based on its predicted function (Figures 7.10-7.13). The endogenous protein was detected in the cytoplasm of SH-SY5Y cells. Again, no nuclear NFXL1 was detected (Figures 7.14-7.17). These results may indicate that NFXL1 is cytoplasmic and requires cellular signals or environmental conditions before it can become active and translocate to the nucleus.

Next, a case-study of a girl with language impairment is discussed. The girl has an inversion on chromosome 7, and part of the inverted sequence translocated (through a reciprocal translocation) to chromosome 11. The inverted – and translocated – part from chromosome 7 contains *FOXP2*, a gene implicated in speech and language disorders. In fact, *FOXP2* is the last gene in the 3' end of the inverted fragment. I performed a mutation screen of the coding regions of *FOXP2* and did not find any mutations. I then performed a gene expression analysis and compared *FOXP2* expression levels in the proband with those in the other family members. One trend was found with certain *FOXP2* transcripts but not with others: it appeared that the proband had higher expression levels for those transcripts than the rest of the family members, followed by the father, who also had some form of language impairment. However, these differences were not significant. The inversion breakpoints were mapped to regions on chromosomal band 7p21 and chromosomal band 7q31 (Figure 8.5). The breakpoint on 7p21 disrupts a lincRNA gene. It has been shown that lincRNAs have a role in gene regulation and can act as both cis- and trans-regulators. One flanking gene of the disrupted lincRNA is involved in brain development, and the other flanking gene has been implicated in neuropsychiatric disorders.

One of the strengths of the work described in this thesis is the use of two distinct cohorts of SLI families. When one gene was implicated in one population, I examined the gene in the other population, where suitable methods were available. In this manner, I found that a SNP in *CNTNAP2* obtained the lowest p-value in a targeted association analysis which focused on the linkage region on chromosome 7, which had previously been identified in the genome-wide linkage study of the Robinson Crusoe Island population. This is the first evidence from an SLI cohort other than the SLIC cohort for the involvement of *CNTNAP2* in SLI. Similarly, I found two rare mutations in *NFXL1*, the gene implicated in the association study of the Robinson Crusoe Island population, as described in this thesis, in three probands from the SLIC cohort. One of these mutations, found in two SLIC families, was predicted to be disease-causing and co-segregated with low NWR in one family (the other family was not informative).

It should be noted that while having access to two SLI cohorts enabled me to examine SLI candidate genes in two populations, the two cohorts were very different from each other in terms of their pedigree structure and the tests administered to them. For that reason, I could not use the exact same methods and/or phenotypes when studying them. In that sense, the SLIC cohort was unsuitable as a true replication cohort for the Robinson Crusoe Island cohort, and *vice versa*. The functional work described in this thesis was also subject to some technical limitations: due to the lack of high-performance antibodies for *NFXL1*, I could not do all the experiments that I had originally planned to do; similarly, due to the lack of published functional work on *NFXL1*, the choice of cell line was not straightforward, and, ideally, more cell lines and experimental conditions should have been used in the experiments. The *FOXP2* expression analyses were also limited in the sense that the RNA was extracted from fibroblast cells, which might not be suitable for the study of this gene, but due to ethical and technical considerations, this was the only viable approach at the time.

The general picture which emerges from my investigations supports the notion that SLI has a complex genetic architecture. While strong associations with variants in new candidate genes as well as in genes previously implicated in SLI were identified, deleterious non-synonymous coding changes in new and old SLI candidate genes were also observed in SLI families. This is in line with the notion that a mutational load across genes relevant to language development (and/or brain development in general) may increase the risk of having a language disorder (or a more general neurodevelopmental disorder), considering that none of the mutations identified here is likely to cause SLI by itself. As can be learned from the results of the functional part of my work, the gap between genetic association and biological mechanism influencing a complex phenotype is sometimes very wide. I believe that a combination of genetic and functional work is the key for gaining an understanding of the genetic basis of complex traits.

In summary, this work has contributed to the identification of new SLI candidate genes, including *NFXL1*, *NOP9*, *ZNF423*, and several HLA genes. Furthermore, the functional study of *NFXL1* described here identified a novel splice variant and found

differential expression levels of the gene in various brain regions. Further work will be required to determine how variants in these genes contribute to SLI susceptibility. Another interesting finding concerns the mapping of two chromosomal breakpoints, one of which disrupts a lincRNA gene, and another, which is distal to the 3' end of *FOXP2*. Functional work on the sequences disrupted by these breakpoints will have the potential to identify a possible regulatory problem that results in language impairment and/or identify regulatory elements important to *FOXP2* function. The results of the work presented here provide new and exciting directions for future research into the genetic basis of developmental language disorders.

Part of the work described in this thesis has been published, and a list of publications is available in Appendix E.

Appendix A: Primers

The primer sequences for the amplifying and the sequencing of FOXP2 exons were the ones used by MacDermot *et al.* (2005),except for the primers used only for sequencing and not for the PCR. F/R (forward/reverse) refers to the strand the primer sequence is on, relative to the sequence used in designing the primer.

Project	ID	Primer sequence	F/R
FOXP2 exons	1	GTAAAGCAATTGCCAAATCTACC	F
FOXP2 exons	1	GATCGGGCAGAGGTGTACTCAC	R
FOXP2 exons	10	GAGGCAAGCTCAATGATAAGATG	F
FOXP2 exons	10	CACCCATGGTGTAGATTGGATAG	R
FOXP2 exons	11	TTGATTCAGCTACAGTTTTCCTC	F
FOXP2 exons	11	AATTAGTTGAGATTGGCTGCTTC	R
FOXP2 exons	12	ATGACACTTTCTCATCCAAGCC	F
FOXP2 exons	12	CTTCCTCACTGAATCACTTTACC	R
FOXP2 exons	13	CCTTGTGCTATCTGTAAGGAAC	F
FOXP2 exons	13	AGTAAGGCTTGGATGAGAAAGTG	R
FOXP2 exons	14	CATTGCCACGAGAATGTTAGC	F
FOXP2 exons	14	TCAATAATGTAGTATGTTGGGCTG	R
FOXP2 exons	15	GTGTGAGACAAGCCAGAACATAC	F
FOXP2 exons	15	CTTGAATTAGAACCAGTGCTGTC	R
FOXP2 exons	15_alt_end_R (sequencing only)	GCCCTAAACATTCAAATACAACA	R
FOXP2 exons	15_alt_start_F (sequencing only)	TTTCAGAAGTCCAACCTTAGTAAAA	F
FOXP2 exons	16	CTGCCACAAGTAGCCAGTTAGG	F
FOXP2 exons	16	CAACTTATACAACAGTAAAAACTTCG	R
FOXP2 exons	17	TGACCTCTTCACTGCAAAGTTGG	F
FOXP2 exons	17	GTCAAATATTCATGGTTGTGGAG	R
FOXP2 exons	2	TAGCTTAACACAACATGCTCAG	F
FOXP2 exons	2	GAGAGGGACATCTTGATAATG	R
FOXP2 exons	2b	GGTGGTTGGCTTTGTTTCAT	F
FOXP2 exons	2b	GTTGGTGAGAGTTATGAGGCTAT	R
FOXP2 exons	3	TGGGTCTGCACATCTGTTATC	F
FOXP2 exons	3	GAAAGGAATATGGGAGTTCTTG	R
FOXP2 exons	3a	ACAATGAAGGAAGTGATTACAATGC	F
FOXP2 exons	3a	GCACATACACACATGCACACTTCC	R
FOXP2 exons	3b	CATATAATCTATTTAGTGGGACTGG	F
FOXP2 exons	3b	CTTCCCTATTTTGACCTACCAAG	R
FOXP2 exons	4	GATAACATACTATTTGTGAAGTTG	F
FOXP2 exons	4	TCTAGCACGCTAATAGGTTGTCC	R

Project	ID	Primer sequence	F/R
FOXP2 exons	4a	CAACTGGTAGAGTATAGCCTAGTTTTT	F
FOXP2 exons	4a	TGTTGTTGCTGCTGCTGT	R
FOXP2 exons	5	TGAATCTTAATGGATACTCTGCC	F
FOXP2 exons	5	TCTAAGACTATTCTTGCCGCTC	R
FOXP2 exons	6	AAGGAGTGTGCATTTCCCTG	F
FOXP2 exons	6	CAGAAAGGCCATGAAATGGTAG	R
FOXP2 exons	7	AATTTATGCAGGTAACATCACTG	F
FOXP2 exons	7	CTTTTTCATTGTCTCAATGGTG	R
FOXP2 exons	8	GTGCCTAAAATGCCCATATAATCC	F
FOXP2 exons	8	TGTTTGTCACTGATCGTAACCTG	R
FOXP2 exons	9	GCTTTTTAAGTGTAGCCTATGCC	F
FOXP2 exons	9	CAGAATCTGCTCAGTACTCAATG	R
FOXP2 exons	s2	CAGAAAATAATAGGAAGCTTCAACC	F
FOXP2 exons	s2	AAACCTGAATAGCAATAAAAACTGAG	R
FOXP2 exons	s3	TGAAGTCACTGCACTAATTGTGAT	F
FOXP2 exons	s3	AAGCACAATGACCAAAGGAG	R
Long-range sequencing	CMIP-4	TCGGCTGATGTACAGCTTTG	F
Long-range sequencing	CMIP-4	CCAAGAGGTCGCTTGAAGAC	R
Long-range sequencing	CNTNAP2-10	GCCCATCCATTCCAGTTCTA	F
Long-range sequencing	CNTNAP2-10	ACCATGGGCAACCACTGTAT	R
Long-range sequencing	CNTNAP2-12	TACCCTTGGGCAAAGTCAGT	F
Long-range sequencing	CNTNAP2-12	CAGAAACGTTGGAGACTGGA	R
Long-range sequencing	CNTNAP2-2	AAGTTGATGCATTCCCTTGG	F
Long-range sequencing	CNTNAP2-2	TTGGGGCAAACAGAGAAAAC	R
Long-range sequencing	CNTNAP2-7	AGGCTCCAATAAGCAGGACA	F
Long-range sequencing	CNTNAP2-7	GGCAGACCCTCAGCTTACAG	R
Long-range sequencing	CNTNAP2-8	CCTTGTTGCAGTGACCATAGC	F
Long-range sequencing	CNTNAP2-8	CTGGCACTGTAGGTGGGAAT	R
Long-range sequencing	New_ATP2C2_1	GCTCAGTCTGGGAAAACAGC	F
Long-range sequencing	New_ATP2C2_1	CCTGTGATGGGCAGATAGGT	R
Long-range sequencing	New_ATP2C2_3	GTGCATCCAATCCCTGTCTT	F
Long-range sequencing	New_ATP2C2_3	CACAGGCAGCACTTCCTACA	R
Long-range sequencing	New_CMIP_2	TGTGGTCAGTGGCTCTGAAG	F
Long-range sequencing	New_CMIP_2	GCTCAGTCCCCTCTGCTATG	R
Long-range sequencing	New_CMIP_5	TGCGGGAAAATAATGGAGAG	F

Project	ID	Primer sequence	F/R
Long-range sequencing	New_CMIP_5	AGCCAGAAGGCAGGTTCATA	R
Long-range sequencing	New_CNTNAP2_1	CATAGAAATGCAGCCAGCAA	F
Long-range sequencing	New_CNTNAP2_1	GGCAGTGCACCATCATATTG	R
Long-range sequencing	New2_ATP2C2_2a	GGATGGCAAGGTGAAGACAT	F
Long-range sequencing	New2_ATP2C2_2a	CCAGAAAGCCCCAATTCACT	R
Long-range sequencing	New2_ATP2C2_2b	TGGCTCTGGTGCATAGTCAG	F
Long-range sequencing	New2_ATP2C2_2b	GGGGGAGGAGAGATTTTGAG	R
Long-range sequencing	New2_ATP2C2_4b	ATGAACCAAGGCTTCCAGTG	F
Long-range sequencing	New2_ATP2C2_4b	TCCTTTCTCACCCCACACTC	R
Long-range sequencing	New2_CMIP1a	CTGTGCTGCACTCACCTGAT	F
Long-range sequencing	New2_CMIP1a	AGGTAATGCCGTTCAGATGG	R
Long-range sequencing	New2_CMIP3a	ATTGGCCTGTTTCTGTGTCC	F
Long-range sequencing	New2_CMIP3a	GAGTCTGTTCCTGCCTCCAG	R
Long-range sequencing	New2_CMIP3b	CCAAACTGCGAAGGAAGAAG	F
Long-range sequencing	New2_CMIP3b	GCGTTAGGTGCTGCTTTAGG	R
Long-range sequencing	New2_CNTNAP2_11a	ATCTCTTTGCCAGGGTTGTG	F
Long-range sequencing	New2_CNTNAP2_11a	CATGGCTTGGACAGAACAGA	R
Long-range sequencing	New2_CNTNAP2_3b	GGACGGCATGAAGACAGATT	F
Long-range sequencing	New2_CNTNAP2_3b	TCTTTCCCTCACAGGAGGTG	R
Long-range sequencing	New2_CNTNAP2_4a	GGTAATCGCCAACAAACACA	F
Long-range sequencing	New2_CNTNAP2_4a	GCCTGGCATAAGTGATGTGA	R
Long-range sequencing	New2_CNTNAP2_4b	TTGTTGACAGTTTGGGAGGA	F
Long-range sequencing	New2_CNTNAP2_4b	ACCTTCAATGTTGCCTCCAC	R
Long-range sequencing	New2_CNTNAP2_5a	AATGAAGGGGGAGAAGGAGA	F
Long-range sequencing	New2_CNTNAP2_5a	CGGTTTTCAAAGCTTCCTTG	R
Long-range sequencing	New2_CNTNAP2_6a	TGAATCAACCCTGGAAGAGG	F
Long-range sequencing	New2_CNTNAP2_6a	GGGGAATATCACTCCCCCTA	R
Long-range sequencing	New2_CNTNAP2_6b	TTCGAATGCGTGAGTGTCTC	F
Long-range sequencing	New2_CNTNAP2_6b	TCTGTTAAAGCCTGGCTCGT	R
Long-range	New2_CNTNAP2_9a	AACGCCACATGCAGATGTTA	F

Project	ID	Primer sequence	F/R
Long-range sequencing	New2_CNTNAP2_9a	CACATGCTTTGGGGCTAAAT	R
Long-range sequencing	New2_CNTNAP2_9b	TGGGGAACAATCACTTGACA	F
Long-range sequencing	New2_CNTNAP2_9b	TCTAGCTGATGCTTGCTGGA	R
Long-range sequencing	New3_CMIP1b	ATTTCCGTCAGGGGAGGTAG	F
Long-range sequencing	New3_CMIP1b	GATGGGACTACCACCTGCAT	R
Long-range sequencing	New3_CNTNAP2_11b	GCATCTCCATCGTTCAACTG	F
Long-range sequencing	New3_CNTNAP2_11b	CCACCCAAGGTCTAGCTCTG	R
Long-range sequencing	New3_CNTNAP2_5b	TCAGACCATGCTTTCAGCAG	F
Long-range sequencing	New3_CNTNAP2_5b	GCAAAACCACAACTACAGTCATT	R
Long-range sequencing	New4_ATP2C2_primer4a2A2	ATCTGCCAGAGGTGGTTTGA	F
Long-range sequencing	New4_ATP2C2_primer4a2A2	ATCCCTTCTAAGCCCCTGAA	R
Long-range sequencing	New4_ATP2C2_primer4a2B2	TTTTCTCCACCCCTCCTCTT	F
Long-range sequencing	New4_ATP2C2_primer4a2B2	CAGGCAGAATGACCTCTGCT	R
Long-range sequencing	New4_CNTNAP2_3a1_other	TGGAACTATCCCAGAATTCTATGTC	F
Long-range sequencing	New4_CNTNAP2_3a1_other	CGGGTGCATGAGACTCTTTT	R
Long-range sequencing	new5_CNTNAP2_3a2a	CCACATGGTCCCCAGTAAAG	F
Long-range sequencing	new5_CNTNAP2_3a2a	TCTATGCCATCAAGAGATAAGATGA	R
Long-range sequencing	new6_CNTNAP2_3a2b_alt/ new6_CNTNAP2_3a2b_other_R	TCATCTTATCTCTTGATGGCATAGA	F
Long-range sequencing	new6_CNTNAP2_3a2b_alt/ new6_CNTNAP2_3a2b_other_R	TGTGGAAAGAAAAAGTAACCCATA	R
Long-range sequencing	NFXL1_1	GGGGCTTCTGGGAGTTGTAG	F
Long-range sequencing	NFXL1_1	GCCACTGCCATTCTATTTGC	R
Long-range sequencing	NFXL1_2	CCTCAGGGTCTTTTCATTCAAC	F
Long-range sequencing	NFXL1_2	GGGAGGATAAAACTGGGAATCT	R
Long-range sequencing	NFXL1_3	CAAAGATGGAGCAGAGTTCCTT	F
Long-range sequencing	NFXL1_3	GGGCCCAAAATGGAGAAA	R
Long-range sequencing	NFXL1_4	CCCACTGGCAAAGTGATAGG	F
Long-range sequencing	NFXL1_4	GTGTTGGTGAGAGCATACCTGA	R
Long-range sequencing	NFXL1_5	TTGACTGTGCTTGCTACTAACC	F
Long-range sequencing	NFXL1_5	CAACTGGGGGTAAAGACTCAAA	R

Project	ID	Primer sequence	F/R
Long-range sequencing	NFXL1_6	ATTGGTCACACAGTCAAGTTGG	F
Long-range sequencing	NFXL1_6	CATGTCTTCCTTCCAGGTCTCT	R
Long-range sequencing	NFXL1_7	TAGTAAAGGGCTGGATGGTTTG	F
Long-range sequencing	NFXL1_7	GATTGGCCAAATGTAGAAGGAC	R
Long-range sequencing	NFXL1_8	GCTAGGATTGATTGACCCTTTG	F
Long-range sequencing	NFXL1_8	ATGGTCTCCAGAAGATGCCTTA	R
Long-range sequencing	NFXL1_9	CTCCCAGTCTCCTTTCTGCTTA	F
Long-range sequencing	NFXL1_9	TCCATCAATGACTGTCTTCCAC	R
NFXL1 cDNA	NFXL1_cDNA1	AGTTCCTTCTGGCACCAGTC	F
NFXL1 cDNA	NFXL1_cDNA1	AGGCCAGGGACAATCTTTCT	R
NFXL1 cDNA	NFXL1_cDNA2	CTTTTCAAGCAGGGGCTATG	F
NFXL1 cDNA	NFXL1_cDNA2	TTATGTTGCCCACAAAGCAA	R
NFXL1 cDNA	NFXL1_cDNA3	GCCAAGTATGTGAGCGTGAA	F
NFXL1 cDNA	NFXL1_cDNA3	GCGACAATGCTTTTCCACTT	R
NFXL1 cDNA	NFXL1_cDNA4	TTTCTTTGCCTTGTACAGAAGATG	F
NFXL1 cDNA	NFXL1_cDNA4	GCAAGGTTGATGACATGGTG	R
NFXL1 cDNA	NFXL1_cDNA5	TGTGGCAATACAAAGGTGACA	F
NFXL1 cDNA	NFXL1_cDNA5	TGGGTGAAGACAACCTAGTGG	R
NFXL1 cDNA	NFXL1_cDNA6	GCCACAAAGTAACCAAAACTGA	F
NFXL1 cDNA	NFXL1_cDNA6	CTGGTTGCAGTTAAAGGGACA	R
NFXL1 cDNA	NFXL1_cDNA6mid_alt1	TGATTTTTGCAACAACTGAGGA	R
NFXL1 cDNA	NFXL1_cDNA7	GTTGCAAAAATCAGTGCCCTA	F
NFXL1 cDNA	NFXL1_cDNA7	TGTCCTTCTAACAACAGCTGAG	R
NFXL1 cDNA	NFXL1_cDNA8alt1	CAGTGTGTGGAGTTGTGGTTG	F
NFXL1 cDNA	NFXL1_cDNA8alt1	ACACGCCCCATGACATTTAT	R
NFXL1 expression	NFXL1_E (exons 3 and 4)	TAACCAAGCTGCAGCCAGAA	F
NFXL1 expression	NFXL1_E (exons 3 and 4)	TGGTTTCTCTTCACCGAAGC	R
RC exome sequencing variants	NM_000418	TGTTGCTGTAGCCTGGTCAC	F
RC exome sequencing variants	NM_000418	CACAGGGGAAGAATGGAGAG	R
RC exome sequencing variants	NM_001004690	CTCAGTCAACTGTTCCTCATGG	F
RC exome sequencing variants	NM_001004690	TGATGATTGCAACAGGAAAAAC	R
RC exome sequencing variants	NM_001042678	TTACTGTGGTAACTGGCCCTCT	F
RC exome sequencing	NM_001042678	CCTCCCTCACCAGGAGATATTA	R

Project variants	ID	Primer sequence	F/R
RC exome sequencing variants	NM_001130141	CTCCTACAGGAGTGGGTAGCAG	F
RC exome sequencing variants	NM_001130141	CTCTCAGAGACCCTTGGGAGAT	R
RC exome sequencing variants	NM_001143766	CCAATTGCATTTTTCCATGA	F
RC exome sequencing variants	NM_001143766	ATGGGACCACAAGAGAAACG	R
RC exome sequencing variants	NM_006044	GGGGTCCTCACTCTCTCTCA	F
RC exome sequencing variants	NM_006044	TCAAAAAGCTTTCTCCAGGTTT	R
RC exome sequencing variants	NM_139160	TTTGAATAATGTCCCATGTCC	F
RC exome sequencing variants	NM_139160	TGCACTGAAGATGGTAAGCA	R
RC exome sequencing variants	NM_148174	AATGCAGCCCTCTCTGACAT	F
RC exome sequencing variants	NM_148174	TGGCTTATGGAATGTGAAGC	R
RC exome sequencing variants	NM_152995	CCTCTTGACAGCTCATTGGTT	F
RC exome sequencing variants	NM_152995	CAAACTAAGGTATTTTGGGGAGA	R
Variants in ATP2C2 and NFXL1	ATP2C2-00	CATTCAGCACAGCCAAGAAA	F
Variants in ATP2C2 and NFXL1	ATP2C2-00	TCTCGTTCCACTGTCTGTCG	R
Variants in ATP2C2 and NFXL1	ATP2C2-1718	TCCCCGTGTGTGTCTTTGTA	F
Variants in ATP2C2 and NFXL1	ATP2C2-1718	TGGGGGTCACTAAGGAACTG	R
Variants in ATP2C2 and NFXL1	ATP2C2-32	GAGATTCCACAGCCTTTTCC	F
Variants in ATP2C2 and NFXL1	ATP2C2-32	AGCTTACAAAAGTCCCGAAGC	R
Variants in ATP2C2 and NFXL1	ATP2C2-50	TCAGTAAACCGTCTCCAGCA	F

Project	ID	Primer sequence	F/R
Variants in ATP2C2 and NFXL1	ATP2C2-50	TGTTGCTGGCAACTTCAGAG	R
Variants in ATP2C2 and NFXL1	ATP2C2-77	TCCTTACTCCCCCTCTCTCC	F
Variants in ATP2C2 and NFXL1	ATP2C2-77	CTTCCCTTTTGCAGAACCTG	R
Variants in ATP2C2 and NFXL1	ATP2C2-84	TTGGCGTAATTGGAAGCTCT	F
Variants in ATP2C2 and NFXL1	ATP2C2-84	TATTTCCCTCATGGCGCTTT	R
Variants in ATP2C2 and NFXL1	ATP2C2-84alt	TGGAAGCTCTCTGAATGTGC	F
Variants in ATP2C2 and NFXL1	ATP2C2-84alt	AATGCCTATTTCCCTCATGG	R
Variants in ATP2C2 and NFXL1	NFXL1-52	TTTGCACAAGCTTCCATCTG	F
Variants in ATP2C2 and NFXL1	NFXL1-52	TACAACCCACAGCAATGCAC	R
Variants in ATP2C2 and NFXL1	NFXL1-65	TTTATTTCCATGCTGTTTTTCAG	F
Variants in ATP2C2 and NFXL1	NFXL1-65	AGGCATCTAGAGTTTAAGCAAAGT	R
Variants in ATP2C2 and NFXL1	NFXL1-6876	TCTCTCATGTTGCTTTTAAAGGTG	F
Variants in ATP2C2 and NFXL1	NFXL1-6876	GAAGAAATAATCACCCACAAACAA	R
Variants in ATP2C2 and NFXL1	NFXL1-75	AGCCTGAAGAACATGGCACT	F
Variants in ATP2C2 and NFXL1	NFXL1-75	AAACAACCACAGGACAATCTCA	R
Variants in ATP2C2 and NFXL1	NFXL1-75 (used for C89.4)	AGCCTGAAGAACATGGCACT	F
Variants in ATP2C2 and NFXL1	NFXL1-75 (used for C89.4)	AAACAACCACAGGACAATCTCA	R
ZNF423	znf423-23alt	ATGTGTCCCATCTGTGGTGA	F
ZNF423	znf423-23alt	ACGTCAAGCTTCACCAGGTC	R
ZNF424	znf423-8177	GTCCGATGTGGAGGTGTCTT	F
ZNF424	znf423-8177	TCTCATTGCTGTGCTTCACC	R
ZNF425	znf423-96	TGCTCCATGGGTTTCCTTAC	F
ZNF425	znf423-96	GGCCAGTGGAATGTTCTTGT	R

Project	ID	Primer sequence	F/R
ZNF426	znf423-99	CCTTCATCCGCTTGAGCTAC	F
ZNF426	znf423-99	CAGGTGGATCTTGAGGTGGT	R

Appendix B: HLA Allele Counts and Allele Frequencies in Probands and in Controls

This appendix presents the HLA Allele Counts and Allele Frequencies in Probands and in Controls as well as 95% confidence intervals using the Wilson score for the allele frequencies. The Wilson score intervals were calculated using the following formula:

$$\frac{1}{1+\frac{1}{n}z^2}\left[\hat{p}+\frac{1}{2n}z^2 \pm z\sqrt{\frac{1}{n}\hat{p}(1-\hat{p})+\frac{1}{4n^2}z^2}\right]$$

Where $\hat{p}$ is the allele frequency of the allele in question, n is the total allele count for the gene in question, and z is the $1-\frac{\alpha}{2}$ percentile of a standard normal distribution. For 95% confidence intervals, z=1.96.

Locus	Allele	Allele count in probands	Total allele count in probands	Frequency in probands	95% CI lower boundary	95% CI upper boundary
HLA-A	A1	72	439	0.164	0.132	0.202
HLA-A	A11	35	439	0.08	0.058	0.109
HLA-A	A2	130	439	0.296	0.255	0.34
HLA-A	A23	10	439	0.023	0.013	0.042
HLA-A	A24	45	439	0.103	0.078	0.135
HLA-A	A26	1	439	0.002	0	0.012
HLA-A	A29	21	439	0.048	0.032	0.072
HLA-A	A3	63	439	0.144	0.114	0.18
HLA-A	A30	16	439	0.036	0.022	0.058
HLA-A	A31	13	439	0.03	0.018	0.05
HLA-A	A32	12	439	0.027	0.015	0.047
HLA-A	A33	1	439	0.002	0	0.012
HLA-A	A66	1	439	0.002	0	0.012
HLA-A	A68	19	439	0.043	0.028	0.066
HLA-B	B13	12	443	0.027	0.015	0.047
HLA-B	B18	21	443	0.047	0.031	0.071
HLA-B	B27	15	443	0.034	0.021	0.055
HLA-B	B35	34	443	0.077	0.056	0.106
HLA-B	B37	7	443	0.016	0.008	0.033
HLA-B	B38	7	443	0.016	0.008	0.033
HLA-B	B39	0	443	0	0	0.009
HLA-B	B3901	2	443	0.005	0.001	0.017
HLA-B	B41	2	443	0.005	0.001	0.017
HLA-B	B44	73	443	0.165	0.133	0.202

Locus	Allele	Allele count in probands	Total allele count in probands	Frequency in probands	95% CI lower boundary	95% CI upper boundary
HLA-B	B45	1	443	0.002	0	0.012
HLA-B	B47	1	443	0.002	0	0.012
HLA-B	B49	6	443	0.014	0.007	0.03
HLA-B	B50	3	443	0.007	0.002	0.02
HLA-B	B51	21	443	0.047	0.031	0.071
HLA-B	B52	4	443	0.009	0.004	0.023
HLA-B	B53	0	443	0	0	0.009
HLA-B	B55	16	443	0.036	0.022	0.058
HLA-B	B56	4	443	0.009	0.004	0.023
HLA-B	B57	10	443	0.023	0.013	0.042
HLA-B	B58	5	443	0.011	0.005	0.026
HLA-B	B60	27	443	0.061	0.042	0.087
HLA-B	B61	3	443	0.007	0.002	0.02
HLA-B	B62	31	443	0.07	0.05	0.098
HLA-B	B63	1	443	0.002	0	0.012
HLA-B	B64	3	443	0.007	0.002	0.02
HLA-B	B65	10	443	0.023	0.013	0.042
HLA-B	B7	65	443	0.147	0.117	0.183
HLA-B	B71	3	443	0.007	0.002	0.02
HLA-B	B72	0	443	0	0	0.009
HLA-B	B8	56	443	0.126	0.098	0.16
HLA-C	Cw1	20	395	0.051	0.033	0.077
HLA-C	Cw10	38	395	0.096	0.071	0.129
HLA-C	Cw2	14	395	0.035	0.021	0.058
HLA-C	Cw4	51	395	0.129	0.099	0.166
HLA-C	Cw5	50	395	0.127	0.098	0.163
HLA-C	Cw6	34	395	0.086	0.062	0.118
HLA-C	Cw7	141	395	0.357	0.311	0.405
HLA-C	Cw8	14	395	0.035	0.021	0.058
HLA-C	Cw9	33	395	0.084	0.06	0.116
HLA-DQA	0101	79	453	0.174	0.142	0.212
HLA-DQA	0102	93	453	0.205	0.17	0.245
HLA-DQA	0103	23	453	0.051	0.034	0.075
HLA-DQA	0201	59	453	0.13	0.102	0.164
HLA-DQA	0301	79	453	0.174	0.142	0.212
HLA-DQA	0401	6	453	0.013	0.006	0.028
HLA-DQA	0501	114	453	0.252	0.214	0.294
HLA-DQB	DQ2	73	419	0.174	0.141	0.213
HLA-DQB	DQ4	6	419	0.014	0.006	0.03
HLA-DQB	DQ5	83	419	0.198	0.163	0.239
HLA-DQB	DQ6	124	419	0.296	0.254	0.341
HLA-DQB	DQ7	67	419	0.16	0.128	0.198
HLA-DQB	DQ8	46	419	0.11	0.084	0.144
HLA-DQB	DQ9	20	419	0.048	0.031	0.073

Locus	Allele	Allele count in probands	Total allele count in probands	Frequency in probands	95% CI lower boundary	95% CI upper boundary
HLA-DRB	DR1	42	395	0.106	0.079	0.14
HLA-DRB	DR10	7	395	0.018	0.009	0.036
HLA-DRB	DR103	5	395	0.013	0.006	0.03
HLA-DRB	DR11	12	395	0.03	0.017	0.052
HLA-DRB	DR12	2	395	0.005	0.001	0.018
HLA-DRB	DR13	55	395	0.139	0.108	0.177
HLA-DRB	DR14	12	395	0.03	0.017	0.052
HLA-DRB	DR15	77	395	0.195	0.159	0.237
HLA-DRB	DR16	3	395	0.008	0.003	0.023
HLA-DRB	DR17	64	395	0.162	0.129	0.202
HLA-DRB	DR4	43	395	0.109	0.082	0.144
HLA-DRB	DR7	57	395	0.144	0.113	0.182
HLA-DRB	DR8	7	395	0.018	0.009	0.036
HLA-DRB	DR9	9	395	0.023	0.012	0.043

Locus	Allele	Allele count in controls	Total allele count in controls	Frequency in controls	95% CI lower boundary	95% CI upper boundary
HLA-A	A1	202	1040	0.194	0.171	0.219
HLA-A	A11	72	1040	0.069	0.055	0.086
HLA-A	A2	329	1040	0.316	0.288	0.345
HLA-A	A23	15	1040	0.014	0.008	0.023
HLA-A	A24	85	1040	0.082	0.067	0.1
HLA-A	A26	5	1040	0.005	0.002	0.011
HLA-A	A29	36	1040	0.035	0.025	0.048
HLA-A	A3	161	1040	0.155	0.134	0.178
HLA-A	A30	27	1040	0.026	0.018	0.038
HLA-A	A31	32	1040	0.031	0.022	0.043
HLA-A	A32	24	1040	0.023	0.015	0.034
HLA-A	A33	0	1040	0	0	0.004
HLA-A	A66	2	1040	0.002	0.001	0.007
HLA-A	A68	50	1040	0.048	0.037	0.063
HLA-B	B13	19	1055	0.018	0.012	0.028
HLA-B	B18	44	1055	0.042	0.031	0.056
HLA-B	B27	28	1055	0.027	0.019	0.039
HLA-B	B35	61	1055	0.058	0.045	0.074
HLA-B	B37	17	1055	0.016	0.01	0.026
HLA-B	B38	6	1055	0.006	0.003	0.013
HLA-B	B39	2	1055	0.002	0.001	0.007
HLA-B	B3901	3	1055	0.003	0.001	0.009
HLA-B	B41	7	1055	0.007	0.003	0.014

Locus	Allele	Allele count in controls	Total allele count in controls	Frequency in controls	95% CI lower boundary	95% CI upper boundary
HLA-B	B44	191	1055	0.181	0.159	0.205
HLA-B	B45	6	1055	0.006	0.003	0.013
HLA-B	B47	0	1055	0	0	0.004
HLA-B	B49	13	1055	0.012	0.007	0.021
HLA-B	B50	10	1055	0.009	0.005	0.017
HLA-B	B51	41	1055	0.039	0.029	0.052
HLA-B	B52	7	1055	0.007	0.003	0.014
HLA-B	B53	2	1055	0.002	0.001	0.007
HLA-B	B55	25	1055	0.024	0.016	0.035
HLA-B	B56	4	1055	0.004	0.002	0.01
HLA-B	B57	50	1055	0.047	0.036	0.061
HLA-B	B58	4	1055	0.004	0.002	0.01
HLA-B	B60	64	1055	0.061	0.048	0.077
HLA-B	B61	8	1055	0.008	0.004	0.015
HLA-B	B62	90	1055	0.085	0.07	0.103
HLA-B	B63	1	1055	0.001	0	0.005
HLA-B	B64	12	1055	0.011	0.006	0.019
HLA-B	B65	30	1055	0.028	0.02	0.04
HLA-B	B7	163	1055	0.155	0.134	0.178
HLA-B	B71	6	1055	0.006	0.003	0.013
HLA-B	B72	1	1055	0.001	0	0.005
HLA-B	B8	140	1055	0.133	0.114	0.155
HLA-C	Cw1	42	966	0.043	0.032	0.058
HLA-C	Cw10	79	966	0.082	0.066	0.101
HLA-C	Cw2	38	966	0.039	0.029	0.053
HLA-C	Cw4	90	966	0.093	0.076	0.113
HLA-C	Cw5	126	966	0.13	0.11	0.153
HLA-C	Cw6	100	966	0.104	0.086	0.125
HLA-C	Cw7	369	966	0.382	0.352	0.413
HLA-C	Cw8	45	966	0.047	0.035	0.062
HLA-C	Cw9	77	966	0.08	0.065	0.099
HLA-DQA	0101	156	1078	0.145	0.125	0.167
HLA-DQA	0102	191	1078	0.177	0.155	0.201
HLA-DQA	0103	50	1078	0.046	0.035	0.06
HLA-DQA	0201	158	1078	0.147	0.127	0.169
HLA-DQA	0301	250	1078	0.232	0.208	0.258
HLA-DQA	0401	26	1078	0.024	0.016	0.035
HLA-DQA	0501	247	1078	0.229	0.205	0.255
HLA-DQB	DQ2	167	979	0.171	0.149	0.196
HLA-DQB	DQ4	25	979	0.026	0.018	0.038
HLA-DQB	DQ5	174	979	0.178	0.155	0.203
HLA-DQB	DQ6	247	979	0.252	0.226	0.28
HLA-DQB	DQ7	173	979	0.177	0.154	0.202
HLA-DQB	DQ8	125	979	0.128	0.109	0.15

Locus	Allele	Allele count in controls	Total allele count in controls	Frequency in controls	95% CI lower boundary	95% CI upper boundary
HLA-DQB	DQ9	68	979	0.069	0.055	0.087
HLA-DRB	DR1	93	912	0.102	0.084	0.123
HLA-DRB	DR10	2	912	0.002	0.001	0.008
HLA-DRB	DR103	6	912	0.007	0.003	0.015
HLA-DRB	DR11	25	912	0.027	0.018	0.04
HLA-DRB	DR12	18	912	0.02	0.013	0.031
HLA-DRB	DR13	96	912	0.105	0.087	0.127
HLA-DRB	DR14	21	912	0.023	0.015	0.035
HLA-DRB	DR15	162	912	0.178	0.155	0.204
HLA-DRB	DR16	9	912	0.01	0.005	0.019
HLA-DRB	DR17	140	912	0.154	0.132	0.179
HLA-DRB	DR4	144	912	0.158	0.136	0.183
HLA-DRB	DR7	156	912	0.171	0.148	0.197
HLA-DRB	DR8	24	912	0.026	0.017	0.038
HLA-DRB	DR9	16	912	0.018	0.011	0.029

Appendix C: Association Analysis Using Known Variants in SLI Candidate Genes

Associations with q≤0.05 are shown.

Chr	Position (hg19)	SNP	p-value	Reference	Variant	Variant Frequency in SLIC probands	Variant Frequency in ALSPAC	Gene
7	147642038	rs2707581	4.24E-75	T	C	0.0714	0.606124	*CNTNAP2*
7	147645164	rs2204923	6.97E-74	A	C	0.075	0.606124	*CNTNAP2*
7	147646762	rs2710088	7.08E-56	A	G	0.0714	0.521536	*CNTNAP2*
7	147641720	rs2707580	7.97E-55	T	G	0.0742	0.522055	*CNTNAP2*
7	147563860		2.45E-32	G	T	0.1179	0.002076	*CNTNAP2*
7	147645765	rs13225138	8.39E-27	G	A	0.5711	0.255838	*CNTNAP2*
7	147645706	rs13242781	3.65E-26	T	G	0.5693	0.255838	*CNTNAP2*
7	147643987	rs1922884	1.12E-24	G	T	0.5571	0.255838	*CNTNAP2*
7	147641383	rs2707579	2.55E-23	A	G	0.6633	0.896212	*CNTNAP2*
7	147643119	rs2952670	3.48E-22	T	C	0.6718	0.896212	*CNTNAP2*
7	147643602	rs1922887	2.80E-14	A	C	0.0714	0.252724	*CNTNAP2*
7	147647187	rs1882687	2.85E-14	C	A	0.0714	0.252465	*CNTNAP2*
7	147643923	rs1922885	2.85E-14	G	A	0.0714	0.252465	*CNTNAP2*
7	147646687	rs2710089	4.35E-14	G	A	0.0714	0.251168	*CNTNAP2*
7	147643746	rs1922886	4.35E-14	G	A	0.0714	0.251168	*CNTNAP2*
7	147642166	rs2708276	4.40E-14	C	T	0.0714	0.250908	*CNTNAP2*
7	147643046	rs2952824	2.32E-13	G	A	0.0748	0.250389	*CNTNAP2*
7	147519901		2.50E-10	G	A	0.0321	0.000259	*CNTNAP2*
7	147587826	rs186992976	4.59E-08	A	G	0.025	0.000259	*CNTNAP2*
7	147624760	rs144050752	5.55E-08	A	C	0.0285	0.000778	*CNTNAP2*
7	147595750	rs148180430	1.35E-07	T	C	0.0428	0.004411	*CNTNAP2*
7	147597599	rs188212772	1.95E-07	G	A	0.0249	0.000519	*CNTNAP2*
7	147628065	rs150921414	2.83E-06	A	G	0.0392	0.005189	*CNTNAP2*
7	147641341	rs13230414	4.50E-06	C	T	0.3852	0.255838	*CNTNAP2*
16	84458675	rs146957637	1.38E-05	G	A	0.025	0.001816	*ATP2C2*
7	147645279	rs2178770	1.51E-05	G	A	0.075	0.167359	*CNTNAP2*
7	147549011	rs1723441	1.76E-05	G	T	0.5678	0.695381	*CNTNAP2*
4	47892115	rs192315691	2.05E-05	G	A	0.0321	0.004152	*NFXL1*
7	147519223	rs851658	2.58E-05	C	A	0.0178	0.000519	*CNTNAP2*
16	84471642	rs429790	4.93E-05	A	C	0.5162	0.39206	*ATP2C2*
16	81687864	rs11863265	5.11E-05	A	G	0.079	0.028282	*CMIP*
7	147590210		5.98E-05	A	G	0.022	0.001557	*CNTNAP2*
16	84472552	rs247906	6.82E-05	A	G	0.5172	0.395693	*ATP2C2*
7	147527567	rs851714	9.76E-05	T	G	0.0142	0.000259	*CNTNAP2*
16	81659665		9.76E-05	G	A	0.0138	0.000259	*CMIP*
7	147639426		9.76E-05	A	G	0.0142	0.000259	*CNTNAP2*

Chr	Position (hg19)	SNP	p-value	Reference	Variant	Variant Frequency in SLIC probands	Variant Frequency in ALSPAC	Gene
16	84473985	rs4782626	0.000105	G	A	0.3265	0.443176	*ATP2C2*
16	84473192	rs931077	0.000111	A	G	0.3387	0.457706	*ATP2C2*
16	84473942	rs4782625	0.000156	T	C	0.3067	0.422159	*ATP2C2*
16	84454267	rs11647343	0.000172	A	C	0.2539	0.362999	*ATP2C2*
16	84454581	rs2875890	0.00019	C	T	0.3574	0.472237	*ATP2C2*
16	81666955		0.000214	T	G	0.0285	0.00467	*CMIP*
7	147562582	rs188640067	0.000214	T	G	0.0285	0.00467	*CNTNAP2*
16	84467915	rs34284536	0.000232	G	A	0.4995	0.38713	*ATP2C2*
16	84455961	rs2288573	0.00025	A	G	0.3716	0.484432	*ATP2C2*
16	84463915	rs247839	0.000267	G	A	0.5249	0.412299	*ATP2C2*
16	81673683		0.000277	G	A	0.0142	0.000519	*CMIP*
7	147648847	rs2246605	0.000319	A	C	0.8282	0.90192	*CNTNAP2*
16	84458648	rs8051403	0.000348	A	G	0.2535	0.358588	*ATP2C2*
16	84482559	rs2278299	0.000353	G	A	0.0503	0.114946	*ATP2C2*
16	84453850	rs66715009	0.000398	A	C	0.2216	0.322522	*ATP2C2*
16	81675733	rs7205166	0.000415	G	A	0.1219	0.062273	*CMIP*
16	84466062	rs7193536	0.000507	G	T	0.2437	0.342761	*ATP2C2*
16	84473768	rs11149654	0.000527	T	C	0.3141	0.420083	*ATP2C2*
16	84455249	rs11866442	0.000612	C	T	0.0142	0.000778	*ATP2C2*
16	84454763	rs17741623	0.000614	T	C	0.2178	0.3137	*ATP2C2*
16	84472972	rs247905	0.00064	G	A	0.4987	0.395174	*ATP2C2*
16	84446676	rs247812	0.000683	C	T	0.1494	0.085106	*ATP2C2*
7	147647777	rs1882688	0.000743	C	T	0.0931	0.167618	*CNTNAP2*
16	84465150	rs4782960	0.000813	A	G	0.3829	0.48547	*ATP2C2*
16	84460578	rs16973771	0.000868	T	C	0.3234	0.42657	*ATP2C2*
16	84458049	rs11641976	0.00091	G	A	0.3379	0.441619	*ATP2C2*
16	81679151	rs4889356	0.000954	A	G	0.5468	0.443435	*CMIP*
16	84454643	rs57454078	0.000994	T	G	0.2216	0.315257	*ATP2C2*
16	81675947	rs78733067	0.001018	C	T	0.1181	0.062273	*CMIP*
16	84447057	rs247811	0.001058	A	G	0.1465	0.085106	*ATP2C2*
16	84482012	rs34189707	0.001096	T	C	0.1184	0.063051	*ATP2C2*
16	84473632	rs11867035	0.001105	G	A	0.3276	0.427608	*ATP2C2*
16	84473874	rs11149656	0.001119	G	A	0.2884	0.38739	*ATP2C2*
4	47892102		0.00116	A	G	0.0142	0.001038	*NFXL1*
7	147520197		0.001169	T	C	0.0107	0.000259	*CNTNAP2*
16	81686864		0.001169	C	T	0.0107	0.000259	*CMIP*
16	81683368	rs116413284	0.001169	C	T	0.0107	0.000259	*CMIP*
7	147587819	rs77935737	0.001169	C	T	0.0107	0.000259	*CNTNAP2*
16	84454734	rs56964733	0.001238	C	T	0.2227	0.313181	*ATP2C2*
16	84468068	rs60535774	0.001274	A	G	0.0857	0.153866	*ATP2C2*
7	147533190	rs149042465	0.001693	A	G	0.0391	0.012455	*CNTNAP2*
16	84446423	rs79934118	0.001724	C	T	0.0143	0.05371	*ATP2C2*
16	81673241	rs76679721	0.00185	T	C	0.0749	0.035029	*CMIP*

Chr	Position (hg19)	SNP	p-value	Reference	Variant	Variant Frequency in SLIC probands	Variant Frequency in ALSPAC	Gene
16	84456705	rs185063	0.001974	C	T	0.6817	0.587961	*ATP2C2*
16	84454321	rs11642620	0.001975	T	C	0.2249	0.313181	*ATP2C2*
16	84454903	rs11861710	0.001977	T	C	0.0142	0.001297	*ATP2C2*
16	84454712	rs61576794	0.001982	G	A	0.2252	0.312143	*ATP2C2*
16	84473478	rs11866988	0.001999	G	A	0.3201	0.416191	*ATP2C2*
16	84466314	rs11646863	0.002342	A	G	0.0895	0.154645	*ATP2C2*
16	84484992	rs3826149	0.002357	T	C	0.0535	0.109237	*ATP2C2*
16	84462354	rs17814858	0.002524	A	G	0.2071	0.290088	*ATP2C2*
7	147533720	rs117645742	0.002869	A	G	0.0428	0.015568	*CNTNAP2*
16	84455734	rs8044383	0.003121	C	T	0.0142	0.001557	*ATP2C2*
16	84450519	rs182964529	0.003121	G	A	0.0142	0.001557	*ATP2C2*
7	147522117	rs142652606	0.00319	G	A	0.0249	0.005968	*CNTNAP2*
16	84463909	rs2875891	0.003225	C	T	0.2941	0.381422	*ATP2C2*
16	81677448	rs7201632	0.003475	T	C	0.3788	0.470161	*CMIP*

Appendix D: A Preliminary GWAS for the Robinson Crusoe Island Cohort

This GWAS was carried out using the SNPs genotyped for the linkage study, as well as eight of the exome sequencing variants (only markers from autosomal chromosomes could be used with MQLS). It was done using a tool which employed the MQLS method (downloaded from http://www.sph.umich.edu/csg/liang/MQLS), but, at the time of this analysis, the pedigree structure was not as accurate as it was when the analyses reported in Chapter 5 were done. The results are given in Figure D1 and Table D1. In this analysis seven SNPs were above the "expected" line in the Q-Q plot. One of these SNPs fell within *ZNF423*, and this was what first drew my attention to this gene.

Table D1: Top SNPS in the Preliminary GWAS for the Robinson Crusoe Island Cohort.

SNP	Gene	p-value
rs751266	*SCFD2*	0.0000302266
NM_152995/rs144169475	*NFXL1*	0.0000448322
rs1875084	-	0.000103334
rs1861659	*ZNF423*	0.000108952
rs757528	*ATP8B3*	0.000502759
rs1559510	-	0.000593829
rs1906255	-	0.000629051

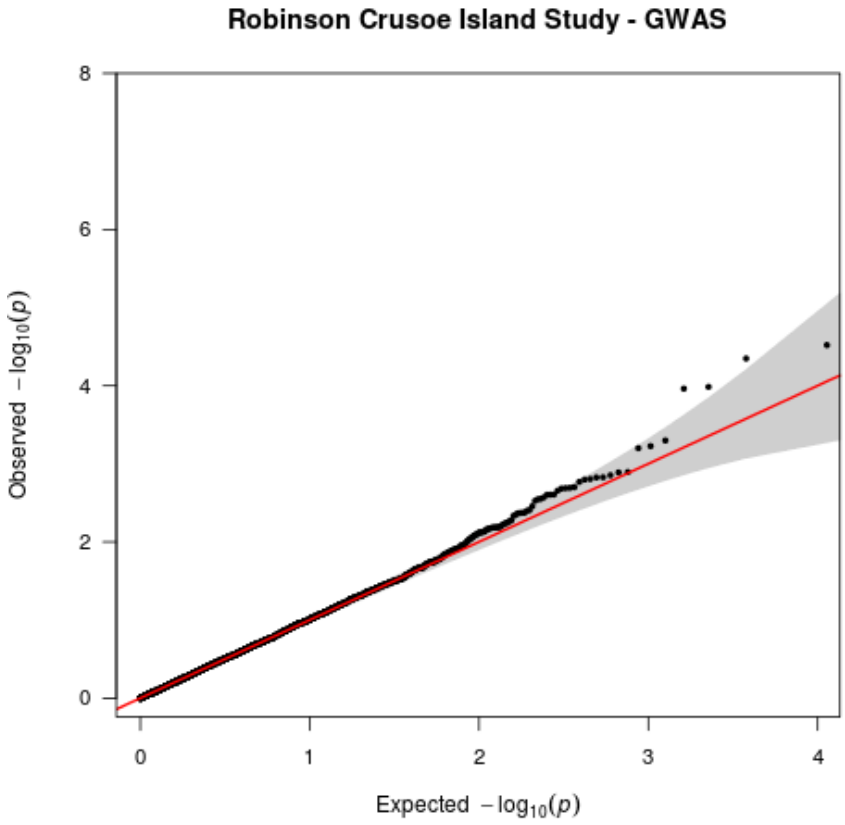


Figure D1: A Preliminary GWAS for the Robinson Crusoe Island Cohort, Manhattan and Q-Q plots (with 95% confidence intervals in grey).

Appendix E: Publications Arising from this Work

Part of the work described in this thesis has been published:

R. Nudel, N. H. Simpson, G. Baird, A. O'Hare, G. Conti-Ramsden, P. F. Bolton, E.R. Hennessy, The SLI Consortium, S. M. Ring, G. Davey Smith, C. Francks, S. Paracchini, A. P. Monaco, S. E. Fisher, D. F. Newbury. Genome-wide association analyses of child genotype effects and parent-of-origin effects in specific language impairment. Genes, Brain and Behav. 2014 13(4):418–429.

R. Nudel, N. H. Simpson, G. Baird, A. O'Hare, G. Conti-Ramsden, P. F. Bolton, E.R. Hennessy, The SLI Consortium, A. P. Monaco, J. C. Knight, B. Winney, S. E. Fisher, D. F. Newbury. Associations of HLA alleles with specific language impairment. J Neurodev Disord. 2014 6(1):1.

Part of the work from Chapters 5 and 6 was included in a manuscript recently accepted for publication:

P. Villanueva*, **R. Nudel***, A. Hoischen, M. A. Fernández, N. H. Simpson, C. Gilissen, R. H. Reader, L. Jara, M. M. Echeverry, C. Francks, G. Baird, G. Conti-Ramsden, A. O'Hare, P. F. Bolton, E. R. Hennessy, The SLI Consortium, H. Palomino, L. Carvajal-Carmona, J. A. Veltman, J.-B. Cazier, Z. De Barbieri, S. E. Fisher, D. F. Newbury. Exome sequencing in isolated population indicates NFXL1 variants confer a risk for Specific Language Impairment. Accepted by PLoS Genetics.

* Denotes joint first authorship.

Bibliography

ABECASIS, G. R., CARDON, L. R. & COOKSON, W. O. 2000. A general test of association for quantitative traits in nuclear families. *Am J Hum Genet,* 66**,** 279-92.

ABECASIS, G. R., CHERNY, S. S., COOKSON, W. O. & CARDON, L. R. 2002. Merlin--rapid analysis of dense genetic maps using sparse gene flow trees. *Nat Genet,* 30**,** 97-101.

ABOU JAMRA, R., WOHLFART, S., ZWEIER, M., UEBE, S., PRIEBE, L., EKICI, A., GIESEBRECHT, S., ABBOUD, A., AL KHATEEB, M. A., FAKHER, M., HAMDAN, S., ISMAEL, A., MUHAMMAD, S., NOTHEN, M. M., SCHUMACHER, J. & REIS, A. 2011. Homozygosity mapping in 64 Syrian consanguineous families with non-specific intellectual disability reveals 11 novel loci and high heterogeneity. *Eur J Hum Genet,* 19**,** 1161-6.

ABRAHAMS, B. S. & GESCHWIND, D. H. 2008. Advances in autism genetics: on the threshold of a new neurobiology. *Nat Rev Genet,* 9**,** 341-55.

ABRAHAMS, B. S., TENTLER, D., PEREDERIY, J. V., OLDHAM, M. C., COPPOLA, G. & GESCHWIND, D. H. 2007. Genome-wide analyses of human perisylvian cerebral cortical patterning. *Proc Natl Acad Sci U S A,* 104**,** 17849-54.

ADAMS, A. M. & GATHERCOLE, S. E. 2000. Limitations in working memory: implications for language development. *Int J Lang Commun Disord,* 35**,** 95-116.

ADZHUBEI, I. A., SCHMIDT, S., PESHKIN, L., RAMENSKY, V. E., GERASIMOVA, A., BORK, P., KONDRASHOV, A. S. & SUNYAEV, S. R. 2010. A method and server for predicting damaging missense mutations. *Nat Methods,* 7**,** 248-9.

AINSWORTH, H. F., UNWIN, J., JAMISON, D. L. & CORDELL, H. J. 2011. Investigation of maternal effects, maternal-fetal interactions and parent-of-origin effects (imprinting), using mothers and their offspring. *Genet Epidemiol,* 35**,** 19-45.

ALARCON, M., ABRAHAMS, B. S., STONE, J. L., DUVALL, J. A., PEREDERIY, J. V., BOMAR, J. M., SEBAT, J., WIGLER, M., MARTIN, C. L., LEDBETTER, D. H., NELSON, S. F., CANTOR, R. M. & GESCHWIND, D. H. 2008. Linkage, association, and gene-expression analyses identify CNTNAP2 as an autism-susceptibility gene. *Am J Hum Genet,* 82**,** 150-9.

ALCARAZ, W. A., GOLD, D. A., RAPONI, E., GENT, P. M., CONCEPCION, D. & HAMILTON, B. A. 2006. Zfp423 controls proliferation and differentiation of neural precursors in cerebellar vermis formation. *Proc Natl Acad Sci U S A,* 103**,** 19424-9.

AMOS, C. I. 1994. Robust variance-components approach for assessing genetic linkage in pedigrees. *Am J Hum Genet,* 54**,** 535-43.

ANDERSON, C. A., PETTERSSON, F. H., CLARKE, G. M., CARDON, L. R., MORRIS, A. P. & ZONDERVAN, K. T. 2010. Data quality control in genetic case-control association studies. *Nat Protoc,* 5**,** 1564-73.

ANNEY, R., KLEI, L., PINTO, D., REGAN, R., CONROY, J., MAGALHAES, T. R., CORREIA, C., ABRAHAMS, B. S., SYKES, N., PAGNAMENTA, A. T., ALMEIDA, J., BACCHELLI, E., BAILEY, A. J., BAIRD, G., BATTAGLIA, A., BERNEY, T., BOLSHAKOVA, N., BOLTE, S., BOLTON, P. F., BOURGERON, T., BRENNAN, S., BRIAN, J., CARSON, A. R., CASALLO, G., CASEY, J., CHU, S. H., COCHRANE, L., CORSELLO, C., CRAWFORD, E. L., CROSSETT, A., DAWSON, G., DE JONGE, M., DELORME, R., DRMIC, I., DUKETIS, E., DUQUE, F., ESTES, A., FARRAR, P., FERNANDEZ, B. A., FOLSTEIN, S. E., FOMBONNE, E., FREITAG, C. M., GILBERT, J., GILLBERG, C., GLESSNER, J. T., GOLDBERG, J., GREEN, J., GUTER, S. J., HAKONARSON, H., HERON, E. A., HILL, M., HOLT, R., HOWE,

J. L., HUGHES, G., HUS, V., IGLIOZZI, R., KIM, C., KLAUCK, S. M., KOLEVZON, A., KORVATSKA, O., KUSTANOVICH, V., LAJONCHERE, C. M., LAMB, J. A., LASKAWIEC, M., LEBOYER, M., LE COUTEUR, A., LEVENTHAL, B. L., LIONEL, A. C., LIU, X. Q., LORD, C., LOTSPEICH, L., LUND, S. C., MAESTRINI, E., MAHONEY, W., MANTOULAN, C., MARSHALL, C. R., MCCONACHIE, H., MCDOUGLE, C. J., MCGRATH, J., MCMAHON, W. M., MELHEM, N. M., MERIKANGAS, A., MIGITA, O., MINSHEW, N. J., MIRZA, G. K., MUNSON, J., NELSON, S. F., NOAKES, C., NOOR, A., NYGREN, G., OLIVEIRA, G., PAPANIKOLAOU, K., PARR, J. R., PARRINI, B., PATON, T., PICKLES, A., PIVEN, J., POSEY, D. J., POUSTKA, A., POUSTKA, F.*, et al.* 2010. A genome-wide scan for common alleles affecting risk for autism. *Hum Mol Genet,* 19**,** 4072-82.

ANNEY, R. J., HAWI, Z., SHEEHAN, K., MULLIGAN, A., PINTO, C., BROOKES, K. J., XU, X., ZHOU, K., FRANKE, B., BUITELAAR, J., VERMEULEN, S. H., BANASCHEWSKI, T., SONUGA-BARKE, E., EBSTEIN, R., MANOR, I., MIRANDA, A., MULAS, F., OADES, R. D., ROEYERS, H., ROMMELSE, N., ROTHENBERGER, A., SERGEANT, J., STEINHAUSEN, H. C., TAYLOR, E., THOMPSON, M., ASHERSON, P., FARAONE, S. V. & GILL, M. 2008. Parent of origin effects in attention/deficit hyperactivity disorder (ADHD): analysis of data from the international multicenter ADHD genetics (IMAGE) program. *Am J Med Genet B Neuropsychiatr Genet,* 147B**,** 1495-500.

ARCHIBALD, L. M. & ALLOWAY, T. P. 2008. Comparing language profiles: children with specific language impairment and developmental coordination disorder. *Int J Lang Commun Disord,* 43**,** 165-80.

ARCHIBALD, L. M. & GATHERCOLE, S. E. 2006. Short-term and working memory in specific language impairment. *Int J Lang Commun Disord,* 41**,** 675-93.

ARKING, D. E., CUTLER, D. J., BRUNE, C. W., TESLOVICH, T. M., WEST, K., IKEDA, M., REA, A., GUY, M., LIN, S., COOK, E. H. & CHAKRAVARTI, A. 2008. A common genetic variant in the neurexin superfamily member CNTNAP2 increases familial risk of autism. *Am J Hum Genet,* 82**,** 160-4.

ASHWOOD, P. & VAN DE WATER, J. 2004. A review of autism and the immune response. *Clin Dev Immunol,* 11**,** 165-74.

BADDELEY, A. & WILSON, B. A. 1993. A developmental deficit in short-term phonological memory: implications for language and reading. *Memory,* 1**,** 65-78.

BADDELEY, A. D. & HITCH, G. J. 1974. Working Memory. *In:* BOWER, G. A. (ed.) *Recent Advances in Leaning and Motivation.* New York: Academic Press.

BAIRD, J., STEVENSON, J. C. & WILLIAMS, D. C. 2000. The evolution of ADHD: a disorder of communication? *Q Rev Biol,* 75**,** 17-35.

BAKKALOGLU, B., O'ROAK, B. J., LOUVI, A., GUPTA, A. R., ABELSON, J. F., MORGAN, T. M., CHAWARSKA, K., KLIN, A., ERCAN-SENCICEK, A. G., STILLMAN, A. A., TANRIOVER, G., ABRAHAMS, B. S., DUVALL, J. A., ROBBINS, E. M., GESCHWIND, D. H., BIEDERER, T., GUNEL, M., LIFTON, R. P. & STATE, M. W. 2008. Molecular cytogenetic analysis and resequencing of contactin associated protein-like 2 in autism spectrum disorders. *Am J Hum Genet,* 82**,** 165-73.

BALLIF, B. C., THEISEN, A., MCDONALD-MCGINN, D. M., ZACKAI, E. H., HERSH, J. H., BEJJANI, B. A. & SHAFFER, L. G. 2008. Identification of a previously unrecognized microdeletion syndrome of 16q11.2q12.2. *Clin Genet,* 74**,** 469-75.

BARBER, R. D., HARMER, D. W., COLEMAN, R. A. & CLARK, B. J. 2005. GAPDH as a housekeeping gene: analysis of GAPDH mRNA expression in a panel of 72 human tissues. *Physiol Genomics,* 21**,** 389-95.

BARRY, G. 2014. Integrating the roles of long and small non-coding RNA in brain function and disease. *Mol Psychiatry,* 19**,** 410-6.

BARRY, J. G., YASIN, I. & BISHOP, D. V. 2007. Heritable risk factors associated with language impairments. *Genes Brain Behav,* 6**,** 66-76.

BARTLETT, C. W., FLAX, J. F., FERMANO, Z., HARE, A., HOU, L., PETRILL, S. A., BUYSKE, S. & BRZUSTOWICZ, L. M. 2012. Gene x gene interaction in shared etiology of autism and specific language impairment. *Biol Psychiatry,* 72**,** 692-9.

BARTLETT, C. W., FLAX, J. F., LOGUE, M. W., SMITH, B. J., VIELAND, V. J., TALLAL, P. & BRZUSTOWICZ, L. M. 2004. Examination of potential overlap in autism and language loci on chromosomes 2, 7, and 13 in two independent samples ascertained for specific language impairment. *Hum Hered,* 57**,** 10-20.

BARTLETT, C. W., FLAX, J. F., LOGUE, M. W., VIELAND, V. J., BASSETT, A. S., TALLAL, P. & BRZUSTOWICZ, L. M. 2002. A major susceptibility locus for specific language impairment is located on 13q21. *Am J Hum Genet,* 71**,** 45-55.

BARTLETT, C. W., HOU, L., FLAX, J. F., HARE, A., CHEONG, S. Y., FERMANO, Z., ZIMMERMAN-BIER, B., CARTWRIGHT, C., AZARO, M. A., BUYSKE, S. & BRZUSTOWICZ, L. M. 2014. A genome scan for loci shared by autism spectrum disorder and language impairment. *Am J Psychiatry,* 171**,** 72-81.

BAUMAN, M. L. & KEMPER, T. L. 2005. Neuroanatomic observations of the brain in autism: a review and future directions. *Int J Dev Neurosci,* 23**,** 183-7.

BEG, A. A., RUBEN, S. M., SCHEINMAN, R. I., HASKILL, S., ROSEN, C. A. & BALDWIN, A. S., JR. 1992. I kappa B interacts with the nuclear localization sequences of the subunits of NF-kappa B: a mechanism for cytoplasmic retention. *Genes Dev,* 6**,** 1899-913.

BENITO-CUADRADO, M. M., ESTEBA-CASTILLO, S., BOHM, P., CEJUDO-BOLIVAR, J. & PENA-CASANOVA, J. 2002. Semantic verbal fluency of animals: a normative and predictive study in a Spanish population. *J Clin Exp Neuropsychol,* 24**,** 1117-22.

BENJAMINI, Y. & HOCHBERG, Y. 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological),* 57**,** 289-300.

BISHOP, D. V. M. 1998. Development of the Children's Communication Checklist (CCC): a method for assessing qualitative aspects of communicative impairment in children. *J Child Psychol Psychiatry,* 39**,** 879-91.

BISHOP, D. V. M. 2000. Pragmatic language impairment: A correlate of SLI, a distinct subgroup, or part of the autistic continuum? . *In:* BISHOP, D. V. M. & LEONARD, L. B. (eds.) *Pragmatic language impairment: A correlate of SLI, a distinct subgroup, or part of the autistic continuum? .* UK: Psychology Press.

BISHOP, D. V. M. 2003. Autism and specific language impairment: categorical distinction or continuum? *Novartis Found Symp,* 251**,** 213-26; discussion 226-34, 281-97.

BISHOP, D. V. M. 2006. What Causes Specific Language Impairment in Children? *Curr Dir Psychol Sci,* 15**,** 217-221.

BISHOP, D. V. M. 2010. Overlaps between autism and language impairment: phenomimicry or shared etiology? *Behavior Genetics,* 40**,** 618-29.

BISHOP, D. V. M. & BAIRD, G. 2001. Parent and teacher report of pragmatic aspects of communication: use of the children's communication checklist in a clinical setting. *Dev Med Child Neurol,* 43**,** 809-18.

BISHOP, D. V. M., BISHOP, S. J., BRIGHT, P., JAMES, C., DELANEY, T. & TALLAL, P. 1999. Different origin of auditory and phonological processing problems in children with language impairment: evidence from a twin study. *J Speech Lang Hear Res,* 42**,** 155-68.

BISHOP, D. V. M. & HAYIOU-THOMAS, M. E. 2008. Heritability of specific language impairment depends on diagnostic criteria. *Genes Brain Behav,* 7**,** 365-72.

BISHOP, D. V. M., NORTH, T. & DONLAN, C. 1995. Genetic basis of specific language impairment: evidence from a twin study. *Dev Med Child Neurol,* 37**,** 56-71.

BISHOP, D. V. M., NORTH, T. & DONLAN, C. 1996. Nonword repetition as a behavioural marker for inherited language impairment: evidence from a twin study. *J Child Psychol Psychiatry,* 37**,** 391-403.

BISHOP, D. V. M. & SNOWLING, M. J. 2004. Developmental dyslexia and specific language impairment: same or different? *Psychol Bull,* 130**,** 858-86.

BOUWMEESTER, T., BAUCH, A., RUFFNER, H., ANGRAND, P. O., BERGAMINI, G., CROUGHTON, K., CRUCIAT, C., EBERHARD, D., GAGNEUR, J., GHIDELLI, S., HOPF, C., HUHSE, B., MANGANO, R., MICHON, A. M., SCHIRLE, M., SCHLEGL, J., SCHWAB, M., STEIN, M. A., BAUER, A., CASARI, G., DREWES, G., GAVIN, A. C., JACKSON, D. B., JOBERTY, G., NEUBAUER, G., RICK, J., KUSTER, B. & SUPERTI-FURGA, G. 2004. A physical and functional map of the human TNF-alpha/NF-kappa B signal transduction pathway. *Nat Cell Biol,* 6**,** 97-105.

BOYD, A., GOLDING, J., MACLEOD, J., LAWLOR, D. A., FRASER, A., HENDERSON, J., MOLLOY, L., NESS, A., RING, S. & DAVEY SMITH, G. 2013. Cohort Profile: the 'children of the 90s'--the index offspring of the Avon Longitudinal Study of Parents and Children. *Int J Epidemiol,* 42**,** 111-27.

BRADFORD, Y., HAINES, J., HUTCHESON, H., GARDINER, M., BRAUN, T., SHEFFIELD, V., CASSAVANT, T., HUANG, W., WANG, K., VIELAND, V., FOLSTEIN, S., SANTANGELO, S. & PIVEN, J. 2001. Incorporating language phenotypes strengthens evidence of linkage to autism. *Am J Med Genet,* 105**,** 539-47.

BRAMEIER, M., KRINGS, A. & MACCALLUM, R. M. 2007. NucPred--predicting nuclear localization of proteins. *Bioinformatics,* 23**,** 1159-60.

BRUNET, A., SWEENEY, L. B., STURGILL, J. F., CHUA, K. F., GREER, P. L., LIN, Y., TRAN, H., ROSS, S. E., MOSTOSLAVSKY, R., COHEN, H. Y., HU, L. S., CHENG, H. L., JEDRYCHOWSKI, M. P., GYGI, S. P., SINCLAIR, D. A., ALT, F. W. & GREENBERG, M. E. 2004. Stress-dependent regulation of FOXO transcription factors by the SIRT1 deacetylase. *Science,* 303**,** 2011-5.

BURGEMEISTER, B. B., BLUM, L. H. & LORGE, I. 1972. *The Columbia Mental Maturity Scale,* San Antonio, Psychological Corporation.

BUXBAUM, J. D. 2009. Multiple rare variants in the etiology of autism spectrum disorders. *Dialogues Clin Neurosci,* 11**,** 35-43.

CARADEC, J., SIRAB, N., KEUMEUGNI, C., MOUTEREAU, S., CHIMINGQI, M., MATAR, C., REVAUD, D., BAH, M., MANIVET, P., CONTI, M. & LORIC, S. 2010. 'Desperate house genes': the dramatic example of hypoxia. *Br J Cancer,* 102**,** 1037-43.

CARDON, L. R. & BELL, J. I. 2001. Association study designs for complex diseases. *Nat Rev Genet,* 2**,** 91-9.

CARLSON, J. M., HECKERMAN, D. & SHANI, G. 2009. Estimating false discovery rates for contingency tables. *Microsoft Research technical report, MSR-TR-2009-53*.

CARROW-WOOLFOLK, E. 1999. *Comprehensive assessment of spoken language (CASL),* Circle Pines, MN, American Guidance Service.

CERONI, F., SIMPSON, N. H., FRANCKS, C., BAIRD, G., CONTI-RAMSDEN, G., CLARK, A., BOLTON, P. F., HENNESSY, E. R., DONNELLY, P., BENTLEY, D. R., MARTIN, H., PARR, J., PAGNAMENTA, A. T., MAESTRINI, E., BACCHELLI, E., FISHER, S. E. & NEWBURY, D. F. 2014. Homozygous microdeletion of exon 5 in ZNF277 in a girl with specific language impairment. *Eur J Hum Genet,* 22**,** 1165–1171.

CHAERKADY, R., LETZEN, B., RENUSE, S., SAHASRABUDDHE, N. A., KUMAR, P., ALL, A. H., THAKOR, N. V., DELANGHE, B., GEARHART, J. D., PANDEY, A. & KERR, C. L. 2011. Quantitative temporal proteomic analysis of human embryonic stem cell differentiation into oligodendrocyte progenitor cells. *Proteomics,* 11**,** 4007-20.

CHAMBERLAIN, S. J. & LALANDE, M. 2010. Neurodevelopmental disorders involving genomic imprinting at human chromosome 15q11-q13. *Neurobiol Dis,* 39**,** 13-20.

CHOTAI, J. 1984. On the lod score method in linkage analysis. *Ann Hum Genet,* 48**,** 359-78.

CINGOLANI, P., PLATTS, A., WANG LE, L., COON, M., NGUYEN, T., WANG, L., LAND, S. J., LU, X. & RUDEN, D. M. 2012. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. *Fly (Austin),* 6**,** 80-92.

COHEN, N. J., BARWICK, M. A., HORODEZKY, N. B., VALLANCE, D. D. & IM, N. 1998. Language, achievement, and cognitive processing in psychiatrically disturbed children with previously identified and unsuspected language impairments. *J Child Psychol Psychiatry,* 39**,** 865-77.

COLHOUN, H. M., MCKEIGUE, P. M. & DAVEY SMITH, G. 2003. Problems of reporting genetic associations with complex outcomes. *Lancet,* 361**,** 865-72.

CONNOLLY, A. M., CHEZ, M. G., PESTRONK, A., ARNOLD, S. T., MEHTA, S. & DEUEL, R. K. 1999. Serum autoantibodies to brain in Landau-Kleffner variant, autism, and other neurologic disorders. *J Pediatr,* 134**,** 607-13.

CONTI-RAMSDEN, G., BOTTING, N. & FARAGHER, B. 2001. Psycholinguistic markers for specific language impairment (SLI). *J Child Psychol Psychiatry,* 42**,** 741-8.

COOK, E. H., JR. & SCHERER, S. W. 2008. Copy-number variations associated with neuropsychiatric conditions. *Nature,* 455**,** 919-23.

COPE, N., HAROLD, D., HILL, G., MOSKVINA, V., STEVENSON, J., HOLMANS, P., OWEN, M. J., O'DONOVAN, M. C. & WILLIAMS, J. 2005. Strong evidence that KIAA0319 on chromosome 6p is a susceptibility gene for developmental dyslexia. *Am J Hum Genet,* 76**,** 581-91.

CURRAN, S., BOLTON, P., ROZSNYAI, K., CHIOCCHETTI, A., KLAUCK, S. M., DUKETIS, E., POUSTKA, F., SCHLITT, S., FREITAG, C. M., LEE, I., MUGLIA, P., POOT, M., STAAL, W., DE JONGE, M. V., OPHOFF, R. A., LEWIS, C., SKUSE, D., MANDY, W., VASSOS, E., FOSSDAL, R., MAGNUSSON, P., HREIDARSSON, S., SAEMUNDSEN, E., STEFANSSON, H., STEFANSSON, K. & COLLIER, D. 2011. No association between a common single nucleotide polymorphism, rs4141463, in the MACROD2 gene and autism spectrum disorder. *Am J Med Genet B Neuropsychiatr Genet,* 156B**,** 633-9.

DALTON, P., DEACON, R., BLAMIRE, A., PIKE, M., MCKINLAY, I., STEIN, J., STYLES, P. & VINCENT, A. 2003. Maternal neuronal antibodies associated with autism and a language disorder. *Ann Neurol,* 53**,** 533-7.

DARNELL, J. E., JR. 1997. STATs and gene regulation. *Science,* 277**,** 1630-5.

DAWN TEARE, M. & BARRETT, J. H. 2005. Genetic linkage studies. *Lancet,* 366**,** 1036-44.

DE BARBIERI, Z., MAGGIOLO L & Y, A. 1999. Trastornos de la comunicación oral en niños que asisten a control de salud en un consultorio de atención primaria. *Rev. Chil. Pediatr.,* 70**,** 36-40.

DE RENZI, E. & VIGNOLO, L. 1962. The Token Test: a sensitive tests to detect receptive disturbances in aphasics. *Brain,* 85**,** 665-678.

DE SMET, H. J., PAQUIER, P., VERHOEVEN, J. & MARIEN, P. 2013. The cerebellum: its role in language and related cognitive and affective functions. *Brain Lang,* 127**,** 334-42.

DEFRIES, J. C. & FULKER, D. W. 1985. Multiple regression analysis of twin data. *Behav Genet,* 15**,** 467-73.

DEFRIES, J. C. & FULKER, D. W. 1988. Multiple regression analysis of twin data: etiology of deviant scores versus individual differences. *Acta Genet Med Gemellol (Roma),* 37**,** 205-16.

DILTHEY, A. T., MOUTSIANAS, L., LESLIE, S. & MCVEAN, G. 2011. HLA*IMP--an integrated framework for imputing classical HLA alleles from SNP genotypes. *Bioinformatics,* 27**,** 968-72.

DISIMONI, F. G. 1978. *The Token Test for Children,* Boston, Teaching Resources.

DOVER, C. J. & LE COUTEUR, A. 2007. How to diagnose autism. *Arch Dis Child,* 92**,** 540-5.

DUAN, S., HUANG, R. S., ZHANG, W., BLEIBEL, W. K., ROE, C. A., CLARK, T. A., CHEN, T. X., SCHWEITZER, A. C., BLUME, J. E., COX, N. J. & DOLAN, M. E. 2008. Genetic architecture of transcript-level variation in humans. *Am J Hum Genet,* 82**,** 1101-13.

DUNN, L. M. & DUNN, L. M. 1981. *Peabody Picture Vocabulary Test, Revised,* Circle Pines, American Guidance Service.

DUNN, L. M., DUNN, L. M. & WILLIAMS, K. T. 1997. *Peabody Picture Vocabulary Test III,* Circle Pines, American Guidance Service.

ELIA, J., GAI, X., XIE, H. M., PERIN, J. C., GEIGER, E., GLESSNER, J. T., D'ARCY, M., DEBERARDINIS, R., FRACKELTON, E., KIM, C., LANTIERI, F., MUGANGA, B. M., WANG, L., TAKEDA, T., RAPPAPORT, E. F., GRANT, S. F., BERRETTINI, W., DEVOTO, M., SHAIKH, T. H., HAKONARSON, H. & WHITE, P. S. 2009. Rare structural variants found in attention-deficit hyperactivity disorder are preferentially associated with neurodevelopmental genes. *Mol Psychiatry,* 15**,** 637-646.

EWENS, W. J., LI, M. & SPIELMAN, R. S. 2008. A review of family-based tests for linkage disequilibrium between a quantitative trait and a genetic marker. *PLoS Genet,* 4**,** e1000180.

FABBRO, F., MORETTI, R. & BAVA, A. 2000. Language impairments in patients with cerebellar lesions. *Journal of Neurolinguistics,* 13**,** 173-188.

FAIRFAX, B. P., MAKINO, S., RADHAKRISHNAN, J., PLANT, K., LESLIE, S., DILTHEY, A., ELLIS, P., LANGFORD, C., VANNBERG, F. O. & KNIGHT, J. C. 2012. Genetics of gene expression in primary immune cells identifies cell type-specific master regulators and roles of HLA alleles. *Nat Genet,* 44**,** 502-10.

FALCARO, M., PICKLES, A., NEWBURY, D. F., ADDIS, L., BANFIELD, E., FISHER, S. E., MONACO, A. P., SIMKIN, Z. & CONTI-RAMSDEN, G. 2008. Genetic and phenotypic effects of phonological short-term memory and grammatical morphology in specific language impairment. *Genes Brain Behav,* 7**,** 393-402.

FANTES, J., REDEKER, B., BREEN, M., BOYLE, S., BROWN, J., FLETCHER, J., JONES, S., BICKMORE, W., FUKUSHIMA, Y., MANNENS, M. & ET AL. 1995. Aniridia-associated cytogenetic rearrangements suggest that a position effect may cause the mutant phenotype. *Hum Mol Genet,* 4**,** 415-22.

FARAONE, S. V., PERLIS, R. H., DOYLE, A. E., SMOLLER, J. W., GORALNICK, J. J., HOLMGREN, M. A. & SKLAR, P. 2005. Molecular genetics of attention-deficit/hyperactivity disorder. *Biol Psychiatry,* 57**,** 1313-23.

FERNANDO, M. M., STEVENS, C. R., WALSH, E. C., DE JAGER, P. L., GOYETTE, P., PLENGE, R. M., VYSE, T. J. & RIOUX, J. D. 2008. Defining the role of the MHC in autoimmunity: a review and pooled analysis. *PLoS Genet,* 4**,** e1000024.

FEUK, L., KALERVO, A., LIPSANEN-NYMAN, M., SKAUG, J., NAKABAYASHI, K., FINUCANE, B., HARTUNG, D., INNES, M., KEREM, B., NOWACZYK, M. J., RIVLIN, J., ROBERTS, W., SENMAN, L., SUMMERS, A., SZATMARI, P., WONG, V., VINCENT, J. B., ZEESMAN, S., OSBORNE, L. R., CARDY, J. O., KERE, J., SCHERER, S. W. & HANNULA-JOUPPI, K. 2006. Absence of a paternally inherited FOXP2 gene in developmental verbal dyspraxia. *Am J Hum Genet,* 79**,** 965-72.

FISHER, S. E., VARGHA-KHADEM, F., WATKINS, K. E., MONACO, A. P. & PEMBREY, M. E. 1998. Localisation of a gene implicated in a severe speech and language disorder. *Nat Genet,* 18**,** 168-70.

FLICEK, P., AMODE, M. R., BARRELL, D., BEAL, K., BRENT, S., CARVALHO-SILVA, D., CLAPHAM, P., COATES, G., FAIRLEY, S., FITZGERALD, S., GIL, L., GORDON, L., HENDRIX, M., HOURLIER, T., JOHNSON, N., KAHARI, A. K., KEEFE, D., KEENAN, S., KINSELLA, R., KOMOROWSKA, M., KOSCIELNY, G., KULESHA, E., LARSSON, P., LONGDEN, I., MCLAREN, W., MUFFATO, M., OVERDUIN, B., PIGNATELLI, M., PRITCHARD, B., RIAT, H. S., RITCHIE, G. R., RUFFIER, M., SCHUSTER, M., SOBRAL, D., TANG, Y. A., TAYLOR, K., TREVANION, S., VANDROVCOVA, J., WHITE, S., WILSON, M., WILDER, S. P., AKEN, B. L., BIRNEY, E., CUNNINGHAM, F., DUNHAM, I., DURBIN, R., FERNANDEZ-SUAREZ, X. M., HARROW, J., HERRERO, J., HUBBARD, T. J., PARKER, A., PROCTOR, G., SPUDICH, G., VOGEL, J., YATES, A., ZADISSA, A. & SEARLE, S. M. 2012. Ensembl 2012. *Nucleic Acids Res,* 40**,** D84-90.

FRADIN, D., CHESLACK-POSTAVA, K., LADD-ACOSTA, C., NEWSCHAFFER, C., CHAKRAVARTI, A., ARKING, D. E., FEINBERG, A. & FALLIN, M. D. 2010. Parent-of-origin effects in autism identified through genome-wide linkage analysis of 16,000 SNPs. *PLoS One,* 5**,** e12513.

FRANCKS, C., PARACCHINI, S., SMITH, S. D., RICHARDSON, A. J., SCERRI, T. S., CARDON, L. R., MARLOW, A. J., MACPHIE, I. L., WALTER, J., PENNINGTON, B. F., FISHER, S. E., OLSON, R. K., DEFRIES, J. C., STEIN, J. F. & MONACO, A. P. 2004. A 77-kilobase region of chromosome 6p22.2 is associated with dyslexia in families from the United Kingdom and from the United States. *Am J Hum Genet,* 75**,** 1046-58.

FRASER, A., MACDONALD-WALLIS, C., TILLING, K., BOYD, A., GOLDING, J., DAVEY SMITH, G., HENDERSON, J., MACLEOD, J., MOLLOY, L., NESS, A., RING, S., NELSON, S. M. & LAWLOR, D. A. 2013. Cohort Profile: the Avon Longitudinal Study of Parents and Children: ALSPAC mothers cohort. *Int J Epidemiol,* 42**,** 97-110.

FUJITA, E., TANABE, Y., SHIOTA, A., UEDA, M., SUWA, K., MOMOI, M. Y. & MOMOI, T. 2008. Ultrasonic vocalization impairment of Foxp2 (R552H) knockin mice related to speech-language disorder and abnormality of Purkinje cells. *Proc Natl Acad Sci U S A,* 105**,** 3117-22.

FULKER, D. W., CARDON, L. R., DEFRIES, J. C., KIMBERLING, W. J., PENNINGTON, B. F. & SMITH, S. D. 1991. Multiple Regression Analysis of Sib-Pair Data on Reading to Detect Quantitative Trait Loci. *In:* PENNINGTON, B. (ed.) *Reading Disabilities.* Springer Netherlands.

GAMAGE, T. H., MISCEO, D., FANNEMEL, M. & FRENGEN, E. 2013. A balanced de novo inv(7)(p14.3q22.3) disrupting PDE1C and ATXN7L1 in a 14-year old developmentally delayed boy. *Eur J Med Genet,* 56**,** 361-4.

GAMAZON, E. R., ZHANG, W., KONKASHBAEV, A., DUAN, S., KISTNER, E. O., NICOLAE, D. L., DOLAN, M. E. & COX, N. J. 2010. SCAN: SNP and copy number annotation. *Bioinformatics,* 26**,** 259-62.

GARCÍA-BELLIDO, P., BENÍTEZ-BURRACO, A., ROSELLÓ, M., MONFORT, S., MARTÍNEZ, F., OLTRA, S. & ORELLANA, C. 2009. *A case of Spanish language disorders with a rare genetic cause,* Madrid, UNED-Euphonia Ediciones.

GATHERCOLE, S. E. & BADDELEY, A. D. 1990. Phonological memory deficits in language disordered children: Is there a causal connection? *Journal of Memory and Language,* 29**,** 336-360.

GATHERCOLE, S. E., WILLIS, C. S., BADDELEY, A. D. & EMSLIE, H. 1994. The Children's Test of Nonword Repetition: a test of phonological working memory. *Memory,* 2**,** 103-27.

GAUB, S., GROSZER, M., FISHER, S. E. & EHRET, G. 2010. The structure of innate vocalizations in Foxp2-deficient mouse pups. *Genes Brain Behav,* 9**,** 390-401.

GEURTS, H. M. & EMBRECHTS, M. 2008. Language profiles in ASD, SLI, and ADHD. *J Autism Dev Disord,* 38**,** 1931-43.

GLASER, T., JEPEAL, L., EDWARDS, J. G., YOUNG, S. R., FAVOR, J. & MAAS, R. L. 1994. PAX6 gene dosage effect in a family with congenital cataracts, aniridia, anophthalmia and central nervous system defects. *Nat Genet,* 7**,** 463-71.

GLATT, S. J., STONE, W. S., NOSSOVA, N., LIEW, C. C., SEIDMAN, L. J. & TSUANG, M. T. 2011. Similarities and differences in peripheral blood gene-expression signatures of individuals with schizophrenia and their first-degree biological relatives. *Am J Med Genet B Neuropsychiatr Genet,* 156B**,** 869-87.

GOLDMAN, R. & FRISTOE, M. 2000. Goldman Fristoe Test of Articulation. *American Guidance Service*.

GOPNIK, M. 1990a. Feature-blind grammar and dysphasia. *Nature,* 344**,** 715.

GOPNIK, M. 1990b. Genetic basis of grammar defect. *Nature,* 347**,** 26.

GOPNIK, M. 1997. Language deficits and genetic factors. *Trends in Cognitive Sciences,* 1**,** 5-9.

GOPNIK, M. & CRAGO, M. B. 1991. Familial aggregation of a developmental language disorder. *Cognition,* 39**,** 1-50.

GOTTESMAN, II & GOULD, T. D. 2003. The endophenotype concept in psychiatry: etymology and strategic intentions. *Am J Psychiatry,* 160**,** 636-45.

GRIGORENKO, E. L., WOOD, F. B., MEYER, M. S., PAULS, J. E., HART, L. A. & PAULS, D. L. 2001. Linkage studies suggest a possible locus for developmental dyslexia on chromosome 1p. *Am J Med Genet,* 105**,** 120-9.

GROSZER, M., KEAYS, D. A., DEACON, R. M., DE BONO, J. P., PRASAD-MULCARE, S., GAUB, S., BAUM, M. G., FRENCH, C. A., NICOD, J., COVENTRY, J. A., ENARD, W., FRAY, M., BROWN, S. D., NOLAN, P. M., PAABO, S., CHANNON, K. M., COSTA, R. M., EILERS, J., EHRET, G., RAWLINS, J. N. & FISHER, S. E. 2008. Impaired synaptic plasticity and motor learning in mice with a point mutation implicated in human speech deficits. *Curr Biol,* 18**,** 354-62.

GUTTMAN, M. & RINN, J. L. 2012. Modular regulatory principles of large non-coding RNAs. *Nature,* 482**,** 339-46.

HABIB, M. 2000. The neurological basis of developmental dyslexia: an overview and working hypothesis. *Brain,* 123 Pt 12**,** 2373-99.

HAESLER, S., ROCHEFORT, C., GEORGI, B., LICZNERSKI, P., OSTEN, P. & SCHARFF, C. 2007. Incomplete and inaccurate vocal imitation after knockdown of FoxP2 in songbird basal ganglia nucleus Area X. *PLoS Biol,* 5**,** e321.

HAMMILL, D. D., BROWN, V. L., LARSEN, S. C. & WIEDERHOLT, J. L. 1987. *Test of adolescent language-2,* Austin, TX, Pro-Ed.

HAMMILL, D. D. & NEWCOMER, P. L. 1988. *Test of Language Development–2, Intermediate*, Pro-ed.

HASEMAN, J. K. & ELSTON, R. C. 1972. The investigation of linkage between a quantitative trait and a marker locus. *Behav Genet,* 2**,** 3-19.

HAWI, Z., KENT, L., HILL, M., ANNEY, R. J., BROOKES, K. J., BARRY, E., FRANKE, B., BANASCHEWSKI, T., BUITELAAR, J., EBSTEIN, R., MIRANDA, A., OADES, R. D., ROEYERS, H., ROTHENBERGER, A., SERGEANT, J., SONUGA-BARKE, E., STEINHAUSEN, H. C., FARAONE, S. V., ASHERSON, P. & GILL, M. 2010. ADHD and DAT1: further evidence of paternal over-transmission of risk alleles and haplotype. *Am J Med Genet B Neuropsychiatr Genet,* 153B**,** 97-102.

HAWI, Z., SEGURADO, R., CONROY, J., SHEEHAN, K., LOWE, N., KIRLEY, A., SHIELDS, D., FITZGERALD, M., GALLAGHER, L. & GILL, M. 2005. Preferential transmission of paternal alleles at risk genes in attention-deficit/hyperactivity disorder. *Am J Hum Genet,* 77**,** 958-65.

HILL, E. L. 2001. Non-specific nature of specific language impairment: a review of the literature with regard to concomitant motor impairments. *Int J Lang Commun Disord,* 36**,** 149-71.

HILLIS, A. E. 2007. Aphasia: progress in the last quarter of a century. *Neurology,* 69**,** 200-13.

HIRSCHHORN, J. N. & DALY, M. J. 2005. Genome-wide association studies for common diseases and complex traits. *Nat Rev Genet,* 6**,** 95-108.

HOGGART, C. J., CLARK, T. G., DE IORIO, M., WHITTAKER, J. C. & BALDING, D. J. 2008. Genome-wide significance for dense SNP and resequencing data. *Genet Epidemiol,* 32**,** 179-85.

HORII, T., MORITA, S., KIMURA, M. & HATADA, I. 2009. Epigenetic regulation of adipocyte differentiation by a Rho guanine nucleotide exchange factor, WGEF. *PLoS One,* 4**,** e5809.

HOWEY, R. & CORDELL, H. J. 2012. PREMIM and EMIM: tools for estimation of maternal, imprinting and interaction effects using multinomial modelling. *BMC Bioinformatics,* 13**,** 149.

HRESKO, W. P., REID, D. K. & HAMMILL, D. D. 1999. *Test of Early Language Development (TELD-3),* Austin, TX, Pro-Ed.

HURST, J. A., BARAITSER, M., AUGER, E., GRAHAM, F. & NORELL, S. 1990. An extended family with a dominantly inherited speech disorder. *Dev Med Child Neurol,* 32**,** 352-5.

KAHRIZI, K., NAJMABADI, H., KARIMINEJAD, R., JAMALI, P., MALEKPOUR, M., GARSHASBI, M., ROPERS, H. H., KUSS, A. W. & TZSCHACH, A. 2009. An autosomal recessive syndrome of severe mental retardation, cataract, coloboma and kyphosis maps to the pericentromeric region of chromosome 4. *Eur J Hum Genet,* 17**,** 125-8.

KAMAL, M., VALANCIUTE, A., DAHAN, K., ORY, V., PAWLAK, A., LANG, P., GUELLAEN, G. & SAHALI, D. 2009. C-mip interacts physically with RelA and inhibits nuclear factor kappa B activity. *Mol Immunol,* 46**,** 991-8.

KAMHI, A. G. & CATTS, H. W. 1986. Toward an understanding of developmental language and reading disorders. *J Speech Hear Disord,* 51**,** 337-47.

KENT, W. J., SUGNET, C. W., FUREY, T. S., ROSKIN, K. M., PRINGLE, T. H., ZAHLER, A. M. & HAUSSLER, D. 2002. The human genome browser at UCSC. *Genome Res,* 12**,** 996-1006.

KIM, J. W., WALDMAN, I. D., FARAONE, S. V., BIEDERMAN, J., DOYLE, A. E., PURCELL, S., ARBEITMAN, L., FAGERNESS, J., SKLAR, P. & SMOLLER, J. W. 2007. Investigation of parent-of-origin effects in ADHD candidate genes. *Am J Med Genet B Neuropsychiatr Genet,* 144B**,** 776-80.

KIM, K. K., NAM, J., MUKOUYAMA, Y. S. & KAWAMOTO, S. 2013. Rbfox3-regulated alternative splicing of Numb promotes neuronal differentiation during development. *J Cell Biol,* 200**,** 443-58.

KJELGAARD, M. M. & TAGER-FLUSBERG, H. 2001. An Investigation of Language Impairment in Autism: Implications for Genetic Subgroups. *Lang Cogn Process,* 16**,** 287-308.

KONG, A., STEINTHORSDOTTIR, V., MASSON, G., THORLEIFSSON, G., SULEM, P., BESENBACHER, S., JONASDOTTIR, A., SIGURDSSON, A., KRISTINSSON, K. T., FRIGGE, M. L., GYLFASON, A., OLASON, P. I., GUDJONSSON, S. A., SVERRISSON, S., STACEY, S. N., SIGURGEIRSSON, B., BENEDIKTSDOTTIR, K. R., SIGURDSSON, H., JONSSON, T., BENEDIKTSSON, R., OLAFSSON, J. H., JOHANNSSON, O. T., HREIDARSSON, A. B., SIGURDSSON, G., FERGUSON-SMITH, A. C., GUDBJARTSSON, D. F., THORSTEINSDOTTIR, U. & STEFANSSON, K. 2009. Parental origin of sequence variants associated with complex diseases. *Nature,* 462**,** 868-74.

KOTZOT, D. 2001. Comparative analysis of isodisomic and heterodisomic segments in cases with maternal uniparental disomy 14 suggests more than one imprinted region. *Clin Genet,* 60**,** 226-31.

KRAWCZAK, M. 2006. Multiple Significance Testing in Genetic Epidemiology. *eLS.* Chichester: John Wiley & Sons, Ltd.

KURT, S., FISHER, S. E. & EHRET, G. 2012. Foxp2 mutations impair auditory-motor association learning. *PLoS One,* 7**,** e33130.

KUTSCHE, K., YNTEMA, H., BRANDT, A., JANTKE, I., NOTHWANG, H. G., ORTH, U., BOAVIDA, M. G., DAVID, D., CHELLY, J., FRYNS, J. P., MORAINE, C., ROPERS, H. H., HAMEL, B. C., VAN BOKHOVEN, H. & GAL, A. 2000. Mutations in ARHGEF6, encoding a guanine nucleotide exchange factor for Rho GTPases, in patients with X-linked mental retardation. *Nat Genet,* 26**,** 247-50.

LAI, C. S., FISHER, S. E., HURST, J. A., LEVY, E. R., HODGSON, S., FOX, M., JEREMIAH, S., POVEY, S., JAMISON, D. C., GREEN, E. D., VARGHA-KHADEM, F. & MONACO, A. P. 2000. The SPCH1 region on human 7q31: genomic characterization of the critical interval and

localization of translocations associated with speech and language disorder. *Am J Hum Genet,* 67**,** 357-68.

LAI, C. S., FISHER, S. E., HURST, J. A., VARGHA-KHADEM, F. & MONACO, A. P. 2001. A forkhead-domain gene is mutated in a severe speech and language disorder. *Nature,* 413**,** 519-23.

LAI, C. S., GERRELLI, D., MONACO, A. P., FISHER, S. E. & COPP, A. J. 2003. FOXP2 expression during brain development coincides with adult sites of pathology in a severe speech and language disorder. *Brain,* 126**,** 2455-62.

LAKICS, V., KARRAN, E. H. & BOESS, F. G. 2010. Quantitative comparison of phosphodiesterase mRNA distribution in human brain and peripheral tissues. *Neuropharmacology,* 59**,** 367-74.

LANDER, E. & KRUGLYAK, L. 1995. Genetic dissection of complex traits: guidelines for interpreting and reporting linkage results. *Nat Genet,* 11**,** 241-7.

LANDER, E. S. & SCHORK, N. J. 1994. Genetic dissection of complex traits. *Science,* 265**,** 2037-48.

LAUDERDALE, J. D., WILENSKY, J. S., OLIVER, E. R., WALTON, D. S. & GLASER, T. 2000. 3' deletions cause aniridia by preventing PAX6 gene expression. *Proc Natl Acad Sci U S A,* 97**,** 13755-9.

LEADHOLM, B. J. & MILLER, J. F. 1992. *Language Sample Analysis: The Wisconsin Guide,* Milwaukee, Wisconsin Department of Public Instruction.

LEE, L. C., ZACHARY, A. A., LEFFELL, M. S., NEWSCHAFFER, C. J., MATTESON, K. J., TYLER, J. D. & ZIMMERMAN, A. W. 2006. HLA-DR4 in families with autism. *Pediatr Neurol,* 35**,** 303-7.

LESCH, K. P., TIMMESFELD, N., RENNER, T. J., HALPERIN, R., ROSER, C., NGUYEN, T. T., CRAIG, D. W., ROMANOS, J., HEINE, M., MEYER, J., FREITAG, C., WARNKE, A., ROMANOS, M., SCHAFER, H., WALITZA, S., REIF, A., STEPHAN, D. A. & JACOB, C. 2008. Molecular genetics of adult ADHD: converging evidence from genome-wide association and extended pedigree linkage studies. *J Neural Transm,* 115**,** 1573-85.

LEVY, Y. & EBSTEIN, R. P. 2009. Research review: crossing syndrome boundaries in the search for brain endophenotypes. *J Child Psychol Psychiatry,* 50**,** 657-68.

LEWIS, B. A., SHRIBERG, L. D., FREEBAIRN, L. A., HANSEN, A. J., STEIN, C. M., TAYLOR, H. G. & IYENGAR, S. K. 2006. The genetic bases of speech sound disorders: evidence from spoken and written language. *J Speech Lang Hear Res,* 49**,** 1294-312.

LI, H., HANDSAKER, B., WYSOKER, A., FENNELL, T., RUAN, J., HOMER, N., MARTH, G., ABECASIS, G. & DURBIN, R. 2009. The Sequence Alignment/Map format and SAMtools. *Bioinformatics,* 25**,** 2078-9.

LIEGEOIS, F., BALDEWEG, T., CONNELLY, A., GADIAN, D. G., MISHKIN, M. & VARGHA-KHADEM, F. 2003. Language fMRI abnormalities associated with FOXP2 gene mutation. *Nat Neurosci,* 6**,** 1230-7.

LIN, P. I., VANCE, J. M., PERICAK-VANCE, M. A. & MARTIN, E. R. 2007. No gene is an island: the flip-flop phenomenon. *Am J Hum Genet,* 80**,** 531-8.

LISOWSKI, P., PIERZCHALA, M., GOSCIK, J., PAREEK, C. S. & ZWIERZCHOWSKI, L. 2008. Evaluation of reference genes for studies of gene expression in the bovine liver, kidney, pituitary, and thyroid. *J Appl Genet,* 49**,** 367-72.

LIU, J., NYHOLT, D. R., MAGNUSSEN, P., PARANO, E., PAVONE, P., GESCHWIND, D., LORD, C., IVERSEN, P., HOH, J., OTT, J. & GILLIAM, T. C. 2001. A genomewide screen for autism susceptibility loci. *Am J Hum Genet,* 69**,** 327-40.

LIVAK, K. J. & SCHMITTGEN, T. D. 2001. Analysis of relative gene expression data using real-time quantitative PCR and the 2(-Delta Delta C(T)) Method. *Methods,* 25**,** 402-8.

MACDERMOT, K. D., BONORA, E., SYKES, N., COUPE, A. M., LAI, C. S., VERNES, S. C., VARGHA-KHADEM, F., MCKENZIE, F., SMITH, R. L., MONACO, A. P. & FISHER, S. E. 2005. Identification of FOXP2 truncation as a novel cause of developmental speech and language deficits. *Am J Hum Genet,* 76**,** 1074-80.

MAGGIOLO, M. & PAVEZ, M. 2000. *Test para evaluar los procesos fonológicos de simplificación (TEPROSIF),* Santiago, Escuela de Fonoaudiología, Facultad de Medicina, Universidad de Chile.

MANLY, K. F. 2005. Reliability of statistical associations between genes and disease. *Immunogenetics,* 57**,** 549-58.

MARCHMAN, V. A., WULFECK, B. & WEISMER, S. E. 1999. Morphological productivity in children with normal language and SLI: a study of the English past tense. *J Speech Lang Hear Res,* 42**,** 206-19.

MARLOW, A. J., FISHER, S. E., FRANCKS, C., MACPHIE, I. L., CHERNY, S. S., RICHARDSON, A. J., TALCOTT, J. B., STEIN, J. F., MONACO, A. P. & CARDON, L. R. 2003. Use of multivariate linkage analysis for dissection of a complex cognitive trait. *Am J Hum Genet,* 72**,** 561-70.

MAYCOX, P. R., KELLY, F., TAYLOR, A., BATES, S., REID, J., LOGENDRA, R., BARNES, M. R., LARMINIE, C., JONES, N., LENNON, M., DAVIES, C., HAGAN, J. J., SCORER, C. A., ANGELINETTA, C., AKBAR, M. T., HIRSCH, S., MORTIMER, A. M., BARNES, T. R. & DE BELLEROCHE, J. 2009. Analysis of gene expression in two large schizophrenia cohorts identifies multiple changes associated with nerve terminal function. *Mol Psychiatry,* 14**,** 1083-94.

MCARTHUR, G. M. & BISHOP, D. V. 2004. Which People with Specific Language Impairment have Auditory Processing Deficits? *Cogn Neuropsychol,* 21**,** 79-94.

MCARTHUR, G. M., HOGBEN, J. H., EDWARDS, V. T., HEATH, S. M. & MENGLER, E. D. 2000. On the "specifics" of specific reading disability and specific language impairment. *J Child Psychol Psychiatry,* 41**,** 869-74.

MCCARTHY, M. I., ABECASIS, G. R., CARDON, L. R., GOLDSTEIN, D. B., LITTLE, J., IOANNIDIS, J. P. & HIRSCHHORN, J. N. 2008. Genome-wide association studies for complex traits: consensus, uncertainty and challenges. *Nat Rev Genet,* 9**,** 356-69.

MEABURN, E., DALE, P. S., CRAIG, I. W. & PLOMIN, R. 2002. Language-impaired children: No sign of the FOXP2 mutation. *Neuroreport,* 13**,** 1075-7.

MEMET, S. 2006. NF-kappaB functions in the nervous system: from development to disease. *Biochem Pharmacol,* 72**,** 1180-95.

MICK, E. & FARAONE, S. V. 2008. Genetics of attention deficit hyperactivity disorder. *Child Adolesc Psychiatr Clin N Am,* 17**,** 261-84, vii-viii.

MISSIAEN, L., DODE, L., VANOEVELEN, J., RAEYMAEKERS, L. & WUYTACK, F. 2007. Calcium in the Golgi apparatus. *Cell Calcium,* 41**,** 405-16.

MONACO, A. P. & THE SLI CONSORTIUM 2007. Multivariate Linkage Analysis of Specific Language Impairment (SLI). *Ann Hum Genet,* 71**,** 660–673.

MORRIS, J. A. & GARDNER, M. J. 1988. Calculating confidence intervals for relative risks (odds ratios) and standardised ratios and rates. *Br Med J (Clin Res Ed),* 296**,** 1313-6.

MORTON, N. E. 1955. Sequential tests for the detection of linkage. *Am J Hum Genet,* 7**,** 277-318.

MOTT, R., YUAN, W., KAISAKI, P., GAN, X., CLEAK, J., EDWARDS, A., BAUD, A. & FLINT, J. 2014. The Architecture of Parent-of-Origin Effects in Mice. *Cell,* 156**,** 332-42.

NAJMABADI, H., MOTAZACKER, M. M., GARSHASBI, M., KAHRIZI, K., TZSCHACH, A., CHEN, W., BEHJATI, F., HADAVI, V., NIEH, S. E., ABEDINI, S. S., VAZIFEHMAND, R., FIROUZABADI, S. G., JAMALI, P., FALAH, M., SEIFATI, S. M., GRUTERS, A., LENZNER, S., JENSEN, L. R., RUSCHENDORF, F., KUSS, A. W. & ROPERS, H. H. 2007. Homozygosity mapping in consanguineous families reveals extreme heterogeneity of non-syndromic autosomal recessive mental retardation and identifies 8 novel gene loci. *Hum Genet,* 121**,** 43-8.

NEVILLE, H. J., COFFEY, S. A., HOLCOMB, P. J. & TALLAL, P. 1993. The neurobiology of sensory and language processing in language-impaired children. *Journal of Cognitive Neuroscience,* 5**,** 235-253.

NEVSIMALOVA, S., TAUBEROVA, A., DOUTLIK, S., KUCERA, V. & DLOUHA, O. 1992. A role of autoimmunity in the etiopathogenesis of Landau-Kleffner syndrome? *Brain Dev,* 14**,** 342-5.

NEWBURY, D. F., BISHOP, D. V. M. & MONACO, A. P. 2005. Genetic influences on language impairment and phonological short-term memory. *Trends Cogn Sci,* 9**,** 528-34.

NEWBURY, D. F., BONORA, E., LAMB, J. A., FISHER, S. E., LAI, C. S., BAIRD, G., JANNOUN, L., SLONIMS, V., STOTT, C. M., MERRICKS, M. J., BOLTON, P. F., BAILEY, A. J. & MONACO, A. P. 2002. FOXP2 is not a major susceptibility gene for autism or specific language impairment. *Am J Hum Genet,* 70**,** 1318-27.

NEWBURY, D. F., PARACCHINI, S., SCERRI, T. S., WINCHESTER, L., ADDIS, L., RICHARDSON, A. J., WALTER, J., STEIN, J. F., TALCOTT, J. B. & MONACO, A. P. 2011. Investigation of dyslexia and SLI risk variants in reading- and language-impaired subjects. *Behav Genet,* 41**,** 90-104.

NEWBURY, D. F., WINCHESTER, L., ADDIS, L., PARACCHINI, S., BUCKINGHAM, L. L., CLARK, A., COHEN, W., COWIE, H., DWORZYNSKI, K., EVERITT, A., GOODYER, I. M., HENNESSY, E., KINDLEY, A. D., MILLER, L. L., NASIR, J., O'HARE, A., SHAW, D., SIMKIN, Z., SIMONOFF, E., SLONIMS, V., WATSON, J., RAGOUSSIS, J., FISHER, S. E., SECKL, J. R., HELMS, P. J., BOLTON, P. F., PICKLES, A., CONTI-RAMSDEN, G., BAIRD, G., BISHOP, D. V. M. & MONACO, A. P. 2009. CMIP and ATP2C2 modulate phonological short-term memory in language impairment. *Am J Hum Genet,* 85**,** 264-72.

NEWCOMER, P. L. & HAMMILL, D. D. 1988. The Test of Language Development–2, Primary. Austin, TX: Pro-Ed.

NEWCOMER, P. L. & HAMMILL, D. D. 1999. The Test of Language Development–Primary, Third edition. Austin, TX: Pro-Ed.

NEWEY, S. E., VELAMOOR, V., GOVEK, E. E. & VAN AELST, L. 2005. Rho GTPases, dendritic structure, and mental retardation. *J Neurobiol,* 64**,** 58-74.

NITHART, C., DEMONT, E., MAJERUS, S., LEYBAERT, J., PONCELET, M. & METZ-LUTZ, M. N. 2009. Reading disabilities in SLI and dyslexia result from distinct phonological impairments. *Dev Neuropsychol,* 34**,** 296-311.

NORBURY, C. F. & BISHOP, D. V. M. 2002. Inferential processing and story recall in children with communication problems: a comparison of specific language impairment, pragmatic language impairment and high-functioning autism. *Int J Lang Commun Disord,* 37**,** 227-51.

O'BRIEN, E. K., ZHANG, X., NISHIMURA, C., TOMBLIN, J. B. & MURRAY, J. C. 2003. Association of specific language impairment (SLI) to the region of 7q31. *Am J Hum Genet,* 72**,** 1536-43.

O'ROAK, B. J., DERIZIOTIS, P., LEE, C., VIVES, L., SCHWARTZ, J. J., GIRIRAJAN, S., KARAKOC, E., MACKENZIE, A. P., NG, S. B., BAKER, C., RIEDER, M. J., NICKERSON, D. A., BERNIER, R., FISHER, S. E., SHENDURE, J. & EICHLER, E. E. 2011. Exome sequencing in sporadic autism spectrum disorders identifies severe de novo mutations. *Nat Genet,* 43**,** 585-9.

ODELL, J. D., WARREN, R. P., WARREN, W. L., BURGER, R. A. & MACIULIS, A. 1997. Association of genes within the major histocompatibility complex with attention deficit hyperactivity disorder. *Neuropsychobiology,* 35**,** 181-6.

OGDIE, M. N., BAKKER, S. C., FISHER, S. E., FRANCKS, C., YANG, M. H., CANTOR, R. M., LOO, S. K., VAN DER MEULEN, E., PEARSON, P., BUITELAAR, J., MONACO, A., NELSON, S. F., SINKE, R. J. & SMALLEY, S. L. 2006. Pooled genome-wide linkage data on 424 ADHD ASPs suggests genetic heterogeneity and a common risk locus at 5p13. *Mol Psychiatry,* 11**,** 5-8.

OGDIE, M. N., FISHER, S. E., YANG, M., ISHII, J., FRANCKS, C., LOO, S. K., CANTOR, R. M., MCCRACKEN, J. T., MCGOUGH, J. J., SMALLEY, S. L. & NELSON, S. F. 2004. Attention deficit hyperactivity disorder: fine mapping supports linkage to 5p13, 6q12, 16p13, and 17p11. *Am J Hum Genet,* 75**,** 661-8.

OGDIE, M. N., MACPHIE, I. L., MINASSIAN, S. L., YANG, M., FISHER, S. E., FRANCKS, C., CANTOR, R. M., MCCRACKEN, J. T., MCGOUGH, J. J., NELSON, S. F., MONACO, A. P. & SMALLEY, S. L. 2003. A genomewide scan for attention-deficit/hyperactivity disorder in an extended sample: suggestive linkage on 17p11. *Am J Hum Genet,* 72**,** 1268-79.

PALMER, C. G., HSIEH, H. J., REED, E. F., LONNQVIST, J., PELTONEN, L., WOODWARD, J. A. & SINSHEIMER, J. S. 2006. HLA-B maternal-fetal genotype matching increases risk of schizophrenia. *Am J Hum Genet,* 79**,** 710-5.

PARACCHINI, S., SCERRI, T. & MONACO, A. P. 2007. The genetic lexicon of dyslexia. *Annu Rev Genomics Hum Genet,* 8**,** 57-79.

PAVEZ, M. 2003. *Test exploratorio de Gramática española de A. Toronto. Aplicación en Chile,* Santiago, Ediciones Universidad católica de Chile.

PENNINGTON, B. F. & BISHOP, D. V. M. 2009. Relations among speech, language, and reading disorders. *Annu Rev Psychol,* 60**,** 283-306.

PETER, B., RASKIND, W. H., MATSUSHITA, M., LISOWSKI, M., VU, T., BERNINGER, V. W., WIJSMAN, E. M. & BRKANAC, Z. 2011. Replication of CNTNAP2 association with nonword repetition and support for FOXP2 association with timed reading and motor activities in a dyslexia family sample. *J Neurodev Disord,* 3**,** 39-49.

PETERSEN, S. E., FOX, P. T., POSNER, M. I., MINTUN, M. & RAICHLE, M. E. 1989. Positron emission tomographic studies of the processing of singe words. *J Cogn Neurosci,* 1**,** 153-70.

PINACHO, R., VILLALMANZO, N., LALONDE, J., HARO, J. M., MEANA, J. J., GILL, G. & RAMOS, B. 2011. The transcription factor SP4 is reduced in postmortem cerebellum of bipolar disorder subjects: control by depolarization and lithium. *Bipolar Disord,* 13**,** 474-85.

POLAN, M. B., PASTORE, M. T., STEINGASS, K., HASHIMOTO, S., THRUSH, D. L., PYATT, R., RESHMI, S., GASTIER-FOSTER, J. M., ASTBURY, C. & MCBRIDE, K. L. 2013. Neurodevelopmental disorders among individuals with duplication of 4p13 to 4p12 containing a GABA receptor subunit gene cluster. *Eur J Hum Genet*.

POLANCZYK, G., DE LIMA, M. S., HORTA, B. L., BIEDERMAN, J. & ROHDE, L. A. 2007. The worldwide prevalence of ADHD: a systematic review and metaregression analysis. *Am J Psychiatry,* 164**,** 942-8.

POLIAK, S. & PELES, E. 2003. The local differentiation of myelinated axons at nodes of Ranvier. *Nat Rev Neurosci,* 4**,** 968-80.

PRATT, S. C., DALY, M. J. & KRUGLYAK, L. 2000. Exact multipoint quantitative-trait linkage analysis in pedigrees by variance components. *Am J Hum Genet,* 66**,** 1153-7.

PRUIM, R. J., WELCH, R. P., SANNA, S., TESLOVICH, T. M., CHINES, P. S., GLIEDT, T. P., BOEHNKE, M., ABECASIS, G. R. & WILLER, C. J. 2010. LocusZoom: regional visualization of genome-wide association scan results. *Bioinformatics,* 26**,** 2336-7.

PURCELL, S., CHERNY, S. S. & SHAM, P. C. 2003. Genetic Power Calculator: design of linkage and association genetic mapping studies of complex traits. *Bioinformatics,* 19**,** 149-50.

PURCELL, S., NEALE, B., TODD-BROWN, K., THOMAS, L., FERREIRA, M. A., BENDER, D., MALLER, J., SKLAR, P., DE BAKKER, P. I., DALY, M. J. & SHAM, P. C. 2007. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet,* 81**,** 559-75.

RABIN, M., WEN, X. L., HEPBURN, M., LUBS, H. A., FELDMAN, E. & DUARA, R. 1993. Suggestive linkage of developmental dyslexia to chromosome 1p34-p36. *Lancet,* 342**,** 178.

RAPIN, I. 1996. Practitioner review: developmental language disorders: a clinical update. *J Child Psychol Psychiatry,* 37**,** 643-55.

RAVEN, J., RAVEN, J. & VCOURT, J. 2003. *Manual for Raven's Progressive Matrices and Vocabulary Scales,* San Antonio, TX., Harcourt assessment.

RICE, G. M., RACA, G., JAKIELSKI, K. J., LAFFIN, J. J., IYAMA-KURTYCZ, C. M., HARTLEY, S. L., SPRAGUE, R. E., HEINTZELMAN, A. T. & SHRIBERG, L. D. 2012. Phenotype of FOXP2 haploinsufficiency in a mother and son. *Am J Med Genet A,* 158A**,** 174-81.

RICE, M. & WEXLER, K. 2001. *Test of Early Grammatical Impairment,* San Antonio, Psychological Corporation.

RICE, M. L., SMITH, S. D. & GAYAN, J. 2009. Convergent genetic linkage and associations to language, speech and reading measures in families with Specific Language Impairment. *J Neurodev Disord,* 1**,** 264-282.

RICE, M. L., WEXLER, K. & CLEAVE, P. L. 1995. Specific language impairment as a period of extended optional infinitive. *J Speech Hear Res,* 38**,** 850-63.

RIMMER, A., MATHIESON, I., LUNTER, G. & MCVEAN, G. 2012. Platypus: An Integrated Variant Caller. *www.well.ox.ac.uk/platypus*.

RISCH, N. & MERIKANGAS, K. 1996. The future of genetic studies of complex human diseases. *Science,* 273**,** 1516-7.

RISCH, N. J. 2000. Searching for genetic determinants in the new millennium. *Nature,* 405**,** 847-56.

RITVO, E. R., FREEMAN, B. J., SCHEIBEL, A. B., DUONG, T., ROBINSON, H., GUTHRIE, D. & RITVO, A. 1986. Lower Purkinje cell counts in the cerebella of four autistic subjects: initial findings of the UCLA-NSAC Autopsy Research Report. *Am J Psychiatry,* 143**,** 862-6.

RIVAS, M. A., BEAUDOIN, M., GARDET, A., STEVENS, C., SHARMA, Y., ZHANG, C. K., BOUCHER, G., RIPKE, S., ELLINGHAUS, D., BURTT, N., FENNELL, T., KIRBY, A., LATIANO, A., GOYETTE, P., GREEN, T., HALFVARSON, J., HARITUNIANS, T., KORN, J. M., KURUVILLA, F., LAGACE, C., NEALE, B., LO, K. S., SCHUMM, P., TORKVIST, L., DUBINSKY, M. C., BRANT, S. R., SILVERBERG, M. S., DUERR, R. H., ALTSHULER, D., GABRIEL, S., LETTRE, G., FRANKE, A., D'AMATO, M., MCGOVERN, D. P., CHO, J. H., RIOUX, J. D., XAVIER, R. J. & DALY, M. J. 2011. Deep resequencing of GWAS loci identifies independent rare variants associated with inflammatory bowel disease. *Nat Genet,* 43**,** 1066-73.

ROMANOS, M., FREITAG, C., JACOB, C., CRAIG, D. W., DEMPFLE, A., NGUYEN, T. T., HALPERIN, R., WALITZA, S., RENNER, T. J., SEITZ, C., ROMANOS, J., PALMASON, H., REIF, A., HEINE, M., WINDEMUTH-KIESELBACH, C., VOGLER, C., SIGMUND, J., WARNKE, A., SCHAFER, H., MEYER, J., STEPHAN, D. A. & LESCH, K. P. 2008. Genome-wide linkage analysis of ADHD using high-density SNP arrays: novel loci at 5q13.1 and 14q12. *Mol Psychiatry,* 13**,** 522-30.

ROZEN, S. & SKALETSKY, H. 2000. Primer3 on the WWW for general users and for biologist programmers. *Methods Mol Biol,* 132**,** 365-86.

RUST, J. 1996. *WOLD Wechsler Objective Language Dimensions Manual,* London, United Kingdom, The Psychological Corporation.

RUST, J., GOLOMBOK, S., TRICKEY, G. 1993. *Wechsler objective reading dimensions,* Sidcup, Kent, Psychological Corporation.

SAUVAGEAU, M., GOFF, L. A., LODATO, S., BONEV, B., GROFF, A. F., GERHARDINGER, C., SANCHEZ-GOMEZ, D. B., HACISULEYMAN, E., LI, E., SPENCE, M., LIAPIS, S. C., MALLARD, W., MORSE, M., SWERDEL, M. R., D'ECCLESSIS, M. F., MOORE, J. C., LAI, V., GONG, G., YANCOPOULOS, G. D., FRENDEWEY, D., KELLIS, M., HART, R. P., VALENZUELA, D. M., ARLOTTA, P. & RINN, J. L. 2013. Multiple knockout mouse models reveal lincRNAs are required for life and brain development. *Elife,* 2**,** e01749.

SCERRI, T. S., MORRIS, A. P., BUCKINGHAM, L. L., NEWBURY, D. F., MILLER, L. L., MONACO, A. P., BISHOP, D. V. & PARACCHINI, S. 2011. DCDC2, KIAA0319 and CMIP are associated with reading-related traits. *Biological Psychiatry,* 70**,** 237-45.

SCHANEN, N. C. 2006. Epigenetics of autism spectrum disorders. *Hum Mol Genet,* 15 Spec No 2**,** R138-50.

SCHMITT, A., LEONARDI-ESSMANN, F., DURRENBERGER, P. F., PARLAPANI, E., SCHNEIDER-AXMANN, T., SPANAGEL, R., ARZBERGER, T., KRETZSCHMAR, H., HERRERA-MARSCHITZ, M., GRUBER, O., REYNOLDS, R., FALKAI, P. & GEBICKE-HAERTER, P. J. 2011. Regulation of immune-modulatory genes in left superior temporal cortex of schizophrenia patients: a genome-wide microarray study. *World J Biol Psychiatry,* 12**,** 201-15.

SCHREUDER, G. M., HURLEY, C. K., MARSH, S. G., LAU, M., MAIERS, M., KOLLMAN, C. & NOREEN, H. J. 2001. The HLA dictionary 2001: a summary of HLA-A, -B, -C, -DRB1/3/4/5, -DQB1 alleles and their association with serologically defined HLA-A, -B, -C, -DR, and -DQ antigens. *Hum Immunol,* 62**,** 826-49.

SCHWARZ, J. M., RODELSPERGER, C., SCHUELKE, M. & SEELOW, D. 2010. MutationTaster evaluates disease-causing potential of sequence alterations. *Nat Methods,* 7**,** 575-6.

SEIDMAN, L. J. 2006. Neuropsychological functioning in people with ADHD across the lifespan. *Clin Psychol Rev,* 26**,** 466-85.

SEMEL, E. M., WIIG, E. H. & SECORD, W. 1992. *Clinical Evaluation of Language Fundamentals - Revised,* San Antonio, Phychological Corporation.

SEMEL, E. M., WIIG, E. H. & SECORD, W. 1995. *Clinical Evaluation of Language Fundamentals - Third edition,* San Antonio, Phychological Corporation.

SHIINA, T., INOKO, H. & KULSKI, J. K. 2004. An update of the HLA genomic region, locus information and disease associations: 2004. *Tissue Antigens,* 64**,** 631-49.

SHOUKIER, M., WICKERT, J., SCHRODER, J., BARTELS, I., AUBER, B., ZOLL, B., SALINAS-RIESTER, G., WEISE, D., BROCKMANN, K., ZIRN, B. & BURFEIND, P. 2012. A 16q12 microdeletion in a boy with severe psychomotor delay, craniofacial dysmorphism, brain and limb malformations, and a heart defect. *Am J Med Genet A,* 158A**,** 229-35.

SHRIBERG, L. D., TOMBLIN, J. B. & MCSWEENY, J. L. 1999. Prevalence of speech delay in 6-year-old children and comorbidity with language impairment. *J Speech Lang Hear Res,* 42**,** 1461-81.

SHU, W., CHO, J. Y., JIANG, Y., ZHANG, M., WEISZ, D., ELDER, G. A., SCHMEIDLER, J., DE GASPERI, R., SOSA, M. A., RABIDOU, D., SANTUCCI, A. C., PERL, D., MORRISEY, E. & BUXBAUM, J. D. 2005. Altered ultrasonic vocalization in mice with a disruption in the Foxp2 gene. *Proc Natl Acad Sci U S A,* 102**,** 9643-8.

SHU, W., YANG, H., ZHANG, L., LU, M. M. & MORRISEY, E. E. 2001. Characterization of a new subfamily of winged-helix/forkhead (Fox) genes that are expressed in the lung and act as transcriptional repressors. *J Biol Chem,* 276**,** 27488-97.

SHYN, S. I., SHI, J., KRAFT, J. B., POTASH, J. B., KNOWLES, J. A., WEISSMAN, M. M., GARRIOCK, H. A., YOKOYAMA, J. S., MCGRATH, P. J., PETERS, E. J., SCHEFTNER, W. A., CORYELL, W., LAWSON, W. B., JANCIC, D., GEJMAN, P. V., SANDERS, A. R., HOLMANS, P., SLAGER, S. L., LEVINSON, D. F. & HAMILTON, S. P. 2011. Novel loci for major depression identified by genome-wide association study of Sequenced Treatment Alternatives to Relieve Depression and meta-analysis of three studies. *Mol Psychiatry,* 16**,** 202-15.

SILVERI, M. C., LEGGIO, M. G. & MOLINARI, M. 1994. The cerebellum contributes to linguistic production: a case of agrammatic speech following a right cerebellar lesion. *Neurology,* 44**,** 2047-50.

SINGH, L. N. & HANNENHALLI, S. 2008. Functional diversification of paralogous transcription factors via divergence in DNA binding site motif and in expression. *PLoS One,* 3**,** e2345.

SINSHEIMER, J. S., PALMER, C. G. & WOODWARD, J. A. 2003. Detecting genotype combinations that increase risk for disease: maternal-fetal genotype incompatibility test. *Genet Epidemiol,* 24**,** 1-13.

SMITH, C. A. 1963. Testing for Heterogeneity of Recombination Fraction Values in Human Genetics. *Ann Hum Genet,* 27**,** 175-82.

SMITH, S. D., PENNINGTON, B. F., BOADA, R. & SHRIBERG, L. D. 2005. Linkage of speech sound disorder to reading disability loci. *J Child Psychol Psychiatry,* 46**,** 1057-66.

SONG, Z., KRISHNA, S., THANOS, D., STROMINGER, J. L. & ONO, S. J. 1994. A novel cysteine-rich sequence-specific DNA-binding protein interacts with the conserved X-box motif of the human major histocompatibility complex class II genes via a repeated Cys-His domain and functions as a transcriptional repressor. *J Exp Med,* 180**,** 1763-74.

SPIELMAN, R. S., MCGINNIS, R. E. & EWENS, W. J. 1993. Transmission test for linkage disequilibrium: the insulin gene region and insulin-dependent diabetes mellitus (IDDM). *Am J Hum Genet,* 52**,** 506-16.

SPITERI, E., KONOPKA, G., COPPOLA, G., BOMAR, J., OLDHAM, M., OU, J., VERNES, S. C., FISHER, S. E., REN, B. & GESCHWIND, D. H. 2007. Identification of the transcriptional targets of FOXP2, a gene linked to speech and language, in developing human brain. *Am J Hum Genet,* 81**,** 1144-57.

ST LOUIS, K. & RUSCELLO, D. 1987. Oral Speech Mechanism Screening Exam. Austin, TX: Pro-Ed.

STEIN, C. M., MILLARD, C., KLUGE, A., MISCIMARRA, L. E., CARTIER, K. C., FREEBAIRN, L. A., HANSEN, A. J., SHRIBERG, L. D., TAYLOR, H. G., LEWIS, B. A. & IYENGAR, S. K. 2006. Speech sound disorder influenced by a locus in 15q14 region. *Behav Genet,* 36**,** 858-68.

STEIN, C. M., SCHICK, J. H., GERRY TAYLOR, H., SHRIBERG, L. D., MILLARD, C., KUNDTZ-KLUGE, A., RUSSO, K., MINICH, N., HANSEN, A., FREEBAIRN, L. A., ELSTON, R. C., LEWIS, B. A. & IYENGAR, S. K. 2004. Pleiotropic effects of a chromosome 3 locus on speech-sound disorder and reading. *Am J Hum Genet,* 74**,** 283-97.

STOREY, J. D. 2002. A direct approach to false discovery rates. *Journal of the Royal Statistical Society: Series B (Statistical Methodology),* 64**,** 479-498.

STRANGER, B. E., MONTGOMERY, S. B., DIMAS, A. S., PARTS, L., STEGLE, O., INGLE, C. E., SEKOWSKA, M., SMITH, G. D., EVANS, D., GUTIERREZ-ARCELUS, M., PRICE, A., RAJ, T., NISBETT, J., NICA, A. C., BEAZLEY, C., DURBIN, R., DELOUKAS, P. & DERMITZAKIS, E. T. 2012. Patterns of cis regulatory variation in diverse human populations. *PLoS Genet,* 8**,** e1002639.

STRAUSS, K. A., PUFFENBERGER, E. G., HUENTELMAN, M. J., GOTTLIEB, S., DOBRIN, S. E., PAROD, J. M., STEPHAN, D. A. & MORTON, D. H. 2006. Recessive symptomatic focal epilepsy and mutant contactin-associated protein-like 2. *N Engl J Med,* 354**,** 1370-7.

STROMSWOLD, K. 1998. Genetics of spoken language disorders. *Hum Biol,* 70**,** 297-324.

STROMSWOLD, K. 2001. The Heritability of Language: A Review and Metaanalysis of Twin, Adoption, and Linkage Studies. *Language,* 77**,** 647-723.

SUBRAMANIAM, P. S., LARKIN, J., 3RD, MUJTABA, M. G., WALTER, M. R. & JOHNSON, H. M. 2000. The COOH-terminal nuclear localization sequence of interferon gamma regulates STAT1 alpha nuclear translocation at an intracellular site. *J Cell Sci,* 113 (Pt 15)**,** 2771-81.

SUTTON, V. R. & SHAFFER, L. G. 2000. Search for imprinted regions on chromosome 14: comparison of maternal and paternal UPD cases with cases of chromosome 14 deletion. *Am J Med Genet,* 93**,** 381-7.

TALKOWSKI, M. E., MAUSSION, G., CRAPPER, L., ROSENFELD, J. A., BLUMENTHAL, I., HANSCOM, C., CHIANG, C., LINDGREN, A., PEREIRA, S., RUDERFER, D., DIALLO, A. B., LOPEZ, J. P., TURECKI, G., CHEN, E. S., GIGEK, C., HARRIS, D. J., LIP, V., AN, Y., BIAGIOLI, M., MACDONALD, M. E., LIN, M., HAGGARTY, S. J., SKLAR, P., PURCELL, S., KELLIS, M., SCHWARTZ, S., SHAFFER, L. G., NATOWICZ, M. R., SHEN, Y., MORTON, C. C., GUSELLA, J. F. & ERNST, C. 2012. Disruption of a large intergenic noncoding RNA in subjects with neurodevelopmental disabilities. *Am J Hum Genet,* 91**,** 1128-34.

TALLAL, P. & PIERCY, M. 1973a. Defects of non-verbal auditory perception in children with developmental aphasia. *Nature,* 241**,** 468-9.

TALLAL, P. & PIERCY, M. 1973b. Developmental aphasia: impaired rate of non-verbal processing as a function of sensory modality. *Neuropsychologia,* 11**,** 389-98.

TANEGASHIMA, K., ZHAO, H. & DAWID, I. B. 2008. WGEF activates Rho in the Wnt-PCP pathway and controls convergent extension in Xenopus gastrulation. *EMBO J,* 27**,** 606-17.

TANG, W., YUAN, J., CHEN, X., SHAN, Y., LUO, K., GUO, Z., ZHANG, Y., WAN, B. & YU, L. 2005. Cloning and characterization of the CDZFP gene which encodes a putative zinc finger protein. *DNA Seq,* 16**,** 391-6.

THE 1000 GENOMES PROJECT CONSORTIUM 2012. An integrated map of genetic variation from 1,092 human genomes. *Nature,* 491**,** 56-65.

THE SLI CONSORTIUM 2002. A genomewide scan identifies two novel loci involved in specific language impairment. *Am J Hum Genet,* 70**,** 384-98.

THE SLI CONSORTIUM 2004. Highly significant linkage to the SLI1 locus in an expanded sample of individuals affected by specific language impairment. *Am J Hum Genet,* 74**,** 1225-38.

THE UNIPROT CONSORTIUM 2012. Reorganizing the protein space at the Universal Protein Resource (UniProt). *Nucleic Acids Res,* 40**,** D71-5.

THE WELLCOME TRUST CASE CONTROL CONSORTIUM 2007. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature,* 447**,** 661-78.

THIELE, H. & NURNBERG, P. 2005. HaploPainter: a tool for drawing pedigrees with complex haplotypes. *Bioinformatics,* 21**,** 1730-2.

THOMSON, E., RAPPSILBER, J. & TOLLERVEY, D. 2007. Nop9 is an RNA binding protein present in pre-40S ribosomes and required for 18S rRNA synthesis in yeast. *RNA,* 13**,** 2165-74.

THOMSON, G. 1981. A review of theoretical aspects of HLA and disease associations. *Theor Popul Biol,* 20**,** 168-208.

THORNTON, T. & MCPEEK, M. S. 2007. Case-control association testing with related individuals: a more powerful quasi-likelihood score test. *American Journal of Human Genetics,* 81**,** 321-37.

THORNTON, T., ZHANG, Q., CAI, X., OBER, C. & MCPEEK, M. S. 2012. XM: association testing on the X-chromosome in case-control samples with related individuals. *Genet Epidemiol,* 36**,** 438-50.

TOMBLIN, J. B., O'BRIEN, M., SHRIBERG, L. D., WILLIAMS, C., MURRAY, J., PATIL, S., BJORK, J., ANDERSON, S. & BALLARD, K. 2009. Language features in a mother and daughter of a chromosome 7;13 translocation involving FOXP2. *J Speech Lang Hear Res,* 52**,** 1157-74.

TOMBLIN, J. B., RECORDS, N. L., BUCKWALTER, P., ZHANG, X., SMITH, E. & O'BRIEN, M. 1997. Prevalence of specific language impairment in kindergarten children. *J Speech Lang Hear Res,* 40**,** 1245-60.

TORRES, A. R., MACIULIS, A., STUBBS, E. G., CUTLER, A. & ODELL, D. 2002. The transmission disequilibrium test suggests that HLA-DR4 and DR13 are linked to autism spectrum disorder. *Hum Immunol,* 63**,** 311-6.

TORRES, A. R., SWEETEN, T. L., CUTLER, A., BEDKE, B. J., FILLMORE, M., STUBBS, E. G. & ODELL, D. 2006. The association and linkage of the HLA-A2 class I allele with autism. *Hum Immunol,* 67**,** 346-51.

TZENOVA, J., KAPLAN, B. J., PETRYSHEN, T. L. & FIELD, L. L. 2004. Confirmation of a dyslexia susceptibility locus on chromosome 1p34-p36 in a set of 100 Canadian families. *Am J Med Genet B Neuropsychiatr Genet,* 127**,** 117-24.

ULLMAN, M. T. 2001. A neurocognitive perspective on language: the declarative/procedural model. *Nat Rev Neurosci,* 2**,** 717-26.

ULLMAN, M. T. & PIERPONT, E. I. 2005. Specific language impairment is not specific to language: the procedural deficit hypothesis. *Cortex,* 41**,** 399-433.

UNDERWOOD, J. G., BOUTZ, P. L., DOUGHERTY, J. D., STOILOV, P. & BLACK, D. L. 2005. Homologues of the Caenorhabditis elegans Fox-1 protein are neuronal splicing regulators in mammals. *Mol Cell Biol,* 25**,** 10005-16.

VAAGS, A. K., LIONEL, A. C., SATO, D., GOODENBERGER, M., STEIN, Q. P., CURRAN, S., OGILVIE, C., AHN, J. W., DRMIC, I., SENMAN, L., CHRYSLER, C., THOMPSON, A., RUSSELL, C., PRASAD, A., WALKER, S., PINTO, D., MARSHALL, C. R., STAVROPOULOS, D. J., ZWAIGENBAUM, L., FERNANDEZ, B. A., FOMBONNE, E., BOLTON, P. F., COLLIER, D. A., HODGE, J. C., ROBERTS, W., SZATMARI, P. & SCHERER, S. W. 2012. Rare deletions at the neurexin 3 locus in autism spectrum disorder. *Am J Hum Genet,* 90**,** 133-41.

VAN AELST, L. & D'SOUZA-SCHOREY, C. 1997. Rho GTPases and signaling networks. *Genes Dev,* 11**,** 2295-322.

VAN DER AA, N., VANDEWEYER, G., REYNIERS, E., KENIS, S., DOM, L., MORTIER, G., ROOMS, L. & KOOY, R. F. 2012. Haploinsufficiency of CMIP in a girl with autism spectrum disorder and developmental delay due to a de novo deletion on chromosome 16q23.2. *Autism Res,* 5**,** 277-81.

VAN DER LELY, H. K. 1994. Canonical linking rules: forward versus reverse linking in normally developing and specifically language-impaired children. *Cognition,* 51**,** 29-72.

VAN DER LELY, H. K. 2003. Do Heterogeneous Deficits Require Heterogeneous Theories? SLI Subgroups and the RDDR Hypothesis. *In:* LEVY, Y. & SCHAEFFER, J. C. (eds.) *Language competence across populations: Toward a definition of specific language impairment.*

VAN DER LELY, H. K. & STOLLWERCK, L. 1997. Binding theory and grammatical specific language impairment in children. *Cognition,* 62**,** 245-90.

VARGHA-KHADEM, F., WATKINS, K., ALCOCK, K., FLETCHER, P. & PASSINGHAM, R. 1995. Praxic and nonverbal cognitive deficits in a large family with a genetically transmitted speech and language disorder. *Proc Natl Acad Sci U S A,* 92**,** 930-3.

VARGHA-KHADEM, F., WATKINS, K. E., PRICE, C. J., ASHBURNER, J., ALCOCK, K. J., CONNELLY, A., FRACKOWIAK, R. S., FRISTON, K. J., PEMBREY, M. E., MISHKIN, M., GADIAN, D. G. & PASSINGHAM, R. E. 1998. Neural basis of an inherited speech and language disorder. *Proc Natl Acad Sci U S A,* 95**,** 12695-700.

VERNES, S. C., NEWBURY, D. F., ABRAHAMS, B. S., WINCHESTER, L., NICOD, J., GROSZER, M., ALARCON, M., OLIVER, P. L., DAVIES, K. E., GESCHWIND, D. H., MONACO, A. P. & FISHER, S. E. 2008. A functional genetic link between distinct developmental language disorders. *N Engl J Med,* 359**,** 2337-45.

VERNES, S. C., NICOD, J., ELAHI, F. M., COVENTRY, J. A., KENNY, N., COUPE, A. M., BIRD, L. E., DAVIES, K. E. & FISHER, S. E. 2006. Functional genetic analysis of mutations implicated in a human speech and language disorder. *Hum Mol Genet,* 15**,** 3154-67.

VERNES, S. C., OLIVER, P. L., SPITERI, E., LOCKSTONE, H. E., PULIYADI, R., TAYLOR, J. M., HO, J., MOMBEREAU, C., BREWER, A., LOWY, E., NICOD, J., GROSZER, M., BABAN, D., SAHGAL, N., CAZIER, J. B., RAGOUSSIS, J., DAVIES, K. E., GESCHWIND, D. H. & FISHER, S. E. 2011. Foxp2 regulates gene networks implicated in neurite outgrowth in the developing brain. *PLoS Genet,* 7**,** e1002145.

VERNES, S. C., SPITERI, E., NICOD, J., GROSZER, M., TAYLOR, J. M., DAVIES, K. E., GESCHWIND, D. H. & FISHER, S. E. 2007. High-throughput analysis of promoter occupancy reveals direct neural targets of FOXP2, a gene mutated in speech and language disorders. *Am J Hum Genet,* 81**,** 1232-50.

VIELAND, V. J. 1998. Bayesian linkage analysis, or: how I learned to stop worrying and love the posterior probability of linkage. *Am J Hum Genet,* 63**,** 947-54.

VILLANUEVA, P. 2000. *Pauta de Examen en Habla y motricidad Orofacial. ,* Santiago, Escuela de Fonoaudiología, Facultad de Medicina, Universidad de Chile.

VILLANUEVA, P., DE BARBIERI, Z., PALOMINO, H. M. & PALOMINO, H. 2008. Alta prevalencia de trastorno específico de lenguaje en isla Robinson Crusoe y probable efecto fundador. *Rev Med Chil,* 136**,** 186-92.

VILLANUEVA, P., NEWBURY, D. F., JARA, L., DE BARBIERI, Z., MIRZA, G., PALOMINO, H. M., FERNANDEZ, M. A., CAZIER, J. B., MONACO, A. P. & PALOMINO, H. 2011. Genome-wide analysis of genetic susceptibility to language impairment in an isolated Chilean population. *Eur J Hum Genet,* 19**,** 687-695.

WAGNER, R. K., RASHOTTE, C. A. & TORGESEN, J. K. 1999. *Comprehensive Test of Phonological Processing,* Austin, Pro-Ed.

WANG, K. S., LIU, X., ZHANG, Q., ARAGAM, N. & PAN, Y. 2012. Parent-of-origin effects of FAS and PDLIM1 in attention-deficit/hyperactivity disorder. *J Psychiatry Neurosci,* 37**,** 46-52.

WANG, X., ARAI, S., SONG, X., REICHART, D., DU, K., PASCUAL, G., TEMPST, P., ROSENFELD, M. G., GLASS, C. K. & KUROKAWA, R. 2008a. Induced ncRNAs allosterically modify RNA-binding proteins in cis to inhibit transcription. *Nature,* 454**,** 126-30.

WANG, Y.-P., TIAN, Y., ZHU, J.-H., YANG, Y.-F., ZHANG, H.-B., WANG, C.-H., LIU, L., LV, Y. & XIONG, L.-P. 2008b. Study on the association between HLA-DRB1 genes and ADHD in Xi'an. *Chinese Journal of Child Health Care,* 16**,** 010.

WANG, Y., SUZUKI, H., YOKOO, T., TADA-IIDA, K., KIHARA, R., MIURA, M., WATANABE, K., SONE, H., SHIMANO, H., TOYOSHIMA, H. & YAMADA, N. 2004. WGEF is a novel RhoGEF expressed in intestine, liver, heart, and kidney. *Biochem Biophys Res Commun,* 324**,** 1053-8.

WATKINS, K. E., DRONKERS, N. F. & VARGHA-KHADEM, F. 2002a. Behavioural analysis of an inherited speech and language disorder: comparison with acquired aphasia. *Brain,* 125**,** 452-64.

WATKINS, K. E., VARGHA-KHADEM, F., ASHBURNER, J., PASSINGHAM, R. E., CONNELLY, A., FRISTON, K. J., FRACKOWIAK, R. S., MISHKIN, M. & GADIAN, D. G. 2002b. MRI analysis of an inherited speech and language disorder: structural brain abnormalities. *Brain,* 125**,** 465-78.

WECHSLER, D. 1974. *Wechsler Intelligence Scale for Children, revised,* San Antonio, TX, Psychological Corporation.

WECHSLER, D. 1981. *Wecshler Adult Intelligence Scale, revised,* San Antonio, TX, Psychological Corporation.

WECHSLER, D. 1989. *Wechsler Pre-school and Primary Scale of Intelligence, revised,* San Antonio, TX, Psychological Corporation.

WECHSLER, D. 1991. *Wechsler Intelligence Scale for Children - Third edition,* San Antonio, TX, Psychological Corporation.

WECHSLER, D. 1992. *Wechsler Intelligence Scale for Children - Third UK Edition,* London, Psychological Corporation.

WECHSLER, D. 1993. *Escala Wechsler de Inteligencia para niños revisada,* Madrid, TEA ediciones.

WECHSLER, D. 1997. *Wecshler Adult Intelligence Scale - Third edition,* San Antonio, TX, Psychological Corporation.

WECHSLER, D. 1999. *Wechsler Abbreviated Scale of Intelligence,* San Antonio, TX, Psychological Corporation.

WEINBERG, C. R., WILCOX, A. J. & LIE, R. T. 1998. A log-linear approach to case-parent-triad data: assessing effects of disease genes that act either directly or through maternal effects and that may be subject to parental imprinting. *Am J Hum Genet,* 62**,** 969-78.

WEXLER, K. 1994. Optional infinitives, head movement and the economy of derivations. *In:* LIGHTFOOT, D. & HORNSTEIN, N. (eds.) *Verb movement.*

WEYN-VANHENTENRYCK, S. M., MELE, A., YAN, Q., SUN, S., FARNY, N., ZHANG, Z., XUE, C., HERRE, M., SILVER, P. A., ZHANG, M. Q., KRAINER, A. R., DARNELL, R. B. & ZHANG, C. 2014. HITS-CLIP and integrative modeling define the Rbfox splicing-regulatory network linked to brain development and autism. *Cell Rep,* 6**,** 1139-52.

WHITEHOUSE, A. J., BARRY, J. G. & BISHOP, D. V. M. 2007. The broader language phenotype of autism: a comparison with specific language impairment. *J Child Psychol Psychiatry,* 48**,** 822-30.

WHITEHOUSE, A. J., BISHOP, D. V. M., ANG, Q. W., PENNELL, C. E. & FISHER, S. E. 2011. CNTNAP2 variants affect early language development in the general population. *Genes Brain Behav,* 10**,** 451-6.

WIEDERHOLT, J. L. & BRYANT, B. R. 1992. *Gray Oral Reading Tests, Third edition,* Austin, TX, Pro-ed.

WIGGINTON, J. E. & ABECASIS, G. R. 2005. PEDSTATS: descriptive statistics, graphics and quality assessment for gene mapping data. *Bioinformatics,* 21**,** 3445-7.

WILKINSON, L. S., DAVIES, W. & ISLES, A. R. 2007. Genomic imprinting effects on brain development and function. *Nat Rev Neurosci,* 8**,** 832-43.

WILLIAMS, J. & O'DONOVAN, M. C. 2006. The genetics of developmental dyslexia. *Eur J Hum Genet,* 14**,** 681-9.

WINNEY, B., BOUMERTIT, A., DAY, T., DAVISON, D., ECHETA, C., EVSEEVA, I., HUTNIK, K., LESLIE, S., NICODEMUS, K., ROYRVIK, E. C., TONKS, S., YANG, X., CHESHIRE, J., LONGLEY, P., MATEOS, P., GROOM, A., RELTON, C., BISHOP, D. T., BLACK, K., NORTHWOOD, E., PARKINSON, L., FRAYLING, T. M., STEELE, A., SAMPSON, J. R., KING, T., DIXON, R., MIDDLETON, D., JENNINGS, B., BOWDEN, R., DONNELLY, P. &

BODMER, W. 2012. People of the British Isles: preliminary analysis of genotypes and surnames in a UK-control population. *Eur J Hum Genet,* 20**,** 203-10.

WISZNIEWSKI, W., HUNTER, J. V., HANCHARD, N. A., WILLER, J. R., SHAW, C., TIAN, Q., ILLNER, A., WANG, X., CHEUNG, S. W., PATEL, A., CAMPBELL, I. M., GELOWANI, V., HIXSON, P., ESTER, A. R., AZAMIAN, M. S., POTOCKI, L., ZAPATA, G., HERNANDEZ, P. P., RAMOCKI, M. B., SANTOS-CORTEZ, R. L., WANG, G., YORK, M. K., JUSTICE, M. J., CHU, Z. D., BADER, P. I., OMO-GRIFFITH, L., MADDURI, N. S., SCHARER, G., CRAWFORD, H. P., YANATATSANEEJIT, P., EIFERT, A., KERR, J., BACINO, C. A., FRANKLIN, A. I., GOIN-KOCHEL, R. P., SIMPSON, G., IMMKEN, L., HAQUE, M. E., STOSIC, M., WILLIAMS, M. D., MORGAN, T. M., PRUTHI, S., OMARY, R., BOYADJIEV, S. A., WIN, K. K., THIDA, A., HURLES, M., HIBBERD, M. L., KHOR, C. C., VAN VINH CHAU, N., GALLAGHER, T. E., MUTIRANGURA, A., STANKIEWICZ, P., BEAUDET, A. L., MALETIC-SAVATIC, M., ROSENFELD, J. A., SHAFFER, L. G., DAVIS, E. E., BELMONT, J. W., DUNSTAN, S., SIMMONS, C. P., BONNEN, P. E., LEAL, S. M., KATSANIS, N., LUPSKI, J. R. & LALANI, S. R. 2013. TM4SF20 ancestral deletion and susceptibility to a pediatric disorder of early language delay and cerebral white matter hyperintensities. *Am J Hum Genet,* 93**,** 197-210.

WOODCOCK, R. 1987. *Woodcock Reading Mastery Tests, Revised,* Circle Pines, MN, American Guidance Service.

WOODCOCK, R. 1998. *Woodcock Reading Mastery Tests, Revised,* Circle Pines, MN, American Guidance Service.

WRAY, G. A., HAHN, M. W., ABOUHEIF, E., BALHOFF, J. P., PIZER, M., ROCKMAN, M. V. & ROMANO, L. A. 2003. The evolution of transcriptional regulation in eukaryotes. *Mol Biol Evol,* 20**,** 1377-419.

WRIGHT, B. A., LOMBARDINO, L. J., KING, W. M., PURANIK, C. S., LEONARD, C. M. & MERZENICH, M. M. 1997. Deficits in auditory temporal and spectral resolution in language-impaired children. *Nature,* 387**,** 176-8.

WRIGHT, P., DONALDSON, P. T., UNDERHILL, J. A., CHOUDHURI, K., DOHERTY, D. G. & MURRAY, R. M. 1996. Genetic association of the HLA DRB1 gene locus on chromosome 6p21.3 with schizophrenia. *Am J Psychiatry,* 153**,** 1530-3.

ZEESMAN, S., NOWACZYK, M. J., TESHIMA, I., ROBERTS, W., CARDY, J. O., BRIAN, J., SENMAN, L., FEUK, L., OSBORNE, L. R. & SCHERER, S. W. 2006. Speech and language impairment and oromotor dyspraxia due to deletion of 7q31 that involves FOXP2. *Am J Med Genet A,* 140**,** 509-14.

ZEMBRZYCKI, A., GRIESEL, G., STOYKOVA, A. & MANSOURI, A. 2007. Genetic interplay between the transcription factors Sp8 and Emx2 in the patterning of the forebrain. *Neural Dev,* 2**,** 8.

ZEREM, A., VINKLER, C., MICHELSON, M., LESHINSKY-SILVER, E., LERMAN-SAGIE, T. & LEV, D. 2011. Mosaic marker chromosome 16 resulting in 16q11.2-q12.1 gain in a child with intellectual disability, microcephaly, and cerebellar cortical dysplasia. *Am J Med Genet A,* 155A**,** 2991-6.

ZHENG, J. Q. & POO, M. M. 2007. Calcium signaling in neuronal motility. *Annu Rev Cell Dev Biol,* 23**,** 375-404.

ZHOU, X., TANG, W., GREENWOOD, T. A., GUO, S., HE, L., GEYER, M. A. & KELSOE, J. R. 2009. Transcription factor SP4 is a susceptibility gene for bipolar disorder. *PLoS One,* 4**,** e5196.

ZIATS, M. N. & RENNERT, O. M. 2013. Aberrant expression of long noncoding RNAs in autistic brain. *J Mol Neurosci,* 49**,** 589-93.

ZIMMERMAN, I. L., STEINER, V. G. & POND, R. E. 1992. *Preschool Language Scale-3,* San Antonio, Texas, Psychological Corporation.

Ron Nudel

Lincoln College, University of Oxford

Molecular Genetics of Language Impairment

Corrigenda October 30 2015

Chapter 1 Section 1.4.4: "…but it is not clear whether those differences could account for the opposite association trends with CMIP SNPs…" should read: "…but it is not clear whether other population differences may exist that could account for the opposite association trends with CMIP SNPs…".

Chapter 2 Section 2.1.1: The cut-off value for sample exclusion based on the heterozygosity rate was ±3 SD from the mean, and not ±2 SD from the mean.

Chapter 2 Section 2.6.1: The section dealing with affection statuses should read:
Affection status was based on the expressive and receptive language scores from the CELF: probands and any siblings with ELS or RLS of 1.5 SD below the general-population mean for their age (or ELS or RLS lower than that value) were defined as cases (mean=100, and SD=15); siblings whose ELS and RLS were higher than the score with a value of 0.5 SD below the mean were defined as controls; all other individuals were given an "unknown" affection status. EMIM uses cases, case duos (case and one parent), and case trios (case and two parents) in its analyses. EMIM uses every individual only once, and, therefore, the subsets of the families used in the analyses are independent. Individuals who had "unknown" affection statuses, controls, and their parents (unless they were in a subset with a case) were not used in the analyses due to the fact that EMIM assumes that individuals with a missing phenotype may be considered controls, which holds true only for rare diseases.

Chapter 2 Section 2.6.1: The section dealing with quality control steps for the genotype data should read:
The raw genotype data were processed with the Genome Studio software (Illumina). The quality control measures for the SNPs used in the EMIM analyses are as follows: 47 samples

were duplicated across plates (concordance rate of 0.98968), and all the samples were randomised across plates with cases and controls being spread evenly across plates, as were different sample types (saliva, blood, or mouth swab samples) and samples of different origins (the various UK centres which supplied the samples). SNPs with a Gentrain score below 0.5 were removed. Before performing the association analyses, SNPs and samples were filtered based on several quality control measures, as described in (Anderson *et al.*, 2010): the success thresholds for genotyping for both SNPs and samples were 95%, and SNPs with a minor allele frequency of less than 1% were removed. SNPs and samples with an error rate of 1% or higher, as estimated by bad inheritances, were removed (this was confirmed using PEDSTATS). SNPs were filtered with PLINK based on a HWE p-value exclusion threshold of $P \leq 10^{-6}$. Genotypes were compared with existing SNP genotype data for 700 individuals and 734 SNPs, with an error rate of 0.00967. The total number of SNPs used in the EMIM analyses was 614,937. Following the generation of p-values, SNPs that showed trends of association were manually checked for good clustering with Genome Studio i.e. SNP clustering was checked to make sure there were three distinct clusters (for genotypes AA, AB, and BB). SNPs with association p-values ≤0.0001 were checked for HWE using PEDSTATS, which selects unrelated samples in a different way from that used by PLINK, resulting in a more powerful test. SNPs with HWE P≤0.001 were then removed.

Chapter 2 Section 2.6.6: The section dealing with quality control steps for the genotype data should read:

The raw genotype data were processed with the Affymetrix Genotyping Console software. As recommended by Affymetrix, samples with a dish quality score of less than 0.82 were removed. The dish quality score is a measure of the extent of which signal values are separated from background values across pre-selected sites with no polymorphisms, with 0 indicating no separation, and 1 indicating a perfect separation. Markers and samples with call rates below 95% (per either the LAT array or the custom array) were removed. The genotype data were examined with the Affymetrix SNPolisher script, which provides several quality scores for the clustering, including: Fisher's linear discriminant, heterozygous cluster strength offset, and homozygote ratio offset. As recommended in the Affymetrix guidelines, markers were removed if their scores did not reach the following threshold: 3.6 (Fisher's linear

discriminant), -0.1 (heterozygous cluster strength offset), 0.3 (homozygote ratio offset, if the marker had fewer than three clusters), or -0.9 (homozygote ratio offset, if the marker had three clusters). In the custom array, several markers had more than one probe designed for them. In such cases, one probe was selected based on information supplied by Affymetrix. Markers that were SNPs were used in further quality control procedures with PLINK. Samples and SNPs with a Mendelian error rate of more than 1% were removed. SNPs with HWE $P \leq 10^{-6}$ as estimated with PLINK were removed. Samples were checked for IBD and discordant sex information, and the pedigree was amended where necessary.

In the targeted association analysis of the linkage region on chromosome 7, SNPs were tested in PEDSTATS for their HWE p-values, and SNPs with HWE p-values of less than 0.001 (based on 19 unrelated individuals from the large pedigree) were removed. QVALUE was used to generate q-values for the p-values.

Chapter 5 Introductory Section: The section dealing with affection statuses should read: The affection statuses used in the targeted association analysis described in this chapter were the same as the ones used in the linkage study, as was the sample. In the association analysis for the exome sequencing variants, also described in this chapter, affection statuses were determined in the following manner: probands were diagnosed with SLI strictly on the basis of their language test scores and exclusionary criteria (as described in Chapter 1); children aged 6 or younger with scores of >2 SD below those expected on the phonology test and children above the age of 6 with performance of two years below that expected for their chronological age on the phonology test, as well as children who performed below the 10th percentile on either the expressive or the receptive scales of the grammar test were classified as having SLI, if their non-verbal IQ was not below the 10th percentile, and they had normal hearing, oral-motor skills and neurological development. Children whose performance on the phonology test was not 2 SD below expected or two years below expected for their chronological age, and whose performance on the grammar test was above the 10th percentile on both the expressive and receptive scales were classified as unaffected i.e. as having a typical language development. Other children e.g. children having language problems and low IQ or other developmental problems, were excluded from this study.

Parents and siblings of probands were diagnosed with SLI if they performed poorly (below the 10th percentile) on one of the language tests administered to them (as outlined in Chapter 1) and/or if they reported a clinical history of language problems. They were considered unaffected if they performed above the 10th percentile on both language tests and had no history of language problems.

Chapter 7: Throughout this chapter, “temporal frontal lobe” should be “temporal lobe”.