Genetic studies of bipolar disorder and recurrent major depression in a large Scottish family

Lorna M. Houlihan

BA (Genetics) *Trinity College Dublin, Ireland*

MSc (Science by Research) *University of Edinburgh, UK*

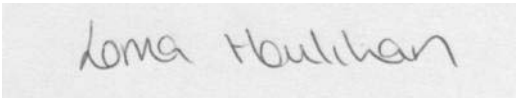
A thesis submitted for the degree of Doctor of Philosophy

The University of Edinburgh

2008

Declaration

I hereby declare that all work described in this thesis, unless otherwise acknowledged, has been performed by myself and has not been accepted for any previous application for a degree. The information obtained from sources other than this study is acknowledged in the text or included in the references.



_____

Lorna M. Houlihan

For my family

with thanks

Acknowledgements

My primary thanks go to the patients and the family, who have participated in this study. I would also like to thank the Wellcome Trust for their funding and the Wellcome Trust board in Edinburgh for their encouragement.

Firstly, in the laboratory, I would like to thank Susan Anderson, Helen Torrance and Dr. Margaret McLean for help in preparing DNA samples, protein samples and technical advice. I would like to thank Dr. Margaret McLean for her help in establishing the Taqman allele-specific assays and Angie Fawkes and Lee Murphy at the Wellcome Trust-Clinical Research Facility for advice on Taqman assays and Illumina linkage genotyping. For my protein work, I am grateful to Dr. Shane Minogue and Dr. Tamas Balla for *PI4K2B* material. Furthermore, I would like to thank Dr. Pippa Thomson, Dr. Shaun Mackie, Dr. Jennifer Chubb, Dr. Ben Pickard, Dr. Fumiaki Ogawa, Helen Knight, Sarah Whittal and all members of the Medical Genetics Section for discussion, advice and molecular biology materials. Of course, I would like to thank Rosemary, Susan, Helen and Heather for the smooth running of my studies.

Secondly, for the statistical and computational aspect of my project, I would like to acknowledge authors of the different programmes who helped me immensely in executing their software: Dr. Dan Weeks, Dr. Goncalo Abecasis, Dr. Mario Falchi and Dr. Bill Duren. I would also like to thank Stewart Morris for formatting data, troubleshooting software programmes and running the linkage permutation analysis. Advice on SNP selection and association study analysis was greatly received from Dr. Andrea Christoforou.

Finally, I would like to thank my supervisors, Dr. Kathy Evans and Prof. Douglas Blackwood, for their direction, help and encouragement and Prof. David Porteous for his support, advice and persuasion. My friends have been fantastic, at home in Ireland: Gwyneth, Máire-Bríd, Sharon, Cathy & Micky, here in Edinburgh: Jenn & Andrea (the greatest PhD pals ever!), Vicky, Helen Miranda, Josefin, Katie, Linsay, Christine, Elaine, Stefano & Ita and my new office-mates: Michelle, Lars & Dave. I have dedicated this thesis to my family: Mum, Dad, Nuala, Donal and Elaine, and my Granny, who have loved, supported and inspired me. And, I am eternally indebted to Manuel for motivating me, with strength and courage in my studies and making me very happy!

Abstract

Genetic studies of bipolar disorder and recurrent major depression in a large Scottish family

Bipolar disorder and recurrent major depression are complex psychiatric illnesses with a substantial, yet unknown genetic component. Genetic studies have identified linkage of bipolar disorder and recurrent major depression with markers on chromosome 4p15-p16 in a large Scottish family and three smaller families. To focus the search for genetic factors for susceptibility to illness two approaches were adopted: a chromosome 4p15-p16 candidate gene study and a whole-genome linkage scan.

In the first instance, phosphatidylinositol 4-kinase type-II beta (*PI4K2B*) was selected as a candidate gene. Analysis of haplotypes in the four linked families identified two regions, both of which were shared by three families. *PI4K2B* lies within one of these regions. *PI4K2B* is also a worthy functional candidate as it is a member of the phosphatidylinositol pathway, which is targeted by lithium for therapeutic effect in bipolar disorder. Expression studies at the allele-specific mRNA and protein level were performed in lymphoblastoid cell lines from the large Scottish family. There was no evidence for expression differences between affected and non-affected family members. However, a case-control association study showed preliminary evidence for association of schizophrenia but not bipolar disorder, with tagging single nucleotide polymorphisms from the *PI4K2B* genomic region.

Second, the linkage evidence for bipolar disorder and recurrent major depression in the large Scottish family was re-examined. This was important because additional family members had been recruited and advances in technology made it feasible to cover all chromosome regions more densely than had been possible ten years ago. Stringent genotyping and pedigree error checks were performed to ensure an optimised dataset for analysis. Furthermore, the large family was divided in an informative manner for ease of analysis using both parametric and non-parametric methods, supplemented by haplotype analysis. Genome-wide significant evidence for linkage was observed on chromosome 4p15-p16 and genome-wide suggestive evidence was observed on chromosomes 8p21 and 1p36. The analysis clearly supports the evidence for a susceptibility locus of bipolar disorder and recurrent major depression on chromosome 4p15-p16, while identifying other genetic loci that may confer risk to psychiatric illness.

Publications from this Thesis

Papers

Le Hellard S, Lee AJ, Underwood S, Thomson PA, Morris SW, Torrance HS, Anderson SM, Adams RR, Navarro P, Christoforou A, **Houlihan LM**, Detera-Wadleigh S, Owen MJ, Asherson P, Muir WJ, Blackwood DH, Wray NR, Porteous DJ, Evans KL (2007) "Haplotype analysis and a novel allele sharing method refines a chromosome 4p locus linked to bipolar affective disorder." Biological Psychiatry **61**(6):797-805

Abstracts and oral presentations

Lorna M Houlihan, David Porteous, Douglas Blackwood, Kathryn Evans (2006) "Investigations into a chromosome 4p bipolar disorder susceptibility gene; Allelic expression assays and an association study in PI4K2B" American Journal of Medical Genetics 2006, 141B (1) P83

Lorna Houlihan, Margaret McLean, Walter Muir, David Porteous, Douglas Blackwood, Kathryn Evans "Analysis of a phosphoinositol kinase as a candidate bipolar disorder susceptibility gene on chromosome 4p" (2006) Bipolar Disorders 8 (S1) 63

Lorna Houlihan, Margaret McLean, Pippa Thompson, Andrea Christoforou, Walter Muir, David Porteous, Douglas Blackwood, Kathryn Evans "Analysis of PI4KIIβ, a Candidate Gene from the Bipolar Disorder Susceptibility Locus on Chromosome 4p15-16". (2005) American Journal of Medical Genetics 138B (1) 86

Abbreviations

+ve	Positive
-ve	Negative
bp	Base Pairs
BP	Bipolar Disorder
cM	CentiMorgans
Ctrl	Control
DF	Degrees of Freedom
DSM	Diagnostic and Statistical Manual of Mental Diseases
FS	Full-Sibling
gDNA	Genomic DNA
H_2O	Water
HLOD	Heterogeneous LOD
HS	Half-Sibling
HWE	Hardy Weinberg Equilibrium
IBD	Identical by Descent
IBS	Identical by State
IC	Information Content
LCL	Lymphoblastoid Cell Lines
LD	Linkage Disequilibrium
LOD	Logarithm of Odds
LRS	Likelihood Ratio Score
MDA	Multiple Displacement Action
Min	Minutes
MZ	Mono-zygotic twin
NAD	Non Affected with Psychiatric Illness
NPL	Non Parametric Linkage Analysis
OPA	Oligo Pool All
PBS	Phosphate Buffered Saline
PI	Phosphoinositide
PL	Parametric Linkage Analysis
PO	Parent Offspring
qs	*quantum sufficiat,* indicates that liquid is added to a given final volume.
RACE	Rapid Amplification of cDNA Ends
RMD	Recurrent Major Depression
RT	Room Temperature (15-25°C)
RT-PCR	Reverse Transcription-Polymerase Chain Reaction
SADS-L	Schedule for Affective Disorder and Schizophrenia Lifetime version
Soln	Solution
TE	10mM Tris-HCl, 1mM EDTA, pH7.5 Buffer
UCSC	University of California Santa Cruz
µl	Microlitre
µM	Micromolar
WGA	Whole Genome Amplified
WHO	World Health Organisation
WTCCC	Wellcome Trust Case Control Consortium
WT-CRF	Wellcome Trust Clinical Research Facility

Contents Page

List of Figures

List of Tables

Chapter 1
Introduction

1. Introduction

1.1. *Bipolar Disorder & Recurrent Major Depression*

1.1.1. Diagnosis of the illness

Bipolar disorder is a distressing psychiatric illness. It was originally termed "manic-depression" to encompass the primary features of the illness: separate episodes of mood elevation "mania" and deep disabling depression. The periods of mania are marked by feelings of elation, euphoria, irritation and increased activity, whereas feelings of sadness, anxiety, guilt, anger and hopelessness are common in depression episodes. The duration of the depression periods are longer and more frequent than the manic episodes. The illness is defined by these episodes of extreme behaviour, which are often inappropriate and unpredictable in nature. Accompanying characteristics of the mood disturbances are psychosis, cognitive impairment, increasing creative energy, sleep and other periodicity problems (ISBD 2006). Undoubtedly, bipolar disorder impacts negatively on quality of life (World Health Organization 2005).

The criteria for diagnosing mental disorders are defined in the DSM-IV (Diagnostic and Statistical Manual of Mental Diseases) from the American Psychiatric Association (American Psychiatric Association 2000) and the International Classification of Diseases (ICD-10) of the World Health Organisation (WHO) (World Health Organization. 1994). In DSM-IV, bipolar disorder is subdivided into two classes, bipolar I disorder and bipolar II disorder. In bipolar I disorder, the mania episodes consist of mood elevation persisting for longer than one week with social impairment. In bipolar II disorder, patients suffer shorter periods of hypomania with mood elevation for at least four days. Using DSM-IV or ICD-10 criteria ensures that diagnoses of mental disorders are reliable. However, the validity of the diagnoses is unknown as there are no biological diagnostic markers for the illness.

In particular there are no diagnostic markers that will be considered for inclusion in the newly proposed DSM-V (Hyman 2007).

Patients with recurrent major depression, sometimes referred to as unipolar disorder, experience episodes of depression but no mania symptoms. Individuals diagnosed with recurrent major depression suffer from the same characteristics associated with the depression episodes in individuals with bipolar disorder.

Yet another mental illness, namely schizophrenia is also discussed in this study. It is regarded as a separate illness to bipolar disorder and recurrent major depression, with positive symptoms which are psychological phenomena that do not occur in healthy individuals, such as hallucinations and delusions, and negative symptoms which arise from deficits in normal functions, such as lack of motivation and blunt emotional responses. Cognitive deficits, including impairments in working memory and executive functions, are found in both schizophrenia and bipolar disorder (Hyman 2007).

1.1.2. Prevalence of the illness

The lifetime prevalence of bipolar disorder in the general population is 0.5-1.5% for both males and females, as shown in Table 1.1 and the mean age of onset is estimated to be 21 years (Smith and Weissmann 1992). Recurrent major depression is more common in the population, but there are varying reports of the prevalence. The large US multi-site Epidemiological Catchment Area (ECA) study reported a lifetime population prevalence for DSM-III major depression of 4.4% (Weissman, Leaf et al. 1988) whereas the US National Comorbidity Survey estimated the lifetime prevalence of DSM-III major depression to 17% (Kessler, McGonagle et al. 1994). In addition, the lifetime rate for recurrent major depression in women (21.3%) is twice that for men (12.7%) (Kessler, McGonagle et al. 1994). Nevertheless, it is widely

regarded that recurrent major depression is common in the population (McGuffin, Owen et al. 2002).

RELATIONSHIP TO AFFECTED BIPOLAR DISORDER	RISK OF BIPOLAR DISORDER (%)	ADDITIONAL RISK OF RECURRENT MAJOR DEPRESSION
Monozygotic co-twin	40-70	15-25
First degree relation	5-10	10-20
General population	0.5-1.5	5-10

Table 1.1 The approximate lifetime rates for bipolar disorder and recurrent major depression. This table is adapted from (McGuffin, Owen et al. 2002). The lifetime risk for a mood disorder in a relation is obtained by adding the risk of bipolar disorder and the risk of recurrent major depression.

Table 1.1 summarises classical genetic studies of bipolar disorder to show that there is an increase in risk of bipolar disorder and recurrent major depression with an affected relation; the risk of a bipolar disorder diagnosis is 5-10% when the individual has an affected first degree relation and 40-70% for an affected monozygotic twin. One particular study of 2.1 million individuals from the Danish population, of whom 2,229 were diagnosed with bipolar disorder, showed a nearly 14-fold increase in risk of bipolar disorder, when an individual had a first-degree relation with bipolar disorder (Mortensen, Pedersen et al. 2003). The risk of a recurrent major depression is 10-20% for those with a first degree relation with bipolar disorder and 15-25% for a monozygotic twin with bipolar disorder.

Twin and adoption studies also demonstrate heritability of bipolar disorder, and to some degree in recurrent major depression. The largest bipolar disorder twin study used the Danish Twin Register and found that the probandwise concordance (the proportion of proband twins with bipolar disorder who had a twin with bipolar

disorder) in monozygotic twins was 0.62, and for dizygotic twins was 0.08 (Bertelsen, Harvald et al. 1977). As there was not full concordance (100%) in monozygotic twins, it suggests the existence of "environmental" risk factors in addition to genetic risk factors. While the estimation of a genetic component in recurrent major depression is not so clear-cut due to unknown diagnostic validity, ascertainment problems and gender differences, a meta-analysis of twin studies illustrates a significant additive genetic effect and an estimated heritability or recurrent major depression of 37% (95%CI 31-42%) (Sullivan, Neale et al. 2000).

Further evidence supporting genetic risk factors for bipolar disorder has been provided by adoption studies. Adoption studies are useful to distinguish post-natal environment and genetic risk factors, as the affected individuals inherit genetic risk factors from their biological parents but their environmental exposure from a different family. A bipolar disorder adoption study showed that the biological relations of bipolar disorder adoptees were at a greater risk for bipolar disorder than were the adoptive relations. The risk in biological relations of bipolar adoptees was similar to that in the biological relations of bipolar non-adoptees (Mendlewicz and Rainer 1977).

1.1.1. Treatment for the illness

The WHO publishes a list of essential medicines that satisfy the priority health care needs of the population (WHO Model List of Essential Medicines, 15th edition, revised March 2007). The core list presents an index of the minimum medicine needs for a basic health care system, cataloguing the most efficacious, safe and cost-effective medicines for priority conditions. Priority conditions are selected on the basis of current and estimated future public health relevance, and potential for safe and cost-effective treatment. For the treatment of bipolar disorder, the WHO recommends lithium carbonate and the anticonvulsants: carbamazepine and valproate. For the treatment of depression, the tricyclic antidepressant, amitriptyline

and the selective serotonin reuptake inhibitor (SSRI) class antidepressant, fluoxetine are recommended. However, the mood-stabilising treatments are often not fully effective, have many reported adverse side-effects and can be toxic at doses close to the therapeutic range (Moller and Nasrallah 2003).

1.2. Genetic Analysis

In the following section, I would like to introduce basic genetic concepts. These are important to understand when genetic mapping studies for bipolar disorder are discussed. Ensuing this, I will present a more detailed explanation of case-control association analysis and linkage analysis. While it is not essential reading to introduce this study, readers will find it helpful as it directly introduces methods used in chapters 4, 5 and 6 of this thesis.

1.2.1. Genetic concepts

The concept behind the two genetic mapping approaches, linkage and association, is shown in Figure 1.1. Association compares the frequency of the markers between the cases with the disease and the unaffected controls. For example, marker "C" occurs more frequently in cases than the controls in Figure 1.1a. Linkage analysis searches for two loci that are inherited together, for example the disease and marker "C" in Figure 1.1b. Both approaches search for the same genetic determinant for disease, however there are notable differences between the two methods. Association is observed in unrelated individuals, whereas linkage is only in related individuals. Also, association only occurs when the marker is extremely close to the actual disease locus, whereas linkage is detectable with markers far away from the disease locus.

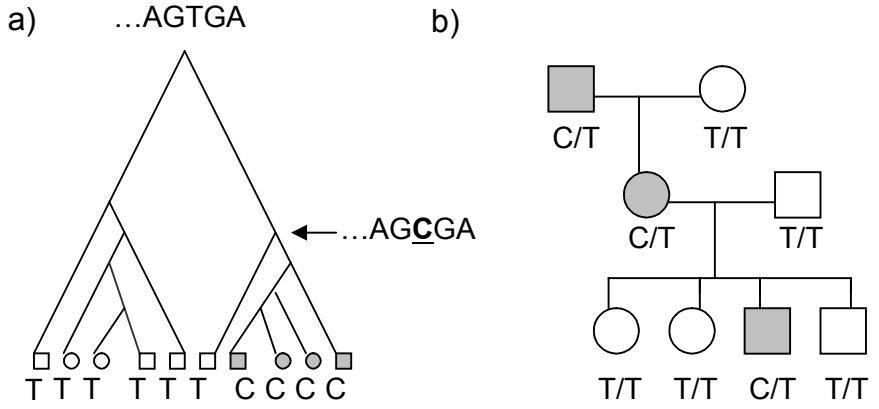


Figure 1.1 Association and linkage mapping. Association mapping is illustrated in a. The affected individuals share the C allele identical by descent (IBD) from a common ancestor in the past marked by the arrow. Linkage mapping is illustrated in (b) and shows that the affected individuals share the C allele IBD from a common ancestor, three generations in the past (Personal communication, Chad Garner, 23rd July 2006, University of California, US (Weeks, Lathrop et al. 2006)).

1.2.2. Hardy-Weinberg equilibrium

An important feature of genetic studies is Hardy-Weinberg equilibrium, which ensures the markers under investigation for linkage or association analysis behave in a non-biased manner. Hardy–Weinberg distribution describes the simple relationship between gene frequencies and genotype frequencies that is found in a population under certain conditions. In the presence of random mating, the alleles at a neutral locus are independent such that the genotype frequencies are given by the product of the allele frequencies. For example, in a biallelic marker having alleles A and a,

- the allele frequencies are Pr(A)=p and Pr(a)=q=(1-p)
- the genotype frequencies are Pr(AA)=p^2, Pr(Aa)=2pq, Pr(aa)=q^2
- which sum to one p^2+2pq+q^2=1

Hardy Weinberg distribution can break down due to a variety of reasons including systematic genotyping errors, inbreeding, random sampling, sampling of affected individuals under certain disease models or chance occurrence in small populations (Sham 1998).

1.2.3. Linkage disequilibrium

Another important characteristic is linkage disequilibrium (LD). LD, also called allelic association, is the non-random association of markers at different loci, whether or not the loci are physically linked. Markers are in LD if they are inherited together more often than expected by chance. LD can be created through mutation, drift in small populations, population bottlenecks, founder effects or selection. Conversely, LD decays as a function of the recombination rate, the effective population size, time and mutation. There are conventionally three ways to measure LD between markers: D', r^2 and LOD. The co-efficient D is, in short, the equivalent to the co-variance between loci (Lewontin 1964). D' is the normalised measure of Lewontin and is the absolute value of the difference between observed and expected probabilities. An advantage of this method is that it is related to recombination rate; however a disadvantage is that it is biased in small sample sizes. A D' value of one means complete LD between two markers, such that there is no evidence for an ancestral recombination event between two markers. r^2 looks at the statistical association between markers by the correlation coefficient between markers. One advantage of r^2 is that it is more reliable than D' for markers with low allele frequencies. An r^2 of one means perfect LD between two markers, no recombination between the markers or they have the same allele frequencies. Also, either of the markers are redundant and can act as "perfect proxies" to each other (Pritchard and Przeworski 2001). The third measure is the logarithm of the odds (LOD score) for linkage disequilibrium between a given marker pair, and is commonly used in HapMap (section 1.3.6) as an LD measure (www.hapmap.org).

1.2.4. Identity by descent

The labelling of genetic markers as identical by descent (IBD) or identical by state (IBS) is an important concept in genetics. Related individuals are similar to each other because they have variants that are IBD, which are copies of the same variant from a common ancestor, as illustrated with the red circles in Figure 1.2. IBD

variants are of the same allelic type whereas non-IBD variants are of independent types and are illustrated by the purple circles in Figure 1.2. For alleles that are the same, but it is unclear whether they are IBD, are termed identical by state (IBS), shown by the green circles in Figure 1.2. To date, many useful applications of IBD estimation have been devised, for example: relatedness analysis, case-control association studies and non-parametric linkage analysis and are discussed in the upcoming sections.

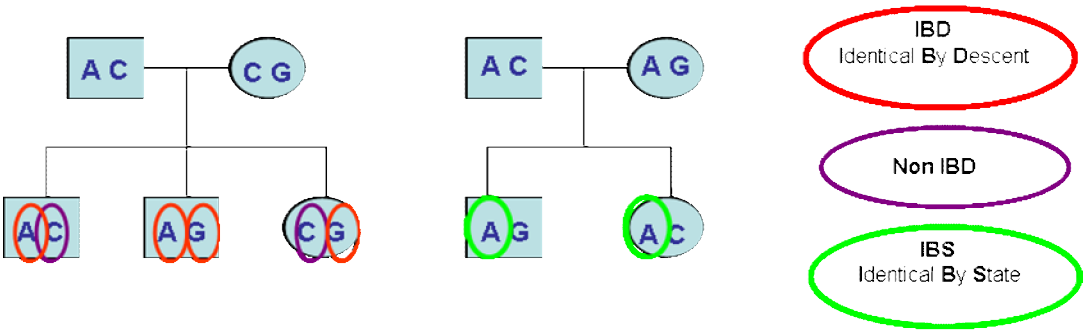


Figure 1.2 Allele-sharing. Definition of identity by descent (IBD) and identity by state (IBS) is illustrated with two pedigrees on a three-allele (A, C, G) marker. IBD alleles are highlighted in red, IBS alleles in green and non-IBD alleles in purple.

1.2.5. Relatedness analysis

The estimation of IBD and IBS is central to relatedness analysis between individuals in a family. A pedigree or relationship determines probabilities of IBD, which then determines probabilities of joint genotypes. This is equivalent to similarity among relations. The identification of alleles that are IBS is also informative over a large number of markers, shared by a pair of individuals to detect full-siblings, half-siblings and unrelated individuals (Ehm and Wagner 1998). IBD sharing between related individuals is measured as the probability of a relationship-pair sharing 0, 1 or 2 markers IBD ($\kappa 0$, $\kappa 1$, $\kappa 2$). The values for each relationship-pair are detailed in Table 1.2.

PAIRWISE RELATIONSHIP	K_0	K_1	K_2	Ψ
Unrelated	1.00	0	0	0
Parent-Offspring	0	1.00	0	0.25
Monozygous twin	0	0	1.00	0.5
Full sibling	0.25	0.5	0.25	0.25
Half sib, grandparent, aunt	0.5	0.5	0.00	0.125
First cousin	0.75	0.25	0	0.0625
Double first cousin	0.5625	0.375	0.0625	0.125
Quadruple half first cousin	0.5312	0.4375	0.0312	0.125

Table 1.2 Identity by descent sharing relationship between relations. κ is the probability of 0, 1, or 2 markers shared identical by descent (IBD) (pi = Pr (i genes IBD), $\kappa_2 + \kappa_1 + \kappa_0 = 1$.ψ is the kinship coefficient which measures IBD between two markers. This table is adapted from (Thompson 2000; Weir, Anderson et al. 2006)

ψ is the kinship coefficient which measures IBD between two markers. Kinship coefficients can be theoretical, derived from assumed levels of marker sharing in a pedigree as shown in Table 1.2, or conditional, which use marker data. The theoretical kinship coefficient ψ ij is the probability that a randomly sampled gene from j is IBD to a randomly sampled gene from the same arbitrary locus of i. For example, ψ ij = 1/2 if i = j and ψ ij = 1/4 if i and j are first degree relatives. In both cases, no inbreeding is allowed.

1.3. Case-Control Association Studies

1.3.1. Concept of association studies

In an influential paper, Risch and Merikangas advised that the "future of genetics requires large-scale testing by association analysis" (Risch and Merikangas 1996). Case-control association studies are commonly used to detect variants causing susceptibility to common complex disorders. In the present day literature,

association studies fall into two categories; candidate gene studies and whole-genome association. Candidate gene studies are hypothesis-driven, that focus on a particular biological function or area of the genome. Whole genome association studies are performed without a prior hypothesis, as susceptibility genes are often not obvious biological candidates, and comprehensively cover the whole genome with a dense marker set. However, those whole genome association studies performed searching for common bipolar disorder variants have met with a mixed success rate, as will be described in section 1.5.3.

1.3.2. Methods of association analysis

Figure 1.3 depicts the calculations used for association study analysis methods, utilising χ^2 contingency models at the allele level and genotype level, incorporating the dominant, recessive and additive models (Lewis 2006). An additive model (Armitage's test for trend) is specified by assigning to the heterozygous genotype, a value that is halfway between the values of the two homozygous genotypes.

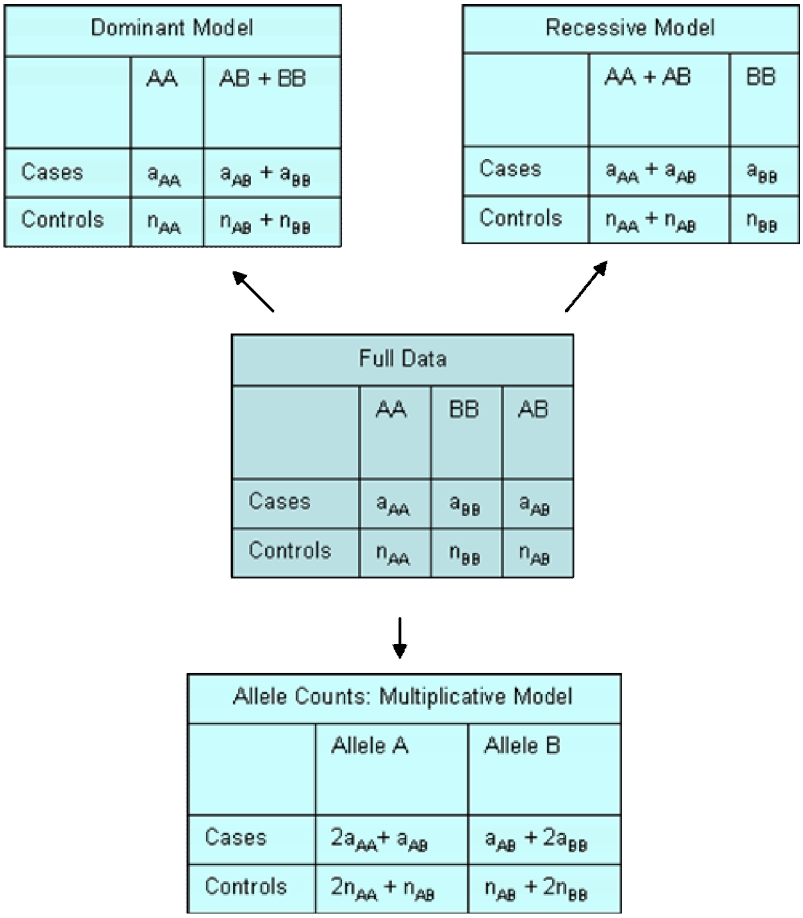


Figure 1.3 Contingency table calculations for case-control association study analysis. AA, BB and AB represent the genotypes at certain SNPs. a is the number of cases with a particular genotype. N is the number of controls with a particular genotype.

The χ^2 statistic tests the null hypothesis that there is no difference between observed and expected results and is calculated as follows

- $\chi^2=\Sigma(obs_i-exp_i)^2/exp_i$
- Degrees of Freedom, df = (number of columns-1)*(number of rows-1) where the observed counts are the number of alleles/genotypes and the expected results in each cell
- Exp_i = column total*row total/total

The Odds Ratio (OR) is a measure of relative risk derived from case-control association studies; it is the ratio of odds of disease in the exposed group over the

non-exposed group (Zondervan and Cardon 2007). The OR can also be defined as the ratio of the odds of an event occurring in one group (cases) to the odds of it occurring in another group (controls) is also calculated with 95% confidence intervals.

The Odds Ratio of the AB genotype

- OR AB genotype baseline AA = $(n_{AA}*a_{AB})/(n_{AB}*a_{AA})$
- OR BB genotype baseline AA = $(n_{AA}*a_{BB})/(n_{AA}*n_{BB})$
- OR AB genotype baseline BB = $(n_{BB}*a_{AB})/(n_{AB}*a_{BB})$
- OR AA genotype baseline BB = $(n_{BB}*a_{AA})/(n_{AA}*a_{BB})$

The 95% Confidence Intervals are calculated as

- 95%CI = OR/EF ⟶ OR*EF
- Error Factor (EF) = exp[1.96*s.e.(logOR)]
- s.e.(logOR) = $\sqrt{[1/n_{AA} + 1/a_{AB} + 1/n_{AB} + 1/a_{AA}]}$

as an example for OR AB (Kirkwood and Sterne 2003).

The expected genotype frequencies, as calculated above, are compared to the observed genotype frequencies, using the χ^2 statistic and corresponding *P*-value.

1.3.3. Haplotype analysis in association studies

Analytic methods for association studies can be extended beyond the allele level and genotype level. Association analysis performed on haplotypes can improve the information content of the genomic region under investigation (Ott and Rabinowitz 1997; Chapman and Wijsman 1998), with an aim to detect chromosomal segments that carry the true functional variant. UNPHASED (Dudbridge 2003) is a selection of programmes for association analysis of multilocus haplotypes from unphased genotype data. UNPHASED includes a program COCAPHASE v2.43 which performs case-control association study analysis and is used here to first, perform haplotype frequency estimation and then perform association analysis of such

haplotypes to illness. This software uses the expectation maximisation (EM) algorithm, which is a numerical method to find the maximum likelihood estimates of parameters and is applicable in situations where there are more categories than can be distinguished, such as known genotype data but incomplete haplotype data. The EM algorithm estimates the haplotype frequencies of unphased genotype data and standard unconditional logistic regression analysis, applying the likelihood ratio test under a log-linear model, to ultimately compare haplotype frequencies between cases and controls. Potential problems caused by rare haplotypes are circumvented by declaring all haplotypes rare, that have a frequency less than or equal to 5% in both the cases and the controls, and clumping them together for testing the null hypothesis. The maximum-likelihood frequencies are re-computed after identification of rare haplotypes.

1.3.4. Reporting evidence for association studies

To report positive association results, the number of tests performed must be taken into account. The range of tests must be considered; at the genotype level, from multiple polymorphisms in candidate gene studies to whole genome association studies; at the phenotype level, from multiple phenotypes definitions to multiple endophenotypes and; at the analysis level, from single allele analysis to multiple-marker haplotype analysis. The traditional critical significance level (α<0.05) is not thought to be sufficient to avoid false positive results. There are many options to account for the number of tests, such as Bonferroni correction, that corrects the critical significance level by the number of tests (n) performed (α=0.05/n) and Nyholts, which is based on spectral decomposition (Nyholt 2004). However, they may be both overly conservative for non-independent tests (Perneger 1998; Salyakina, Seaman et al. 2005). A preferred option is permutation analysis, which permutes the affection status of both cases and controls within the particular study, recalculates the tests for association, records the smallest *P*-value and repeats this to obtain empirical distribution of the smallest *P*-values. This method is

computationally intensive but provides a clearer picture to the degree of significance for the association results.

1.3.5. Features of association studies

There are many ways to improve the power to detect a real positive association result. For example, increasing sample size and marker density, improving accuracy of phenotype and genotype measurements, rigorous quality control and error checking and selecting individuals with highest genetic loading as cases, and individuals with lowest genetic loading as controls can provide substantial improvements. At the analysis stage, power can be improved with a through analysis plan; incorporating all phenotype and genotype information, allowing for heterogeneity and population substructure (Sham 1998). Cautious interpretation of results and a replication study are also recommended.

Additionally, there are many population genetic assumptions made in association analysis. Firstly, Hardy-Weinberg equilibrium is assumed. Secondly, the same mating patterns are assumed between cases and controls and this is especially important if genotype frequencies are tested. Thirdly, it is assumed that the search is for common alleles, which arose from old mutations in founder individuals, as illustrated in Figure 1.1. Linkage disequilibrium is also assumed, as is no population admixture or stratification. Finally, it is assumed there is little or no allelic heterogeneity (where a single trait is caused by different mutations within one gene) within the disease gene. The disease gene will have one common mutation, where this functional variant, within the gene will influence the phenotype.

Indeed, case-control association studies have many advantages over linkage analysis. One major benefit of case-control association studies is the ease of collection of large sample sizes. Also, there is no correlation between individuals so the statistics are simpler. Association studies are also more efficient than linkage

studies in narrowing down the region of interest, as linkage studies do not localise the phenotype and genotype to a small chromosomal location. However, there are many factors that influence the power of association studies such as the mode of inheritance of the illness, population sub-structure, the age of the disease mutation, allele frequency, distance of marker to the trait, the mutation rate and genotyping errors. In addition, the risk allele must be common to have high power in association studies and hence rare risk variants are extremely difficult to detect.

There are also disadvantages related to association studies. One potential bias in case-control association studies, which is not present in linkage studies, is the presence of confounding in the control population. Confounding exists when an unmeasured confounder variable is associated with both the risk factor and the outcome, causing an apparent association between the measured risk factor and the outcome. Alcohol consumption was reported to be risk factor for lung cancer because a significant association was observed. However, this association was not true and was due to a confounding factor, smoking, as the fact that individuals who smoked were also more likely to drink larger amounts of alcohol (Personal communication, Chad Garner, 23rd July 2006, University of California, US (Weeks, Lathrop et al. 2006)).

It is important that the control population is representative of the same population that the cases are from, to reduce false positive findings due to information and selection biases. However, the most important bias is confounding due to population stratification, which is related to the ethnic origin of cases and controls. Population stratification occurs when the comparison of the frequency of the genetic variant between cases and controls shows a significant difference due to the underlying sampling of the population and not due to a disease risk variant. This spurious association can occur in a population with different proportions of ethnicity to cause the confounding effects of population admixture. For example, population admixture affected the outcome of an association study which detected a

mutation that was protective against diabetes in an Indian population. The association did not hold when population stratification was taken into account (Cardon and Palmer 2003). This can be avoided by carefully matching the ethnicity of the control population to that of the case population.

1.3.6. HapMap

The International HapMap Project is a very important resource. The aim of the HapMap is to determine the common patterns of DNA sequence variation in the human genome, by characterising sequence variants, their frequencies and correlations between them (HapMap 2003). Phase II of HapMap provides information on the location of ~4 million common SNPs across the genome in four populations of different ethnic origin (Caucasians of Northern and Western European origin, Japanese from Tokyo, Han Chinese from Beijing and Yoruba from Nigeria) (Frazer, Ballinger et al. 2007). Furthermore, within each population HapMap provides information on the allelic association of SNPs in the same genomic region. This is advantageous because knowledge of the LD structure of a specific region helps to select SNPs that capture the majority of all common genetic variation, because one SNP can predict the allelic status of other nearby SNPs without having to genotype these variants themselves. A method available to visualise haplotypes is "Haploview" (Barrett, Fry et al. 2005). LD between markers can be visualised using the standard LD measurements D', r^2 and LOD. One method in Haploview is the "Solid Spine of LD". This method searches for a "spine" of strong LD running from one marker to another along the triangle in the figure of LD. This implies that the first and last markers in a block are in strong LD with all intermediate markers but that the intermediate markers are not necessarily in LD with each other.

1.4. Family-Based Linkage Studies

Genetic linkage analysis is a method to study the inheritance of traits in families, by mapping the relative positions of two or more loci using genetic markers. Linkage was described by Morgan in 1911 in the model organism, the fruit fly (*Drosophila melanogaster)*, and the first gene mapping paper was written by Sturtevant in 1913. Both reports demonstrated the basis of genetic linkage; when a recombination event occurs between two genetic loci, it does so at a rate related to the distance between the two loci, on the same chromosome. Thus, loci that are physically close together, tend to be inherited together. The aim of linkage analysis is to determine whether two loci tend to cosegregate more often than they should if they are not physically close together on the same chromosome (Ott 1999).

There are many approaches to detect linkage to a phenotype in a family. The main two methods are i) the traditional model-based parametric linkage analysis and ii) the model-free non-parametric linkage analysis method based on allele sharing. Parametric and non-parametric linkage methods follow the same experimental procedure by examining the pedigree structure, determining the expected pattern (of recombination fraction or allele sharing respectively) and comparing the difference to that expected by chance. Both methods can provide the approximate location of the disease marker, the placement of the disease loci relative to other loci and the exclusion of other loci that do not contain the disease marker.

1.4.1. Parametric linkage analysis

Traditional linkage analysis is based on Mendel's second law of independent assortment that states "genes controlling different characters segregate independently." For example as shown in Figure 1.4, if a person is heterozygous at two loci, A/a at one loci and B/b at the other loci, there are four possible gametes: AB, Ab, aB and ab and all are equally likely, if under independent assortment.

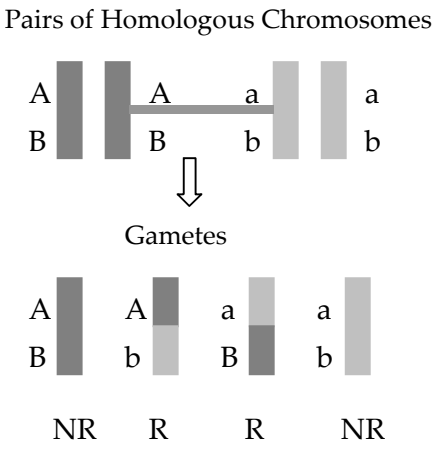


Figure 1.4 Illustration of Mendel's second law of independent assortment. There are two loci shown on each chromosome, A and B. There are four possible gametes. Two are non-recombinants (NR) and two are recombinants (R).

However, if the two loci, one a marker and the other one a hypothetical disease locus, are tightly linked, and recombination between loci does not take place, then the only chromosomes will be AB and ab which are the parental, non-recombinant (NR) gametes. The recombinants (R) or non-parental gametes Ab and aB will be rare. Meioses are informative for linkage, if the gamete can be identified as a recombinant or non-recombinant. If θ is the recombination fraction, which is a measure of the genetic distance between two loci as discussed in more detail in 1.4.1, between locus A and B, then for a person with AB|ab phased genotypes, the gamete probabilities are as Table 1.3.

GAMETE	PROBABILITY	TYPE
AB	(1-θ)/2	NR
Ab	(1-θ)/2	NR
Ab	(θ)/2	R
aB	(θ)/2	R

Table 1.3 Gamete probabilities. Here are the gamete probabilities from the phased genotypes AB|ab. There are four possible gametes. Θ is the recombination fraction. NR is non-recombinant and R is recombinant.

The possibility of crossover between non-sister chromatids is illustrated in Figure 1.4. One crossover results in 50% recombinant chromosomes or two crossovers results in either no recombinants or all recombinants. The maximum recombination fraction is 50%, when the two loci are not linked, as a consequence of independent assortment.

These calculations can be applied to pedigrees to test for linkage; the likelihood (L) of recombination event occurring between two loci (θ^R) or not [$(1-\theta)^{NR}$]

- $L(data|\theta)= \theta^R(1-\theta)^{NR}$

and compared to the null hypothesis that loci A and B are not linked, such that $\theta_0 = 0.5$. A likelihood ratio test compares a general θ_1 against the null θ_0:

- $L(data| \theta_1)/L (data | 0.5)$.

This is the odds for θ_1 compared to the Null hypothesis. The next step is to calculate the likelihood ratio test (odds) at various values for θ. The value θ that maximizes the odds is the Maximum Likelihood Estimate (MLE) of θ. This is the best estimate for θ, given the data. LOD is the logarithm base 10 of the odds

- $LOD (\theta) =\log_{10} [L (\theta) / L (0.5)]$

In general, the Null hypothesis is rejected when $LOD (\theta_1)>3$ for whole genome scans and there is evidence for linkage between two loci. Conversely, linkage can be excluded at θ_1 when $LOD (\theta_1) <-2$. The thresholds for significant and suggestive

linkage are discussed later in section 1.4.6.3. LOD scores can be added across independent families, also discussed in section 1.4.1.3.

1.4.1.1. Genetic distance

The recombination fraction (θ) is a measure of the genetic distance between two loci, such that it is the expected number of crossovers between the two loci per gamete. Recombination fractions define genetic distance rather than physical distance. Two loci that show 1% recombination are defined as being 1 centiMorgan (cM) apart on a genetic map. Recombination fractions are not additive across a genetic map. To describe the relationship between recombination fractions and genetic map distance, a mapping function is used. There are several different map functions such as the Morgan map function, the Haldane map function and the Kosambi map function. The simplest one is the Morgan map function that assumes a maximum of one crossover but which only works for small genetic distances. The Haldane map function assumes any number of random, independent crossovers but does not allow for interference. The Kosambi map function takes into account that over short distances a crossover at one location reduces the chance of a second crossover nearby due to physical proximity, namely interference. A STR (short tandem repeat) high resolution genetic map by deCODE genetics (Kong, Gudbjartsson et al. 2002) is now commonly used in linkage studies and also in this study.

For small distances, genetic distance (Morgans) is similar to recombination fraction (θ). Also, genetic distance is additive whereas recombination fractions are not additive. There is a complex relationship between recombination and genetic distance, and further again in relation to physical distance measured in base pairs (bp). The female human genetic map is generally longer that the male map. On some small regions ~10:1, and more rarely the converse ~1:10 for example on the pseudo-autosomal region on chromosome X that is 19cM in males but 2.7cM in females. Sex-specific maps should ideally be used for linkage analysis but although

this area is undergoing continuous research, this is not yet possible for the whole genome.

1.4.1.2. Genetic model of inheritance

Traditional parametric linkage analysis requires a model of inheritance to specify how the genotype at the trait locus influences the trait phenotype. The model is specified as a penetrance function, the probability of a genotype given a certain phenotype: Pr (phenotype|genotype). To explain penetrance further, it is useful to consider the disease allele, D, and the wild-type allele, d. For a simple, fully penetrant recessive disease the genotypes Dd and dd have penetrances of 0 and the genotype DD has penetrance 1. The set of penetrances for the genotypes at a locus are referred to as the penetrance vector, (f_{dd}, f_{Dd}, f_{DD}). For the example of a recessive disease, the penetrance vector is (0, 0, 1). For a simple dominant disease the penetrance vector is (0, 1, 1). A value of f_{DD} less than one indicates that some individuals homozygous for the disease allele do not express the disease, and such is termed reduced penetrance. A greater than zero value for f_{dd} allows for phenocopies, which are affected individuals without the disease allele, where disease is presumed to have a different cause. The number of phenocopies assumed in a genetic model is a function of the penetrance vector and the gene frequency of the disease allele in the population. The relationship between f_{Dd} and f_{DD} specifies the dominance of the disease gene (Terwilliger and Ott 1994).

In some diseases, including bipolar disorder and recurrent major depression, penetrance varies with age. To account for this, linkage analysis can include age-dependent penetrances where individuals are grouped according to age and assigned liability classes depending on age. A penetrance vector is then applied to each liability class. The simplest method is directly from age at onset within the family. These values can be estimated approximately from cumulative age of onset distributions for sets of affected individuals (Ott 1999).

In short, linkage analysis tests how close two loci are on a chromosome. Parametric linkage analysis is based on a model of inheritance that defines parameters to describe genotypes at a trait locus influencing a trait phenotype. The resulting output estimates the degree of linkage to show how far apart the loci are from each other.

1.4.1.3. Features of parametric linkage analysis

As with any statistical test, there are many important assumptions to be considered in parametric linkage analysis. It is assumed that all markers are in Hardy-Weinberg equilibrium and in linkage equilibrium, which assumes that two alleles at different loci are independent of each other. Random mating is also assumed. Additionally, in the search for a single locus, it is assumed there is no interaction between alleles at different loci and, in fact, there is less power to detect if such epistasis (genetic interaction) is present.

There are many advantages to parametric linkage analysis. Firstly, the method is not sensitive to allelic heterogeneity or population history. Also population substructure can be problematic in population case-control association studies but not in family linkage studies. Additionally, there are no candidate genes, regulatory regions or transcription factor binding sites to propose, as one might need to focus on if resources are restricted in performing association studies. Thirdly, linkage analysis is performed within pedigrees and results from many pedigrees can be combined to obtain an overall result. Finally, a linkage signal with a marker can be detected up to 20cM away from the proposed disease locus, so markers separated by 10cM (microsatellites) or 2cM (SNPs) have a good chance of detecting linkage (Terwilliger and Ott 1994; Ott 1999).

There are however, disadvantages to linkage mapping. The first deterrent is that families are difficult to collect for late-onset traits and the power of linkage analysis is low to detect genes of modest effect (Risch and Merikangas 1996). A second complication in parametric linkage analysis is sensitivity to bilinearity, where illness introduced into the family from outside, can mask the true inheritance pattern of the disease within the family. A third limitation is that generally, the localised region will not be finely-defined unless the family is very large but narrowed to a region >3 Mb wide with wide confidence intervals, dependent on the number of meioses (Terwilliger and Ott 1994). Linkage analysis employs the concept of linkage disequilibrium (LD). However, LD regions in families are large, but between unrelated individuals it is smaller and this allows the fine-mapping of LD regions in case-control association studies. A final obstacle is locus heterogeneity. Locus heterogeneity occurs where more than one trait loci is suspected when examining linkage to a disease in a group of families. The disease may be due to one locus in one family and a different locus in other families. This can be overcome by testing for locus heterogeneity (α<1) as the probability that a pedigree is segregating a disease gene at any given θ. θ and α can be simultaneously estimated in an unbiased fashion using parametric linkage analysis. In addition, at each θ a Heterogeneous LOD (HLOD) can be obtained which is maximised over all value of α. It is sometimes advised to use this HLOD score for complex traits. However, this is not relevant when a single family is analysed.

1.4.2. Non-parametric linkage analysis

An alternative method to parametric linkage analysis is non-parametric linkage analysis. The aim of non-parametric linkage analysis is to identify the disease locus, by identifying chromosomal markers where affected individuals with a disease are more alike at a location, than expected by chance. The main advantage of this linkage analysis method is that it is a model-free test, with no model-dependent inferences about underlying disease genotypes. Furthermore, this analysis can

increase power to detect linkage if the model is not known, or imprecisely specified, which can be the case in complex diseases (Clerget-Darpoux, Bonaiti-Pellie et al. 1986).

Essentially, non-parametric linkage analysis is based on the affected sib pair method which looks at a measure of sharing IBD (as explained in Figure 1.2) and compares the observed and expected sharing, over the sharing standard deviation (SD) with a Mean Sharing Test, Z = Mean sharing – Expected sharing / SD (sharing). The best method of detecting genetic sharing between relations within a pedigree would be to completely specify the gene flow throughout the whole pedigree, such that an inheritance vector would completely specify the inheritance pattern of all the alleles in a family. The number of these inheritance vectors for a specific family is a function of the number of meioses and increases exponentially with pedigree size. To circumvent this problem, there are computational methods that estimate IBD sharing. The MERLIN programme was chosen to estimate the S_{all} sharing statistic, which sums the numbers of non-trivial permutations of all possible sets consisting of one allele from each affected individual. This favours the sharing of a single allele by a large number of affected individuals as hypothesised in this study. The SALL statistic is commonly used in linkage studies of complex disease and performs well (Davis, Schroeder et al. 1996; Sengul, Weeks et al. 2001).

In order to evaluate the level of IBD sharing, Kong and Cox have a single parameter alternative model using the δ parameter. The δ parameter measures the amount of deviation of the inheritance vector distribution from its null distribution, in one direction, such that the null hypothesis, H_0: $\delta = 0$ vs. $\delta \geq 0$. There are two types of alternative directions; linear (upper bound of δ) and exponential. The linear model is the standard method to identify small increases in allele sharing spread across a large number of families. The exponential model is designed to identify a large increase in allele sharing in a small number of families, and is a better test if a large increase in allele sharing among affected individuals is expected. It is, however,

more computationally intensive (Abecasis, Cherny et al. 2002). The Kong and Cox approach generates more accurate *P*-values than those obtained by other methods and is commonly reported (McGuffin, Knight et al. 2005; Shugart, Samuels et al. 2006).

There are certain limitations to non-parametric linkage analysis in a single pedigree. There is a decrease in power compared to parametric linkage analysis if the model is correctly specified in parametric linkage analysis (Terwilliger and Ott 1994). Also, the results are particularly sensitive to marker allele frequencies. Furthermore, genetic heterogeneity is not taken into account and recombination fraction is not estimated (Kong and Cox 1997; Sengul, Weeks et al. 2001; Shih and Whittemore 2001; Abecasis, Cherny et al. 2002; Weeks, Lathrop et al. 2006). A suggestion in the literature is to weight different sub-pedigrees according to their size (Sengul, Weeks et al. 2001). However, with multiple pedigrees, each of a different size, there is no single optimal way to weight the contributions from different pedigrees. Furthermore, if this is performed then the non-parametric test is no longer truly "model-free" (Sham 1998).

1.4.3. Linkage analysis programmes

Currently, there are many general linkage analysis programmes. These programmes can be classified into three different groups depending on the algorithm used for analysis: i) Elston-Stewart algorithm on which the classic linkage analysis programmes are based, for example FastLink, Linkage, Mendel and Vitesse (Elston and Stewart 1971). ii) Lander-Green algorithm on which Allegro, GeneHunter, Mendel and MERLIN are based (Lander and Green 1987) iii) Markov chain Monte Carlo (MCMC) algorithm on which Simwalk2 and Loki are based (Sobel and Lange 1996). There are drawbacks for each algorithm. The Elston-Stewart is limited by marker number (~8 loci), the Lander-Green is restricted by pedigree size (~20 people) and the MCMC provides only an estimated solution (Weeks, Lathrop et al.

2006). Although the MCMC methods are not restricted by pedigree size or marker number the computational time is a huge burden and is unsuitable for analysis of a large number of markers.

Since its publication in 2002, MERLIN (Multipoint Engine for Rapid Likelihood INference) has been highly cited and widely-used because it permits quicker analysis of larger pedigrees, while reducing memory constraints. The algorithms that MERLIN uses are based on sparse binary trees that summarise the gene flow within pedigrees. The reasons for MERLIN's speed are twofold i) it explores the symmetries in the pedigree structures and data to derive a family of related algorithms for examining the pedigree and ii) reduces the inheritance-vector space by eliminating classes of vectors that are not compatible with observed genotypes (Nicolae and Cox 2002). Furthermore, MERLIN can perform many applications such as rapid genotype error detection, parametric and non-parametric linkage analysis and haplotyping, with a large number of markers (Abecasis, Cherny et al. 2002).

Since the advent of dense marker sets there has been development of linkage programmes that can accommodate analysis of large families for linkage on a large number of markers, for example Superlink (Fishelson and Geiger 2002) and a multipoint non-parametric technique (Thomson, Quinn et al. 2007). However, none match the range of features, the ease of execution and the reliability offered by MERLIN.

1.4.4. Multipoint linkage analysis

As linkage analysis in anything other than a simple small pedigree with many markers requires a computer, algorithms have been developed, as described in section 1.4.3. One such algorithm is the Lander-Green algorithm which considers an entire pedigree one locus at a time and combines this information to produce a multipoint LOD score by combining information from all loci (Lander and Green

1987). This is an extension of 2-point analysis in which linkage of a disease marker is tested not just to a single marker, but to an entire map of markers and is reported in this study.

The Lander-Green method computes the probability of marker genotypes, given an inheritance vector. For example, at a single marker, the linkage information of each non-founding member can be summarised by two binary digits, one for the allele in the parental haplotype and one for the allele in the maternal haplotype. A pedigree with *n-f* non-founders contains 2(*n-f*) potentially informative gametes so that the pattern of inheritance at a single locus can be described by a vector of 2(*n-f*) elements. Each element is the inheritance vector, v, which describes the parental status of the allele at the locus in one gamete. There are $2^{2(n-f)}$ possible inheritance vectors, each describing a different pattern of allele transmissions at the locus (Sham 1998). For each possible inheritance vector at the locus, P(Mi|Vi) at locus I, where P is the probability, Mi is the marker data at this locus (evidence) and Vi is a certain inheritance vector. In brief, if "a" is a vector of alleles assigned to founders of the pedigree. There are restrictions imposed by the observed marker genotypes on the vector "a" that can be assigned to the founder genes. The Lander-Green algorithm extracts only vectors "a" compatible with the marker data. P [m|v] is obtained by summing all compatible vectors "a".

Multipoint analysis uses binary digits, 0/1 notation, to describe the parental origin of an allele in relation of the first allele, where 0 is no recombination with respect to the first allele, 1 is recombination with respect to the first allele, the first allele is not coded (0). These values are then related to haplotype frequencies via θ and the probability of each haplotype determined under the Haldane map function. The next step is the calculation of haplotype counts to obtain the maximum likelihood estimate of haplotype probabilities. Following that, the haplotype probabilities are transformed into recombination fractions (θ) where the sizes suggest the order of loci. In order to map the disease, fixed marker loci are moved through test positions.

The multilocus likelihood at each test position L_x is then compared to the multilocus likelihood at unlinked L_{inf}.

- $2(\ln L_x - \ln L_{inf})$=location score=4.6*multipoint LOD score.

There are many advantages to multipoint analysis. Firstly, it is a more informative measure as genotype information can be imputed at an originally uninformative locus via haplotype information. Secondly, the location of the disease gene can be fine-mapped by evaluating many locations along a chromosome. However, it is critical that the order of markers and the distances between the markers are correctly specified (Haines and Pericak-Vance 1998). Finally, multipoint linkage analysis can increase power to detect to detect linkage and uncover minor gene effects (Baron 2001).

1.4.5. Potential sources of error in linkage analysis

Despite the availability of robust methods that test for linkage, their success depends on high-quality data. There are numerous potential sources of error in the experimental flow-chart from pedigree collation, DNA preparation, genotype calling and data file preparation. Errors in these areas can greatly affect the power of linkage analysis. Pedigree errors can decrease the power to detect linkage or produce false positive results (McPeek and Sun 2000). Genotyping errors can also have a severe consequence in linkage analysis. Errors can affect haplotype frequencies (Kirk and Cardon 2002) and lead to an inflation of genetic map length (Hackett and Broadfoot 2003). A 1% error rate can generate a loss of 53-58% of the linkage information for a trait locus (Douglas, Boehnke et al. 2000).

Genotyping errors can occur due to DNA of poor quality or quantity. For example, a small quantity of DNA molecules remaining in a sample may have resulted from a very high dilution of DNA or degradation. This can lead to both allelic dropout, which is the stochastic non-amplification of an allele; such that one of the two alleles

present at a heterozygous locus is amplified and a false allele, which is an allele-like artefact that is generated by PCR (Taberlet, Griffin et al. 1996). Low quantity of DNA also increases the risk of contamination because there is a higher probability of amplifying contaminant molecules when the DNA sample is of low concentration (Taberlet, Griffin et al. 1996; Pompanon, Bonin et al. 2005). Recently, whole genome amplification of DNA is regarded as a practical method to conserve DNA, while also maintaining the integrity of the DNA. Amplified DNA samples have been used successfully in whole-genome scans for complex genetic diseases, for example in the neurodegenerative disease amyotrophic lateral sclerosis (ALS) (Kasperaviciute, Weale et al. 2007).

1.4.6. Reporting evidence for linkage

1.4.6.1. Power

Power is defined as the probability of a study to obtain a significant result, if this result is true in the underlying population from which the study subjects were sampled (Zondervan and Cardon 2007). The power to detect linkage depends strongly on the magnitude of the contribution that the disease locus makes to the genetic variation of the trait. There are other complicating factors that influence power of both parametric and non-parametric linkage analysis. The distance of markers from disease loci, phenocopy rate, genetic heterogeneity and magnitude of the genetic effect cannot be accounted for and may affect the power of the study. A study can be optimised by seeking a pedigree of a suitable size and structure tested with informative markers.

Simulation can also be used to compute the power of a linkage study, where replicates of datasets are simulated under the alternative hypothesis of linkage. However, the simulation process is very time-consuming, the simple assumptions are non-realistic for complex models and there is difficulty in determining an appropriate alternative hypothesis.

1.4.6.2. LOD support intervals

Support intervals for LOD scores are difficult to obtain analytically (Terwilliger and Ott 1994). Originally a 1-LOD support interval was recommended (Conneally, Edwards et al. 1985) which is obtained by constructing a horizontal straight line at 1 unit LOD score below the maximum LOD score, provided that LOD max ≥ 3 (otherwise, no support interval is constructed). The endpoints are the intersection of this horizontal line with the LOD score curve (Ott 1999). The width of the 1-LOD support interval is R-L, where L and R are the first positions at which the LOD score is less than the maximum LOD score minus 1 to the left and right of the position of the maximum LOD score respectively. However, an inconsistency occurs when constructing 1-LOD-unit support intervals between the statistical test and the support intervals when $1<Z_{MAX}<3$, where no support interval should be constructed when $Z_{MAX}<3$ (Terwilliger and Ott 1994). Terwilliger and Ott recommend the use of 3-LOD-unit support intervals. However, 1-LOD unit rule is widely accepted and is the current method of reporting linkage support intervals.

1.4.6.3. Significance

Thresholds of significance are an important consideration to avoid reporting false positive linkage results (Strachan and Read 1999). The higher the LOD scores the greater the evidence for linkage. Traditionally, a score of 3 was regarded as significant evidence of linkage in a classical two-point linkage study of simple Mendelian traits corresponding to a $P=10^{-4}$. LOD -2 was regarded as evidence against linkage.

Widely acceptable significance thresholds are based on the classic paper by Lander and Kruglyak in 1995, who proposed a series of thresholds as criteria for suggestive and significantly linkage for complex diseases. Suggestive linkage is a LOD score or *P*-value that would be expected to occur once by chance in a whole genome scan. Significant linkage is statistical evidence that would be expected to occur by chance 0.05 times in a whole genome scan. Highly suggestive linkage is a LOD score or *P*-

value that would be expected to occur by chance 0.001 times in a whole genome scan. Confirmed linkage is the confirmation of significant linkage in a further sample, preferably by an independent group of investigators. Table 1.4 shows the thresholds calculated by Lander and Kruglyak for mapping loci underlying complex traits. For suggestive linkage, a LOD of 1.9 is recommended and for significant linkage, a LOD of 3.3 is recommended for LOD score analysis.

MAPPING METHOD	SUGGESTIVE LINKAGE		SIGNIFICANT LINKAGE	
	***P* value**	**LOD**	***P* value**	**LOD**
LOD score analysis in human	1.7×10^{-3}	1.9	4.9×10^{-5}	3.3
Allele-sharing methods in human (sibs and half-sibs)	7.4×10^{-4}	2.2	2.2×10^{-5}	3.6

Table 1.4 Thresholds for evidence of linkage in complex traits. This table is adapted from Lander and Kruglyak, 1995. It shows the classification based on the number of times that one would expect to see a results at random in a dense genome scan.

The thresholds for allele-sharing methods depend on the relationship measures. A weighted average for a mixture of different relative types should be used. At the sib-pair level, the *P*-value and for suggestive linkage are 7.4×10^{-4} and 2.2, and significant linkage are 2.2×10^{-5} and 3.6. The LOD scores for genome-wide significance should be in the range 3.3-4.0, depending on study design according to Lander and Kruglyak criteria. Suggestive linkage results may be wrong, but are fworth reporting if accompanied with a warning. Lander and Kruglyak also advocate reporting all regions with a nominal *P* value of *P*=0.05 encountered in a complete genome scan, but without any claims of linkage.

There are some difficulties in estimating significance levels for example using multiple diagnostic schemes to define affection status. Strictly, a Bonferroni

correction should be applied so the *P* values are multiplied by the number of models. This is, however, overly stringent as the models are closely related and are not statistically independent.

Another difficulty with estimating linkage thresholds is the addition of extra markers to regions of linkage, leading to an increase in the false positive rate. Lander and Kruglyak initial criteria in 1995 were based on a dense map of markers covering the whole genome. Sawcer *et al* in 1997 argued this point stating that such stringent measures were not required for initial linkage scans with sparse coverage of the genome, where linkage peaks were followed up with increased marker coverage. Their simulations showed that follow up of interesting linkage peaks with more markers did not inflate the false positive rate (Sawcer, Jones et al. 1997). However, problems were found in their simulation method (Kruglyak and Daly 1998). When increasing the marker density to simulate adding extra markers around linkage peaks, they did this on the peak of their interest, not on the simulated false-positive peaks. In conclusion, the original recommendations to use dense-map thresholds still holds as shown in Table 1.4 and are regularly reported for whole-genome linkage scans; for example in a bipolar disorder whole-genome linkage scan (Park, Juo et al. 2004). However, simulations are sometimes required to take into account specific features of studies, such as the family structure. These simulations must model the experimental parameters precisely, including the follow-up of interesting regions. This should avoid under-estimating the false positive rate (Kruglyak and Daly 1998).

1.4.6.4. Simulation

Computer simulation is important to assess the significance of a result by generating realistic artificial data. This is achieved by obtaining random numbers from a deterministic computer by a "seed". This seed determines a starting point in a sequence, then uses a different seed for generating each dataset and thus creates

random datasets. A genotype can be simulated while taking account of the allele frequency. This can be extended by a process called "gene-dropping" (Jean W. MacCluer 1986). Gene-dropping generates random founder genotypes according to allele frequencies at each marker, using random numbers as explained above. Once all founders are assigned genotypes, the offspring's alleles are "dropped" down the pedigree at random from the parental genotypes. These random genotypes are then segregated through the pedigree, using the relationships as specified in the pedigree. If the random seed is changed, a different dataset of founder genotypes and segregation pattern is generated. Furthermore, the simulated genotype data mirrors the pattern of the original data in terms of missing data, marker informativeness, marker spacing and phenotype status.

The advantages to "gene-dropping" are many. The method is fast, conditional on a known pedigree structure and unconditional on any known disease status. Consequently, this method is useful for this study, as the pedigree structure was known and limited by available DNA. A disadvantage of the method is with larger pedigrees, the proportion of simulated outcomes compatible with the data becomes very small and most of the generated observations are discarded. There are many assumptions in the "gene-dropping" method. For example, it is assumed that the parents are chosen randomly from population, the markers are in Hardy-Weinberg equilibrium, there is no segregation distortion such that there are 50:50 transmission probabilities, there are no mutations and no genotyping error.

MERLIN is a useful programme for simulating data. It generates random datasets based on the gene-dropping method described above. The genotype data is simulated under the null hypothesis of no linkage to the observed phenotypes. The resulting data is a random genome that is unlinked to the trait of interest and is useful for examining false positive rates due to peculiarities in marker informativeness, marker spacing and trait distribution (Abecasis, Cherny et al. 2002).

Simulation of random datasets allows the calculation of empirical threshold to find the desired significance level (Sawcer, Jones et al. 1997). To estimate the threshold T, a replicate dataset is simulated under the null hypothesis, linkage analysis is computed and a LOD is recorded. This is performed numerous times. The genome-wide threshold of significance T is then estimated so that the number of times the LOD score is equalled or exceeded divided by the number of simulations (#>T)/(Total #) = 0.05 (Weeks, Lathrop et al. 2006). The threshold for highly suggestive linkage is a LOD score that would be expected to occur by chance 1/1,000 times in a whole genome scan.

1.4.7. Haplotype analysis

A haplotype is a combination of genetic markers such as SNPs or microsatellites that are located closely together on the same chromosome and tend to be inherited together. Haplotype analysis aims to describe the inheritance of genetic information descending through a family. More specifically, a haplotype for an individual at a certain loci is defined as the set of alleles inherited from one parent at a series of markers. Each individual has two haplotypes, one of maternal origin and the other of paternal origin. Traditionally haplotype analysis aimed to estimate the "best" haplotype for each of the two haplotypes for every individual in a family. This analysis is improved by considering multiple generations in a family to construct haplotypes that segregate through the family.

Haplotype analysis is important for many reasons. Firstly, it can help to identify genotyping errors that comply with Mendelian inheritance, but when considered as a haplotype that would require a double recombination event over a close distance, which is very rare. Secondly, it can make linkage analysis more informative by combining loci that are uninformative as individual markers to use as a single point. This can often improve standard linkage analysis. Haplotype analysis can narrow down a region of a putative trait locus. Examination of haplotypes under linkage

peaks can be used to check for inheritance of a consistent chromosomal region by all affected individuals. This is particularly important in single large families, such as the family in this study.

It can be hypothesised in an isolated population that a trait is introduced by mutation for example, in a founder individual and the proceeding generations have inherited this mutation from the single founder. Haplotype analysis of all affected individuals could reveal a conserved haplotype inherited with the trait through many generations. This conserved haplotype will be surrounded by areas not inherited from the founder individual. These areas are evidence of recombination events that have occurred with transmission of the haplotype through generations. The section of the haplotype that is common to all affected individuals may contain the trait locus. This localisation technique has been successful in isolating the gene "ATM" for the autosomal disorder ataxia-telangiectasia (A-T) by supplementing classical linkage analysis in 176 pedigrees, with pedigrees from Britain and Costa Rica (Savitsky, Bar-Shira et al. 1995).

Undoubtedly, defining the origin of haplotypes is a complex process. The main problem is missing genotype data, which can substantially increase the number of haplotypes that are consistent with the observed data. This problem is confounded by uninformative markers, such as biallelic SNPs. However, these problems can be reduced in large pedigrees when the majority of people are genotyped, by providing information on the homozygosity in the founders and informative matings. There are many computer programs available for haplotype analysis, each with their own advantages and drawbacks. For this study, the chosen method of haplotype analysis is based on the Lander-Green algorithm that searches for gene flow representation in a pedigree using MERLIN.

It is important to note that haplotype construction is based on estimations and may not reveal the true haplotype. This is particularly important with respect to widely

spaced markers and large amounts of missing data. Manuel inspection of haplotypes at key regions can resolve some of these problems. Furthermore, the most likely haplotype may not in fact be the actual true haplotype (Sobel and Lange 1996). However, in the search for the true haplotype, adding more information may actually uncover the true haplotype.

1.4.8. Importance of single large families in genetic studies

There are many advantages to studying single large families. Firstly, the mode of inheritance of a phenotype can be hypothesised with a genetic model and this model can be refined, depending on how the trait segregates through the family. The risk of the illness can also be estimated within the family which makes the genetic model more accurate. Linkage analysis can help decipher whether the signal detected is either indicating the presence of a causative factor in a genomic region or the actual disease causing variant. This was discussed in Clerget-Darpoux and Elston 2007, specifying two examples: the *VNTR* locus flanking the insulin gene in type-1 diabetes and the *PTP22* gene in rheumatoid arthritis. In both cases, examining IBD sharing in affected siblings, showed the variant under investigation to be the causal variant (Hodge 1993).

Secondly, the study of single families is recommended to avoid allelic heterogeneity (Clerget-Darpoux and Elston 2007). In fact, a disease that appears to be multifactorial may be due to single gene effects. One example is the identification of *BRCA1* by segregation analysis in large sample of families. For example, a four-fold reduction in sample size was achieved by enriching a cohort with cases who had two or more other family members with the illness, making a more efficient and cost-effective study of a risk allele *BRCA1* for breast cancer (Easton, Pooley et al. 2007). Thirdly, another advantage is that large families can be investigated to assess complex genetic effects, for example imprinting, anticipation and epigenetics. Finally, genotyping errors, genomic mutations and copy number variants are easier

to detect or validate in family data. Certainly, the larger the family, the more power is available to detect any of the above effects. However, studies in single families also have drawbacks as the larger the family, the greater the chance that more than one genetic risk factor is represented, but this is negligible in comparison to a large collection of small families or population-based cohorts.

1.4.9. Benefits of re-analysis

Many experts in the field advocate the re-analysis of linkage studies, due to the current generation of newer, high-resolution data (Evans and Cardon 2004). A map of closely spaced SNPs may offer many advantages over low-density microsatellite maps, including increased information content, greater power to detect linkage, and may improve the localisation of the disease locus (Evans and Cardon 2004; John, Shephard et al. 2004; Thalamuthu, Mukhopadhyay et al. 2005). Evans and Cardon (2004) suggest that "previous linkage studies that employed sparse microsatellite maps could benefit substantially from reanalysis by use of a denser map of markers" (Evans and Cardon 2004). For example, reanalysing the 1996 UK multiple sclerosis whole genome linkage screen showed that high density SNP linkage mapping sets can extract significantly more information and achieve higher rates of genotyping success and accuracy than previous studies (Sawcer, Maranian et al. 2004). Other technological advancements are a more accurate marker map and greater genotyping success rate. These improvements are vital, as missing and incorrect data can severely reduce the power to detect linkage (Pompanon, Bonin et al. 2005). Hence, re-analysis of single large families in genetic studies, with high-resolution data is then certainly recommended twofold.

1.5. Genetic Studies of Bipolar Disorder

As the characteristics of both bipolar disorder and genetic mapping have been introduced separately, I would like to discuss genetic mapping studies pertinent to bipolar disorder.

1.5.1. Genetic mechanism of bipolar disorder

Many methods have been undertaken to elucidate the genetic mechanism of bipolar disorder: collecting single affected individuals for cytogenetic studies, families for linkage studies and population cohorts for association studies. Each method has a specific advantage, cytogenetic and linkage studies may detect single, rare variants that could highlight candidate genes or pathways or association studies can detect multiple, common susceptibly variants. Early linkage studies assumed single gene inheritance and large, apparently autosomal dominant, pedigrees were studied. Unfortunately, consistent evidence of single, major gene effects in bipolar disorder has not been forthcoming (Segurado, Detera-Wadleigh et al. 2003).

The primary deterrent in the gene search is that the precise mode of the inheritance of bipolar disorder is unknown. Segregation analyses on large pedigrees have produced mixed results, where some studies are consistent with single gene models (Rice, Reich et al. 1987) and others unable to demonstrate major locus transmission. The observed very rapid decrease in recurrence risk, as shown in Table 1.1, from identical co-twins (40-70%) to first degree relatives (5-10%) and to the general population (0.5-1%) is not consistent with a single gene mode of inheritance (Craddock, Khodel et al. 1995). It may be suggestive of a polygenic pattern, where a number of genes with small, additive effects provide an underlying genetic predisposition to disease, an epistatic interaction of multiple genes or a more complex genetic mechanism. Complex genetic mechanisms suggested in bipolar disorder include locus heterogeneity, imprinting, anticipation, mitochondrial inheritance (McGuffin, Owen et al. 2002) or disruption of alternative splicing mechanisms (Wang and Cooper 2007). Albeit, the lack of agreement in segregation analyses, to determine a single gene model of inheritance or complex mechanisms, could be due to the difficulties in defining a phenotype and a reflection of the number of assumptions required for segregation analysis. It is evident, undoubtedly, that the pattern of inheritance for bipolar disorder in some large pedigrees follows simple or quasi-Mendelian inheritance, consistent with a single

gene contributing to susceptibility to bipolar disorder in that family, and may yet prove fruitful in revealing a single genetic disease-susceptibility variant. Nonetheless, it must be assumed that no single pattern of inheritance underlies all of bipolar disorder. It is likely that cases may be sporadic or inherited, and that heritable forms of illness may include single and multiple genes.

1.5.2. Cytogenetic & linkage studies of bipolar disorder

Rare variants of large effect have indeed been identified as contributing to bipolar disorder and related disorders. Cytogenetic studies examine individual chromosomes for evidence of abnormalities which may cause gene disruption and linkage studies search for greater than expected, cosegregation of an illness and a chromosomal region. In a linkage study of a large Scottish family with bipolar disorder (1 case), schizophrenia (7 cases) and recurrent major depression (10 cases), a balanced translocation was found to segregate with illness (maximum LOD=7.1) (St Clair, Blackwood et al. 1990; Blackwood, Fordyce et al. 2001). The translocation between chromosomes 1 and 11 lead to the identification of a candidate susceptibility gene "Disrupted In Schizophrenia" (DISC1) (Millar, Wilson-Annan et al. 2000). Molecular genetic studies have also begun to achieve success in uncovering susceptibility genes for complex diseases by their chromosomal location. Significant linkage regions for bipolar disorder have been reported in extended pedigrees on chromosomes 1q, 4p, 6p, 10p, 10q, 11p, 12q, 13q, 18p, 18q, 21q, 22q and Xp (Baron 2002). A meta-analysis performed on whole-genome linkage scans of bipolar disorder found the strongest evidence for susceptibility loci on chromosome 13q ($P<6x10^{-6}$) and 22q ($P<1x10^{-5}$) (Badner and Gershon 2002).

1.5.3. Association studies of bipolar disorder

In contrast to linkage and cytogenetic studies that identify rare variants for disease, allelic-association studies have made important genetic contributions by identifying common susceptibility variants for bipolar disorder and other related psychiatric

illnesses (McGuffin, Owen et al. 2002). Allelic association refers to the co-occurrence of an allele at a particular locus and a disease, above the level of chance. This co-occurrence can suggest causation or that the marker is in linkage disequilibrium (LD) with the causative variant. The case-control association study compares individuals with a psychiatric illness (cases) with unaffected subjects from the same population (controls) for differences in allele, genotype or haplotype frequencies. Association analysis considering multiple, linked markers can often be more informative, as illness may be associated with a particular haplotype, more strongly than a single allele.

The results of the first three whole-genome association studies in bipolar disorder have been recently reported. The first whole-genome association study for bipolar disorder was performed by genotyping over 550,000 single nucleotide polymorphisms (SNPs), on the Illumina HumanHap550, in two independent case-control samples of European origin. The first cohort was 461 cases and 563 controls individuals of European origin from the NIMH Genetics Initiative and the second collection was 772 cases and 876 controls from the German population. The study incorporated a two-stage study design by performing the initial association screen using pooled DNA, and then the selected SNPs were confirmed by individual genotyping. There were 88 SNPs that met the criteria for replication in both samples. There were no reports of a single SNP of large effect. The strongest association signal was detected at a SNP in the first intron of *diacylglycerol kinase eta* (*DGKH*; P=1.5x10^{-8}, experiment-wide P<0.01, Odds ratio (OR) =1.59). An association signal was also reported at a marker on chromosome 4p16.1 in intron 2 of a VPS10 domain-containing receptor (*SORCS2*) that is predominantly expressed in the developing brain (*SORCS2*; P=1.4x10-5, OR=1.67). As this association study shows evidence for several genes each of modest effect, the authors advocate a polygenic disease model influencing bipolar disorder risk (Baum, Akula et al. 2007).

A second whole genome association study was reported by the Wellcome Trust Case-Control Consortium (WTCCC) by genotyping 500,000 SNPs on 2,000 individuals affected with bipolar disorder and 3,000 controls from Great Britain. Again, there were no reports of powerful associations. The strongest association signal was to a marker on chromosome 16p12 (genotypic test $P=6.3\times10^{-8}$, OR=2.1). There were four other regions showing association at $P<5\times10^{-7}$ on chromosomes 1p31, 2q31, 12q21 and 22q12. The associated markers were not within previously studied candidate genes for bipolar disorder (WTCCC 2007).

A third genome wide association scan was reported in 1,461 patients with bipolar I disorder and 2,008 controls from the US and UK with genotyping for >370,000 SNPs (Sklar 2008). The genotyping was performed on the same Affymetrix Gene Chip Human Mapping 500K Array platform as the WTCCC. Once more, there was no evidence for association that did not meet either the criteria for genome-wide significance (5×10^{-8}) nor that replicated in two independent sample sets. The strongest results were in two brain-expressed genes; *myosin5B* (*MYO5B*; $P=1.66\times10^{-7}$, OR=1.51) and *tetraspanin-8* (*TSPAN8*; $P=6.11\times10^{-7}$, OR=0.58). No associations were replicated from the other two genome-wide scans for bipolar disorder. The authors hypothesise the reason for the lack of agreement between studies was because the susceptibility alleles for bipolar disorder are likely to be modest in effect size and therefore require larger cohorts to detect such effects. Other reasons for non-replication of significant association results between studies may include phenotypic and genotypic heterogeneity, population specific alleles or epistatic interactions of multiple modest effect genes.

1.5.4. Candidate susceptibility genes for bipolar disorder

Studies of rare chromosomal rearrangements also reveal candidate genes for bipolar disorder and schizophrenia. For example, *Glutamate Receptor, Ionotropic, Kainate, type 4* (*GRIK4)* originally found to be disrupted at a chromosomal breakpoint in a patient

with schizophrenia, was also found to be associated with bipolar disorder (*P*=0.0002) and schizophrenia (*P*=0.0005) in a Scottish population (368 bipolar disorder cases, 386 schizophrenia cases and 458 controls) (Pickard, Malloy et al. 2006). Another example of a gene identified at a balanced chromosomal translocation, t(9,14)(q34.2;q13) is a neuronal transcription factor called *NPAS3*, initially associated with schizophrenia and learning difficulties (Pickard, Malloy et al. 2005). A further study showed that *NPAS3* is associated with bipolar disorder (*P*=0.0000010) and schizophrenia (*P*=0.0000012) at the haplotype level by calculating the "net genetic load" incorporating genetic heterogeneity, in the same Scottish population (Pickard, Christofororou et al. 2008).

Linkage studies, followed by meta-analyses have identified and confirmed candidate regions, which were refined by linkage disequilibrium mapping to identify promising candidate genes for bipolar disorder and also to schizophrenia. These include: *G72* (a brain expressed protein) on chromosome 13q34, *DISC1* on chromosome 1q42, *neuregulin-1* (*NRG1*) on chromosome 8p21, a regulator of G-protein signalling (*RGS4)* on chromosome 1q, a cell adhesion molecule (*NCAM1*) on chromosome 11q23.1, a *D-Amino acid oxydase* (*DAO*) on chromosome 12q24, metabotropic glutamate receptor 3 (*GRM3*) and *GRM4* on chromosomes 7q21 and 6p21 respectively, an *N-methyl-D-asparate receptor subunit 2B* (*GRIN2B*) on chromosome 12p12, a WKL1, cation channel (*MLC1*) on chromosome 22q13, a *synaptogyrin 1* (*SYNGR1*) on chromosome 22q13, a potassium chloride co-transported (*SLC12A6*) on chromosome 15q13 and *catechol-O-methyltransferase* (*COMT*) on 22q11 (Kato 2007). There are other genes also found using the candidate gene approach that show association to bipolar disorder, including a G-protein coupled receptor (*GPR50*) on chromosome Xq28 (Thomson, Wray et al. 2005). Emerging evidence has also supported an association between *methylenetetrahydrofolate reductase* (*MTHFR*) on chromosome 1p36 and bipolar disorder, recurrent major depression and schizophrenia (Gilbody, Lewis et al. 2007). In addition, there are several candidate genes in the circadian rhythm pathway that

are associated with bipolar disorder; *TIMELESS* on chromosome 12q12 and *PERIOD3* on chromosome 1p36. Further, genes that have been implicated in more than one study are the dopamine D1 receptor (*DRD1)* on chromosome 5q35 and *inositol(myo)-1-monophosphatase 2* (*IMPA2*) on chromosome 18p11 (reviewed in Kato 2007). Despite the number of candidate genes studied, there has been no causative gene or genetic risk factor implicit for bipolar disorder and recurrent major depression.

1.5.5. Genetic overlap between bipolar disorder & schizophrenia

Despite clear differences in the definition of the phenotype, there is evidence for a genetic overlap between bipolar disorder and schizophrenia. A twin study examining the relationship between bipolar disorder, schizophrenia and schizoaffective (an illness with features of both schizophrenia and bipolar disorder symptoms) suggested that there is a set of genes that contribute to all three syndromes (Cardno, Rijsdijk et al. 2002). Other studies also include families with both bipolar disorder and schizophrenia such as that which identified *DISC1* (Blackwood, Fordyce et al. 2001) and suggested the candidate genes *NRG1* and *GRIK4* (reviewed in (Blackwood, Pickard et al. 2007). In addition, the loci highlighted for bipolar disorder in a meta-analysis of genome-wide linkage scans, suggested evidence for linkage to chromosome 13q and 22q, regions previously implicated in schizophrenia (Badner and Gershon 2002). Also, promising candidate genes for bipolar disorder were originally identified in molecular genetic studies of schizophrenia are a positional G-protein receptor kinase 3 on 22q, the G72/G30 locus on 13q34 and *brain-derived neurotrophic factor (BDNF)* on 11p13 (McGuffin, Owen et al. 2002). Furthermore, there are many signalling pathways and cellular mechanisms where evidence converges for bipolar disorder and schizophrenia: phosphoinositide 3-kinase (PI3K) and serine/threonine-specific protein kinase (AKT) signalling, growth factors, N-methyl-D-aspartic acid (NMDA) and glutamate-related,

dopaminergic and serotonergic pathways, circadian genes, cytokines, oxidative and other stress pathways and endoplasmic reticulum stress (Carter 2006) (www.polygenicpathways.co.uk). Collating this evidence suggests common genetic risk factors for bipolar disorder and schizophrenia.

1.6. Large Scottish Family in a Genetic Study of Bipolar Disorder

In the following section, I would like to describe the family that is investigated in genetic studies of bipolar disorder in chapters 3, 5 and 6 of this thesis.

1.6.1. Family description

The Scottish family in this study is a valuable and rare genetic resource, particularly with respect to the size of the pedigree and the clear diagnosis of bipolar disorder and recurrent major depression in the family. The pedigree was originally reported in 1996 as F22, with 11 individuals with bipolar disorder and 16 with recurrent major depression in a 132 member pedigree (Blackwood, He et al. 1996). The family was recruited through two probands: individuals 21 and 29. A follow-up of this family was subsequently reported with 32 cases, 12 of which were bipolar disorder and 20 of which were diagnosed with recurrent major depression (Le Hellard, Lee et al. 2007). Further follow-up of this family is reported here.

Figure 1.5 clearly illustrates the expanse of the present day pedigree. Table 1.5 lists the number of affected individuals at the time of the original study in 1996, and at the present day. There are 180 members that span five generations. Of the 180 members, 45 are founders and 135 are non-founders. Founders are individuals that are married into the family, have no parents specified and are assumed to be unrelated (Ott 1999).

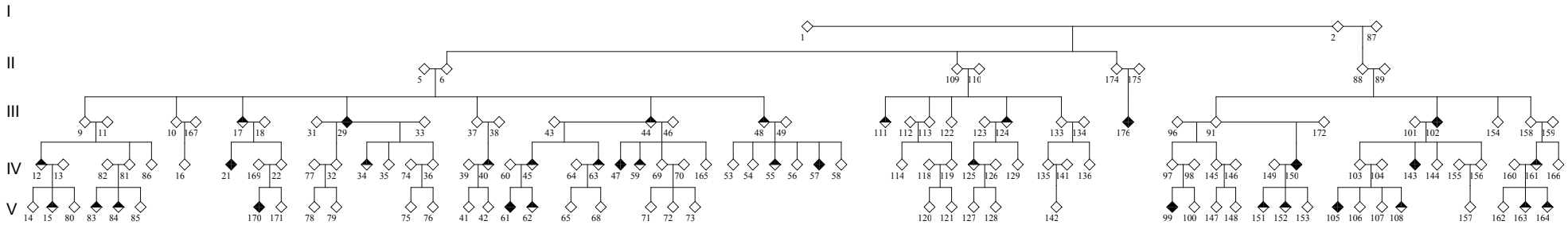


Figure 1.5 The pedigree The pedigree is a representation of the large Scottish family in this study**.** For confidentiality, a selection of extraneous individuals was removed and the genders are hidden. The filled symbols denote bipolar disorder and half-filled symbols represent recurrent major depression. The numbers under the symbols is the numbering system used in this thesis. The generations are labelled with roman numerals on the left.

Prior to this study, the diagnosis for all affected individuals were re-evaluated and reached by consensus between two psychiatrists (Prof. D. Blackwood, Dr. W. Muir) using clinical information collected by direct interview based on the Schedule for Affective Disorders and Schizophrenia-Life Time Version (SADS-L) (Endicott and Spitzer 1978) supplemented by case-note review, further information from general practice or hospital case records, and collateral information from relations. Twelve individuals have been diagnosed with bipolar disorder and 24 with recurrent major depression. The key numbers are shown in Table 1.5. Eight individuals were diagnosed with other psychiatric illness including anxiety states, alcoholism, single episode depression or minor depression.

DIAGNOSIS	1996 STUDY	PRESENT STUDY
Bipolar disorder	11	12
Recurrent major depression	16	23
Other psychiatric diagnosis	12	8
Unaffected	72	92
Unknown	21	45
Total	132	180

Table 1.5 Number of affected individuals in the family in 1996 and in this study For each psychiatric diagnosis, the number of affected individuals in the family is listed.

The changes to diagnoses are listed in Table 1.6. The diagnosis for 18 individuals was changed from the original classification. A change in diagnosis occurred for example, from bipolar disorder (at initial assessment) to recurrent major depression (at review ten years later) because the information about manic symptoms obtained at the initial assessment was considered inconclusive and was not supported by subsequent follow-up, when the case was reviewed.

INDIVIDUAL	PREVIOUS DIAGNOSIS	UPDATED DIAGNOSIS
12	Other psychiatric illness	Recurrent major depression
15	Other psychiatric illness	Recurrent major depression
20	Other psychiatric illness	Unknown
27	Other psychiatric illness	Unaffected
30	Recurrent major depression	Other psychiatric illness
35	Other psychiatric illness	Unaffected
55	Bipolar disorder	Recurrent major depression
65	Recurrent major depression	Other psychiatric illness
69	Other psychiatric illness	Unaffected
71	Unaffected	Unknown
83	Other psychiatric illness	Recurrent major depression
84	Other psychiatric illness	Recurrent major depression
97	Other psychiatric illness	Unaffected
124	Other psychiatric illness	Recurrent major depression
130	Unaffected	Recurrent major depression
157	Recurrent major depression	Other psychiatric illness
165	Other psychiatric illness	Unknown
166	Recurrent major depression	Other psychiatric illness

Table 1.6 Diagnoses updates. The diagnoses for individuals in the Scottish family were updated as above in February 2006. Other psychiatric diagnosis includes single episode depression, anxiety or alcoholism.

One individual was changed from bipolar disorder to recurrent major depression. Four individuals were updated from recurrent major depression to other psychiatric illness. Six individuals were altered in the opposite direction from other psychiatric illness to recurrent major depression. Three individuals' other psychiatric diagnosis became unaffected and two individuals became unknown. Two individuals whose diagnoses were previously unaffected were updated: one as unknown and the other as recurrent major depression.

1.7. Evidence for a Bipolar Disorder Susceptibility Region on Chromosome 4

1.7.1. Linkage evidence on chromosome *4p15-p16* from a large Scottish family

A genome-wide linkage study was performed on the aforementioned large Scottish family. There were 11 individuals diagnosed with bipolar disorder and 16 with recurrent major depression, both classified together as major affective disorder. Significant linkage of bipolar disorder to chromosome 4p15-16 was found with a maximum LOD of 4.1 (Blackwood, He et al. 1996). Additional statistical evidence is based on a significant result (LOD=3.7) using a variance component method of analysis (Visscher, Haley et al. 1999). Genetic analysis was updated by re-evaluating family members and increasing the number of markers in the chromosome 4p15-p16 region, resulting in a maximum LOD score of 4.4 (Le Hellard, Lee et al. 2007). This significant LOD score implies this region is likely to contain a susceptibility locus for bipolar disorder.

1.7.2. Further evidence for chromosome *4p15-p16* from other families

Evidence for psychiatric illness, including bipolar disorder, schizophrenia and related phenotypes to this region on chromosome 4 has been confirmed in other linkage, population and affected individual studies as detailed in Table 1.7.

Origin	Sample	Psychiatric Diagnosis	Genetic Evidence	Reference
Scotland	F22	BP & RMD	LOD 4.4	(Blackwood, He et al. 1996; Le Hellard, Lee et al. 2007)
Wales	F50	SCZ & SA	LOD 2	(Asherson, Mant et al. 1998)
Scotland	F59	BP & RMD	LOD 0.9	(Blackwood, He et al. 1996)
US (Ashkenazi Jewish)	F48	BP, SCZ, RMD & others	LOD 3.2	(Detera-Wadleigh, Badner et al. 1999)
Danish	Families (2)	BP	LOD 2	(Ewald, Degn et al. 1998)
US	Families (154)	BP with psychosis & suicidal behaviour	LOD 1.8	(Cheng, Juo et al. 2006)
Wales	Sib-pairs	SCZ	LOD 1.7	(Williams, Rees et al. 1999)
Arab Israeli	Families (21)	SCZ & SA	LOD 2.2	(Lerer, Segman et al. 2003)
Italy	Families (16)	SCZ & BP	LOD 1.5	(Vazza, Bertolin et al. 2007)
Japan	Individual	SCZ	t(4; 13) (p16.1; q21.31)	(Itokawa, Kasuga et al. 2004)
Canadian	Individual	SCZ	[inv 4 (p15.2;q21.3)]	(Palmour, Miller et al. 1994)
Brazil	Individual	Mental retardation & SCZ features	t(1;4) (p21;p14)	(Cordeiro, Zung et al. 2007)
Faroe Islands	Individuals (11 SCZ & 17 BP)	BP & SCZ	P=0.00007	(Als, Dahl et al. 2004)
Scottish	Population	BP & SCZ	P=0.044 (SCZ)	(Underwood, Christoforou et al. 2006)
Scottish	Population	BP & SCZ	$P \leq 0.0005$, $P \leq 0.0003$	(Christoforou, Le Hellard et al. 2007)

Table 1.7 Genetic evidence for psychiatric illness on chromosome 4p14-p16. This table lists the studies that show genetic evidence for a mental illness locus on chromosome 4p14-p16. The table is grouped according to the type of genetic evidence: linkage studies, cytogenetic studies and association studies, respectively. Details of the study origin, the study samples, their diagnosis and the genetic evidence from the study and related references are listed. BP is bipolar disorder, RMD is recurrent major depression, SCZ is schizophrenia, SA is schizoaffective disorder. US is United States of America. Inv is a chromosomal inversion.

The primary evidence for linkage for bipolar disorder on chromosome 4p14-p16 stems from the Scottish family in this study, a maximum LOD 4.1 with marker D4S394 and a maximum multipoint LOD 4.8 with markers D4S431, D4S394 and D4S403 on chromosome 4p15.33-4p16.1. There are eight additional sources of linkage evidence to chromosome 4p14-16; i) in a Welsh family, a maximum LOD 1.97 with marker D4S403 on chromosome 4p15.33 was reported (Asherson, Mant et al. 1998) ii) in another Scottish family, a LOD 0.9 with marker D4S394 on chromosome 4p16.1 was reported, (LOD 0.9, (Blackwood, He et al. 1996)) iii) in a large family from the United States (one family of the 22 families studies) a LOD 3.2 with marker D42632 on chromosome 4p16 was reported (Detera-Wadleigh, Badner et al. 1999)) iv) in two Danish families a LOD 2 with marker D4S394 on chromosome 4p16 was reported (Ewald, Degn et al. 1998) v) in a 154 pedigree sample set with psychosis and suicidal behaviour a LOD 1.8 with D4S2366 on chromosome 4p16.1 was reported (Cheng, Juo et al. 2006)) vi) an increase sharing to chromosome 4p16.1 in Welsh schizophrenia sib-pairs [LOD 1.73, (Williams, Rees et al. 1999)] vii) in Arab-Israeli families with schizophrenia and schizoaffective disorder (non-parametric LOD 2.2 to D4S394, (Lerer, Segman et al. 2003)) viii) in an Italian sample of 16 families affected by bipolar disorder and schizophrenia (non-parametric LOD 1.5 to D4S405 on chromosome 4p14 (Vazza, Bertolin et al. 2007))

Three studies of single individuals provide further support for chromosome 4p14-16; two schizophrenia affected individuals, one with a balanced translocation t(4;13) (p16.1;q21.31) (Itokawa, Kasuga et al. 2004) and another with a chromosomal inversion, [inv 4 (p15.2;q21.3)] (Palmour, Miller et al. 1994). A third study on an individual with mild mental retardation and physical anomalies has a balanced translocation t(1;4) (p21;p14).

From population based cohorts, there are three studies that support this genomic region. A report on individuals from the Faroe Islands with bipolar disorder (17) and schizophrenia (11) describes excess haplotype sharing to chromosome 4p16.1

(best *P* value, *P*=0.00007, Als, Dahl et al. 2004). Preliminary evidence suggests association of the orphan *G-protein coupled receptor (GPR78)* on chromosome 4p16.1, in females with bipolar disorder and schizophrenia from a total of 377 bipolar disorder, 392 schizophrenia cases and 470 controls from the Scottish population (Underwood, Christoforou et al. 2006). The other association study will be described in section 1.7.3, as it was involved in refining the chromosome 4p15-p16 locus.

1.7.3. Refinement of chromosome 4p15-16 region

1.7.3.1. High-resolution haplotype analysis

Figure 1.6 illustrates recombination mapping in the abovementioned large Scottish family and three other families which are listed in Table 1.7. The aim was to narrow the linkage region that segregates with illness in three of the four families. There were two families of Scottish origin [F22 & F59 (Blackwood, He et al. 1996)], a Welsh family [F50 (Asherson, Mant et al. 1998)] and a large US family of Ashkenazi Jewish origin [F48 (Detera-Wadleigh, Badner et al. 1999)]. Both F22 and F48 are large families and the linkage regions are broad (22Mb for F22 and 10Mb for F48) with endpoints defined by recombination breakpoints in single affected individuals. Families F59 and F50 are smaller and do not have significant LOD scores. Therefore, they do not necessarily supply reliable data, as the possibility that affected members share common haplotypes, by chance across the genome is increased. However, their inclusion can be helpful if they are used to prioritise sub-regions of the large regions identified in the other families.

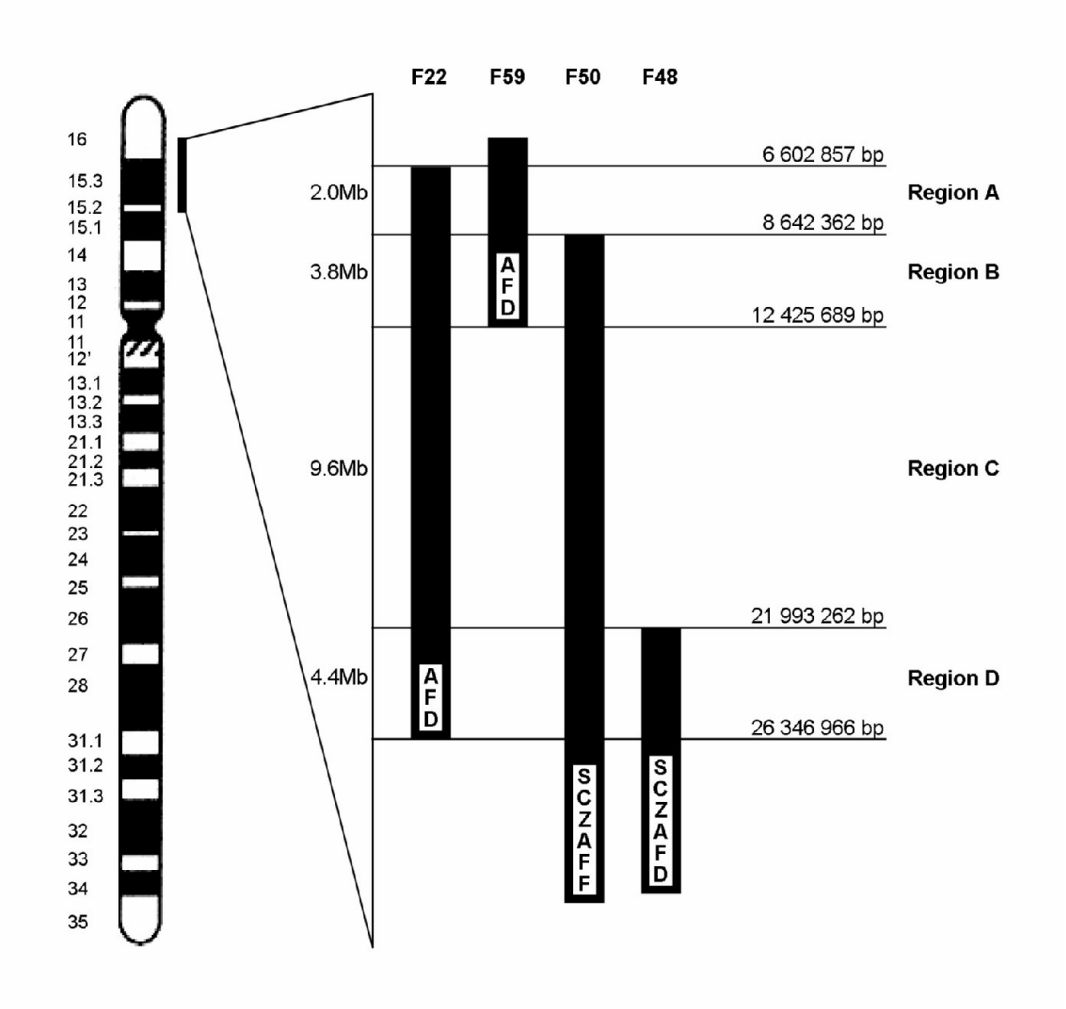


Figure 1.6 Definition of haplotypes shared by affected members in four families on chromosome 4. This diagram is taken from (Le Hellard, Lee et al. 2007). It illustrates the extent of the disease related haplotype on chromosome 4p15-p16 that is linked to affective illness. F22 is the large Scottish family under investigation in this study with individuals affected with bipolar disorder and recurrent major depression, combined as major affected disorder (AFD). F59 is another Scottish family with bipolar disorder and recurrent major depression (AFD). F50 is a Welsh family with schizophrenia and schizoaffective disorder (SCZAFF). F48 is a large family from the United States of Ashkenazi Jewish origin with major mental illness including bipolar disorder and schizophrenia (SCZAFD). Regions of overlap in four families are Region B and Region D. The co-ordinates are from UCSC May 04, NCBI build 35.

To date, haplotype analysis and allele sharing analysis highlight two regions in the 20Mb linked region, where three of the four linkage regions overlap, namely region B and region D (Le Hellard, Lee et al. 2007). Haplotypes that segregated with illness in the four families were determined at high-resolution, and are represented by the black bars in Figure 1.6 that overlap at region B and region D. In these two regions, allele sharing between linked haplotypes from the four families, which were assumed to be distantly related, was compared to that between control chromosomes from the four families. The proposed disease-causing locus will be shared in common between the families, as the original mutation occurred in a common ancestor. This is derived from the assumption in population genetics that affected individuals from a founder population will share a mutation at the disease locus, with a large genomic haplotype surrounding the disease locus. Recombination events specific to each family, will reduce the length of the genomic haplotype, surrounding the disease locus shared between the families. The number of consecutive markers that comprised a region of sharing was counted. A region of significant excess allele sharing was discovered in region B (P=0.009).

In addition, as listed in Table 1.7, an association study of region B and region D was performed (Christoforou, Le Hellard et al. 2007). 408 haplotype tagging SNPs were selected on a block-by-block basis and tested on 368 bipolar disorder samples, 386 schizophrenia samples and 458 control samples. In this study, a tagging SNP was one chosen to represent a haplotype block, as it is in high LD with other SNPs in the block. There were two clusters of markers and/or haplotypes that met significant levels: two in region B and another two in region D. In region B, overlapping SNPs and haplotypes met the criteria for significance specific to that region (P<0.0005) at the global and individual haplotype test level and clustered in two regions, chromosome 4p16.1 and chromosome 4p15.33. For region D, there were no individual significant SNPs, but certain global and individual haplotypes were associated with bipolar disorder and/or schizophrenia (region-wide threshold, P<0.0003). These overlapping haplotypes fell into two regions on chromosome

4p15.2. There were no reported known RefSeq genes in these four regions, there were however predicted transcripts and genes marked nearby. In addition to the four clusters, there were other less significant associations, but potentially true associations, that did not exceed the nominal significance thresholds determined by principal component analysis. Thus, this study was successful in identifying significant associations between bipolar disorder and markers on chromosome 4p15-p16 that are worthy of further investigation, but which of course, await replication.

Overall, region D appeared an attractive region for further study, from the strength of the linkage evidence emerging from the larger families of F22 (LOD 4.4) and F48 (LOD 3.2). Region B was also favourable as the three families, two Scottish and one Welsh, share a common Celtic heritage and may potentially share a common ancestral origin. To investigate the regions further, candidate genes for susceptibility to bipolar disorder were considered. There were 20 genes in regions B and D, as listed in Table 1 (Le Hellard, Lee et al. 2007). There were seven known genes in region B, including *GPR78* that has shown preliminary association in females with bipolar disorder and schizophrenia (Underwood, Christoforou et al. 2006). Of the thirteen known genes in region D (Le Hellard, Lee et al. 2007), there is biological and neuropharmacological evidence, to prioritise *PI4K2B* as a candidate gene. This evidence is discussed below.

1.8. PI4K2B

As I have introduced genetic studies of bipolar disorder, a large Scottish family where bipolar disorder segregates and the genetic evidence originating from this family to chromosome 4p15, I would now like to introduce a candidate gene study on *PI4K2B*, using expression analysis in a large Scottish family.

1.8.1. Candidate Susceptibility Gene for Bipolar Disorder

Popular candidate genes for bipolar disorder are selected based on positional evidence, from linkage and association studies or on functional evidence, often focussing on neurotransmitter systems. Functional candidate genes have been chosen for several reasons: involvement in the serotonin, dopamine and norepinephrine/noradrenaline neurotransmitter pathways, the GABAergic system, circadian rhythm-related pathways or cytokines. Dysfunction of these systems has been proposed as possible biochemical mechanisms in bipolar disorder. Additionally, popular functional candidates are target molecules of lithium, such as the phosphoinositide (PI) signalling pathway as lithium and an alternative drug for bipolar disorder, valproate, act on this pathway for therapeutic effect (reviewed in (Farmer, Elkin et al. 2007; Kato 2007)). A phosphatidylinositol-4 kinase type 2 beta, *PI4K2B* fulfils both positional and functional criteria as a candidate susceptibility gene for bipolar disorder, based on its position on chromosome 4p15, in region D shown in Figure 1.6, and its involvement in the PI pathway, illustrated in Figure 1.7.

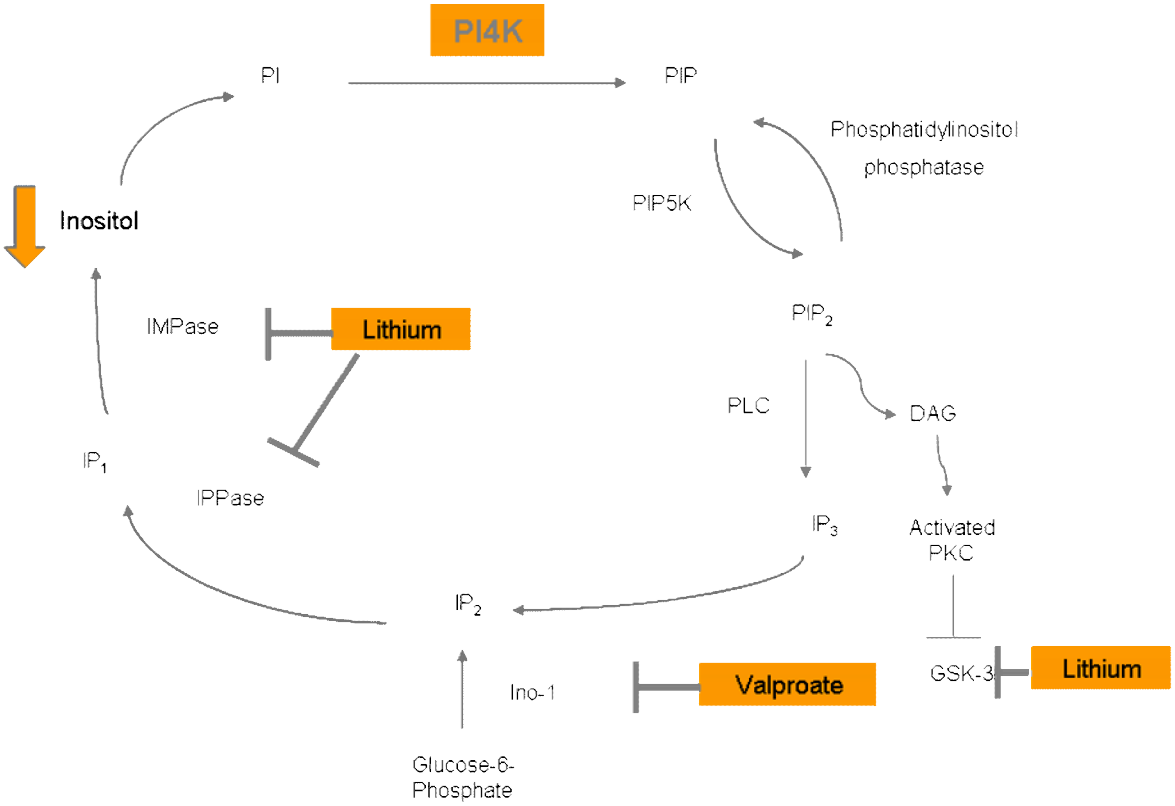


Figure 1.7 Functional evidence for *PI4K2B* in inositol signalling. This simplified figure highlights the action of bipolar disorder treatment drugs; lithium and valproate on the phosphatidylinositol (PI) signalling pathway. Activation of some G proteins induces hydrolysis of phosphoinositide-4,5-bisphosphate (PIP_2) to diacylglycerol (DAG) and inositol-1,4,5-triphosphate (IP_3) via phospholipase C (PLC). DAG activates protein kinase C (PKC) which inhibits glycogen synthase kinase-3 (GSK-3). IP_3 is recycled back to PIP_2 by inositol monophosphatase (IMPase) and inositol polyphosphatase phosphatase (IPPase); both of which are inhibited by lithium (Majerus 1992). Valproate inhibits inositol synthase (Ino-1) that is involved in the production of inositol-4,5-diphosphate (IP_2), alongside inositol-4 phosphate (IP_1). The recycling steps also involve the phosphorylation of phosphatidylinositol (PI) to phosphatidylinositol-4-phosphate (PIP) via phosphatidylinositol 4-kinase (PI4K), and the downstream phosphorylation of PIP to PIP_2 via phosphatidylinositol-4-phosphate 5-kinase (PIP5K).

As the genetic evidence supporting the role of a chromosome 4p15 locus in bipolar disorder has been discussed, I will now consider the genetic evidence to support genes involved in PI signalling in bipolar disorder, the functional evidence that implicates abnormal PI signalling in bipolar disorder and then focus specifically on the characteristics, function and cellular location of PI4K2B.

1.8.2. Genetic evidence for phosphoinositide genes in psychiatric illness

Table 1.8 lists a number of genes in the PI pathway that map to candidate regions for susceptibility to bipolar disorder, schizophrenia and other major affective disorders. It is notable that including *PI4K2B*, there are four phosphatidylinositol kinases in candidate susceptibility regions (PIP5K2A, PIK3C3 and PIK4CA).

CHROMOSOME	*PI GENE*	DESCRIPTION	REFERENCE
3q13	*GSK3β*	Glycogen synthase kinase 3, beta	(Benedetti, Bernasconi et al. 2004; Lachman, Pedrosa et al. 2007; Rowe, Wiest et al. 2007)
4p15	*PI4K2B*	Phosphatidylinositol-4 kinase type 2 Beta	(Blackwood, He et al. 1996; Le Hellard, Lee et al. 2007)
10p12	*PIP5K2A*	Phosphatidylinositol-4-phosphate 5-kinase, type II A	(Stopkova, Saito et al. 2003; Schwab, Knapp et al. 2006; Bakker, Hoogendoorn et al. 2007)
13q14	*DGKH*	Diacylglycerol kinase eta	(Baum, Akula et al. 2007)
15q13	*cPLA2*	Phospholipase A2 Beta (cytosolic)	(Craddock and Lendon 1999)
18p11.2	*IMPA2*	Inositol monophosphatase 2	(Sjoholt, Ebstein et al. 2004; Ohnishi, Yamada et al. 2007)
18q12	*PIK3C3*	Phosphatidylinositol 3-kinase, class III	(Stopkova, Saito et al. 2004; Duan, Gao et al. 2005)
21q22	*Synaptojanin 1 (SYNJ1)*	Inositol 5-phosphatase	(Saito, Guan et al. 2001; Stopkova, Saito et al. 2004)
22q11	*PIK4CA*	Phosphatidylinositol 4-kinase a type II	(Saito, Stopkova et al. 2003)
22q12	*PIB5PA*	Phosphatidylinositol 4,5-bisphosphate 5-phosphatase A	(Lachman, Kelsoe et al. 1997)
22q12	*PITPNB*	Phosphatidylinositol transfer protein beta	
22q12	*Synapsin III*	synaptic vesicle membrane phosphoprotein	(Lachman, Stopkova et al. 2005; WTCCC 2007)

Table 1.8 Phosphoinositide genes and psychiatric illness. Genes involved in the phosphoinositide pathway that are mapped to chromosomal regions linked to psychiatric illness, including bipolar disorder and schizophrenia (Bennett and Horrobin 2000; Saito, Stopkova et al. 2003) (www.polygenicpathways.co.uk). The references for each gene are listed, except for *PIB5PA* and *PITPNB* on 22q12 where the linkage evidence to the region is referenced. Further detail on each study in provided in the text.

1.8.2.1. GSK3β

The bipolar disorder treatment, lithium, inhibits Glycogen Synthase Kinase 3, beta (*GSK3β*), as illustrated in Figure 1.7. This suggests that manipulating GSK3β may have a therapeutic value in treating bipolar disorder. Concomitant with functional evidence reviewed for this candidate gene (Rowe, Wiest et al. 2007), there is also genetic evidence from separate studies. In a sample of 185 Italian patients with bipolar disorder, a SNP in the promoter region of *GSK3β* did not show association. However in the same study, homozygotes for the wild-type variant (T/T) showed an earlier age at onset than carriers of the mutant allele (F=5.53, d.f.=2,182, *P*=0.0047) (Benedetti, Bernasconi et al. 2004) A second study on bipolar disorder cohorts from the Czech Republic and the US, has reported evidence for a submicroscopic copy number variations (CNVs) in the *GSK3β* locus that appears to disrupt the gene's 3'-coding elements. The CNV also affects two other annotated genes. They report that patients with bipolar disorder have an increased frequency of this CNV, primarily the duplication variant, compared with controls (*P*=0.002) (Lachman, Pedrosa et al. 2007).

1.8.2.2. PIP5K2A

A gene in the phosphatidylinositol 4-phosphate 5-kinase family, *PIP5K2A* was screened for polymorphisms in a cohort of bipolar disorder samples (118) and schizophrenia (96) from the US, Czech Republic and Israel. This study revealed the existence of an imperfect CT repeat polymorphism (disruption of CT repeat) located in an intronic region. The distribution of alleles from this highly polymorphic variant was modestly different between bipolar disorder and schizophrenia patients (*P*=0.03) (Stopkova, Saito et al. 2003). Another polymorphism screening of *PIP5K2A* revealed that a SNP in the same intronic region shows evidence of association (Stopkova, Saito et al. 2003). Furthermore, in a sample of 65 sib-pair families from Germany (56) and Israel (9) affected with schizophrenia, evidence for an association was reported for SNP rs10828317 in exon seven of *PIP5K2A*. This SNP, which causes

a non-synonymous amino-acid exchange (asparagine/serine) produced a *P*-value of 0.001 (experiment-wide significance level 0.00275) for over-transmission of the major allele coding for serine, analysed by transmission disequilibrium test (Schwab, Knapp et al. 2006). In addition, association of this SNP with schizophrenia has been also described in a sample of 273 Dutch schizophrenic patients and 580 controls (*P*=0.0004) (Bakker, Hoogendoorn et al. 2007). A negative association to *PIP5K2A* has also been reported in 260 bipolar disorder cases, 268 schizophrenia cases and 325 controls (Jamra, Klein et al. 2006). Aside from the positional genetic evidence, there is also alternative functional evidence to implicate *PIP5K2A* in bipolar disorder. In a microarray study that profiled lithium-modulated gene expression in human neuronal cells, *PIP5K2A* was upregulated upon lithium treatment (1.2 fold change, *P*=0.02) (Seelan, Khalyfa et al. 2007)

1.8.2.3. DGKH

One genome-wide bipolar disorder association study in two independent case-control cohorts of European origin revealed association to *DGKH*; $P=1.5\times10^{-8}$, experiment-wide $P<0.01$, OR=1.59) (Baum, Akula et al. 2007). The DGKH protein has an important role in the PI pathway by metabolising diacylglycerol (DAG). Figure 1.7 illustrates the position of DAG in the PI pathway.

1.8.2.4. cPLA2

Evidence for the cytosolic phospholipase A2 Beta (cPLA2) and psychiatric illness is provided by a survey of linkage analysis to the chromosomal region 15q12 (Craddock and Lendon 1999), which was originally identified in two genome scans of autism, with modest linkage reports (LOD ~1).

1.8.2.5. IMPA

Other genes involved in the PI pathway have undergone association analysis for susceptibility to bipolar disorder, namely the myo-inositol monophosphatase genes, *IMPA1* (Sjoholt, Molven et al. 1997) and *IMPA2* (Yoshikawa, Kikuchi et al. 2001). *IMPA2* maps to 18p11.2, a genomic locus for which evidence of linkage to bipolar disorder has been supported by several reports. A case-control association study was reported in a large Japanese cohort with three *IMPA2* SNPs. All three SNPs showed significant genotypic association (nominal *P*=0.031-0.0001) with schizophrenia, but not with bipolar disorder (Yoshikawa, Kikuchi et al. 2001). Another study investigated DNA variants in *IMPA1* and *IMPA2* genes in Norwegian bipolar disorder patients (44) and controls (48), followed by examination of selected polymorphisms and haplotypes in a family-based bipolar sample of Palestinian Arab proband-parent trios (75 nuclear families, 95 affected offspring). Two SNPs in the *IMPA2* promoter sequence and their corresponding haplotypes showed transmission disequilibrium in the Palestinian Arab trios. The -461C and -207T alleles were significantly more often transmitted than not transmitted to bipolar offspring (*P*=0.006 and 0.002 respectively, uncorrected). No association was found between the *IMPA1* polymorphisms and bipolar disorder in the Norwegian cohort, neither with respect to disease susceptibility nor with variation in lithium treatment response (Sjoholt, Ebstein et al. 2004). Evidence for *IMPA2* was strengthened by a report of a replication study in a Japanese cohort (496 patients with bipolar disorder and 543 control subjects). Association of the same *IMPA2* promoter SNPs was detected (-461C (*P*-value 0.042) and -207T (*P*-value=0.046)) with bipolar disorder in a Japanese cohort (496 bipolar disorder cases and 543 controls) (Ohnishi, Yamada et al. 2007). However, this association was not supported for bipolar disorder, in a study involving 237 parent-offspring trios and in 174 cases and controls from Bulgaria and UK (Dimitrova, Milanova et al. 2005).

1.8.2.6. PIK3C3

The chromosome 18q12 gene, *PIK3C3* encodes phosphoinositide-3-kinase, class three, which is highly expressed throughout the brain. A promoter variant in this gene has been associated with bipolar disorder and schizophrenia in two studies to date. Firstly, studies on 83 bipolar disorder patients from the Czech Republic showed a statistically significant difference in -432T allele distribution (P=0.008) and in 124 schizophrenia patients from Israel (P=0.0003) (Stopkova, Saito et al. 2004). Secondly, in a Chinese family-based study of schizophrenia in a sample of 235 trios there was significant evidence of preferential transmission at the same position of -432C allele (P=0.0036, global significance P=0.0092) (Duan, Gao et al. 2005). A negative association in the Danish population (310 schizophrenia cases and 880 controls) has also been reported (Jungerius, Hoogendoorn et al. 2007). In addition, this region on chromosome 18q12 has been highlighted in a whole-genome homozygosity association study where the gene is located adjacent to a homozygosity run that is overrepresented in schizophrenia individuals (P=0.0012) (Lencz, Lambert et al. 2007).

1.8.2.7. SYNJ1

Several rare promoter and splice junction polymorphisms were found only in bipolar disorder patients in the *SYNJ1* gene but this finding was not considered significant due to the small sample number (Saito, Guan et al. 2001). An ensuing study in a cohort of 84 bipolar disorder patients did not detect any rare variants, but showed in increase of an intron 12 polymorphism that was again not significant but may have been underpowered (Stopkova, Vevera et al. 2004).

1.8.2.8. PIK4CA

A screening of *PIK4CA* for polymorphisms identified rare variants at a splice site and the promoter region in three patients from the United States with bipolar disorder, three patients with schizophrenia and one control. There was no difference

between patients with a psychiatric diagnosis and controls for the variants, however the authors report a trend towards significance in the distribution of promoter genotypes in bipolar disorder patients (Saito, Stopkova et al. 2003).

1.8.2.9. Synapsin III

Another PI candidate gene is *synapsin III*, which encodes for an intrinsic synaptic vesicle membrane protein. *Synapsin III* was highlighted in the genome-wide association study, previously described in section 1.5.3, performed by the WT-CCC on 2,000 bipolar disorder cases and 3,000 controls, as SNP rs11089599 in *synapsin III* was among one of the higher ranked signals (P=7.2x10^{-5}) (WTCCC 2007). Previous linkage studies in families with schizophrenia have shown evidence for chromosome 22q12-p13, where *synapsin III* is located (reviewed in (McGuffin, Owen et al. 2002) but this result was not supported by a multicentre linkage study (Mowry, Holmans et al. 2004). An association study has reported an increase in the number of African American individuals (124) with schizophrenia that are homozygous for a SNP in *synapsin III*, compared to homozygous controls (P=0.04) (Lachman, Stopkova et al. 2005).

1.8.2.10. PI4K2B

Finally, genetic evidence suggesting *PI4K2B*, the focus of this research, as a candidate gene in psychiatric illness was provided by results of a large-scale association study, as described in section 1.7.3 (Christoforou, Le Hellard et al. 2007). Tagging SNPs from the genomic region containing *PI4K2B* were used in the case-control association study on 368 bipolar disorder individuals, 386 individuals diagnosed with schizophrenia and 458 control individuals. The association study identified SNP rs1093903, which is located in the same haplotype block as *PI4K2B*, 100kb upstream of the coding region, as a potentially important variant in schizophrenia (allele P=0.006, odds ratios: 1.314, 95% CI 1.08-1.59), but not bipolar

disorder. This result did not meet the region-wide nominal significance threshold, $P \leq 0.0003$ but warranted further investigation.

Incompatible with the above genetic evidence suggesting a role of PI genes in susceptibility to bipolar disorder, a recent family-based association study of lithium-related and other candidate genes in bipolar disorder has not supported this notion. This study examined a dense set of haplotype-tagging SNPs using a gene-based test of association in 379 US trios. No genes that were specifically chosen to probe lithium were associated with bipolar disorder and hence, found that no variants in lithium-related genes contribute to susceptibility to bipolar disorder. However, only two of the genes listed in Table 1.8 were tested, namely *GSK3β* and *IMPA2* and did not show association (Perlis, Purcell et al. 2008).

In summary, the above positional genetic evidence for *PI4K2B, IMPA2, PIP5K2A, SYNJ1* and *PIK4CA* support a possible role for PI genes as bipolar disorder susceptibility genes. Further evidence from functional studies also supports this notion.

1.8.3. Functional evidence for phosphoinositide genes in bipolar disorder

1.8.3.1. Phosphoinositide expression differences in bipolar disorder

Several investigations in bipolar disorder patients have suggested abnormalities in the PI pathway. One study reported that levels of the scaffolding molecule PIP_2 are significantly increased in platelet membranes of drug-free depressed bipolar disorder patients. However, there was no difference in the levels of PI and PIP in bipolar disorder patients and healthy individuals (Soares, Dippold et al. 2001). A second study showed that a decrease in PIP_2 resulted in synaptic vesicle defects (Di Paolo, Moskowitz et al. 2004). Other studies show altered inositol monophosphatase

(IMPase) enzyme activity and *IMPA1* and *IMPA2* expression levels in lymphoblastoid cell lines of bipolar patients, particularly in lithium responders (Shamir, Ebstein et al. 1998; Nemanov, Ebstein et al. 1999; Yoon, Li et al. 2001). Notably, these studies are in platelet membranes (Soares, Dippold et al. 2001) and lymphoblastoid cell lines (Shamir, Ebstein et al. 1998) and may not truly represent neuronal function.

1.8.3.2. Effect of lithium on phosphoinositide signalling

Lithium is recommended by the WHO as a prophylaxis against bipolar disorder and can be effective in reducing the frequency and severity of manic and depressive episodes. Despite its chemical simplicity, the complex pharmacology and molecular mechanisms underlying the therapeutic actions of lithium have not been elucidated. The primary hypothesis of the lithium effect is the inositol hypothesis (Berridge, Downes et al. 1989). This idea proposes that lithium affects the PI signal transduction pathway to deplete inositol, as illustrated in Figure 1.7. It has been shown that lithium acts as a potent non-competitive inhibitor of myo-inositol monophosphatase (IMPase) (the rate limiting enzyme in inositol recycling) leading to inositol depletion (Hallcher and Sherman 1980; Berridge, Downes et al. 1989). This possibly causes abnormal activation of Calcium (Ca) Ca^{2+}-mobilizing receptors. The upstream inositol polyphosphate-1 phosphatase (IPPase) has also been identified as an additional target for lithium (Majerus 1992; Acharya, Labarca et al. 1998). This in turn, decreases the amount of phosphatidylinositol-4, 5-bisphosphate (PIP_2), available for signalling cascades that rely upon this pathway, for example synaptic vesicle trafficking. The signalling lipid, $PI(3,5)P_2$ is also critical for survival of neural cells (Zhang, Zolov et al. 2007). A recent study has shown that lithium regulates Akt/glycogen synthase kinase 3 (GSK3) signalling by disrupting a signalling complex composed to Akt, β-arrestin 2 and protein phosphatase 2A. This is interesting as β-arrestin 2 promotes formation of signalling complexes allowing G protein-coupled receptors (GPCR) to signal independently from G proteins, and as

the authors suggest, the β-arrestin complex may represent a target for drug intervention aimed at the regulation of GPCR signalling (Beaulieu, Marion et al. 2008)

Valproate, also a mood stabiliser recommended by the WHO, is proposed to act by the similar mechanism of inositol depletion, as illustrated in Figure 1.7. Functional evidence shows that chronic valproate treatment decreases inositol and increases inositol monophosphate (IP_1) concentration in the rat brain (O'Donnell, Rotzinger et al. 2000), by inhibiting inositol synthase (*ino-1*) (Williams, Cheng et al. 2002).

Additionally, lithium, valproate and another mood stabilising drug, carbamazepine, have been shown to inhibit the collapse of growth cones and increase the growth cone area in sensory neurons. These effects were reversed by inositol supplementation, again suggesting inositol depletion, as an important target for the treatment of bipolar disorder (Williams, Cheng et al. 2002).

1.8.3.3. Phosphoinositide signalling in neuronal physiology

The phosphorylation of inositol phospholipids is important in cellular regulation and synaptic vesicle trafficking for normal neuronal physiology. The generation of PIP from PI is an important precursor step. On synaptic vesicles, the membrane associated α isoform of PI4K, PI4KIIα is responsible for the majority of PI 4-kinase activity in the brain and is concentrated at the synapse and in the region of the Golgi complex in neuronal perikarya (Guo, Wenk et al. 2003). PI4KIIα is thought to be important in synaptic vesicle trafficking and downstream PI intracellular signalling (Guo, Wenk et al. 2003). The important signalling and scaffolding molecule, PIP_2, has been shown to regulate synaptic transmission by modulating Ca^{2+} signalling through its metabolic products, inositol triphosphate (IP_3) and diacylglycerol (DAG). A decrease in PIP_2 levels in the brain and impairment of PIP_2 synthesis in nerve terminals has been shown to cause postnatal lethality and synaptic defects in

mice, including enhanced synaptic depression and delayed endocytosis (Di Paolo, Moskowitz et al. 2004).

Therefore, the above positional and functional evidence support a possible role for PI genes as bipolar disorder susceptibility genes. In addition, the studies on PI genes and the lithium hypothesis suggest a model where functional mutations in genes affecting PI signalling pathway may cause an abnormality in neuronal PI4K homeostasis. On the basis of this hypothesis, further research in the genes involved in PI metabolism as candidates for psychiatric illness is reasonable, in particular the positional and functional candidate *PI4K2B*. In addition, it validates my approach to perform expression analysis, to examine the polymorphisms and resulting haplotypes for association analysis on this positional candidate gene involved in the PI pathway.

1.8.4. Phosphatidylinositol 4 kinases

1.8.4.1. Characteristics of PI4K2B

Four phosphatidylinositol 4 kinases (PI4K) have been identified. The four PI4K RefSeq genes are *PI4K2B* (NM_018323), *PI4K2A* (NM_018425), *PI4KA* (NM_058004) and *PI4KB* (NM_002651). Figure 1.8 illustrates *PI4K2B* as a gene of 45,000 base pairs (bp) with ten exons. The mRNA is 3458 bases long, the protein size is 481 amino acids and 54,754 Daltons (Da). Figure 1.8 also depicts the position of the PI3_PI4_kinase domain from amino acid 167 to amino acid 415.

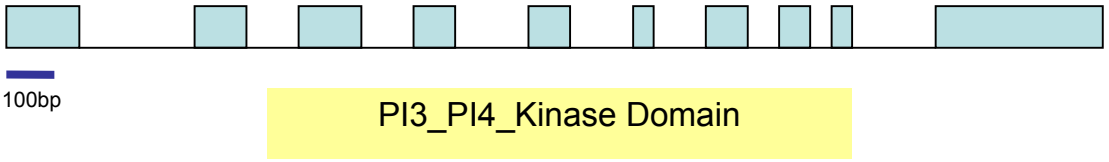


Figure 1.8 *PI4K2B* gene. The blue bars represent the 10 exons of *PI4K2B* in proportion to their actual size. The dark blue scale bar represents 100bp. The phosphatidylinositol (PI) kinase domain is highlighted.

PIK42B is a type II kinase and is highly homologous (58% identical and 75% homologous) to PI4KIIA, apart from a unique N-terminal 100 amino acid sequence.

1.8.4.2. Function of *PI4K2B*

PI4K2B is phosphatidylinositol 4-kinase and a member of the PI signal transduction pathway, as illustrated in Figure 1.7. Its primary function is phosphorylation of phosphatidylinositol (PI) to generate phosphatidylinositol 4-phosphate (PIP). PIP is an immediate precursor of important signalling and scaffolding molecules, such as phosphatidylinositol 4,5-bisphosphate (PIP_2) (Wei, Sun et al. 2002). The phosphoinositide cascade mediates the transmission of numerous hormonal signals through second messengers, diacylglycerol (DAG) and inositol-1,4,5-triphosphate (IP_3), stimulating the release of the calcium ion (Ca^{2+}). This in turn, modulates the activity of many intracellular events (Gould, Quiroz et al. 2004) to spatially and temporally regulate signalling, cytoskeletal dynamics and membrane trafficking (Wei, Sun et al. 2002). However, there is controversy over the enzymatic ability of PI4K2B as "membrane-associated PI4K2B is as active as membrane-associated PI4K2α and has essentially identical kinetic properties" (Wei, Sun et al. 2002) to the contrary "most striking thing about PI4K2β is that it seems to be virtually inactive enzymatically compared to the α isoform." (Shane Minogue, Personal Communication, 31st January 2006, University College London)

1.8.4.3. Location of PI4K2B

Balla *et al,* 2002 have cloned *PI4K2B* and showed that over-expression of *PI4K2B* co-localises with overexpressed *PI4KII*α in endosomal vesicles, but not in the plasma membrane or the Golgi apparatus (Balla, Tuymetova et al. 2002). Their data also indicate the existence of multiple forms of PI4K2B in mammalian cells (Balla, Tuymetova et al. 2002). Contrary to this, Wei *et al* have shown that PI4K2B is a predominantly cytosolic protein and is recruited to the plasma membrane by the platelet-derived growth factor (PDGF) and mediated through the small GTPase, Rac in a Guanosine-5'-triphosphate (GTP)-dependent manner (Wei, Sun et al. 2002). Rac simultaneously recruits and activates PI4K2B and PIP5K, a phosphatidylinositol 4-phosphate 5-kinase, which converts PIP to PIP_2, at the membrane. This pathway is very important in intracellular second messenger signalling, for example through Ca^{2+} and for receptor dependent phospholipase C (PLC) and PI3K signalling. This suggests that PI4K2B synthesizes PIP and PIP_2 to increase cellular trafficking in membrane ruffling, endocytosis and exocytosis (Wei, Sun et al. 2002).

1.9. Expression Analysis

1.9.1. Allele-specific expression

Allele-specific differences in gene expression have been reported in epigenetic phenomena of X-chromosome inactivation and genomic imprinting. Investigations on the extent of gene expression in populations, have suggested that allele-specific expression is also common in non-imprinted autosomal genes and that these allele-specific differences are heritable (Lo, Wang et al. 2003; Pastinen, Sladek et al. 2004). Allele-specific differences in gene expression can be explained by *cis*-acting genetic variations causing differential expression between alleles. A *cis*-regulatory region is a segment of DNA that regulates transcription of a gene. For example, segments can typically lie immediately 5'- of the start site of transcription, or can be discontinuous or individual segments within introns, 5'- and 3'-untranslated regions (UTRs) or tens of kilobases on either side of the gene they regulate (Wray 2007). The search for

these genomic variations is helped when the haplotype structure of the locus has been defined (Pastinen, Ge et al. 2005). The differences can be determined using variation within the coding regions, to distinguish the relative abundance of transcript arising from the two alleles. For example, when cells from an individual who is heterozygote for a SNP are analysed, the allelic origin of the transcript can be identified. It is unlikely that the coding SNP will be functional but it may be used as a marker to distinguish the relative abundance of the two alleles.

A screen for allele-specific differences in transcript abundance in human fetal liver and kidney tissue was reported using a microarray platform (Lo, Wang et al. 2003). This high-throughput analysis showed a twofold difference in transcript abundance between alleles in at least one individual, out of seven individuals, in 54% of the 602 genes studied, and a fourfold difference in 28% of the genes. The majority of the differentially expressed genes were not in known imprinting domains and were distributed throughout the genome. The authors suggested that variation of gene expression between alleles is common, and this variation may contribute to human variability (Lo, Wang et al. 2003).

Pastinen et al, 2003 similarly reported differential expression of transcripts in heterozygous samples by quantifying intragenic marker alleles in messenger (mRNA) or heteronuclear RNA (hnRNA). In the study, they used 193 SNPs from 129 genes expressed in lymphoblastoid cell lines, to identify 23 genes (18%) with common transcripts whose expression from the two alleles deviated from the expected equimolar ratio. Further study of three genes, in samples from a three-generation pedigree showed patterns of transmission, including co-segregation of allelic skewing across generations, compatible with Mendelian inheritance and random monoallelic expression (Pastinen, Sladek et al. 2004). Thus, the authors provide evidence for widespread allele-specific differential gene expression and inheritance of this variation in some instances.

Allelic differences in expression have been reported in candidate genes for bipolar disorder and schizophrenia. First, the bipolar disorder candidate gene encoding the winged-helix transcription factor *RFX2*, has shown evidence for significant allelic differences in expression ($P<0.001$) in post-mortem brain tissue cDNA (Glaser, Kirov et al. 2005). Second, assays of allelic expression on the schizophrenia susceptibility gene *dystrobrevin binding protein 1, DTNBP1,* have shown a relative reduction in mRNA expression in human cerebral cortex ($P<0.0001$), possibly mediated through *cis*-acting variants tagged by defined schizophrenia risk haplotypes (Bray, Preece et al. 2005). Last, allelic expression analysis on the schizophrenia candidate gene *GNB1L,* a G-protein beta-subunit-like polypeptide, showed evidence for differential allelic expression of *GNB1L* transcripts ($P=0.0006$) in post-mortem brain tissue cDNA. The study also found evidence that markers associated with psychosis are also correlated with the presence of *cis*-acting influences on alterations in *GNB1L* expression (Williams, Glaser et al. 2008).

Moreover, expression differences can also be detected at the protein level. For example, a microarray study which investigated the effect of lithium on gene expression in a human neuronal cell line identified as *Peroxiredoxin 2 (PRDX2)*, an antioxidant enzyme, as the most upregulated gene and *tribbles homolog 3 (TRB3)*, a pro-apoptotic protein, as the most down-regulated gene upon lithium treatment. This difference was also detected at the protein level by Western blotting technique, ensuring greater confidence in the result (Seelan, Khalyfa et al. 2007).

1.9.2. Lymphoblastoid cell lines as a cellular model

The ability to perform expression analysis studies depends on the availability of pertinent biological tissue. Fortunately, lymphoblastoid cell lines were available from 50 individuals in the aforementioned large Scottish family. There are many advantages to using lymphoblastoid cell lines as an experimental resource. Firstly, this resource is exceptional because of the number of cell lines available within one

family. Secondly, they provide immense experimental potential, including a limitless supply of DNA. Thirdly, a postulated advantage is minimising the influence of medication, as the cell lines can be cultured for more than one month (Washizuka, Kakiuchi et al. 2005).

Importantly, expression analysis on these blood-derived cell lines can be extended to model brain tissue. There is evidence for altered expression of neuronal receptor proteins and neurotransmitter systems in lymphocytes, and for similarities of hormonal effects on nervous system processes and lymphocyte physiology in schizophrenia, recurrent major depression and related phenotypes (Gladkevich, Kauffman et al. 2004). Microarray expression analysis between brain and whole blood samples shows that whole blood shares significant gene expression similarities with multiple CNS tissues, which is useful for gene expression studies. For example, from a total of 45 schizophrenia candidate genes studied, the expression was correlated in 21 genes between the blood and prefrontal cortex of the brain. This research also showed that correlation between blood and brain tissues was good for certain processes: including axon guidance, neurogenesis, G-protein coupled receptor protein signalling and intracellular signalling cascade and would be useful (Sullivan, Fan et al. 2006). Finally, lymphoblastoid cell lines have been used as a cellular model to study complex human diseases such as bipolar disorder, hypertension, diabetes mellitus, Alzheimer's disease and Huntington's disease [as reviewed in (Gladkevich, Kauffman et al. 2004; Iwamoto, Kakiuchi et al. 2004).

However, there are disadvantages associated with lymphoblastoid cell lines such as a risk of structural genomic alterations due to the Epstein Barr Virus (EBV) immortalisation process, lymphoblastoid cell lines creation or prolonged cell culturing (Simon-Sanchez, Scholz et al. 2007). Another drawback is not all genes are expressed in lymphoblastoid cell lines or behave as they would *in vivo*. Furthermore, lymphoblastoid cell lines are non-neuronal cells so it is difficult to

extend any finding in lymphoblastoid cell lines to brain disorders such as bipolar disorder. The aforementioned microarray expression study between brain and whole blood samples also showed that correlation in gene expression was not good for certain processes. For example, the study of brain oxygen transport, RNA binding and DNA binding in blood tissue would not reflect the processes in brain tissue (Sullivan, Fan et al. 2006). Thus, when the relevant gene is expressed in both blood and brain, any interpretation of experimental data from lymphoblastoid cell lines should be cautious.

Several intermediate phenotypes of bipolar disorder have been reported at the lymphoblastoid cell level. Many reports relate to abnormal intracellular calcium metabolism, phosphoinositides and cAMP signalling (Kato, Ishiwata et al. 2003). This suggests that certain biochemical pathways involved in bipolar disorder can be studied using non-neuronal lymphoblastoid cell lines. Furthermore, differentially expressed genes (*IMPA2, TRPC7, NDUFV2*) have been reported in lymphoblastoid cell lines from patients with bipolar disorder (Washizuka, Kakiuchi et al. 2003; Washizuka, Kakiuchi et al. 2005). Also, candidate genes *HSPF1* and *LIM* showed altered gene expression in both post-mortem brains and lymphoblastoid cell lines from Japanese patients with bipolar disorder (Iwamoto, Bundo et al. 2004). A reported association of a mitochondrial complex I subunit gene, *NDUFV2* at chromosome 18p11 with bipolar disorder (Washizuka, Kakiuchi et al. 2003), was substantiated by a decrease in *NDUFV2* and other mitochondria-related genes mRNA expression in lymphoblastoid cell lines from patients with bipolar disorder (21) compared with controls (11) (Washizuka, Kakiuchi et al. 2005). Lymphoblastoid cell lines can also be used to investigate the effects of bipolar disorder treatment such as lithium. A examination of *in vitro* activity of a target of lithium, IMPase activity in lymphoblastoid cell lines derived from individuals with bipolar disorder with a positive response to lithium (56), with a poor response to lithium (11) and control samples (29) found that levels of IMPase indicate susceptibility to bipolar disorder and response to lithium treatment (Shamir, Ebstein et al. 1998). A study on

bipolar disorder affected individuals currently treated with lithium (9), lithium-free individuals (13) and healthy control subjects (23) found that levels of the lithium target, GSK3β, previously discussed in section 1.8.2.1, were eightfold higher in peripheral blood mononuclear cells from lithium-treated bipolar disorder individuals than healthy control individuals (Li, Friedman et al. 2007).

Investigations of the allelic expression of *PI4K2B* in lymphoblastoid cell lines were used to test the hypothesis of this study, that impairment of phosphatidylinositol phosphorylation by altered expression levels may lead to disturbances in cell signalling implicated in development of bipolar disorder.

1.10. Aim of the Study

This study tested the hypothesis that there is a genetic risk factor for susceptibility to the psychiatric illnesses, bipolar disorder and recurrent major depression in a large Scottish family. The aims were twofold i) investigation of a candidate gene, *PI4K2B*, for susceptibility to bipolar disorder using expression studies in the family and in a case-control association study and ii) whole genome linkage analysis of the family with increased marker coverage and updated diagnoses.

- In Chapter 3, expression analysis of *PI4K2B* was conducted in lymphoblastoid cell lines from the family, at the allele-specific RNA level and the protein level.

- In Chapter 4, a case-control association study tested association of markers in the *PI4K2B* region with bipolar disorder and schizophrenia in the Scottish population

- In Chapter 5, the data for the whole genome linkage scan was prepared and verified, by performing relatedness analysis on the family, error-checking the genotyping data and sub-dividing the large family to allow linkage analysis.

- In Chapter 6, the linkage evidence for bipolar disorder and recurrent major depression in the family was evaluated by parametric linkage analysis, non-parametric linkage analysis and haplotype analysis.

Chapter 2
Material & Methods

2. Material & Methods

2.1. General Molecular Biology Material

A collection of cell lines and DNA samples were available for this study and are detailed in section 2.1.1. For the molecular biology methods detailed in section 2.2, the general chemicals were obtained from Sigma, unless otherwise noted. All PCR primers, cell culture media and supplements were supplied by Invitrogen. Section 2.1.2 details the protocol for preparing solutions, buffers and materials. Table 2.1 lists the kits used for molecular biology techniques.

2.1.1. DNA

The DNA samples available to this study are detailed in this section. There were DNA samples from a large Scottish family, lymphoblastoid cell lines from the same Scottish family, a panel of DNA samples from four families, a selection of trios and a cohort of unrelated cases and controls. The details of the collection method are listed below. The phenotyping methods for all patients was through interview using the semi-structured instrument "Schedule for Affective Disorder and Schizophrenia-Lifetime" version (SADS-L) (Endicott and Spitzer 1978) and supplementary clinical information was also available from hospital case notes and from relations of cases. Diagnosis was determined by criteria from the Diagnostic and Statistical Manual of Mental Disorders (4th Edition) (DSM-IV) for bipolar disorder and schizophrenia. All diagnoses were reviewed independently by two experienced psychiatrists and a consensus diagnosis was reached where necessary. A peripheral venous blood sample (20ml) was obtained for DNA extraction and separation of lymphocytes to establish transformed cell lines by European Collection of Cell Cultures (ECACC). DNA samples were stored on 96-well plates and archived in the Wellcome Trust Core Research Facility (WT-CRF), Western General Hospital, Edinburgh.

2.1.1.1. DNA samples from Scottish family

DNA samples, extracted from blood samples prior to this study, were available from the large Scottish family, which has previously named as F22 (Blackwood, He et al. 1996). The family was ascertained at a local psychiatric hospital where an aunt and niece had been admitted at the same time, each with bipolar disorder. The researchers obtained informed consent from both probands to contact other family members. The whole family was gradually recruited over several years. Some individuals were interviewed on more than one occasion. Linkage analyses were performed under two definitions of the disease phenotype: broad and narrow. For the broad definition, the phenotype included bipolar I disorder, bipolar II disorder and recurrent major depression cases. The narrow definition included only cases with a bipolar I disorder and bipolar II disorder diagnosis and all other psychiatric diagnoses were labelled as unknown

2.1.1.2. Lymphoblastoid cell lines

Lymphoblastoid cell lines were established from members of the large Scottish family. The cell lines were available from ECACC. All individuals gave consent to take part in these studies.

2.1.1.3. Allele sharing panel

The Allele Sharing Panel was a collection of 46 DNA samples from individuals in the four families that show linkage of bipolar disorder on chromosome 4. There were 31 samples from the large Scottish family (F22), seven samples from F50, three samples from F48, five samples from F59 (Blackwood, He et al. 1996) a control DNA pool and a negative control. This represented multiple copies of the linked haplotype to chromosome 4p15-16. This enabled unambiguous determination of the linked haplotype and 38 control chromosomes in region B and 37 control chromosomes in region D (defined as all non-disease chromosomes). Please refer to Figure 1.6.

2.1.1.4. Trio bipolar panel

The Trio Bipolar Panel is a collection of DNA samples from 96 individuals that comprise 32 proband-parent trios. A trio is an affected offspring (bipolar disorder) and both parents. The panel were recruited from the south Scottish population and are all of North European ancestry.

2.1.1.5. Association study sample

The study sample consists of DNA from 444 control individuals, 362 bipolar disorder patients and 383 schizophrenia patients. These studies were approved by the Scottish MultiCentre Research Ethics Committee (MREC) and patients and controls gave written informed consent. Individuals suffering from bipolar disorder and schizophrenia were recruited from The Royal Edinburgh Hospital and other Scottish psychiatric hospitals by trained psychiatrists. Samples were also obtained from Fyffe, The Borders and Lanackshire through collaborations with psychiatrists based in these regions.

Control subjects were drawn from the same population in South East and South Central Scotland. The majority (391) were recruited through the Scottish National Blood Transfusion service. Although the blood donors were not screened by interview for personal or family history of psychiatric illness, donors are only allowed to donate blood if they are not currently on medication and had no chronic illness. The remaining controls (79) were recruited from the local population and from hospital staff. These controls were briefly screened by interview to exclude anyone currently on medication or with a history of treatment for psychiatric illness. The samples were drawn from the Scottish population.

2.1.2. Solutions, buffers and gel-loading dyes

10% Acrylamide gel

Separating gel	
30% Acrylamide	2.6ml
1.5M Tris pH8.8	2ml
dH_2O	3.34ml
20% SDS	40µl
TEMED	20µl
*leave to set	
Stacking gel	
30% acrylamide	870µl
0.5M Tris pH6.8	1.5ml
dH_2O	3.75ml
20% SDS	30µl
TEMED	3µl
25% APS	30µl

25% APS

Ammonium persulfate	0.25g
Qs 5ml dH_2O	

Cell freezing medium

RPMI or DMEM depending on cell type
10% Fetal Bovine Serum
10% DMSO
Aliquot and store at -20°C.

Coomassie Brilliant Blue protein stain

Coomassie stain	0.25g
Methonal: Water (1:1 v/v)	90ml
Glacial acetic acid*	10ml

Filter solution and store at room temperature (RT)
*Always add acid to water

Coomassie de-stain (Methanol:Acetic acid)

Glacial acetic acid	100ml
Methanol	300ml

Make to 1L with distilled water, store at RT

1M DTT

Dithiothreitol	0.7g
Qs 5ml dH_2O	

95% Ethanol

5ml of deionised water (dH_20) added to 95ml ethanol (99.7-100% v/v), mixed and stored at room temperature (RT).

70% Ethanol
30ml dH_20 added to 70ml ethanol (99.7-100% v/v), mixed and stored at RT.

10µg/100ml Ethidium Bromide
1µl of 10mg/ml ethidium bromide solution was added to 100ml dH_20. The solution was mixed and stored at RT.

Orange G Loading Dye

Ficoll 400	3g (15%)
SDS (20% solution)	50µl (0.05%)
EDTA (0.5M)	800µl (20mM)
Orange G	0.125%

Qs (Quantity sufficient) to 20ml dH_20, store at RT.

PBS/BSA
0.06g Bovine Serum Albumin in 300ml PBS

Ponceau stain

Ponceau S	0.5g
Acetic Acid*	2ml

Qs 100ml, store at RT.

PBST blocking buffer (Western Blotting)
1x PBS, 5% semi-skimmed milk (Marvel) 0.2% Tween20.

Protease inhibitors (Complete)
1 tablet dissolved in 1 ml dH_20, aliquot and store at -20°C.

Protein sample buffer (2X Laemmli)

Tris (hydroxymethyl)-methylamine (0.5M, pH6.8)	2ml (100mM)
Glycerol	2ml (20%)
SDS (20% solution)	2ml (4%)
Bromophenol Blue	0.2g (0.02%)

Qs 10ml dH_20, store at RT. Add 100mM DTT before use.

RIPA buffer

Tris (hydroxymethyl)-methylamine (50mM, pH7.4)	0.61g
Sodium chloride (NaCl, 150mM)	0.9g
dH_20	80ml
Stirred to dissolve, pH7.4 with 5M HCl	
1% Trition	1ml
Sodium deoxycholate (10% solution)	2.5ml
Sodium dodecyl sulphate (20% solution)	250µl
Stirred to dissolve	
1mM EGTA (100mM solution)	1ml

Qs 100ml dH_20, 5ml aliquots, store at 4°C. Add 100µl protease inhibitors to 5ml RIPA lysis buffer before use.

REPA lysis buffer

10% Sodium deoxycholate	2.5ml
Nonidet p40 (aka IGEPAL CA-630)	0.5ml
20% SDS	250µl

Qs PBS 50ml
Add 1 tablet of complete protease inhibitors, aliquot and freeze at -20°C.

Semi-dry transfer buffer

Tris(hydrosymethyl)-methylamine	(48mM) 5.8g
Glycine	(39mM) 2.9g
20% Sodium dodecyl sulphate	(0.04%) 2ml
Methanol	(20%) 200ml

Qs to 1L dH_20, pH9-9.4, store at 4°C.

3M Sodium Acetate
8.16g sodium acetate *qs* 20ml dH_20. *adjust the pH to 5.2.

Sodium azide
5g NaN_3 in 50ml dH_2O.
Use at 1:100 PBS

1x TBST (Tris buffered saline with Tween)

1MTris, pH 7.5	50ml
Sodium chloride	8.75g

Qs 1L dH_20, add 1ml Tween20 (0.1%)

TBST blocking buffer
1X TBST with 1% Marvel

1% Trition Lysis Buffer

PBS	10ml
20% Trition	500µl
Protease Inhibitors	200µl

1.5M Tris pH 8.8
181.5g Tris (hydroxymethyl)–methylamine in 1L dH_20, *adjust pH and store at RT.

0.5M Tris pH 6.8
60.5g of Tris (hydroxymethyl)–methylamine in 1L dH_20, *adjust pH, store at RT

20X TBE Buffer

Tris	242.0g
Boric Acid	123.4g
Disodium EDTA	14.88g

Qs 1L ml dH_20. The buffer was autoclaved to sterilise before use and stored at RT.

1X TBE Buffer

20X TBE	500ml
dH_20	9.5L

The solution was mixed and stored at RT.

2.1.3. Kits

METHOD	KIT	SOURCE
cDNA preparation	First Strand cDNA Kit	Roche
gDNA preparation	DNeasy Tissue Kit	Qiagen
Gel drying	DryEase® Mini-Gel Drying System	Invitrogen
Gel extraction	QIAquick gel extraction kit	Qiagen
Gel staining	SimplyBlue™ SafeStain	Invitrogen
Immunodetection	ECL Plus	Amersham Biosciences
Membrane Transfer	X-Cell SureLock™MiniCell	Invitrogen
PCR Amplification	10X PCR Enhancer	Invitrogen
PCR Amplification	PCR Core Kit	Sigma
PCR Amplification	PCR Buffer System	Applied Biosystems
Plasmid purification	EndoFree Plasmid Maxi Kit	Qiagen
Protein concentration estimation	BCA Analysis	Pierce
Protein electrophoresis	NuPAGE ®Novex Bis Tris gels	Invitrogen
RACE	FirstChoice® RNA-Ligase-Mediated	Ambion
RNA extraction	RNeasy Mini Kit	Qiagen
Whole genome amplification of DNA	REPLI-g Midi Kit	Qiagen

Table 2.1 Kits used in molecular biology techniques.

2.2. General Molecular Biology Methods

Molecular biology techniques were developed from Sambrook et al, 1989, (Sambrook, Fritsch et al. 1989) unless otherwise stated.

2.2.1. Cell culture

2.2.1.1. General cell culture

All mammalian cells were classified as containment level one and were manipulated using aseptic technique in Envair Bio2+ Class II safety cabinets. Cells were grown under humid conditions (37°C, 5% Carbon dioxide) in the appropriate growth medium. T.25, T.75 and T.175 tissue culture flasks with filter caps were used to maintain cell cultures (CellStar®, Greiner Bio-One).

2.2.1.2. Lymphoblastoid cell culture

Lymphoblastoid cell lines were cultured with RPMI 1640 (Sigma) containing 10% Fetal Calf Serum (Invitrogen), 1:100 penicillin/streptomycin (Invitrogen), 1:100 L-Glutamine, 1:500 1.25M MOPS (Sigma). Cells were maintained by feeding every three-four days. It is recommended that lymphoblastoid cell cultures be maintained at a density of 3-9x10^4 cells per ml (www.ecacc.org.uk). For three days prior to harvesting cells, the culture medium was changed every day in order to randomise the cell cycle (Washizuka, Kakiuchi et al. 2005). Healthy lymphoblastoid cells grew in aggregates which settled at the bottom of the tissue culture flask. To replenish medium or passage cells, the flasks were removed from the incubator into the safety cabinet with minimum agitation to avoid disrupting the cell aggregates. The medium was removed using a sterile glass pipette. The medium was then either replenished or the cells were split by removing an appropriate amount of cells and placing them into a new flask of fresh medium. Once in fresh medium, the cells were agitated to dissociate clumps and promote growth. All cell lines were

subjected to the same cell culture and experimental conditions, to minimise any changes to gene or protein expression.

2.2.1.3. Adherent cell culture

COS7 cells were cultured. These adherent cells were from monkey kidney fibroblasts (Medical Genetics, Molecular Medicine Centre). Cell medium (DMEM, 10% FBS) was replenished every three-four days as required. Cells were passaged at 70-90% confluency. To passage cells, growth medium was aspirated using a sterile glass pipette and cells were washed in D-PBS (GIBCO®) to remove any remaining serum. Cells were incubated with a 1:1 solution of versene plus trypsin (GIBCO) at 37°C to lift them from the flask surface. The versene:trypsin solution was inactivated by adding an equal volume of growth medium containing 10% fetal bovine serum. Cells were centrifuged gently at 1,000rpm for five minutes (CR312 Jouan centrifuge) at RT and then re-seeded into new medium. Cells were re-seeded at 1 in 10 to 1 in 50 dilutions.

2.2.1.4. Preparation and recovery of cell stocks

Frozen cell stocks were prepared by centrifuging an appropriate amount of cells (4-9x10^6 lymphoblastoid cells or 1-2x10^6 adherent) for five minutes at 1,000rpm. Cell pellets were resuspended in 1ml cell freezing medium and placed in a CryoTube™ (NUNC™). Cells were frozen using a Cryo 1°C freezing container (NALGENE). The freezing chamber was filled with 100% ethanol or propan-2-ol and the CryoTube was placed in the holder provided. This was stored at -70°C overnight. Using this container, the cells were frozen slowly at a rate of -1°C/minute. The cells were then placed in liquid nitrogen for long term storage. Frozen cells were recovered by thawing quickly at 37°C and placing into a T.25 tissue culture flask with 5ml growth media. Cells were incubated overnight at 37°C, 5% CO_2 to recover. Following this, cells were resuspended in fresh medium and grown as normal.

2.2.1.5. Haemocytometer counting to determine cell number

The haemocytometer (Improved Neubar) was cleaned with 70% ethanol. A coverslip was placed over the haemocytometer chamber. Confluent cell suspensions were diluted 1 in 10 for counting. Each chamber of the haemocytometer was filled by capillary action using a Pasteur pipette to place a drop of the cell suspension on the chambers edge. A minimum of four 1mm square (divided into 16 smaller squares) were counted and the average cell number was calculated. Each square represented the number of cells $x10^4$/ml.

2.2.1.6. Mycoplasma test with Hoechst 33258

Lymphoblastoid cell lines were incubated with an equal volume of MeOH/Acetic Acid fix for five minutes. The cells were spun down, washed, resuspended in a small quantity of fix and placed on a glass slide. The slide was stained with 1µg/ml Hoechst 33258 (Invitrogen, Spectra) for 10 minutes and washed with dH_20. A coverslip with Vectashield mounting media (Vector labs, H-1400) covered the glass slide, secured with rubber sealant. The stained slides were analysed by fluorescent microscopy. Uncontaminated cultures showed only cell nuclei staining and there was no evidence of small cocci or filaments of mycoplasma-positive cultures.

2.2.1.7. Chemical transfection of mammalian cell lines

Transfection of Cos7 cells was performed by Lipofectamine™ 2000 (Invitrogen) according to the manufacturer's instructions, in duplicate, in 6-well COSTAR® flat bottom plates. There were controls included in the transfection procedure of PI4K2B-HA; no DNA, HA 14-3-3, MYC 14-3-3 and HA. PCR amplification of these preparations did not produce a product with PI4K2B specific primers.

2.2.2. Nucleic acid preparation

2.2.2.1. Genomic DNA preparation

Genomic DNA (gDNA) was isolated using the DNeasy Tissue Kit (Qiagen) as per the manufacturer's instructions. A maximum of $5x10^6$ cells were used. Briefly, cells were lysed using proteinase K, ethanol was added to the lysate, which was then applied to a spin column with a specialised silica-gel membrane that binds DNA. The contaminants were removed by washing the column and the DNA was eluted. DNA was eluted twice using 2x200µl volumes of AE buffer provided. The extracted gDNA was stored temporarily at 4°C or at -20°C for long-term storage.

2.2.2.2. RNA preparation

RNA was prepared using RNeasy Mini Kit (Qiagen) as per the manufacturer's instructions. Briefly, samples were lysed and homogenized in the presence of a highly denaturing guanidine-thiocyanate–containing buffer, which immediately inactivates RNases to ensure purification of intact RNA. Ethanol was added to provide appropriate binding conditions, and the sample was then applied to an RNeasy Mini spin column, where the total RNA bound to the membrane and contaminants were washed away. RNA was then eluted in 30–100µl dH_20. With this protocol, all RNA molecules longer than 200 nucleotides were purified, enriching for mRNA since most RNAs <200 nucleotides (such as 5.8S rRNA, 5S rRNA, and tRNAs, which together comprise 15–20% of total RNA) were excluded. The RNA was DNase treated to remove genomic DNA, by incubating DNaseI directly on the membrane at RT for 15 minutes. RNA preparations were immediately frozen on dry ice and stored at -70°C.

2.2.2.3. cDNA preparation

mRNA was reverse transcribed to cDNA with First Strand cDNA Kit (Roche) using random primers. This method produces single stranded RNA using AMV (Avian

myeloblastosis virus) reverse transcriptase. A typical cDNA synthesis reaction is described in Table 2.2.

REAGENT	VOLUME	FINAL CONCENTRATION
10X reaction buffer	4µl	1X
25mM Magnesium chloride	8µl	5mM
10mM dNTPs	4µl	1mM
Random Primers $p(dN)_6$	4µl	3.2µg
RNase inhibitor	2.5µl	50 units
AMV reverse transcriptase	1.6µl	20 units
Total RNA	X µl	<1µg total RNA
dH_20	X µl	
	20µl	

Table 2.2 cDNA synthesis reaction

A master mix containing the appropriate reagents was prepared, mixed and centrifuged before aliquoting into 0.5ml RNase free thin walled tubes. The reactions were then cycled on a thermal cycler (Peltier PTC-225) or separately on a heating block as shown in Table 2.3.

TEMPERATURE	TIME	PURPOSE
25°C	10 minutes	Primer annealing
42°C	60 minutes	Reverse transcription
99°C	5 minutes	AMV inactivation
4°C	5 minutes	

Table 2.3 cDNA incubation times

The cDNA was stored at -20°C. For all cDNA preparations a negative control reaction was prepared. This control included RNA, but did not include the reverse transcriptase enzyme. This was to ensure that no DNA contamination was present in the RNA sample which would interfere with downstream PCR applications. Negative controls containing dH_20 and reverse transcriptase were also prepared to ensure all reagents were free of contamination.

The cDNA was tested for genomic contamination by PCR with intron-spanning housekeeping primers [ANAPC4, WDR1, DDX15 (Invitrogen)] as the Sigma protocol. The cDNA sample without reverse transcriptase was also tested with the three Taqman assays used for PI4K2B expression analysis in chapter 3. Please see Table 2.4 for primer sequences.

2.2.2.4. Gel extraction of DNA

DNA was extracted from agarose gel to isolate PCR products for sequence analysis. PCR products were extracted from low melting point (LMP) agarose gels using the QIAquick gel extraction kit (Qiagen) with the microfuge, according to the manufacturers' instructions. In brief, gel slices containing the desired PCR product were dissolved at 50°C in the appropriate buffer. An appropriate volume of isopropanol was added and the mixture was then applied to a spin column where the DNA bound to a silica membrane. After impurities such as salt and agarose were washed away, the DNA was eluted and ready for use.

2.2.2.5. Transformation of PI4K2B-HA plasmid

The DNA was provided by Tamas Balla on 6th December 2005, NICHD, Bethesda, MD, USA. The DNA was extracted from filter paper with 20µl sterile water (Invitrogen) and incubated for 14 hours. The DNA was transformed into Subcloning Efficiency™ DH5α™ competent cells (Invitrogen) using the heat shock protocol, as per manufacturers' instructions. Briefly, the DH5α cells were thawed on ice, 25µl per reaction. 2µl DNA was added and mixed with the cells. The samples were incubated on ice for 30 minutes. The cells were heat-shocked for 20 seconds in a 42°C water bath, then on ice for 2 minutes. The cells recovered in 1ml L-broth with glucose at 37°C for 30 minutes before spreading the transformed DNA on plates. The transformed DNA was plated on kanamycin (50mg/ml). DNA was prepared from one colony. The DNA plasmid was purified by EndoFree Plasmid Maxi Kit (Qiagen) as per manufacturers' instructions. In brief, the EndoFree plasmid

purification procedure was based on the selectivity of QIAGEN Resin that purified ultrapure supercoiled plasmid DNA with high yields. The plasmid purification protocol was based on a modified alkaline lysis procedure, followed by binding of plasmid DNA to QIAGEN Anion-Exchange Resin under appropriate low-salt and pH conditions. RNA, proteins, dyes and low-molecular weight impurities were removed by a medium-salt wash. Plasmid DNA was eluted in a high salt buffer and then concentrated and desalted by isopropanol precipitation. Finally, the correct plasmid was verified by digest with EcoRI and XmaI. The restriction sites were determined using NEBcutter version 2, Table 2.8.

2.2.3. Polymerase chain reaction

2.2.3.1. Primer design

Genomic sequence information was obtained from the University College Santa Cruz (UCSC) genome browser. Information on features of chromosome 4p14-p16 sequence was obtained from the in-house browser ACeDB. The Primer3 design programme was used (Rozen and Skaletsky 2000). Primers were designed from repeat and SNP free regions of sequence. The repetitive sequences were screened out using Repeat Masker. Specificity of the primer sequence was checked using UCSC In-silico PCR (Kent, Sugnet et al. 2002) which searches a sequence database with a pair of PCR primers. The primers were synthesized by Invitrogen Life Technologies. The primers were tested to establish optimal PCR conditions on DNA samples using Sigma, Invitrogen and Applied Biosystems methods. Primer details and specific PCR conditions are listed in Table 2.4. All primers were diluted to 100μM with dH_20, then to a 10μM working stock and stored at -20°C.

NAME	PRIMER SEQUENCE	CONDITIONS
Microsatellites		
D1S439 F	CACAGACTTCATTAGAGGGG	Sigma, TD55
D1S439 R	GTTGAAATGGTGAATTTGGA	
D1S103 F	ACGAACATTCTACAAGTTAC	Sigma, TD55
D1S103 R	TTTCAGAGAAACTGACCTGT	
D1S225 F	GCCTGGGTGACAAAGCA	Sigma,15uM (normal is 20μM) TD55
D1S225 R	TGGCCTGAATAGACCATAAAAA	
D1S229 F	GCTTGTTTCCATTTATTGTG	Sigma, TD55
D1S229 R	ACTCTAGTTGTGTGTGAATGTATG	
D1S1621 F	TATAACCACTGCATCCTGACC	Sigma, TD55
D1S1621 R	TTTCTCACCTTTAAATGTCATCA	
D4S1599 F	CCTTAAAAGTATCCAGTAAAGCACA	Sigma, TD55
D4S1599 R	CAAGGTTGTCCTGTGTCTGCST	
D4S1533 F	TCTCTCTCTTGCTCTTCCC	Sigma, TD55ext x32
D4S1533 R	TCAGGCTTGTATGTGTGTTG	
STB131K9 F	GCTGAGCTTGCTCACTCTGTTA	Sigma, TD 55ext x32
STB131K9 R	GTGGCTTAAGGGAATTGTGTTC	
D4S1609 F	TCTGAAAATGCCCTTGACC	Sigma, TD55ext x32
D4S1609 R	CATCATTACTGCTGGGATGC	
D4S2408 F	AATAAACTTCAACTTCAATTCATCC	Sigma, TD55ext x30
D4S2408 R	AGGTAAAGGCTCTTCTTGGC	
D4S1546 F	CTACCGACTAGAATCACATGGG	Sigma, TD55ext x30
D4S1546 R	GACCTGTCCAAACGCCT	
3′ RACE		
3' RACE Outer Primer	GCGAGCACAGAATTAATACGACT	As 3′ RACE Protocol
3' RACE Inner Primer	CGCGGATCCGAATTAATACGACTCACTATAGG	As 3′ RACE Protocol
3' RACE Control Primer	CAGGGACATTTTCCAGCAAATTC	As 3′ RACE Protocol
5' PCR Control Primer	CAAGTCTGGTTCTTCTCCTCTT	As 3′ RACE Protocol
PI4K2B-specific for detection of alternative splicing		
stcPI4K2B.ex1aF	GAGGTGGCGTCCGTTCTAC	Invitrogen/AB/Sigma, TD55
stcPI4K2B.ex1bF	GGTCTCCCGGAGTTCGTC	Invitrogen/AB/Sigma, TD55
stcPI4K2B.ex4R	CTCACTGACAAGCCAAACCA	Invitrogen/AB/Sigma, TD55
stcPI4K2B.exon10_outer (aka 3start_primer)	AACGCAGTCAAGGTGGAAGT	As 3′ RACE Protocol
stcPI4K2B.ex10_inner (aka exon10_inner)	GTCAATTGCAGGAAGCCATT	As 3′ RACE Protocol
End of 3′ 1kb sequence (aka 3′UTRend)	AGCAATGTTGCCTTAATTTCAAAAGCT	As 3′ RACE Protocol

Table 2.4 Primer list. Continued on the next page.

PI4K2B cDNA SPECIFIC		
PI4K2B F	GGGGGTCTCCCGGAGTTCG	Invitrogen/Sigma, TD55
PI4K2B R	CCATAAGGCTCTTCTGATTTGG	
Intron spanning primers		
WDR1 F	AAGGCCCACGACGGTGGGATTTAC	Sigma, TD65
WDR1 R	GGAGTTCACGCTGACGTCCCAAAT	
DDX15 F	AACACCTATCCTGAGATTTTGCGTTCT	AB, TD65
DDX15 R	TTCCAGGGCTCTCATCAGAGTTTC	
ANAPC4 exon 27_30 F	GCTAGATGAACAGTGTAGTGCTATTCC	AB, TD55
ANAPC4 exon 27_30 R	ACACACACAATAATGGCAAGC	
PeakPicker primers		
stePI4K2B.ex10eF	CCTTTGAGGATGCCTACGTC	Sigma, TD55
stePI4K2B.ex10eR	TGAACAAGGCCAGTACTCTGAA	
stPI4K_ex2 F	AGAAGTGGGGGTTGTTAGGC	Sigma, TD55
stPI4K_ex2 R	CCTTTGTTTTGGAAAAGAGGTT	
stePI4K2B.int1 F	TGATGGATGCATTGACGTTT	Invitrogen, TD55
stePI4K2B.int1 R	ATGGGGATCTCGCTATGTTG	

Table 2.4 Primer list Abbreviations; AB is Applied Biosystems, F is forward primer, R is reverse primer, TD is touchdown PCR amplification conditions.

2.2.3.2. Non-quantitative PCR

PCRs were carried out using Sigma, Invitrogen or Applied Biosystems methods, in a total volume of 20µl as follows: Sigma; 2µl 10xSigma Buffer, 0.2µl Sigma Taq Polymerase, 0.6µl 10mM dNTPs (Sigma), 0.5µl 20mM Primers, 1µl DNA, 15.7µl dH_2O as per Sigma protocol: Invitrogen; 2µl 10x Invitrogen Buffer, 0.08µl Invitrogen Taq Polymerase, 0.6µl 10mM dNTPs, 0.5µl 20mM Primers, 0.52µl $MgSO_4/MgCl_2$, 0.6µl 10%BSA (New England Biolabs), 0.6µl DMSO, 2µl Invitrogen PCR Enhancer, 1µl DNA, 14.7µl dH_2O as per Invitrogen protocol: and Applied Biosystems; 2µl 10x ABI Buffer, 0.2µl In-house Taq Polymerase, 0.6µl 10mM dNTPs, 0.5µl 20mM Primers, 1.2µl MgCl2, 1-2µl DNA (20ng), 13µl dH_2O as per Applied Biosystems protocol. For whole genome amplified products that failed initial genotyping, a HotStar Taq PCR method was attempted (+/- 2µl $MgCl_2$ and 4µl 5X Q Solution) as recommended by Qiagen for higher PCR specificity. HotStarTaq Qiagen protocol was 2µl 10x Buffer, 0.1µl Taq Polymerase, 0.6µl 10mM dNTPs, 0.5µl 20mM Primers, 2µl MgCl2, 1-2µl DNA (20ng), 4µl Q-solution, 8.8µl dH_2O. The amplification

occurred in a Peltier Thermal Cycler, PTC-225 (MJ Research) using TD55 & TD58 programmes at an initial denaturation of 93°C for 1 min, followed by 10 cycles of 93°C for 20s, touch down annealing from 65°C to 55°C (or 65°C to 58°C) for 30s over 10 cycles (-1°C/cycle) and 72°C for 1min followed by 30 cycles of 93°C for 20s, 55°C for 30s, 72°C for 1min and a final extension of 72°C for 10mins. 3µl PCR products were resolved on 2% agarose gel stained with SYBR Stain (Invitrogen). The primers for PCR amplification are shown in Table 2.4.

2.2.3.3. Rapid amplification of cDNA Ends (RACE)

The 3′ RACE FirstChoice® RNA-Ligase-Mediated (RLM) RACE-Ready cDNA method was used according to manufacturers' instructions (Ambion 3200). RACE ready human heart cDNA was used (Ambion). Both inner and outer PCR was performed. SUPER TAQ plus (HT Biotechnology) was used with this method. Direct sequencing of the PCR products was chosen as the optimum method for sequencing products. To perform this, RACE products were excised from a low melting point gel, purified and sequenced as previously described in section 2.2.2.4.

2.2.4. Sequencing

All DNA intended for sequencing was subjected to agarose gel electrophoresis to check for integrity of DNA or PCR products. Occasionally, the DNA was gel extracted before use in sequencing procedures.

2.2.4.1. ExoSAP-IT treatment of DNA

Exo-SAP-IT™ (GE Healthcare) contains exonuclease I and alkaline shrimp phosphatase which removes all excess primers and dNTPs respectively before sequencing. 1-2µl PCR product, 2µl dH2O and 1µl ExoSAP-IT™ were incubated on the thermal cycler for 60 minutes at 37°C followed by 20 minutes at 80°C.

2.2.4.2. PCR direct sequencing

The PCR Sequencing reactions were performed in a thermo-fast non-skirted plate, using BigDye® Terminator Ready Reaction Mix v3.1 with 1µl BDv3.1, 1µl sequencing buffer, 1µl primer 3.2pmol, 3µl dH_20, added to the ExoSAP-IT treated PCR. The primers used for sequencing are listed in Table 2.4. Sequencing reactions were performed on a Peltier PTC-225 thermal cycler. The thermal profile of the sequencing reaction is detailed in Table 2.5.

TEMPERATURE	TIME	CYCLE
96°C	1 minute	1
96°C	10s	25
50°C	5s	1
60°C	4 minutes	1
4°C	∞	1

Table 2.5 Thermal profile of BDv3.1 sequencing reaction

2.2.4.3. Ethanol/EDTA precipitation of sequencing reactions

EDTA stabilised the extension products during precipitation and washed out unincorporated dyes from the completed reaction. After sequencing, the DNA was briefly spun at 3,000rpm for 3 minutes. 2.5µl of 125mM EDTA and 30µl 100% BDH absolute ethanol were added to each sample and mixed four times. This was incubated at RT for 15 minutes followed by centrifugation at 3,000rpm for 30 minutes. All ethanol was removed and a further 30µl 70% ethanol was added. This was centrifuged at 3,000rpm for 15 minutes and the ethanol was removed. The samples were stored at -20°C. Samples were analysed using Applied Biosystems 3730 DNA Analyser by Alison Condie at the WT-CRF and sequences were examined using CONSED (Gordon, Abajian et al. 1998), Chromas (McCarthy 1998) or Geneious (Drummond, Ashton et al. 2007).

2.2.5. Genotyping

2.2.5.1. Microsatellite

Genotyping of microsatellites was performed with 1μl of 1:10 dilution (PCR product with dH_2O) added to 10μl of 1:200 dilution of Tamra-350 (Applied Biosystems) with HiDi (Applied Biosystems) and stored at -20°C. Genotyping was performed on Applied Biosystems 3730 DNA Analyser at the WT-CRF and analysed with Genemapper v3 (Applied Biosystems). The results were scored by two different individuals.

2.2.5.2. Linkage study

96 DNA samples from the family were genotyped in the linkage study. The DNA samples were stored at the Medical Genetics Section, Molecular Medicine Centre and the WT-CRF. I prepared and quality control tested the 33 samples from the Medical Genetics Section. Two of these DNA samples were whole genome amplified. Replicates of whole genome amplified sample from each original aliquot were pooled for genotyping to avoid allelic-amplification bias. The concentration and quality of each DNA sample was determined by DNA gel electrophoresis. The preparation of DNA aliquots and dilutions were performed under supervision to minimise sample mix-up. The samples from the WT-CRF were originally extracted from blood sample by GenVar (56 samples) or in-house at the WT-CRF (7 samples). The final DNA concentration was determined by the WT-CRF using picogreen analysis. The concentration of DNA used for genotyping was 50ng/μl, except for sample 37 @ 10ng/μl, sample 44 @ 20ng/μl and sample 160 @ 30ng/μl. As these samples were important for linkage analysis, they were included in the study despite the reduced concentration of DNA.

The DNA samples were genotyped by the Illumina Linkage IVb panel using Illumina BeadArray technology on an Illumina BeadStation in the WT-CRF. The GoldenGate assay protocol was used incorporating Linkage IVb panel. This panel is

a set of 6,008 SNPs, split into four OPAs (Oligo Pool All) containing all of the SNP-specific primers.

2.2.6. Whole genome amplification

2.2.6.1. Template preparation for amplification

Genomic DNA was available as previously published (Blackwood, He et al. 1996). In tubes with no visible DNA solution, 7.5µl RNAase free water was added to the eppendorf, incubated at 37°C for 15 min and 4°C for 48 hours. Genomic DNA from lymphoblastoid cell lines sample 34 or sample 29 were prepared as described in section 2.2.2.1 and used as positive controls.

2.2.6.2. Whole genome amplification

Whole genome amplification was carried out as per manufacturer's instructions for "Amplification of Purified Genomic DNA using the REPLI-g Midi Kit" (REPLI-g Midi Kit, Qiagen, UK). In brief, Buffer D1 (denaturation buffer) and Buffer N1 (neutralization buffer) were prepared by adding sufficient nuclease-free water for the required number of whole genome amplification reactions. 2.5µl (10ng) of template DNA was placed into a microcentrifuge tube. 2.5µl Buffer D1 was added to the DNA, prior to vortexing and brief centrifugation. The samples were incubated at RT for 3 min. 5µl of Buffer N1 was added to the samples to stop the denaturing process, mixed by vortexing and centrifuged briefly. The REPLI-g Midi DNA Polymerase was thawed on ice and the other components were thawed at RT, mixed by vortexing and centrifuged briefly. A master mix was prepared, mixed by vortexing and centrifuged briefly. The master mix comprised of 10µl nuclease-free water, 29µl REPLI-g Midi reaction buffer and 1µl REPLI-g Midi DNA Polymerase per reaction. 40µl of the master mix was then added to 10µl of denatured DNA and incubated at 30°C for 16 hours for the isothermal amplification reaction. The REPLI-g Midi DNA polymerase was inactivated by heating the sample for 3 minutes at 65°C. The amplified products were aliquoted and stored at -20°C. For samples that

failed, the amplification procedure was repeated with more DNA (20ng) or with DNA dissolved in H_2O at 65°C for 1 hour prior to amplification.

2.2.7. Analysis of amplification products

2.2.7.1. Preparation of DNA for quantification

Before quantifying DNA, it must be fully in solution. A 1 in 10 dilution of DNA with TE in a total volume 100µl was dissolved at RT for ~1 hour, and incubated for 48 hours at 4°C to dissolve DNA. Cut-off pipette tips were used to avoid shearing the DNA.

2.2.7.2. DNA quantification by picogreen

"Quant-iT PicoGreen dsDNA Reagent and Kit" (Molecular Probes) was used to quantify DNA. A high-range standard curve was prepared with λ DNA standard to detect a final DNA concentration from 1µg/ml to 1ng/ml. This was diluted 50 fold in TE to give 2µg/ml DNA stock (18µl of 100µg/ml λ DNA stock and 882µl TE). The standards were prepared in duplicate, as described in Table 2.6.

VOLUME OF 2μG/ML DNA STOCK (μL)	VOLUME OF TE (μL)	FINAL DNA CONCENTRATION WHEN MIXED 1:1 WITH PICOGREEN REAGENT (NG/ML)
0	500	Blank
50 (of 10ng/ml stock)	450	1
5	495	10
50	450	100
250	250	500
500	0	1000

Table 2.6 Preparation of λ DNA standards for DNA quantification by picogreen reagent.

For each sample, a 200 fold dilution of 100μl PicoGreen Reagent was prepared using TE and protected from light by covering with foil. A final DNA dilution (10μl diluted DNA and 90μl TE) was prepared in duplicate for each sample in a Microfluor plate (Dynex Microfluor 1, 96 well black plate, flat-bottom, Thermo Life Sciences). 100μl PicoGreen reagent was added to the diluted DNA and mixed by inverting. The fluorescence was measured on a Cytofluor fluorimeter (Cytofluor Series 4000 fluorescence Multi-Well Reader (Applied Biosystems)) with the following settings; Excitation = 485/20nm, emission = 530/25nm, centre only, 4 readings/well, mix time = 10s, plate type = DynaTech 96 well, gain = 45. These settings allowed accurate reading by ensuring the sample with the highest DNA concentration gave a reading close to fluorimeters maximum. The data were analysed by firstly subtracting the fluorescence value of the reagent blank from each of the standards and samples. A standard curve was generated from the corrected values, and the equation and correlation (r^2) of the line were calculated. This enabled the concentration of the samples and the validity of the line to be determined respectively.

2.2.7.3. DNA quantification by agarose gel electrophoresis

DNA was quantified by agarose gel electrophoresis with 1µl DNA (1µl DNA, 8µl H2O, 1µl 10X Orange G loading buffer) compared to various concentrations of λ*Hin*dIII on the same gel, 500ng was 10µl of 50ng/µl (Invitrogen). The gel was run for five minutes. The concentration of the DNA sample was determined by comparing the band of the DNA sample with a band of most similar intensity from the various concentrations of λ*Hin*dIII.

2.2.7.4. Agarose gel electrophoresis

Standard agarose gels of 1-1.5% were used to separate fragments of size 0.2kb-0.5kb and 0.8-1% gels were used to resolve fragments of size 0.5kb or greater. Gels were prepared and run using 1XTris buffer. Samples were loaded using 10X Orange buffer. Electrophoresis was carried out using Hybaid Electro-4 gel tanks. Voltage applied (usually 80-100V) creates an electric current causing the DNA to migrate through the agarose from the cathode (-ve) to the anode (+ve) due to the negatively charged phosphate backbone of the DNA. Amplified DNA products were diluted to 100ng/µl as calculated from expected yields and 1µl was visualised on a 0.8% agarose gel, stained with SyberStain in 0.5X TE buffer. This allowed a rough estimation of DNA quantity.

In general, DNA was stained with SYBR Safe™ (1µl/100ml molten agarose, Invitrogen) or occasionally with Ethidium Bromide (0.5µg/ml molten agarose, SIGMA). Fragments were visualised using a UV transilluminator (Thistle Scientific) with an uvidoc (Uvitec, DOC-CF08.XD). The molecular weight of each fragment was determined by comparing its mobility to a 1Kb DNA ladder (Invitrogen) or a λ*Hin*dIII (Invitrogen) marker for larger fragments of up to 23Kb in size.

DNA fragments were also run on low melting point (LMP) agarose gels to allow for recovery of the DNA after electrophoresis. LMP agarose gels were prepared using 1XTAE and usually run at 80V.

2.2.8. Taqman Assays

2.2.8.1. Taqman assay design

Two SNPs, rs313548 and rs313567 were available by pre-designed Taqman SNP genotyping assays (c_764549, c_764538 respectively) while rs6834255 was a custom designed Taqman SNP genotyping Assay (Applied Biosystems). The Taqman SNP genotyping assay uses a 5' nuclease assay to discriminate between two alleles of a specific SNP. Assays for rs313548 and rs313567 were 20x mix and a 40x mix for rs6834255, of forward primer, reverse primer, FAM™ dye - MGB labelled probe, and VIC® dye - MGB labelled probe. Each probe binds preferentially to one of the two alleles.

2.2.8.2. gDNA standard curve preparation

A standard curve of dilutions from homozygous lymphoblastoid cell line samples was prepared to evaluate the contribution of each allele in the heterozygote samples. Genomic DNA prepared from homozygote samples was used at a concentration of 5ng/µl. gDNA was mixed from two homozygotes of the SNP, e.g. genotype AA was mixed in dilutions 90:10 with genotype GG, and so on to make a standard curve 90:10, 80:20, 70:30, 60:40, 50:50, 40:60, 30:70, 20:80, 10:90. The optimal reaction was 10µg of mixed gDNA in a total volume of 5µl, 2ul gDNA, 0.25µl Taqman SNP Genotyping Assay, 2.5µl Taqman® Universal PCR Master Mix, 0.25µl dH_2O. Three standard curves with six different homozygotes were prepared to test the precision of each assay. Two standard curves with four different homozygotes were prepared to test the precision of rs313548 and rs6834255 assays. Three standard curves with four different homozygotes in different combinations were prepared to test rs313567 assay; 132 (CC) & 86 (TT), 132 (CC) & 35 (TT), T3445 (CC) & 86 (TT). For assays rs313548 and rs6834255, there were an insufficient number of lymphoblastoid cell line sample homozygote individuals so it was necessary to use homozygote individuals typed from the Trio Bipolar Panel. Each homozygote mix was assayed in quadruplicate.

2.2.8.3. Taqman SNP genotyping assay

For the heterozygote DNA, 4ng cDNA or 10ng gDNA was amplified; 2µl cDNA (2ng/µl) or 2µl gDNA (5ng/µl), 0.25µl Taqman SNP Genotyping Assay, 2.5µl Taqman® Universal PCR Master Mix, 0.25µl dH_2O. Each sample was prepared in quadruplicate and each assay was performed in triplicate from the same RNA preparation. The assays were performed in ABgene Thermo Fast 384 well PCR plates (Applied Biosystems) and covered with a Clear Seal Diamond heat seal (Applied Biosystems). The Allelic Discrimination PCR reaction (initial step is hold for 10 minutes 95°C, denaturing step is 15s 92°C X 40 cycles, and annealing step is 1 minute 60°C X 40 cycles) was performed in the ABI PRISM 7900HT sequence detection system (Applied Biosystems) available at the WT-CRF.

2.2.9. Protein preparation

2.2.9.1. Protein lysate preparation

All lymphoblastoid cell line protein extractions were performed at the same time to optimise protein concentration estimation and minimise varying degradation effects. The lymphoblastoid cell line pellets were centrifuged 2,000rpm for 5 minutes, the supernatant removed and washed in PBS. The cell pellet was resuspended in ice-cold lysis buffer (50mM Tris, 150mM NaCl, 1% Trition, 0.25% Sodium deoxycholate, 1mM EDTA, 2% Complete Protease Inhibitors) for 30 minutes on a rotor. Lysates were sonicated three times for three seconds at 20 revs (MSC Sonicator), centrifuged and supernatant saved. The protein concentration was determined immediately, the samples were aliquoted and/or prepared to 10µg aliquots in SDS buffer and stored at -80°C until use.

2.2.9.2. Protein concentration estimation

The protein concentrations of lymphoblastoid cell line lysates were estimated using the BCA Analysis (Pierce) as per manufacturers' instructions. In brief, protein samples of unknown concentration were prepared in duplicate at 1:20 dilution

(100µl Total Volume = 5µl protein + 95µl dH_2O) and a 1:10 dilution (100µl Total Volume = 10µl protein + 90µl dH_2O). The diluted albumin (BSA) standards were prepared as Table 2.7.

BCA PROTEIN ASSAY			
Vial	*Volume of Diluent (H_2O)*	*Volume & Source of BSA*	*Final BSA Concentration*
A	0	250µl Stock	2,000µg/ml
B	125µl	375µl Stock	1,500µg/ml
C	325µl	325µl Stock	1, 000µg/ml
D	175µl	175µl of vial B dilution	750µg/ml
E	325µl	325µl of vial C dilution	500µg/ml
F	325µl	325µl of vial E dilution	250µg/ml
G	325µl	325µl of vial F dilution	125µg/ml
H	400µl	100µl of vial G dilution	25µg/ml
I	400µl	0	0µg/ml (Blank)

Table 2.7 Standard curve for protein concentration estimation

The BCA Working Reagent (WR) was prepared following this equation;

- (# standards +# unknowns) x (# replicates) x (volume of WR per sample) = total volume WR required, adding 50 parts Reagent A to 1 part Reagent B.

2ml working reagent was added to each sample tube and standard tube and mixed well. The samples were incubated for 30 minutes in a 37°C water bath and then cooled to RT. The absorbance of the samples was read at OD 562nm from a spectrophotometer (GeneQuant) in plastic cuvettes (1cm path length, Fischerbrand). A standard curve was prepared using a regression line from the duplicate standard samples. The concentrations of the samples were estimated from the standard curve.

2.2.10. Quantitative protein assays

2.2.10.1. Protein gel electrophoresis

The Western blotting procedure was performed on 5μg protein lysates using NuPAGE® Electrophoresis System, as per manufacturer's instructions with the 17-well 4-12% Bis-Tris Gels. Western blotting was also performed using Mini-PROTEAN® 3 Cell (Bio-Rad) according to manufacturers instructions, with the following specifications: separating protein samples with gel electrophoresis on a 10% SDS-PAGE gel at 100V for 10 minutes, 150V for approximately 90 minutes.

2.2.10.2. Staining & drying protein gels

Proteins were stained after electrophoresis with SimplyBlue™ SafeStain as per manufacturers' instructions (Invitrogen) or Coomassie Brilliant Blue stain (as Section 2.1.2). This method checked the equal protein loading in each well and also determined the successful transfer of proteins after western blotting. The stained gels were dried using DryEase® Mini-Gel Drying System as per manufacturer's instructions.

2.2.10.3. Immobilising proteins on membrane

The protein gels were transferred with X-Cell SureLock™ MiniCell (Invitrogen) as manufacturer's instructions to PVDF membrane (Hybond-P PVDF Membrane, Amersham Biosciences). Semi-dry transfer was also performed at 20V, 250A for 60 minutes with Trans-blot SD Semi Dry Transfer Cell (Bio-Rad) according to manufacturer's instructions. This method was used in conjunction with 10% acrylamide gels, transferred to PVDF membrane with blotting paper either side of the PVDF membrane (Whatman® paper).

2.2.10.4. Ponceau S staining membranes

Ponceau's stain was used to determine successful protein transfer onto PVDF membranes. After transfer, membranes were rinsed in dH_20 followed by five minutes incubation in Ponceau S. The membranes were destained in dH_20.

2.2.10.5. Immunoblotting

All membranes were incubated in 10% Marvel PBS overnight at 4°C to block non-specific sites on the protein-containing membrane. For fluorescent antibody detection, the membranes were blocked in 5% Gelatin from cold water fish skin (Sigma-Aldrich) overnight at 4°C. The antibodies were used at the following conditions; PI4K2B (all three antibodies are discussed in section 3.7.1) at 1:1000 for two hours, α-tubulin (Sigma) at 1:3000 for one hour, GAPDH (Abcam) at 1:3000 for one hour, β-actin (Sigma) at 1:1000 for one hour, R47, Pig α-Rabbit (Dako) at 1:3000 for twenty minutes and Rabbit α-mouse (Sigma) at 1:1000 for 30 minutes. The membranes were detected as per manufacturers' instructions (Amersham Biosciences ECL Plus Western Blotting Detection System) and visualised on X-ray film.

2.2.10.6. Stripping immunoblots

Immunoblots were stripped with Restore Western Blotting Stripping Buffer (Pierce), according to manufacturer's instructions.

2.3. Analysis of Molecular Biology Data

2.3.1. Sequence trace analysis

Sequence trace was visualised using Consed (Gordon, Abajian et al. 1998), Chromas or Geneious (Drummond, Ashton et al. 2007). The peak heights of the sequence trace were measured using PeakPicker software as per authors' instructions (Ge, Gurd et al. 2005).

2.3.2. Analysis of allele-specific assays

After the PCR amplification and end point read performed by the ABI 7900HT, it was necessary to assess three parameters for each assay; baseline, threshold, Ct value, in order for accuracy and precision to be optimal. The baseline read may fluctuate from assay to assay, due to changes in the reaction mix so it was important to ensure the amplification was sufficiently higher than the background signal. The threshold is a numerical value assigned for each run and shows a statistically significant point above the calculated baseline. The threshold cycle (Ct) reflects the cycle number at which the fluorescence generated within a reaction crosses the threshold. The Ct value assigned to each well is the point during the reaction where a sufficient number of amplifications have accumulated to be at a statistically significant point above the baseline. The same settings were used in repeats of the same assay, to improve precision and to make the data comparable. Outliers of the four replicates in each experiment were determined by displaying the data as CT vs. Well position and they were removed from further analysis. A reading was constituted as an outlier if one of the four replicates did not cluster with the other three replicates. After the appropriate settings were determined, the allele types were scored by ABI PRISM® SDS software v. 1.7 using the fluorescence readings (Rn) from the endpoint read. Rn is the reporter signal normalized to the Passive Reference for a given reaction. The delta Rn value is the Rn value minus the Rn value for the No Template Control. The Rn values from the FAM™ reporter dye and VIC® reporter dye tagged to each allele were used for statistical analysis.

After the PCR amplification, an end point read was analysed and the allele types called using ABI PRISM® SDS software v. 1.7. From the endpoint read, the ratio of the fluorescence reading from the FAM™ reporter dye and VIC® reporter dye was calculated. The mean of all replicates was calculated for each DNA sample and the range of values is shown by errors bars of 95% Confidence Intervals (CI). A standard curve (linear regression line) was established for the LOG of the fluorescent intensity ratio (FAM™/VIC®) versus the LOG of the allele ratio i.e. the

LOG_2 ratio of the dilution of each homozygote, from which the ratio of cDNA samples from heterozygote individuals were deduced. The transformation of the data to LOG_2 was performed to normalise the data, enabling further analysis. This allowed the deduction of the ratios of gene expression between the two heterozygote alleles from the standard curve, by measuring the fluorescent intensity (LOG_2 FAM™/VIC® reporter dye intensity) of the two alleles in cDNA samples.

The sensitivity of the assays and variation between results was checked with basic statistical analysis using the mean of the replicates, measuring the variability by standard deviation, measuring the amount of variability in the sample mean by standard error and measuring the range of plausible values for population mean with a 95% and 99% confidence interval. The confidence interval was added or subtracted to each mean to have a low and high confidence interval range and was plotted with results from standard curve replicates (please refer to Appendix A (Kirkwood and Sterne 2003)).

Experimental consistency was tested within each assay looking at the coefficient of variation between replicates and reproducibility between assays was tested with an average coefficient of variation.

As it was assumed that genomic DNA samples have an allele ratio of 1:1, but because of differences in fluorescent emissions from the two dyes tags (FAM and VIC) the measured ratios for gDNA samples deviated from 1:1 sample. Thus, cDNA values were calculated by the cDNA/gDNA ratio from the mean value of DNA from the same individual. Variation within and between experiments was measured by the coefficient of variation (standard deviation/mean). To test a difference between groups, the Students T-test was used.

2.3.3. Analysis of sequence data with PeakPicker

The amplification and sequencing of duplicate cDNA and gDNA samples was attempted for the PI4K2B genomic region surrounding the three SNPs. For each sample heterozygous at the particular SNP, two cDNA and two gDNA sequences surrounding the SNP were imported into the PeakPicker programme and analysis was performed according to the author's instructions (Ge, Gurd et al. 2005). In brief, each heterozygote was analysed separately. Firstly a gDNA sequence was set as the reference sequence for alignment and peak selection. PeakPicker performed a multiple pairwise alignment of the imported sequences starting from the first base of each. A sequence length of 250bp-300bp and a 70% identity cut-off was used. Only sequence traces of good quality were analysed.

The peaks comprising the SNP were selected in the reference sequence automatically. Reference peaks of the same genotype of the SNP, were chosen from the same region at positions in the reference sequence where there was 100% consensus amongst all imported sequences.

Once the appropriate set of reference peaks had been chosen, PeakPicker calculated the trace heights at these positions in all sequences. It then normalised the heights of the peaks relative to the mean height of their corresponding reference peaks within each sequence. The output was a normalized peak ratio which could be compared across sequences.

2.3.4. Analysis of protein quantity

Densitometry analysis was performed using Scion Image or ImageJ v 1.37 (Abramoff 2004) to quantify the immunoreactive protein in each lane. The pixel value of the peak corresponded to the intensity of the band from the image in greyscale. PI4K2B expression was analysed as PI4K2B intensity:house-keeping protein intensity ratio. If more than one housekeeping gene or Ponceau S stained membrane was quantified, an average of the control intensities was used.

2.3.5. Investigation of SNP effects

The SIFT programme, as detailed in Table 2.8, was used to obtain positional and coding information about particular SNPs. This incorporates a BLAST search with the protein sequence to align with similar proteins, and utilising the Sorting Intolerant from Tolerant (SIFT) (Ng and Henikoff 2003) algorithm to find functional domains of the protein, and if the amino acid change falls in an important regulatory region.

2.4. Bioinformatics

The online resources used for bioinformatic analysis are listed in Table 2.8.

RESOURCE	WEBSITE
Allen Brain Atlas	http://www.brainatlas.org/aba/
Chromas – Chromatogram Viewer	http://www.technelysium.com.au/chromas_lite.html
Database of Chromosomal Imbalance & Phenotype in Humans using Ensembl Resources (DECIPHER)	http://decipher.sanger.ac.uk
DNA & Protein Sequence Analysis	http://www.geneious.com/
Haploview	http://www.broad.mit.edu/mpg/haploview/
HapMap	http://www.hapmap.org/
Multiple Sequence Alignment	http://www.ebi.ac.uk/Tools/clustalw2/index.html
NEBcutter v2	http://tools.neb.com/NEBcutter2/index.php
Primer Design	http://wwwgenome.wi.mit.edu/genome_software/other/primer3.html
Repeat Masker	http://repeatmasker.genome.washington.edu/cgi-bin/RM2_req.pl
Sequence Alignment	http://bioweb.pasteur.fr/docs/EMBOSS/prettyplot.html
SIFT (Sorting intolerant from tolerant)	http://blocks.fhcrc.org/sift/SIFT.html
UCSC Human Genome Browser	http://genome.ucsc.edu/index.html?org=Human
UCSC In silico primer	http://genome.ucsc.edu/cgi-bin/hgPcr

Table 2.8 Bioinformatic resources

2.5. Statistical Methods

Standard statistics were calculated using MS Excel 2003, SPSS 14.0 (SPSS_Inc. 2005), R version 2.4.0 (R_Development_Core_Team 2006) or GraphPad Prism (GraphPad_Software version 4.03). The statistical analysis programmes used in this thesis are detailed in Table 2.9. Please also refer to appendix A for statistical equations.

LINKAGE PROGRAMME	FULL NAME	AIM	SOURCE	VERSION	REFERENCE
CARTOGR	**Cartogr**apher	Estimates genetic position of markers by interpolation from neighbouring markers based on your map of choice (deCODE)	https://apps.bioinfo.helsinki.fi/software/cartographer.aspx	1.1.0	(Knuuttila 2007)
GREFFA	**G**enetic **R**elationship **E**xplorer for **F**amiliality **A**ggregation	Pedigree splitting	Provided by M. Falchi, KCL, UK.	0.2	(Falchi, Forabosco et al. 2004)
Haploview	**Haplo**type **vi**sualisation	LD and haplotype block analysis single SNP association tests.	http://www.broad.mit.edu/mpg/haploview/	3.2	(Barrett, Fry et al. 2005)
MEGA2	**M**anipulation **E**nvironment for **G**enetic **A**nalyses	Data manipulation and file preparation	http://watson.hgen.pitt.edu/register/	3, 4	(Mukhopadhyay, Almasy et al. 2005)
MERLIN	**M**ultipoint **E**ngine for **R**apid **L**ikelihood **I**nference	Linkage analysis, error detection, haplotype estimation & simulation	http://www.sph.umich.edu/csg/abecasis/merlin/	1.0.1, 1.1.alpha	(Abecasis, Cherny et al. 2002)
PedCheck	**Ped**igree **Check**	Detecting marker typing incompatibilities in pedigree data	http://watson.hgen.pitt.edu/register/	1.1	(O'Connell and Weeks 1998)

Table 2.9 Continued on the next page.

Pedstats	**Ped**igree **Stat**istic**s**	Summary statistics and basic quality assessments for gene-mapping	http://www.sph.umich.edu/csg/abecasis/Pedstats/download/	0.6.8.	(Wigginton and Abecasis 2005)
PREST	**P**edigree **RE**lationship **S**tatistical **T**est	Detection of pedigree errors by use of genome-screen data	http://www.utstat.utoronto.ca/sun/Software/PREST/	3.0	(McPeek and Sun 2000)
Relpair	**Rel**ationship **pair**s	Infers the relationships of pairs of individuals based on genetic marker data, either within families or across an entire sample	http://csg.sph.umich.edu/boehnke/relpair.php	2.0.1	(Boehnke and Cox 1997; Epstein, Duren et al. 2000; Duren W.L. June 2004)
Unphased (CoCaPhase)	Association analysis	Association analysis of multilocus haplotypes from unphased genotype data, including case-control data.	http://portal.litbio.org/Registered/Option/unphased.html	2.43	(Dudbridge 2003)

Table 2.9 Software programmes used for statistical genetic data preparation & analysis.

2.5.1. Case-control association study

2.5.1.1. Selection of markers

SNP genotype data was downloaded for the 30 CEPH trios (Utah residents with ancestry from northern and western Europe) from 100kb upstream and downstream of the *PI4K2B* genomic region (24,745,440bp-24,986,687bp, NCBI build 35, May 2004 UCSC) from HapMap Phase II build 125 (http://www.hapmap.org) on 7th February 2006. This HapMap data was uploaded to the Haploview version 3.2 programme (http://www.broad.mit.edu/mpg/haploview/). The parameters chosen for construction of the LD map were as follows: pair-wise comparisons of markers more than 500kb apart were ignored, minor allele frequency ≥0.1, Hardy Weinberg Equilibrium >0.01, genotyping success rate of >85%, and a maximum number of Mendelian errors =1. The haplotype blocks were defined by the solid spine of LD (D'>0.8). Adjacent LD blocks were then merged if MAD (Hedrick's Multi Allelic D') ≥0.95 (Hedrick 1987). Tagging SNPs were then selected to tag the common haplotypes (>5%) within each haplotype block using Haploview's internal tagging programme. Tagging SNPs were replaced by other SNPs that equally tag the haplotypes if the SNPs had been genotyped previously, if pre-designed assays were available from the supplier or if the SNPs were included in the PI4K2B expression study.

2.5.1.2. Genotyping

Genotyping was performed using pre-designed Taqman assays on demand or assays by design from Applied Biosystems. The genotyping was performed on the ABI PRISM 7900HT sequence detection system at the WT-CRF.

2.5.1.3. Data quality control

The quality of the SNP assays was validated by checking Mendelian inheritance and genotype success rate on DNA from 30 trios (father, mother and affected offspring).

The results of the genotyping were processed by the Computing Support Department (Euan Adie and Stewart Morris). This involved formatting the results as per standard linkage format described in section 2.5.3, checking duplicates, confirming diagnoses and merging results from original study with the newly chosen SNPs.

2.5.1.4. Analysis

The standard X^2 test of independence was used to examine all markers for deviations from Hardy Weinberg Equilibrium (HWE) in the control samples. The Haploview default HW significance threshold of p=0.001 was selected as a cut-off to exclude potentially erroneous markers due to for example, genotyping error. Single-marker analysis was performed using in-house software developed by Naomi Wray that calculates basic association analysis "BasicAS", with the X^2 test of independence to look for differences in allele and genotype frequencies between cases and controls. Haplotype analysis was performed separately on bipolar disorder and schizophrenia cases using CoCaphase, with a sliding-windows approach to test haplotypes of 2 marker length. Rare haplotypes with a frequency ≤0.05 in both cases and controls were clumped together. The maximum-likelihood frequencies are re-computed after identification of rare haplotypes.

2.5.1.5. Permutation

To estimate the significance of the association study result, taking the number of markers and haplotypes tested into account, permutation analysis was performed using CoCaphase. CoCaphase reassigns the diagnosis labels (case versus control) of the individuals in 1,000 simulations of the data. All of the markers within the specified window size are tested and the most significant *P*-value is stored. Based on the distribution of the most significant *P*-values resulting from each test of the permuted data, a significance level is provided for the *P*-value of interest.

2.5.1.6. Bioinformatic analysis

Analysis of human genome sequence data was performed using the University College of Santa Cruz (UCSC) genome browser (Kent, Sugnet et al. 2002). For bioinformatic analysis, the most updated version March 2006 human reference sequence (NCBI Build 36.1) was used. Otherwise, the May 2004 human reference sequence (NCBI Build 35) was used, as these were the co-ordinates used by Illumina to compute the genetic position for the markers tested in the linkage study.

2.5.2. Pedigree drawing

Pedigrees were drawn using Progeny Desktop Version 6 (Progeny Software LLC, South Bend, IN, www.progenygenetics.com) or HaploPainter v.029.5 (Thiele and Nurnberg 2005) according to the authors' instructions. Standard pedigree symbols were used (Bennett and Steinhaus 1995). Each diamond symbol represents an individual from the family. Individuals with a diagnosis of bipolar disorder were designated a filled symbol, those with recurrent major depression by a half-filled symbol and those with other related psychiatric diagnosis including single episode depression, anxiety states or alcoholism are marked by a quarter-filled symbol. Individuals that were genotyped for the linkage study have a central blue circle. To preserve anonymity, individuals who were known to the research team and who had been interviewed but made no contribution to linkage analysis, were not included in the pedigree drawing.

2.5.3. Linkage data preparation & analysis

The software programmes used for linkage analysis are listed in Table 2.9.

2.5.3.1. Linkage file preparation

The raw genotyping data for each of the 96 samples genotyped with 6,008 SNPs from the Illumina Linkage IVb panel were processed into standard linkage format by Stewart Morris. Standard linkage format (Terwilliger and Ott 1994) is required

for most statistical genetics programmes and can easily be converted to other formats. The following files necessary for linkage analysis were prepared. The "pedigree" file contains the pedigree structure, trait phenotypes and genetic marker data in a numerical file, tab-delimited, with a separate line for each individual, with the information as follows: Pedigree number, Individual ID, Father ID, Mother ID, Gender, Affection status, Liability class, Genotype allele 1, Genotype allele 2, etc. The "names" file lists the names of all genetic loci in map order. The "map" file lists the marker names with their NCBI's Build 35 chromosomal location and the deCODE genetic distance in centiMorgans (cM). To adjust data in the "pedigree" file, an "omit" file was prepared. This is a list of all erroneous genotypes which are removed from the linkage analysis files. This bypasses the error-prone method of adjusting the large "pedigree" file.

2.5.3.2. Genetic map

The genetic map position for each SNP was provided by Illumina and was determined by linear interpolation using NCBI's build 35 physical map position and a high-resolution STR genetic map constructed by deCODE genetics (Kong, Gudbjartsson et al. 2002). This map was a sex-averaged map. The marker list and position are available from the Illumina website (http://www.illumina.com/pages.ilmn?ID=191). For supplementary chromosome 4 microsatellite markers, the genetic distance was estimated by deCODE genetics, using the Cartographer programme or by linear interpolation from flanking markers. Table 2.10 lists the name and genetic position of the microsatellite markers.

MICROSATELLITE NAME	PHYSICAL POSITION (JULY 2003, BP)	GENETIC POSITION (DECODE CM)
D4S431	6,480,116	12.46
D4S394	7,024,409	14.79
stCeGAx54.p1	8,925,699	20.16
st149b15_3p24	9,516,083	21.83
stb448G15.ca2	9,638,891	22.17
stb751L19.p1	9,635,534	22.17
D4S615	9,665,580	22.25
st39F5_87N1	9,811,952	22.36
D4S2928	10,363,119	22.78
D4S1582	10,452,339	23.06
st452b6_30m19	10,480,697	23.1
D4S1605	10,803,013	23.18
D4S2281	10,705,420	23.38
D4S3009	10,887,666	24.01
D4S2949	11,274,349	25.34
stb473m.ca5b	11,344,063	25.38
stbA473m13.2n	11,345,436	25.38
stb74m11.p1	11,526,286	25.49
stb122E22.ca	11,609,487	25.55
D4S3036	11,969,826	25.77
stb4e12.p1	12,003,222	25.79
D4S403	13,501,710	26.71
D4S1091	26,348,843	45.67
D4S2397	27,008,683	46.75

Table 2.10 Chromosome 4 microsatellite markers. The 24 markers were used in the linkage study. The physical position of each marker is according to July 2003 human reference sequence (NCBI Build 34), the genetic position was estimated by deCODE genetics (not shaded), using the Cartographer programme (grey diagonal lines) or by linear interpolation from flanking markers (dark-grey shaded).

2.5.3.3. Preparation of input files for statistical analysis programmes

MEGA2 was used to prepare input files in various formats for different statistical packages (Mukhopadhyay, Almasy et al. 2005; Mukhopadhyay N 2006). The pedigree, names, map and omit files described in section 2.5.3 were input into MEGA2 and transformed to the desired format. Allele frequencies for the locus data files were determined using the RECODE feature in MEGA2, which determined the allele frequencies from all genotyped members of the family. With the help of Stewart Morris, allele frequency files were also prepared from the Caucasian population that were generated from 60 unrelated samples from the HapMap CEPH population (Jim Acierno, Personal Communication, 21st August 2007, Illumina (http://www.hapmap.org)).

2.5.3.4. Genotyping data description

The quality of the genotyping data and Hardy Weinberg Equilibrium testing was checked and described using Pedstats according to the author's instructions (Wigginton and Abecasis 2005). Mendelian inconsistencies in the pedigree data were identified by PedCheck according to the authors instructions (O'Connell and Weeks 1998). Further genotypes that implied excessive and unlikely recombination events between tightly linked markers were detected as erroneous by MERLIN according to the authors' instructions (Abecasis, Cherny et al. 2002). These errors were collected in the "omit" file, which is an input file for MEGA2 to exclude the erroneous genotypes from further analysis, as described in section 2.2.3.1.

2.5.3.5. Detection of marker linkage disequilibrium

LD between markers from the Illumina Linkage IVb panel was detected in the family using Haploview (Barrett, Fry et al. 2005). Haploview chose a maximum unrelated subset of individuals to detect LD. These were 12 unrelated individuals (samples 11, 13, 20, 38, 46, 49, 98, 104, 146, 149, 160, 180) and one trio (samples 29, 33,

35). LD was detected using default parameters: solid spine of LD D'≥0.8 for markers ≤ 500kb apart. SNPs were said to be in LD, if D'≥0.8 and LOD ≥ 2.

2.5.4. Pedigree splitting

As the pedigree was too large to be analysed for whole genome linkage analysis, it was split into sub-pedigrees using GREFFA (Genetic Relationship Explorer for Familiality Aggregation) software, according to the authors' instructions (Falchi, Forabosco et al. 2004). The input files were in standard linkage format, and the genotyped individuals to be clustered were labelled "2" and those not genotyped were labelled "0". The pedigree was split before and after adjustment for relationship errors. Splitting was based on a kinship coefficient pairwise-relationship measure at the level of first-cousin relationships (1/16), $\psi \geq 0.0625$. The software automatically determined the minimum number of generations to reconstruct the pedigrees and extracted a pedigree for the largest set of individuals as the most informative pedigree. It also extracted more pedigrees for all sets of individuals. The number of permutations was 999, which was the number of maximum trials for the shuffling algorithm to search for the best clique, which is a cluster of individuals, according to the selected parameters.

2.5.5. Relatedness analysis

2.5.5.1. PREST

The programme PREST was used to detect pedigree errors (McPeek and Sun 2000). PREST estimates the probabilities κ_0, κ_1 and κ_2 of two individuals sharing 0, 1 and 2 alleles IBD respectively. The input files were converted from standard linkage format to PREST format using MEGA2, with setting of option 25 "PREST" and all defaults, except genotyping errors were not set to zero. This ensured that Mendelian inconsistencies were not removed before performing relationship testing. Analysis with PREST was performed as suggested, calculating IBD over eleven relationship pairs (parent-offspring, full-sibs, half-sib, avuncular, first-cousins, grandparent-

grandchild, half-avuncular, half-first cousin, half-sib plus first cousin, monozygotic twins and unrelated) within the family. Pedigree errors were detected by PREST's statistical tests: conditional estimated identity by descent (EIBD), adjusted identity by state (AIBS) and IBS, in that order and where applicable, at $\alpha = 0.0001$, to highlight more significant problems. The pedigree errors that were detected were examined again using PREST's accompanying programme ALTERTEST that tests two individuals for each of the eleven relationship classes, using a default 100,000 replicates to calculate the empirical *P*-values for the EIBD, AIBS, IBS and MLLR test. PREST output data was displayed using an "R" script written by Prof. Dan Weeks, which was provided with the programme package. This script was adapted with the help of Dr. Andy McLeod to create a scatter diagram of IBDs on a relationship triangle, similar to a diagram showing pedigree errors in the Framingham Heart Study (Brush and Almasy 2003). Nine of the eleven tested relationships were applicable to the family.

2.5.5.2. Relpair

Relpair checks for pedigree errors by inferring relationships between individuals in a pedigree (Epstein, Duren et al. 2000). The input files required for Relpair were a "control" file specifying output file names and parameter settings, a "locus" file specifying marker names, positions and allele frequencies and a "pedigree" file specifying pedigree structure and individual genotypes. The input files were converted from standard linkage format to Relpair format using MEGA2; with option 2 "Relpair" using all MEGA2 default options, except that genotyping errors were not set to zero. The default parameters for the Relpair analysis were used with the following exceptions: within the family, with a minimum of 100 shared genotypes, a genotyping error rate of 0.1% and a critical value of 1000 to be reported. The default settings were used for compiling the programme Relpair, with the following exceptions to the parameter file as described in Table 2.11. This changed the allocation of memory in the programme to accommodate running large number of SNPs.

PARAMETER	CHROMOSOME 1	CHROMOSOME 13 & 16	CHROMOSOME X
MAXALL	2	2	2
cMAXFAM	1	1	1
MAXLOC	500	200	300
MAXPEO	182	182	182
MAXREL	8	8	8
MXPTOT	182	182	182

Table 2.11 Adjustments to parameter files for Relpair programme. The following adjustments were made to the source code for Relpair to change the memory allocation such that it would analyse a large number of SNPs. The markers were biallelic SNPs thus the number of alleles was two. The analysis was performed on one pedigree with 182 individuals on four different chromosomes, chromosome 1, 13, 16 and X. Each chromosome has a different number of SNPs (chr1=489 SNPs, chr13=191 SNPs, chr16=197 SNPs, chrX=301 SNPs). Eight SNPs with zero allele frequency were removed from X chromosome analysis. The MAXLOC parameter was changed to accommodate the different number of SNPs per chromosome. The parameters are MAXALL=maximum number of alleles, MAXFAM=maximum number of families, MAXLOC=maximum number of markers (loci), MAXPEO=maximum number of people per family, MAXREL=maximum number of relationships tested, MXPTOT=maximum number of people in total in all families.

2.5.6. Simulation analysis

Simulation of the family genotype data was performed using the –simulate option in MERLIN. The input files were created using MEGA2 for each chromosome for two phenotype classes, broad and narrow. The simulation of repeated datasets was repeated 10,000 (broad model) or 1,000 (narrow model) times with a different seed using the –rerun option. Non-parametric linkage analysis was performed on the data. A histogram was prepared with counts of the highest LOD scores obtained on whole genome analysis per replicate. The 95 and 99 percentiles were calculated. Empirical *P*-values were computed by dividing the number of replicates that exceeded the observed Z-score or LOD score by the number of replicates.

2.5.7. Parametric linkage analysis

Parametric linkage analysis was performed on the sub-pedigrees using MERLIN. The input files were prepared using MEGA2, to create a "pedin", "map", "omit" and a "names" file with option "L" to take into account the binary trait and the liability classes. Each individual was assigned a trait class of unknown affection status, well or affected; equivalent to 0, 1 or 2 respectively. Each individual was also assigned a liability class. For individuals who are well, there were four age-dependent liability classes; 1 = <20 years, 2 = 20-30 years, 3 = 31-40 years and 4 = > 40 years. The other liability classes denote married-in individual as 5, 6 is bipolar disorder, 7 is recurrent major depression and 8 is other minor psychiatric illness. The disease locus parameters were specified in a separate text file and are shown in full in Appendix B, and are explained in section 1.4.1 and 6.2.1. The results of the parametric linkage analysis are the estimated multipoint LOD score at a particular location, followed by the estimate proportion of linked families (α) and the corresponding maximum heterogeneity LOD score (HLOD).

2.5.8. Non-parametric linkage analysis

Non-parametric linkage analysis was performed on the sub-pedigrees using MERLIN, according to the authors' instructions. The input files were prepared using MEGA2. The calculation of both the Whittemore and Halpern non-parametric linkage (NPL) pairs and NPL-all statistics were performed. The evidence for linkage was evaluated using both the Kong and Cox linear and exponential model. The output details the maximum possible scores for the dataset followed by analysis results at each location (cM position, Z score, *P*-value assuming normal approximation, Kong and Cox delta, Kong and Cox LOD score and Kong and Cox *P*-value).

2.5.9. Haplotype analysis

Construction of haplotypes from the family genotype data on regions of suggestive linkage were performed using the –best option in MERLIN. The input files were created using MEGA2. This created an output of the most likely haplotype vector. The haplotypes in each sub-pedigree were visualised using HaploPainter software.

Chapter 3
PI4K2B Expression Studies

3. PI4K2B Expression Studies

3.1. *Preface*

There are solid arguments to suggest *PI4K2B* as a good positional and functional candidate susceptibility gene for bipolar disorder, as described in detail in section 1.8. The reasons are based, in brief, on genetic evidence from a linkage study (Blackwood, He et al. 1996), an association study (Christoforou, Le Hellard et al. 2007) and haplotype analysis (Le Hellard, Lee et al. 2007) that prioritised region D on chromosome 4p15, within which *PI4K2B* is located. The involvement of *PI4K2B* in the phosphoinositide (PI) pathway, that is targeted by lithium and valproate for therapeutic effect in bipolar disorder, provides functional evidence (Berridge, Downes et al. 1989).

The hypothesis of this present study was that *PI4K2B* could confer susceptibility to bipolar disorder and recurrent major depression through altered expression levels. *PI4K2B* expression was measured at the RNA and protein level in lymphoblastoid cell lines from a large Scottish family. Some members of the family have a defined haplotype on chromosome 4p15-16, referred to as "linked haplotype" that segregates with the majority of cases of bipolar disorder and recurrent major depression. At inception of this study, there was no evidence for coding variants in *PI4K2B* in family members that carry the linked haplotype. However, the presence of variants in regulatory regions that alter gene expression can be detected by measuring allele specific expression (Pastinen and Hudson 2004). In this study, it was hypothesised that there was a difference in *PI4K2B* expression related to the linked haplotype, and that this difference could be detected in lymphoblastoid cell lines by allele-specific expression analysis.

RNA expression of *PI4K2B* was assessed in lymphoblastoid cell lines by allele-specific quantitative RT-PCR using Taqman technology. Three SNPs were used to quantify the relative levels of *PI4K2B* alleles in cDNA from linked haplotype carriers. The advantage to this method was the sensitivity obtained by comparing the linked haplotype to a series of different control alleles within the one family. This method could reveal the presence of *cis*-acting regulatory variants, which may contribute to susceptibility for bipolar disorder. *Cis*–acting variation has been reported to account for 25 –35% of inter-individual differences in gene expression throughout the human genome (Pastinen and Hudson 2004). In order to compare methodologies, allele-specific expression was also measured by PeakPicker analysis of sequence data (Ge, Gurd et al. 2005). Furthermore, as any disruption of *PI4K2B* gene regulation may manifest in protein expression abnormalities, it was important to investigate PI4K2B protein expression by quantitative Western blotting.

3.1.1. Lymphoblastoid cell lines available

Table 3.1 lists the 50 lymphoblastoid cell lines that were available for expression analysis. There were 20 samples with the "linked haplotype". Of these 20 samples, six samples were from individuals with a diagnosis of bipolar disorder, five samples were from individuals with a recurrent major depression diagnosis, one had another psychiatric diagnosis and eight samples were from individuals without a psychiatric diagnosis. For the samples without the disease haplotype, there were three samples from individuals with a recurrent major depression diagnosis, one with another psychiatric diagnosis, one sample was unknown, 15 samples were from individuals without a psychiatric diagnosis and 10 samples were from individuals outside of the family and were without a psychiatric diagnosis.

ID	Linked Haplotype	Psychiatric diagnosis	Heterozygote rs313548	rs313567 & rs6834255
29	Yes	Bipolar disorder		
47	Yes	Bipolar disorder	Yes	
57	Yes	Bipolar disorder	Yes	
61	Yes	Bipolar disorder	Yes	
99	Yes	Bipolar disorder	Yes	
143	Yes	Bipolar disorder	Yes	
45	Yes	Recurrent major depression	Yes	
48	Yes	Recurrent major depression	Yes	
55	Yes	Recurrent major depression	Yes	
63	Yes	Recurrent major depression	Yes	
163	Yes	Recurrent major depression		
30	Yes	Other		
36	Yes	Well	Yes	
54	Yes	Well		
58	Yes	Well	Yes	
69	Yes	Well		Yes
73	Yes	Well		Yes
76	Yes	Well	Yes	
97	Yes	Well		
103	Yes	Well	Yes	
153	No	Other		
34	No	Recurrent major depression		Yes
40	No	Recurrent major depression	Yes	Yes
164	No	Recurrent major depression		
71	No	Unknown		
14	No	Well		
32	No	Well		
35	No	Well	Yes	
42	No	Well	Yes	Yes
56	No	Well		
72	No	Well		Yes
86	No	Well	Yes	
113	No	Well	Yes	Yes
127	No	Well	Yes	Yes
128	No	Well		
129	No	Well	Yes	Yes
131	No	Well		
132	No	Well		
133	No	Well		
142	No	Well		
11	No	Well & married-in	Yes	Yes

Table 3.1 Lymphoblastoid cell lines information. Continued on the next page.

13	No	Well & married-in		Yes
33	No	Well & married-in	Yes	Yes
38	No	Well & married-in		
39	No	Well & married-in		
46	No	Well & married-in	Yes	Yes
49	No	Well & married-in		
70	No	Well & married-in		Yes
98	No	Well & married-in		
141	No	Well & married-in		

Table 3.1 Lymphoblastoid cell lines information. The ID number is according to the pedigree. The psychiatric diagnosis is listed for all samples; other psychiatric diagnosis can mean a single episode depression, anxiety etc. The presence or absence of the chromosome 4p15-16 “linked haplotype” is marked. Samples that were heterozygous for the markers used in the allele-specific expression studies are labelled. The same individuals are heterozygote for both SNPs rs313567 and rs6834255. The table is ordered according to “linked haplotype” status and then by diagnosis.

The methods used for RNA, DNA and protein preparations from the lymphoblastoid cell lines are described in section 2.2. Each lymphoblastoid cell line was cultured in the same way and the RNA extractions and cDNA preparations were prepared in a consistent manner. The RNA was quantified using the Agilent 2100 Bioanalyser. This method also determined the quality of the RNA. The Agilent RNA 6000 Nano kit was used. This was performed by Margaret McLean in the WT-CRF. All the lymphoblastoid cell lines were tested and found to be negative for mycoplasma infection, as contamination can affect uptake across cell membranes, interfere with membrane receptor function, cause morphological change, influence amino acid and nucleic acid metabolism and induce cell transformation (www.ecacc.org.uk). Consequently, these could have substantial effects on the reproducibility of results, in particular gene expression.

3.2. *No Sequence Variants in PI4K2B*

Primers were designed from repeat-free intronic regions to incorporate ~100bp either side of the exon, to include possible splice site SNPs and to amplify ~600bp genomic DNA fragments. *PI4K2B* was sequenced in 46 samples from individuals in four linked families, please refer to section 2.1.1.3 allele sharing panel. There was no evidence of a mutation in the coding region of *PI4K2B*.

3.3. *PI4K2B Expressed in Lymphoblastoid Cell Lines*

Figure 3.1 showed that the *PI4K2B* gene was expressed in two lymphoblastoid cell line samples, 58 and 55. *PI4K2B* cDNA was amplified using RT-PCR (reverse transcription-polymerase chain reaction) technique with gene specific primers and sequenced to check PI4K2B specificity. This was an important first step for future experiments, as not all genes are expressed in lymphoblastoid cell lines.

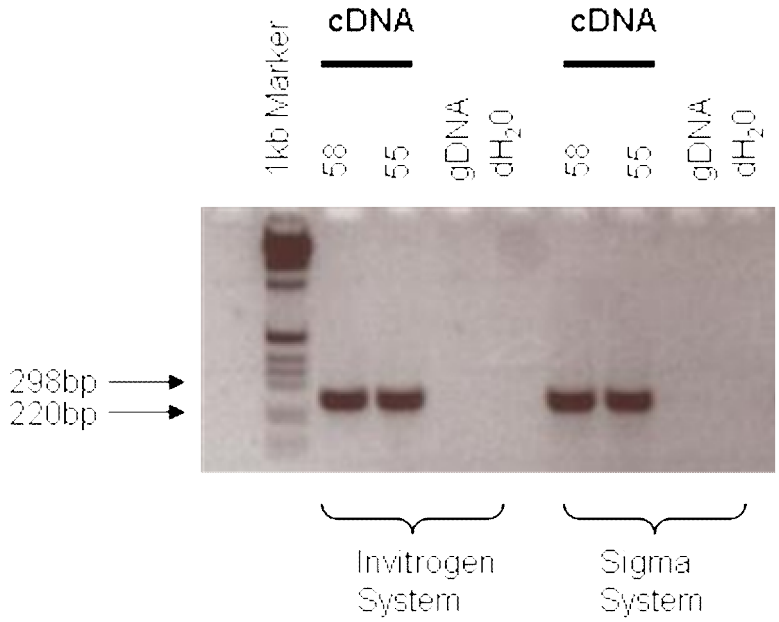


Figure 3.1 PI4K2B is expressed in lymphoblastoid cell lines. *PI4K2B* was amplified by RT-PCR on cDNA from lymphoblastoid cell lines. The intron spanning primers ensured there was no band in the gDNA control. The water control was used to test for contaminants in the PCR reaction mix. Two reaction mixes from Invitrogen and Sigma were tested. The band of expected size 255bp was present.

3.4. *Allele-Specific Expression*

3.4.1. Marker selection

Three SNPs were chosen to measure *PI4K2B* allele expression and Taqman assays were designed to amplify the SNPs. TaqMan® SNP Genotyping Assays consisted of unlabelled PCR primers and TaqMan® MGB probes (FAM™ and VIC® dye-labelled) that were designed for the allelic discrimination of the specific SNPs. Table 3.2 details the features of the TaqMan® SNP Genotyping Assays. Two of the assays were premade, off-the-shelf assays "assays on demand" (c____764549 and c____764538) and one assay was an "assay by design" (rs6834255). SNP assays were chosen for an exonic coding region, an intron and in the 3'UTR region to look for a difference in expression between two alleles. The intronic assay was included as studies have reported the detection of allele-specific expression from heterogeneous nuclear RNA (hnRNA), which is the unprocessed precursor of the mature, functional mRNA (Pastinen, Sladek et al. 2004). Allele specific expression was measured in cDNA prepared from lymphoblastoid cell lines.

HapMap reference	ABI Assay name	Physical Coordinates (May 04)	PI4K2B genomic Location	Strand	Observed	Assay	Fam Labelled Allele	Effect of SNP	Linked haplotype	number Linked Haplotype Heterozygotes	number Control Haplotype Heterozygotes
rs313548	c____764549	24,913,660	Intron 1	Minus	CT	AG	G	None known	G	13	10
rs313567	c____764538	24,930,264	Exon 2	Positive	CT	CT	T	Synonymous, no change in peptide	T	2	12
rs6834255	rs6834255	24,955,589	Exon 10	Positive	AG	AG	A	None known, in UTR, prior to polyadenylation signal	G	2	12

Table 3.2 *PI4K2B* SNPs used in allele-specific assays. The rs number refers to the name of the SNP as www.hapmap.org. The Applied Biosystems (ABI) reference is the name of the assay provided to amplify the specific SNP. rs6834255 does not have a specific ABI reference as it was designed for this study ABD (assay by design). The physical and genomic locations of the SNPs were provided by the UCSC genome browser (May 2004). The strand information, the observed genotypes, the assay genotypes and the particular allele labelled with FAM dye were provided by ABI. The effect of the SNP shows the potential action the SNP may have on peptide sequence. UTR was the Untranslated Region. The genotype of the linked haplotype is listed, as Figure 3.10. The number of samples from lymphoblastoid cell lines that are heterozygote for each particular SNP is listed.

3.4.1.1. Quality control of SNP assays

The three SNP assays were quality controlled by checking for ease and reliability of scoring and Mendelian segregation, using samples from 46 individuals from four linked families and samples from 32 trios, as described in section 2.1.1.3 and 2.1.1.4 respectively. The three assays fully passed these quality control measures. The markers rs313567, rs313548 and rs6834255 were also genotyped in 444 members of the Scottish population. This was the same population used for the control group in the *PI4K2B* case control association study in chapter 4. This allowed additional quality control analysis to be carried out. The genotyping success for the three SNPs was >97%. The three markers also passed the Haploview recommended Hardy-Weinberg equilibrium significance threshold (P=0.001) at P=0.57, 0.18 and 0.15 for rs313567, rs313548 and rs6834255 respectively. Furthermore, as shown in Table 3.3 the three SNPs were in complete LD with each other, based on D'=1, which implied there was no evidence for ancestral recombination event between two SNPs.

MARKER 1	MARKER 2	D'	LOD	R^2	CILOW	CIHI	DIST(BP)
rs313548	rs313567	1	69.66	0.627	0.96	1	16,604
rs313548	rs6834255	1	67.89	0.618	0.96	1	41,929
rs313567	rs6834255	1	115.41	1	0.98	1	25,325

Table 3.3 Linkage disequilibrium between *PI4K2B* SNPs. D' is the value of D prime between the two loci, LOD is the log of the likelihood odds ratio, a measure of confidence in the value of D', r^2 is the correlation coefficient between the two loci, CIlow is 95% confidence lower bound on D', CIhi is the 95% confidence upper bound on D'. Dist is the distance in bases between the loci.

Only rs313567 and rs6834255 are in perfect LD, based on r^2=1. Perfect LD means r^2 = 1, where the SNPs are redundant and act as "perfect proxies". This was reflected by the genotypes observed in all sample sets genotyped: lymphoblastoid cell line

samples, 46 samples from four-linked families and 32 trios. Individuals were consistently heterozygous or homozygous at both rs313567 and rs6834255, but this correlation did not extend to rs313548.

Having passed these quality control checks, the markers were genotyped on all available lymphoblastoid cell lines. There was no information for eight lymphoblastoid cell lines due to either genotyping failure or unavailability of the cell line at that time. These samples were 32, 30, 132, 97, 98, 164, 38 and 153. Of the available 42 samples, those heterozygous for specific SNPs were used, so the relative contribution of the two alleles could be compared. There were 23 samples heterozygote for rs313548, 13 of which have the linked haplotype (five bipolar disorder, four recurrent major depression, four well) and 10 have a control haplotype (one recurrent major depression, six well, three well and married-in). There were 14 samples heterozygote for both rs313567 and rs6834255, two samples had the linked haplotype and are well, 12 samples had the control haplotype (two recurrent major depression, five well, five married-in and well). All of these heterozygote samples were tested for allelic imbalance according to the optimised experimental procedure, as outlined in section 2.2.8.

3.4.2. Taqman Assays

3.4.2.1. Standard curves of genomic DNA dilutions

The initial experimental design was to prepare a standard curve made from a series of genomic DNA (gDNA) homozygote dilutions. This standard curve could then be used to read the expression of each allele from the cDNA for each individual. This approach was first reported by Pastinen *et al*, and has been used by Kimura *et al* and Lo *et al*.(Lo, Wang et al. 2003; Pastinen, Sladek et al. 2004; Kimura, Nishioka et al. 2005). Several stages of the experimental procedure were optimised to improve the quality of the standard curves: pipetting accuracy was improved with electronic pipettes, DNA concentration was estimated by picogreen reagent, steps were taken

to ensure that DNA dilutions were fully dissolved, the DNA concentration required for each particular assay was investigated, the number of replicates was increased and the most appropriate analysis method was sought. Standard curves with different homozygotes mixes were prepared to test the precision of each assay. Three standard curves with four different homozygotes in different combinations were prepared to test rs313567 assay, and two standard curves with four different homozygotes were prepared to test the precision of rs313548 and rs6834255 assays. Each homozygote mix was assayed in quadruplicate.

Figure 3.2 shows the optimised standard curve for rs313567 obtained using the LOG_2 ratio of FAM allele:VIC allele against the LOG_2 Ratio of FAM™ /VIC® reporter dye intensity. The data was transformed for normal distribution (Bland and Altman 1996). A linear regression line was drawn for each standard curve to report on the linearity of the ratios and accuracy of the triplicate assays. The correlation within the four replicates of each standard curve was good, $r^2>0.97$, as was the correlation between the three standard curves, $r^2>0.99$. However, for each homozygote dilution, represented in Figure 3.2, the 95% confidence intervals overlapped. This showed that a difference could not be confidently determined between neighbouring homozygote dilutions. Figure 3.3 and Figure 3.4 show the standard curve data for assays rs313548 and rs6834255 respectively. The correlation within the three replicates for each standard curves was also good, r2>0.95 and $r^2>0.98$ respectively, as was the correlation between two standard curves for both assays, $r^2>0.98$. For both of these assays, the 95% confidence intervals also overlapped between neighbouring homozygote dilutions and a difference could not be confidently determined.

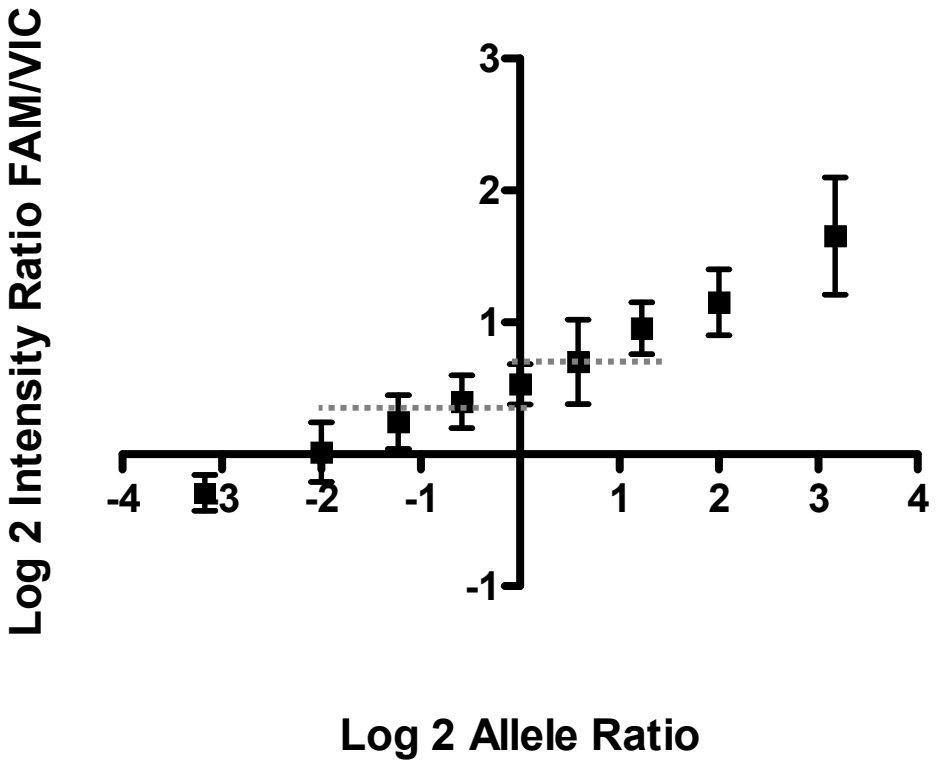


Figure 3.2 Standard curve of gDNA dilutions at rs313567. The graph shows the known homozygote dilutions allele ratio on the x-axis against the intensity value from the assay on the y-axis. FAM and VIC labelled the two alleles. There were nine homozygote dilutions tested, from left to right on the graph diluted according to the following ratios: 90:10, 80:20, 70:30, 60:40, 50:50, 40:60, 30:70, 20:80 and 10:90. The box shows the mean value of three standard curves for each homozygote dilution. There were four different homozygote gDNA samples mixed; 132 (CC) & 86 (TT), 132 (CC) & 35 (TT), T3445 (CC) & 86 (TT). There were four replicates of each mix. The error bars show 95% CI range. The grey dotted lines from the 50:50 homozygote mix (log2=0) show the limit of estimating allelic imbalance from the standard curve.

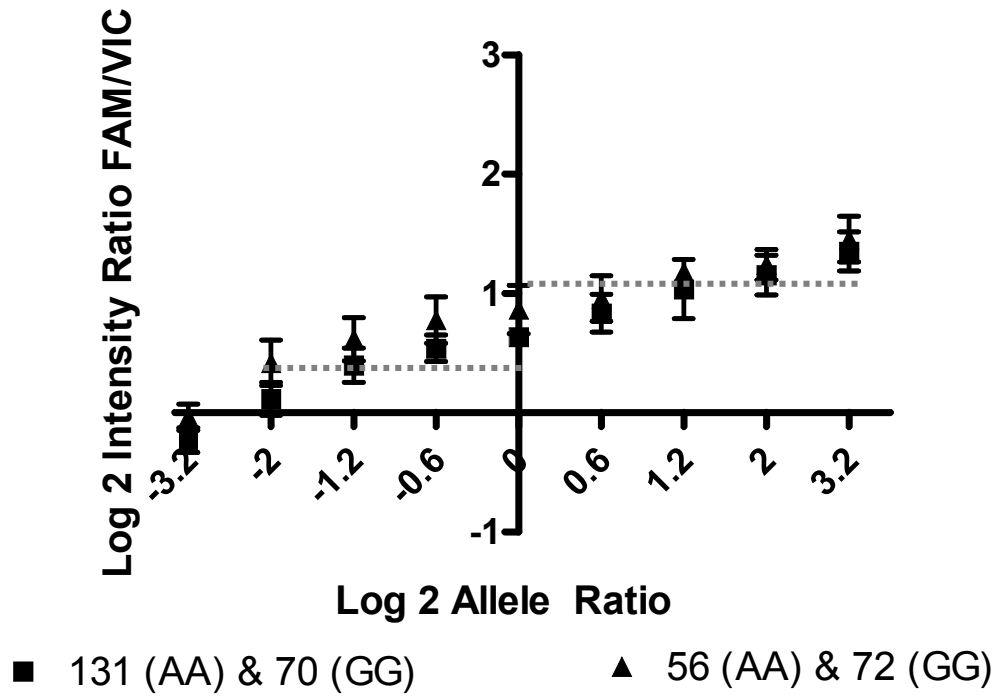


Figure 3.3 Standard curve of gDNA dilutions at rs313548. The graph shows the known homozygotes allele ratio on the x-axis against the intensity value of the assay on the y-axis. FAM and VIC labelled the different alleles. There were nine homozygote dilutions tested, from left to right on the graph diluted according to the following ratios: 90:10, 80:20, 70:30, 60:40, 50:50, 40:60, 30:70, 20:80 and 10:90. The triangle represents the mean value of three replicates for one standard curve; 131 (AA) & 70 (GG). The box represents the mean value of three replicates for another standard curve 56 (AA) & 72 (GG). The error bars show 95% CI range. Each standard curve was correlated, r^2>0.95. The grey dotted lines from the 50:50 homozygote mix (log2=0) show the limit of estimating allelic imbalance from the standard curve.

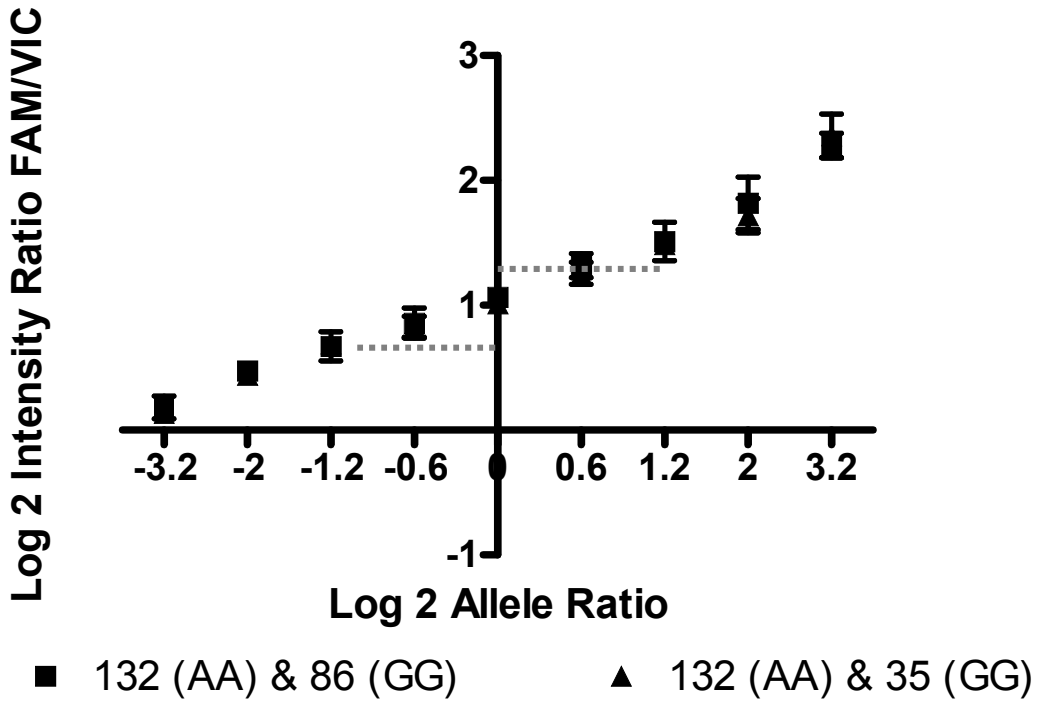


Figure 3.4 Standard curve of gDNA dilutions at rs6834255. The graph shows the known homozygotes allele ratio on the x-axis against the intensity value of the assay on the y-axis. FAM and VIC labelled the different alleles. There were nine homozygote dilutions tested, from left to right on the graph diluted according to the following ratios: 90:10, 80:20, 70:30, 60:40, 50:50, 40:60, 30:70, 20:80 and 10:90. The box represents the mean value of three replicates for one standard curve; samples 132 (AA) & 86 (GG). The triangle represents the mean value of three replicates for another standard curve; samples 132 (AA) & 86 (GG). The error bars show 95% CI range. Each standard curve was correlated, r^2>0.98. The grey dotted lines from the 50:50 homozygote mix (log2=0) show the limit of estimating allelic imbalance from the standard curve.

3.4.2.2. Estimation of allelic imbalance from standard curves

cDNA from each heterozygote lymphoblastoid cell line was assayed with four replicates from the same RNA preparation in each assay. The four replicates within the three assays showed good reproducibility of individual cDNA ratios with an average coefficient of variation (SD/mean) of <0.037. A cut-off average coefficient of variation <0.05 is often used (Bray, Preece et al. 2005). Each assay was repeated three times with the same preparation of cDNA and these assays showed good reproducibility with average coefficient of variation <0.034. The intensity of the cDNA heterozygotes was then compared to the relevant standard curve. The grey dotted lines on Figure 3.2, Figure 3.3 and Figure 3.4 show the only difference that could be reliably detected from equal expression from both alleles (50:50) was a 80:20 allele ratio at rs313567, a 90:10 allele ratio at rs313548 and a 70:30 allele ratio at rs6834255, respectively. This was not as sensitive, as previous studies have shown a threshold for detecting allelic imbalance at a 63:37 allelic ratio (Pastinen, Sladek et al. 2004)

The allelic imbalances observed for the heterozygote cDNA samples did not reach the detection threshold. The mean LOG_2FAM/VIC intensity for 14 heterozygotes at rs313567 = 0.5 (minimum 0.47 - maximum 0.55, 95%CI 0.01), for 23 heterozygotes at rs313548 = 0.64 (minimum 0.57 - maximum 0.69, 95%CI 0.01) and for 14 heterozygotes rs6834255 = 1.15 (maximum 1.24 – minimum 1.05, 95% CI 0.02). Figure 3.2, Figure 3.3 and Figure 3.4 show that these mean cDNA allelic values lie in the middle of each of the standard curves and do not extend past the confidence intervals for allelic imbalance beyond 50:50. Thus, no allelic over-expression was observed to this degree for any of the three assays. Moreover, the standard curves were not sensitive enough to detect a small difference between disease haplotype carriers and controls.

Proof of this is shown in Figure 3.5, which illustrates a comparison with gDNA from heterozygotes to that of gDNA equal homozyogote mixes, 50:50. The gDNA equal homozygote mixes should mimic the real heterozygotes. The gDNA heterozygotes cluster whereas the homozygote mixes have a greater spread. This illustrates the technical difficulties with creating accurate mixtures of DNA, and thus the lack of sensitivity of the standard curve method.

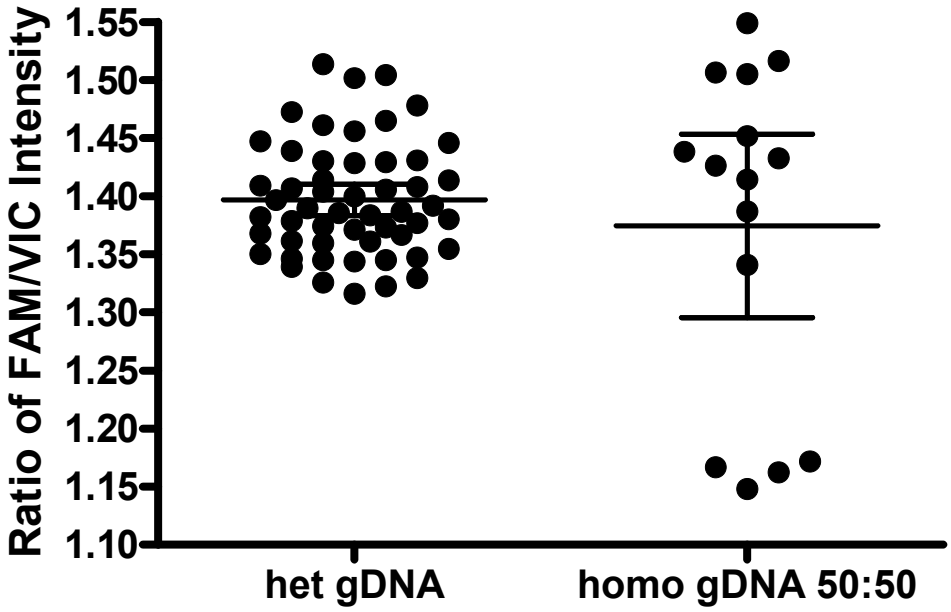


Figure 3.5 Comparison of gDNA from heterozygotes with 50:50 homozygote mixes for rs313567. There were four replicates for each sample. There were fourteen heterozygote (het) gDNA and four different standard curves. The standard curve homozygote mixes were 132 & 86, 132 & 35, T3445 & 86, F119 T3195 & 56. The mean value is represented by the horizontal bar and the error bars show the 95% confidence interval.

3.4.2.3. Direct comparison method

As the creation of accurate standard curves was unsuccessful, I adopted another method, which incorporated gDNA measurements from heterozygotes, which represented equal expression from each allele, as a control. This method has been used in many studies (Hoogendoorn, Norton et al. 2000; Bray, Jehu et al. 2004; Hannula-Jouppi, Kaminen-Ahola et al. 2005; Li, Grupe et al. 2006; Williams, Glaser et al. 2008). Each of the three assays was performed. gDNA and cDNA samples were assayed from 23 heterozygote samples for rs313548 and 14 heterozygotes samples for both rs313567 and rs6834255. The data was analysed as described in section 2.3.2. In brief, the ratio of cDNA:gDNA was calculated, as genomic DNA represents a 1:1 ratio of the two alleles. The ratio allows any assay specific and PCR amplification bias to be taken into account. For example, the FAM label was detected at a greater intensity than the VIC label intensity (Angie Fawkes, Personal Communication, 30th May 2005, Wellcome-Trust Clinical Research Facility) and incorporating both gDNA and cDNA in the same assay will cancel any intensity bias. Figure 3.6 and Figure 3.7 (a and b) showed the cDNA:gDNA ratios for rs313548, rs313567 and rs6834255 assays respectively.

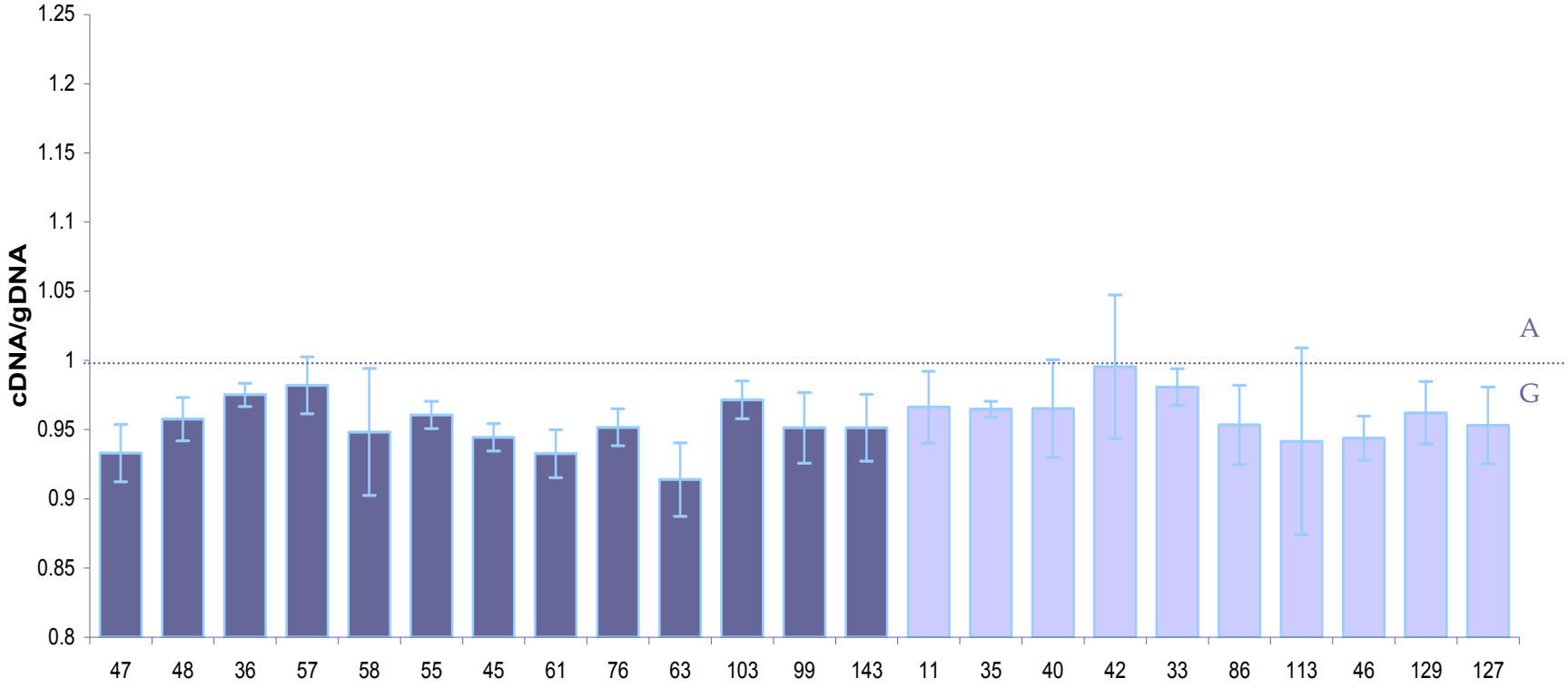


Figure 3.6 PI4K2B allelic expression at rs313548 between linked haplotype carriers and controls. The heterozygote lymphoblastoid cell line samples are shown on the x-axis. The cDNA/gDNA ratio for each sample is shown on the y-axis. This ratio is the mean of 3 assays, each with 4 replicates, calculated as intensity FAM/intensity VIC. The error bars show 95% Confidence Interval. The dark blue bars represent samples with the *PI4K2B* linked haplotype that are heterozygote at this SNP. The light blue bars represent samples without the linked haplotype. A dotted line is drawn at 1, which represents equal expression from both alleles.

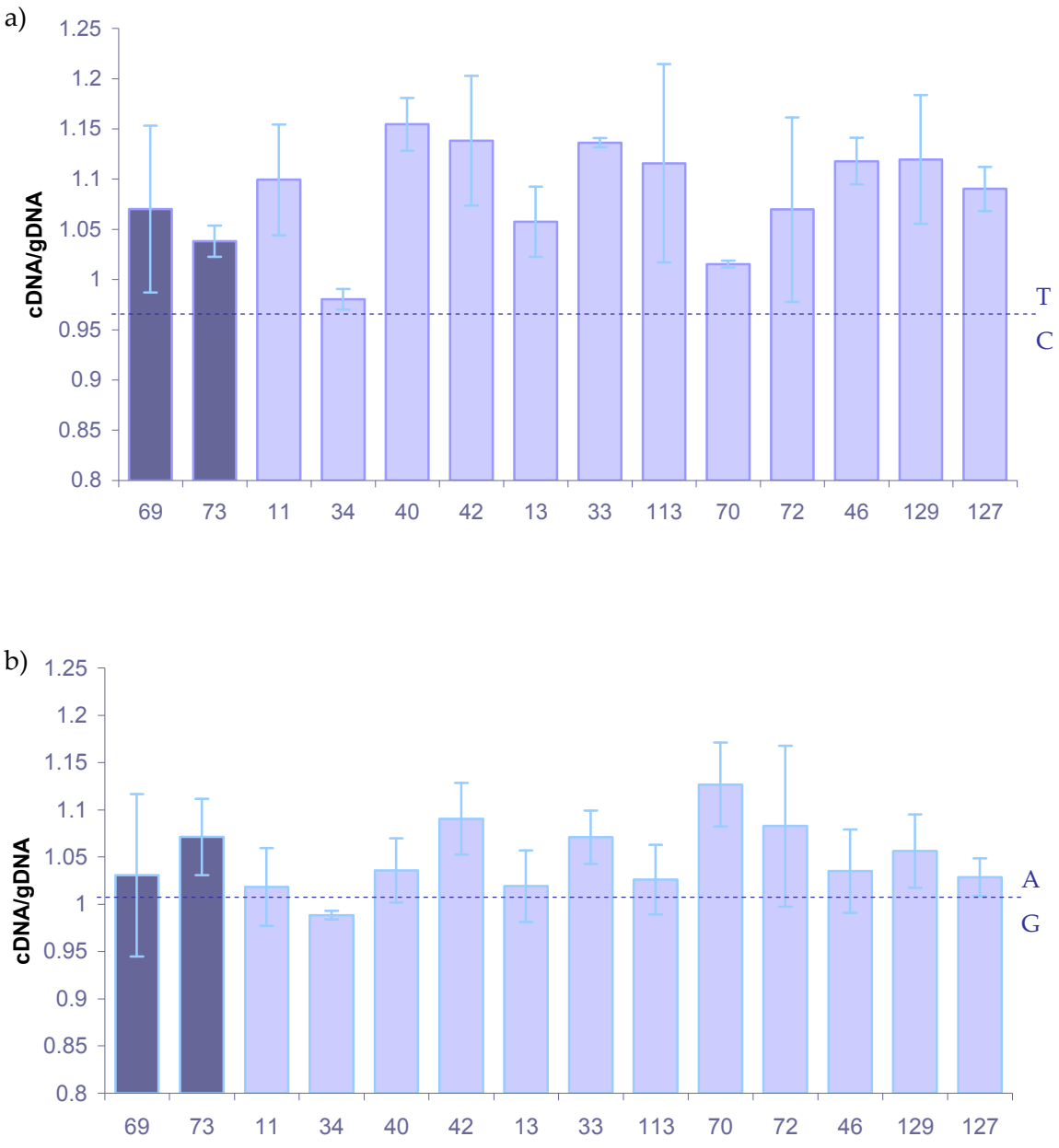


Figure 3.7 Allelic expression between linked haplotype carriers and controls at rs313567 and rs6834255. The heterozygote lymphoblastoid cell line samples are shown on the x-axis. The cDNA/gDNA ratio is shown on the y-axis for rs313567 (a) and rs6834255 (b). This ratio is the mean of 3 assays, each with 4 replicates, calculated as intensity FAM/intensity VIC. Error bars show 95% Confidence Interval. The dark blue bars represent samples with the linked haplotype that are heterozygote at these SNPs. The light blue bars represent samples without the linked haplotype. A dotted line is drawn at 1 which represents equal expression from both alleles.

It was possible to test for a difference in allelic expression between samples that are heterozygote for rs313548. There were 13 heterozygote individuals with the *PI4K2B* linked haplotype and 10 individuals without. Allelic expression was compared between the two groups using a T Test. The Student-T test, with a two-tailed distribution, determined whether two samples are likely to have come from the same two underlying populations that have the same mean. The result was non-significant, *P*=0.28, thus I was unable to reject the Null Hypothesis that there was no difference in allele expression between individuals with the *PI4K2B* linked haplotype and those who do not have the haplotype.

A difference in allelic expression was also tested in the other two SNP assays, rs313548 and rs6834255. It was evident from Figure 3.7 that the cDNA/gDNA ratio of the samples with the linked haplotype, 69 and 73, did not differ from the other 12 samples without. However, it would be difficult to detect a difference in only two samples with the linked haplotype. The T-test did not suggest a significant difference between *PI4K2B* linked haplotype samples and the control samples, *P*=0.36 (rs313548) and *P*=0.98 (rs6834255).

3.4.3. PeakPicker assays

A second technique for measuring allele expression, namely the "PeakPicker" method was also attempted (Ge, Gurd et al. 2005). This method had previously shown to be successful in detecting allelic expression differences (Pastinen, Ge et al. 2005). This technique investigated the sequence trace produced by amplification of the DNA surrounding the SNP and measured the peak height ratios of the two alleles of the SNP, and used the peak heights of the surrounding sequence to calculate the level of expression of the two alleles. The peak height ratios for both gDNA and cDNA were measured and compared. Figure 3.8 illustrates an example on samples 69 and 70.

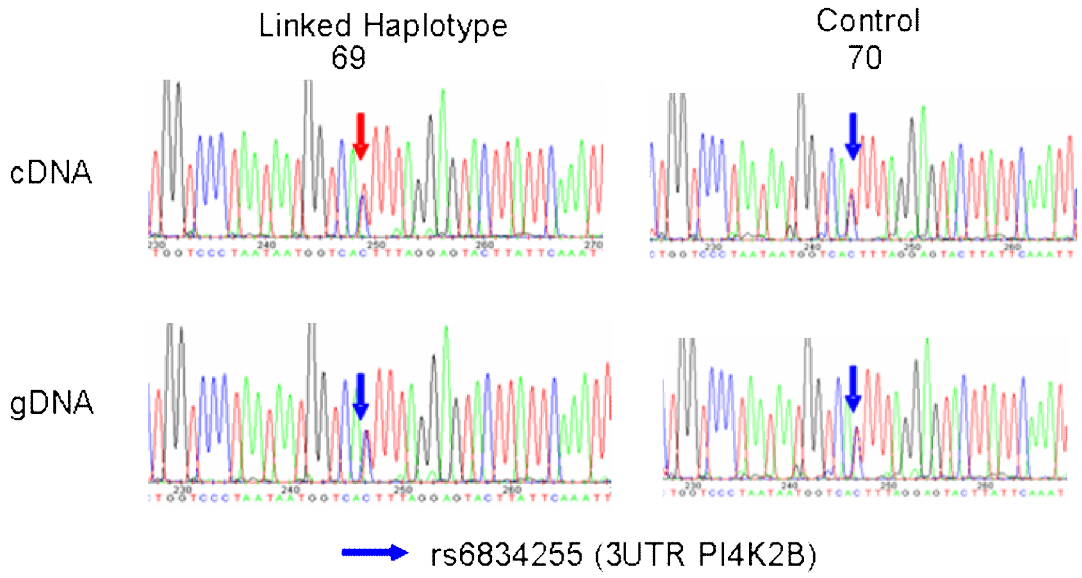


Figure 3.8 Peak heights of the SNP in cDNA & gDNA using PeakPicker software. Sequence trace for cDNA and gDNA for two lymphoblastoid cell line samples are shown surrounding the marker rs6834255. The arrows point to the two alleles of rs6834255. The peaks correspond to different bases: green peaks are adenine (A), blue peaks are cytosine (C), black peaks are guanine (G) and red peaks are thymine (T).

This method was attempted for the three SNPs in duplicate for both cDNA and gDNA from all heterozygote lymphoblastoid cell line samples. The cDNA was prepared from a single RNA sample that was also used in the Taqman assays. Each DNA sample was sequenced twice. Peak height ratios of the SNP from the sequence traces were measured and normalised based on the surrounding sequence using PeakPicker software, according to the authors instructions (Ge, Gurd et al. 2005). This method was successfully completed for SNP rs6834255. However, reliable peak estimates surrounding rs313548 in *PI4K2B* intron 1 could not be obtained as there was double sequence in both cDNA and gDNA. Also, the sequence surrounding rs313567 was only of good quality in three out of fourteen samples of gDNA and in thirteen out of fourteen samples of cDNA, thus a valid comparison between cDNA and gDNA could not be made. The problems in obtaining good quality sequence traces may have occurred due to poor DNA quality, poor primer design which may have caused the primer to bind to two or more sites on the template, unsuitable PCR

amplification conditions or insufficient clean-up reaction. The sequence surrounding rs6834255 was successfully obtained for 13 gDNA samples (12 samples in duplicate and one singly) and 13 cDNA samples (nine in duplicate and four singly). Analysis was performed on this data.

Figure 3.9a shows level of allelic imbalance at rs6834255 for 13 heterozygotes. The samples with the linked haplotype, 69 and 73, lie in the range of middle of allelic expression ratios of samples with the control haplotype. A student T-test, with a two-tailed distribution, was performed and showed no evidence for a difference between the two samples with the linked haplotype and those samples without (P=0.91).

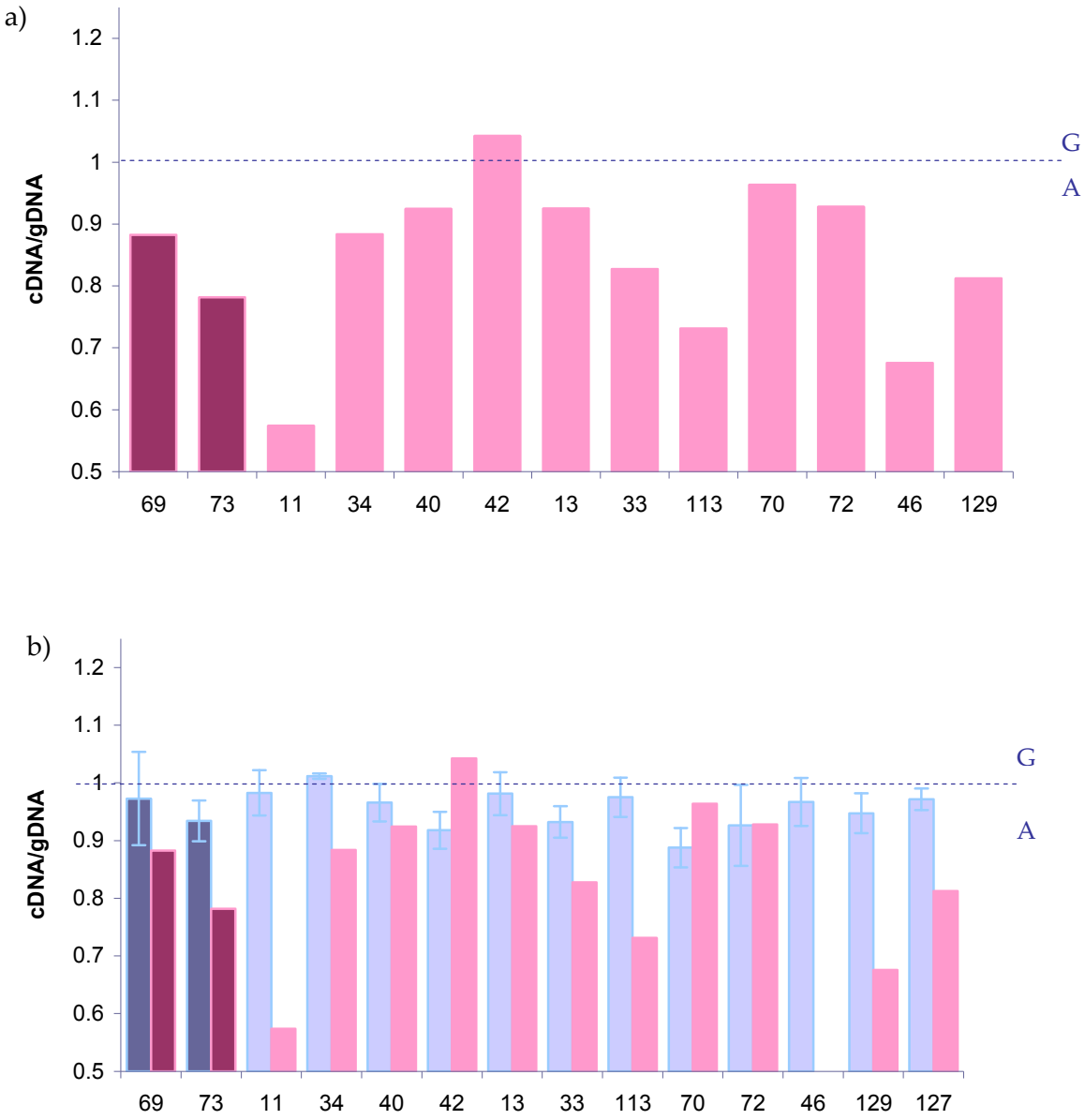


Figure 3.9 *PI4K2B* allelic imbalance by Taqman and PeakPicker methods. The x-axis shows the lymphoblastoid cell line samples that were heterozygous for the *PI4K2B* 3'UTR SNP, rs6834255. The y-axis shows the cDNA/gDNA ratio. Figure (a) shows the DNA ratio estimated by the PeakPicker method by the pink bars. This ratio was the mean of two amplifications from each individual, sequenced once. Figure (b) shows the DNA ratio as the PeakPicker method (pink bars) beside the ratio obtained by the Taqman method (blue bars). The Taqman ratio was the mean of 3 assays, each with 4 replicates, calculated as VIC/FAM. Error bars show 95% Confidence Interval for the replicate Taqman Assays. The mean coefficient of variation experiment-wide is 0.03 (Taqman assays) and 0.06 (PeakPicker analysis).The dark bars represent samples with the linked haplotype (69 and 73) that are heterozygous at rs6834255. The dotted line at 1 on the y-axis represents equal expression of both alleles and labels the direction of expression of each allele [Guanine (G) or Adenine (A)]

Figure 3.9b compares the allelic expression at this marker for the PeakPicker data to the Taqman data. In general, the allelic expression detected by both the PeakPicker and Taqman methods showed the same direction to the A allele, except for sample 34 measured by the Taqman method and for sample 42 measured by the PeakPicker method. However, the measurements of expression were not the same between the PeakPicker and Taqman method. In fact, the majority of the PeakPicker measurements did not overlap with the 95% confidence intervals from the Taqman measurements. Sample 46 was the only one that displayed the same expression levels by both methods.

There were many caveats to this analysis. Firstly, the number of replicates in the PeakPicker method was less than the Taqman assays. Technical difficulties curbed further experiments to increase the number of replicates and so confidence intervals could not be calculated. Secondly, the mean experiment-wide coefficient of variation for the PeakPicker method, CV=0.06, was greater than both the Taqman assays, CV=0.03, and the recommended cut-off, CV=0.05. Finally, experiments by other members in the laboratory showed unacceptable levels of variability in the PeakPicker method (Susan Anderson and Helen Torrance, Personal communication, 26th July 2006, Molecular Medicine Centre, University of Edinburgh). Thus, the PeakPicker method did not prove a reliable manner to measure allele-specific expression levels for this study.

3.4.4. Haplotype analysis

To investigate *PI4K2B* expression at the haplotype level, the allele-specific analysis by Taqman of the three SNPs was combined. In theory, the linked allele at each of the three SNPs should show the same strength and direction of allelic expression. Firstly, it was possible, through segregation analysis to determine haplotypes at the three SNPs in the individuals tested. Secondly, the data was combined to determine the expression of the haplotype.

3.4.4.1. Determining phase of the SNPS

From Table 3.2 the possible genotypes for SNPs rs313548, rs313567 and rs6834255 are A/G, C/T and A/G respectively. The number of possible haplotypes is 2^3 haplotypes (8 haplotypes). It was firstly determined that the markers showed Mendelian segregation through the family. Previous work completed by our group identified the individuals that carry the linked haplotype, as listed in Table 3.1. This information was used to identify the linked haplotype for the three *PI4K2B* SNPs. For example, Figure 3.10 shows that for SNP, rs313548 the genotype on the linked haplotype must be G, and for rs313567 and rs6834255 the genotype must be TG. Thus, the linked haplotype was GTG. All genotyped samples with the linked haplotype have the haplotype G-T-G.

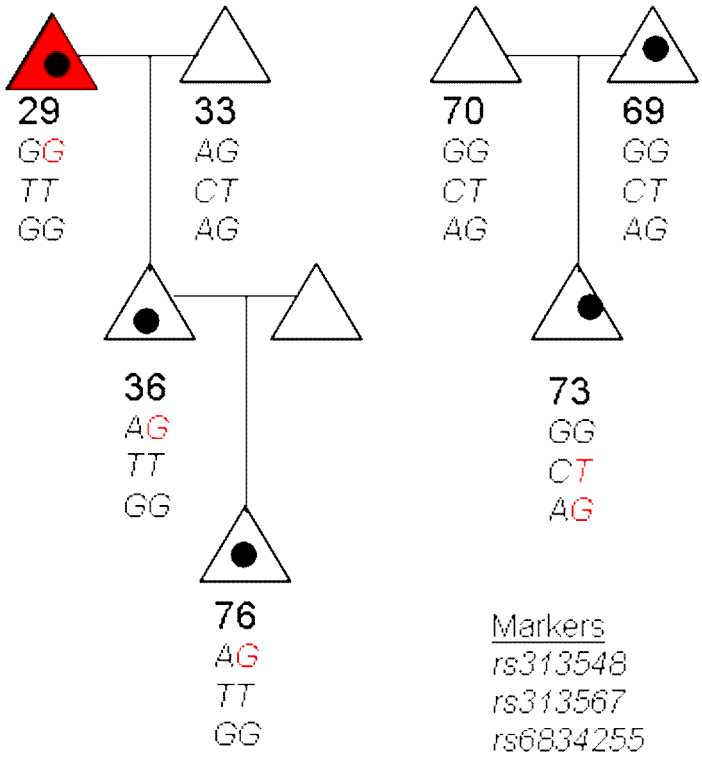


Figure 3.10 Determination of haplotype phase of *PI4K2B* SNPs. Two branches of the large family are shown here. The genotypes for each sample are listed in the following order; rs313548, rs313567 and rs6834255. The red-filled triangle shows a sample from an individual with a bipolar disorder diagnosis. The small black circle means the sample was from an individual with the linked haplotype.

Figure 3.11 illustrates the aim of combining the analysis to compare expression at the haplotype level of the linked haplotype versus the control haplotype.

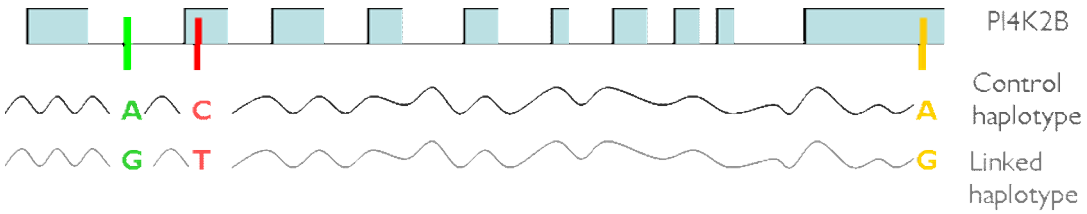


Figure 3.11 Illustration of haplotype of SNPs used in allele-specific expression assays. The PI4K2B gene structure is shown here with boxes that represent exons and lines that represent introns. The SNP markers are rs313548, rs313567 and rs6834255 in intron 1, exon 2 and 3'UTR region of PI4K2B, respectively. The control haplotype is defined as A-C-A and the disease haplotype is defined as G-T-G.

3.4.4.2. Haplotype analysis of Taqman data

The expression data was combined for the three assays to measure the *PI4K2B* linked haplotype; G-T-G for rs313548-rs313567-rs6834255. There are eight samples who were heterozygous at all three SNPs and for whom the haplotype could therefore, be investigated using mean cDNA/gDNA ratios. Although sample 11 was heterozygous for the three markers, this sample was not included, as one haplotype was A-T-G and the other haplotype was G-C-A for rs313548, rs313567 and rs6834255, respectively, as determined by segregation analysis. For the other seven samples, their haplotypes were the same, namely G-T-G and A-C-A for rs313548, rs313567 and rs6834255, respectively.

Figure 3.12 shows the haplotype analysis for seven individuals who were heterozygous for the three SNP markers. In general, SNPs rs313567 and rs313548 show the same direction of allelic expression, while SNP rs6834255 shows the opposite direction. Thus, the same expression pattern was not detected at all three SNPs. This analysis was not used to investigate linked haplotype versus control as none of the eight samples that were heterozygous at all three markers had the linked haplotype.

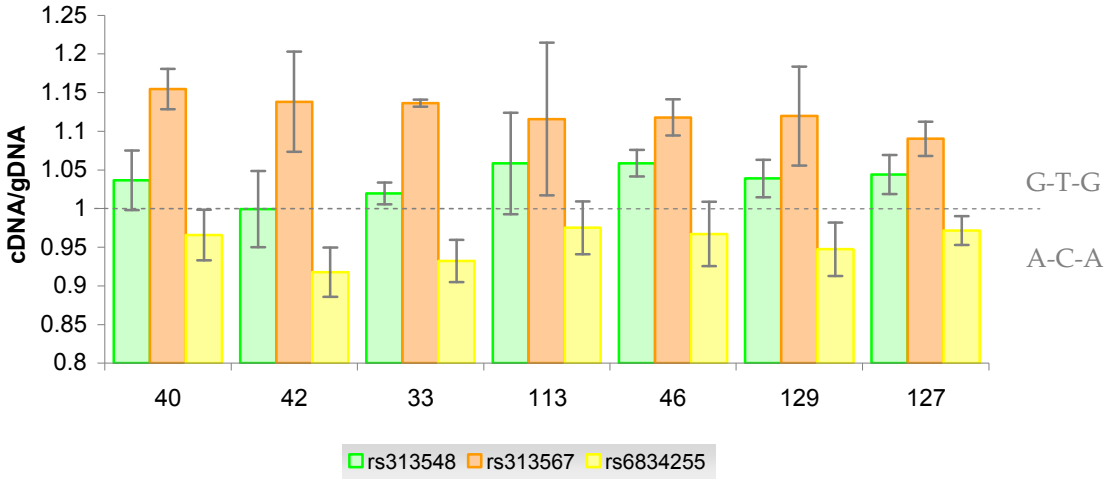


Figure 3.12 Haplotype analysis of *PI4K2B* allelic expression. The x-axis shows the lymphoblastoid cell line samples. The y-axis shows the cDNA/gDNA ratio. This ratio is the mean of three assays, each with four replicates, calculated as VIC/FAM (rs3131548 A/G and rs6834255 G/A) and FAM/VIC (rs313567 T/C). Error bars show 95% Confidence Interval. The dotted line that crosses the y-axis at one represents equal expression from both alleles.

3.5. *Alternative Splicing of PI4K2B*

3.5.1. Preface

A hypothesis to explain the discordance of *PI4K2B* allele-expression at the haplotype level shown in Figure 3.11 was alternative splicing of *PI4K2B*. In theory, there may be several *PI4K2B* isoforms, some of which may not contain all three SNPs in this study. Alternatively, the intronic SNP may be contained in a pseudoexon. A pseudoexon is an intronic sequence that resembles an exon because it matches 3′ and 5′ splice sites that are not normally spliced. The hypothesised variants could regulate *PI4K2B* transcript levels in an isoform-specific manner. Conversely, there may be differential regulation of particular isoforms that could include the SNPs in question.

Here, the three SNPs used in measuring allele-specific expression in *PI4K2B* were examined to see if they were contained in different isoforms. Two approaches were employed i) bioinformatics using the UCSC genome browser and ii) PCR techniques using both RT-PCR and 3′ Rapid Amplification of cDNA Ends (RACE) on cDNA from lymphoblastoid cell lines. The second experimental approach was undertaken because, despite thorough analysis using EST, cDNA and gene-prediction approaches on genome browsers, a significant fraction of alternative 5′- or 3′-terminal exons and internal alternative splicing can be missed (Wang and Cooper 2007).

The UCSC human genome browser was used to search for evidence of alternative transcripts, such as multiple transcription initiation sites (first exons), polyadenylation sites and alternatively spliced internal and terminal exons. Figure 3.13 illustrates the CpG islands and polyadenylation sites that suggest the 5′ and 3′ boundary of the *PI4K2B* gene, respectively.

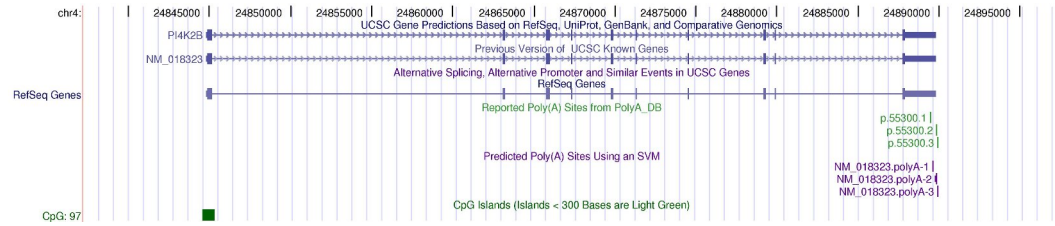


Figure 3.13 Genomic structure of *PI4K2B*. The graphic was produced from the UCSC Genome Browser, using the March 2006 build and the custom tracks option. The physical position on chromosome 4 in base pairs is noted on the top. The tracks show the position of PI4K2B (NM_018323). The thickest parts of the track indicate the coding exon regions within the gene. The slightly thinner portions at the leading and trailing ends of the gene track show the 5' and 3' UTRs. Introns are depicted as lines with arrows indicating the direction of transcription. The PI4K2B genomic region is shown by four tracks; Known Genes Based on Swiss-Prot, TrEMBL, mRNA and RefSef, RefSeq Genes, Mammalian Gene Collection Full ORF mRNAs and Aceview Gene Models with Alt-splicing. The 5'- start of the PI4K2B is shown by the CpG Islands track. The 3'- end of PI4K2B is shown by two tracks: Reported Poly(A) sites from PolyA_DB and Predicted Poly(A) Sites using an SVM.

3.5.2. No evidence for alternative transcripts at rs313548

Analysis of the human genome browser data showed that there was no evidence for alternative transcripts that would include the intronic SNP, rs313548. The investigations carried out on the UCSC genome browser included all gene prediction tracks from human and non-human sources, six human mRNAs from GenBank and 22 human ESTs that covered this region. None showed evidence for rs313548 in a coding region.

The intronic sequence containing rs313548 may be spliced into the exon. This was checked by amplifying and sequencing *PI4K2B* exon 1 to exon 4 and searching for extra sequence, double sequence or sequence mis-match. Primers were designed to amplify exon1 and exon4 of *PI4K2B* in cDNA samples, namely stcPI4K2B.ex1aF and stcPI4K2B.ex4R as listed in Table 2.4. This would detect a sequence if rs313548 was spliced into a transcript. Amplification of cDNA produced a single band of expected size, 756bp in the two cDNA samples tested (47 and 49). Two of these cDNA samples were sequenced with both forward and reverse primers. Figure 3.14 shows analysis of the sequence data, with no evidence of an alternative transcript that would contain rs313548.

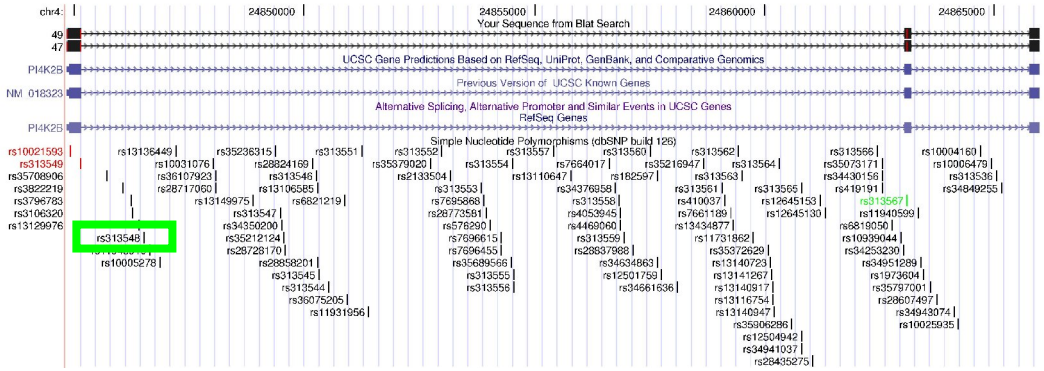


Figure 3.14 No evidence for alternative transcripts in lymphoblastoid cell lines of rs313548. The graphic was produced from the UCSC Genome Browser, using the March 2006 build and the custom tracks option. The physical position on chromosome 4 in base pairs is noted on the top. Two cDNA samples (47 and 49) were amplified and sequenced with forward (F) and reverse (R) primers, stcPI4K2B.ex1aF and stcPI4K2B.ex4R respectively. Their transcript is coloured in black. The thickest parts of the track indicate the coding exon regions within the gene. The slightly thinner portions at the leading and trailing ends of the gene track show the 5' and 3' UTRs. Introns are depicted as lines with arrows indicating the direction of transcription. The position of PI4K2B according to the UCSC genome browser is shown by blue lines. The position of SNPs in the region is shown below. The SNP rs313548 is highlighted with a green square. It is located outside of the exon.

3.5.3. Evidence for alternative splicing at rs313567

There were two sources of evidence for alternative splicing of exon 2 in *PI4K2B*. Firstly, collating information on the UCSC genome browser from human genes, non-human genes, gene-prediction tracks, mRNA tracks and EST tracks showed evidence of alternative splicing. In the non-human RefSeq gene track, exon 2 was shorter in the mouse, rat, frog and fruit fly sequence, although it covered rs313567. Also, of the 25 human ESTs covering exon 2, one human EST, DA744687 (from NT2 cell line, a neuronally committed human teratocarcinoma cell line) does not cover exon 2 and of the 21 non-human mRNAs, two mouse mRNAs, AK106754 and AY148879 do not cover exon 2. Also, of >100 non-human ESTs, several showed transcripts of exon 2 without rs313567. In particular, out of 20 mouse ESTs, five spliced out rs313567: BY710975, BF162501, BI691138, BI691138 and BF162501.

Secondly, evidence for exon 2 alternative splicing came from sequence data. Lymphoblastoid cell line cDNA was amplified using primers to exon 1 and exon 4 *PI4K2B*, stcPI4K2B.ex1bF and stcPI4K2B.ex4R respectively, as described in Table 4. Figure 3.15 illustrates amplification of *PI4K2B* using the Invitrogen and Sigma methods. Using the Invitrogen system, a band of approximately 750bp was detected, which was most likely derived from the most common *PI4K2B* isoform at the expected size of 756bp. There were two different sized products in the two different lymphoblastoid cell line samples amplified using the Sigma system, sample 47 with a band at ~400bp and 49 at ~520bp.

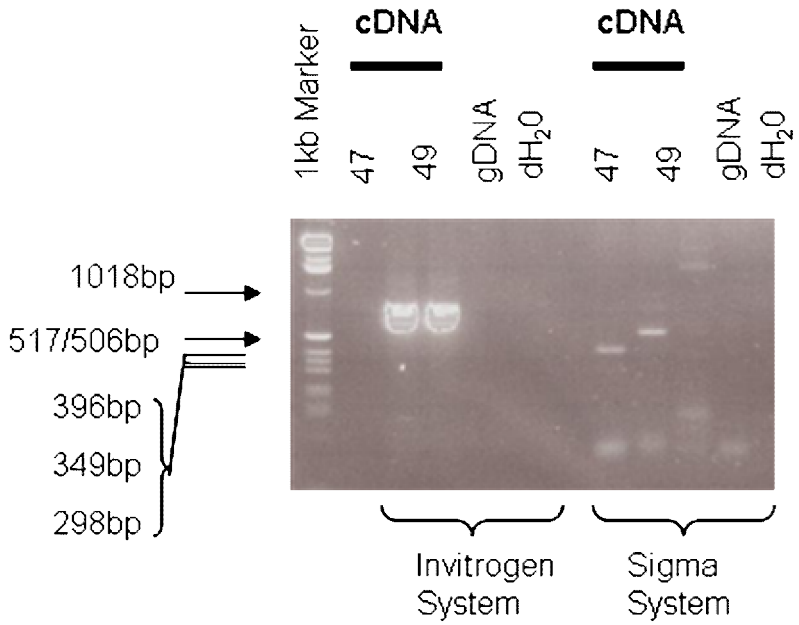


Figure 3.15 Alternative isoforms at *PI4K2B* exon2. The picture shows the amplification of *PI4K2B* exon2 from lymphoblastoid cell line cDNA samples, 47 and 49. gDNA from lymphoblastoid cell line sample, 29 and distilled water were used as controls. Both the Invitrogen and Sigma methods were used for PCR amplification. The bands at ~750bp amplified with the Invitrogen system were the expected size. The bands amplified by the Sigma system at ~400bp and ~500bp were not expected.

All lymphoblastoid cell line cDNA samples were amplified using primers to exon 1 and exon 4 *PI4K2B*, stcPI4K2B.ex1bF and stcPI4K2B.ex4R respectively (as Table 4) with Invitrogen, Sigma and Applied Biosystems methods. In all cDNA samples, a band of ~750bp was detected using the Invitrogen system. Sequence analysis of the ~750bp product from the Invitrogen system showed the expected coverage of exons one to four of *PI4K2B*.

Products of various sizes (350bp, 400bp and 520bp) were obtained with the Sigma and Applied Biosystems amplification methods. PCR products were extracted from the agarose gel and sequenced. Figure 3.16 illustrates the sequence obtained from four cDNA samples, using stcPI4K2B.ex4R primer. There were three different alternative exons that did not contain the SNP rs313567. For sample 49, the full 155bp of exon 2 was covered. The other three isoforms amplified were shorter: sample 47 (99bp) sample 70 (31bp) and sample 133 (78bp).

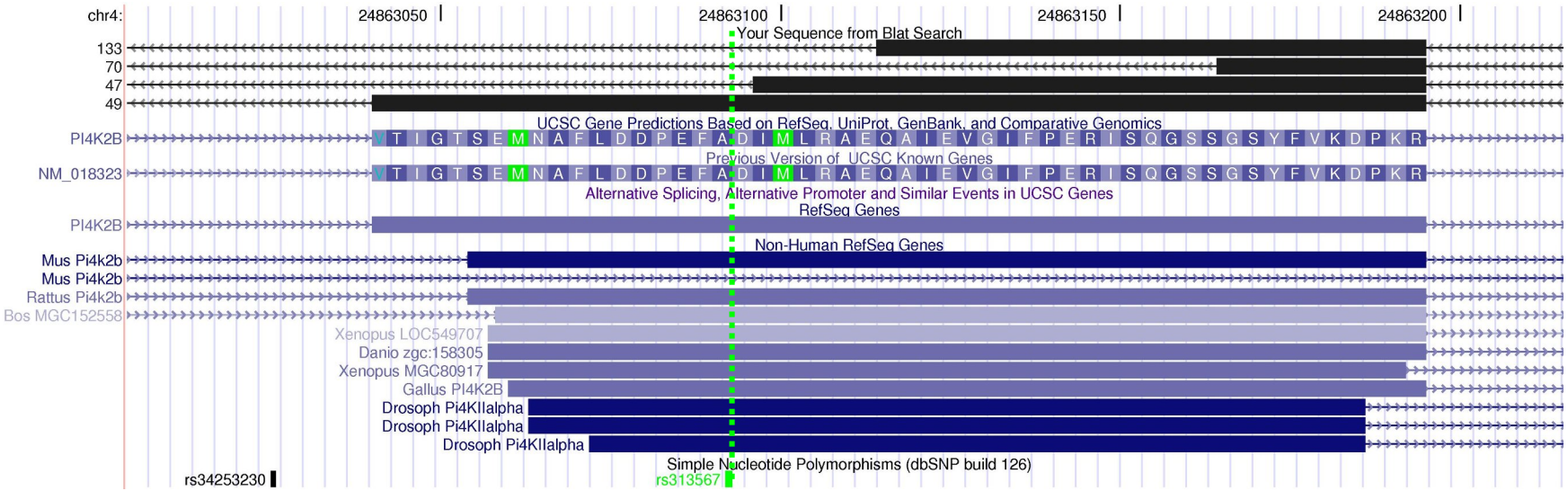


Figure 3.16 Alternative splicing of exon 2 of *PI4K2B.* The graphic was produced from the UCSC Genome Browser, using the March 2006 build and the custom tracks option. The physical position on chromosome 4 in base pairs is noted on the top. The four black lines were sequence data from amplified products from three lymphoblastoid cell lines: 47, 49, 70 and 133. The thickest parts of the line indicate the coding exon regions within the gene. Introns are depicted as lines with arrows indicating the direction of transcription. The position of PI4K2B according to the UCSC gene prediction tracks are shown by blue lines. The position of SNPs in the region is shown below. The SNP used in the allele-specific expression studies is in green, rs313567. A green dotted horizontal line shows the SNP is not covered by the exons of sample 133, 70 and 47 but it is covered by sample 49.

The products amplified from samples 47 and 49 by the Sigma system, as shown in Figure 3.15, suggest they are different splice isoforms due to their different size (400bp and 520bp respectively). Upon analysis of sequence data, there was evidence for exon 2 splice forms for 47 and not for 49. However, the difference in size, from the expected 750bp in sample 49, was due to splicing of *PI4K2B* exon 1.

Pertinent to this study, was that the SNP, rs313567 used in *PI4K2B* allelic expression assays, was spliced out of exon 2 in some isoforms, as illustrated in Figure 3.17. Identification of alternative transcribed transcripts of *PI4K2B* explains why the allelic imbalance data was not consistent across three SNPs on the same haplotype. In addition, alternative splice isoforms were detected in cDNA samples that were used in the Taqman expression assays. Sample 47 was used as a heterozygote for the allelic assay in PI4K2B intron 1 and sample 70 was also heterozygous for another SNP in the 3'UTR region that was used in the haplotype analysis. Finally, it was notable that the splicing of exon2 in *PI4K2B* was in-frame, as examined at the base-level on the UCSC genome browser, and consequently would not affect protein production.

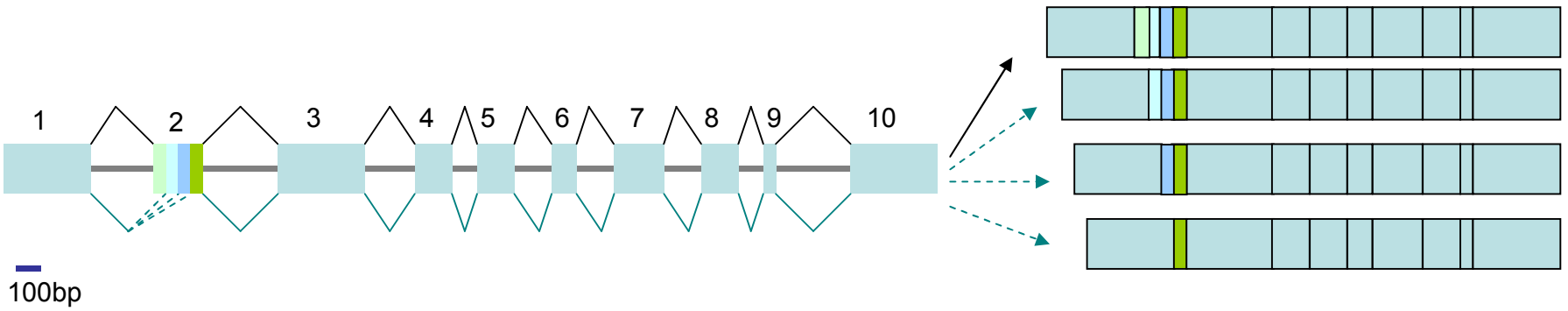


Figure 3.17 Illustration of alternative splicing of exon 2 of PI4K2B. This schematic shows four different PI4K2B isoforms, one with all of exon 2 and three others with exon 2 spliced out to different degrees as shown in Figure 3.16.

3.5.4. No evidence for alternative transcripts at the 3'end

The search of alternative transcripts at the 3' end of PI4K2B incorporated two methods. Firstly, searching the UCSC genome browser for current sequence data and ESTs showed there was no evidence for alternative splicing at the 3' end of PI4K2B. From the predicted gene sequence track on the human genome browser there was no evidence for an alternative 3' end of *PI4K2B* from human data. Also, all five human mRNAs covered the SNP rs6834255. However, the mouse *Pi4k2b* Non-human RefSeq track did not cover rs6834255

Secondly, 3' RACE was used to look for alternative transcripts at the 3' end of *PI4K2B.* Human heart RACE ready cDNA was tested, as described in section 2.2.3.3. *PI4K2B* had previously been shown to be expressed in the heart by Northern Blot analysis (Balla, Tuymetova et al. 2002) and by microarray expression analysis (GNF Expression Atlas 2 Data from U133A and GNF1H Chips, UCSC human genome browser). This validated the use of heart-specific cDNA for this experiment. Figure 3.18 shows the amplification of the 3' end of *PI4K2B* by RACE.

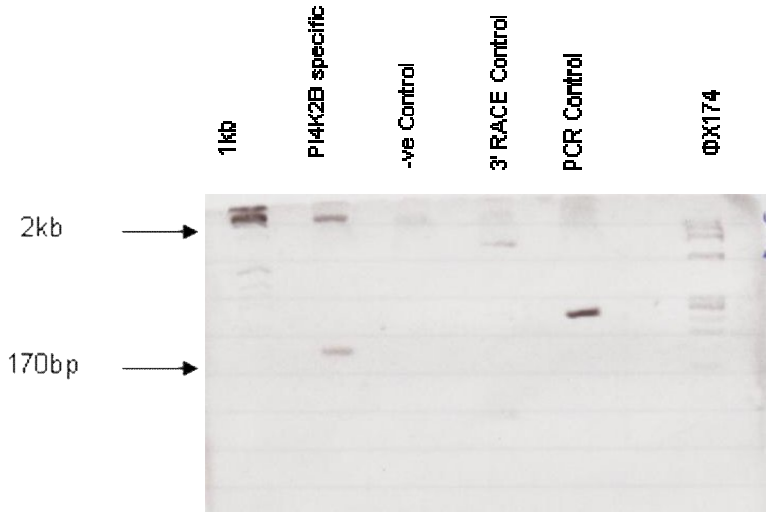


Figure 3.18 3'RACE amplification. Human heart RACE ready cDNA was amplified using a PI4K2B specific assay (stcPI4K2B.ex10_inner and 3' RACE inner), 3' RACE control primers (3' RACE outer and 3' RACE control) and PCR control primers (5' RACE Inner and 5' PCR Control). The products with the *PI4K2B* specific primers were bands at 2kb and 170bp. There was no clear band in the negative control after amplication with the PI4K2B specific primers. The expected products for the 3' RACE control and PCR control were amplified, 733bp and 236bp respectively.

Both the 2kb and 200bp products from the PI4K2B specific primers were gel extracted and sequenced. The following primers were used to amplifiy the products; stcPI4K2B.exon10_outer, stcPI4K2B.ex10_inner and End of 3' 1kb sequence as listed in Table 2.4. The 2kb band was successfully sequenced and matched the known 3'-sequence of PI4K2B as shown by the black bar in Figure 3.19. Importantly, the sequence coverage included the SNP rs6834255, as highlighted in yellow in Figure 3.19. Reliable sequence was not obtained from the 200bp product. Thus it was likely a spurious result and was not considered further.

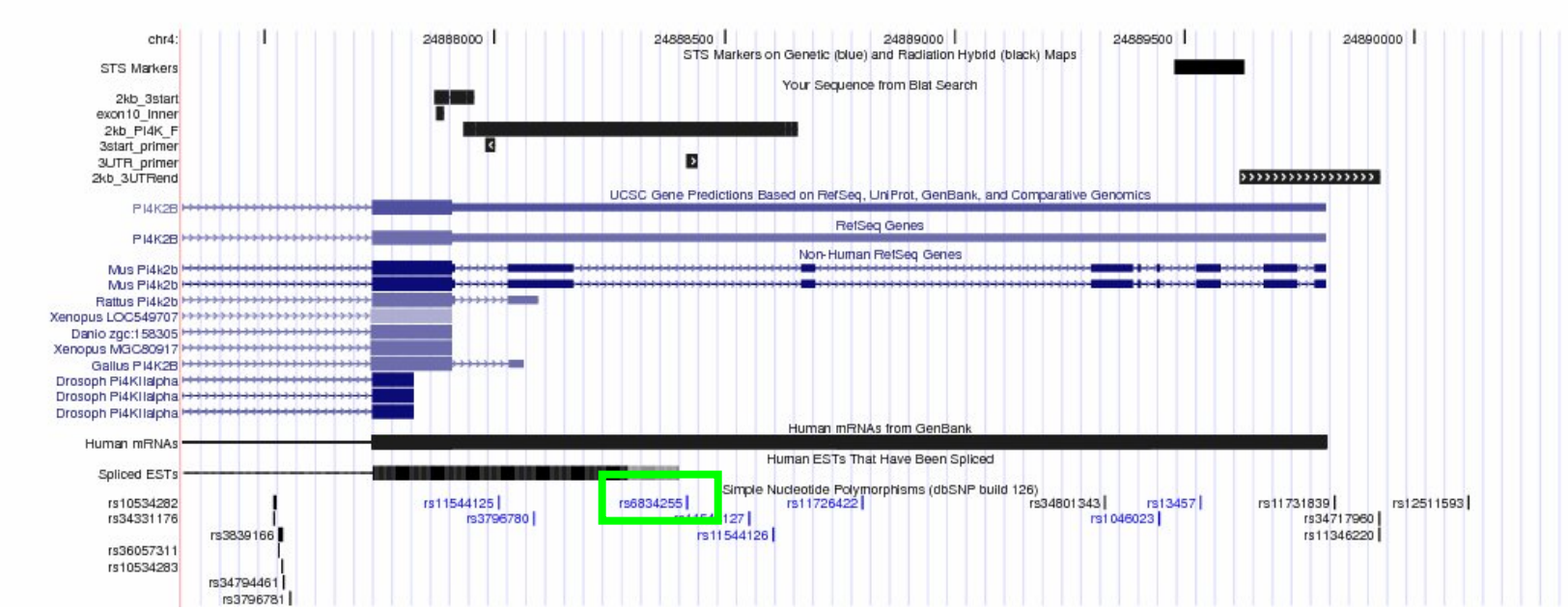


Figure 3.19 No evidence for alternative splicing of PI4K2B at 3'end. The graphic was produced from the UCSC Genome Browser, using the March 2006 build and the custom tracks option. The physical position on chromosome 4 in base pairs is noted on the top. Under “Your Sequence from Blat Search” is the sequence obtained from 3’ RACE amplification of PI4K2B, shown by three black boxes: 2kb_3start, 2kb_PI4K_F and 2kb_3UTRend. The primers used for sequencing are also shown by smaller black boxes; exon10_inner, 3start_primer and 3’UTRend respectively. The position of PI4K2B according to the UCSC gene prediction tracks for human and non-human are shown by blue lines. They show exon 10 of PI4K2B as the thick blue line, which is extended by a slightly thinner blue line to show the 3’UTR end. Introns are depicted as lines with arrows indicating the direction of transcription. The position of SNPs in the region is shown below. The SNP rs6834255 is highlighted by a green box.

3.6. *Protein Expression*

The aim of this study was to determine if there was a difference in PI4K2B protein expression, between those individuals with the chromosome 4p15-p16 linked haplotype and those without, in protein lysate samples from lymphoblastoid cell lines. A difference in protein expression could indicate abnormalities in protein translation or protein modification, due to altered translational or posttranslational processing, which could consequently affect signalling pathways or synaptic vesicle trafficking. First, the choice of PI4K2B antibody to detect protein expression and the quality-control measures to determine the specificity of the antibody are described. Second, the optimisation of the protein detection and quantification techniques is detailed. Last, comparison of PI4K2B protein expression in samples with and without the chromosome 4p15-p16 linked haplotype, are reported.

3.6.1. PI4K2B antibodies

There were three antibodies available that were reported to be specific for PI4K2B detection; one raised against the entire protein (Minogue, Anderson et al. 2001), one specific to the N-terminus and one specific to the C-terminus. Table 3.4 lists the specificity and source of the three antibodies.

ANTIBODY	SPECIFICITY	SOURCE
PI4K2B	raised to *E. coli* expressed GST-PI4K2B, cross-adsorbed against GST	Shane Minogue, UCL
PI4K2B N-Terminal	1-17 αα	Abgent
PI4K2B C-Terminal	319-336 αα	Abcam

Table 3.4 PI4K2B antibodies. There are three PI4K2B specific antibodies available. αα are the amino-acids to epitopes which the antibody is raised, GST is glutathione-S-transferase.

The first PI4K2B antibody listed in Table 3.4, was raised against a recombinant GST fusion protein expressed in *Escherichia coli* (*E.coli*) and will be referred to here as "PI4K2B (Minogue)". Figure 3.20 highlights the peptide epitopes used to raise the antibodies specific to the N-terminal and C-terminal PI4K2B, on a protein alignment of PI4K2B and PI4K2A. PI4K2B is highly homologous (58% identical and 75% homologous) to PI4K2A, apart from a unique N-terminal 100 amino acid sequence. PI4K2B protein (NP_060793) is 481 amino acids in length and encoded on chromosome 4p15.2, whereas the slightly smaller PI4K2A (NP_060895) is 479 amino acids in length and is encoded on chromosome 10q24. The phosphatidylinositol 3- and 4- kinase domain was identified from NCBI conserved domain analysis and spanned amino acids 167-415 in PI4K2B and 172-438 in PI4K2A. The three antibodies were specific for PI4K2B.

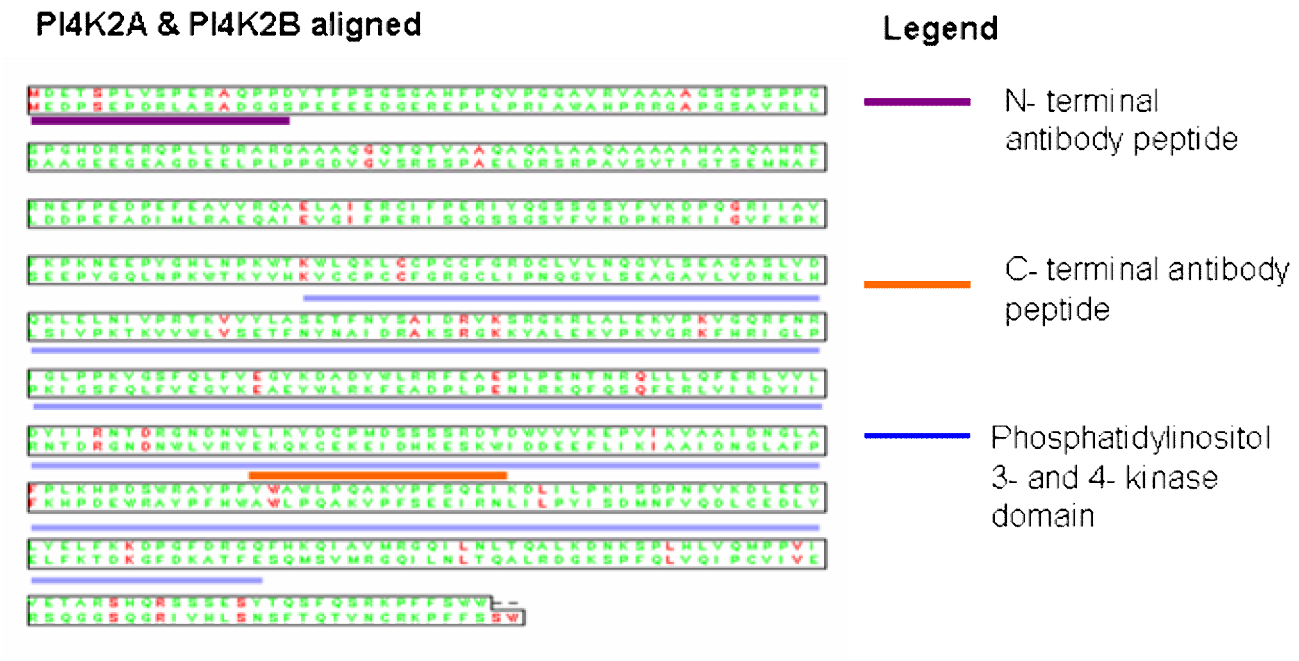


Figure 3.20 PI4K2B Antibodies. The alignment of PI4K2A and PI4K2B is shown. The residues in red are different between the alpha and beta isoform. The epitopes used to raise the antibodies are highlighted in purple (N-terminal) and orange (C-terminal). The phosphatidylinositol 3- and 4- kinase domain is shown in light blue. The alignment was performed using Clustal W (1.82) and the diagram was made using “prettyplot” (Ian Longden, http://bioweb.pasteur.fr/docs/EMBOSS/prettyplot.html)

3.6.2. Quality control of PI4K2B antibodies

The specificity of the three antibodies to PI4K2B was tested using two different positive control protein lysates. One positive control sample was prepared in this study as described in section 2.2.2.5, by transforming the N-terminal *PI4K2B* specific HA-epitope tagged DNA (Balla, Tuymetova et al. 2002), into DH5α competent cells, purifying the plasmid, which was then chemically transfected into COS-7 cells. This positive control was named "Tfn PI4K2B" and a negative control for the transfection procedure was named "Tfn No DNA". A second control sample was provided for this study and consisted of human HT1080 cells stably expressing GFP-PI4K2B (Minogue, Anderson et al. 2001). This control was named "+ve Ctrl". Detection of PI4K2B in two lymphoblastoid cell line protein lysates was also tested. The protein lysates were prepared as section 2.2.9. The gel electrophoresis and immunoblotting of the samples was performed as section 2.2.10. Figure 3.21 shows three replicate immunoblots of the above described control samples and lymphoblastoid cell line samples. Each immunoblot was detected with one of the three PI4K2B antibodies as listed in Table 3.4, according to the conditions detailed in section 2.2.10.5.

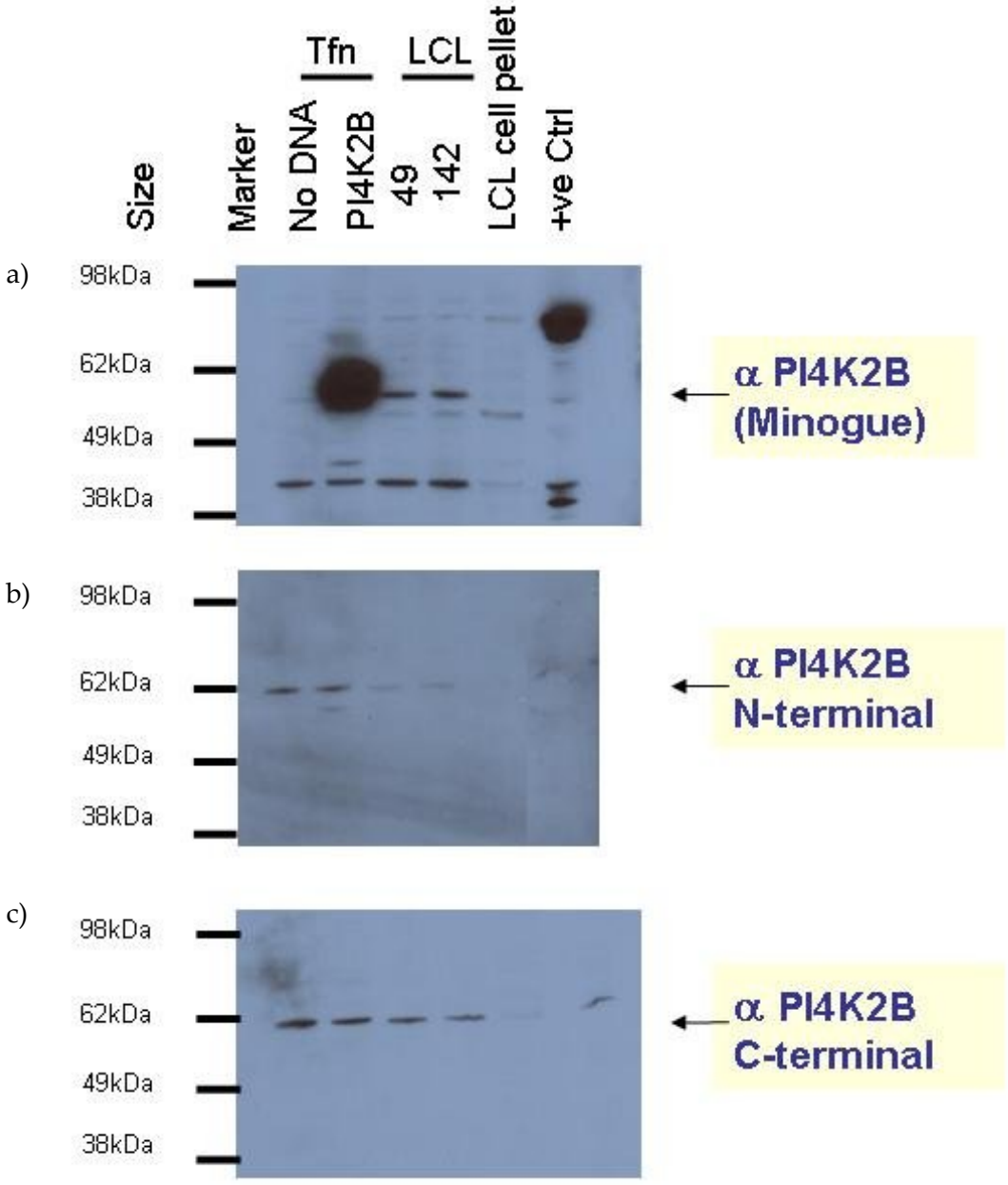


Figure 3.21 Quality control of PI4K2B antibodies by Western blot analysis. Three replicate gels were processed by gel electrophoresis, followed by immunoblotting. Each membrane was probed with one of the three antibodies to PI4K2B: α PI4K2B Minogue (a), α PI4K2B N-terminal (b), α PI4K2B C-terminal (c). There are two positive controls in each panel; i) protein extract from COS-7 cells that were transfected with PI4K2B-HA plasmid which was specific to N-terminal PI4K2B, labelled "Tfn PI4K2B" and a related negative transfection control labelled "Tfn No DNA" and ii) total cell lysate from human HT1080 cells stably expressing GFP-PI4K2B, labelled "+ve Ctrl". There are two protein lysates from lymphoblastoid cell lines (LCL), labelled "LCL 49" and "LCL 142" and one protein lysate prepared from an "LCL cell pellet".

A band size of ~55kDa was expected for PI4K2B. Detection by the PI4K2B (Minogue) antibody of the first positive control, shown in Figure 3.21a, produced a band of greater intensity in the "PI4K2B" transfected sample compared to the "no DNA" transfected sample. This was the expected result, of increased PI4K2B expression from the PI4K2B transfection control sample. However, a differential signal was not detected between the control and the transfected samples, by antibodies specific to the N-terminal or C-terminal domain of PI4K2B, in Figure 3.21 b and c. However, a slight band was detected by the N-terminal antibody in the "PI4K2B" transfected sample ~50kDa, in Figure 3.21b panel 2.

A band at 60kDa was also detected in two lymphoblastoid cell line protein lysates by the three antibodies to PI4K2B, but was barely detected in the lymphoblastoid cell line pellet. This shows that the protein lysate preparation extracted PI4K2B from the lymphoblastoid cell line sample, which partitioned to the soluble fraction.

For the second PI4K2B positive control (GFP-PI4K2B), a band of expected size of 79kDa was only detected by the PI4K2B (Minogue) antibody. Weak, likely non-specific, bands were detected at ~65kDa, by the N-terminal and C-terminal antibodies and at ~55kDa, by the PI4K2B (Minogue) antibody in this positive control sample.

The other, likely non-specific, bands detected by the PI4K2B (Minogue) antibody may be explained as follows: the band of ~44kDa detected in the PI4K2B transfected sample may be a breakdown product or as a result of translation initiation from internal ATG; the lower band at ~40kDa detected in all five samples (except the LCL cell pellet sample) may be due to non-specific antibody binding interactions, such as an unreported novel protein from the same protein family, a different splice variant that shares a similar epitope or the PI4K2B antibody epitope may be common to another peptide, of a different molecular weight; the band detected at 80kDa may be a multimer of PI4K2B protein.

The three antibodies were also tested for cross-reactivity. Immunoblots were probed with either the primary or secondary antibodies only and no signal was detected.

The aim of this experiment was to identify an antibody that specifically detected PI4K2B. It is clear that the PI4K2B (Minogue) antibody was favourable in two respects; i) it was the only antibody of the three tested, that efficiently detected transfected PI4K2B ii) it correctly detected the positive control signal, GFP-PI4K2B, at 79kDa from the cells stably expressing it. Thus, this antibody was chosen to quantify PI4K2B expression in lymphoblastoid cell line protein lysates.

3.6.3. PI4K2B protein expression in brain tissue

Although the expression analysis in this study was performed in lymphoblastoid cell line samples, it was important to test for PI4K2B expression in brain tissue because both bipolar disorder and recurrent major depression are disorders of the brain. The expression studies were not conducted in brain tissue as post-mortem brain tissues from the family could not be obtained. Figure 3.22 demonstrates that PI4K2B protein was detected in neuronal tissue, by the PI4K2B (Minogue) antibody. The neuronal tissue samples were a whole brain homogenate sample and a sample prepared from SH-SY5Y cells which are human brain neuroblastoma cells (provided by Jennifer Chubb, Molecular Medicine Centre, University of Edinburgh). A band of ~50kDa was detected. This was 5kDa smaller than the species detected in lymphoblastoid cell line samples and control transfection samples. This molecular weight discrepancy possibly indicates alternate splicing, post-translational modification or post-translational cleavage of PI4K2B. Figure 3.22 also has protein lysates from the transfected control samples, from lymphoblastoid cell lines and from the positive control, with the expected band pattern as previously shown in Figure 3.21. No protein signal, including from the GAPDH the loading control, was detected in the PBL sample, which may have degraded during processing. The PBL sample was included to ascertain whether PI4K2B was expressed in blood cells. Additionally, detection of GAPDH, was used as a loading control and is shown below the PI4K2B panel in Figure 3.22.

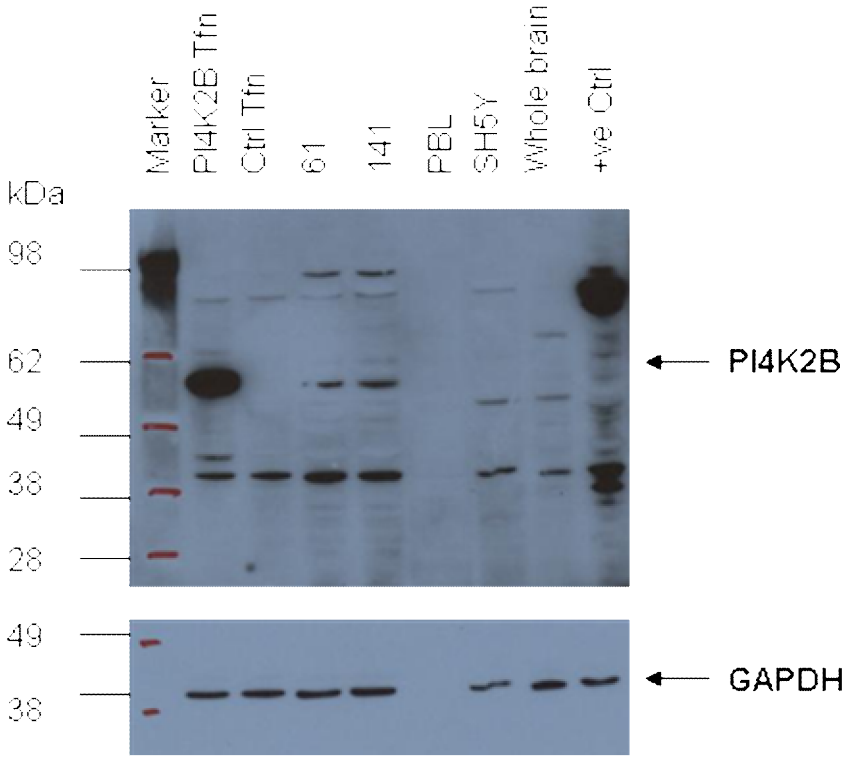


Figure 3.22 PI4K2B detection in brain tissue. Protein lysates prepared from two lymphoblastoid cell lines; 61 and 141, a peripheral blood lymphocyte sample (PBL), SH-SY5Y cells and whole brain homogenate were processed by SDS-PAGE followed by immunoblotting. The positive control is human HT1080 cells, stably expressing GFP-PI4K2B and the transfection samples with and without PI4K2B plasmid DNA are also shown for band size comparison were also processed. The immunoblot was probed with the PI4K2B (Minogue) antibody. The loadings were controlled by probing with the GAPDH antibody.

The detection of PI4K2B protein expression in brain tissue in Figure 3.22 was contrary to other reports. First, a report showed that PI4K2B was not present in the brain (Guo, Wenk et al. 2003) and that PI4K2A accounted for the majority of PI4-kinase activity in the brain. This conflicting result may be due, however, to underexposure of Western Blot pictures in the original report. Second, evidence for PI4K2B protein expression in the brain was not detected by bioinformatic analysis performed in this study using the Allen Brain Atlas (Lein, Hawrylycz et al. 2007). However, non-detection of protein expression in the Allen Brain Atlas resource may be due to experimental variability and should not be considered a definitive negative result.

3.6.4. Optimisation of PI4K2B protein detection technique

The aim of optimising the PI4K2B protein detection technique was to achieve a reproducible and quantifiable pattern of protein detection, which would enable quantifiable expression analysis in protein lysate preparations from lymphoblastoid cell lines. Adjustments were made to every step of the Western blotting procedure as listed in Table 3.5.

In brief, the amount of protein loaded into each lane of a gel and the concentration of antibody used for detection was optimised. This prevented an excess amount of protein masking any detectable difference or an insufficient amount of protein escaping detection by antibody. A selection of sample buffers, electrophoresis gels and gel-transfer methods were trialled. This improved the appearance of the protein detection signal. Gels with more lanes were used, so that all protein lysate samples were on the same gel. This reduced the problem of gel to gel variation, which occurred when samples were divided into two or more gels. Different blocking buffers and antibody diluents were used to minimise signals from non-specific binding. In addition, more than one loading control was used, for detection of proteins that were constitutively expressed in all tissues at high levels, for example

GAPDH, α-tubulin and β-actin. The use of accurate loading controls was imperative for this study to obtain reliable information on any expression level changes in PI4K2B. In addition, the loading controls checked that the lanes in the gels were evenly loaded with protein lysate sample and that there was even-transfer between the gel and membrane across the whole gel. They were used to quantify the amount of protein in each lane by using the density of the loading control band to correct for differences in loading. Various detection methods were also tested, such as infrared fluorescent detection by the Odyssey Infrared Imaging system that provides quantification accuracy and simultaneous two colour detection (http://www.licor.com/bio/odyssey/). However, this strategy was unsuccessful due to the failure of the secondary antibody [IRDye® 800 Conjugated Affinity Purified Anti-RABBIT IgG (H&L) (GOAT)] to bind specifically to the PI4K2B primary antibody. The final technique for optimised PI4K2B protein detection in lymphoblastoid cell line samples by immunoblotting, is listed in Table 3.5 and described in section 2.2.9.

METHOD	INITIAL	OPTIMISED	OUTCOME
Sample preparation	Sample Buffer (0.5M Tris pH6.8, 20% Glycerol, 20% SDS, H_20, bromophenol blue) Protein Sample (Lysate, 1M DTT (10%), Sample Buffer (2x), H_20) 100°C/5 minutes.	Invitrogen LDS Sample Buffer (141mM Tris base, 2%LDS, 10& Glycerol, 0.51mM EDTA, SERVA Blue, Phenol Red, pH8.5) 1M DTT, Lysate, H_20. 72°C/10 minutes.	Complete lysis, ease of loading
Protein concentration	10-15µg	5µg	Sufficient antigen for detection but not overload
Gel electrophoresis	BioRad	Invitrogen 4-12% 10/17 well Bis-Tris gels	Clearer bands, 17 wells, more reproducible
Well position	According to group	According to group & random	No difference in variability
Gel transfer	BioRad semidry & wet	Invitrogen wet transfer	Even transfer across whole membrane, improved consistency between gels and between lanes within a gel.
Gel staining	Coomassie	Simply Blue	Ensured equal loading of gel resulting in increased sensitivity
Membrane	Invitrogen PVDF & nitrocellulose, Hybond-P PVDF membrane	Amersham PVDF	Increased sensitivity, greater transfer of protein
Blocking agent	5% Marvel in PBS-Tween, 5% fish gelatine in PBS, sodium azide	5% Marvel in PBS	Blocked non-specific binding & reduced background signal
Blocking incubation time	2 hours to 64 hours.	16hours	
Antibody storage	Aliquots at -80°C, -20°C, 4°C	0.01% sodium azide	Prolonged effectiveness of antibody
Antibody dilution	PBS-Tween, 5% Marvel in PBS-Tween	PBS-Tween	Avoided cross-reaction between blocking agent & antibody
Antibody concentration	1:100 – 1:2,500	1:1000	Titrated antibody to optimal concentration
Detection	ECL, Odyssey Scanner	ECL+	Varying the time that the film was exposed to membrane
Densitometry	UVITech, Scion Image, ImageJv1.33	ImageJv1.37	User-friendly

Table 3.5 Alterations to immunoblotting procedure for optimised protein detection. Each aspect of the immunoblotting (Western blotting) technique was optimised. This table is ordered according to the sequence of the Western blotting procedure.

3.6.5. No evidence for a difference in PI4K2B protein expression

PI4K2B protein expression was detected in lymphoblastoid cell lines derived from the large Scottish family as listed in Table 3.1. The most informative lymphoblastoid cell line protein samples were used and fell into two groups; twelve samples with a "linked haplotype" and a psychiatric diagnosis [four with bipolar disorder (29, 47, 57, 61), four with recurrent major depression (36, 45, 48, 63), four that were well (58, 69, 73, 89)] and four samples from individuals that were married-in to the family with no psychiatric diagnosis (11, 13, 39, 49). Every effort was made to ensure consistency between samples: all cell lines were cultured in the same manner, all lysates were prepared by the same protocol and all protein concentrations were determined at the same time. Figure 3.23a illustrates an example of PI4K2B protein expression detected by PI4K2B (Minogue) antibody, in lymphoblastoid cell line samples. This procedure was repeated five times under optimised conditions as Table 3.5.

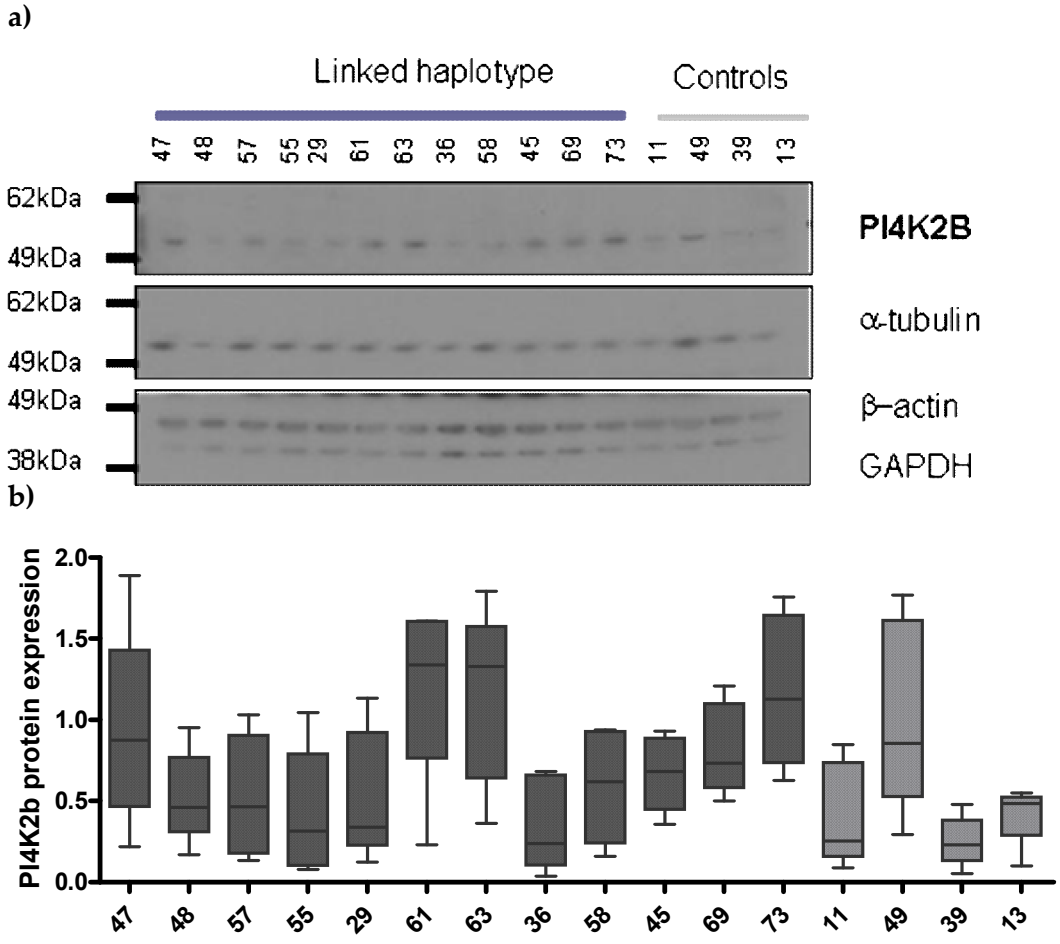


Figure 3.23 No evidence for a PI4K2B protein expression difference between linked haplotype carriers and controls. PI4K2B protein expression (a) was detected at 55kDa in 5µg protein preparations from lymphoblastoid cell lines (Minogue, Anderson et al. 2001). Detection of α-tubulin, β-actin and GAPDH at 50kDa, 42kDa and 40kDa respectively were used as a loading control. Protein expression levels were measured by densitometry (Abramoff 2004). PI4K2B expression level was normalised against the mean expression level from the protein loading controls. The box and whiskers plot (b) shows the five-number summary; lower quartile, median, upper quartile, and largest observation of five replicate gels. The dark grey bars represent individuals with a linked haplotype and light grey bars represent married-in control individuals.

Visual inspection of the immunoblots did not suggest a quantitative difference in PI4K2B expression between the two groups (linked haplotype and control). Therefore, densitometry analysis was performed on the replicate gels for more sensitive quantification of protein expression. This technique measured the intensity of each band representing PI4K2B expression from each sample on each of the five gels. PI4K2B expression was compared to the expression of three loading controls: α-tubulin, GAPDH and β-actin. The expression levels from the three housekeeping genes were correlated within each replicate gel and only used where the correlation value was r2>0.9.Figure 3.23 b shows the result of PI4K2B expression corrected for loading, for each sample from the five replicate gels by a box and whiskers plot. The box-and-whisker plot was a convenient way of graphically depicting groups of numerical data through their five-number summaries; the smallest observation, lower quartile, median, upper quartile, and largest observation.

Further statistical analyses were performed on this data. A student T-test, with a two-tailed distribution, was used to compare between groups within gels, one-way ANOVA (Analysis of Variance) were used to compare between three groups (linked haplotype and ill, linked haplotype and well, and controls) within gels and two way ANOVA were used to compare between gels. The statistical equations are detailed in Appendix B. There was no evidence to suggest a significant difference between groups of linked haplotype and controls on five replicate gels (T test; P=0.51, P=0.33, P=0.16, P=0.33, P=0.79). In addition, there was no evidence to suggest a difference between linked haplotype and ill, linked haplotype and well, and controls groups (One way ANOVA; P=0.83, P=0.64, P=0.07, P=0.69, P=0.78). Combining the five replicate gels, there was no evidence to suggest a difference within the three groups from the five gels (Two way ANOVA; P=0.155).

Another method of quantifying protein expression was to consider only PI4K2B expression in the five replicate gels. This avoided using loading controls which may be unreliable as some physiological factors can alter expression of these proteins in

certain cell types. It also allowed for experimental error that may have occurred during the immunoblotting procedure. From the five replicate gels, detection of PI4K2B from three gels (gel 2, 4 and 5) correlated well, with a Pearson's correlation coefficient, $r^2>0.69$. Using these three sets of data, the PI4K2B expression of each sample was ranked and the position of the rank for each sample in the three gels was noted. This avoided using the absolute PI4K2B expression level that may not be consistent due to inter-experimental variability. Figure 3.24 shows the rank of PI4K2B within each gel for each lymphoblastoid sample. It is evident that the rank of each sample correlated well between the three gels. However, there was no difference in PI4K2B expression between the three groups of linked haplotype and ill, linked haplotype and well and control individuals.

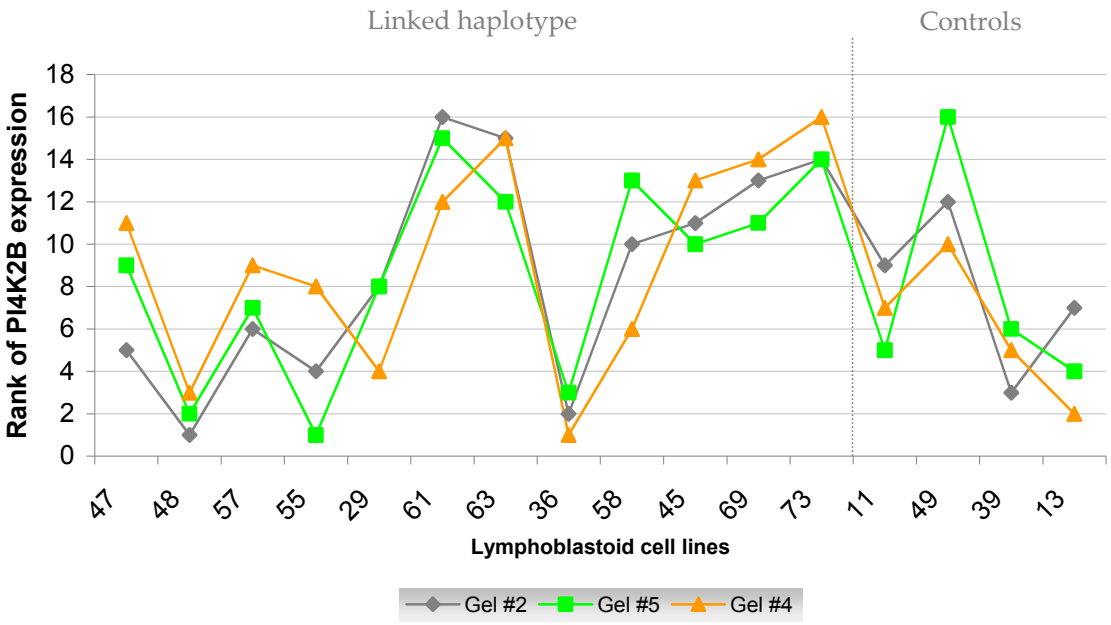


Figure 3.24 Rank analysis of PI4K2B protein gels. Expression of PI4K2B was measured in five replicate gels. The five replicate gels were correlated. Gel 2, 4 and 5 correlated well, r^2 >0.69. For each sample, the rank of PI4K2B expression in each gel was measured and compared between each gel.

3.7. *Discussion*

This study of PI4K2B expression levels has shown that there was no evidence to suggest a difference in *PI4K2B* expression in lymphoblastoid cell line samples with the "linked haplotype" and those without. This was demonstrated at the allele-specific level by Taqman assays and at the protein level by quantitative immunoblotting techniques. One interesting finding from the study was the detection of *PI4K2B* alternate splice isoforms. Initial bioinformatic analysis at the start of this project showed no evidence for alternative splicing of PI4K2B. Over that short time period, substantially more information became available on the UCSC genome browser, some of which suggested *PI4K2B* alternative splicing. Additionally, PCR amplification techniques confirmed PI4K2B splice isoforms that varied in whether they included the exon 2 SNP.

The caveats pertinent to this expression study were primarily technical. Despite every effort to minimise this by rigorous standard protocols, variability may have been introduced during cell culture, RNA preparation and cDNA extraction. As one of the SNP markers was intronic (rs313548) it may have been preferential to purify nuclei to enrich for intronic RNA and then prepare cDNA, as performed in a recent study (Gimelbrant, Hutchinson et al. 2007).

Another source of error may be the technique used for allele-specific expression. While many researchers use this technique for sensitive detection of allelic imbalance, this study has found that many standard curves of gDNA must be constructed to be confident of allelic imbalance. One published study only generated one standard curve and determined their DNA concentration in a less sensitive manner, by the spectrophotometer method. (Maxwell Lee, Personal Communication, 26th April 2005, National Cancer Institute, USA) (Lo, Wang et al. 2003). This study showed that multiple standard curves must be generated and shown to be sufficiently similar before the standard curve can be deemed to be

reliable. Accurate DNA dilutions are a key factor. The analytical approach in this study, to compare heterozygous cDNA samples to gDNA samples, was supported by the recent study of the schizophrenia candidate gene, *dystrobrevin binding protein 1 (DTNBP1).* Specific susceptibility variants were not defined, but a risk haplotype tagged one or more variants that resulted in a relative reduction in *DTNBP1* mRNA expression in human cerebral cortex, conferring susceptibility to schizophrenia (Bray, Preece et al. 2005).

Quantification of protein expression was also a technique not amenable for sensitive detection. Variability may be introduced at many levels of the immunoblotting procedure and so the technique can only be confident of detecting large translational effects. The use of loading controls for protein expression was also uncertain. In proteomics experiments, the widely-used loading control antibody, GAPDH, bound not only to itself, GAPDH, but to three other non-related proteins (Dolores Cahill, Personal Communication, 14th May 2007, 2nd Paris Workshop on Molecular and Statistical Genomic Epidemiology), casting doubts on its specificity.

An important caveat to this study was the use of lymphoblastoid cell lines. Although they supply large amounts of DNA, RNA and protein, it is questionable whether they effectively model pathogenic processes in the brain (Gladkevich, Kauffman et al. 2004) (Tsuang, Nossova et al. 2005). Studies have shown that whole blood does share significant gene expression similarities with multiple central nervous system tissues (Sullivan, Fan et al. 2006). There is also evidence to suggest that gene expression profiles and functional effects in blood lymphocytes do correlate with psychiatric illness (Gladkevich, Kauffman et al. 2004; Iwamoto, Kakiuchi et al. 2004; Vawter, Ferran et al. 2004; Tsuang, Nossova et al. 2005; Marazziti, Dell'Osso et al. 2006)

Other drawbacks of lymphoblastoid cell lines are that prolonged culturing of cells is known to alter gene expression profiles, while establishment of lymphoblastoid cell

lines and culture passage can lead to the appearance of structural genomic variation (Simon-Sanchez, Scholz et al. 2007). In theory, these disadvantages could be overcome by validating assays in lymphoblasts from fresh blood, in different accessible tissue types, such as adipose tissue or in extracts from the same blood samples that were used in the immortalisation to lymphoblastoid cell lines. However, certainty in the accuracy of allele-specific gene expression would be difficult to obtain due to the extreme context-specificity of differential allelic expression, as it cannot be assumed that allelic expression is conserved across different tissues, or even across different cell types of the same tissue (Wilkins, Southam et al. 2007).

An important limitation to this study was the small number of lymphoblastoid samples that were heterozygote for two of the three *PI4K2B* SNPs. This reduced the ability of the study to detect expression differences. Future studies should choose SNPs based on the maximum number of heterozygotes available or use another allele measuring method such as Real-time RT PCR or microarray analysis.

This study was not a complete survey of PI4K2B alternative transcripts, as it was only concerned with the markers involved in the expression assay. Further investigations may yield evidence of additional PI4K2B transcripts. This may be a potentially interesting approach to locate the bipolar disorder susceptibility causative factor in the future. To date, susceptibility mediated through splicing effects has been reported for bipolar disorder and schizophrenia. One example is in the candidate gene, *NCAM1,* where two SNPs have shown a *cis*-effect associated with bipolar disorder. One of the SNPs is within a cluster of alternative exons that shows association to decreased expression of secreted splice variants in post-mortem brain tissue (Atz, Rollins et al. 2007). Another candidate gene, *ErbB4* has a reported *cis*-effect in schizophrenia for one SNP in intron 12 and SNPs near exon 3 that are linked with splicing of exons 16 and 26 respectively, and leads to increased

use of exons 16 and 26 and underlies the genetic association of the gene to schizophrenia (Law, Kleinman et al. 2007).

Although there is no difference detected at the allele and protein level for PI4K2B, it cannot be implied that there would be no difference at the activity level. A case in point is the *MDR1* gene product, where a multidrug resistance synonymous SNP does not change the coding sequence, but results in altered drug and inhibitor interactions. In this case, there were normal mRNA and protein levels, but altered conformations which affected the co-translational folding and insertion of P-gp into the membrane, thereby altering the structure of substrate and inhibitor interaction sites (Kimchi-Sarfaty, Oh et al. 2007). Another example are haplotypes of synonymous SNPs in the human Catechol-*O*-Methyltransferase (COMT) that have a regulatory function, stabilising RNA transcripts and decreasing the amount of translated protein (Nackley, Shabalina et al. 2006). A third example are synonymous mutations in the human dopamine receptor D2 (DRD2) that affect mRNA stability and synthesis of the receptor (Duan, Wainwright et al. 2003). However, this level of investigation was not a priority for *PI4K2B* as: "The most striking thing about *[PI4K2] beta* is that it seems to be virtually inactive enzymatically compared to the alpha isoform." (Shane Minogue, Personal Communication, 31st January 2006, University College London)

The hypothesis of this study was that there was PI4K2B expression differences between chromosome 4p15-p16 linked haplotype carriers and controls that would be a cause or consequence of bipolar disorder and recurrent major depression. However, the techniques utilised in this expression study did not detect a difference in *PI4K2B* expression levels between those with and without the linked haplotype.

Chapter 4
PI4K2B Association Study

4. PI4K2B Association Study

4.1. *Preface*

Before this study, a case-control association study was performed on 362 bipolar disorder cases, 383 schizophrenia cases and 444 controls from the Scottish population, using 408 haplotype tagging SNPs from chromosome 4p15-p16. This association study identified SNP rs10939038, as a potentially important variant in schizophrenia cases (allele *P*=0.006, odds ratio (OR)=1.314, 95%CI:1.08-1.59) (Christoforou, Le Hellard et al. 2007). This result was of interest as the marker, rs10939038 was located in the same linkage disequilibrium (LD) block as the gene *PI4K2B*, 120kb upstream from the coding region. In addition, *PI4K2B* was a worthy functional candidate gene based on pharmacological theory. *PI4K2B* is implicated in the therapeutic effect of the bipolar disorder drug, lithium, as illustrated in Figure 1.7. Furthermore, there was prior statistical evidence of linkage and association of phosphoinositol genes to bipolar disorder, as discussed in section 1.8.2. Due to the strength of the positional and functional evidence for *PI4K2B*, further study was warranted. Other benefits for re-examining the *PI4K2B* genomic region were the increase in the number of variants available on HapMap and the increase in the information on the human genome sequence, since the original study.

The premise of this study was that rs10939038 was in LD with one or more *PI4K2B* functional variants that influence susceptibility to disease, justifying further detailed association analysis of the region, on the same sample set as the original study. The cohort comprised of 362 bipolar disorder cases, 383 schizophrenia cases and 444 controls and was described in full in section 2.1.1. In brief, diagnoses were made according to DSM-IV criteria based on case note review and personal interview using SADS-L. Final diagnoses were reached by consensus between two experienced psychiatrists. Control subjects were drawn from the same population in

South East and South Central Scotland. Genomic DNA was extracted from venous blood samples using standard protocols.

This association study is described below: by first estimating the power of the study, then describing the selection of SNP markers to tag predicted haplotypes, presenting the results of association analysis at the allele, genotype and haplotype levels and finally by determining the validity of those results using permutation analysis.

4.2. *Results*

4.2.1. Power

It is wise to estimate the power of a study before embarking upon what might be a valueless exercise, to determine if a potential benefit can be achieved from the study. The power of this study was calculated by estimating the probability of rejecting the null hypothesis when the specified alternative hypothesis is true. Here, power was investigated as the chance of observing significant association when there is true LD between a disease locus and a SNP marker (alternative hypothesis). The calculation of power depends on the threshold (significance level), the underlying model assumptions, the specification of the alternative, the test statistic and the sample size.

Genetic power calculations were performed on the Genetic Power Calculator (Purcell, Cherny et al. 2003), using the "case-control for discrete traits program." Table 4.1 shows the parameters of the tests for multiplicative (allele counts), dominant and recessive models. Power was calculated in two respects i) the level of power attainable, given the parameters of the study as Table 4.1 and ii) the sample size required for 80% power.

PARAMETER	VALUE		
	Multiplicative	Dominant	Recessive
Frequency of the high-risk allele	0.1	0.1	0.1
Prevalence of disease	0.01	0.01	0.01
Genotype-relative risks of Aa	2	2	1
Genotype-relative risks of AA	4	2	2
LD (D') between tested marker and disease allele	1	1	1
Marker allele frequency*	0.1	0.1	0.1
Number of cases	362	362	362
Control:case ratio	1.227	1.227	1.227
Accepted type I error rate	0.05	0.05	0.05
Power to detect a true effect	0.8	0.8	0.8

Table 4.1 Parameters for estimation of power in the case-control association study. This table lists the case-control parameters used in the power calculations for an association study with 362 cases and 444 controls. The "Genetic Power Calculator" (Purcell, Cherny et al. 2003) was used to determine power. *As generally assumed, the disease allele equals the marker allele.

The parameters chosen for power calculations were assumptions based on population data. It was assumed that the high risk allele frequency for bipolar disorder or schizophrenia were typically rare, with a frequency below 10%. The lifetime prevalence in the general population for bipolar disorder is 0.5-1.5% (Smith and Weissmann 1992) and the lifetime morbid risk in the general population for schizophrenia is 1% (Gottesman 1991), so a value of 1% was used. To specify power at the test locus, the LD measure was D'=1 and the frequencies of the disease allele and the marker allele were equal. This combination, D'=1 and equal allele frequencies between the disease and the marker allele, implies an $r^2=1$, ensuring that the power to search for the actual disease variant, or any marker in complete LD with the disease variant, was calculated.

Table 4.2 shows the results of the power calculations. The available power was estimated for various levels of significance that may be required. These power calculations incorporated the sample size in question here and other sample sizes at

specified alpha levels (α=0.05). The test also assumes that the marker tested is the true disease variant.

Multiplicative Model		
Alpha	Power	N cases for 80% power
0.1	0.9992	97
0.05	0.9978	123
0.01	0.9873	183
0.001	0.9358	267
0.05	0.9978	123
Dominant Model		
Alpha	Power	N cases for 80% power
0.1	0.9926	135
0.05	0.983	171
0.01	0.9338	254
0.001	0.7853	372
0.05	0.983	171
Recessive Model		
Alpha	Power	N cases for 80% power
0.1	0.1583	6,450
0.05	0.0906	8,189
0.01	0.02425	12,190
0.001	0.003504	17,820
0.05	0.0906	8,189

Table 4.2 Power estimations for the *PI4K2B* association study. The power calculations for this association study were performed using the Genetic Power Calculator (Purcell, Cherny et al. 2003). Three models were tested. Alpha is the required significance level. N is the number of cases required for 80% power.

The power calculations showed that this study had sufficient cases of bipolar disorder (362) and schizophrenia (383) to reach a significance level α=0.01, with >93% power for both dominant and multiplicative models but not for a recessive model. There was only 9% power to detect a recessive model at a significance level α=0.05. An extremely large cohort of over 8,224 individuals would be required to detect any variants that would follow a recessive model at a significance level α=0.05.

Power calculations are dependent upon the use of correct parameters to model the disease. Variation in the frequency of the high-risk allele can change the power of the study and the sample size required. For example, if the *PI4K2B* variant in question caused 1% of the disease, and the disease itself had a prevalence of 1%, the frequency of the high-risk allele would be 0.01. Using this value for a rare high-risk allele, while maintaining the other assumptions regarding the disease as Table 4.1, power was calculated for the multiplicative model using the Genetic Power Calculator. Unsurprisingly, there was little power (9%) to detect a significant result (α =0.05). Also, a greater number of cases (8,189) would be required to have 80% power to detect a result (α=0.05). Thus, power calculations were useful to investigate the limits of this study, that although there was ample power to detect a relatively frequent variant, this power decreases if a very rare variant is the causative disease variant.

4.2.2. SNP selection

The SNP selection procedure was described in section 2.5.1.1, the final selection of chosen markers are illustrated in Figure 4.1 and listed in Table 4.3.

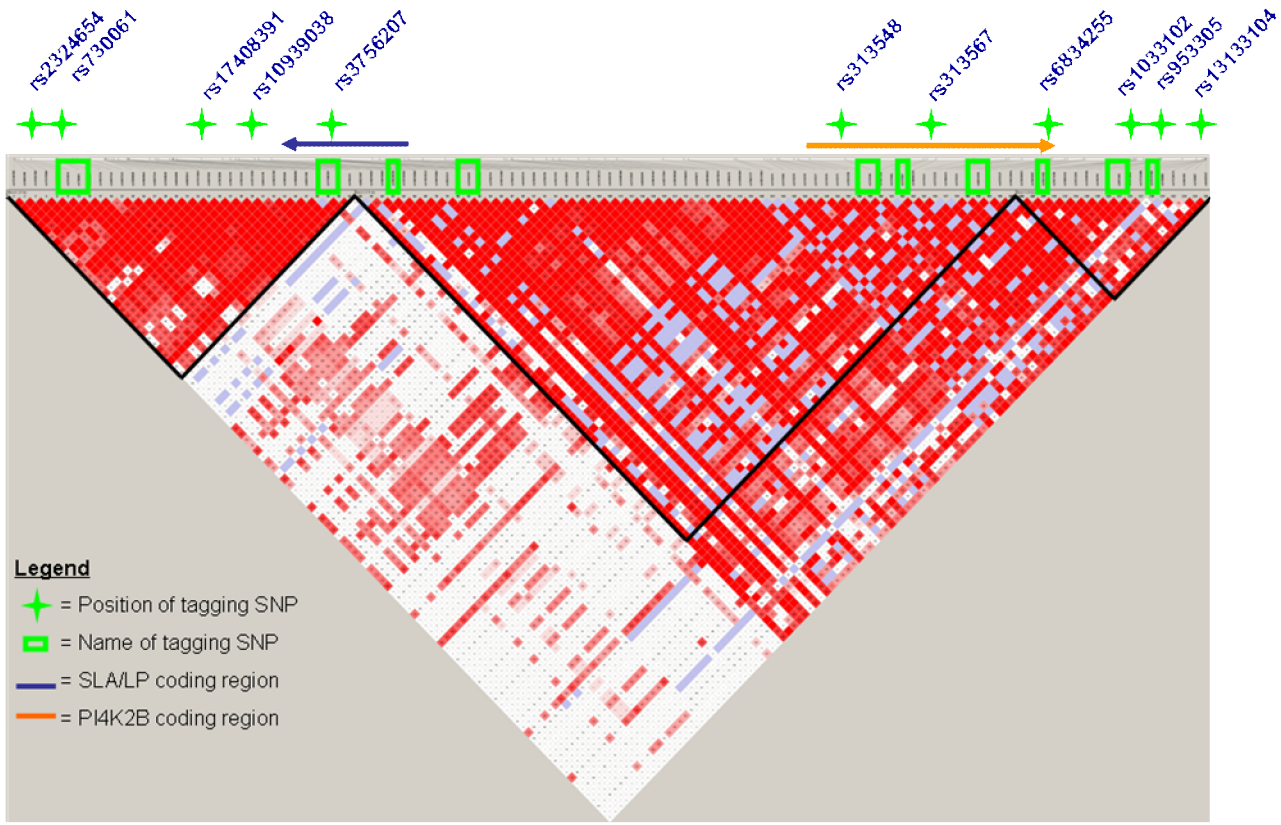


Figure 4.1 Location of tagging SNPs on linkage disequilibrium map of PI4K2B genomic region. LD map of the SNPs in the genomic region of *PI4K2B* on chromosome 4p15.2 from rs13148169 to rs6448347 (24,745,440bp-24,986,687bp). The locations of the tagging SNPs are marked by green stars and their names are marked by the green squares. The blue bar marks the location of *SLA/LP* (24,799,991bp-24,838,254bp) and the orange bar marks the *PI4K2B* [24,911,944bp-24,956,979bp (UCSC May 2004, NCBI build 35)]. The LD map of the region was defined by solid spine of LD (D'>0.8) using HapMap CEU trios downloaded on 7th February 2006, phase II b125. Haplotype blocks were merged at MAD≥0.95. Tagging SNPs were then selected to tag the common haplotypes (>5%) within each haplotype block. The three LD blocks are marked by black lines. The LD map was created using Haploview and shows the pairwise LD statistics for each marker in this region. The colour scheme is the standard Haploview colour scheme: when D'<1, the squares are white for LOD <2 and pink/red for LOD≥ 2, when D'=1, the squares are blue for LOD <2 and bright red with LOD≥2.

The SNPs were chosen from HapMap Data Release 20/Phase II Jan06 (February 2006). This was an updated version compared with HapMap Release 7 (May 2004), which was used in the original study. For this study, the SNPs were selected to cover haplotypes frequencies ≥0.05, than the original study which represented haplotypes of frequencies ≥0.10. Figure 4.1 shows the three LD blocks surrounding *PI4K2B* and the location of the markers in relation to the *PI4K2B* genomic region. The position of eleven SNP markers is marked by stars on the LD plot of the *PI4K2B* region. These markers tag common haplotypes >5% in each of the three haplotype blocks. The three LD blocks in the *PI4K2B* genomic region are evident in Figure 4.1: block one is 37,944bp long (rs13148169-rs759243, 24,745,440-24,783,384bp), block two is 174,202bp (rs6852497-rs10433834, 24,787,656-24,961,858bp) and block three is 24,505bp (rs12649921-rs6448347, 24,962,182-24,986,687bp) covering a total area of 236,651bp, which is ~166kbp upstream and ~29kbp downstream of *PI4K2B* genomic region (24,911,943-24,956,979bp; an area of 45,036bp). There was one other RefSeq gene (NM_153825) in this region, a soluble liver antigen/liver pancreas antigen (*SLA/LP*), also known as SEPSECS [Sep (O-phosphoserine) tRNA:Sec (selenocysteine) tRNA synthase] at 24,799,991bp-24,838,254bp. The function of this gene is to convert O-phosphoseryl-tRNA (Sec) to selenocysteinyl-tRNA (Sec) which is required for selenoprotein biosynthesis. This may be relevant to the phenotype under study, as it has been shown that the main functional sites of selenium in mammals are restricted to specific neurons in the brain (Zhang, Zhou et al. 2007). However, there was little known about this specific gene apart from a proposed role in autoimmune hepatitis (Strassburg and Manns 2002).

Table 4.3 shows the eleven SNPs that tag haplotypes in the *PI4K2B* region.

NUMBER	NAME	PHYSICAL POSITION (BP)	LD BLOCK	NOTE	HWE P-VALUE	MAF
1	rs2324654	24,748,920	1	Illumina SNP	0.5	0.29
2	rs730061	24,754,154	1	Illumina SNP	0.3	0.26
3	rs17408391	24,777,853	1	Taqman SNP	0.52	0.21
4	rs10939038	24,790,933	2	Illumina SNP	0.71	0.47
5	rs3756207	24,809,633	2	Illumina SNP	0.25	0.19
6	rs313548	24,913,660	2	*Taqman SNP	0.57	0.24
7	rs313567	24,930,264	2	*Taqman SNP	0.18	0.17
8	rs6834255	24,955,589	2	*Taqman SNP	0.15	0.17
9	rs1033102	24,971,539	3	Taqman SNP	0.46	0.3
10	rs953305	24,977,611	3	Taqman SNP	0.16	0.42
11	rs13133104	24,986,084	3	Taqman SNP	0.02	0.1

Table 4.3 Marker selection for *PI4K2B* association study. This is a list of the eleven SNPs used in the case-control association study. Their physical position is according to NCBI build 35 of the May 2004 UCSC genome browser. Their position according to the LD blocks relates to Figure 4.1 and the three LD blocks around PI4K2B. The note for each marker gives further information; if the marker was included in the original study "Illumina SNP", a newly chosen "Taqman SNP" or a SNP used in the allele-specific expression assays in chapter 3 " *Taqman SNP". The markers tag all haplotypes >5% frequency in HapMap population. The deviation from Hardy-Weinberg equilibrium (HWE) and the minor allele frequency (MAF) were measured from the Scottish control population (444 individuals).

Seven of these SNPs were selected specifically for this study and the remaining four SNPs were used in the original association study. All of these markers are tagging haplotypes. In Table 4.3, if the marker was also used in the expression study as described in Chapter 3 they are labelled "*Taqman SNP". Otherwise the markers are labelled "Taqman SNP", as genotyping the seven extra markers was performed using Taqman assays, as explained in section 2.5.1.2. The remaining four markers are labelled "Illumina SNP" as they were genotyped by Illumina Inc. (San Diego, CA, USA), using the high-throughput Bead-Array platform technology, for the original study. This genotyping information was included in the association analysis. All markers passed the Haploview recommended Hardy-Weinberg significance threshold (P =0.001).

Table 4.4 shows the LD between the markers as measured in Haploview from the 444 individuals in the control population. Both measures of LD are displayed i) the r^2 values are on upper diagonal of the table and ii) the D' values are on the lower diagonal the table. It is clear that the five markers in LD block 2 are in LD with the three markers in LD block 3, D>0.77.

MARKER	RS2324654	RS730061	RS17408391	RS10939038	RS3756207	RS313548	RS313567	RS6834255	RS1033102	RS953305	RS13133104
rs2324654	-	0.12	0.113	0.095	0.073	0.041	0.059	0.058	0.065	0.014	0.01
rs730061	0.93	-	0.082	0.028	0.042	0.067	0.049	0.049	0.041	0.007	0
rs17408391	1	0.949	-	0.178	0.184	0.175	0.17	0.165	0.231	0.127	0.02
rs10939038	0.447	0.274	0.859	-	0.21	0.189	0.169	0.167	0.279	0.49	0.105
rs3756207	0.857	0.728	0.458	1	-	0.55	0.79	0.787	0.45	0.268	0.026
rs313548	0.563	0.801	0.456	0.824	0.86	-	0.627	0.618	0.66	0.355	0.022
rs313567	0.836	0.855	0.48	0.973	0.969	1	-	1	0.462	0.278	0.023
rs6834255	0.836	0.859	0.473	0.972	0.969	1	1	-	0.457	0.277	0.022
rs1033102	0.599	0.539	0.609	0.859	0.899	0.948	1	1	-	0.563	0.029
rs953305	0.215	0.167	0.588	0.873	0.905	0.903	1	1	0.974	-	0.077
rs13133014	0.192	0.019	0.792	0.885	1	0.778	1	1	0.771	0.948	-

Table 4.4 Linkage disequilibrium indices for association study markers. Above the diagonal are r^2 values and below the diagonal are D' values. The LD was calculated in Haploview from 444 individuals used in the control cohort for this study. The table is divided by dark grey gridlines according to the three LD blocks which were defined by solid spine of LD, D'>0.8 and merged at MAD≥0.95, as described in Figure 1.

4.2.3. Quality control measures

Genotyping error can have an appreciable effect on association results, where random genotyping error can reduce the power of an association study and systematic genotyping error can falsely attribute association (Hattersley and McCarthy 2005). Thus, several measures were undertaken in this study to ensure genotyping quality. Initial tests were performed to determine the quality of the seven Taqman assays for this study. They were checked for Mendelian segregation on 32 proband-patient trios, as described in section 2.1.1.4. . All seven markers segregated correctly in all trios. Each Taqman assay output was also visually inspected for correct differentiation of three clusters: two outer homozygote clusters and one central heterozygote cluster. All but one of the markers clustered into three groups. The genotyping assay for rs13133104 had only two clusters; homozygote CC and heterozygote CT and may be suspicious, however the genotypes segregated in a Mendelian fashion in the 32 trios. After the initial quality control tests, the case-control DNA, which was randomly divided between four plates to avoid bias in genotype calling, were genotyped with seven Taqman SNP assays. The seven SNPs were successfully genotyped at an average locus success rate of 95% (range: 89-99%) in a total of 1212 individuals (93% sample success rate). The genotyping quality of the case-control sample set was checked by two independent people blind to each others results. The rs13133104 assay gave three different clusters when the larger number of samples was genotyped. Furthermore, duplicates were double-checked and one discrepant duplicate was removed and excluded from analysis. Table 4.3 shows that Hardy-Weinberg equilibrium was verified for all markers >0.01 cut-off, as another control measure.

4.2.4. Single allele results

Association analysis was performed using standard χ^2 test, as shown in Figure 1.3. Single marker *P*-values were obtained for all cases against controls and then cases separated by diagnosis against controls. Figure 4.2 shows the single marker *P*-values

for the 11 markers in the PI4K2B region, seven of which were genotyped specifically for this study. Figure 4.2 shows there was no association between these seven SNPs and bipolar disorder or schizophrenia, or both illnesses combined, *P*-values>0.05. The only significant *P*-value (*P*<0.05), was for rs10939038 in the schizophrenia cases. This was a repeat association result of the previously reported finding (Christoforou, Le Hellard et al. 2007). Analysis performed in this study showed that the allelic association to this marker rs10939038 withstood permutation testing with 1,000 simulations in schizophrenia cases (permutated *P*= 0.015, standard error (SE) = 0.0038)

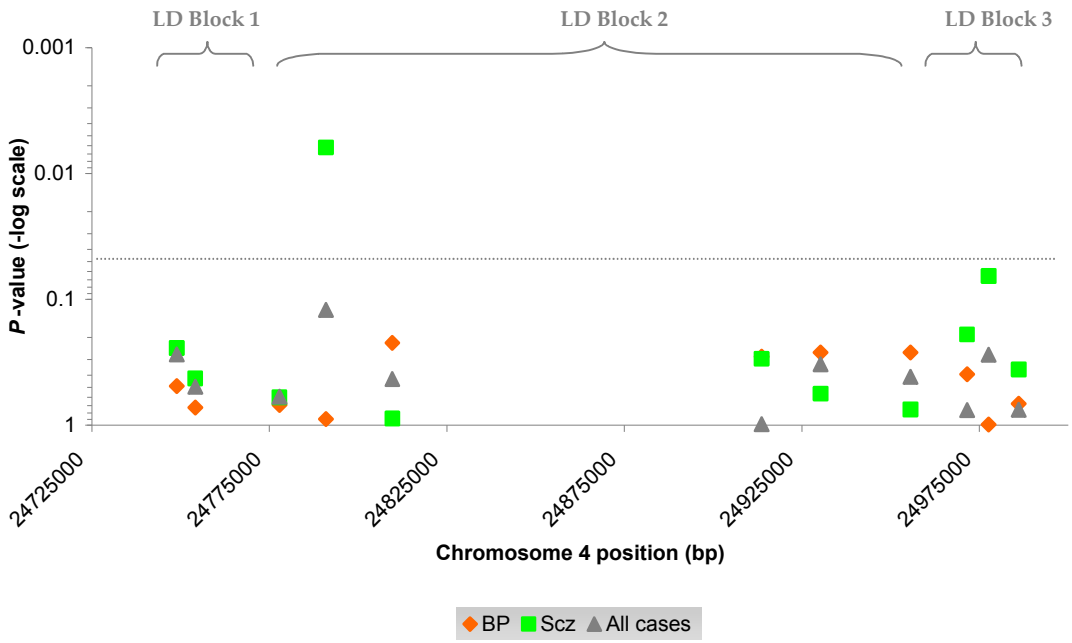


Figure 4.2 Distribution of single-marker allele *P*-values in *PI4K2B* region. The x-axis is the chromosome 4 physical position in base pairs (bp) according to the human genome map of UCSC May 2004, NCBI build 35. The y-axis is the *P*-values on a –log scale. The results are shown for bipolar disorder (BP), schizophrenia (SCZ) and both illness (All cases). The standard significance threshold is indicated with a dashed horizontal line at *P*=0.05. The LD blocks are indicated by a grey bracket and labelled by their respective number according to Table 4.3.

4.2.5. Genotype results

Figure 4.3 shows the results at the genotype level. There were two markers associated with schizophrenia above the provisional significance threshold ≥ 0.05, rs10939038 and rs953305 with *P*-values 0.025 and 0.038 respectively. The association result of rs953305 with schizophrenia was potentially interesting as only *PI4K2B* lies within this LD block, as opposed to the other two LD blocks with *SLA/LP* and *PI4K2B*. However, rs953305 did not withstand permutation testing with 1,000 simulations (permuted *P*=0.069, SE=0.008). Permutation analysis strengthened the result for rs10939038 with 1,000 simulations (permuted *P*=0.039, SE=0.006). Thus, the association of rs10939038 was further investigated in this study.

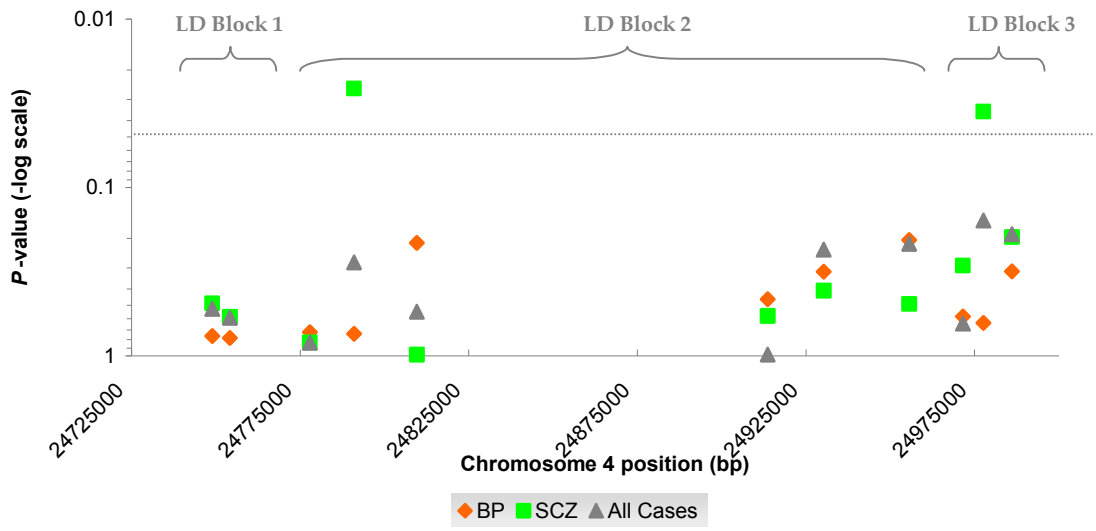


Figure 4.3 Distribution of genotype *P*-values in *PI4K2B* region. The x-axis is the chromosome 4 physical position in base pairs (bp) according to the human genome map of UCSC May 2004, NCBI build 35. The y-axis is the *P*-values on a –log scale for each of the eleven markers. The results are separated according to illness, bipolar disorder (BP), schizophrenia (SCZ) and both illnesses (All cases). The standard significance threshold is indicated with a dashed horizontal line at *P*=0.05.

4.2.6. Details of significant marker

Table 4.5 presents further analyses of rs10939038, taking gender and both bipolar disorder and schizophrenia diagnoses into account. Analysis by gender was justified in bipolar disorder and schizophrenia, as there is reported evidence for gender differences in other candidate gene associations; for example a sex-specific association between bipolar disorder and the G-protein coupled receptor, *GPR50* (Thomson, Wray et al. 2005). From Table 4.5 it is evident that rs10939038 is associated with schizophrenia (*P*=0.006, OR=1.314, 95% CI: 1.08-1.598). This significant result was stronger for male patients with schizophrenia, *P*=0.009, than females, *P*=0.173.

SNP Name	Diagnosis	Gender	Allele (%)		*P*-value	$OR_{T/C}$	Genotype (%)			Total	*P*-value	$OR_{TT/CC}$
			C	T	(X^2)	(95%CI)	CC	CT	TT		(X^2)	(95%CI)
rs10939038	BP	M	171 (55.1)	139 (44.9)	0.348	1.139 (0.85-1.52)	40 (25.5)	91 (58.7)	24 (15.3)	155	0.03	1.46 (0.79-2.71)
rs10939038	BP	F	214 (51.7)	200 (48.3)	0.3	0.866 (0.659-1.137)	60 (28.8)	94 (45.4)	53 (25.5)	207	0.262	1.37 (0.8-2.36)
rs10939038	BP	M&F	385 (53.2)	339 (46.8)	0.899	0.986 (0.811-1.2)	100 (27.4)	185 (51.1)	77 (21.1)	**362**	0.818	1.01 (0.68-1.51)
rs10939038	SCZ	M	327 (59.9)	219 (40.1)	0.009	1.394 (1.08-1.789)	101 (37)	125 (45.8)	47 (17.2)	273	0.04	1.89 (1.15-3.1)
rs10939038	SCZ	F	134 (60.9)	86 (39.1)	0.173	1.261 (0.905-1.759)	39 (35.5)	56 (50.9)	15 (13.6)	110	0.362	1.68 (0.82-3.43)
rs10939038	SCZ	M&F	461 (60.2)	305 (39.8)	0.006	1.314 (1.08-1.598)	140 (36.6)	181 (47.3)	62 (16.2)	**383**	0.025	1.72 (1.15-2.55)
rs10939038	All Cases	M	498 (58.1)	358 (41.9)	0.024	1.288 (1.03-1.615)	141 (32.8)	216 (50.5)	71 (16.5)	428	0.03	1.75 (1.11-2.75)
rs10939038	All Cases	F	348 (54.9)	286 (45.1)	0.905	0.986 (0.769-1.263)	99 (31.1)	150 (47.3)	68 (21.4)	317	0.664	1.06 (0.64-1.76)
rs10939038	All Cases	M&F	846 (56.8)	644 (43.2)	0.119	1.142 (0.966-1.349)	240 (32.2)	366 (49.1)	139 (18.7)	**745**	0.282	1.31 (0.94-1.83)
rs10939038	Controls	M	243 (52)	225 (48)	N/A	N/A	67 (28.6)	109 (46.6)	58 (24.8)	234	N/A	N/A
rs10939038	Controls	F	232 (55.2)	188 (44.8)	N/A	N/A	62 (30)	108 (51)	40 (19)	210	N/A	N/A
rs10939038	Controls	M&F	475 (53.5)	413 (46.5)	N/A	N/A	129 (29)	217 (49)	98 (22)	**444**	N/A	N/A

Table 4.5 Allele and genotype analysis of the significant SNP, rs109390387. The association results are separated by diagnosis, bipolar disorder (BP) and schizophrenia (SCZ) and by gender, male (M) and female (F). OR is the odds ratio. CI is the confidence interval. The highlighted *P*-value represents a significant result *P*-value >0.05. N/A is non-applicable.

4.2.7. Results of model analysis

Each marker was tested for dominant and recessive inheritance models. The results for rs10939038 are shown in Table 4.6. The most significant *P*-value was the dominant model in the schizophrenia group, *P*=0.022.

	X^2	DF	*P*-VALUE
Genotype Counts			
Bipolar Disorder	0.56	2	0.756
Schizophrenia	7.347	2	0.025
All cases	2.545	2	0.280
Dominant Model			
Bipolar Disorder	0.271	1	0.602
Schizophrenia	5.269	1	0.022
All cases	1.198	1	0.274
Recessive Model			
Bipolar Disorder	0.113	1	0.737
Schizophrenia	4.562	1	0.033
All cases	2.129	1	0.145

Table 4.6 rs10939038 results table for different inheritance models. The test results for each group compared to the control group are shown. DF is the degrees of freedom. *P*-values ≤0.05 are highlighted in yellow.

The other markers were also tested. The results are not shown, but there were borderline significant association results for markers in LD block 2, rs313567 and rs6834255 in the bipolar disorder male group for the recessive model, *P*=0.04. Also, rs953305 in LD block 3 was associated in the schizophrenia group for the dominant model (*P*=0.01). However, these results were borderline significant and may not withstand permutation testing. Permutation analysis was not performed due to computational restraints, and the results were not considered further.

4.2.8. Association of haplotypes

Association analysis was performed on haplotypes to improve the information content of the *PI4K2B* genomic region (Ott and Rabinowitz 1997; Chapman and Wijsman 1998), with an aim to detect chromosomal segments that carry the true functional variant. This approach of estimating haplotypes was previously successful in identifying APOE predisposing haplotypes despite two obstacles i) without genotyping the true disease locus and ii) without positive association from the single SNP analysis to Alzheimer's disease (Fallin, Cohen et al. 2001). In this study, the sliding windows analysis was limited to two-SNP haplotypes, separately for bipolar disorder and schizophrenia cases, due to computational restraints and to minimise the multiple-testing burden. Figure 4.4 illustrates the haplotype analysis of the markers in the *PI4K2B* genomic region with the global *P*-value, while Figure 4.5 illustrates the best individual *P*-values. The global test *P*-value assessed the significance of the overall difference in the distribution of haplotype frequencies between cases and controls, which was tested using the Likelihood Ratio Test (LRT). The *P*-value from the individual test represented the significance of the difference in frequency of an individual haplotype between cases and controls and was based on X^2 test.

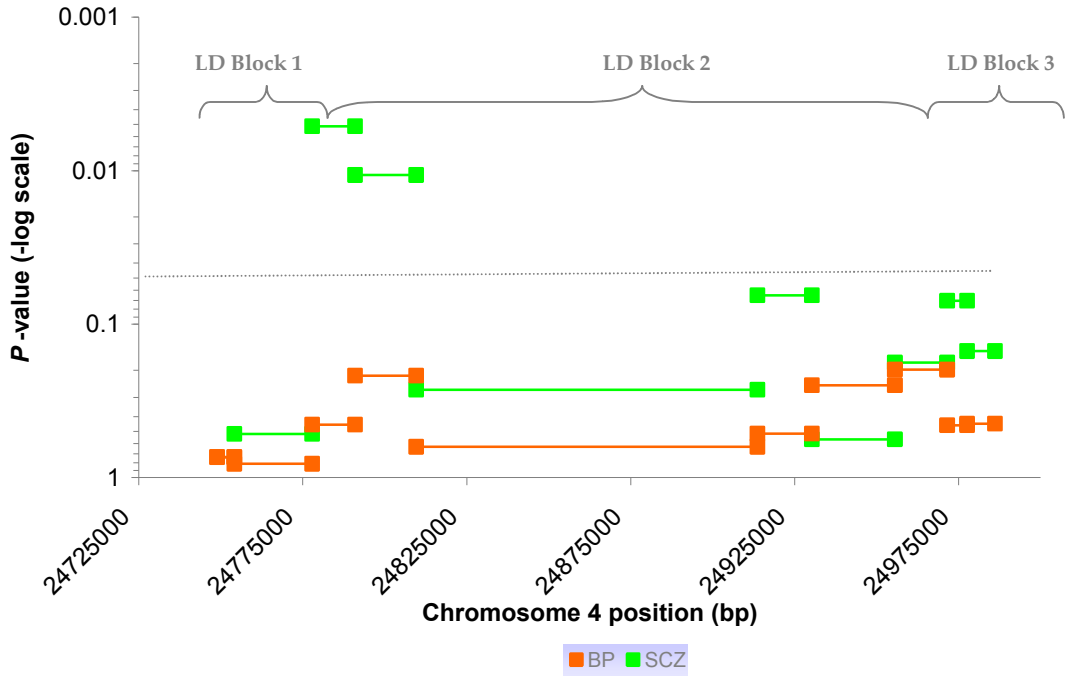


Figure 4.4 Distribution of global haplotype *P*-values from two-marker sliding window haplotype analysis. The x-axis is the chromosome 4 position according to UCSC genome browser, May 04 co-ordinates. The y-axis shows the *P*-values for two-marker sliding windows haplotype analysis using CoCaphase. The provisional significance threshold is indicated with a dashed horizontal line at *P*=0.05. The *P*-value for the first two markers was the same for both bipolar disorder and schizophrenia and overlap so cannot be seen.

Figure 4.4 clearly shows two haplotypes from the two marker sliding-window analyses, which were above the provisional threshold level, *P*-value≥0.05. First, the haplotype of marker 3, rs17408391, in LD block 1 and marker 4, rs10939038, in LD block 2 were determined as significant in the schizophrenia cohort (global *P*= 0.005, permutated global *P*=0.039, SE:0.006). Second, the haplotype identified was from the original data; a two SNP haplotype of markers 4 and 5 in LD block 2; rs10939038 and rs3756207 with a significant global test (*P*=0.01) in the schizophrenia cohort. Table 4.4 shows that marker 3, rs17408391 and marker 4, rs17408391 were in tight LD with each other; D' 0.859 and also that marker 4, rs17408391 and marker 5, rs3756207 were in tight LD with each other; D' 1.0. However, Table 4.4 also shows that markers 3 and markers 5, rs3756207 were not in LD with each other D' 0.458.

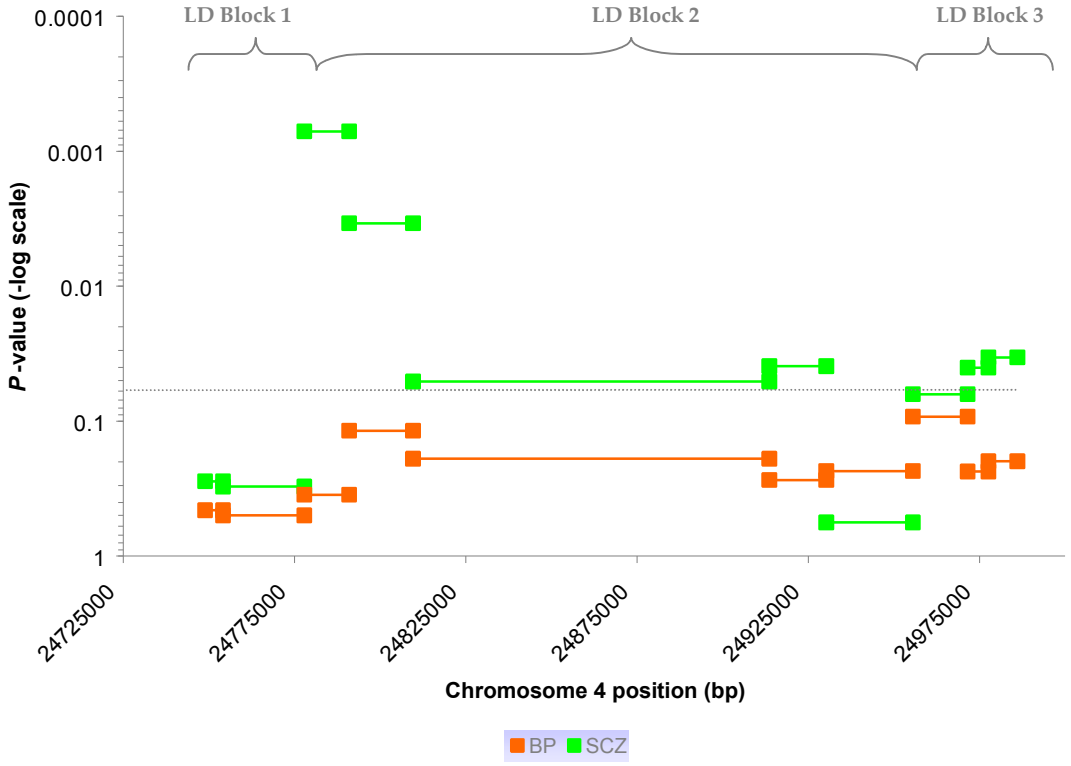


Figure 4.5 Distribution of individual haplotype *P*-values from two-marker sliding window haplotype analysis. The x-axis is the chromosome 4 position according to UCSC genome browser, May 04 co-ordinates. The y-axis shows the *P*-values for two-marker sliding windows haplotype analysis using CoCaphase. The provisional significance threshold is indicated with a dashed horizontal line at *P*=0.05.

Figure 4.5 clearly shows six haplotypes with individual *P*-values that were above the provisional threshold level, *P*-value≥0.05. The best individual haplotype was for marker 3, rs17408391 in LD block 1 and marker 4, rs10939038 in LD block 2 (*P*=0.0007; OR=1.52,1.21-1.91). The next significant haplotype was for neighbouring haplotype markers 4 and 5 in LD block 2; rs10939038 and rs3756207 (*P*=0.004; OR=1.41, 1.13-1.75). These corresponded to the significant global haplotypes reported in Figure 4.4.

The other four individual *P*-values ≥0.05 were as follows; for markers 5 (rs3756207) and 6 (rs313548) in LD block 2 (*P*=0.05, OR=1.45; 95%CI: 1.00-2.11); for markers 6 (rs313548) and 7(rs313567) in LD block 2 (*P*=0.04, OR=1.46; 95%CI: 1.01-2.10); for markers 9 (rs1033102) and 10 (rs953305) in LD block (*P*=0.04); for markers 10 (rs953305) and 11 (rs13133104) in LD block 3 (*P*=0.03, OR=0.81; 95%CI: 0.65-1.01). The most significant haplotype for markers 9 and 10 was very rare (frequency <0.5% in both cases and controls). Odds ratios were not estimated for this haplotype as they are unreliable for rare haplotypes.

4.3. *Discussion*

This study successfully reported that increasing marker coverage in the *PI4K2B* region identified another significant haplotype, which showed greater significance than the original study and importantly, withstood permutation testing.

It is worth noting that the information on the human genome and the understanding of association studies have improved greatly and continue to do so. The association results reported here could reflect association to any feature within the large >240kb genomic region investigated; such as either RefSeq gene *SLA/LP* or *PI4K2B* or any predicted, but yet confirmed gene or in fact, any regulatory or control region. The single marker that is significantly associated with schizophrenia,

rs10939038 is not located in a gene. It is, however, 1,614bp upstream of a human spliced EST isolated from adult cerebellum (AA323552), which potentially could be an alternative 5' UTR for SLA/LP. Upon analysis of the genomic region on the UCSC genome browser (Kent, Sugnet et al. 2002), there was no evidence to suggest the SNP is involved in gene regulation as it was not in an area of cross-species conservation or regulatory potential area. Although, chromosome 4p15-16 is a hot-spot for copy number variation, as will be discussed in section 6.5, there was no evidence for copy number variation or structural variation in the associated region. The other chromosome 4p15.2 marker, rs953305 is located within a human spliced EST isolated from adult testis (DB523474). It is also in an area of cross-species conservation with monkey, mouse, dog and horse sequence. The marker is >20kb downstream of PI4K2B and >13kb upstream of *ZCCHC4* (zinc finger CCHC domain containing 4). Marker rs17408391, is also located outside of any gene coding region. The nearest unspliced EST which was isolated from human fetal liver spleen tissue (H54394), is 3,858bp away. It is in a highly conserved region, as illustrated by the "Vertebrate Multiz Alignment and Conservation" track on the UCSC genome browser. It shows sequence conservation with the rabbit, dog, armadillo, elephant and opossum but not with the mouse, rat, chicken or frog sequence. The other marker that is significant in the haplotype analysis, rs3756207, is located in intron 3 of the *SLA/LP* gene. This marker is also in a highly conserved region with respect to rabbit, dog, armadillo, elephant and opossum sequence but not to mouse, rat, chicken or frog sequence.

The significant result for association between rs10939038 and schizophrenia may be due to one of three possibilities; i) direct association, where the SNP allele directly affects the disease risk ii) indirect association, where rs10939038 is in LD with the true functional disease mutation or iii) spurious due to population stratification or random chance. Firstly, it is highly unlikely that direct association is involved, as bioinformatic analysis showed that the marker does not have any structural features to suggest such a role.

Secondly, the result may be spurious due to genotyping error or population sub-structure. Indeed, genotyping error is a possibility which unfortunately cannot be tested comprehensively in a population study. The genotyping appears to have been robust as the average genotyping success rate was 95%. The success rate for rs17408391 was 98% and the average locus success rate was 95%. Duplication of the genotyping by another method would exclude this possibility.

In addition, presence of population sub-structure can be problematic. Population stratification is due to difference in genotypes between cases and controls reflecting their population origin, not their disease status. Population stratification can only arise if different proportions of cases/controls are from each population and if populations differ in SNP allele frequency. There are many ways to deal with population stratification, for example genomic control (Devlin and Roeder 1999), structured association (Pritchard, Stephens et al. 2000) or to analyse by geographic origin (Clayton, Walker et al. 2005). Although the control group was matched carefully to have the same geographical background as the case group, hidden population sub-structure or relatedness may still be present. Initial studies on the control population using the structured association approach showed no evidence of population stratification in the control group, discussed in (Christoforou, Le Hellard et al. 2007).

Finally, indirect association is possible. Figure 4.6 illustrates the proposed idea that the three markers, rs17408391, rs10939038 and rs3756207 are in fact, in LD with the actual functional variant, labelled in pink. Figure 4.6 also depicts the results of the individual haplotypes from 2-marker sliding window haplotype analysis for the significant SNPs.

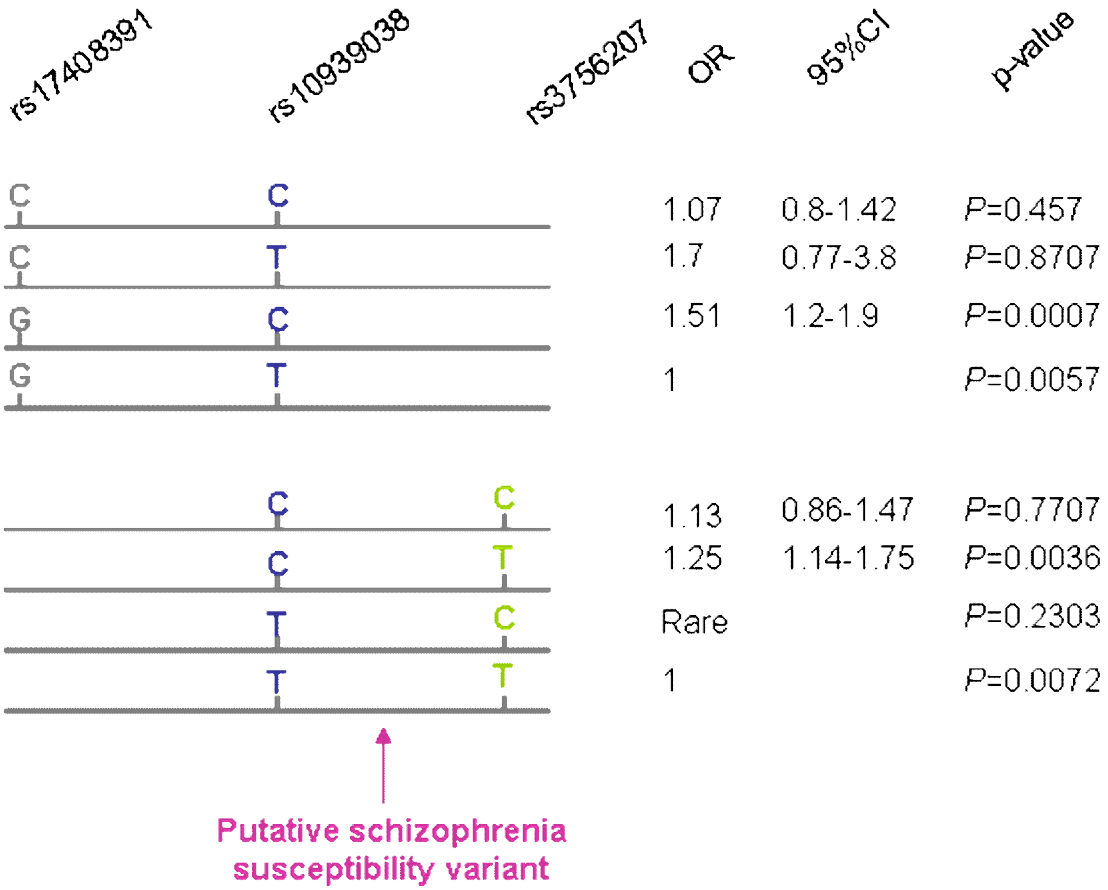


Figure 4.6 Hypothesised indirect association of schizophrenia susceptibility variant. Results of the individual haplotypes from 2-marker sliding window haplotype analysis are shown for the significant markers. The odds ratio (OR), 95% confidence interval (CI) and the X^2 *P*-value are listed.

The indirect association result can be explained in two ways i) although the markers are not physically close to the candidate gene *PI4K2B*, markers rs10939038 and rs3756207 are in the same haplotype block and ii) despite the lack of association in bipolar disorder cases, association cannot be ruled out as there may have been a lack of power or differences in haplotype frequencies of the actual functional variant in the bipolar disorder cases. In the past, examination of haplotype information and LD in a region have confirmed association between the APOE locus and Alzheimer's disease, without testing the actual E4 polymorphism (Fallin, Cohen et al. 2001).

A particular caveat to this study is multiple testing. This study tested association at the gender, phenotype, allele, genotype and haplotype level. The significant association result was examined with respect to the gender of the cases and controls, as there is reported evidence for gender differences in candidate gene studies (Thomson, Wray et al. 2005) and in the age-of-onset, symptoms and treatment of both illnesses (Holden 2005).

To confirm an association, replication is highly recommended. A follow-up case-control association study was performed by others, from a total of 391 schizophrenia cases, 397 bipolar disorder cases and 397 controls in the German population, with the aim of replicating the positive association findings on chromosome 4p15-p16 (Christoforou et al., unpublished). There were over 300 SNPs chosen that tagged significant haplotypes in the Scottish population for the replication study. Of these SNPs, four of them cover the *PI4K2B* genomic region; SNP number 1, 2, 4 and 5 according to Table 4.3 which are rs2324654, rs730061, rs10939038 and rs3756207. However, the significant marker for this study, rs10939038, was not found to be associated with schizophrenia in the German population (allele P=0.55, OR 1.18, 95% CI 0.79-1.77). Furthermore, there was no evidence for association, P-value <0.01, of any of these four markers with bipolar disorder or schizophrenia in this German

population, when testing for gender as a covariate and different inheritance patterns. This implies that the significant association reported here in the Scottish population is less likely to be real.

Although there was no replication of association in the German population and association does not rank highly with the other chromosome 4p15-16 results described in section 1.4.3, the result should not be completely discarded. Replication should be attempted in a different population, preferably one of Celtic origin.

Chapter 5
Design & Preparation for Linkage Study

5. Design and Preparation for Linkage Study

5.1. *Preface*

Appropriate design of a linkage study is essential to achieve meaningful results. In short, the genome must be comprehensively covered with informative markers, the family structure must be conducive for linkage and optimal linkage analysis methods are imperative to maximise the chance of detecting genetic risk variants. To achieve this, three steps were taken. First, although in theory a large pedigree is ideal for linkage analysis (Clerget-Darpoux and Elston 2007), computation analysis is demanding and the capacity to handle very large families is limited. The family, as described in section 1.6, was therefore sub-divided in an informative manner. Second, careful error checking of genotype data is essential to prevent misleading results. There are many sources of error that can be encountered in a family-based linkage study, including errors in diagnoses, gender specification, marker allele frequencies, map order, map distances, genotype calling, DNA quality and family relationships. In this study, several approaches were employed to detect and resolve errors that can lead to misleading inferences about marker inheritance, haplotype frequencies and inflate genetic map length (reviewed in (Pompanon, Bonin et al. 2005). In addition, pedigree structure errors were investigated and resolved. Finally, simulations were performed to estimate the threshold required to report suggestive and significant linkage results. The methods used to optimise the linkage study design, to perform the error checking and to validate the genotyping data are detailed below, with the aim to provide an optimal environment for linkage analysis.

5.1.1. Study design

In the initial stages of study design the genotyping platform, the marker panel and

the sample selection were surveyed. The Linkage IVb Panel from Illumina was chosen for three reasons, i) high genotyping quality at a level of 99.9% reproducibility and genotyping call rate 99.8% (Murray, Oliphant et al. 2004), ii) the 6,008 SNP panel was informative for linkage analysis as the average genetic map spacing is 0.64cM, the average genotyping information content is 97.1%, the average minor allele frequency in the Caucasian population is 37% and the average heterozygosity is 43% (Murray, Oliphant et al. 2004) and iii) the panel has shown previous linkage analysis successes in complex disease (Amos, Chen et al. 2006). The genotyping was performed using a GoldenGate Assay on an Illumina BeadStation platform, which requires a 96-well plate configuration. Cost considerations, therefore, limited the number of samples to 96 in this study. The 96 family samples were chosen by a number of criteria: to maximise the linkage information, to cover all branches of the pedigree, to include all affected individuals, DNA availability and to include a replicate sample of whole genome amplified and genomic DNA. Thus, 95 individuals from the family were genotyped with 6,008 SNPs from the Illumina Linkage IVb Panel.

Technical replicates are often included in large-scale genotyping experiments. These replicates are the same sample, included multiple times in the same experiment, to test variability in the experiment. However, technical replicates were not included in this linkage study. Instead, as this was a family study, errors could be easily detected by checking for deviation from Mendelian inheritance. Furthermore, Illumina have internal control measures that show confidence in their genotyping reproducibility at 99.9% and a low rate of Mendelian inconsistencies at 0.007% (Murray, Oliphant et al. 2004). Nevertheless, a replicate sample with genomic and whole genome amplified DNA was included as a test of reliability of amplification.

No formal calculations of power were performed for this study due to computational burden and family complexity. However, the original detection of

linkage using microsatellite markers proves the ability of the pedigree to generate a significant LOD score and was an *a priori* reason to reinvestigate this family for linkage (Blackwood, He et al. 1996). An increase in the number of affected individuals available for linkage analysis and the comprehensive coverage of the genome by the Illumina IVb Linkage Panel should increase the power of this analysis, in comparison with the first study (Evans and Cardon 2004).

5.2. *Pedigree Preparation*

5.2.1. Phenotype definition

Linkage analysis was performed under two definitions of the phenotype i) a narrow phenotype model that included only individuals with bipolar I disorder and bipolar II disorder as "affected" and ii) a broad model that included individuals with bipolar I and II disorder and recurrent major depression. Individuals with a diagnosis of single episode depression, anxiety or alcoholism, or individuals with potential bilineal inherited illness were coded as unknown for linkage analysis. Thus, there were 12 individuals with bipolar disorder in the narrow phenotype category and 35 individuals (12 bipolar disorder and 23 recurrent major depression) in the broad phenotype category available for linkage analysis, as described in Table 1.5.

5.2.2. Splitting the pedigree

A clear advantage of this project was the availability of a large family. Software such as Simwalk2, that can perform linkage analysis on large pedigrees, was very slow and limited to a very small number of markers. The size restriction for the highly regarded MERLIN programme was ($2n-f \leq 20$, where n was the number of individuals in the pedigree and f was the number of founders). The large pedigree exceeded the size restriction; $2*180–52 = 308$ and was thus computationally complex.

In order to overcome this calculation burden and simultaneously maintain the useful information for linkage mapping, the family was split in a well-defined manner. This method was preferable to splitting the family by eye, which can be an arbitrary approach that may impact on power (Dyer, Blangero et al. 2001; Bourgain and Genin 2005). Sub-pedigrees were, therefore, extracted using GREFFA (Genetic Relationship Explorer for Familiality Aggregation) (Falchi, Forabosco et al. 2004). GREFFA clustered individuals using pair-wise kinship coefficients, partitioned the individuals into optimal sub-groups and reconstructed the sub-pedigrees of each sub-group. To extract the optimal sub-pedigrees, the pairwise kinship coefficient chosen was the level of first-cousin relationship 1/16 (0.0625) and the maximum number of individuals was the limit for MERLIN analysis, <20, as described in section 2.5.4. This approach extracted one large informative family and five others, depicted in Figure 5.1.

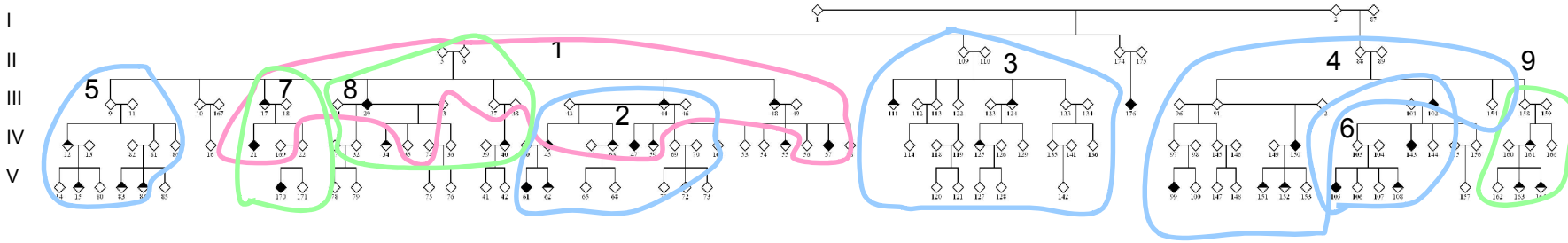


Figure 5.1 Sub-pedigrees. Sub-pedigrees from the main family were extracted using GREFFA software. The most informative sub-pedigree is highlighted in pink, and the other GREFFA extracted pedigrees are highlighted in blue. Further sub-pedigrees highlighted in green were constructed to include additional individuals that had been genotyped. Each sub-pedigree is reproduced in Figure 5.2 to Figure 5.10.

The most informative sub-pedigree is reproduced again in Figure 5.2. The other sub-pedigrees extracted by GREFFA were illustrated from Figure 5.3 to Figure 5.7. As there was genotyping information available for other members, three further sub-pedigrees were manually added, giving a total of nine sub-pedigrees. The additional three sub-pedigrees are shown from Figure 5.8 to Figure 5.10. To provide phase information, it was sometimes necessary to include affected individuals in more than one sub-pedigree. This was the case for three samples with bipolar disorder diagnoses: sample 21 in sub-pedigrees 1 and 7, sample 29 in sub-pedigrees 1 and 8 and sample 102 was in sub-pedigrees 4 and 6 and for five samples with recurrent major diagnoses: sample 44 and 63 in sub-pedigrees 1 and 2, sample 17 and 19 in sub-pedigrees 1 and 7 and sample 34 in sub-pedigrees 1 and 8. In order to avoid bias, the affection status for individuals in more than one sub-pedigree was set to unknown in every subsequent pedigree, and thus the individuals were not considered more than once for linkage analysis.

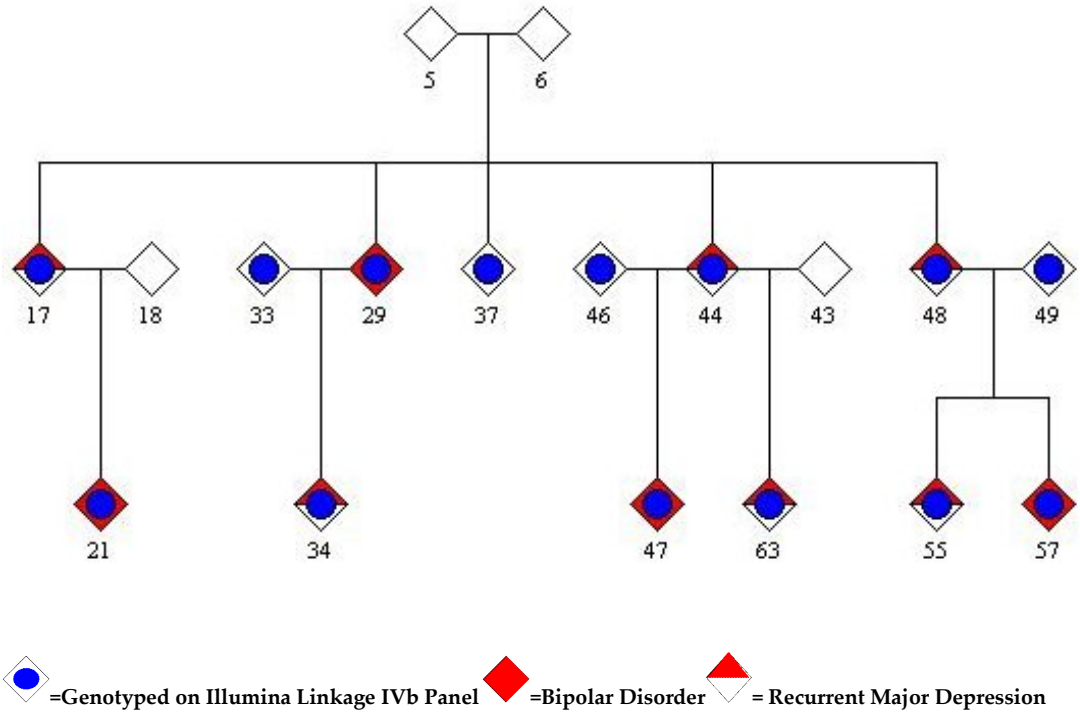


Figure 5.2 Sub-pedigree 1. This is the most informative pedigree for linkage analysis under both the broad and narrow phenotypic model. Samples 21, 29, 34, 44 and 63 overlap with other sub-pedigrees.

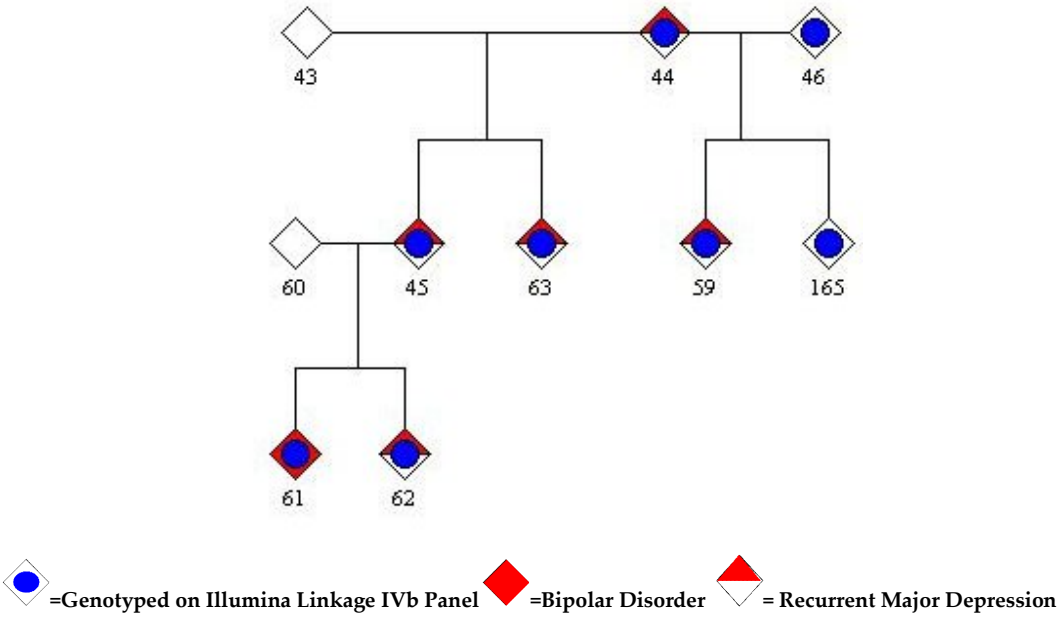


Figure 5.3 Sub-pedigree 2. This pedigree is informative for linkage analysis under the broad phenotypic model. Samples 44 and 63 overlap with sub-pedigree 1.

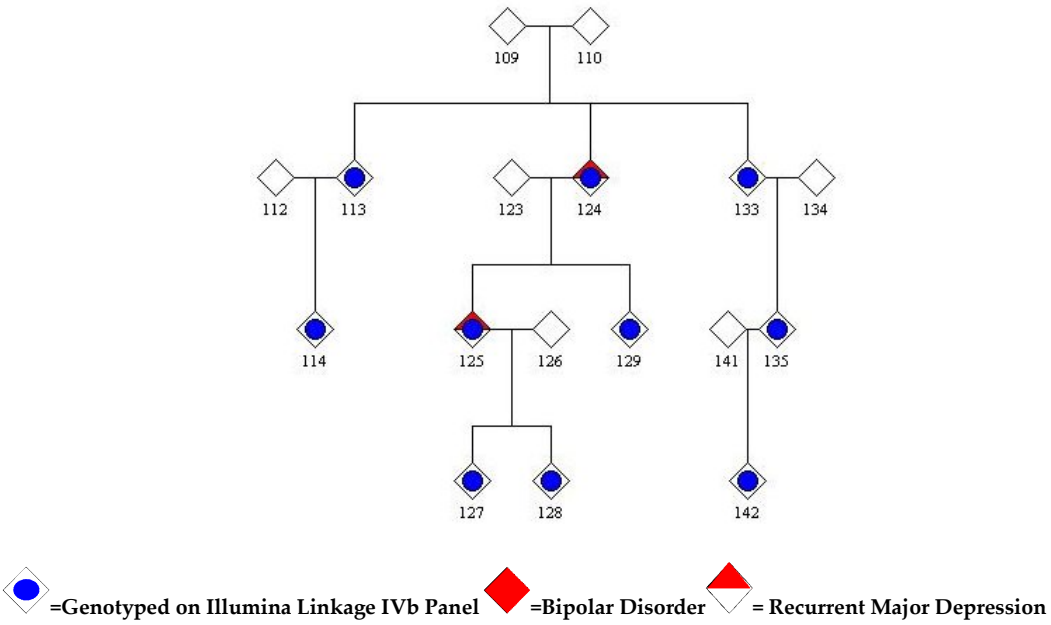


Figure 5.4 Sub-pedigree 3. This pedigree is informative for linkage analysis under the broad phenotypic model.

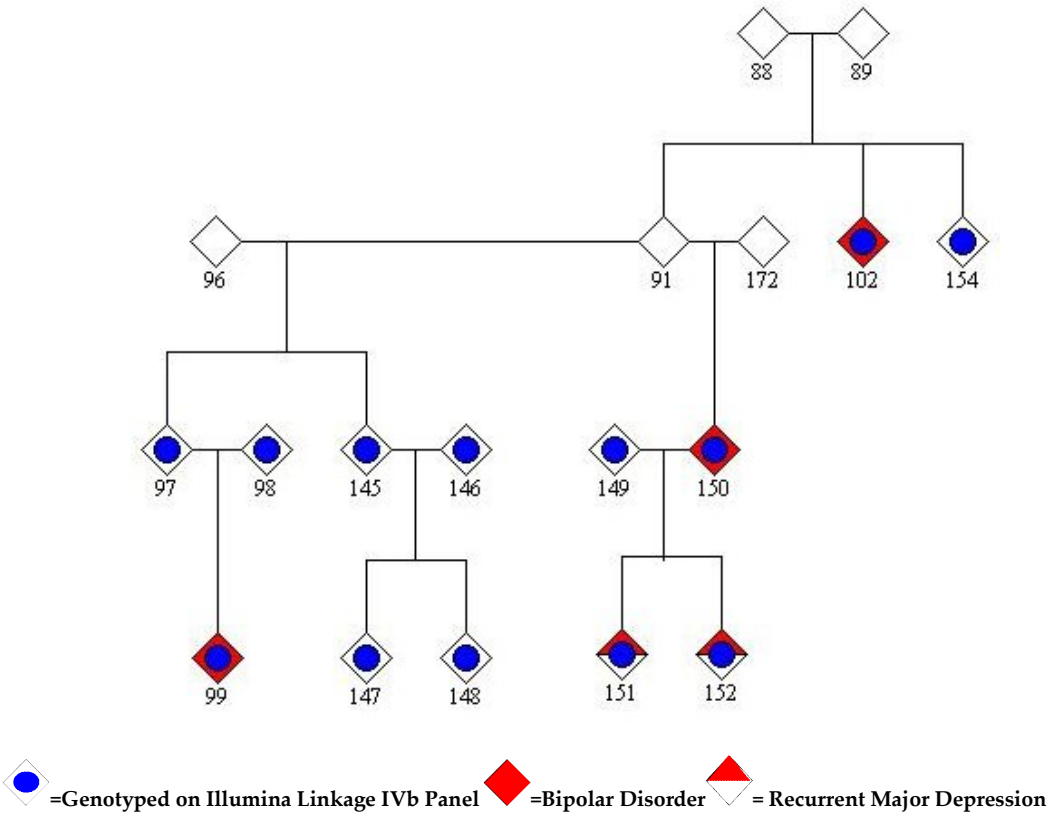


Figure 5.5 Sub-pedigree 4. This pedigree is informative for linkage analysis under the broad and narrow phenotypic model. Sample 102 overlaps with sub-pedigree 6.

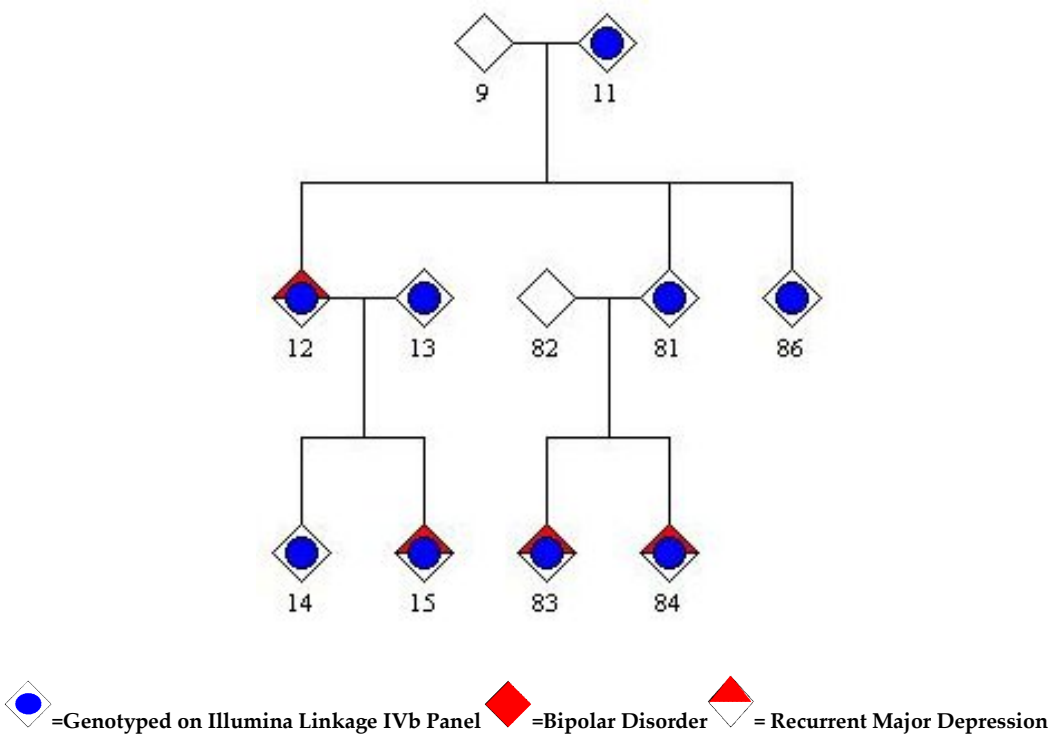


Figure 5.6 Sub-pedigree 5. This pedigree is informative for linkage analysis under the broad phenotypic model.

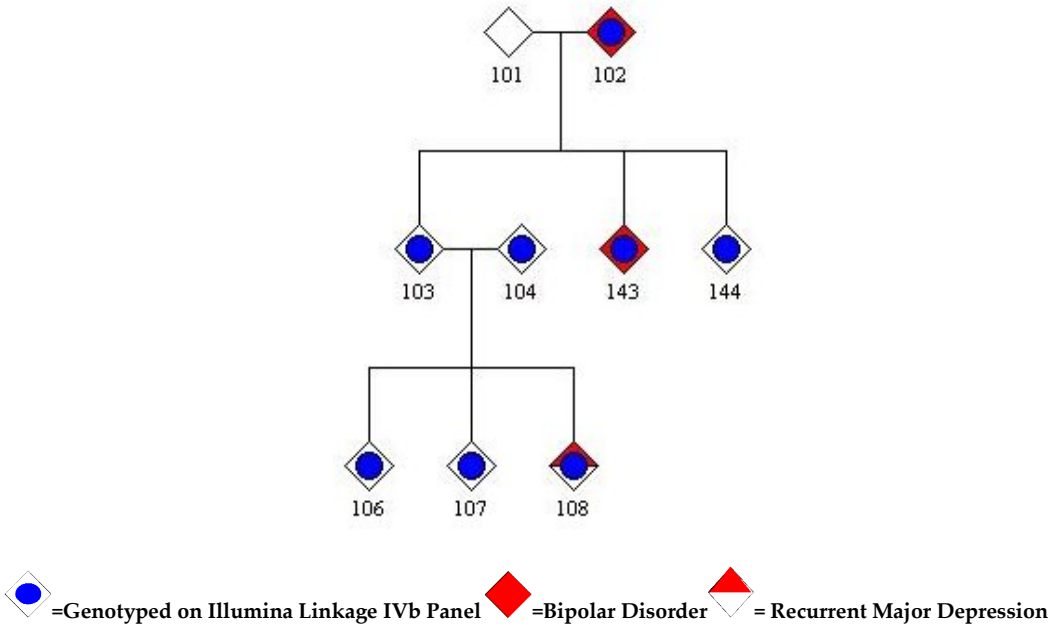


Figure 5.7 Sub-pedigree 6. This pedigree is informative for linkage under the broad phenotypic model. Sample 102 overlaps with sub-pedigree 4.

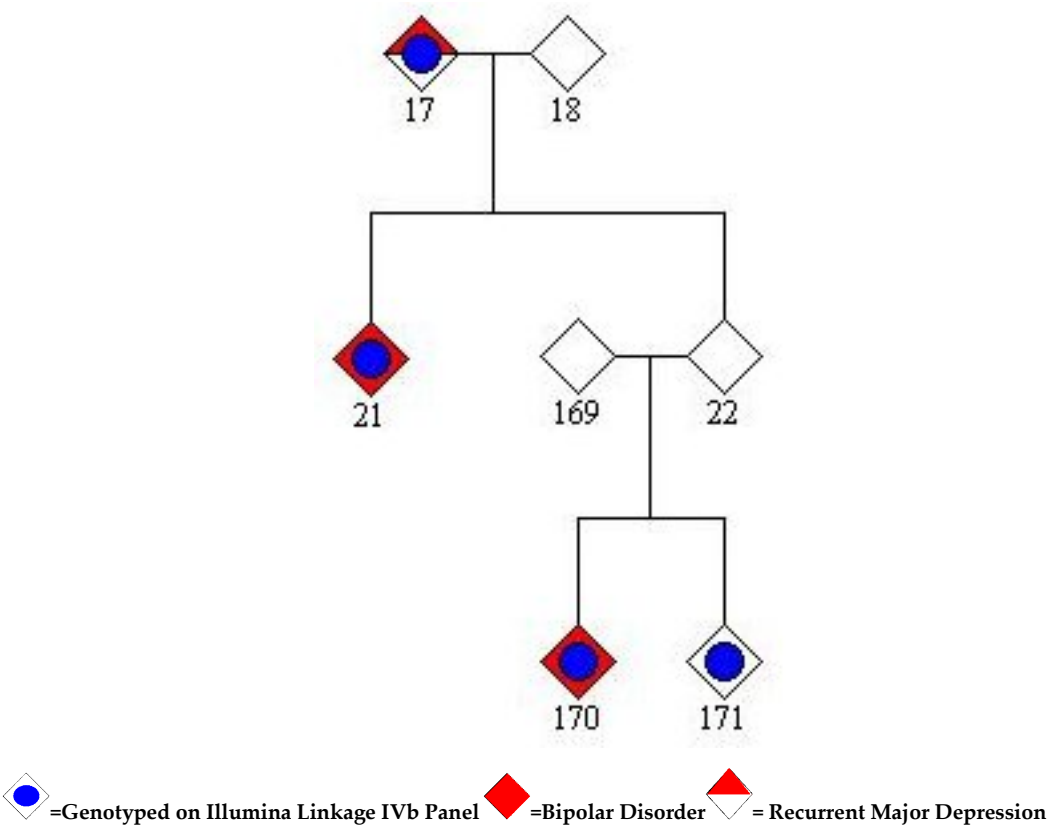


Figure 5.8 Sub-pedigree 7. Samples 17 and 21 overlap with sub-pedigree 1.

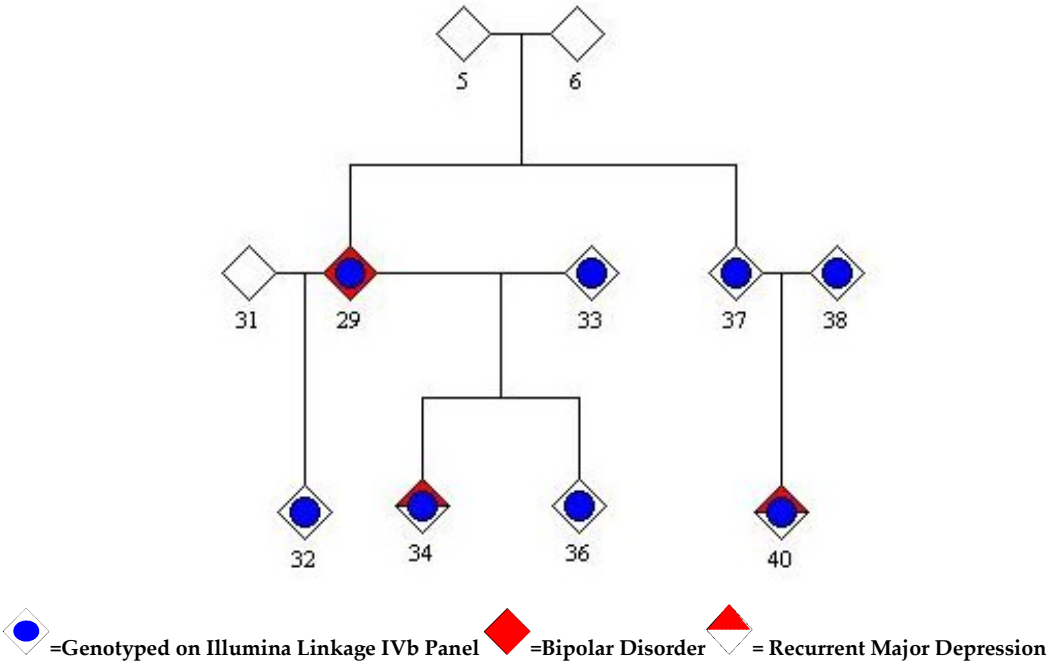


Figure 5.9 Sub-pedigree 8. Samples 29 and 34 overlap with sub-pedigree 1.

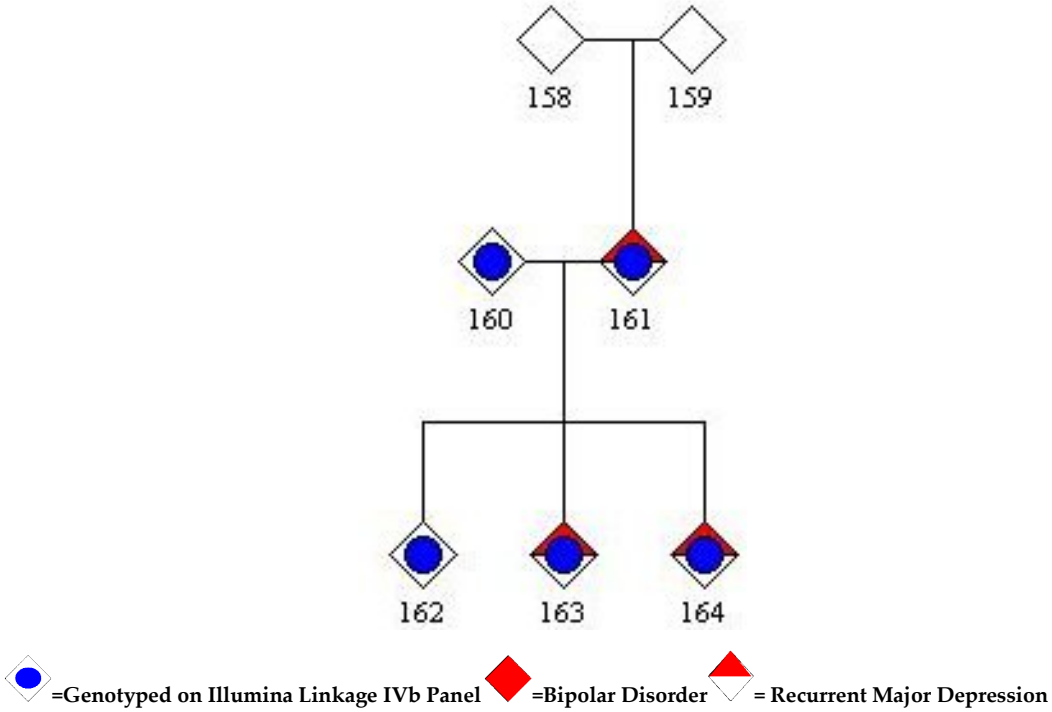


Figure 5.10 Sub-pedigree 9. This pedigree is informative for linkage analysis on the broad phenotypic model.

5.3. *Quality Control of Whole Genome Amplified DNA*

5.3.1. Preface

Certain genomic DNA samples were subjected to whole genome amplification for the purpose of the linkage study and for future use. There were three important reasons to preserve these DNAs i) firstly, there was only a small quantity of DNA remaining, which would limit future experiments ii) secondly, the DNA was old, degraded and would benefit from preservation and iii) finally, certain DNA samples were from individuals now deceased. An established method to amplify DNA is whole genome amplification (Hosono, Faruqi et al. 2003). The amplification method used was based on multiple displacement amplification technology, which performed isothermal genome amplification utilizing a processive DNA polymerase capable of replicating up to 100kb without dissociating from the genome DNA template. The DNA polymerase had a 3′ to 5′ exonuclease proofreading activity to maintain high fidelity during replication and was used in the presence of exonuclease-resistant primers to achieve high yields of DNA product. This method was previously shown to provide highly uniform DNA amplification across the entire genome, with minimal sequence bias (Hosono, Faruqi et al. 2003).

5.3.2. Whole genome amplification results

Whole genome amplification of DNA from two samples (samples 37 and 44) was attempted using the Qiagen REPLI-g Midi Kit, as described in section 2.2.6, in duplicate for each sample from different aliquots. Each amplification reaction was performed in duplicate, so there were four amplification samples for each individual. The duplication was important as any stochastic bias in amplification in one sample could be overcome by combining the two samples. The starting material is detailed in Table 5.1. The starting material was older than ten years, were either apparently empty tubes or had a volume less than 2µl remaining.

						Microsatellite Markers			
	Original [DNA]	DNA remaining	WGA [DNA]		Yield	Chr1		Chr4p15-16	
WGA Sample	Agarose Gel	Volume	Agarose Gel	Picogreen	ug	Successful PCR	Successful Genotype	Successful PCR	Successful Genotype
37a #1			<100ng/µl			5/5	Yes	4/4	Yes
37a #2	>500ng/µl (smear)	~2µl	<100ng/µl	211ng/µl	11	5/5	Yes	4/4	Yes
37b #1	5ng/µl	~2µl	<100ng/µl	212ng/µl	11	5/5	Yes	4/4	Yes
37b #2			<100ng/µl	157ng/µl	8	5/5	Yes	4/4	Yes
Positive Ctrl (29)				151ng/µl		5/5	No	5/5	Yes
Positive Ctrl (34)	5ng/µl			126ng/µl		5/5	Yes	4/4	Yes
44a #1	5ng/µl	~2.5µl	100ng/µl	212ng/µl		None	No	1/4	Yes
44a #2			100ng/µl	219ng/µl		None	No	1/4	Yes
44b #1	10ng/µl	~1µl	100ng/µl	40ng/µl		None	No	0/4	No
44b #2			100ng/µl			None	No	0/4	No
44a	5ng/µl	~2.5µl	75ng/µl			None	No	2/5	No
44b	10ng/µl	~1µl	75ng/µl	207ng/µl		None	No	2/5	No
Negative Ctrl			100ng/µl			None	No	None	No
Negative Ctrl			75ng/µl			None	No	None	No

Table 5.1 Quantification and genotyping results for whole genome amplified (WGA) samples. The amplification of three DNA samples was attempted: 37 and 44. There were two genomic DNA samples, a and b, amplified with two replicates, 1 and 2. Positive controls of good-quality genomic DNA and negative controls of dH_2O were included in the experiment. The DNA concentration was determined by agarose gel electrophoresis for the original DNA sample and also by picogreen for the WGA DNA. Microsatellite markers were genotyped to determine the integrity of the DNA. The success of the PCR amplification step and genotyping step is shown for each sample. There is no information for certain samples when there was a negative result or the test was not performed.

One sample was successfully whole genome amplified, as determined by DNA qualitative and quantitative analysis, supplemented by microsatellite genotyping. One sample, 44, did not amplify successfully. The quality of the DNA was assessed by agarose gel electrophoresis and the DNA yield was determined by picogreen analysis. The details of the DNA quantification and genotyping of the samples and replicates are also in Table 5.1. These quality controls tests were important factors in ensuring optimal DNA for downstream procedures.

Firstly, the initial quantification of the DNA was performed by agarose gel electrophoresis as shown in Figure 5.11. The concentration of the DNA can be estimated by comparing it to the λHindIII marker. The whole genome amplified DNA concentration ranged from 75ng to 200ng. The DNA concentration was also estimated by picogreen reagent and gave similar results. There was evidence of a band in the negative control for whole genome amplification in Figure 5.11. However, PCR amplification of this non-template control was unsuccessful. This may be DNA generated during REPLI-g reaction by random extension of primer dimers and should not effect the quality of actual samples or downstream assays (www.qiagen.com).

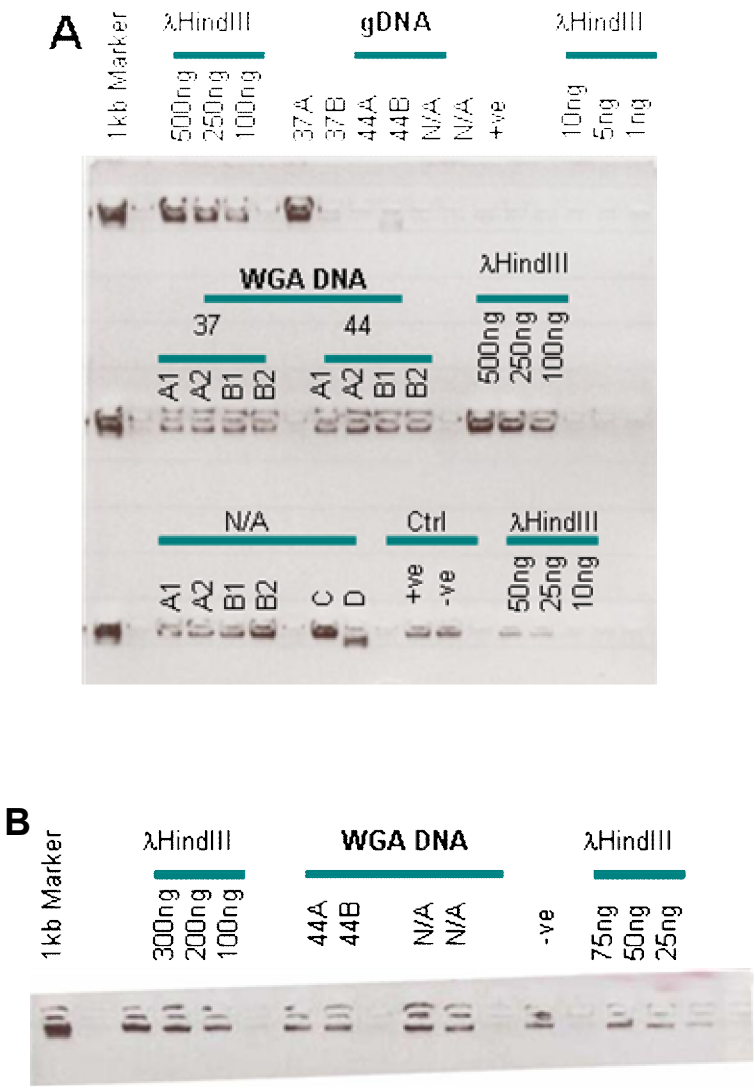


Figure 5.11 Agarose gel of whole genome amplified products to quantify DNA. The amplification products of samples 37 and 44 are shown in (a) and a repeated attempt for sample 44 in (b). 1µl of a 1/10 dilution of the WGA DNA was run on a 0.8% agarose gel stained with EtBr for 5 minutes at 40V. λHindIII is a marker for comparison to estimate the DNA concentration. There were two genomic samples for WGA from each individual, labelled A and B. Each was amplified in duplicate, labelled 1 and 2. The positive controls (+ve) were amplification of lymphoblastoid cell line sample 34 on gel A and 29 on gel B. The non-template control (-ve) was water. N/A means non-applicable and does not pertain to this study. Gel B shows the repeat amplification attempt for sample 44 by encouraging the remaining DNA in the eppendorf to dissolve at 65°C for 1 hour.

Secondly, the quality of the whole genome amplified DNA was assessed by running the agarose gel for a longer time as shown in Figure 5.12. All samples had the expected genomic DNA band 10kb-12kb. There was smearing in all samples which indicated DNA degradation and/or shearing but as the bulk of DNA was >10kb it was sufficient quality for use. The yield for the successfully amplified samples ranged from 8-11μg. The manufacturers quoted an expected yield of 40μg (www.qiagen.com), however starting DNA quality may have been a factor in the reduced yield for these samples.

It was not surprising that the genomic DNA aliquots for sample 44, as shown in Figure 5.12a, and the repeated attempt at DNA dissolution as shown in Figure 5.12b did not amplify. The recommended starting genomic DNA fragment size was 2kb to an optimal 10kb (www.qiagen.com), which was not evident in sample 44, Figure 5.12a. Also, the genomic DNA was low in quantity and thus did not serve as a good template for amplification.

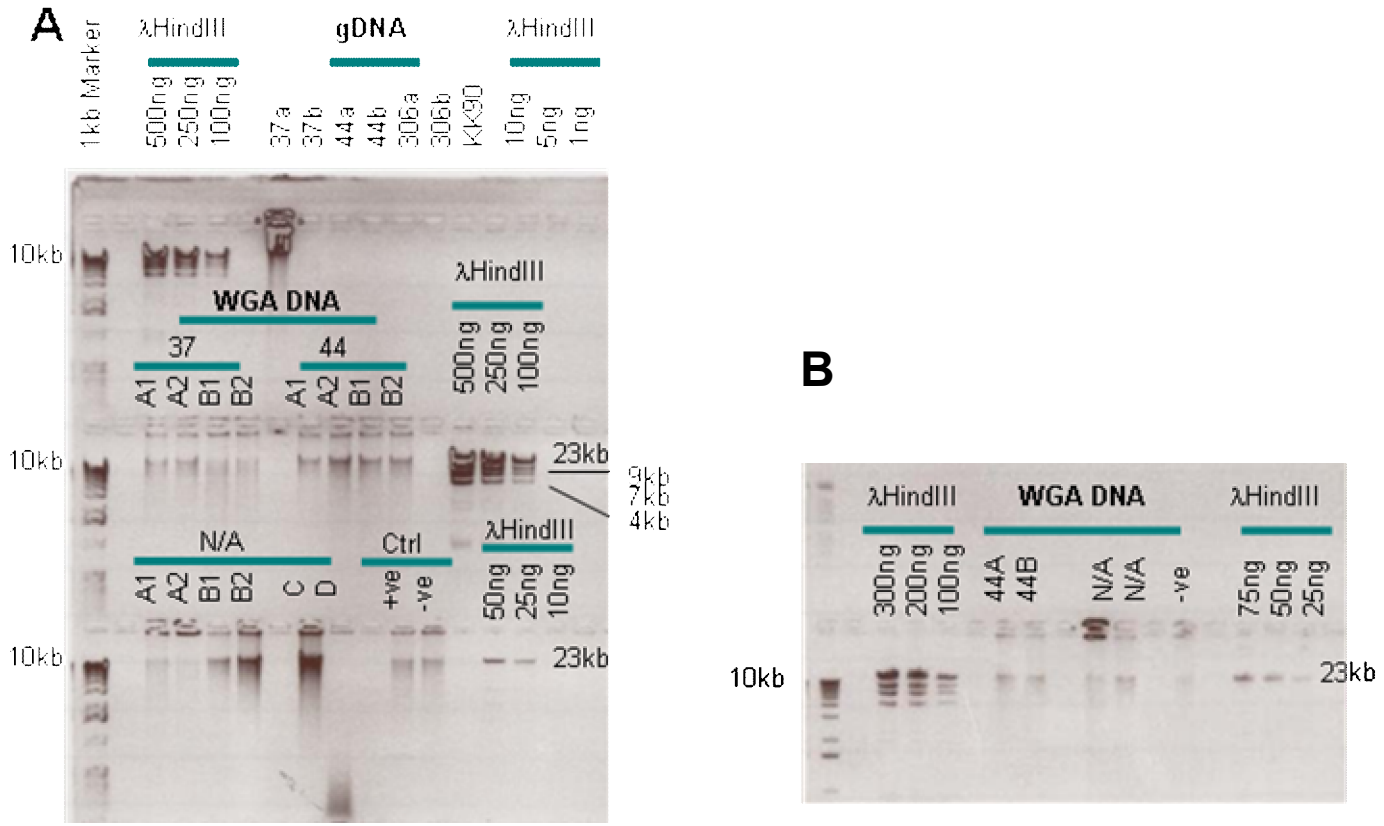


Figure 5.12 Agarose gel of whole genome amplified products to determine DNA quality. The amplification products of samples 37 and 44 are shown in (a) and a repeated attempt for sample 44 in (b).1µl of a 1/10 dilution of the WGA DNA was run on a 0.8% agarose gel stained with EtBr at 40V for 5 minutes, then 80V for 45 minutes (picture A) and 70 minutes (picture B) at 40V. The numbers correspond to ID numbers. There were two genomic samples for WGA from each individual labelled A and B. Each was amplified in duplicate, labelled 1 and 2. The positive controls (+ve) were lymphoblastoid cell line sample 34 on gel A and 29 on gel B. The non-template control (-ve) was water. N/A are samples that do not pertain to this study. Gel B shows the repeat amplification attempt of sample 44 by encouraging the remaining DNA in the eppendorf to dissolve at 65°C for 1 hour.

A further test of whole genome amplified DNA integrity was microsatellite genotyping, as described in section 2.2.5.1. Successful and correct genotyping of polymorphic microsatellites, in all replicates, would lend credence to the whole genome amplification procedure and downstream use of the amplified DNA. The samples were thus genotyped with 11 microsatellite markers from chromosome 1 and chromosome 4.

The microsatellite genotyping results are shown in Table 5.1 and detailed in Table 5.2, showing successful genotyping for sample 37. Importantly, the genotyping matched between all replicate whole genome amplified DNA samples. For further confirmation, there were extra layers of genotype checking. The genotyping matched between genomic and whole genome amplified DNA for sample 37 on four microsatellites, D4S1533, stb131K9, D4S1609 and D4S2408, where data was available for both samples. Family members of sample 37, were also genotyped, and Mendelian segregation patterns were observed. The genotyping results were also checked from previous studies. An important final check was that nine of the microsatellite markers were heterozygous which would have allowed detection of allelic dropout if present. Crucially, the genotyping results matched between different samples, replicates, segregation patterns and previous known results. Thus, there was no allelic drop-out observed and these results were sufficient to include the whole genome amplified samples in the linkage study.

	D1S439		*D1S103*		*D1S225*		*D1S229*		*D1S1621*		*D4S1599*		*D4S1533*		*STB131K9*		*D4S1609*		*D4S2408*		*D4S1546*	
Sample	*1*	*2*	*1*	*2*	*1*	*2*	*1*	*2*	*1*	*2*	*1*	*2*	*1*	*2*	*1*	*2*	*1*	*2*	*1*	*2*	*1*	*2*
37a #1	249	251	77	89	124	124	196	196	163	169	F	F	186	194	325	329	166	174	NA	NA	NA	NA
37a #2	249	251	77	89	124	124	196	196	163	169	NA	NA	186	194	325	329	166	174	NA	NA	NA	NA
37b #1	249	251	77	89	124	124	196	196	163	169	NA	NA	186	194	325	329	166	174	NA	NA	NA	NA
37b #2	249	251	77	89	124	124	196	196	163	169	NA	NA	186	194	325	?	166	174	NA	NA	NA	NA
37stock gDNA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	150	154	186	194	325	329	F	F	270	274	NA	NA
37 diluted gDNA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	150	154	186	194	325	329	166	174	270	274	NA	NA
Key: F=Failed at genotyping, NA = Not attempted, ? = ambiguous results																						

Table 5.2 Genotyping results microsatellite markers on whole genome amplified DNA samples and genomic DNA samples. Samples 37 were successfully whole genome amplified. The samples were tested for genotyping accuracy by comparing results to replicate samples and other genomic DNA samples from the same individual. Eleven microsatellite markers were genotyped, five from chromosome 1 and six from chromsome 4. The genotyping matched between all replicate WGA DNA samples. The genotyping also matched between genomic and WGA DNA for 37 on four microsatellites, D4S1533, stb131K9, D4S1609 and D4S2408.

5.4. *Metrics of Illumina Linkage IVb Mapping Panel*

For this study, 96 samples were genotyped on 6,008 SNPs from the Illumina Linkage IVb Panel. All samples were from 95 individuals in a large Scottish family. One of the samples was a whole-genome amplified replicate (sample 37). Measurements pertaining to the Illumina Linkage IVb Panel genotyping on the large Scottish family were detailed below.

5.4.1. Genotyping call rate success

The average genotyping success rate was 95%, which is lower than the expected rate of 99.8% published by Illumina (Murray, Oliphant et al. 2004) for the Linkage IVb Mapping Panel. The average percentage of missing genotypes per chromosome is shown in Figure 5.13 by the green line. Genotyping failed in samples 12, 15 and 19 for >20% of markers and samples 25, 35 and 40 for > 30% of markers. The DNA from sample 83 failed for 55% of markers. This level of genotyping failure possibly indicates poor DNA sample quality or quantity. Upon further investigation it was found that the genotyping failures were specific to the primer mix from Illumina, named OPA (oligo pool all); sample 83 failed with OPA 6561, sample 25, 35, 40, 83 failed with OPA 6562 and sample 12, 15, 19 failed with OPA 6564. Thus, poor OPA and DNA quality are the probable reasons for the missing genotypes. The genotyping results for sample 83 were removed from the linkage data. The other samples were included in the linkage analysis, as the genotyping failure was OPA specific. To ensure the accuracy of the genotypes, tests were carried out on the markers to check for Mendelian inheritance and Hardy-Weinberg equilibrium (HWE).

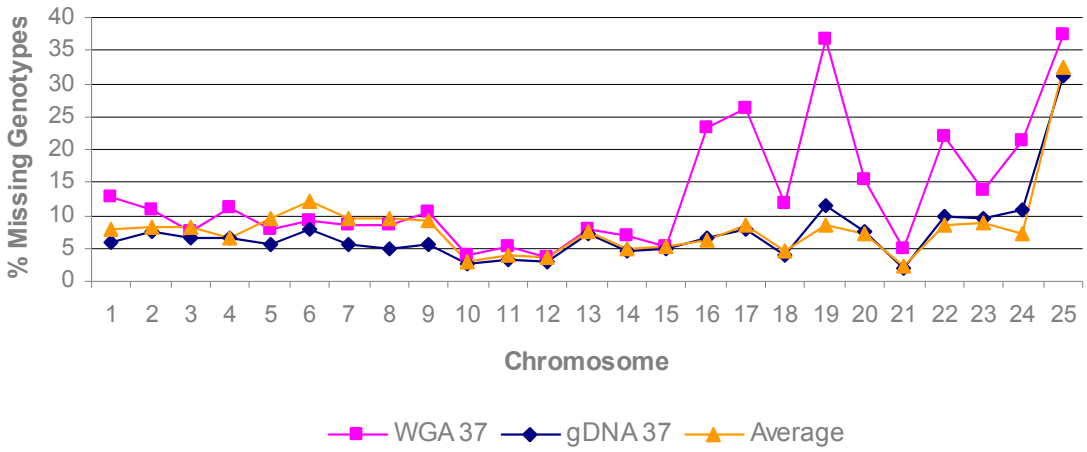


Figure 5.13 Comparison of missing genotypes per chromosome, for individuals genotyped on the Illumina Linkage IVb Panel. The x-axis is the chromosome number; 1-22 are autosomes, 23-25 are X, XY and Y respectively. The y-axis shows the amount of missing data, which was calculated as a percentage of total genotyping performed for each sample that was genotyped on the Illumina Linkage IVb Panel. The whole genome amplified DNA sample is in pink (37). The results for 37 genotyped on genomic DNA are in blue. The average amount of missing data from all 96 samples is shown by the orange line.

5.4.2. Reproducibility success of genomic DNA

From relationship analysis, as explained in section 5.5, there was evidence for a sample duplication of individual 97. The DNA aliquot labelled 105 was in fact, from sample 97. Analysis of the 3,998 markers that genotyped successfully for both samples showed that the data matched exactly and correlated significantly at the 0.01 level, Pearson Correlation Significance (2-tailed) 1.000. This coincides with Illuminas published level of reproducibility of 99.6% (Murray, Oliphant et al. 2004) for the Linkage IVb Mapping Panel.

5.4.3. Success of whole genome amplified DNA

As described in section 5.3.2, one sample (37) was successfully whole genome amplified and genotyped in the linkage study. A genomic DNA sample of 37 was

also genotyped to provide a duplicate data set. Table 5.4 shows that the three samples were genotyped successfully. Investigation of sample 37 genotyping results was performed by checking the correlation between the genomic and whole genome amplified DNA. Table 5.3 showed only three genotyping mismatches out of a total 5,304 genotypes.

SNP	CHROMOSOME	GENOMIC DNA GENOTYPING	WGA DNA GENOTYPING
rs7559853	2	GG	AG
rs1260658	6	AA	AG
rs733342	17	AA	AG

Table 5.3 Mismatch genotyping between a duplicate sample of genomic and whole genome amplified DNA. There were three mismatch genotypes between 37 genomic and whole genome amplified (WGA) replicates. The total number of successful genotyping between both DNA samples was 5,304 SNPs.

The error rate per locus for sample 37 was 0.057%, calculated as shown in Appendix A, using an established method (Pompanon, Bonin et al. 2005). Under the assumption that the genotypes for genomic DNA were correct, it was interesting to note that the amplified DNA genotypes were called heterozygous but were in fact homozygous SNPs. The over-representation of heterozygous erroneous calls have been found by others to be a side-effect of the Illumina clustering algorithm (Personal Communication, Simon Heath, CNG, France, 23rd July 2006). Due to this prior reason, the cluster plots from the Illumina BeadStudio were examined for these three SNPs, and in fact, the whole genome amplified samples were outside of the majority of clustered genotyping and their genotypes were changed to unknown. The genotypes for the genomic DNA samples were positioned in the middle of the homozygote clusters. However, as it was not feasible to re-score all

SNPs it cannot be said that there was complete concordance between the whole genome amplified DNA and the genomic DNA. Nevertheless, the error rate per locus of 0.057% is low.

The genotyping call-rate for the whole genome amplified samples matched the genomic DNA samples. Table 5.4 shows the level of successful genotyping.

SAMPLE	*% SUCCESSFUL GENOTYPING*
All samples (average)	93
37 (genomic DNA)	94
37 (WGA DNA)	89

Table 5.4 Whole genome amplified products were successfully genotyped on the Illumina Linkage IVb Panel. The level of genotyping success for 6,008 SNPs on the Illumina Linkage IVb Panel was calculated for all samples, including whole genome amplified (WGA) samples on all autosomes and sex chromosomes.

This can also be seen in Figure 5.13. Sample 37 has a lower than average success rate of 89% and it is clear from Figure 5.13 that many missing genotypes originate from chromosomes 16, 17, 19 and 22. There are three possible reasons for the greater missing genotypes on chromosomes 16, 17, 19 and 22: i) plate position, ii) OPA specific or iii) position of the failed markers in telomeric or centromeric repetitive sequences which are under-represented by the multiple displacement action (MDA) method (Dean, Hosono et al. 2002). The position of whole genome amplified sample 37 was B9, which was not on the edge on a 96-well plate. All markers for chromosomes 16, 17 and 19 were on OPA 6564 which completely failed for three other individuals, samples 12, 15 and 19. Therefore the OPA did not genotype robustly.

Chromosome	*Number of unknown SNPs in 37 WGA samples*	*Number of SNPs in repetitive regions*
16	37	8
17	33	3
18	16	5
19	44	8
20	15	4
22	20	0
Total	165	28

Table 5.5 The number of failed SNPs in the whole genome amplified DNA sample is not over-represented in genomic repeat regions. The SNPs that did not genotype for sample 37 on the WGA sample from OPA 6564 were investigated to see if the failed genotyping was due to the non-amplification of repetitive regions in WGA samples. However, only a small number of failed SNPs were from genomic repeat regions.

Furthermore, only 17% of the failed markers were located in repetitive regions, as shown in Table 5.5. Also, the missing alleles fall into a wide-range of genomic regions such as repetitive elements and regulatory regions, and do not have a specific function or locus region, as shown in Table 5.6. Thus, the missing genotyping from whole genome amplified sample 37 stems from the poor quality of OPA 6564.

REPETITIVE ELEMENT	*DESCRIPTION*	*# SNPS*
DNA	DNA repeat elements	4
LINE	Long interspersed nuclear elements	12
LTRs	Long terminal repeat elements, including retroposons	3
Satellite	Satellite repeats	1
SINE	Short interspersed nuclear elements (SINEs), which include ALUs	8
Regulatory Regions		
CpG Island	Potential regulatory region	2
Conserved transcription factor binding site (TFBS)	Location & score of TFBS conserved in the human/mouse/rat alignment	2
Conserved region of 8-genome alignment	Human/chimp/mouse/rat/dog/chicken/fugu/zebrafish conserved region from 8-genome multi-alignment	12
Function		
Coding-NonSynon	Change of peptide in contig sequence	3
Coding-Synon	No change of peptide for allele in contig sequence	4
mRNA-UTR	Variation in transcript, but not in coding region interval	22
Locus Region	Variation in region of gene, but not in transcript (within 2kb flanking sequence of the mRNA)	16
Intron	Variation in intron, but not in first 2 or last 2 bases of intron	66
Unknown	No known functional classification	54
		165

Table 5.6 Description of missing alleles in whole genome amplified DNA. The failed SNPs from the whole genome amplified sample (WGA) of 37 were investigated for any specific genomic location or property that may have affected the genotyping failure. The description of the SNPs was obtained from UCSC genome browser (March 2006 version, NCBI Build 36).

5.4.4. Heterozygosity

The average heterozygosity in 5,398 Mendelian consistent markers was 44.6%. Marker heterozygosity was calculated using the descriptive statistics programme Pedstats (Wigginton and Abecasis 2005). This is similar to >43% average heterozygosity in Caucasians published by Illumina (Murray, Oliphant et al. 2004) for the Linkage IVb Panel.

5.4.5. Mendelian inconsistencies

Genotyping errors were detected using Pedcheck, as described in section 2.5.3.4, which identifies all Mendelian inconsistencies in pedigree data (O'Connell and Weeks 1998). Pedcheck inspected data for errors at four different levels. Level 1 checked for simple errors in a nuclear family, such as the alleles of a child and parent that were incompatible, more than four alleles in a sibship, allele out of bounds of any specified range and more than three alleles in a sibship with a homozygous child. Level 2 looked for more subtle errors via a genotype elimination algorithm that, for each untyped individual, constructed the shortest possible list of genotypes that were mutually consistent with all genotypes of their relations. Level 2 identified slight inconsistencies resulting from the elimination of certain genotypes, on the basis of more complex pedigree relations. Level 3 tried to identify the possible error by "untyping" one individual at a time and applied genotype elimination to see if the inconsistency had been eliminated. Level 4 used an odds-ratio algorithm to look at each critical genotype and computed the relative likelihood of different alternative valid genotypes (O'Connell and Weeks 1998). Thirty-eight Level 1 and 2 inconsistencies were found in the pedigree data; that was 27 of the 5663 autosomal SNPs were inconsistent with Mendelian inheritance. No errors were detected by Level 3 and 4 checks. This was a rate of 0.5% Mendelian inconsistencies which was higher than the expected rate of 0.006%, as published by

Illumina. The genotypes contributing to these errors were excluded from the linkage analysis, providing a "cleaned" dataset.

5.4.6. Hardy Weinberg equilibrium testing

Hardy Weinberg disequilibrium can occur for certain reasons including systematic genotyping errors, non-random mating and chance. Therefore, for quality control purposes, it was important to check for deviation from HWE. This was performed using the exact SNP test in the PedStats programme, as described in section 2.5.3.4, on 14 unrelated individuals from the family (individuals 11, 13, 17, 20, 33, 38, 46, 49, 98, 104, 146, 149, 160 and 180). 22 SNPs were excluded as a result of departure from HWE, $P \leq 0.01$ as recommended by the author (Wigginton, Cutler et al. 2005).

5.4.7. Mendelian-consistent genotyping errors

To test the data further for genotyping errors, sensitivity analysis of the likelihood of all genotypes was conducted, using the programme MERLIN, as described in section 2.5.3.4 (Abecasis, Cherny et al. 2002). MERLIN searches for genotypes that provide information about inheritance in a pedigree and that also contradict information provided by other genotyping data. MERLIN considers all of the pedigree data simultaneously (not just that for pairs of individuals) for error detection. MERLIN detected 355 errors in the whole-genome data on the 9 sub-families. These putative erroneous genotypes were omitted. Furthermore, sample 107 proved problematic with an excess of genotyping errors, 13% compared to the average of 7%, detected by Pedcheck and MERLIN. Therefore, this unaffected individual was removed from the linkage analysis.

5.4.8. Linkage disequilibrium between markers

An assumption of most linkage analysis methods is that the markers under investigation are not in linkage disequilibrium (LD). This may not be the case for dense whole genome scans. LD between markers can lead to positive bias in linkage scores when founders are untyped (Josée Dupuis 2007) and if founders are typed, ignoring LD can lead to an inflation of false positive error rate (Yoonhee Kim 2008). For this study, LD between the SNPs on the Illumina Linkage IVb Panel was checked within the family. LD between autosomal markers was investigated in Haploview version 3.2 (Barrett, Fry et al. 2005). The default definition of LD, solid spine LD D'>0.8, for markers 500kb apart was used. Haploview chose 12 singletons and one trio called the "maximum unrelated subset" from the family. These were 12 individuals married-in to the family (samples 11, 13, 20, 38, 46, 49, 98, 104, 146, 149, 160, 180) and one trio (samples 29, 33, 35). There were 456 pair-wise comparisons with D'≥0.8 and LOD≥2. Figure 5.14 shows that the majority of SNPs are not in LD with each (D'<0.8). Figure 5.14 also shows the presence of 456 markers in high D' with each other in close distance <100kb. In each case, one of the two SNPs in pair-wise LD was removed from the analysis. The SNP for removal was chosen on the basis of genotyping success: a lower genotyping error rate and then by lower heterozygosity. For SNPs in a cluster of LD, all but one marker was removed. In total, 455 SNPs were excluded from the linkage analysis.

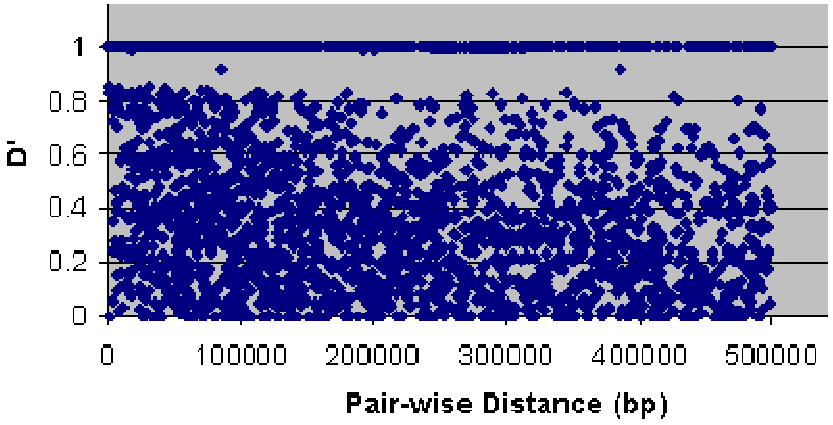


Figure 5.14 Distribution of pairwise D' values according to the pairwise distance of SNPs. The x-axis is the distance between a pair of SNPs in base pairs. The y-axis is the D' value which measures linkage disequilibrium between SNPs, where D'=1 means complete LD.

To summarise, a "cleaned" genotyping dataset was prepared. This excluded many erroneous markers as listed in Table 5.7.

REASON EXCLUDED	NUMBER SNPS EXCLUDED
Call rate per SNP <90%	266
Mendelian errors	27
MERLIN (haplotype)	355 genotypes
High LD with neighbouring SNPs	455
Failed HWE, *P*<0.01	22

Table 5.7 Reduction from 5,663 autosomal SNPs in the Illumina Linkage IVb Panel to 4,893 SNPs used in linkage analyses

5.4.9. Information content

Information content is a measure of the information a panel of markers can provide in a pedigree, which will extract the maximum amount of inheritance information for a linkage analysis. It is a function of marker heterozygosity and the number of meioses in a genetic study. To calculate the information content of the cleaned data, without the erroneous genotypes as shown in Table 5.7, the information content based on the Shannon entropy measure was calculated for all markers, on the sub-families, using MERLIN software. The average information content over all chromosomes was 88% (±0.03 standard deviation) and the minimum value was 69%. This was lower than the published information content for the same marker panel, where the average information content was 97% and the minimum value was 83% (Murray, Oliphant et al. 2004). The information content may be greater if the whole pedigree was taken into account but this was computationally very expensive to calculate.

5.5. *Relatedness Analysis*

5.5.1. Preface

To perform relatedness analysis in the family, three statistical methods were employed. They assessed whether the assumed relationship was likely, and if not, then predict a more likely relationship. The first statistical test determined whether the pattern of allele sharing by each pair of individuals was consistent with the indicated relationship and was performed using "PREST" (Pedigree RElationship Statistical Test) (McPeek and Sun 2000). This method was followed by ALTERTEST (ALTERnative TEST) which determines the probability of a different hypothesised relationship (McPeek and Sun 2000). The third statistical method used a maximum likelihood approach to infer relationships between two individuals based on the genetic marker data provided. For this analysis, the programme "Relpair" was used

(Epstein, Duren et al. 2000). These statistical methods were different approaches to the same question and were chosen to compliment each other.

PREST estimates κ = (κ0, κ1, κ2), the probability of sharing zero, one or two alleles identical by descent (IBD) by a pair of individuals. Related individuals are similar to each other because they have variants that are IBD, which are copies of the same variant from a common ancestor, as illustrated and explained in Figure 1.6. PREST computes the expected IBD (EIBD), adjusted IBS (AIBS), IBS and maximised log-likelihood ratio (MLLR) statistics and performs hypothesis tests for relationship misspecification, considering all relation pairs. The authors of PREST recommend initial screening of genome-wide data using the EIBD and AIBS test, followed by application of the MLLR test on problematic pairs. The EIBD test statistic is the "average of the conditional expected number of alleles shared IBD at each marker, conditional on the data for that marker, the null relationship and the allele frequencies". The AIBS statistic is "an average over all markers m of A_m, where A_m is a sum over each shared allele of its null conditional probability of being shared IBD, given that the allele is shared IBS and under the assumption that the shared alleles results from a random draw of one allele from each of the individuals". The MLLR statistic tests the "null relationship against a specific alternative relationship by calculating the likelihood of the data under the null and alternative relationships specified". The MLLR test is an extension of the likelihood calculations of Goring and Ott (Goring and Ott 1997) and Boehnke and Cox (Boehnke and Cox 1997) to more general relationship pairs, where the IBD test is not based on a Markov chain, but a likelihood-ratio test. ALTERTEST is based on the same statistical tests.

Relpair is based on a maximum likelihood approach; that is to look at a large class of relationship distributions and then chooses the "best" distribution. For each distribution, likelihood is computed, and the best distribution is the one that maximizes this likelihood. This approach infers relationships between two

individuals based on the genetic marker data. This can be done within families, as in this study, or across many pedigrees. The programme calculates and compares the multipoint probability of genetic marker data conditional on different pairwise genetic relationships and infers the relationship that makes the data most likely. Additionally, RELPAIR has two advantages over PREST i) X-linked data can be included to improve distinguishing of second-degree relationships, for example inferring paternal half-sisters and maternal half-brothers (Epstein, Duren et al. 2000) and ii) improves identification of monozygotic twins and parent-offspring pairs by modelling genotyping error when classifying relationships.

5.5.2. Relationship testing results

Statistical tests using PREST and RELPAIR software were performed on the whole-genome linkage data, as described in section 2.5.5, to determine whether the pattern of allele sharing by each pair of individuals was consistent with the indicated relationship provided from the originally published pedigree.

5.5.2.1. PREST results

Pedigree errors were calculated over all relationship pairs known to PREST (parent-offspring, full-siblings, half-siblings, avuncular, first cousins, grandparent-grandchild, half-first-cousins, half-avuncular and unrelated). The results were plotted on a scatter diagram of estimated IBD probabilities κ_0, κ_1 and κ_2 on relationship triangles. An explanation of relationship triangles, first devised by Elizabeth Thompson, is provided in Figure 5.15 (Thompson 2000). The results of all IBD tests are shown in Figure 5.16. For each mis-specified relationship, a MLLR test was performed for simulation using 100,000 replicates for each pair, to assess significance.

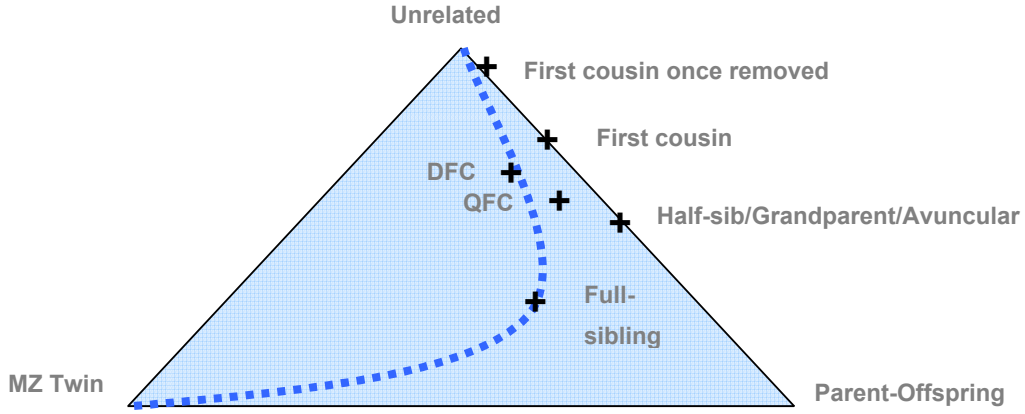


Figure 5.15 The relationship triangle. The relationship triangle depicts the location of the expected results from allele sharing test and was first used to illustrate relationships by Elizabeth Thompson (Thompson 2000). Relationship triangles were used to display results from relationship testing using PREST software. Each vertex of the triangle represents sharing of genes IBD, zero alleles, κ0 =1, for unrelated at the top vertex, one allele, κ1 =1, for parent-offspring on the right vertex and both alleles, κ2 =1, for MZ twin on left vertex. Other levels of sharing are shown by crosses. DFC is double first cousin, QFC is quadruple first cousin. Please refer to section 1.8.4 for further information.

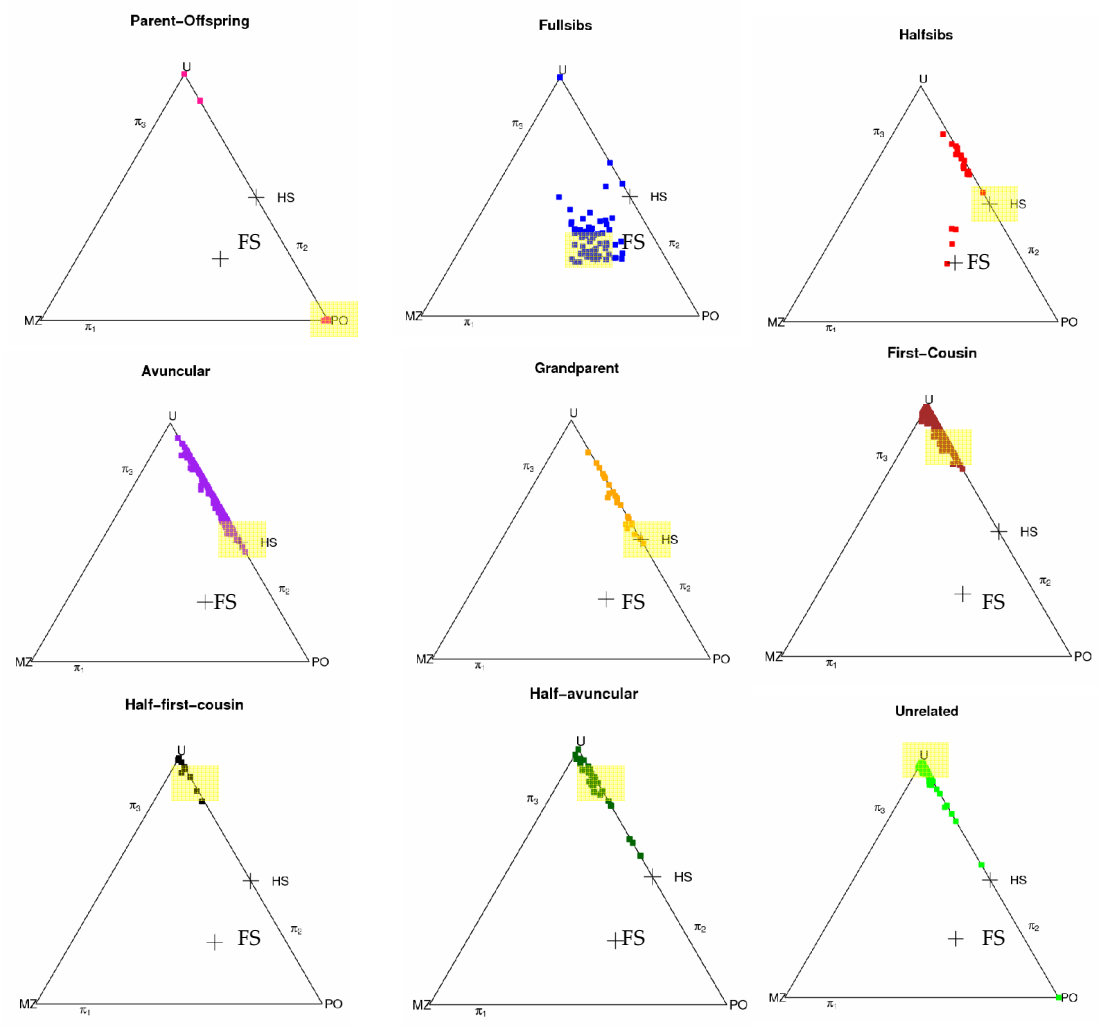
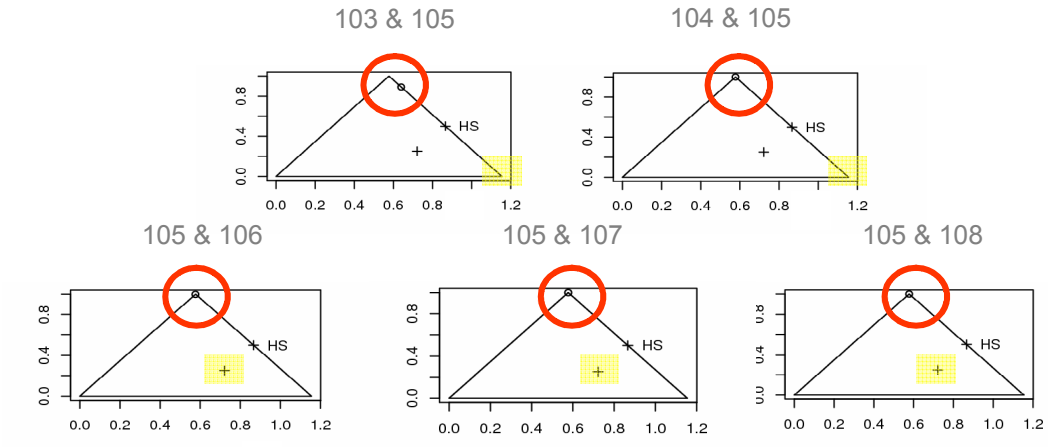


Figure 5.16 Result of relationship analysis based on estimated IBDs using the PREST programme. Each coloured square represents the IBD probability between two individuals. All the tests, for each relationship, are displayed on the relevant relationship triangle. The expected location for each relationship is highlighted in yellow. Deviation from this location suggests an erroneous relationship. There were nine different relationship categories tested in the family. PO = Parent-offspring, FS =Full-sibling, HS = Half-sibling, U = Unrelated, MZ = monozygotic twin, π_0, π_1, π_2 = the probability of sharing zero, one or two alleles IBD by the pair. Please refer to Figure 5.15 for explanation of relationship triangles.

Figure 5.16 shows that the vast majority of relationships fall at the expected location highlighted in yellow in the relationship triangles, thus confirming pedigree structure. The test highlights, however, a number of relationship errors: two parent-offspring errors, three full-sibling errors, four half-sibling errors and two unrelated errors. These errors are explained in detail in the following three figures: Figure 5.17, Figure 5.18 and Figure 5.19.

The PREST results indicate that an incorrect sample was genotyped for sample 105, a duplicate of sample 97 as illustrated in Figure 5.17. Figure 5.17a shows the relationship triangles for each relationship of the nuclear family; parents 103 and 104 and offspring 105, 106, 107 and 108. The actual PREST result, highlighted by the red circle, is not in the same position as the expected result, highlighted in yellow and falls at the "unrelated" position in each case. Figure 5.17b shows the number of markers that were tested for each relationship test, approximately 5,000 SNPs. Figure 5.17c shows the pedigree of the large Scottish family, where the nuclear family, including sample 105 and the position of the duplicated sample 97, can be located within the pedigree.

a)



b)

ID 1	ID 2	Assumed Relationship	Number of Markers	κ0	κ1	κ2
103	105	Parent-offspring	5375	0.892	0.108	0.000
104	105	Parent-offspring	5377	1	0.000	0.000
105	106	Full-sibling	5372	0.998	0.000	0.002
105	107	Full-sibling	4919	0.999	0.000	0.000
105	108	Full-sibling	5363	0.999	0.000	0.000

c)

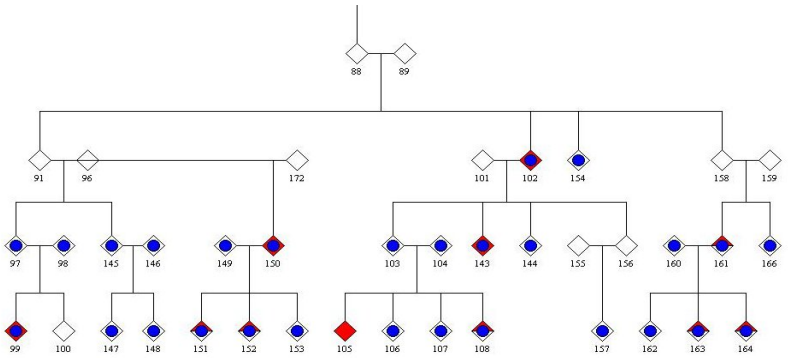
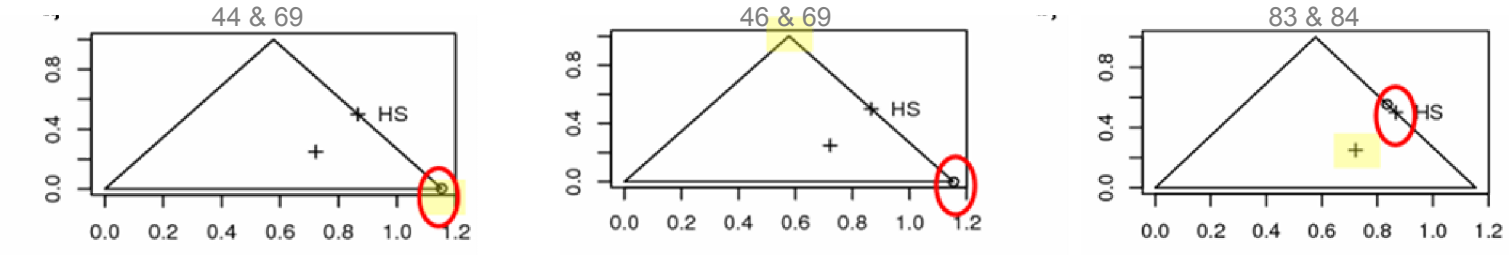


Figure 5.17 Incorrect sample for sample 105 identified by relationship testing in PREST. Analysis by PREST is shown in (a). The expected relationship result of parent-offspring relationships between parents 103, 104 and offspring 105 or full-sibs 106, 107 and 108 are highlighted in yellow, whereas the actual unrelated relationship result from PREST is circled in red. The PREST results used to prepare the relationship triangles are tabulated in (b). κ0, κ1, κ2 is the probability of sharing 0, 1 or 2 alleles IBD result for the specified relationship. The pedigree is a snapshot of the full pedigree showing the original relationship information.

Figure 5.18 illustrates the resolution of paternity in a trio, samples 44, 46, and 69. The initial pedigree structure listed the parents of sample 69 as samples 46 and 173. However, the relationship test results with over 5,000 markers, as shown in the relationship triangles in Figure 5.18a, shows that the parents of sample 69 are in fact, samples 44 and 46 (Figure 5.18b). Furthermore, Figure 5.18 demonstrates relationship testing of samples 83 and 84 and shows that they are not full-siblings as originally assumed, but half-siblings. However, Figure 5.18b shows the number of markers that deduced this relationship was 2,521, only half the number of markers used for testing relationships between all other individuals. This was a consequence of sample 83 failing to genotype with 50% of the markers, and thus all genotypes were set to unknown for this sample and the relationship error was not taken into account.

a)



b)

ID1	ID2	Assumed Relationship	Number of Markers	K 0	K 1	K 2
44	69	Parent-offspring	5253	0.004	0.996	0
46	69	Unrelated	5348	0	1	0
83	84	Full-siblings	2521	0.554	0.446	0

c)

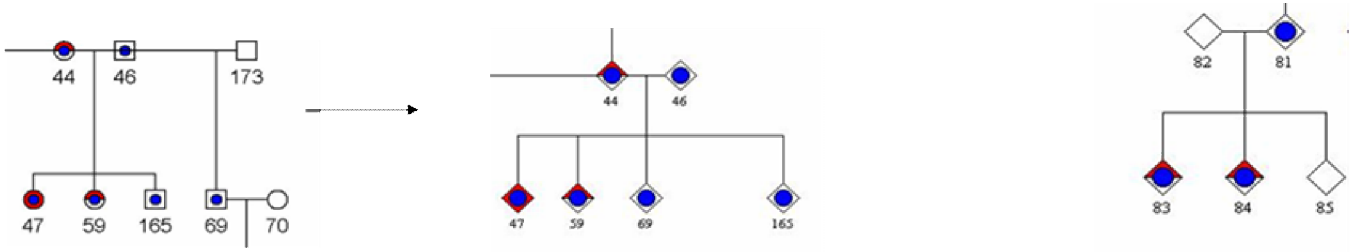
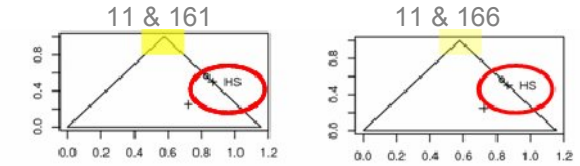


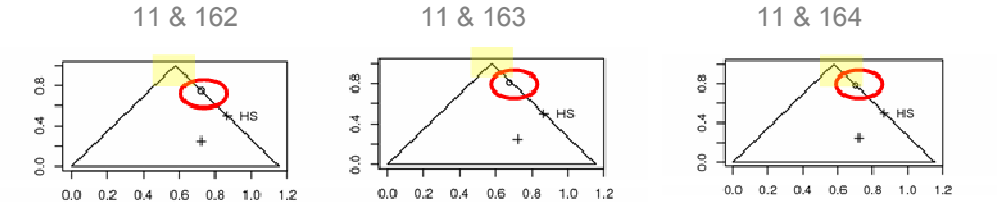
Figure 5.18 Non-paternity resolved. Relationship testing results (a) show that 69 is related to both 44 and 46 at the parent-offspring level circled in red at the bottom right vertex in (a). This differs from putative pedigree (c) where 46 is not a parent. Relationship results also suggest that 83 and 84 are not full sibs. However, sample 83 failed for 50% genotyping so all genotyping was set to unknown. The expected results for each pair of individuals are highlighted in yellow and the actual results are circled in red in the relationship triangles (a). The PREST results are tabulated in (b) κ0, κ1, κ2 is the probability of sharing 0, 1 or 2 alleles IBD result for the specified relationship and number of markers tested. The pedigree is a snapshot of the full pedigree showing the original relationship information and the altered pedigree. Blue circles show the individuals genotyped for this linkage study.

Figure 5.19 demonstrates that sample 11, an assumed unrelated individual, married into the family, may be related to sample 161. The relationship triangles in Figure 5.19a present the PREST results for sample 11 with the assumed unrelated nuclear family, samples 161, 162, 163, 164 and 166, from analysis on over 5,000 markers in Figure 5.19. The results circled in red, deviate from the expected location highlighted in yellow on the relationship triangles. Thus, revealing excess sharing in assumed unrelated individuals (11 and 161), as denoted by the blue circles in Figure 5.19c.

a)



b)



c)

ID1	ID2	Assumed Relationship	Number of Markers	κ0	κ1	κ2
11	161	Unrelated	5352	0.565	0.435	0.000
11	166	Unrelated	5038	0.564	0.436	0.000
11	162	Unrelated	5344	0.75	0.25	0.000
11	163	Unrelated	5351	0.809	0.181	0.010
11	164	Unrelated	5351	0.781	0.219	0.000

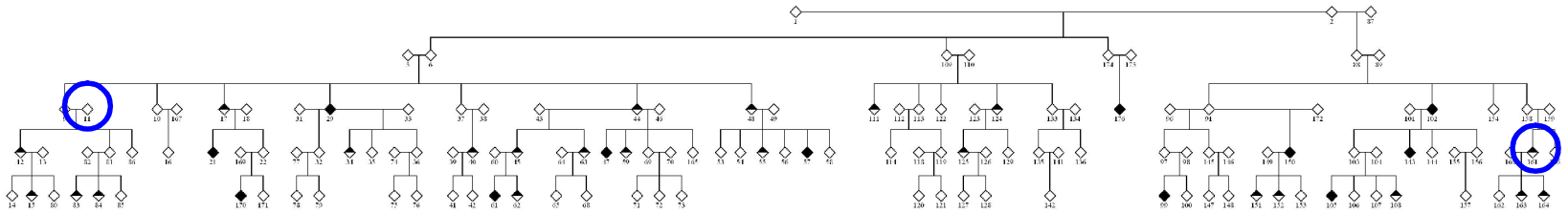


Figure 5.19 Excess sharing detected between unrelated individuals. Analysis by PREST is illustrated in relationship triangles (a).The expected relationship result for the two individuals is highlighted in yellow, whereas the actual relationship result is circled in red. The PREST results are tabulated in (b), κ0, κ1, κ2 is the probability of sharing 0, 1 or 2 alleles IBD result for the specified relationship. The original pedigree is shown with the position of samples 11 and 161 highlighted by blue circles, to the left and right respectively.

5.5.2.2. ALTERTEST results

Alternative relationships were indicated on the relationship triangles for the problematic relationships detected by PREST. These alternative relationships were tested as a null hypothesis using ALTERTEST, performing the EIBD, AIBS and IBS tests on problematic relationship pairs, to see if the alternative was compatible with the data. The results are shown in Table 5.8 and confirm the newly hypothesised relationships. Firstly, samples 97 and 105 share markers to the same degree as an alternative relationship of monozyogotic twins, indicating sample duplication. This was also confirmed by testing sample 105 (duplicated sample 97) with family members of sample 97, where the relationships concur with a sample duplication of 97. Secondly, non-paternity was resolved for parents 44 and 46 and offspring 69. Again, this relationship was confirmed, by alternative relationship testing the offspring of samples 44 and 46 (sample 47, 59 and 165), to show they are full-siblings and not half-siblings. The second case of non-paternity was investigated, by estimating the alternative relationship of half-siblings for samples 83 and 84. Although the number of markers is less than tested for the other alternative relationships, a relationship at the half-sibling level is disclosed. Thirdly, excess sharing was also confirmed by the alternative relationship of half-sibling, grandparent or avuncular, between the assumed unrelated samples 11 and 161, and 11 and 166.

Error			Null Hypothesis	Number of		Estimated		
	ID1	ID2	Relationship	Markers	EIBD	κ0	κ1	κ2
Sample mix-up	97	105	Monozygotic twin	5371	NA	0	0	1
	105	150	Full-sibling	5380	0.8782	0.6431	0.3568	0
	97	145	Full-sibling	5360	0.9808	0.3302	0.449	0.2208
	97	150	Half-sibling	5380	0.4906	0.6419	0.3581	0
Non-paternity	44	69	Parent-offspring	5253	NA	0.0037	0.9963	0
	46	69	Parent-offspring	5348	NA	0	1	0
	47	69	Full-sibling	5355	0.9617	0.3909	0.4314	0.1778
	59	69	Full-sibling	5367	1.0082	0.2471	0.4724	0.2805
	165	69	Full-sibling	5367	0.9641	0.3956	0.4135	0.1909
Non-paternity	83	84	Half-sibling	2521	0.4963	0.5536	0.4464	0
Excess sharing	11	161	Half-sibling	5352	0.4952	0.5646	0.4354	0
	11	166	Half-sibling	5038	0.4955	0.5636	0.4362	0.0002

Table 5.8 Relationship errors detected using the ALTERTEST. The results from testing alternative relationships between individuals where relationship errors were detected by PREST analysis. The test hypothesises the likelihood of a null hypothesis relationship, indicating the estimated IBD (EIBD) value and the estimated probability of sharing 0, 1 or 2 alleles by descent (κ0, κ1, κ2). Please refer to Table 1.2 for theoretical IBD values. NA = Non Applicable.

5.5.2.3. Relpair results

Relpair was used as an alternative method to PREST and helped confirm the results. Due to memory constraints in the use of Relpair, the test was performed on chromosomes 1, 13, 16 and X with 489, 191, 197 and 301 SNPs respectively. The Relpair limit for the number of markers was initially set at 200, thus the choice of chromosomes 13 and 16. Altering the limit constraints, as shown in Table 2.11, allowed the use of chromosome 1 and more importantly the X chromosome, which can help define second-degree relationships. The results were deemed noteworthy if they were confirmed for all four chromosomes under study. The critical value parameter was set to 1000, which is recommended for large data sets. This value reports the likelihood ratio between putative and inferred relationships (INF/PUT). The likelihood ratio is the ratio of two probabilities for the same observations, calculated under alternative hypotheses. In the context of relatedness analysis, the likelihood ratio is formed by dividing the probability of the observed pair of genotypes using the IBD probabilities for one possible relationship by the probability of the genotypes using IBD probabilities for the other possible relationships. The likelihood ratio is a continuous variable that can take any non-negative value and values greater than one support the relationship used for the numerator (Weir, Anderson et al. 2006). Only the results for close relationships of parent-offspring, full-siblings and half-siblings were considered, as this method is not reliable for inferring distantly related individuals. Examining all pairs across a large dataset, the great majority of pairs will be putatively unrelated, so the false positive rate is high (Epstein, Duren et al. 2000).

As shown in Table 5.9, Relpair inferred a monozygotic twin relationship between sample 97 and 105 at INF/PUT ratio of $>10^6$ for chromosome 1, 13, 16 and X. This was confirmed with results from the rest of the family, where sample 105 was unrelated to its putative parents 103 and 104, but related to sample 97. It also showed that the parents of sample 69 are samples 44 and 46. Furthermore it showed

that sample 11 was related to samples 161, 162, 163, 164, 166 to some degree. Thus the results from all three statistical methods using the PREST, ALTERTEST and Relpair programmes were in agreement.

ERROR	ID 1	ID 2	PUTATIVE RELATIONSHIP	INFERRED RELATIONSHIP	INF/PUT RATIO
Sample duplication	97	105	Unrelated	MZ twin	$>10^6$
	99	105	Unrelated	Parent offspring	$>10^6$
	103	105	Parent offspring	Grandparent	$>10^6$
	104	105	Parent offspring	Unrelated	$>10^6$
Non paternity	46	69	Unrelated	Parent offspring	$>10^6$
	47	69	Half-sib	Full-sib	$>10^6$
	59	69	Half-sib	Full-sib	$>10^6$
	69	165	Half-sib	Full-sib	$>10^6$
Excess sharing in unrelated individuals	11	161	Unrelated	AV/HS/CO	$>0.7x10^6$
	11	162	Unrelated	AV/HS	$>10^6$
	11	166	Unrelated	GG/CO	$>10^6$

Table 5.9 Relationship testing results using Relpair software. Relpair detected three relationship errors in the putative pedigree: a sample duplication, non-paternity and excess sharing in unrelated individuals. Furthermore, Relpair inferred a relationship from the level of IBD sharing in the genotype data. AV is avuncular, HS is half-sibling, CO is cousin and GG is grandparent, grandchild relationship. The results for Relpair analysis inferring a relationship between a pair of individuals is reported as a likelihood ratio of inferred relationships compared to the putative relationship (INF/PUT Ratio). Each value above is the average result from markers on all chromosomes analysed (chromosome 1, 13, 16 and X).

The results of the three statistical methods were summarised in Table 5.10. For each relationship error that was detected, the consequence to the pedigree structure was noted.

Error No	ID 1	ID 2	Putative Relationship	Prest Results	ALTERTEST Result	Relpair Results (Chromosome)	Consequence
1	11	161 & 166	Unrelated	Autosomes (permuted)	Confirmed	1, 13, 16, X	No conclusive information
2	83	84	Full sibs	Autosomes (permuted)	Confirmed	13, X	Set genotyping to zero for 83 as failed >50% genotyping
3	46	69	Unrelated	Autosomes	Confirmed	1, 13, 16, X	Pedigree change - Parent
4	105	103	Parent Offspring	Autosomes (permuted)	NA	13, 16	Set genotyping to zero for 105 as duplicate sample
5	97	105	MZ twin	NA	Confirmed	1, 13, 16, X	
6	97	145	Half-sibs	Autosomes	Confirmed	1, 16, X	Pedigree change - Full-sibs
7	97	150	Full-sibs	Autosomes (permuted)	Confirmed	None	Pedigree change - Half sibs

Table 5.10 Summary of Pedigree Errors. There were seven pedigree errors confirmed by three different methods of relationship testing. The putative relationship according to the original pedigree is stated for the individuals in question. The result of relationship test for each putative relationship is shown from PREST, ALTERTEST and Relpair. Those results that withstood permutation testing in PREST are indicated (permutated) in the PREST result column. The consequence to the construction of the pedigree is noted. NA=Non-Applicable.

The appropriate adjustments were made to the pedigree according to Table 5.11. For samples that failed genotyping or a sample error occurred as described in section 5.4, the genotyping information was labelled as unknown. For samples that were shown to deviate from their assumed relationships, the pedigree structure was adjusted.

NUMBER	PEDIGREE STRUCTURE ADJUSTMENT
1	Genotyping for 25, 35, 40, 83, 105 set to unknown
2	The parents of 69 are 46 & 44
3	173 is removed
4	Parents of 97 & 145 are 91 & 96
5	Parents of 150 are 91 & 172

Table 5.11 List of adjustments to pedigree. There were five categories of adjustments made to the pedigree data for linkage analysis.

After correction of the pedigree, both PREST and Relpair were rerun to confirm that family relationships reassignments were consistent with estimated IBD patterns and no other relationship errors were conclusively detected. In brief, all three methods were successful in identifying and resolving several pedigree errors.

There were a number of repercussions from the error checking to the linkage analysis. First, samples with a bipolar disorder diagnosis (105) and with recurrent major depression diagnoses (40 and 83) failed genotyping and were removed from the linkage analysis. A sample duplication of 97 was identified, which was originally thought to be 105. Upon further investigation in the lab, previous tube mislabelling was the source of the error. This loss of information for 105 was unfortunate and the power to detect linkage using the genome screen will be reduced by omission of this individual. Second, non-paternity of 69 was resolved

and the trio of 44, 46 and 69 was resolved from analysis of the whole genome data. The previous finding of non-paternity of sample 69 was due to genotyping of the wrong sample prior to this study. Last, there was a suggestion of excess allele sharing between apparently unrelated sample 11 and 161. It was important to note that the nature of the relationship of 11 to 166 and other samples could not be deciphered. Also, a clear answer was not obtained when the analysis was rerun with the adjustments made to the pedigree. The statistical methods used are not capable of accurately determining distant relationships, so the pedigree was left unchanged. However, this relationship between both these unaffected individuals was borne in mind when examining the results of the linkage analysis in chapter 6.

5.6. *Determination of Significance Thresholds*

5.6.1. Introduction

To determine the criteria for suggestive and significant linkage, simulations of the dataset was performed. Genome-wide significant levels were estimated by simulating replicates of the genotype of each marker on each autosomal chromosome, while maintaining the actual sub-pedigree structures, phenotypes, allele frequencies, marker maps and missing data structure. There were 10,000 replicates performed for the broad phenotype and 1,000 replicates for the narrow phenotype. The marker data was simulated under the null hypothesis of no linkage to observed phenotypes. Each replicate was subject to non-parametric linkage analysis in the same way as the real data. For parametric linkage analysis, the true disease model was unknown, so simulations based upon the assumed disease model may not provide a true reflection of the null distribution. This, together with computational limitations and a redundancy between methods, led to the decision to perform linkage analysis on the replicated datasets using non-parametric methods, in order to determine significance thresholds.

5.6.2. Result for significance thresholds

The peak non-parametric LOD scores from each replicate were noted for each whole genome scan and their distribution analysed. The distribution for all 22 autosomes is shown in Figure 5.20.

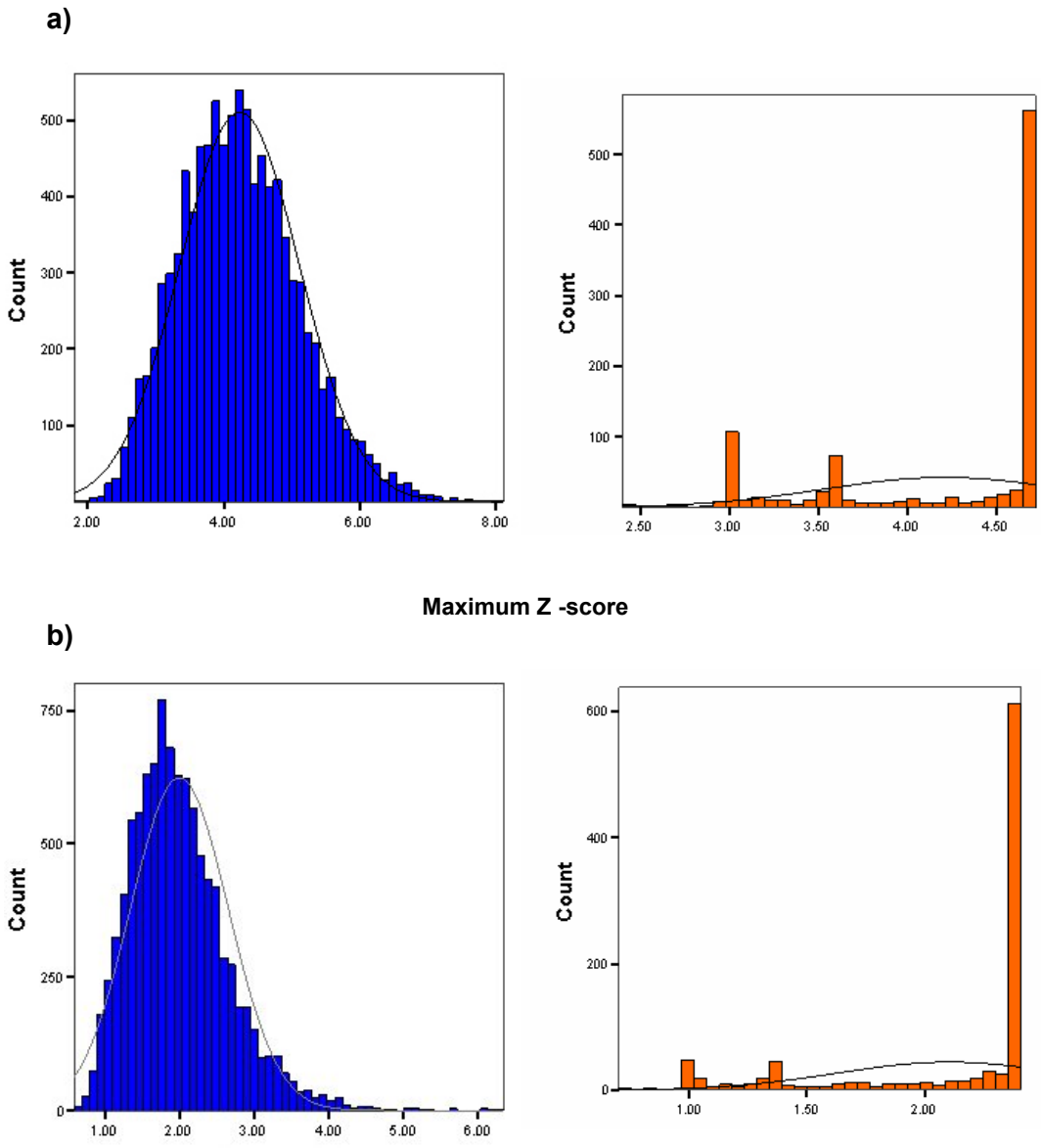


Figure 5.20 Distribution of highest Z and LOD scores for each replicate for simulation of non-parametric linkage analysis on broad and narrow phenotypic model. Genotype data was simulated 10,000 times for the broad phenotype (blue) and 1,000 times for the narrow phenotype (orange) for the sub-pedigrees. Non-parametric linkage analysis was performed on the data. For each whole genome scan, the highest Z (a) and LOD (b) scores were retrieved. The distribution of the Z and LOD scores are shown, where the x-axis is the Z score (a) and LOD score (b) and the y-axis is the counts of the scores. The normal distribution is shown by the curved line.

These distributions were used to determine the threshold at which a finding is observed once in every genome scan which is the mean, once in every 20 genome scans, which is the 95% level and once in every 1000 genome scans for the more stringent 99.9% level. These are the criteria for suggestive, significant or highly significant linkage respectively.

STATISTIC	BROAD		NARROW	
	Z-score	LOD	Z-score	LOD
Mean	4.23	2	4.21	2.11
Std. Error of Mean	0.0088	0.00657	0.02	0.02
95 Percentile	5.79	3.21	4.72	2.41
99.9 Percentile	7.55	5.32	4.72	2.41
Coefficient of Skewness	0.428	1.019	-0.905	-1.34
Std. Error of Skewness	0.024	0.024	0.077	0.077
Kurtosis	0.182	2.234	-0.806	0.238
Std. Error of Kurtosis	0.049	0.049	0.155	0.155

Table 5.12 Statistics of simulated linkage analysis. The relevant statistics on non-parametric linkage analyses on the broad and narrow phenotype models simulated datasets.

It was important that the data should have a normal distribution, as many statistical methods make this assumption. The shape and symmetry of the simulated data was characterised by comparing the shape of the observed frequency distribution with that of the normal distribution, as shown in Figure 5.20. Two measures are used to assess departures from normality: skewness and kurtosis, which are explained in full in Appendix A.

The whole genome simulation results followed a normal distribution. Figure 5.20 showed that the Z-score for the broad phenotype model follows the normal

distribution in shape and symmetry with coefficients of skewness and kurtosis close to zero. The corresponding maximum Kong and Cox exponential LOD score for the broad phenotype was slightly positively skewed (coefficient of skewness = 1.019) with a long right tail (coefficient of kurtosis = 2.234), but fits the normal distribution more than any other distribution, such as X^2 for example. However, it is clear from the histogram that the simulation data for the narrow phenotypic model with both scores, Z- and LOD respectively, was negatively skewed with a long left tail (coefficient of skewness = -0.9, -1.3) and a longer tail than the normal distribution (coefficient of kurtosis = -0.8, 0.2). This negatively skewed distribution, with no values less than zero, follows a X^2 distribution; X^2 test 489.8, degrees of freedom 259, asymptotic *P*-value 0.000, which was expected for LOD scores when N is large (Lander and Kruglyak 1995). The reason for this was the power available in the narrow phenotypic model was limited to a Z-score 4.72 and LOD 2.41, as discussed in section 6.2.2. Therefore, the data from the narrow phenotypic model did not extend on the right tail, limited at LOD 2.41 and Z-score 4.72 and was hence, not ideal for confidently measuring highly significance thresholds.

The thresholds for suggestive, significant and highly significant for the broad model are 2, 3.21, and 5.32 respectively and for the narrow model are 2.11, 2.41 and 2.41 respectively. The threshold levels for significance were as expected and comply with the gold-standard Lander and Krugylak suggestions, as shown in Table 1.8. The simulation result for the broad phenotype model for whole genome analysis followed a normal distribution. The other simulation results for the narrow model should be regarded with caution as their distributions are slightly skewed.

Empirical *P*-values for each reported non-parametric LOD score, under the broad and narrow phenotypic model, were calculated from this data. These *P*-values were computed by dividing the number of replicates that exceeded LOD score by the

number of replicates. This allowed examination of the LOD scores in context with the whole genome scan.

5.7. *Discussion*

The examination of data on 6,008 SNPs for 96 DNA samples led to some interesting observations. The whole genome amplified DNA was consistent with its genomic DNA duplicate, thus ensuring confidence in the other amplified DNA sample. The metrics of the genotyping success were acceptable, although less than the published Illumina success rates (Murray, Oliphant et al. 2004). The failures in genotyping can be attributed to poor quality DNA samples and emphasises the importance of high-quality, intact DNA to generate the best results. Also, information content analysis showed that there was adequate information in the "cleaned" data to detect linkage. In addition, the confirmation of the pedigree structure based on relatedness analysis ensures an accurate platform for linkage analysis. Splitting of the pedigree simplified the error-checking process and the linkage analysis. The LOD scores required for significant linkage were derived empirically for non-parametric linkage analysis and were in keeping with current literature recommendations. It was hoped that these efforts would maximise the power to detect linkage, while not discarding any useful information that may help linkage detection.

Chapter 6
Results of Linkage Analysis

6. Results of Linkage Analysis

6.1. *Preface*

6.1.1. Background

Previous linkage analysis performed on this large Scottish family showed evidence for a bipolar disorder locus on chromosome 4p15-p16 with a maximum LOD score of 4 (Blackwood, He et al. 1996). The initial linkage analysis was subsequently updated with additional chromosome 4 markers and family members confirming linkage with a maximum LOD score of 4.4 (Le Hellard, Lee et al. 2007). The first linkage analysis employed 193 markers that were unevenly spaced across the whole genome, omitting large proportions of chromosomes. The second linkage analysis used 24 markers on chromosome 4p. As these initial linkage studies had incomplete coverage of the genome, a re-evaluation of the linkage evidence was deemed important. This was timely for two reasons i) the recruitment of extra family members, including some affected individuals without the chromosome 4p15-p16 linked haplotype and ii) new technology that enables dense coverage of the whole genome. Here, I describe whole-genome linkage analysis performed on this family, to reassess the chromosome 4p linkage and to test for other genetic loci co-segregating with bipolar disorder and recurrent major depression.

There are several issues that determine the approach to linkage studies of bipolar disorder and recurrent major depression. Firstly, the mode of inheritance is unknown. The analyses should therefore test both parametric (to examine dominant and recessive inheritance) and non-parametric models. Secondly, psychiatric symptoms are assessed by a psychiatrist during a diagnostic interview and a phenotype reached according to the DSM-IV criteria. The lack of diagnostic biological markers means that there is always a subjective element regarding the diagnosis. Also, the overlap between bipolar disorder and recurrent major

depression is unclear. Some patients with depression will develop mania symptoms at a future date and will change diagnosis. Others will have only brief episodes of mania, insufficient to trigger a final diagnosis of bipolar disorder. Therefore linkage analysis was performed under two definitions of phenotype, a narrow model of bipolar disorder and a broad model that includes recurrent major depression. Thirdly, the size of this family and the number of SNP markers used was computationally demanding and only certain linkage programmes could be used. There were two options available; i) analysing the whole family for a few markers using programmes such as Simwalk2 (Sobel and Lange 1996) and Fastlink (Cottingham, Idury et al. 1993) or ii) splitting the family to analyse all markers using the programme MERLIN (Abecasis, Cherny et al. 2002). The attributes of each linkage analysis programme are detailed in section 1.4.3. As the aim of this study was to scan the whole genome for linkage peaks, the latter option was chosen. The family was split, using an informative method performed by the programme GREFFA (Falchi, Forabosco et al. 2004) that extracted the most informative sub-pedigrees for linkage analysis. These sub-pedigrees were then tested for linkage to bipolar disorder and recurrent major depression by a whole genome scan. As LOD scores are additive, linkage evidence from each sub-pedigree was added to provide a LOD score for the whole family. Phenotype information for individuals that overlap between the sub-pedigrees was included in one sub-pedigree and noted as unknown in the other sub-pedigrees, to avoid bias. Thus, there were several layers of complexity to this linkage analysis, with many obstacles circumvented and assumptions made, to increase the power to detect genetic loci for bipolar disorder and recurrent major depression.

Here I report two methods of linkage analysis, parametric and non-parametric for two phenotypic models. For parametric linkage analysis, a genetic model was defined. As the misspecification of the model can lead to a loss of power to detect linkage, non-parametric linkage analysis was also performed to test for increased marker sharing between affected relations. Suggestive regions of linkage to

psychiatric illness are described. Haplotype analysis on these regions is also performed.

6.1.2. Description of phenotypic models

Linkage analysis was performed under two definitions of the disease phenotype. The "narrow" model included only individuals diagnosed with bipolar I disorder or bipolar II disorder as affected individuals. Individuals with recurrent major depression and other minor psychiatric diagnoses were labelled as unknown. The "broad" model included individuals diagnosed with bipolar I disorder, bipolar II disorder and recurrent major depression. Individuals with other psychiatric diagnoses or an uncertain diagnosis were labelled as unknown. A description of the phenotypes is provided in section 1.1. Branches of the family with evidence for bilineal inheritance of psychiatric illness, as assessed by a psychiatrist through interview and family history, were also labelled as unknown.

Division of the family meant that only certain sub-pedigrees were informative for linkage analysis. For the narrow model, the two sub-pedigrees with the greatest information were sub-pedigree 1 with four affected individuals and sub-pedigree 4 with three affected individuals. There was one affected individual in each of sub-pedigrees 2, 6 and 7. For the broad model, the more informative sub-pedigrees were sub-pedigree 1, with ten affected individuals; sub-pedigree 2, with four affected individuals; sub-pedigree 3, with two affected individuals; sub-pedigree 4, with five affected individuals; sub-pedigree 5, with four affected individuals; sub-pedigree 6, with two affected individuals and sub-pedigree 9, with three affected individuals. Sub-pedigree 8 has one affected individual.

An autosomal dominant mode of inheritance was primarily considered for the family because visual inspection of the segregation of the illness did not fit with the features of recessive or X-linked inheritance. However, recessive analysis was

performed as a precautionary measure (Hodge, Durner et al. 1993) as misspecification of a dominant disease model has the greatest effect to reduce power to detect linkage (Clerget-Darpoux, Bonaiti-Pellie et al. 1986).

6.1.3. Genetic marker information

Linkage analysis requires an accurate knowledge of marker allele frequencies in the population. The allele frequencies were estimated from the family itself, as the family was large and all members were from the Scottish population. The allele frequencies were determined by counting the alleles from every genotyped individual in the family, using the RECODE facility provided by the MEGA2 programme, as described in section 2.5.3.3 (Mukhopadhyay, Almasy et al. 2005). Linkage was also analysed using allele frequencies derived from the CEPH pedigrees and under a model of equal allele frequencies, for a marker with n alleles, all frequencies were 1/n. The genetic position of the markers was obtained from a high-resolution sex-averaged published map by deCODE genetics (Kong, Gudbjartsson et al. 2002).

6.1.4. Parametric linkage

The parameters for parametric linkage analysis under the narrow and broad phenotypic model were the same as previously published (Blackwood, He et al. 1996; Le Hellard, Lee et al. 2007). The narrow phenotypic model details inheritance for bipolar disorder and the broad phenotypic model details inheritance for bipolar disorder and recurrent major depression. The models are detailed in full in Appendix B, explained in section 1.4.1.2, and important aspects are discussed below.

As the true disease model is not known for either bipolar disorder or recurrent major depression, the model choice was based on a number of assumptions from population genetic data. The disease allele frequency refers to the population frequency of the allele that causes the psychiatric illness. This was not known for

certain and the aim was to make assumptions about a model that best fit the population genetic data. Thus, the values were based on the assumption of a lifetime risk of 0.005-0.01 for bipolar disorder and 0.05-0.1 for recurrent major depression in the general population (McGuffin, Owen et al. 2002). The disease allele frequency was assumed to be 0.007 for the narrow model, which was midpoint of the lifetime risk for bipolar disorder, and 0.03 for the broad model, which takes into account the lifetime risk for both bipolar disorder and recurrent major depression.

Each individual in good mental health was assigned one of four age dependent liability classes; <20 years, 20-30 years, 31-40 years and >40 years. These liability classes were used to specify variation in disease risk, with respect to age. The age-dependent penetrance values for each the liability class were 4.6%, 19.5%, 23% and 29% for the narrow model and 25%, 39%, 61% and 75% for the broad model. These were calculated for the original study from the family itself, according to established principles described by Jurg Ott (Ott 1999). The use of penetrance measures allowed for reduced penetrance, where some mutation carriers who do not have a psychiatric diagnosis may develop a psychiatric illness post-interview, due to late age-of-onset for bipolar disorder. As explained in section 1.4.1.2, the penetrance vectors for disease allele heterozygotes and homozygotes were equal when a model of autosomal dominant inheritance was assumed.

As linkage tests for one putative disease locus, disease penetrance attributed to any other locus or sporadic mutation must be accounted for with a phenocopy rate. The phenocopy rate is the rate at which affected individuals without a mutation in the region of linkage, occur in the sample. The phenocopy rate relates to population prevalence. The phenocopy rates from the original study were used; a rate of 0.1 for the narrow model and 0.5 for the broad model. These values were calculated according to principles described by Jurg Ott and Joseph Terwilliger (Terwilliger and Ott 1994; Ott 1999).

In summary, the genetic parameters for the parametric linkage analysis were taken from the original study. The calculations are based on the following assumptions i) that the population lifetime risk of 0.005 for bipolar disorder and 0.05 for recurrent major depression ii) age dependent penetrances for both the disease susceptibility genotypes and iii) phenocopy rates of 10% for the narrow model and 50% for the broad model. However, it must be reiterated that this calculated model was based on many assumptions and may not reflect the true nature of psychiatric illness.

Linkage is measured by the LOD score method described in section 1.4. Here, the score reported is the multipoint LOD score for each marker, as described in section 1.4.4, with the corresponding α value which is the proportion families linked to the marker and the HLOD, which is the heterogeneous LOD, taking genetic heterogeneity into account. However, as this study investigated bipolar disorder in a single family, the LOD will be reported in graph form. The tabulated results show the LOD, α and HLOD for interest.

6.1.5. Non-parametric linkage analysis

As previously described in section 1.4.2, the aim of the non-parametric linkage analysis was to identify the disease locus by identifying markers where affected individuals are more alike than expected by chance. In this study, the S_{all} sharing statistic was reported. This favoured the sharing of a single allele by a large number of affected individuals, as hypothesised for the family described here. To evaluate this sharing, the Kong and Cox δ (delta) parameter with the exponential model was used. The exponential model was designed to identify a large increase in allele sharing in a small number of families and is a better test if a large increase in allele sharing among affected individuals is expected, but is more computationally intensive (Abecasis, Cherny et al. 2002). The Kong and Cox exponential model

approach generates more accurate *P*-values than those obtained by other methods, and thus was reported here.

6.1.6. Significance thresholds for non-parametric linkage

The thresholds for suggestive and significant linkage were determined by simulating the broad dataset 10,000 times and the narrow dataset 1,000 times and analyzing the data by non-parametric linkage analysis. The threshold levels are reproduced in Table 6.1.

Phenotype Model	Significance Criteria	Calculation	Z-score	LOD (Kong & Cox Exponential)
Broad	Suggestive	Mean	4.23	2
Broad	Significant	0.05	5.79	3.21
Broad	Highly significant	0.001	7.55	5.32
Narrow	Suggestive	Mean	4.21	2.11
Narrow	Significant	0.05	4.72	2.41
Narrow	Highly significant	0.001	4.72	2.41

Table 6.1 Significance levels for whole genome linkage analysis. These are the levels required for significant linkage analysis in the broad and narrow phenotypic model under non-parametric linkage analysis. They are calculated from the distribution of LOD scores from linkage analysis for 10,000 simulated genotyping datasets on the broad model and 1,000 simulated genotyping datasets on the narrow model, as shown in section 5.6.

6.2. *Results of Whole-Genome Linkage Analysis*

6.2.1. Parametric linkage results

The family was split to accommodate whole-genome linkage analysis, as described in section 5.2.2. The family sub-pedigrees were tested for linkage across all autosomal markers in both the broad and narrow model using the quality controlled dataset, as discussed in chapter 5. The marker list, genetic position, physical position and allele frequencies are available from Illumina (http://www.illumina.com/pages.ilmn?ID=191). Parametric linkage analysis was performed, as defined by the genetic model, Appendix B, using MERLIN software, described in section 2.5.7. Each of the 22 autosomes was scanned for linkage. The results of linkage scan are shown in Figure 6.1 and Figure 6.2. The majority of chromosomes show negative linkage (LOD <-2) or inconclusive linkage (LOD~0).

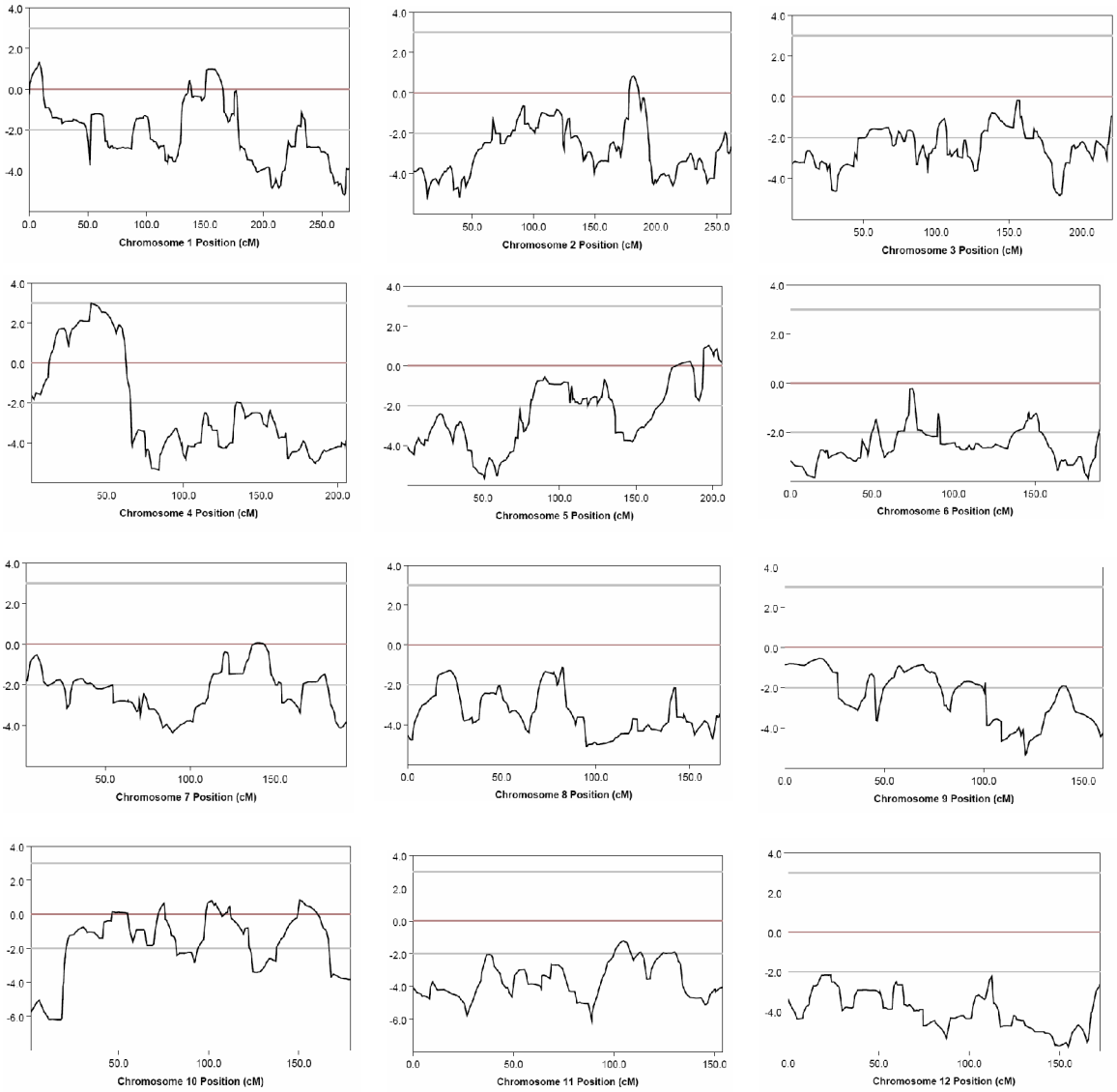


Figure 6.1 Results of parametric linkage analysis under broad phenotypic model. Continued on the next page.

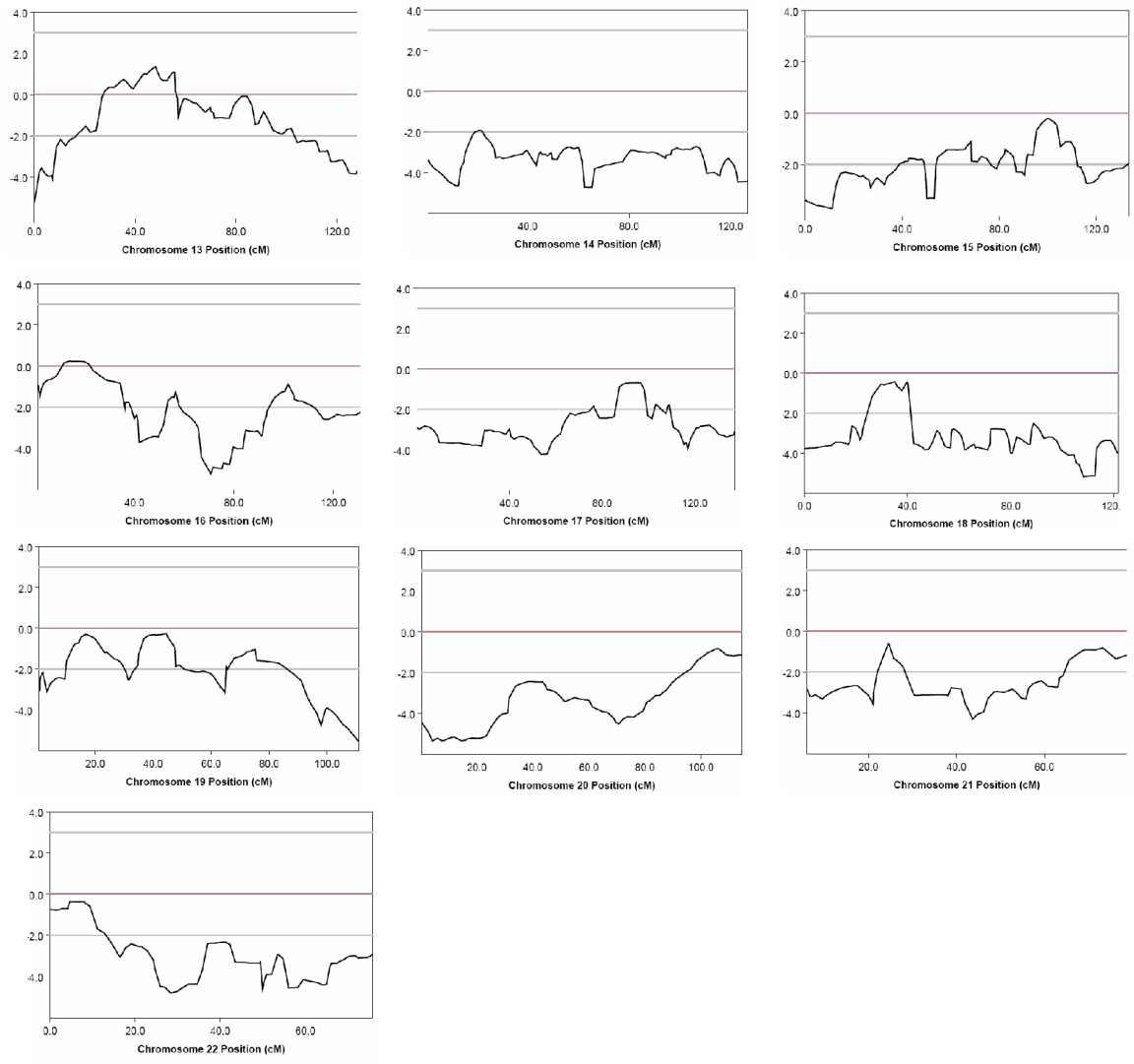


Figure 6.1 Results of parametric linkage analysis under broad phenotypic model. The x-axis shows the chromosome positions according to deCODE cM. The y-axis shows the LOD score. The grey lines show a threshold for significant linkage, LOD=3, and for negative linkage, LOD=-2.

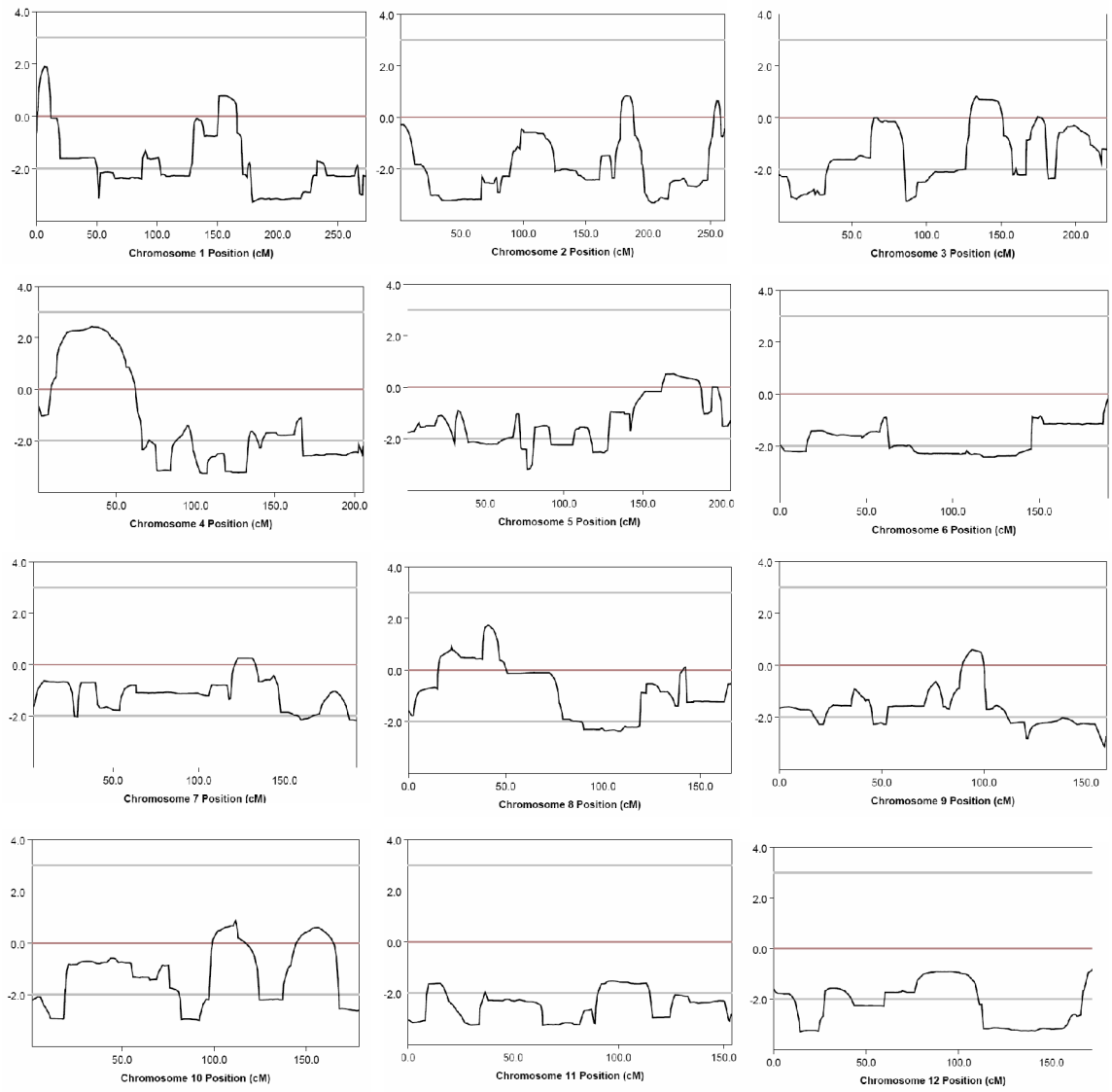


Figure 6.2 Results of parametric linkage analysis under narrow phenotypic model. Continued on the next page.

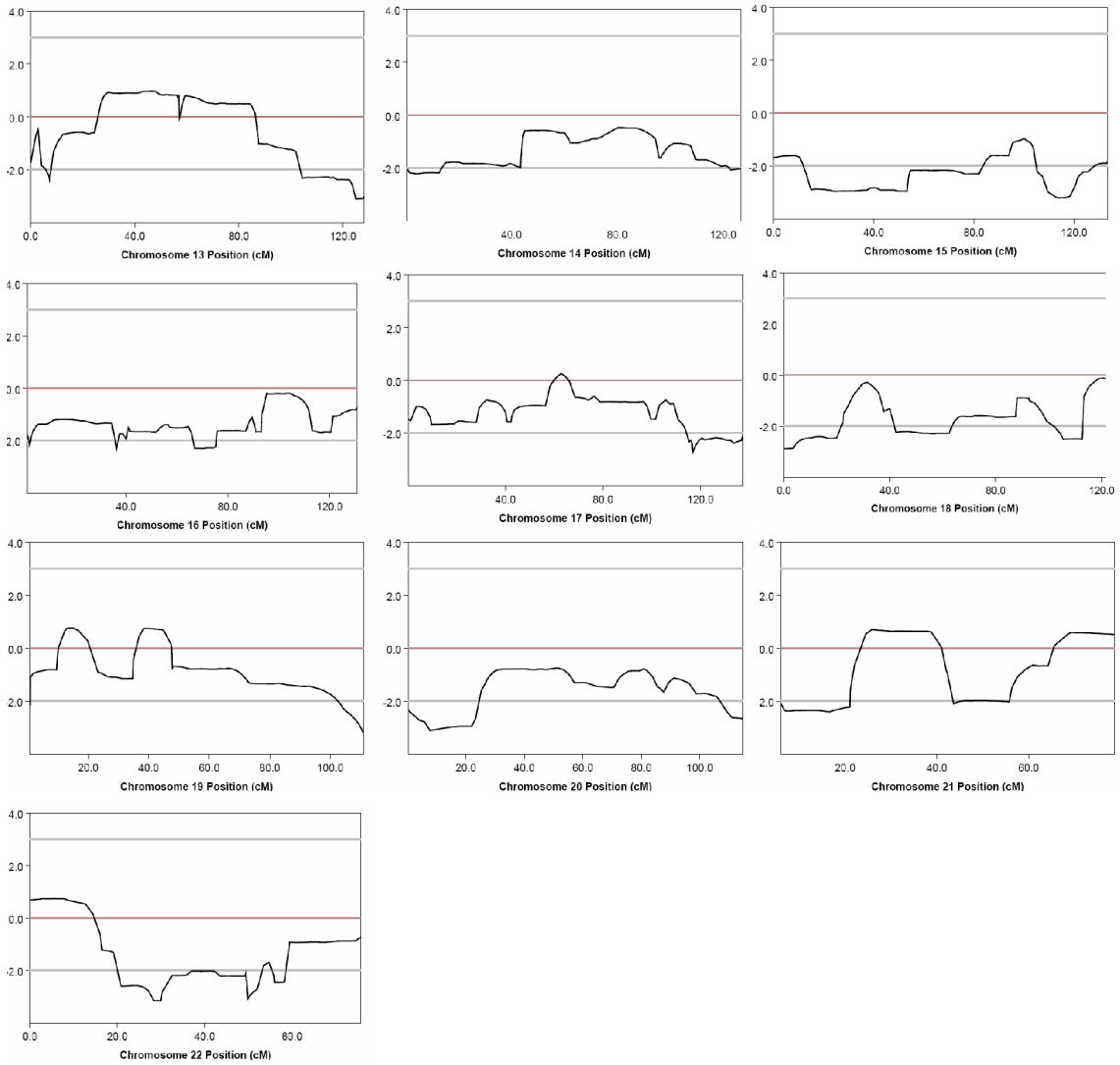


Figure 6.2 Results of parametric linkage analysis under narrow phenotypic model. The x-axis shows the chromosome positions according to deCODE cM. The y-axis shows the LOD score. The grey lines show a threshold for significant linkage, LOD=3, and for negative linkage, LOD=-2.

Regions of suggestive linkage are detailed in Table 6.2. Although simulation studies were not performed for parametric linkage analysis, the cut-offs determined by non-parametric linkage analysis can be applied here as they are comparable with those suggested by Lander and Kruglyak for parametric linkage analysis (Lander and Kruglyak 1995). LODs >2 are deemed suggestive for linkage and warrant further study. In brief, regions of suggestive linkage in the narrow category are chromosomes 4p14-p16 (LOD 2.4) and 1p36 (LOD 1.9) and in the broad diagnostic category chromosome 4p15-p14 (LOD 3) were identified. In Table 6.2 and Table 6.3, the maximum LOD score at the particular genetic marker was noted. The corresponding HLOD and the estimate proportion of linked families (α, alpha) at the particular genetic marker are provided. Additionally, a 1-LOD support interval is reported. The width of the 1-LOD support interval is R-L, where L and R are the first positions at which the LOD score is less than the maximum LOD score minus 1 to the left and right of the position of the maximum LOD score respectively. It is evident from Table 6.2 that the chromosome 4 loci overlap between the narrow and broad phenotypic model. However, the peak was much wider in the narrow model (43cM) than the broad model (25cM) which was consistent with more affected individuals (a greater number of meioses) in the broad model (34) than the narrow model (11). Additionally, genetic heterogeneity was not detected, as the α score was 1 and the HLOD was the same as the LOD score. This result argues against the presence of genetic heterogeneity with this large family. The linkage evidence for chromosome 1p36 in the narrow phenotypic model was reported, as LOD 1.9 approximates to that for suggestive linkage LOD 2. Although it was not suggestive, the results may be of potential interest and worth describing.

Phenotypic Model	Chromosomal Position	Support Interval deCODE cM (Bld 35 Co-ordinates bp)	Number of Markers in Support Interval	Peak at deCODE cM	Marker	LOD	Alpha	HLOD
Narrow	4p14-p16	12.86→55.81 (6,681,015→ 36,965,713)	43	34.78, 34.88	rs1477898, rs1850548	2.4	1	2.4
Narrow	1p36	1.72→11.3 (2,894,007→6,212,680	17	6.43	rs1870509	1.9	1	1.9
Broad	4p15-p14	29.61→55.81 (15,228,692→ 36,965,713)	24	40.15	rs1476880	3	1	3

Table 6.2 Regions of suggestive linkage from parametric linkage analysis on the whole genome SNP scan on sub-pedigrees. Both broad and narrow phenotypic models were analysed. The chromosome position of the linkage peak is from the UCSC genome browser (May04). The support interval is based on the Max 1-LOD rule, $LOD_{MAX} \geq 2$. The highest LOD for each peak is shown at the particular marker position, where more than one marker at the same LOD the both are mentioned. Alpha is the estimated proportion of linked families and the HLOD is the corresponding maximum heterogeneity LOD score. The table is separated by phenotypic model and ordered by descending LOD score.

Regions of nominal interest, LOD≥1 are described in Table 6.3. Although these loci did not reach suggestive levels of significance, they are worth reporting as potential loci linked to illness in the whole family or in individual sub-families. It is worth noting that the potential linkage evidence for chromosome 1p36 (LOD 1.3) for the broad model shown in Table 6.3, overlaps with the region of suggestive linkage (LOD 1.9) in the narrow model as Table 6.2. This lends more support to this locus. The linkage evidence on chromosome 13q was a wide linkage peak (27cM) in both phenotypic models (LOD 1 in the narrow model and LOD 1.4 in the broad model). However, the evidence for linkage comes only from sub-pedigree 4 (LOD 1.77) and 9 (LOD 0.4) suggesting heterogeneity in the broad phenotypic model (LOD 1.4, α 0.7, HLOD1.5). Linkage to a small region of 9cM on chromosome 8p21 (LOD 1.7, α 1, HLOD 1.7) was detected under the narrow phenotypic model, but not under the broad phenotypic model.

Phenotypic Model	Chromosomal Position	Support Interval deCODE cM(Bld 35 Co-ordinates bp)	Number of Markers in Support Interval	Peak at deCODE cM	Marker	LOD	Alpha	HLOD
Narrow	8p21	38.02→46.84 (21,746,352→ 27,647,768)	14	40.75	rs2466216	1.7	1	1.7
Narrow	13q13-q14	24.54→57.09 (29,554,081→ 58,773,607)	53	46.29	rs299344	1	1	1
Broad	13q14	28.71→55.98 (30,950,286→ 53,379,823)	41	48.03	rs1408875	1.4	0.7	1.5
Broad	1p36	0.33→12 (2,458,154→6,647,548)	20	8.51	rs557477	1.3	1	1.3
Broad	1q23	149.91→166.07 (153,127,177 →162,383,383	20	155.97, 156.23	rs1053074, rs10594	1	0.6	1.4

Table 6.3 Regions of nominal linkage from parametric linkage analysis on the whole genome SNP scan on sub-pedigrees. Both broad and narrow phenotypic models were analysed. The chromosome position of the linkage peak is from the UCSC genome browser (May04). The support interval is based on the Max 1-LOD rule, LOD ≥ 1. It is defined by the first and last position at which the LOD score is less than the maximum score minus 1. The highest LOD for each peak is shown at the particular marker position, where more than one marker at the same LOD the both are mentioned. Alpha is the estimated proportion of linked families and the HLOD is the corresponding maximum heterogeneity LOD score. The table is separated by phenotypic model and ordered by descending LOD score.

Linkage analysis was performed under a recessive model for the broad phenotypic model. There were no LOD scores > 0.9 and the recessive mode of inheritance was not considered further.

6.2.2. Non-parametric linkage results

Non-parametric linkage analysis was performed on whole genome SNP data for each sub-pedigree. The preparation of the quality controlled genotyping data and the sub-pedigrees are described in chapter 5. Non-parametric linkage analysis was performed on both the broad and narrow phenotypes, as described in section 2.5.8. The maximum possible scores obtainable, considering the number of genetic markers and the number of affected individuals per sub-pedigree, are shown in Table 6.4.

PHENOTYPE	Z MEAN	*P*-VALUE	EXP DELTA	LOD	*P*-VALUE
Narrow	4.72	0.0000	9.999	2.41	0.0004
Broad	8.68	0.0000	9.999	7.72	0.0000

Table 6.4 Maximum LOD scores obtainable for non-parametric linkage analysis. The Z mean score is mean-sharing score for affected members in a pedigree with the corresponding *P*-value. The exponential delta measures the amount of deviation of the inheritance vector distribution from it's null distribution, with the corresponding LOD score and *P*-value.

The maximum obtainable LOD score was 2.41 for the narrow phenotypic model and 7.72 for the broad model. Linkage analysis on the broad phenotype model had more power as expected, given that there were a greater number of affected individuals. It was important to emphasis that non-parametric linkage analysis cannot achieve LOD scores greater than 2.41, under the narrow phenotypic model.

The analysis was performed using MERLIN software to calculate the Whittemore and Halpern non-parametric linkage analysis S*all* statistics, alongside the Kong and Cox LOD scores. The results for the whole genome scan are presented in Figure 6.3. As the exponential model was a better test for a large increase in allele sharing expected among affected individuals, these results were shown.

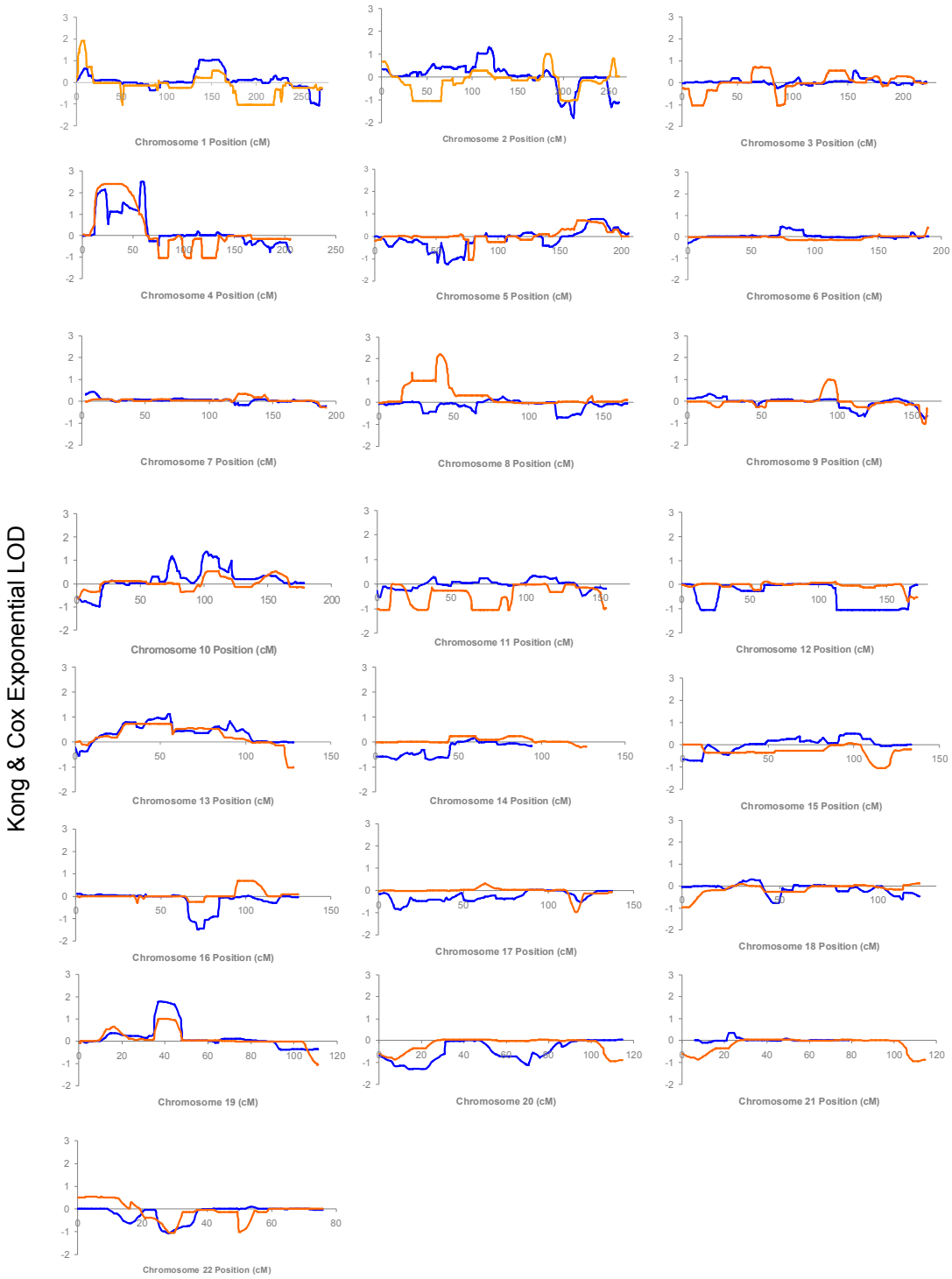


Figure 6.3 Result of non-parametric linkage analysis. The x-axis is the chromosome position according to deCODE cM. The y-axis is Kong and Cox exponential LOD score for the broad phenotype model (blue line) and the narrow phenotype model (orange line).

It is clear from Figure 6.3 that many autosomes show no evidence for linkage (LOD<1) or negative non-parametric LOD scores that indicate less than expected allele sharing in the family. There was evidence for suggestive linkage, as defined by Table 6.1, LOD>2 for the broad phenotype model and LOD>2.11 for the narrow phenotype model. The suggestive linkage results for non-parametric linkage analysis are shown in detail in Table 6.5. There was evidence for suggestive linkage on chromosome 4p14-p16 supported by both the narrow (LOD 2.4) and broad (LOD 2.52) phenotypic model. There was also suggestive linkage on chromosome 8p21 (LOD 2.22). Again, the peak on chromosome 1p36 (LOD 1.9), in the narrow phenotypic model was reported as it is of potential interest. Table 6.5 also details the empirical *P*-values calculated from 10,000 simulations of the broad phenotypic model and 1,000 simulations of the narrow phenotypic model. This assessed the genome-wide significance of the results. None of the linkage results met a genome-wide level of significance (all *P*-values>0.05).

Model	Chromosomal Position	Support Interval deCODE cM (Bld 35 Co-ordinates bp)	Number of Markers in Support Interval	Peak at deCODE cM	Marker with Maximum LOD score	Max non-parametric linkage analysis K&C LOD Score	Z score	Probability of Z score	Delta	Probability of Kong & Cox LOD score	Empirical *P*-value of Kong & Cox LOD score
Narrow	4p14-p16	12.46→ 63.43 (6,539,741→ 42,164,670)	53	22.53→ 38.72	rs881641→ rs216113 (27 SNPs)	2.40	4.70 (mean)	0	9.999	0.0004	0.54
Narrow	8p21	21.74→ 46.84 (8,805,729→ 27,647,768)	48	40.75	rs2466216	2.22	3.83	0.00006	9.999	0.0007	0.686
Narrow	1p36	1.9→ 11.3 (2,949,536→ 6,212,680)	15	6.43	rs1870509, rs2035453, rs557477, rs912991	1.9	3.54	0.0002	9.999	0.002	0.754
Broad	4p14-p16	12.46→ 63.43 (6,539,741→ 42,164,670)	53	60.32	rs278973	2.52	3.59	0.0002	0.842	0.0003	0.187

Table 6.5 Regions of suggestive linkage from non-parametric linkage analysis. The chromosome position of the linkage peak is from the UCSC genome browser (May04). The support interval is based on the Max 1-LOD rule, LOD ≥ 2 which is the threshold for suggestive linkage as determined by simulations. The LOD score shown is the maximum LOD score for the SNP markers listed. The Z score, *P*-value assuming normal approximation and the Kong and Cox *P*-value are also listed. The empirical *P*-values were calculated from 10,000 simulated datasets for the broad model and 1,000 simulated datasets for the narrow model. The data is separated according to phenotypic model and sorted according to descending LOD score.

Figure 6.3 and Table 6.6 display results of nominal linkage, LOD≥1 that did not meet the criteria for suggestive linkage (LOD ≥2.11 for the narrow phenotypic model and LOD≥2 for the broad phenotypic model). It was important to note these positions, without making any substantial claims and to clearly present the LOD scores. It is of interest that the linkage peak on chromosome 1p13-q23, under the broad phenotypic model (LOD 1.03), overlaps with the peak of suggestive linkage under the narrow phenotypic model (LOD 1.9) in Table 6.5. The linkage peak on chromosome 19p13 was also interesting as the peak overlaps in both the broad (LOD 1.77) and narrow (LOD 1.01) phenotypic model.

Model	Chromosomal Position	Support Interval deCODE cM (bld 35 co-ord bp)	Number of Markers in Support Interval	Peak at deCODE cM	Marker with Maximum LOD score	Max non-parametric linkage analysis K&C LOD Score	Z score	Probability of Z score	Delta	Probability of Kong & Cox LOD score	Empirical *P*-value of Kong & Cox LOD score
Narrow	19p13	4.73→ 69.38 (1,979,985→ 48,671,429)	89	38.27→ 39.67	rs1273533→rs754292 (4 SNPs)	1.01	2.96	0.002	0.942	0.02	0.954
Narrow	2q31	177.63→ 192.78 (171,766,525→ 192,224,963)	28	181.66→ 185.63	rs3856434→ rs963854 (7 SNPs)	1	2.98	0.0014	0.931	0.02	0.96
Broad	19p13	34.68→ 47.82 (13,812,830→ 23,229,879)	18	38.27, 38.32	rs1273522, rs2060260	1.77	4.13	0.00002	0.623	0.002	0.601
Broad	10q23	97.15→ 123.03 (78,860,756→ 106,077,502)	36	102.04	rs1336439	1.36	3.43	0.0003	0.567	0.006	0.851
Broad	2p12-q13	47.98→ 138.56 (24,929,008→ 126,678,671)	187	118.9	rs1568261	1.29	3.26	0.0006	0.561	0.007	0.884
Broad	10q21	70.16→ 88.6 (52,464,130→ 71,856,800)	27	74.93	rs1338799	1.18	3.26	0.0006	0.534	0.010	0.925
Broad	13q14	12.3→ 104.36 (24,087,224→ 104,589,808)	134	54.92	rs1058142	1.12	3.14	0.0008	0.524	0.012	0.944
Broad	1p13-q23	129.16→ 169.71 (109,184,855→ 164,849,524)	64	137→ 159	rs2057127→rs1027702 (39 SNPs)	1.03	2.84	0.002	0.522	0.015	0.966

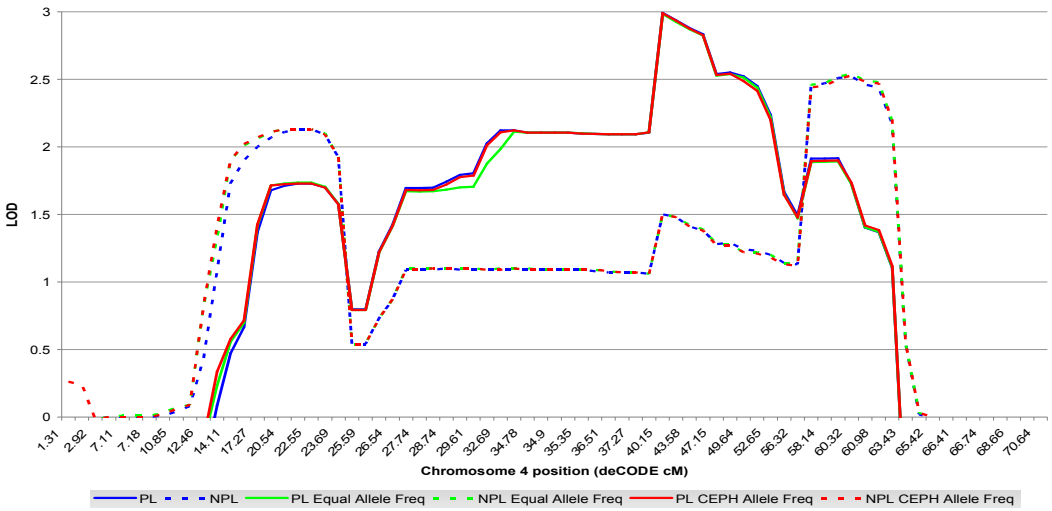
Table 6.6 Regions of nominal linkage from non-parametric linkage analysis. The chromosome position of the linkage peak is from the UCSC genome browser (May04). The support interval is based on the Max 1-LOD rule, LOD ≥ 1. The LOD score shown is the maximum LOD score for the SNP markers listed. The Z score, *P*-value assuming normal approximation and the Kong and Cox *P*-value are also listed. The empirical *P*-values were calculated from 10,000 (broad) and 1,000 (narrow) simulated datasets. The data is sorted according to descending LOD score per phenotypic model.

In conclusion, parametric and non-parametric linkage analysis provided evidence for suggestive linkage on chromosome 4p14-p16 and 1p36 for both the narrow and broad model and chromosome 8p21 for the narrow model. To confirm these results, the following sections test the reliability of the linkage scores.

6.2.3. Robustness of linkage results

The robustness of the linkage results, obtained using family-based allele frequencies, was tested by varying the allele frequencies. Parametric and non-parametric linkage analyses, on both broad and narrow phenotypic models, were performed using allele frequencies from CEPH population and equal allele frequencies. This whole genome analysis did not reveal any chromosomal regions of interest, other than those specified on chromosome 1p36, 4p14-p16 and 8p21 and those shown in Table 6.3 and Table 6.6. The linkage results for chromosome 4p are illustrated in Figure 6.4**,** for chromosome 1 in Figure 6.5 and chromosome 8 in Figure 6.6. In addition, Figure 6.4, Figure 6.5 and Figure 6.6 demonstrate that for parametric and non-parametric linkage analysis, the linkage results based on the various sets of allele frequency data mirrored each other. Importantly, varying the allele frequencies did not alter the chromosomal position of the linkage peaks, the size of the linkage peaks and the position of the maximum LOD scores, thus suggesting the linkage analysis results were robust.

a)



b)

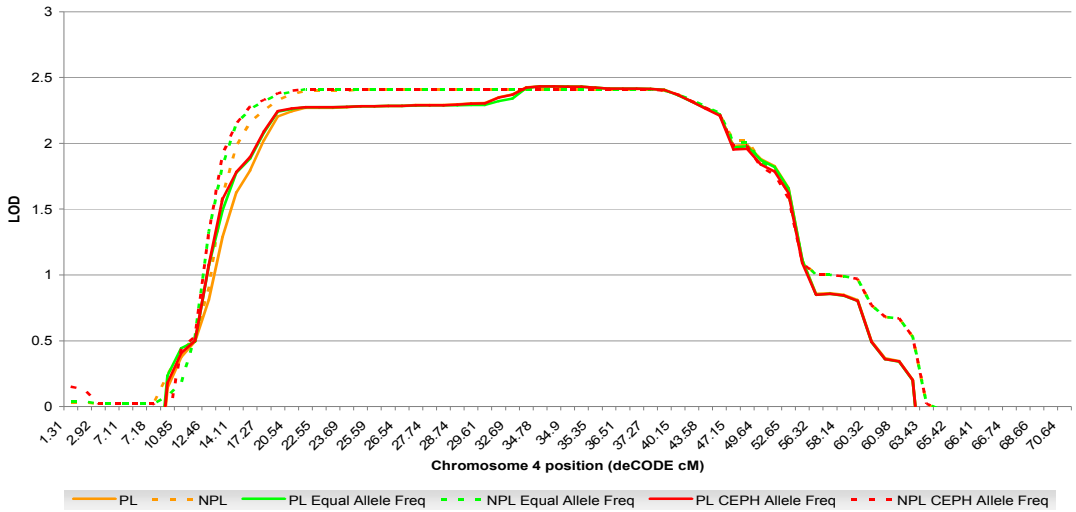
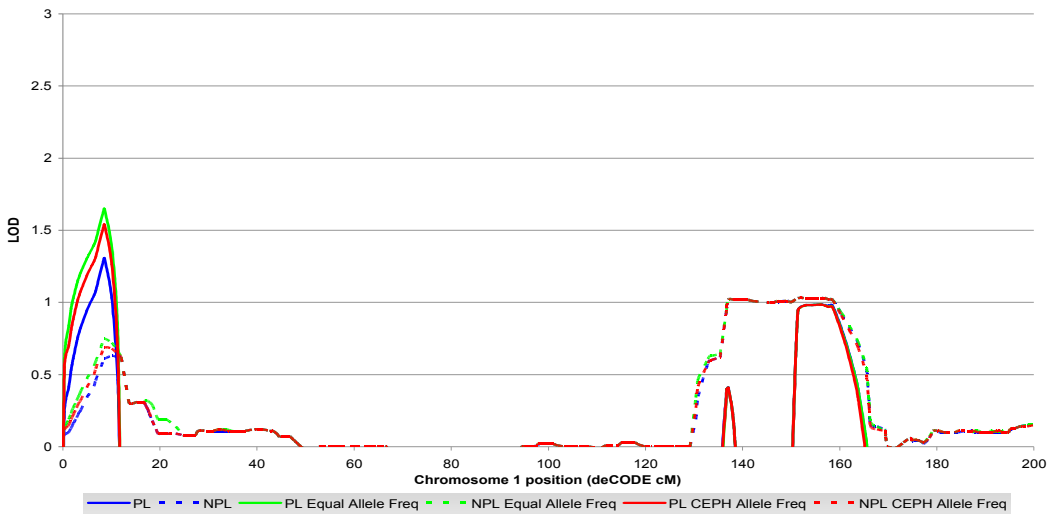


Figure 6.4 Chromosome 4p linkage results are robust to allele frequency testing for the broad phenotypic model and narrow phenotypic model. Linkage analysis was performed with various allele frequencies under the broad (a) and narrow (b) phenotypic model. Chromosome 4, from the start of p-terminal to marker rs895615 at 70.64cM is displayed on the x-axis. There are no linkage peaks after this marker. The y-axis shows the LOD score. Each line represents the linkage result for a particular analysis. Parametric linkage (PL) analysis LOD is shown by a continuous line and non-parametric linkage (NPL) analysis Kong & Cox Exponential LOD score by a dotted line. Each analysis was performed using different allele frequencies, denoted by a different colour; allele frequencies obtained from the family in blue for the broad phenotypic model and orange for the narrow phenotypic model, equal allele frequencies in green and from the CEPH population in red. The blue, orange and green lines are not continuously visible as they are masked by the red line.

a)



b)

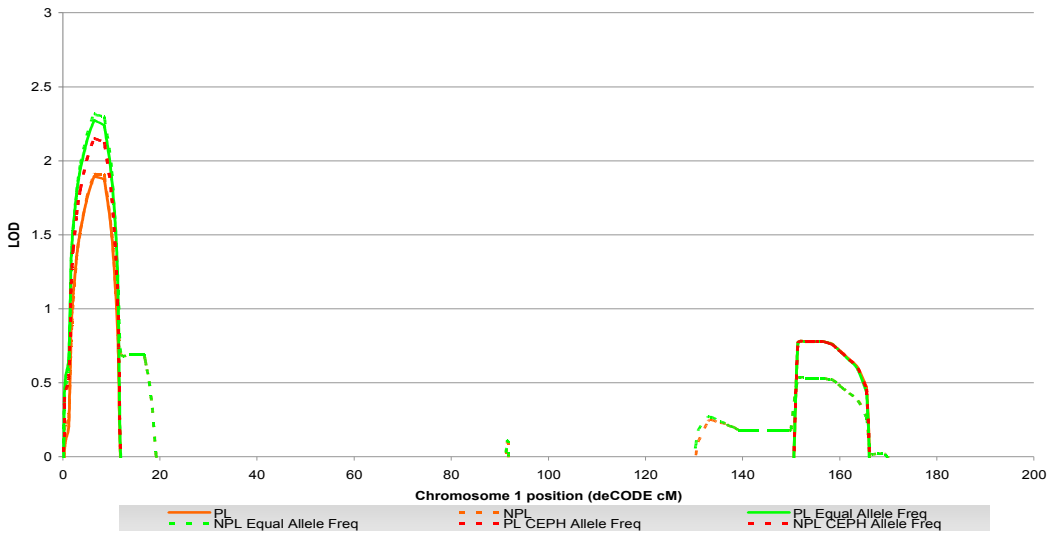


Figure 6.5 Chromosome 1 linkage results are robust to allele frequency testing for the broad phenotypic model and narrow phenotype model. Linkage analysis was performed with different allele frequencies under the broad (a) and narrow (b) phenotypic model. Chromosome 1, from the start of p-terminal to marker rs7513 at 200cM is displayed on the x-axis. The whole chromosome is not shown as there is no LOD >0 past this marker. The y-axis shows the LOD score. Each line represents the linkage result for a particular analysis. Parametric linkage (PL) analysis LOD is shown by a continuous line and non-parametric linkage (NPL) analysis Kong & Cox Exponential LOD score is shown by a dotted line. Each analysis was performed using different allele frequencies, denoted by a different colour; allele frequencies obtained from the family in blue for the broad phenotype and orange for the narrow phenotype. Equal allele frequencies are in green and from the CEPH population in red. The blue, orange and green lines are not continuously visible as they are masked by the red line.

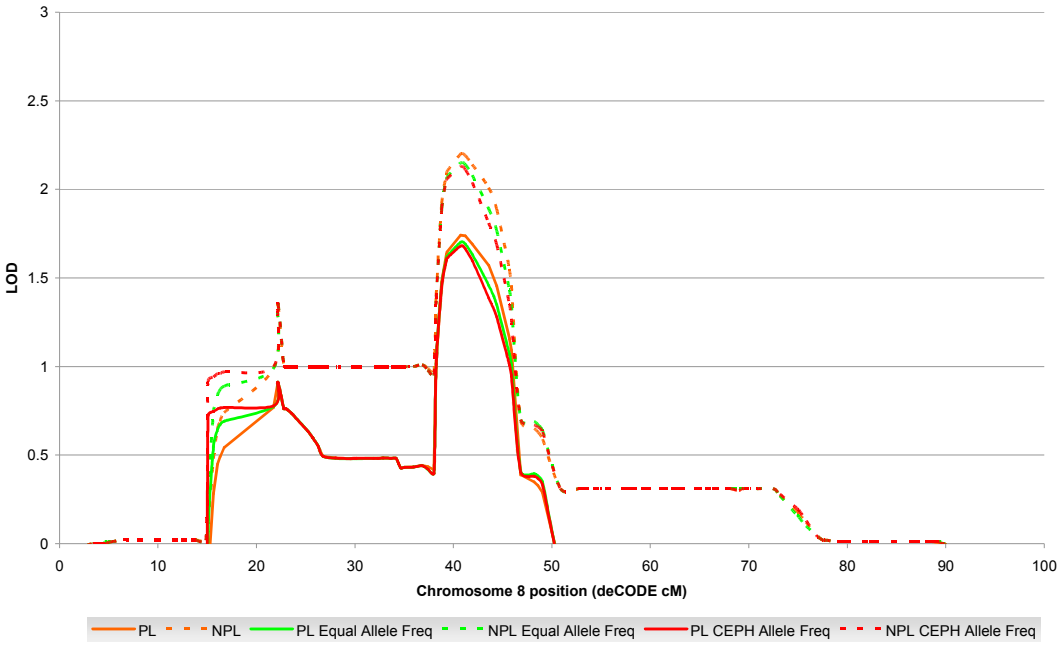


Figure 6.6 Chromosome 8 linkage results are robust to allele frequency testing for the narrow phenotype model. Linkage analysis was performed with different allele frequencies. Chromosome 8, from the start of p-terminal to marker rs1051624 at 99.5cM is displayed on the x-axis. The whole chromosome is not shown as there is no LOD >0 past this marker. The y-axis shows the LOD score. Each line represents the linkage result for a particular analysis. Parametric linkage (PL) analysis LOD is shown by a continuous line and non-parametric linkage (NPL) analysis Kong & Cox Exponential LOD score is shown by a dotted line. Each analysis was performed using different allele frequencies, denoted by a different colour; allele frequencies obtained from the family in orange, equal allele frequencies in green and from the CEPH population in red. The orange and green lines are not continuously visible as they are masked by the red line.

6.3. *Suggestive Linkage Results*

6.3.1. Chromosome 4p15-p16

6.3.1.1. Parametric & non-parametric linkage analyses combined

The chromosome 4p15-p16 linkage regions are illustrated in Figure 6.1 to Figure 6.3 and detailed in Table 6.2 and Table 6.5. Figure 6.7 shows the parametric and non-parametric linkage evidence on chromosome 4p15-p16 illustrated together. Firstly, it is evident that both methods of analysis detected linkage to the same locus, as the full line for parametric analysis is mirrored by the dotted line for non-parametric analysis. There was an exception for the broad phenotypic model at 30-60cM, where the evidence peaked in the parametric analysis and dipped in the non-parametric analysis. As discussed previously in section 1.10, linkage analysis has the ability to detect linkage to a region of ~20cM and thus, the whole region was considered a linkage peak. The information content of the SNP genetic markers was consistent across the region.

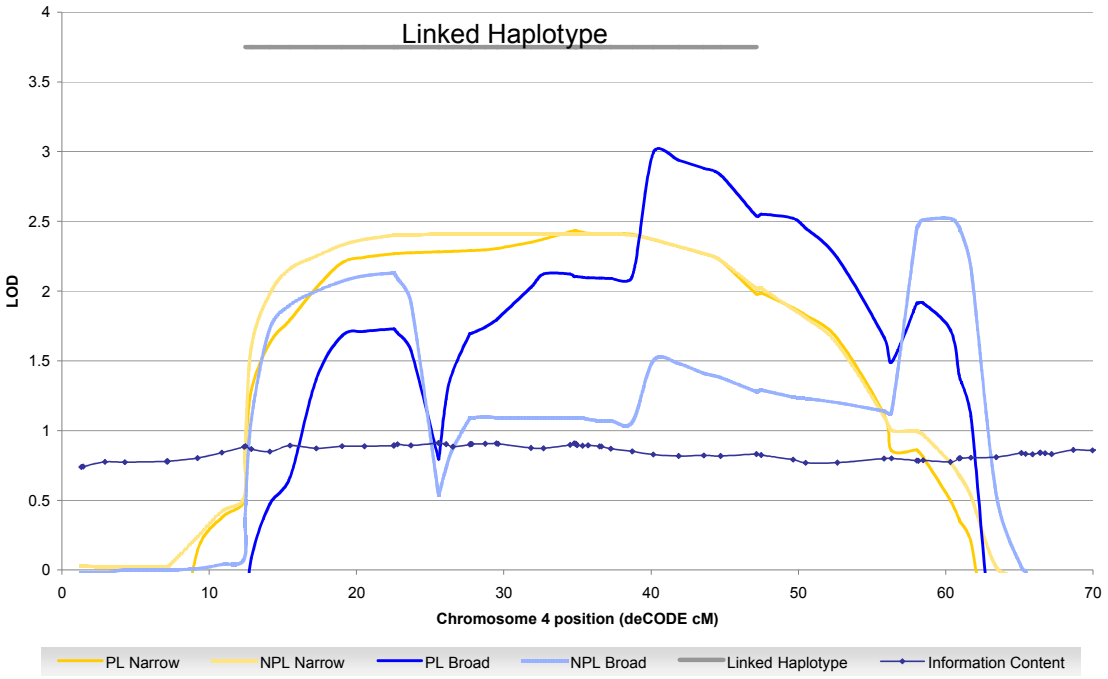


Figure 6.7 Parametric and non-parametric linkage analysis on chromosome 4p14-p16. The x-axis is the genetic position on chromosome 4 from start of p-terminal to marker rs895615 at 70.64cM. There is no linkage evidence LOD>0 beyond this marker. The y-axis shows the LOD score. The result of parametric (PL) and non-parametric analysis (NPL) for the narrow model is shown by the orange and yellow respectively and by the dark blue and light blue line respectively (b) for the broad model. Linkage analysis was performed on all SNP markers. The grey bar shows the linked haplotype region as defined by previous work to highlight the chromosomal region that segregates with bipolar disorder in the family.

6.3.1.2. Increased marker coverage on chromosome 4p14-p16

To improve marker coverage in the chromosome 4p15-p16 region, microsatellites from the previous linkage study were added (Le Hellard, Lee et al. 2007). There were 24 microsatellites available, typed on a range of individuals from the family, as listed in Table 2.10. Figure 6.8 shows that the addition of microsatellite data on chromosome 4 provided greater information, increasing LOD scores in both parametric and non-parametric linkage analysis in the narrow and broad phenotypic model. The LOD scores increased in the narrow model to LOD 2.7 over a wide linkage peak and in the broad model to LOD 3.6 at rs1476880 (40.15cM) for parametric linkage analysis and LOD 3.6 at stb448G15.ca2 (22.17cM) for non-parametric linkage analysis. The dip in linkage at position 22cM in both models and analyses was due to the addition of microsatellite information for some, but not all individuals. These missing genotypes contributed to a loss of linkage information. The position of the linkage peak corresponded to the position of the linked haplotype previously defined by our group for bipolar disorder, shown by the grey bar. In Figure 6.8a, there was a linkage peak outside of the previously linked haplotype at chromosome 4p14, however this was not present in the narrow phenotypic region and was likely due to a different spectrum of recombination events in individuals with recurrent major depression.

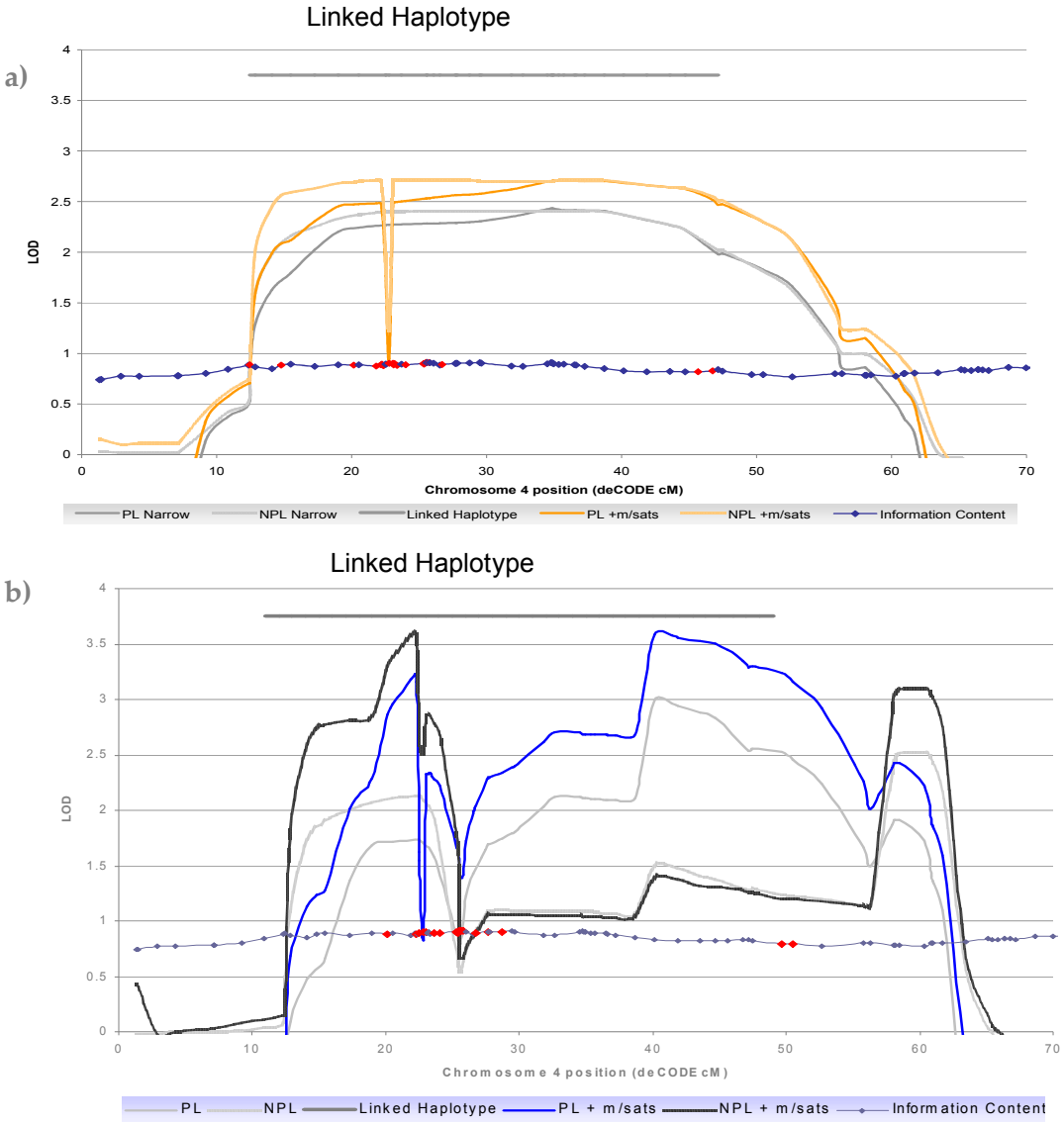


Figure 6.8 Addition of microsatellite markers increases the linkage evidence on chromosome 4p14-p16 under the narrow (a) and broad (b) phenotypic model. The x-axis is the genetic position on chromosome 4 from start of p-terminal to marker rs895615 at 70.64cM. There is no linkage evidence (LOD>0) beyond this marker. The y-axis shows the LOD score. The result of parametric (PL) and non-parametric analysis (NPL) under the narrow model, is shown by the orange and dotted orange line respectively and under the broad model, by the blue and dotted blue line respectively. Linkage analysis was performed on all SNP markers, shown by grey lines and with SNP and microsatellite markers, shown by the coloured lines. The information content for all available markers, both SNPs (dark blue diamond) and microsatellites (red diamond) are marked. The grey bar shows the linked haplotype region, as defined by previous work to highlight the chromosomal region that segregates with bipolar disorder in the family.

Figure 6.9 shows the linkage evidence for both phenotypic models combined. Although, addition of extra markers did not narrow down the linkage region, it does provide support for this locus by increasing the LOD score. Table 6.7 details the scores obtained by inclusion of microsatellite markers.

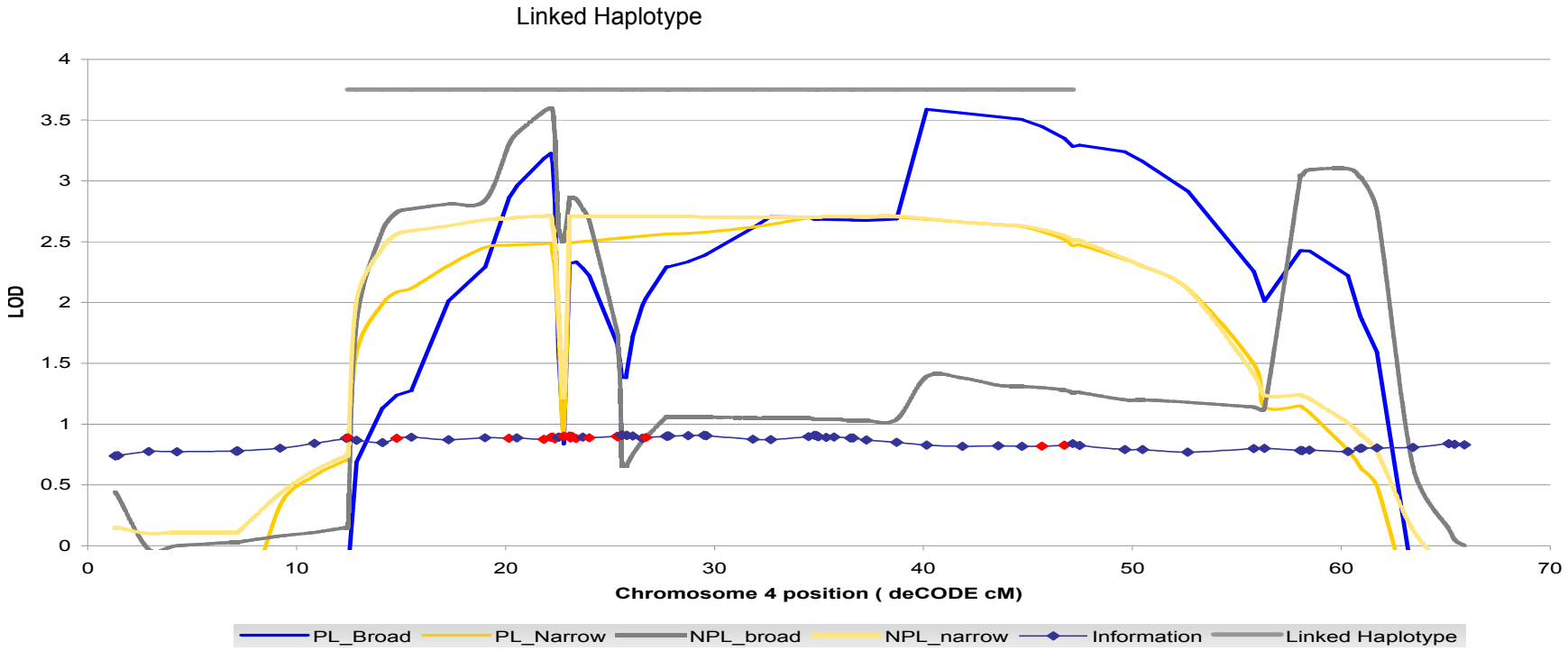


Figure 6.9 Suggestive evidence for linkage on chromosome 4p. The x-axis the genetic position on chromosome 4 from start of p-terminal to marker rs1866989 at 65.9cM. There is no linkage evidence LOD>0 beyond this marker. The y-axis shows the LOD score. The result of parametric (PL) and non-parametric (NPL) analysis for the broad model is shown by the blue and dotted blue line respectively. The result of PL and NPL for the narrow model is shown by the orange and dotted orange line respectively. Linkage analysis was performed on all markers available for chromosome 4p, including SNPs and microsatellites. The information content is shown for SNPs (blue diamonds) and microsatellites (red diamonds). The grey bar shows the linked haplotype region as defined by previous work to highlight the chromosomal region that segregates with bipolar disorder in the family.

a)

Phenotypic Model	Chromosomal Position	Support Interval deCODE cM	Number of Markers in Support Interval	Peak at deCODE cM	Marker with Maximum LOD score	LOD	Alpha	HLOD
Narrow	4p14-p16	14.11→55.81	62	34.78→ 38.72	rs1477898→ rs216113 (10 markers)	2.7	1	2.7
Broad	4p16	17.271→25.34	18	22.17	stb448G15.ca2	3.3	1	3.2
Broad	4p15	29.611→ 55.81	26	40.15	rs1476880	3.6	1	3.6

b)

Model	Chromosomal Position	Support Interval deCODE cM	Number of Markers in Support Interval	Peak at deCODE cM	Marker with Maximum LOD score	Max non-parametric linkage analysis K&C LOD Score	Z score	Probability of Z score	Delta	Probability of Kong & Cox LOD score
Narrow	4p15-p16	12.46→ 55.81	64	12.86→ 38.72	rs881641→rs216113 (41 markers)	2.71	4.40 (mean)	0	9.999	0.0002
Broad	4p16	12.86 →25.34	22	22.17	stb448G15.ca2	3.6	7.99	0	0.789	0.00002
Broad	4p14	56.321→ 63.43	9	60.32	rs278973	3.1	4.55	0	0.835	0.00008

Table 6.7 Regions of suggestive linkage on chromosome 4 scan using both SNP and microsatellite markers. Both broad and narrow phenotypic models were analysed using parametric (a) and non-parametric linkage analysis (b). The chromosome position of the linkage peak is from the UCSC genome browser (May04). The support interval is based on the Max 1-LOD rule, $LOD_{MAX} \geq 2$ which is the threshold for suggestive linkage as determined by simulations. It is defined by the first and last position at which the LOD score is less than the maximum score minus 1. The highest LOD for each peak is shown at the particular marker position, where more than one marker at the same LOD the both are mentioned. Alpha is the estimated proportion of linked families and the HLOD is the corresponding maximum LOD. In table b, the LOD score shown is the maximum LOD score for the SNP markers listed. The Z score, *P*-value assuming normal approximation and the Kong and Cox *P*-value are also listed. The data is sorted according to descending LOD score per phenotypic model.

6.3.1.3. Contribution of each sub-pedigree

The contribution of each sub-pedigree to the linkage result is shown in Figure 6.10. It is clear that sub-pedigree 1 contributed principally to the LOD scores, as shown in Table 6.8. Sub-pedigrees 2 and 4 also contributed to the linkage peak. Table 6.8 shows the LOD scores from each sub-pedigree that were added together to give a LOD score for the whole family.

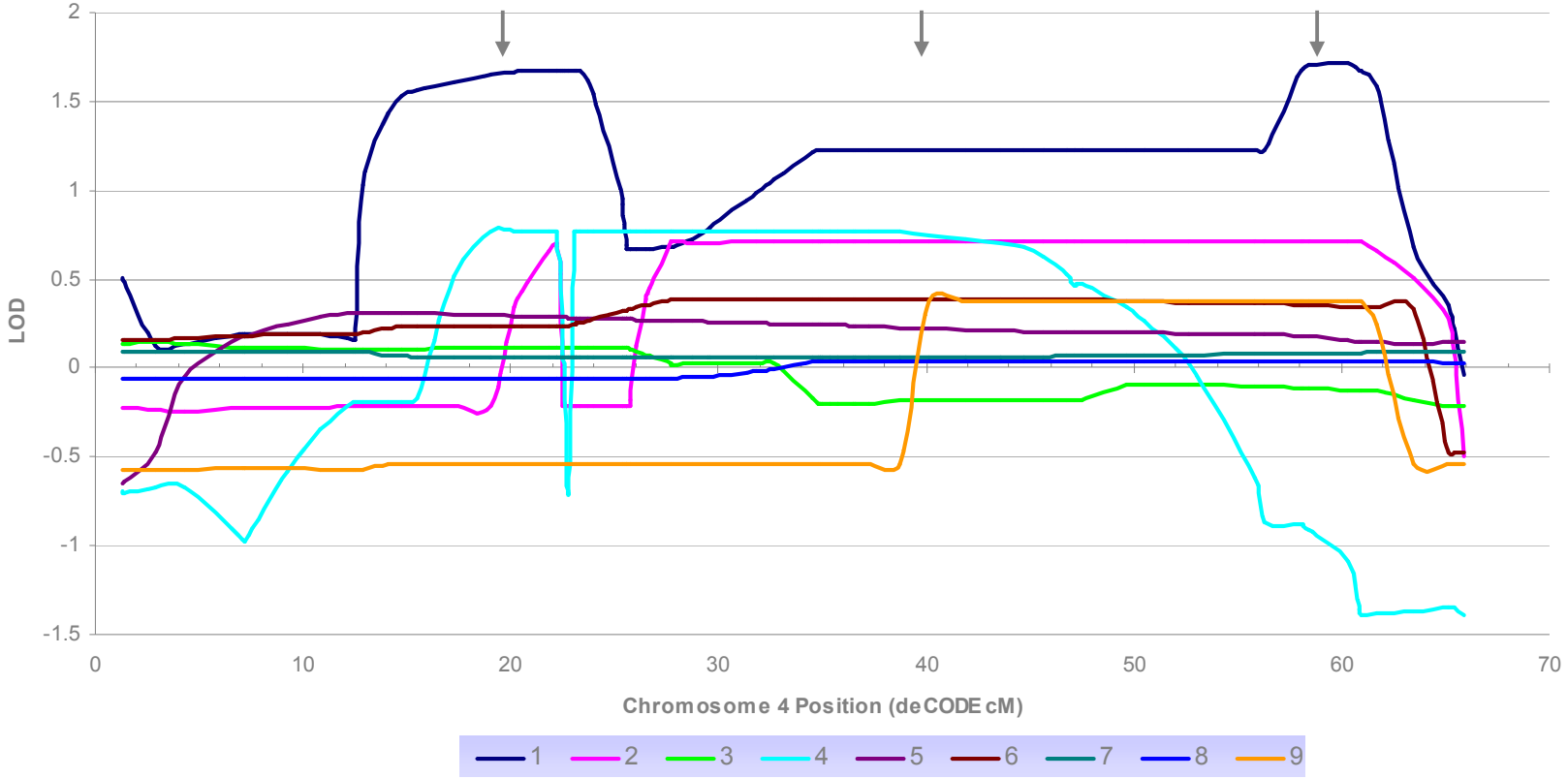


Figure 6.10 Parametric analysis on broad phenotypic model per sub-pedigree on chromosome 4. The x-axis is the position on chromosome 4 position from the p-terminal to rs1866989 at 65.7cM. The linkage result per sub-pedigree is shown by a different coloured line. The grey arrows indicate the linkage peaks at 22.17cM, 40.15cM and 60.32cM.

Sub-pedigree	# Affected Individuals	LOD Score		
		4p16 stb448G15.ca2 (22.17cM)	4p15.2 rs1476880 (40.15cM)	4p14 rs278973 (60.32cM)
1	10	1.7	1.25	1.7
2	4	0.7	0.725	0.7
3	2	0.1	-0.2	-0.1
4	5	0.8	0.725	-1.08
5	4	0.3	0.2	0.15
6	2	0.2	0.4	0.4
7	1	0.1	0.06	0.09
8	1	-0.1	0.04	0.04
9	3	-0.5	0.4	0.4
Total	32	3.3	3.6	2.3

Table 6.8 Chromosome 4p linkage results from sub-pedigree. The linkage results for the sub-pedigrees that are informative for linkage in the broad phenotypic model are listed. There are three linkage peaks in the chromosome 4p14-p16 region. The marker with the highest LOD score in each of the peaks is shown for each sub-pedigree.

6.3.1.4. Contribution of a single affected individual

The linkage results were also tested to see if the peaks were a consequence of one individual only, and therefore an artefact of linkage analysis. There are many reasons for artefactual linkage peaks for example: interval mapping of widely spaced markers can cause problems, as do random fluctuations in marker informativeness and erroneous genotypes that give a false impression of a recombination event (Ott 1999). The analysis chosen to perform this test was non-parametric linkage analysis for the narrow phenotypic model, because non-parametric linkage analysis was not dependent on genetic model assumptions and also, the individuals in the narrow model with a diagnosis of bipolar disorder were most informative. For each linkage analysis, the affection status was set to unknown for one individual with a bipolar disorder diagnosis at a time. The LOD score reached a limit according to the calculated maximum obtainable LOD score as shown in Table 6.9. The maximum LOD score that was obtainable was dependent on the informativeness of each individual. For example, removing samples without parental information; 29, 102 and 150, the maximum obtainable LOD score was 2.11. Conversely, removing more informative samples with parental (inheritance) information; 21, 47, 57 and 99, lead to a decrease in the maximum obtainable LOD score (1.81).

ANALYSIS	SUB-PEDIGREE	MAXIMUM LOD (KONG & COX EXPONENTIAL)
All individuals		2.41
21	1	1.81
29	1	2.11
47	1	1.81
57	1	1.81
99	4	1.81
102	4	2.11
150	4	2.11

Table 6.9 Maximum LOD scores obtainable for analysis when removing selected affected individuals. The maximum obtainable Kong and Cox exponential LOD scores for non-parametric linkage analysis are displayed above. The analysis including all seven individuals with a bipolar disorder diagnosis in the narrow phenotypic model. The other results are for analysis with the affection status changed to unknown for the individuals specified.

Figure 6.11 showed the linkage results for chromosome 4, after removing selected individuals. Each coloured line showed linkage analysis with one bipolar disorder affection status sample removed at a time. These lines can be compared to the orange line which showed the linkage analysis for all individuals with bipolar disorder diagnoses. It is clear that all individuals contribute to the chromosome 4p15-p16 peak, and that each analysis reached their maximum obtainable LOD score in this region, as shown in Table 6.4. This showed that LOD score was robust to removal of single affected individuals. The LOD 1.13 peak at 114cM shows that removal of individual 50 caused an artefact of linkage analysis.

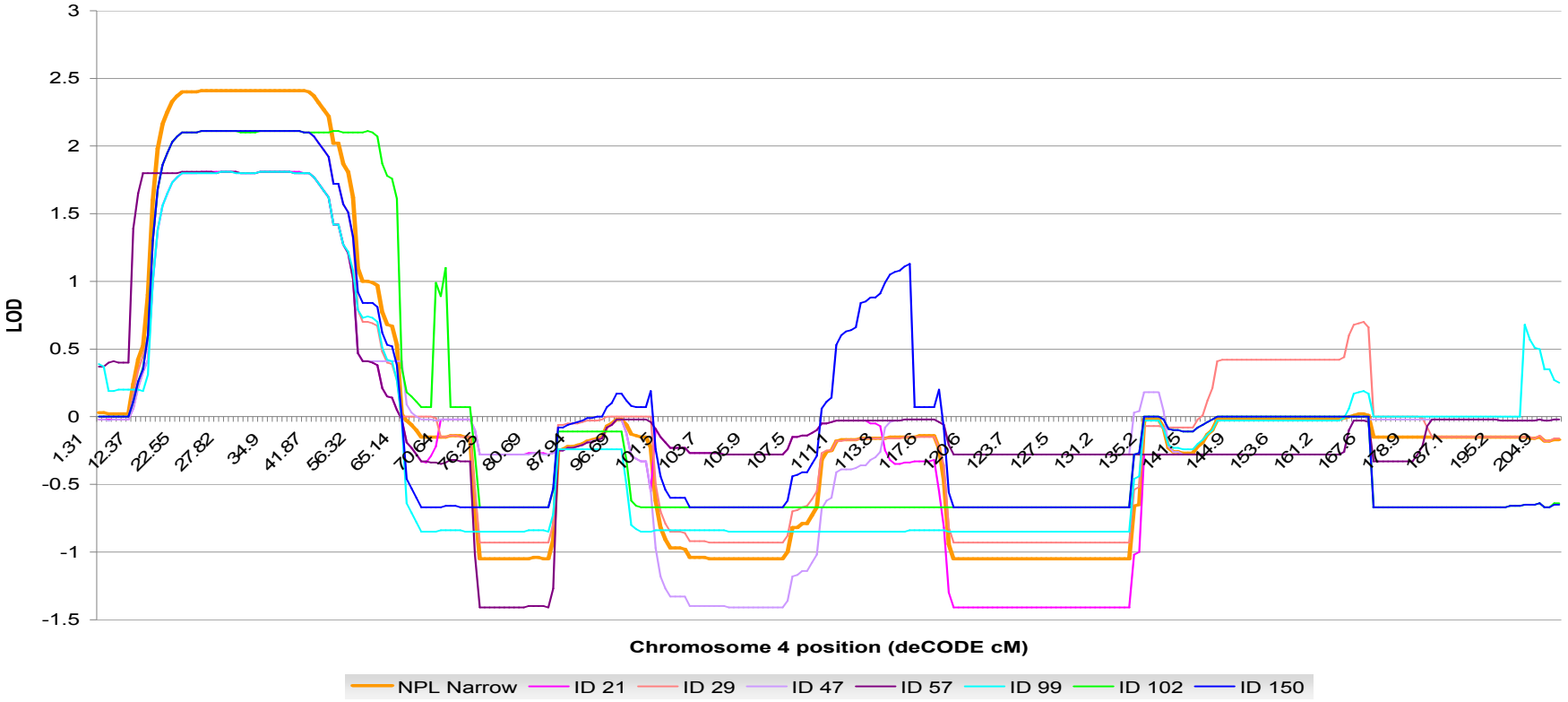


Figure 6.11 Contribution of a single individual to the linkage results on chromosome 4. The x-axis is the chromosome 4 genetic position. The y-axis is the Kong and Cox exponential LOD score from non-parametric linkage analysis. The orange line is the result including all seven affected individuals from the informative sub-pedigrees 1 and 4. The other coloured lines are linkage analysis performed with the individual specified affection status set to unknown. The coral and green line for ID29 and ID102 respectively is masked by blue line for ID150 at the 4p15-p16 peak. The turquoise line for ID99 masks the lines for ID21, ID47 and ID57 at the 4p15-p16 peak. Individuals 21, 29, 47 and 57 are in sub-pedigree 1 and individuals 99, 102 and 150 are in sub-pedigree 4.

6.3.2. Chromosome 1p36

6.3.2.1. Parametric & non-parametric linkage analyses combined

Another instance of suggestive linkage is illustrated in Figure 6.12, which shows a peak on chromosome 1p36 in both narrow (orange) and broad (blue) phenotype in both parametric (full line) and non-parametric linkage analysis (dotted line) analysis methods. The greatest evidence for linkage at chromosome 1p36 stemmed from the narrow phenotypic model. Both methods of analysis, parametric and non-parametric, had the same linkage peak at chromosome 1p36. The peak was in the same location for the broad phenotypic model, but smaller. The bipolar disorder candidate gene on chromosome 1p36.22 *methylenetetrahydrofolate reductase (MTHFR)* (Gilbody, Lewis et al. 2007), the schizophrenia candidate gene *DISC1* (Chubb, Bradshaw et al. 2008) and region 1q21 (Brzustowicz, Hodgkinson et al. 2000) are marked for comparison to the linkage peak.

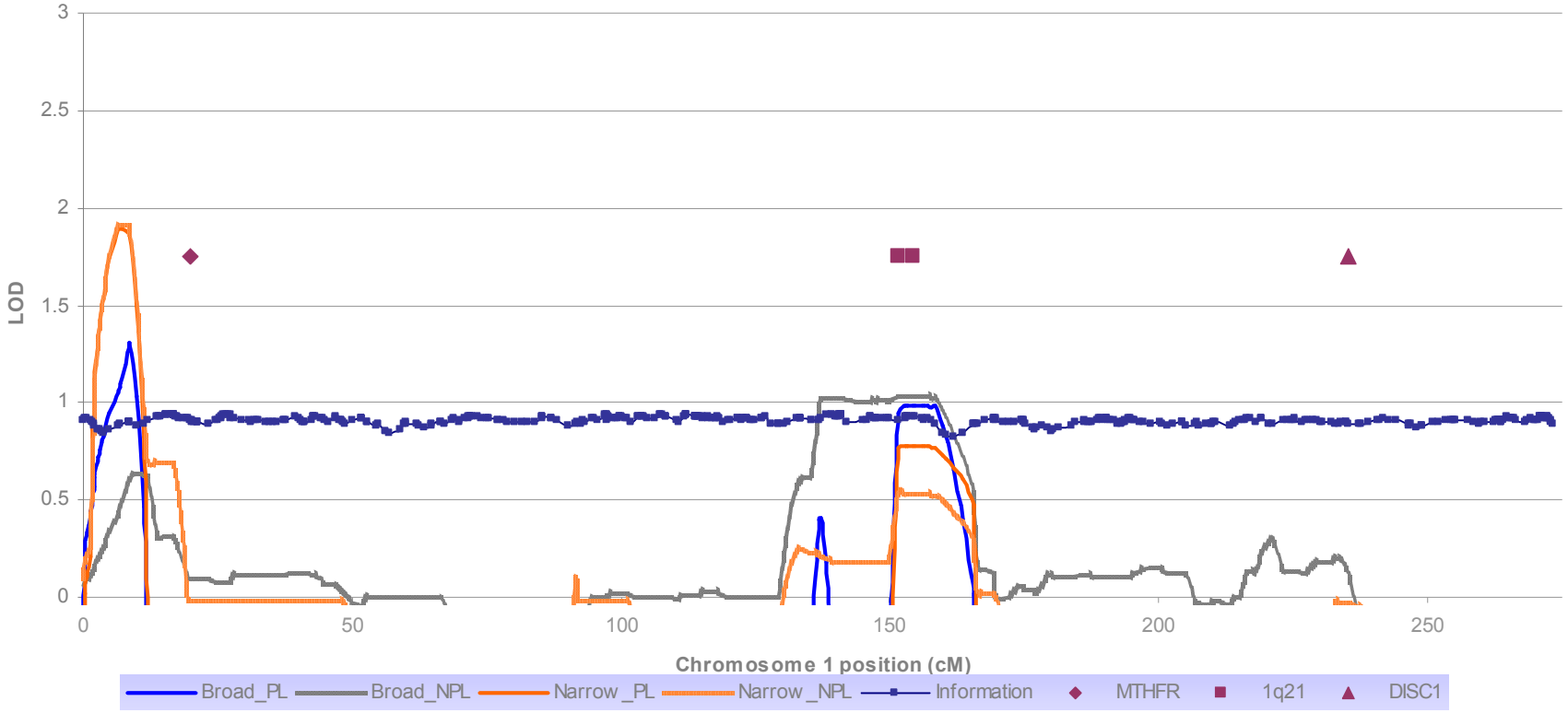


Figure 6.12 Parametric and non-parametric linkage analysis on chromosome 1. The x-axis is the position on chromosome 1. The y-axis is the LOD score from parametric linkage analysis (PL) and the Kong and Cox exponential LOD score from non-parametric linkage analysis (NPL). The information content was calculated for each marker and is shown by a blue-filled square. The results for linkage analysis performed using both PL (full line) and NPL (dotted line) analysis methods on both broad (blue) and narrow (orange) phenotype model. The location of a candidate gene for bipolar disorder (*MTHFR*) and for schizophrenia (1q21 and *DISC1*), are shown.

6.3.2.2. Contribution of each sub-pedigree

Closer examination of the chromosome 1p36 peak from parametric linkage analysis on the narrow phenotype (LOD 1.9) in Figure 6.13 shows the majority of linkage evidence stems from sub-pedigree 1 (LOD 1.45), with a contribution from sub-pedigree 4 (LOD 0.45) at the maximum peak at rs1870509 (6.43cM, 4,477,168bp).

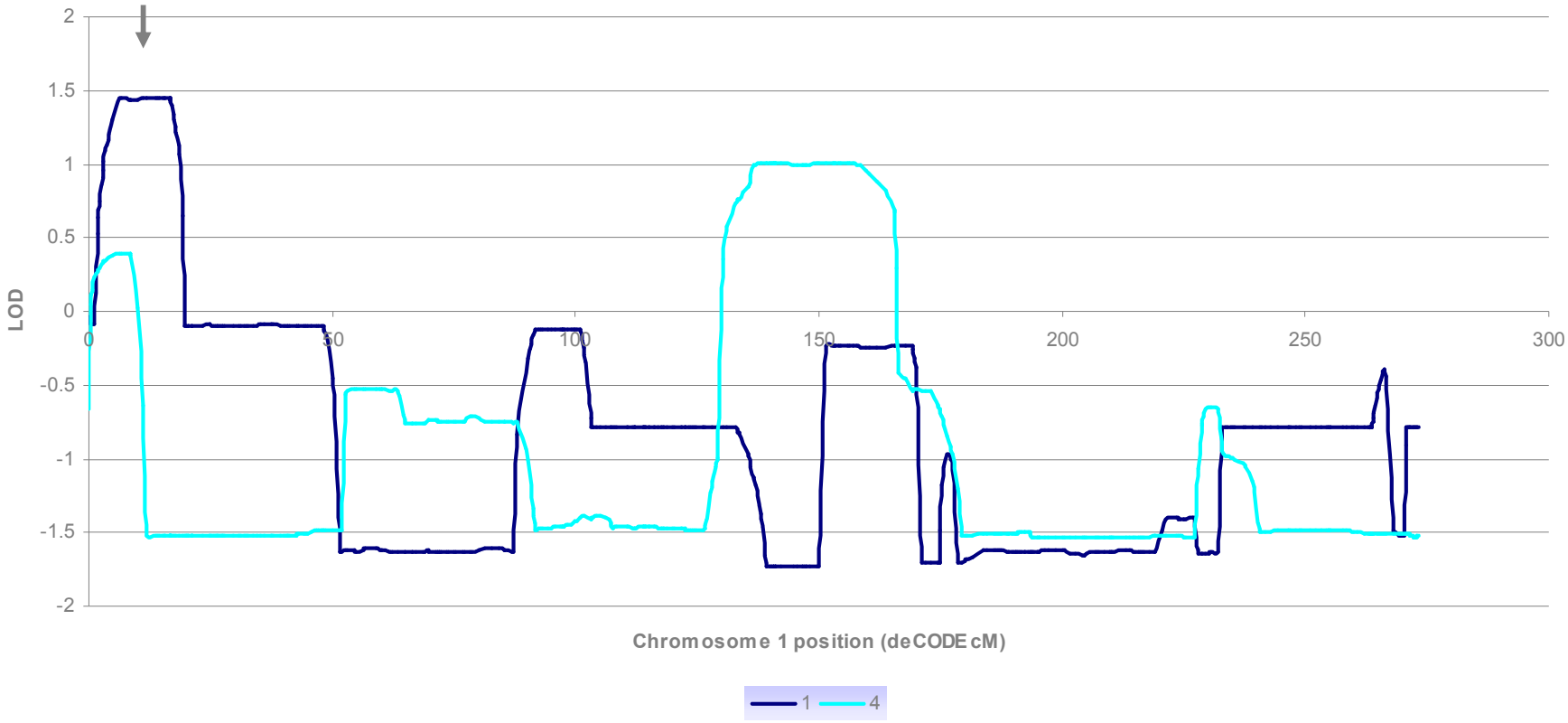


Figure 6.13 Parametric linkage analysis in sub-pedigrees 1 and 4 on chromosome 1p36 for narrow phenotype. The x-axis is the position on chromosome 1 from the p-terminal to the q-terminal. The y-axis is the LOD score at each location on chromosome 1. The sub-pedigrees informative for the narrow model were sub-pedigree 1 and 4. The grey arrow points to the position of the maximum LOD score at rs1870509 (6.43cM).

6.3.2.3. Contribution of a single affected individual

Figure 6.14 illustrates that the linkage peak on chromosome 1p36 was not affected by the removal of any single affected individual from the linkage analysis. Six of the affected individuals contributed to the linkage peak. However, sample 102 did not contribute to the linkage peak at 6.4-8.6cM, as the LOD score increased from LOD 1.9 (orange line) to LOD 2.11 (green line) with removal of sample 102 from the linkage analysis. With the removal of sample 150 (blue line), there was a linkage peak at 136cM [LOD 1.1 rs360636 (136.73cM, 115,132,947bp) - rs2057127 (137.05cM, 115,525,807bp)] and with the removal of sample 47 (violet line), there was a linkage peak at 150cM [LOD 1.8 rs2317232 (151.4cM, 154,501,115bp) – rs1062174 (159.68cM, 159,224,667bp)]. Both peaks were considered linkage artefacts as they were dependent on information from one individual.

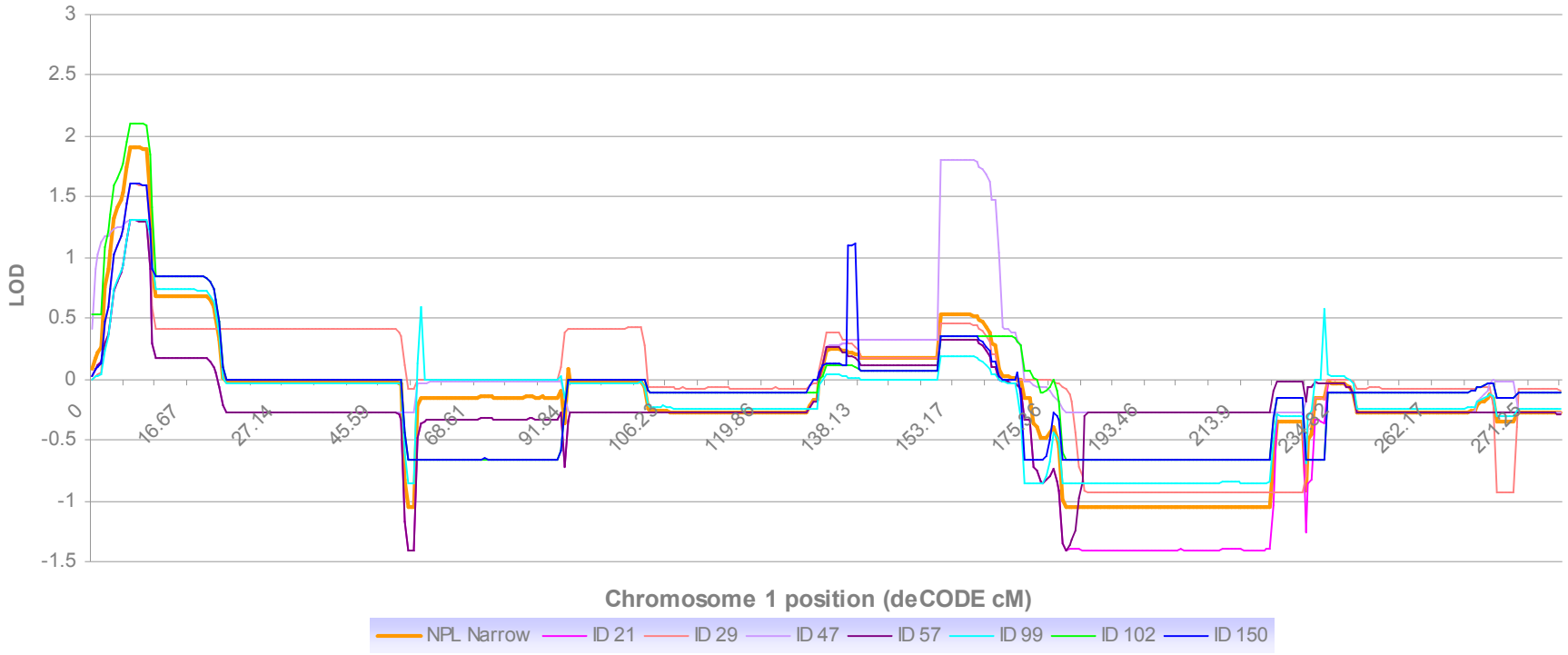


Figure 6.14 Contribution of a single affected individual to the linkage results on chromosome 1. The x-axis is the chromosomal genetic position from p-terminal to q-terminal. The y-axis is the Kong and Cox exponential LOD score from non-parametric linkage analysis (NPL). The orange line is the result including all seven affected individuals from the informative sub-pedigrees 1 and 4. The other coloured lines are linkage analysis performed with the individual specified affection status set to unknown. At the chromosome 1p36 peak the pink line for ID21, the lavender line for ID 47 and the violet line for ID57 are masked by the turquoise line for ID99, the coral line for ID29 is masked by the blue line for ID150. Individuals 21, 29, 47 and 57 are in sub-pedigree 1 and individuals 99, 102 and 150 are in sub-pedigree 4.

6.3.3. Chromosome 8p21

6.3.3.1. Parametric & non-parametric linkage analyses combined

The peak on chromosome 8 overlaps in the parametric and non-parametric linkage analysis for the narrow phenotypic model as shown in Figure 6.15. Greater evidence was provided by the non-parametric linkage analysis (dotted line), with both methods of analysis showing evidence for linkage to the same chromosomal location. The information content was consistent across the maximum linkage peak at 40cM. Within this chromosomal location are three candidate genes for psychiatric illness; a *pericentriolar material 1 gene (PCM1)* and a *calcineurin γ catalytic subunit (PPP3CC)* which are discussed further in section 6.5, and *neuregulin-1 (NRG1)* which was discussed in chapter 1. Their positions are marked on Figure 6.15.

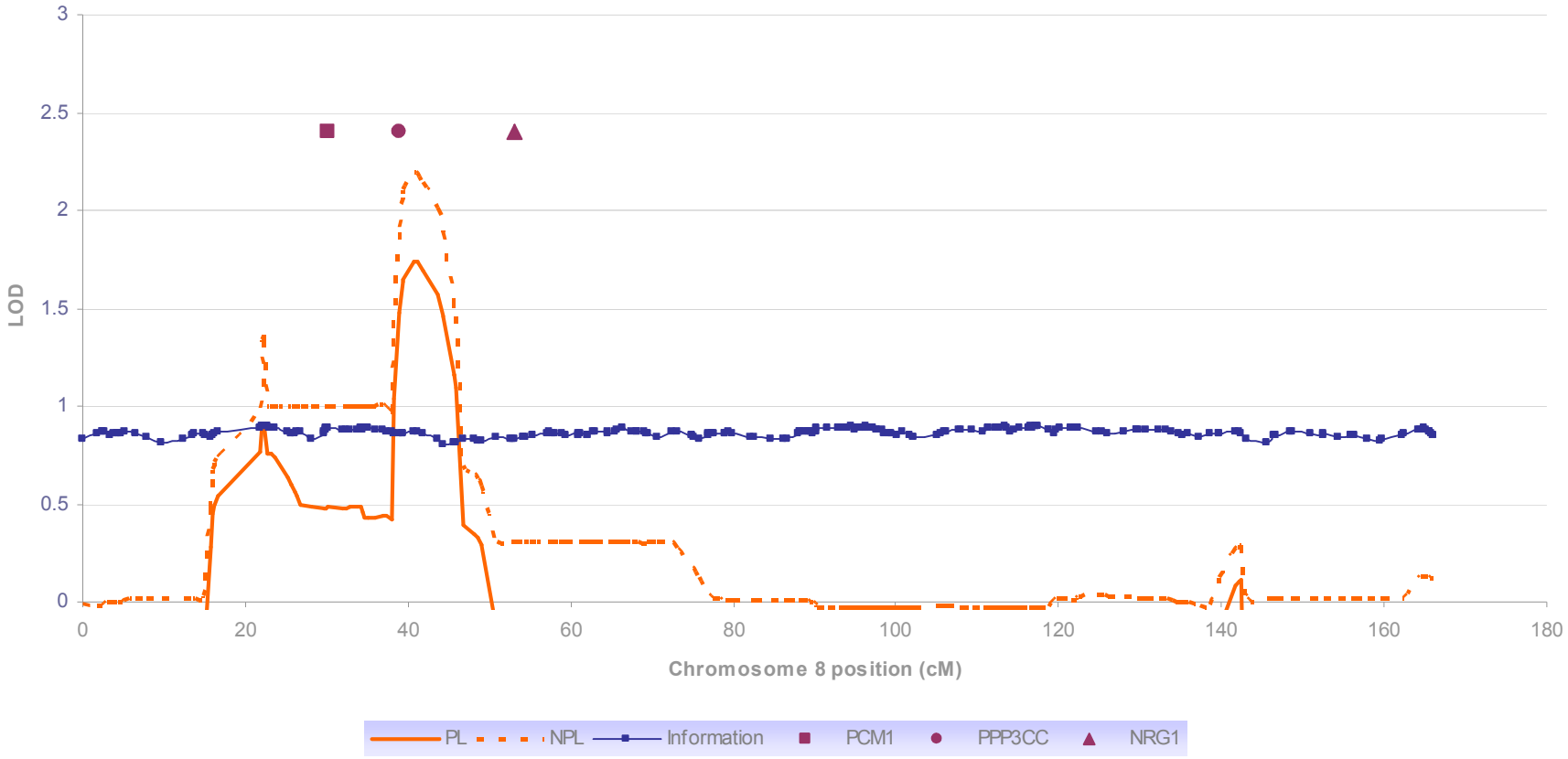


Figure 6.15 Parametric and non-parametric linkage analysis on chromosome 8. The x-axis is the chromosome 8 genetic position. The y-axis is the LOD score from parametric linkage (PL) analysis and the Kong and Cox exponential LOD score from non-parametric linkage (NPL) analysis. The information content was calculated for each marker and is shown by a blue-filled square. The location of the candidate genes for psychiatric illness, PPP3CC, NRG1 and PCM1 are marked by a square, circle and triangle respectively.

6.3.3.2. Contribution of each sub-pedigree

To establish the robustness of the linkage results, two checks were performed by investigating the contribution of each sub-pedigree and each affected individual to the linkage result. Figure 6.16 shows the two sub-pedigrees that contribute to the linkage peak on chromosome 8p21. There were two regions i) 20-50cM for sub-pedigree 1 by the blue line and ii) 40-80cM for sub-pedigree 4 by the turquoise line, that overlap at 40cM to produce the linkage peak. This linkage peak at rs2466216 was produced from a LOD 1.1 from sub-pedigree 1 (four affected individuals) and a LOD score 0.6 from sub-pedigree 4 (three affected individuals).

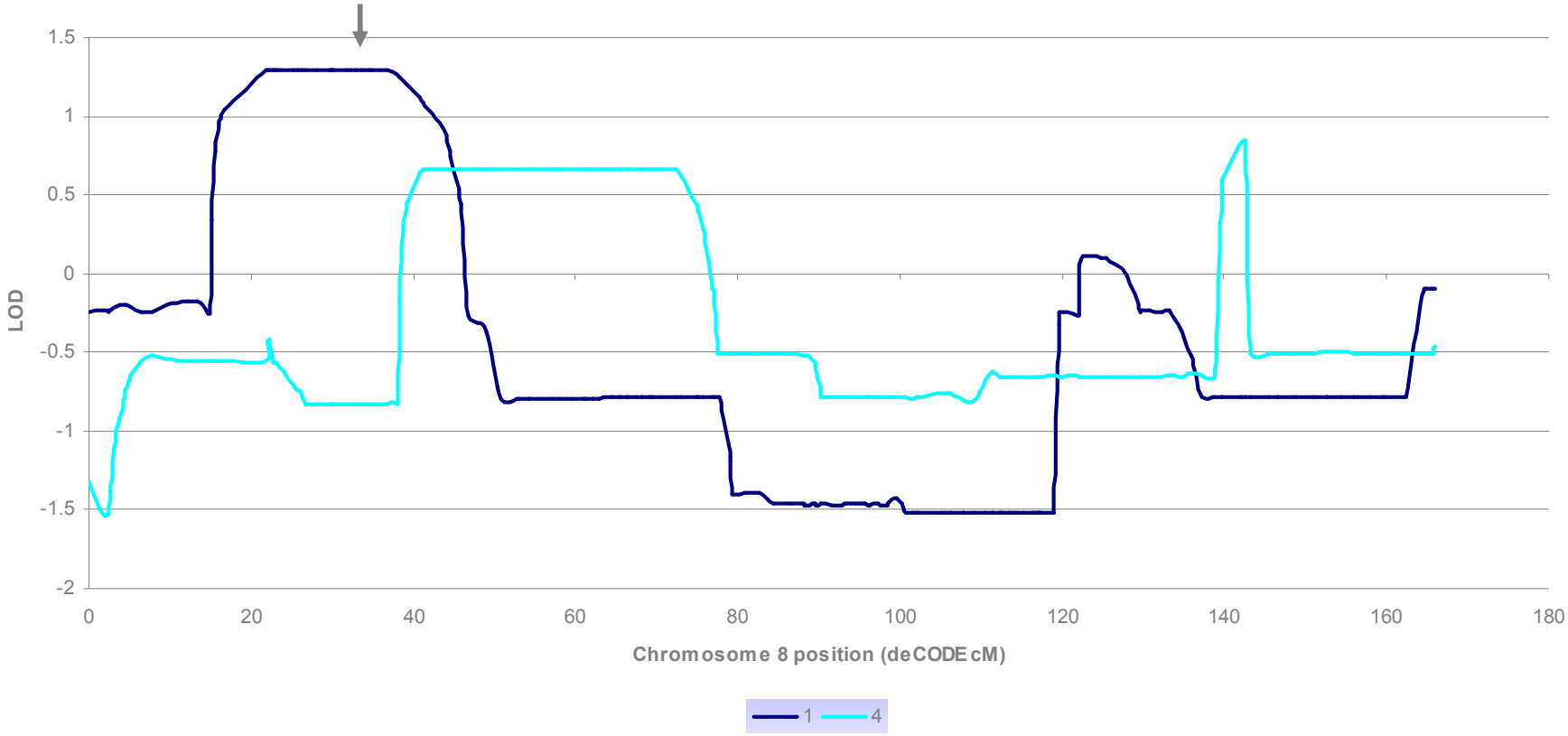


Figure 6.16 Parametric linkage analysis in sub-pedigrees 1 and 4 on chromosome 8 under the narrow phenotype model. Two families were informative for linkage analysis in the narrow model, sub-pedigree 1 (dark blue) and sub-pedigree 4 (turquoise). The x-axis is the position on chromosome 8. The y-axis is the LOD score at each location on chromosome 8. The grey arrow points to the position of the maximum LOD score at rs2466216 (40.75cM, 22,831,661bp).

6.3.3.3. Contribution of a single affected individual

Figure 6.17 shows the contribution of each affected individual to non-parametric linkage analysis on chromosome 8p. Each coloured line shows linkage analysis with one bipolar disorder affection status sample removed at a time. These lines can be compared to the orange line which shows the non-parametric linkage analysis for all individuals with bipolar disorder diagnoses. It is clear that all individuals contributed to the maximum LOD score at rs2466216 (40.75cM), where the LOD score decreased when each individual was removed. This showed that the maximum LOD score was robust to removal of single affected individuals. However, the positional effects were not the same; for example removal of sample 99 (turquoise line) extends the location of the linkage peak by 6cM (15.64cM-21.74cM) to the p-terminus and removal of sample 21 (pink line) extend the location of the linkage peak by 5cM (45.88-50.87cM) in the direction of the q-terminal. The position of the candidate gene NRG1 was marked on Figure 6.17 to compare its position to the linkage peaks. However, the linkage peaks do not extend to the gene. In addition, removal of samples 99 (turquoise line) and 150 (dark blue line), produced a linkage peak of LOD 2.1 at three markers; rs2028806 (22.15cM, 8,967,448bp), rs753012 (22.17cM, 9,019,806bp) and rs735449 (22.22cM, 9,140,857bp) that was greater than the linkage peak when all individuals were included (orange line), showing that this linkage peak was less robust. This is likely to be due to a different spectrum of recombination events in individuals affected with bipolar disorder at this chromosomal location. The LOD 1.05 peak at 164cM shows that removal of sample 29 caused an artefact of linkage analysis.

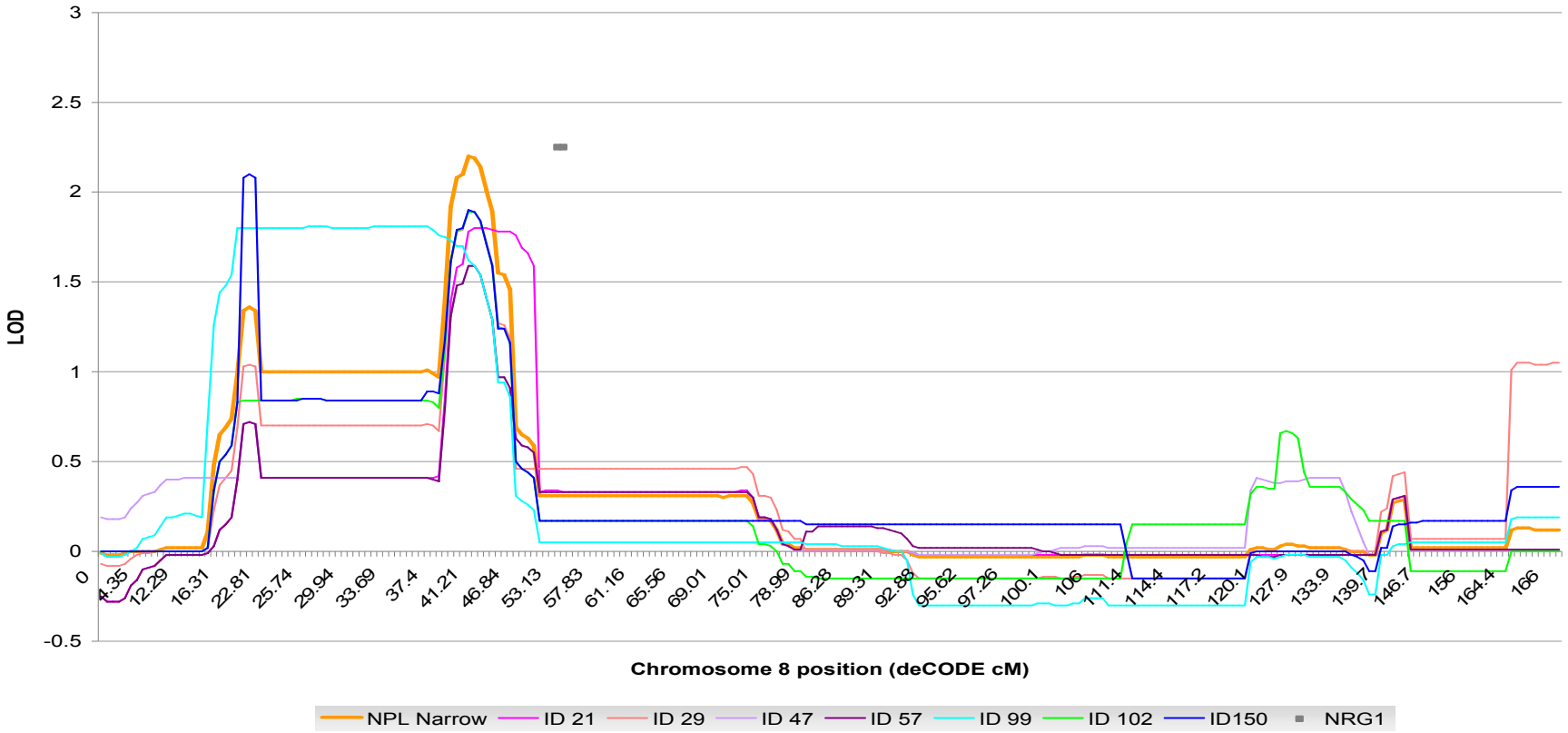


Figure 6.17 Contribution of a single individual to the linkage results on chromosome 8. The x-axis is the chromosomal genetic position from p-terminal to q-terminal. The y-axis is the Kong and Cox exponential LOD score from non-parametric linkage (NPL) analysis. The orange line is the result including all seven affected individuals from the informative sub-pedigrees 1 and 4. The other coloured lines are linkage analysis performed with the individual specified affection status set to unknown. At the chromosome 8p21 peak the coral and green line for ID 29 and 102 respectively is masked by the blue line for ID 150 and the lavender line for ID 47 is masked by the violet line for sample 57. The location of the candidate gene NRG1 is indicated. Individuals 21, 29, 47 and 57 are in sub-pedigree 1 and individuals 99, 102 and 150 are in sub-pedigree 4.

6.4. *Inspection of Haplotypes*

6.4.1. Preface

The aim of haplotype analysis was to estimate the haplotypes segregating with disease through the family, to position recombination breakpoints and thus define the minimum interval that should contain the disease gene. Haplotypes were estimated for the regions with suggestive evidence for linkage on chromosomes 4p14-p16, 1p36 and 8p21. MERLIN estimated the haplotypes corresponding to the most likely pattern of gene flow in the sub-pedigrees, as described in section 2.5.9. Visual inspection of the marker genotypes was then used to check the construction of the estimated haplotype by MERLIN. The haplotype that segregated with bipolar disorder samples was defined as the "linked haplotype" for each locus. The samples with recurrent major depression were not used to define the linked haplotype, as the risk of phenocopies was greater, because recurrent major depression is more frequent than bipolar disorder in the general population. The presence or absence of the linked haplotype was noted for 10 samples with bipolar disorder diagnoses and 22 samples with recurrent major depression diagnoses. Three samples were not included in the haplotype analysis: sample 19 (recurrent major depression) and 105 (bipolar disorder) failed to genotype and sample 176 (bipolar disorder) had no parental or offspring genotyping information that could be used to define haplotype phase.

6.4.2. Definition of chromosome 4p14-p16 haplotype segregating with illness.

Prior to performing the haplotype analysis, there were two types of evidence to consider on chromosome 4p14-p16. Firstly, Figure 6.9 shows the linkage evidence for chromosome 4p15-p16 [LOD=2.7, parametric and non-parametric linkage analysis under the narrow phenotypic model (bipolar disorder samples only)]. As shown in Table 6.7, the predicted location of the disease gene according to the 1-

LOD support interval, for both parametric and non-parametric linkage analysis under the narrow phenotypic model, conservatively, was a region of 43.35cM on 4p14-p15, from rs878283 (6,539,741bp, 12.46cM) to rs725292 (36,965,713bp, 55.81cM). In addition, from the parametric linkage analysis under the narrow phenotypic model, no locus heterogeneity was detected (α=1) for 41.8cM from rs726111 (6,030,055bp, 10.85cM) to rs902659 (32,011,211bp, 52.65cM). Secondly, haplotype analysis with microsatellite markers and visual construction of the haplotypes had been performed before this study (Le Hellard, Lee et al. 2007). This defined a haplotype that segregated with bipolar disorder diagnoses. The telomeric boundary of the linked haplotype was defined by a recombination event in sample 48 (recurrent major depression) that was inherited by samples 55 (recurrent major depression) and 57 (bipolar disorder) distal to rs5017682 (6,633,843bp, 12.6cM). The centromeric recombination event was defined by a recombination event before microsatellite marker, stb131K9.p1 (26,279,795bp, 46cM) in sample 102 (bipolar disorder) and inherited by the bipolar disorder samples 143 and also is samples 99 and 150. The position of the linked haplotype is marked on Figure 6.18 as "previously defined linked haplotype."

Haplotype analysis was performed within the sub-pedigrees for all SNP markers on chromosome 4, as illustrated in Figure 5.1-Figure 5.10. The microsatellite markers were not included in the current haplotype analysis, as data were not available for many of the samples. Missing data causes difficulties when resolving haplotypes, and haplotype analysis with missing data can be error-prone (Schaid, McDonnell et al. 2002). Figure 6.18 illustrates the results of the haplotype analysis and shows the estimated haplotypes for all affected individuals in the sub-pedigrees. The linked haplotype, labelled in orange, segregated with the 10 samples from individuals with bipolar disorder.

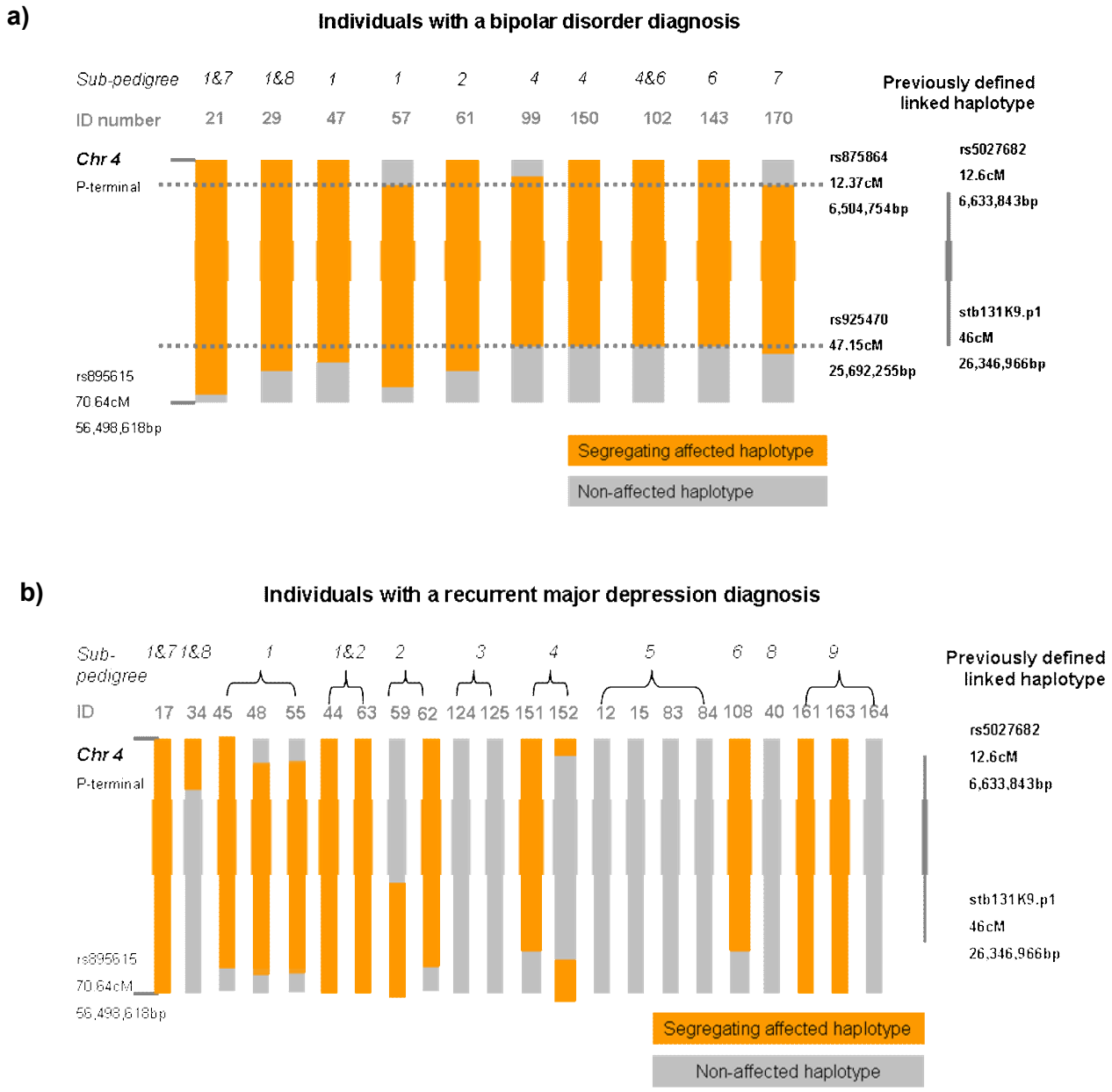


Figure 6.18 Illustration of haplotype analysis on chromosome 4p14-p16 linkage peak. Haplotype analysis was performed on 10 individuals with a diagnosis of bipolar disorder (a) from the p-terminal of chromosome 4 and beyond the linkage peak. The haplotype that segregates with illness in each individual is illustrated by the orange bar. Haplotype analysis was also performed on 22 individuals diagnosed with recurrent major depression (b).

At the telomeric end of the chromosome 4 linkage region, a recombination event after rs875864 (6,504,754bp, 12.37cM), was observed in sample 48, that was inherited by samples 55 and 57. A recombination event at this position was also observed in sample 170. Towards the centromere on chromosome 4, there was a recombination event observed in sample 102 before rs1456087 (27,180,675bp, 47.14cM) that was inherited by sample 143. The recombination event was observed in sample 99 and also, sample 150 that was inherited by sample 151. The recombination events in samples 57 and 102 conservatively defined the minimum interval of the linked haplotype segregating with bipolar disorder. Taking the predicted recombination events from haplotype analysis on the SNP data into account, the haplotype that segregates with bipolar disorder on chromosome 4 spanned a region of 34.78cM (12.37cM-47.15cM).

There were many obstacles relating to the haplotype analysis. One problem was that the recombination events predicted by MERLIN were not entirely correct and required validation by visual inspection. For example, the recombination event that MERLIN predicted at the telomeric region of chromosome 4p15-p16 was not accurate. Visual inspection of the designated haplotypes showed that MERLIN was not sufficiently conservative in predicting the breakpoint. MERLIN predicted a recombination event between markers rs11734660 (15.49cM) and rs1557816 (19.02cM). However, the preceding two markers, rs875579 (12.86cM) and rs13147693 (14.11cM), were not informative for haplotype phase determination and so the location of the recombination breakpoint could only confidently determined after marker rs875864 (12.37cM).

Another difficulty was that samples were contained in more that one sub-pedigree. For example, three samples diagnosed with bipolar disorder were in different sub-pedigrees; sample 21 (sub-pedigrees 1 and 7), sample 29 (sub-pedigrees 1 and 8) and sample 102 (sub-pedigrees four and six). The same haplotype should be estimated regardless of the sub-pedigree. However, this was not necessarily the case. The

same haplotype was estimated for samples 21 and 29 in both their respective sub-pedigrees. However, the haplotype for sample 102 was not the same in sub-pedigrees 4 and 6. A recombination event between 44.72cM and 47.14cM was predicted in sample 102 in sub-pedigree 4, but not in sub-pedigree 6. Visual inspection of this region showed there was indeed a recombination event in sample 102 at this position, which was inherited by its descendents (samples 103, 143, 144 and 156).

As described above, the individuals with recurrent major depression were not used to define the linked haplotype. It was, however, interesting to investigate how many carried the linked haplotype as shown in Figure 6.18 (b). Previous microsatellite information had shown that 11 samples with recurrent major depression carried the linked haplotype (samples 17, 44, 45, 48, 55, 62, 63, 108, 151, 161 and 163) and 11 did not have the linked haplotype (samples 12, 34, 40, 59, 83, 84, 124, 125, 152 and 164). Haplotype analysis in this study showed that eight samples with recurrent major depression carried the linked haplotype. The haplotype analysis could not determine whether the linked haplotype was carried by samples 108 (sub-pedigree 6), 161 and 163 (sub-pedigree 9) due to insufficient information. The samples with recurrent major depression: 17, 44, 45, 48 and 63 transmitted the defined linked haplotype to an affected offspring, ensuring greater confidence in haplotype analysis. There were no recombination events that would help to narrow down the minimum interval of the linked haplotype (12.6-46cM).

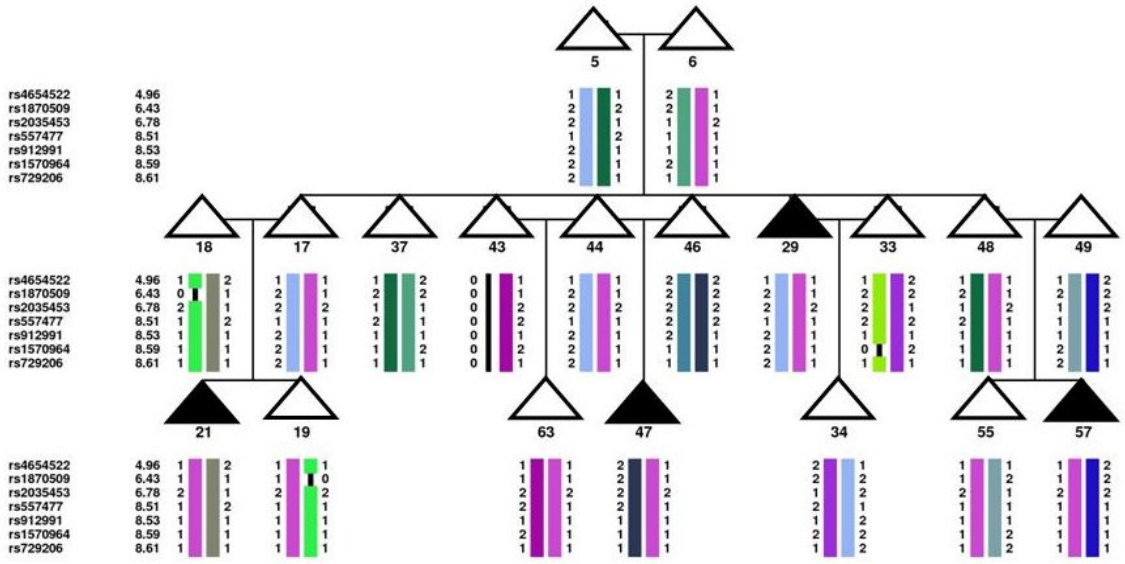
6.4.3. Definition of chromosome 1p36 haplotype segregating with illness

To define the chromosome 1p36 linked haplotype, several pieces of linkage evidence were considered. Firstly, there was a region of suggestive linkage on chromosome 1p36 for the narrow phenotypic model (maximum non-parametric LOD 1.9), as shown in Figure 6.12. Secondly, the 1-LOD support interval, from both parametric

and non-parametric linkage analyses, narrowed the region of interest to 9.6cM (1.72cM-11.33cM). Thirdly, the location of genetic homogeneity under the linkage region (α=1) was 9.75cM (0.33cM-10.08cM). Finally, all but one individual (sample 102) contributed to linkage in this region, as shown in Figure 6.14.

Haplotype analysis covering >60cM on chromosome 1p36, was performed on the sub-pedigrees using MERLIN and by visual inspection. Within each sub-pedigree, the affected individuals shared a haplotype. Figure 6.19 shows the haplotype analysis for seven markers under the chromosome 1p36 linkage peak for three sub-pedigrees. Four of these markers, rs1870509-rs912991, resulted in the maximum LOD 1.9. Figure 6.19a shows that the haplotype which segregates with bipolar disorder in samples 21, 29, 47 and 57 is 1-1-2-1-1-1-1 (coloured in purple). Figure 6.19b shows that the haplotype that segregates with bipolar disorder in sub-pedigree 4, with samples, 102, 150 and 99 is 1-2-2-1-2-2-2 (coloured in pink). However, Figure 6.19c shows that the haplotype that segregates with bipolar disorder in sub-pedigree 6, with sample 102 again and sample 143 is coloured in green, 1-1-2-1-2-2-2. This green haplotype in sub-pedigree 6 is different from the pink haplotype segregating with illness in sub-pedigree 4. It is evident that one chromosome (pink) from 102, segregates with bipolar disorder in sub-pedigree 4 and the other chromosome (green) segregates with bipolar disorder, in sub-pedigree 6. The definition of the haplotypes in sub-pedigrees 4 and 6 for sample 102 rests on one marker, rs1870509, that is heterozygous. Comparison of these haplotypes between the sub-pedigrees, indicates sharing of four consecutive markers (1-1-2-1; rs4654522, rs1870509, rs2035453 and rs557477) between sub-pedigrees 1 and 6 and indicates sharing of the last three markers (2-2-2; rs912991, rs1570964, rs729206) between sub-pedigrees 4 and 6.

a)



b)

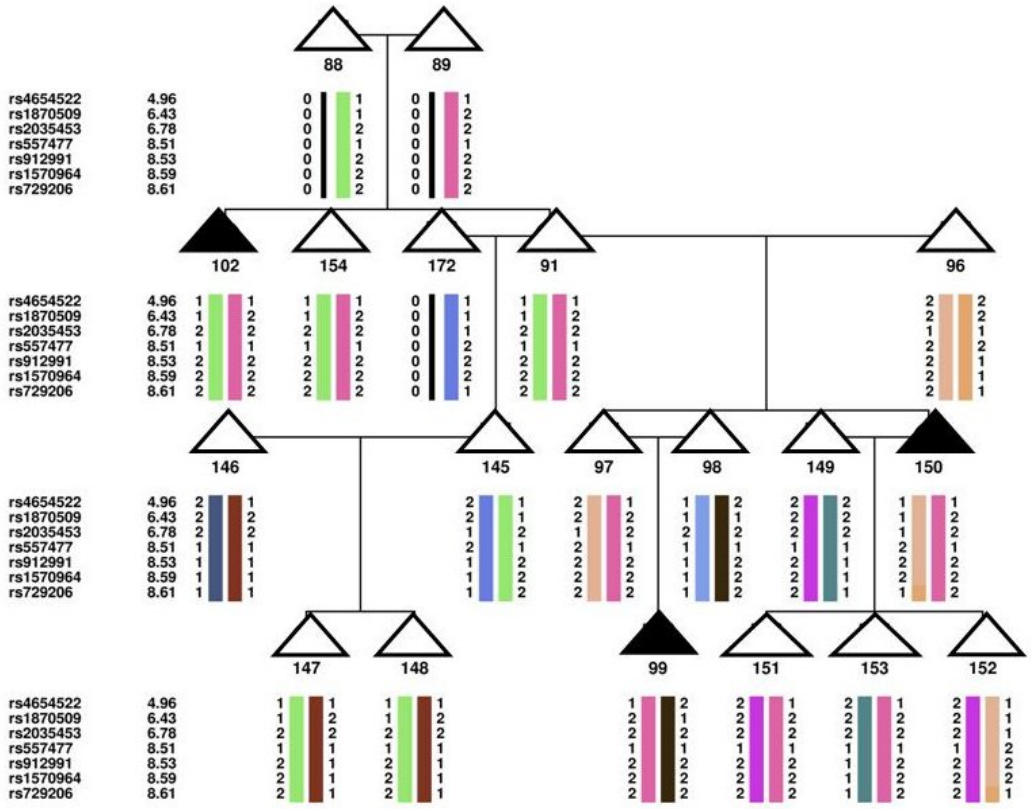


Figure 6.19 Haplotype analysis on chromosome 1p36 linkage peak in sub-pedigree 1 (a), 4 (b) and 6 (c). Continued on the next page.

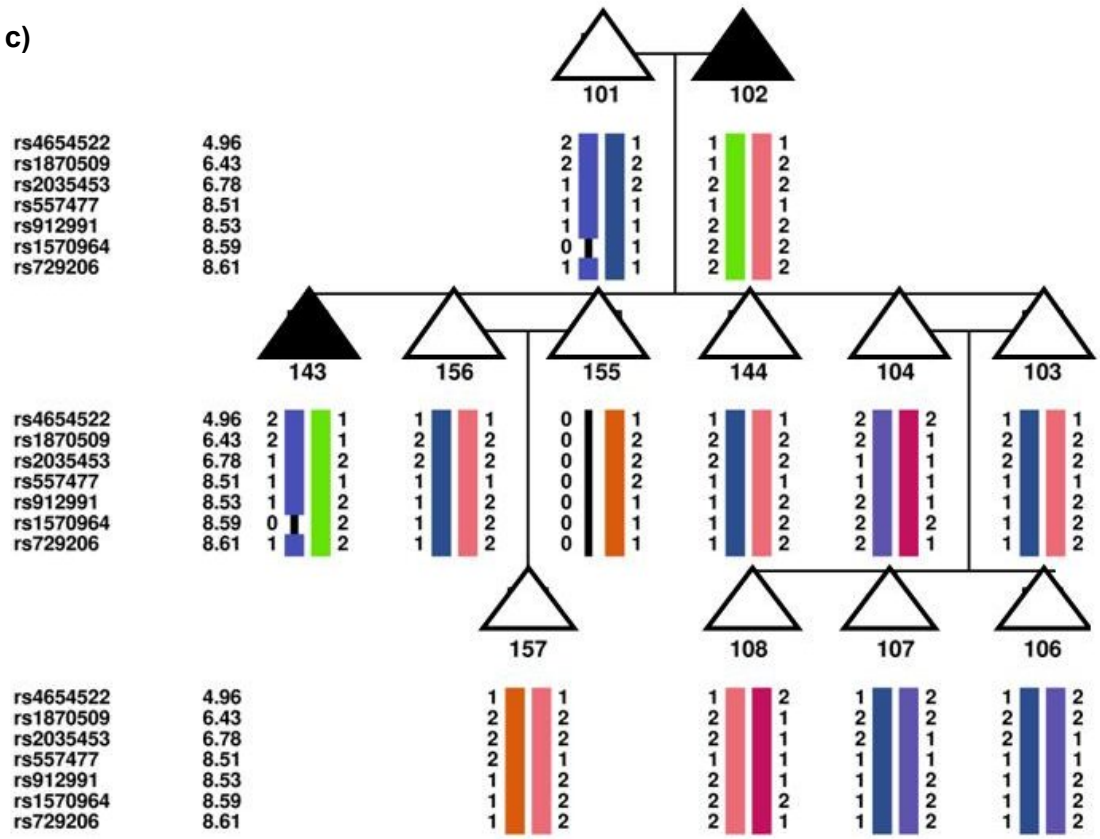


Figure 6.19 Haplotype analysis on chromosome 1p36 linkage peak in sub-pedigree 1, 4 and 6. The name and position in deCODE cM of the seven markers tested are listed to the left of the pictures with sub-pedigree 1 (a), 4 (b) and 6 (c).The haplotypes were predicted using MERLIN software and the pedigrees were drawn with Haplopainter. The gender is disguised by triangles (the usual diamonds were not used as they could not disguise the symbols fully) and Haplopainter requires gender information to draw pedigree. The filled symbols indicate a diagnosis of bipolar disorder.

The other samples with a diagnosis of bipolar disorder do not have haplotypes that segregate with bipolar disorder in sub-pedigrees 1, 4 and 6. Sample 61 in sub-pedigree 2 does not inherit the haplotype segregating with disease from sample 44. Sample 44 is the sample which is contained in both sub-pedigrees 1 and 2. Sample 170 in sub-pedigree 7 does not inherit the haplotype segregating with disease from sample 17. Sample 17 is contained in both sub-pedigree 1 and 7.

The linkage peak on chromosome 1p36 seems to be a result of the four markers in common between sub-pedigree 1 and 6, and the three markers in common between sub-pedigree 4 and 6. However, upon investigation at the chromosome level, the haplotypes segregating with illness are not the same. The most probable explanation is that the markers are identical by state (IBS) and not identical by descent (IBD), possibly resulting in the positive linkage results. This is difficult to establish conclusively with biallelic markers, but future analysis with multiallelic microsatellite markers would resolve this issue.

6.4.4. Definition of chromosome 8p21 haplotype segregating with illness

Several aspects of the linkage investigations were collated to assist in the definition of a haplotype segregating with illness on chromosome 8p21. In the first instance, the linkage evidence for chromosome 8 was solely present in the narrow phenotype for both parametric and non-parametric linkage analysis as shown in Figure 6.15 (maximum non-parametric LOD 2.2). The 1-LOD support interval defined a region of 8.82cM (38.02cM-46.84cM) in parametric linkage analysis as Table 6.3 and a region of 25.1cM (21.74cM-46.84cM) in non-parametric linkage analysis. Secondly, genetic homogeneity (α=1) from the parametric linkage analysis was present for a region of 7.63cM (38.25cM-45.88cM). Finally, all individuals from sub-pedigrees 1 and 4 contributed to the linkage peak on chromosome 8p21 as shown in Figure 6.17.

Haplotype analysis was performed within the sub-pedigrees, to include the linkage peak and beyond using MERLIN and by visual inspection. Haplotypes were also analysed at a finer level, in a region defined by the five markers where non-parametric LOD scores were greater than two: rs900267 (39.34cM), rs2466216 (40.75cM), rs3924018 (41.21cM), rs310319 (41.91cM) and rs879958 (43.58cM). It was difficult to obtain phase for all five markers in many of the bipolar disorder cases. MERLIN did not predict the same haplotype phase for the same individual that was in two sub-pedigrees (sample 21 in sub-pedigrees 1 and 7). Furthermore, examination of the pedigree as a whole did not provide any extra phase information. It did show, however, that three samples were phased: samples 99 and 150 were phased 1-1-1-1-2, and sample 61 was phased 1-1-1-2-1. The remaining seven samples could not be phased. Of these samples, five samples (samples 21, 29, 57, 102 and 143) were compatible with haplotype 1-1-1-1-2, and seven samples (samples 21, 29, 47, 57, 102 and 170) were compatible with haplotype 1-1-1-2-1. However, all ten samples with a diagnosis of bipolar disorder are 1-1-1 for the first three markers (rs900267, rs2466216 and rs3924018). These markers that are IBS possibly contributed to the linkage result. Of the samples that were unable to be phased, four samples (samples 21, 29, 57 and 102) are homozygous at the three markers.

In summary, the establishment of haplotypes segregating with bipolar disorder was inconclusive for chromosome 1p36 and chromosome 8p21. The reason for this was the uninformative nature of biallelic SNPs. Although the markers themselves singly had sufficient informativeness: the average information content (IC) of the seven markers on chromosome 1p36 is IC=0.89 (range: 0.86-0.90) and the average information content over the five markers on chromosome 8p21 is IC=0.86 (range: 0.83-0.87), they did not provide enough diversity when segregated through the family. An ideal situation to establish haplotype phase is when for example, within a trio, where both parents are homozygous, but different to each other, for one marker (and of course, the more markers the better), transmit one chromosome each

to offspring that are heterozygous at this marker. Thus, this enables the source of the chromosome to be established and the phase of said chromosome. Undoubtedly, conclusive definition of the haplotypes could be achieved with more informative microsatellite markers in the future.

6.4.5. Limitations of haplotype analysis

One limitation of this haplotype analysis was the recombination events were detected by MERLIN. Although the haplotypes were visually inspected for consistencies and segregation, the apparent recombination event, may not be the location of the real recombination event, as the surrounding markers may not be informative (Speed and Waterman 1996). Conversely, recombination events may be missed due to uninformative markers (Hodge, Boehnke et al. 1999). Furthermore, MERLIN estimated the most likely haplotype, while this was useful there may have been other haplotypes almost as likely. This was one of the major drawbacks of using biallelic markers for linkage analysis.

A further impediment to haplotype construction for this large family was that its large size prevented haplotype construction as a whole. Naturally, estimating the gene flow in a pedigree would be more informative if the pedigree as a whole was considered. Despite this, the chromosome 4p14-p16 haplotype confirmed knowledge previously obtained through microsatellite analysis. However, appropriate interpretation of this haplotype analysis is important as the linkage peaks on chromosome 1p36 and 8p21 may be an artefact of allele frequencies or alleles shared IBS but not IBD.

The presence or absence of the linked haplotype on chromosome 4p15-p16 was not taken into account for healthy individuals. Bipolar disorder is not a fully penetrant disease so the presence of the linked haplotype in unaffected individuals could not be used to exclude chromosomal regions. Also the sample 176, which was not

included in the haplotype analysis, may carry the linked haplotype on chromosomes 4p15 and either potential haplotypes on chromosomes 1p36 and 8p21.

6.5. *Discussion*

6.5.1. Preface

The linkage analysis was successful in confirming linkage to chromosome 4p15-p16 and suggesting other regions of linkage on chromosome 1p36 and 8p21. Reinforcing the genetic evidence for chromosome 4 was important, considering the recruitment of new individuals to the study and the updated phenotypes. The key benefit of the analysis was to exclude other loci of major effect, while suggesting other minor risk factor loci.

Optimisation of the linkage study was of prime concern. Great care was taken to ensure well-defined affected individuals were used, by thorough review of case-notes and repeat interviews. In addition, a "cleaned" genotyping dataset was prepared, parameters for significant linkage were established and the robustness of reported linkage results was validated. The linkage results were robust to perturbations in the data, such as excluding affected individuals in turn and altering the allele frequencies. This was important as the allele frequencies were obtained by counting all genotyped individuals, which did not take into account family information. This could potentially bias towards one chromosome, if many offspring from one parent were genotyped. The linkage results were robust to changes in allele-frequencies, in turn making the results more credible. This is a noteworthy point as misspecification of a genetic model and allele frequencies can produce a false-positive LOD score (Clerget-Darpoux, Bonaiti-Pellie et al. 1986). Conversely, an over-dogmatic approach to the disease loci hunt may lead to an unfortunate oversight which hopefully has not occurred in this study. Finally, the effectiveness of any particular technique in linkage analysis, parametric and non-parametric linkage analysis will usually depend upon the true, but unknown,

genetic model applicable to the disease. The evidence for linkage to chromosomes 4p15-p16, 1p36 and 8p21 as loci of potential importance was supported by both parametric and non-parametric linkage analyses.

Notably, the linkage evidence was greatest in sub-pedigree 1, which had the maximum linkage information and the linkage evidence was present in the narrow phenotypic model. The linkage evidence for chromosome 4p14-p15 was increased, with the addition of microsatellite markers and has support in independent samples. The litmus test for chromosome 1p36 and 8p21 will be maintaining the linkage evidence and determination of a consistent haplotype that segregates with illness in the family. Firstly, addition of informative microsatellites on chromosomes 1p36 and 8p21 could increase the linkage evidence as shown in this study for chromosome 4 and resolve haplotype construction ambiguities. Secondly, genotyping of additional families would help discern whether the result is spurious or significant. Neither of these chromosomal regions was covered in the original genome scan. There were no markers on the p-terminal arm of chromosome 1 and only three microsatellite markers on the p-terminal arm of chromosome 8, with the nearest marker 20cM away from the linkage peak reported in this study (Blackwood, He et al. 1996).

6.5.2. Chromosome 4p15-p16 linkage evidence

Figure 6.20 illustrates the past and present genetic analyses on this family, which have shown the primary locus for susceptibility to bipolar disorder is chromosome 4p15-p16. The linkage peak defined in this study overlapped with the previously known “linked haplotype”, which includes the candidate genes GPR78 and PI4K2B, the “allele-sharing region” and the regions of association, as previously discussed in section 1.4. Although the LOD score did not reach 4 as originally reported in 1996 and again in 2006, it still met the genome-wide significance criteria (NPL LOD 3.6 under broad phenotypic model, as listed in Table 6.7). The reason for a decrease in

LOD score may be due to many factors including, splitting the family, reduced informativeness of SNPs compared to microsatellite markers and the addition of affected family members without the chromosome 4p haplotype.

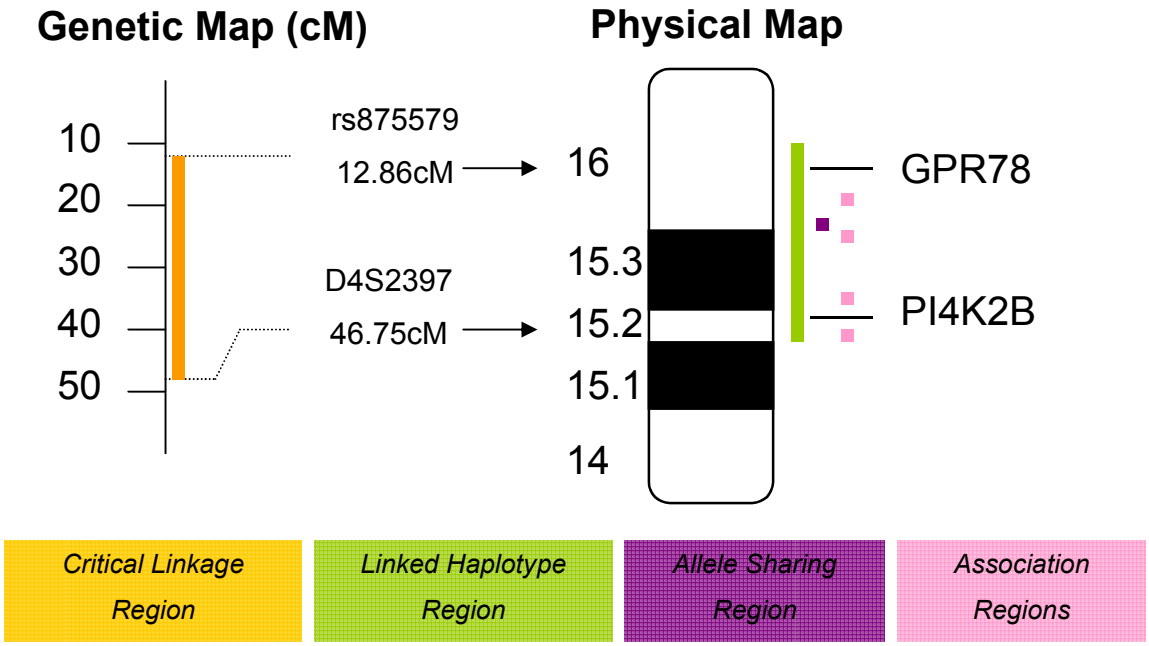


Figure 6.20 Hallmarks of the chromosome 4p15-p16 linkage region. The linkage region is defined by the haplotype analysis in Figure 6.18. The positions of the markers on the boundary of the region are marked by their genetic position with respect to deCODE cM and physical distance with respect to UCSC May 2004, NCBI Build 35 assembly of the human genome. The approximate location of interesting candidate genes, the allele sharing region and associated regions are marked.

In the future, it would be interesting to investigate this family for copy number variants as the chromosome 4p arm has many CNV hotspots (Redon, Ishikawa et al. 2006). Bioinformatic investigation of this region on chromosome 4p16.1 revealed eleven CNV loci on this band from the Database of Genomic Variants (Iafrate, Feuk et al. 2004). Furthermore, investigation of the DECIPHER database revealed four patients with chromosomal abnormalities and with CNV change (https://decipher.sanger.ac.uk/). There was phenotype information for one of the patients (610) with a copy number change indicating a deletion on chromosome 4p16.1;p16.3. This individual has neurological, thorax and stature abnormalities. With respect to the neurological phenotype, there were mental and cognitive abnormalities with mental retardation and developmental delay.

6.5.3. Chromosome 1p36 linkage evidence

In support of these candidate loci on chromosome 1p36 and 8p21, previous work has implicated them in susceptibility to psychiatric illness (Kato 2007). Recent evidence has supported an association between *methylenetetrahydrofolate reductase (MTHFR)* on chromosome 1p36 and bipolar disorder, recurrent major depression and schizophrenia (Gilbody, Lewis et al. 2007). In addition, chromosome 1p36 was implicated in linkage studies for recurrent major depression. For example, McGuffin et al, 2005 performed a genome-wide linkage analysis on 497 sib-pairs concordant for recurrent major depression and found suggestive linkage on chromosome 1p36 (13.8cM-21.8cM) where the LOD score for female-female pairs >3 (McGuffin, Knight et al. 2005). This region is just centromeric to the region reported here. Also, a report of a chromosome 1p36 peak in four European bipolar disorder families is outside of this support interval (LOD 3.3) (Schumacher, Kaneva et al. 2005). However, there are other linkage reports to chromosome 1p36 that overlap with the region, such as those described for mood disorders (maximum LOD 3.6) (Zubenko, Maher et al. 2003) and bipolar disorder in a single family (LOD 3.1) (Curtis, Kalsi et al. 2003).

This region was also identified in IBD haplotype sharing analysis in psychotic patients from Israel (LOD 1.5) (Kohn, Danilovich et al. 2004).

Below the chromosome 1p linkage peak are 22 known genes. The known genes were those genes with RefSeq status that are protein-coding genes taken from the NCBI mRNA reference sequences collection. None of these 22 genes show immediate potential for candidacy to psychiatric illness. The maximum LOD score falls at a predicted gene, *adherens junction associated protein 1 (AJAP1)*, also known by its Ensembl name as *SHREW1*, and the gene product is involved in cell invasion (Schreiner, Ruonala et al. 2007). An interesting candidate may be *KCNAB2* (a potassium voltage-gated channel shaker-related) gene that has putative functions including regulating neurotransmitter release and neuronal excitability. *WDR8* was also noteworthy, as this gene encodes a member of the WD repeat protein family. WD repeats are minimally conserved regions of approximately 40 amino acids, which may facilitate formation of heterotrimeric or multiprotein complexes. Members of this family are involved in a variety of cellular processes; including cell cycle progression, signal transduction, apoptosis and gene regulation. Although this chromosome 1p36 locus did not reach whole-genome wide significance levels, it should be subjected to further investigation.

6.5.4. Chromosome 8p21 linkage evidence

As illustrated in Figure 6.15, the chromosome 8 linkage peak is in the same location as candidate genes *PCM1* and *PPP3CC*. *PCM1* is a pericentriolar material 1 gene that is significantly linked to schizophrenia, is associated with orbitofrontal grey matter volumetric deficits (Gurling, Critchley et al. 2006) and has been shown to bind to kendrin (PCNT) (Li, Hansen et al. 2001) and NUDEL (Guo, Yang et al. 2006), which are both binding partners of DISC1. *PPP3CC* is a calcineurin γ catalytic subunit on 8p21.3 that is associated with schizophrenia in a United States and South African dataset (Gerber, Hall et al. 2003), in a Taiwanese dataset (Liu, Fann et al. 2007) and

associated with bipolar disorder in a French cohort (115 bipolar disorder cases and 97 controls) (Mathieu, Miot et al. 2008). There was a negative association reported with PPP3CC to schizophrenia in a Japanese cohort (457 schizophrenia cases and 429 controls) (Kinoshita, Suzuki et al. 2005) and in a cohort of European descent (1,870 schizophrenia and schizoaffective disorder cases and 2,002 controls) (Sanders, Duan et al. 2008).

This region was also highlighted from linkage analysis on 40 families from the US and Israel on bipolar disorder with psychotic features (multipoint LOD score 3.46) (Park, Juo et al. 2004). Furthermore, many linkage studies have implicated this region in schizophrenia (reviewed in (Blackwood, Pickard et al. 2007). In particular, a recent linkage study on expression of the susceptibility gene for schizophrenia, *dystrobrevin binding protein (DTNBP1)* in the CEPH pedigrees showed genome-wide suggestive evidence for linkage to chromosome 8p (maximum LOD 2.77). The authors suggest this locus on chromosome 8p12-8p21.1 (rs2169315-rs726908), that overlaps with the 1-LOD support interval reported here, exerts a trans-acting effect on *DTNBP1* expression (Bray, Holmans et al. 2008).

In fact, there were 156 known genes under this linkage peak. The maximum LOD score on chromosome 8p21 was at rs2466215 which falls within the *PEBP4* (phosphatidylethanolamine-binding protein 4) gene and the same arguments apply to this gene which is involved in the PI pathway, for a candidacy for susceptibility to bipolar disorder as does *PI4K2B*, described in chapter 1. In short, the PI pathway is inhibited by lithium for therapeutic effect in bipolar disorder.

6.5.5. Caveats

As in any statistical study, there are many caveats. Firstly, the failure of certain samples to genotype reduced the power of the linkage study. For example, individual 105 with a diagnosis of bipolar disorder was not genotyped successfully

on the Illumina IVb panel. Experiments performed in-house showed the individual did not have the chromosome 4p defined linkage haplotype. It would be interesting to determine if the person increased linkage evidence to chromosome 1p36 or 8p21. Individuals 12, 15 and 19 failed genotyping on one quarter of the 6,008 SNP Illumina IVb panels, which meant there was missing genotype information on chromosome 4 for these three individuals with recurrent major depression diagnoses. Thus, the missing samples and genotyping information may cause a reduction in the power to detect linkage and it would be prudent to consider re-genotyping the individuals for further information.

Secondly, the decision to split the family was based on the capacity of current linkage analysis programs. The linkage analysis may be improved considering the family as a whole, as these linkage programmes develop to give exact linkage results for large families (Tong and Thompson 2008), as opposed to the estimated results produced by programmes such as Simwalk2.

Thirdly, the simulation results would benefit from an increased number of simulations, which were not performed due to computational limitations. A desirable number of simulations would be 10,000 replicates for both parametric and non-parametric linkage analysis, which would ensure stringent significant level estimations. It must be noted that the simulations were performed on the initial whole genome screen marker data and do not supply significance levels for linkage analysis when the marker coverage was increased with microsatellites on chromosome 4. According to Lander and Kruglyak "increasing marker density around linkage hits has a strong effect on the false-positive rate"(Kruglyak and Daly 1998). More simulations should be performed to evaluate the level of significance as the addition of microsatellites to chromosome 4p14-p16 pushed the LOD scores over the suggestive linkage threshold into the significant linkage category Furthermore, the results do not reach a Bonferroni significance threshold level for this study where, P=0.05/n, where n is the number of independent associations, such that n (m-

1) for the testing of n loci with m alleles each correction factor. For this study, a *P*-value of $8x10^{-6}$ would be required for evidence of significant linkage and no markers remotely achieved this. However, Bonferroni correction is recognized to be overly conservative (Perneger 1998; Ott 1999).

A final argument involves the widespread use of sex-averaged maps for the genetic distance of markers. The ratio of male and female genetic map distances vary across the human genome. One multipoint linkage study showed that when either all or no parental genotypes are available, there was no important difference in the expected maximum LOD score or location estimates of the disease locus between the sex-averaged map and true sex-specific maps (Fingerlin, Abecasis et al. 2006). However, if there are genotypes available for only one parent and the recombination rate is higher in females, then the LOD score using the sex-averaged map was inflated compared to the sex-specific map if only the mothers were genotyped and deflated if the fathers were genotyped. Conversely when the recombination rate was higher in males, the LOD was inflated if only the fathers were genotyped and deflated if the mothers were genotyped. In this study, the majority of affected trios have either both or neither parents genotyped [13 (3 bipolar disorder) affected individuals with both parents genotypes and 11 (5 bipolar disorder) affected individuals with no parents genotyped]. There are 9 (2 bipolar disorder) individuals where the genotyping information is from the mother and 1 bipolar disorder individual where only the father is genotyped. However, without knowing the recombination rate over the whole-genome, it is not easy to predict the effect on LOD scores.

6.5.6. Future work

As previously discussed, bipolar disorder is thought to be a complex genetic disorder. The primary aim of this study was to investigate this family for a single susceptibility genetic factor for bipolar disorder and recurrent major depression. In keeping with this hypothesis, chromosome 4 remains the most significant region for

susceptibility to bipolar disorder and recurrent major depression. The loci on chromosome 8p21 and 1p36 may contribute to the hypothesis in the following ways: i) they may be false positives, a result that could be confirmed by further linkage and haplotype analysis, using more informative markers ii) the loci may combine to contribute to illness with chromosome 4p15-16 and finally, iii) they are single risk factors. However, the latter is not the case as it appears that the chromosome 4 locus contributes to psychiatric illness in the family.

It would be interesting to investigate the loci revealed in this study for additive, multiplicative or epistatic effects. For example, a linkage study for Hirschsprung disease (a complex genetic disease that is a common heritable cause of intestinal obstruction) have found evidence for oliogenic inheritance with one major loci and two other contributing loci (Gabriel, Salomon et al. 2002). This study investigated the magnitude of the genetic effects from the three loci to estimate how they interact, whether in an additive, multiplicative, mixed multiplicative or epistatic fashion. They found that all loci were involved in the manifestation of the illness. Application of these notions to this study may prove beneficial in dissecting complex mental illness. At present, this analysis requires expensive computational resources. Recently novel score statistics have been developed to simultaneously scan for two disease susceptibility loci in pedigree data and these may be interesting to apply to this dataset in the future (Schaid, McDonnell et al. 2007). It also would be interesting to condition the family for the presence of the chromosome 4p15-p16 linked haplotype and re-analyse the whole genome linkage scan on the basis of this stratification. This method was successful in identifying additional risk loci in a family with DISC1 risk alleles (Hennah, Tomppo et al. 2007).

6.5.7. Summary

To summarise, it is worth reiterating the importance of studying the genetics of large families for complex disorders (Venken and Del-Favero 2007). Any one family

is likely carry a single gene variant of major effect or multiple rare genes. The ability to identify DISC1 from linkage analysis of a single Scottish family was a proof in point. Negative whole-genome association studies and whole-genome linkage meta-analyses on chromosome 4p should not detract from the evidence provided here. These studies by definition, look for common variants whereas, large families lend themselves to identification of a rare variants of large effect, as in the case of the *DISC1* locus (Macgregor, Visscher et al. 2002). This study demonstrated that the focus of the hunt for a disease locus for bipolar disorder and the related phenotype, recurrent major disorder in this family, remains on chromosome 4p15-p16. This study has also hinted at other suggestive genetic risk factors that should also be considered. It remains to be seen whether these statistically suggestive linkage findings yield any findings of biological importance.

Chapter 7
Concluding Remarks

7. Concluding Remarks

7.1. *Preface*

Bipolar disorder and recurrent major depression are complex psychiatric illnesses. Studies to uncover the genetic component of these disorders are in their early stages. Linkage analysis has identified evidence of linkage to bipolar disorder and recurrent major depression with markers on chromosome 4p15-p16 in a large Scottish family and three smaller families (Blackwood, He et al. 1996; Le Hellard, Lee et al. 2007). High-resolution haplotype analysis of this region defined a linked haplotype that segregates with bipolar disorder. This region was also supported by allele-sharing analysis and a case-control association study (Christoforou, Le Hellard et al. 2007; Le Hellard, Lee et al. 2007). This study contributes to this research by adopting two approaches i) a candidate gene study and ii) a high resolution whole-genome linkage scan.

7.2. *Candidate Gene Study*

PI4K2B was a worthy candidate gene in many respects. Positional evidence from linkage, allele sharing and association analyses had implicated this genomic region on chromosome 4p15. Genetic evidence has also suggested other members of the phosphoinositide signalling pathway as susceptibility factors for bipolar disorder, schizophrenia and recurrent major depression (Kato 2007). Additionally, functional evidence supported members of the phosphoinositide signalling pathway as candidate genes, as components of the phosphoinositide pathway are inhibited by lithium for therapeutic effect in bipolar disorder (Berridge, Downes et al. 1989). The salient advantage of this study was the opportunity to investigate this candidate gene, in lymphoblastoid cell lines from a large Scottish family with bipolar disorder.

PI4K2B expression studies at the allele-specific mRNA and protein level were performed using lymphoblastoid cell lines. RNA expression was measured at the allele-specific level using Taqman assays. The advantage of this method was the sensitivity obtained by comparing the linked haplotype to a series of different control alleles from related individuals. Three *PI4K2B* SNPs were used to measure gene expression levels. Initially, standard curves of homozygote gDNA dilutions were created, to look for a deviation from the expected 1:1 allelic ratio in cDNA. Despite optimisation attempts, the standard curves did not achieve the required sensitivity to detect small differences. However, the standard curves were useful to show that a gross deviation in allele-specific expression was not present. A second method, which compared expression in cDNA to that seen in gDNA from heterozygous samples, proved more sensitive. There was, however, no evidence for a difference between samples with the linked haplotype and those without. PI4K2B protein expression was also quantified. This method required much optimisation, to ensure the specificity of antibodies, the sensitivity of protein detection and the reliability of quantification methods. At the protein level, there was no evidence to suggest a difference in PI4K2B expression between samples with the linked haplotype and those without.

Although the expression studies showed no evidence for abnormal *PI4K2B* RNA and protein levels, it must be stated that there were a number of technical limitations to this study. Firstly, the number of heterozygotes available for two of the SNPs used to measure allele-specific expression was small and therefore, not well powered to detect an expression difference between groups. Secondly, regardless of efforts to maintain constant protocols for RNA, DNA and protein preparations throughout the study, experimental variation was an inevitable side-effect. Thirdly, and despite efforts to find alternatives (PeakPicker method in place of Taqman Assays and fluorescent antibody detection to improve Western blotting), no technique was sufficiently sensitive or reliable to measure *PI4K2B* expression differences.

The candidacy of *PI4K2B* was also tested by increasing the marker density in the *PI4K2B* genomic region in a case-control association study. This study showed evidence for association of schizophrenia, but not bipolar disorder, with tagging SNPs from the *PI4K2B* genomic region. Two-marker haplotypes were also associated with bipolar disorder and schizophrenia. The association of SNP rs313548 with schizophrenia withstood permutation analysis and may suggest the presence of disease susceptibility variant in this region, but replication studies are required to confirm or refute this.

7.3. *Whole Genome Linkage Study*

The evidence for a bipolar disorder locus on chromosome 4 arose from linkage analysis on a large Scottish family, performed in 1996. Here, the linkage evidence for bipolar disorder and recurrent major depression was re-examined in the same family. This was important because additional affected family members had been recruited and advances in technology made it feasible to cover all chromosome regions more uniformly and more densely than had been previously possible.

Stringent genotyping and pedigree error checks were performed to ensure an optimised dataset for analysis. Markers that could potentially inflate linkage evidence were removed, such as markers in LD with each other, markers that failed a proportion of genotyping, markers that failed Mendelian segregation checks and markers that failed Hardy-Weinberg equilibrium tests. Pedigree errors, including sample duplication, were also detected and resolved. Additionally, computation constraints necessitated sub-dividing the whole family for linkage analysis.

A whole-genome linkage scan was performed using both parametric and non-parametric methods. Two phenotypic models were tested; a narrow phenotypic model with only bipolar disorder cases and a broad phenotypic model with bipolar

disorder and recurrent major depression cases. Genome-wide suggestive evidence was observed on chromosomes 4p15-p16, 8p21 and 1p36 (non-parametric LOD for narrow phenotypic model 2.4, 1.9 and 2.2 respectively). Following the addition of chromosome 4 microsatellite markers genome-wide evidence for significant linkage was observed on chromosome 4p15-p16 (non-parametric LOD 3.6 for broad phenotypic model). The robustness of the linkage evidence was tested by varying allele frequencies, by removing one affected individual at a time and by performing simulation analysis to determine the significance of the results. The three linkage peaks withstood these tests. Importantly, there are previous linkage and association studies to bipolar disorder, schizophrenia and recurrent major depression that support these three loci. Haplotype analysis demonstrated that a consistent haplotype segregated with bipolar disorder on chromosome 4p15-p16. Haplotypes that segregated with bipolar disorder on chromosome 1p36 and 8p21 were also defined. This analysis provided stronger evidence for the segregation of a consistent haplotype with most cases of bipolar disorder on chromosome 8, than on chromosome 1. However, no firm conclusions can be drawn without further analysis with more informative microsatellite markers.

The analysis clearly supported the evidence for a susceptibility locus of bipolar disorder and recurrent major depression on chromosome 4p15-p16 and identified other genetic loci that may confer risk to psychiatric illness. The lack of evidence supporting chromosome 4p15-p16, 1p36 and 8p21 from whole-genome association studies should not depreciate the data provided here. These studies by definition look for common variants and were successful in identifying association of SNPs for Crohn's disease, diabetes and cancer. One reason for this success was the ability to define reliable and valid phenotypes, which is not yet feasible for bipolar disorder.

There are many caveats to linkage analysis. Firstly, the three linkage peaks did not reach the threshold required for highly significant linkage, as established by the permutation analysis (LOD>5.32) and may be false positive results. Linkage

artefacts can occur for many reasons including errors in diagnoses, gender specification, marker allele frequencies, map order, map distances, genotype calling, DNA quality and family relationships. Despite testing extensively for these inaccuracies, some errors are almost undetectable and may have been present (Pompanon, Bonin et al. 2005). Secondly, the linkage results did not narrow the search for the disease gene as the 1-LOD support regions cover large regions of chromosome 4p15-p16, 1p36 and 8p21.

7.4. *Future Work*

This study showed that there was no direct evidence for a role of *PI4K2B* in susceptibility to bipolar disorder from lymphoblastoid cell lines. However, the technical methods and analyses developed could be applied to other candidate genes. Nevertheless, this candidate gene still merits further investigation to assess its potential role in psychiatric illness, because of the significant case-control association study result that withstood permutation testing. In the future, *PI4K2B* could be tested for association to bipolar disorder and schizophrenia in other Scottish samples, as well as those derived from other populations. A replication study should follow stringent criteria to allow clear and unambiguous interpretation of the results, to establish or refute association (Chanock, Manolio et al. 2007). Furthermore, sequencing of the risk haplotypes in samples with schizophrenia could be applied to search for functional variants.

The linkage evidence could be enhanced by performing linkage analysis using new computational methods, that promise accurate LOD score calculations, when the pedigree size and marker number are large (Tong and Thompson 2008). In addition, it would be interesting to perform a more comprehensive analysis on interacting genetic loci. Also, it would be prudent to re-genotype the samples that failed genotyping as they may provide additional information. Moreover, it is imperative to follow-up the linkage peaks, by performing haplotype analysis using multiallelic

microsatellite markers, to establish inheritance of a consistent haplotype with illness.

Indeed new technologies can be embraced in the search for the genetic basis to bipolar disorder. In particular, the linkage peaks reported in this study are ideal candidates for genome resequencing. Recent technological advances make genome resequencing a salient prospect for the near future, incorporating sequencing technologies such as Solexa from Illumina, SOLiD™ from Applied Biosystems and 454 Life Sciences™ from Roche Applied Science. Sequencing of the large genomic regions could survey protein coding, intronic and intergenic sequences for structural variations and point mutations (Stratton 2008).

The availability of large-scale, high-throughput and high-quality assays should enable the discovery of crucial disease variants. For example, microarray gene expression analysis aims to understand complex functional mechanisms and can be applied to the search of unknown exons, protein-DNA interactions, protein-RNA interactions and structural analysis (Hoheisel 2006). In the future, gene expression profiles for bipolar disorder patients would be tremendous to understand the biological process of the illness, such has been performed previously in studies of cancer analysis where gene expression profiling helped to identify new subtypes and predict clinical outcomes (Nevins and Potti 2007). Access to relevant biological tissue, specific for the illness itself or a reliable proxy, will be key to the long-term applicability of these methods.

Emerging technologies also allow investigations beyond the present capabilities. One example is the ability to assay for chromosomal inversions by single-molecule haplotyping, which is an improvement on current cytogenetic techniques (Turner, Shendure et al. 2006). Another example is the inclusion of copy number variants probes on whole genome SNP arrays, such as the Genome-wide Human SNP Array 6.0 from Affymetrix or the Human 1M BeadChip from Illumina. Both technologies

have greater than one million SNP and CNV probes for whole genome genotyping applications, which is double the coverage than was available 12 months ago. Prudent application of this technology to genetic studies of bipolar disorder will hopefully improve the understanding of the biological basis of the illness.

7.5. *Conclusions*

The search for genetic susceptibility to bipolar disorder and recurrent major depression is still in its infancy. This study has made a contribution to this search in two respects, i) *PI4K2B* is no longer a preferred candidate gene in the large Scottish family, as there was no evidence for a difference in allele-specific and protein expression between samples, with the linked haplotype and those without the linked haplotype and ii) chromosome 4p15-p16 remains the priority location for a genetic susceptibility factor to bipolar disorder in this family, as there are no other loci of major effect, while other minor loci are suggested.

8. References

Abecasis, G. R., S. S. Cherny, et al. (2002). "Merlin--rapid analysis of dense genetic maps using sparse gene flow trees." Nat Genet **30**(1): 97-101.

Abramoff, M. D., Magelhaes, P.J., Ram, S.J. (2004). ""Image Processing with ImageJ"." Biophotonics International **11**(7): 36-42.

Acharya, J. K., P. Labarca, et al. (1998). "Synaptic defects and compensatory regulation of inositol metabolism in inositol polyphosphate 1-phosphatase mutants." Neuron **20**(6): 1219-29.

Als, T. D., H. A. Dahl, et al. (2004). "Possible evidence for a common risk locus for bipolar affective disorder and schizophrenia on chromosome 4p16 in patients from the Faroe Islands." Mol Psychiatry **9**(1): 93-8.

American Psychiatric Association (2000). Diagnostic and statistical manual of mental disorders : DSM-IV-TR. Washington, DC., American Psychiatric Association.

Amos, C. I., W. V. Chen, et al. (2006). "High-density SNP analysis of 642 Caucasian families with rheumatoid arthritis identifies two new linkage regions on 11p12 and 2q33." Genes Immun **7**(4): 277-86.

Asherson, P., R. Mant, et al. (1998). "A study of chromosome 4p markers and dopamine D5 receptor gene in schizophrenia and bipolar disorder." Mol Psychiatry **3**(4): 310-20.

Atz, M. E., B. Rollins, et al. (2007). "NCAM1 association study of bipolar disorder and schizophrenia: polymorphisms and alternatively spliced isoforms lead to similarities and differences." Psychiatr Genet **17**(2): 55-67.

Badner, J. A. and E. S. Gershon (2002). "Meta-analysis of whole-genome linkage scans of bipolar disorder and schizophrenia." Mol Psychiatry **7**(4): 405-11.

Bakker, S. C., M. L. Hoogendoorn, et al. (2007). "The PIP5K2A and RGS4 genes are differentially associated with deficit and non-deficit schizophrenia." Genes Brain Behav **6**(2): 113-9.

Balla, A., G. Tuymetova, et al. (2002). "Characterization of type II phosphatidylinositol 4-kinase isoforms reveals association of the enzymes with endosomal vesicular compartments." J Biol Chem **277**(22): 20041-50.

Baron, M. (2001). "The search for complex disease genes: fault by linkage or fault by association?" Mol Psychiatry **6**(2): 143-9.

Baron, M. (2002). "Manic-depression genes and the new millennium: poised for discovery." Mol Psychiatry **7**(4): 342-58.

Barrett, J. C., B. Fry, et al. (2005). "Haploview: analysis and visualization of LD and haplotype maps." Bioinformatics **21**(2): 263-5.

Baum, A. E., N. Akula, et al. (2008). "A genome-wide association study implicates diacylglycerol kinase eta (DGKH) and several other genes in the etiology of bipolar disorder." Mol Psychiatry **13:** (197-207).

Beaulieu, J. M., S. Marion, et al. (2008). "A beta-arrestin 2 Signaling Complex Mediates Lithium Action on Behavior." Cell **132**(1): 125-36.

Benedetti, F., A. Bernasconi, et al. (2004). "A single nucleotide polymorphism in glycogen synthase kinase 3-beta promoter gene influences onset of illness in patients affected by bipolar disorder." Neurosci Lett **355**(1-2): 37-40.

Bennett, C. N. and D. F. Horrobin (2000). "Gene targets related to phospholipid and fatty acid metabolism in schizophrenia and other psychiatric disorders: an update." Prostaglandins Leukot Essent Fatty Acids **63**(1-2): 47-59.

Bennett, R. L. and K. A. Steinhaus (1995). "Recommendations for standardized human pedigree nomenclature. Pedigree Standardization Task Force of the National Society of Genetic Counselors." Am J Hum Genet **56**(3): 745-52.

Berridge, M. J., C. P. Downes, et al. (1989). "Neural and developmental actions of lithium: a unifying hypothesis." Cell **59**(3): 411-9.

Bertelsen, A., B. Harvald, et al. (1977). "A Danish twin study of manic-depressive disorders." Br J Psychiatry **130**: 330-51.

Blackwood, D. H., A. Fordyce, et al. (2001). "Schizophrenia and affective disorders--cosegregation with a translocation at chromosome 1q42 that directly disrupts

brain-expressed genes: clinical and P300 findings in a family." Am J Hum Genet **69**(2): 428-33.

Blackwood, D. H., L. He, et al. (1996). "A locus for bipolar affective disorder on chromosome 4p." Nat Genet **12**(4): 427-30.

Blackwood, D. H., B. J. Pickard, et al. (2007). "Are some genetic risk factors common to schizophrenia, bipolar disorder and depression? Evidence from DISC1, GRIK4 and NRG1." Neurotox Res **11**(1): 73-83.

Bland, J. M. and D. G. Altman (1996). "The use of transformation when comparing two means." British Medical Journal **312**: 1153.

Boehnke, M. and N. J. Cox (1997). "Accurate inference of relationships in sib-pair linkage studies." Am J Hum Genet **61**(2): 423-9.

Bourgain, C. and E. Genin (2005). "Complex trait mapping in isolated populations: Are specific statistical methods required?" Eur J Hum Genet **13**(6): 698-706.

Bray, N. J., P. A. Holmans, et al. (2008). "Cis- and Trans- Loci Influence Expression of the Schizophrenia Susceptibility Gene DTNBP1." Hum Mol Genet **17** (8): 1169-74.

Bray, N. J., L. Jehu, et al. (2004). "Allelic expression of APOE in human brain: effects of epsilon status and promoter haplotypes." Hum Mol Genet **13**(22): 2885-92.

Bray, N. J., A. Preece, et al. (2005). "Haplotypes at the dystrobrevin binding protein 1 (DTNBP1) gene locus mediate risk for schizophrenia through reduced DTNBP1 expression." Hum Mol Genet **14**(14): 1947-54.

Brush, G. and L. Almasy (2003). "Pedigree and genotype errors in the Framingham Heart Study." BMC Genet **4 Suppl 1**: S41.

Brzustowicz, L. M., K. A. Hodgkinson, et al. (2000). "Location of a major susceptibility locus for familial schizophrenia on chromosome 1q21-q22." Science **288**(5466): 678-82.

Cardno, A. G., F. V. Rijsdijk, et al. (2002). "A twin study of genetic relationships between psychotic symptoms." Am J Psychiatry **159**(4): 539-45.

Cardon, L. R. and L. J. Palmer (2003). "Population stratification and spurious allelic association." Lancet **361**(9357): 598-604.

Carter, C. J. (2006). "Schizophrenia susceptibility genes converge on interlinked pathways related to glutamatergic transmission and long-term potentiation, oxidative stress and oligodendrocyte viability." Schizophr Res **86**(1-3): 1-14.

Chanock, S. J., T. Manolio, et al. (2007). "Replicating genotype-phenotype associations." Nature **447**(7145): 655-60.

Chapman, N. H. and E. M. Wijsman (1998). "Genome screens using linkage disequilibrium tests: optimal marker characteristics and feasibility." Am J Hum Genet **63**(6): 1872-85.

Cheng, R., S. H. Juo, et al. (2006). "Genome-wide linkage scan in a large bipolar disorder sample from the National Institute of Mental Health genetics initiative suggests putative loci for bipolar disorder, psychosis, suicide, and panic disorder." Mol Psychiatry **11**(3): 252-60.

Christoforou, A., S. Le Hellard, et al. (2007). "Association analysis of the chromosome 4p15-p16 candidate region for bipolar disorder and schizophrenia." Mol Psychiatry **12**(11): 1011-25.

Chubb, J. E., N. J. Bradshaw, et al. (2008). "The DISC locus in psychiatric illness." Mol Psychiatry **13**(1): 36-64.

Clayton, D. G., N. M. Walker, et al. (2005). "Population structure, differential bias and genomic control in a large-scale, case-control association study." Nat Genet **37**(11): 1243-6.

Clerget-Darpoux, F., C. Bonaiti-Pellie, et al. (1986). "Effects of misspecifying genetic parameters in lod score analysis." Biometrics **42**(2): 393-9.

Clerget-Darpoux, F. and R. C. Elston (2007). "Are linkage analysis and the collection of family data dead? Prospects for family studies in the age of genome-wide association." Hum Hered **64**(2): 91-6.

Conneally, P. M., J. H. Edwards, et al. (1985). "Report of the Committee on Methods of Linkage Analysis and Reporting." Cytogenet Cell Genet **40**(1-4): 356-9.

Cordeiro, Q., S. Zung, et al. (2007). "Chromosomal translocation t(1;4) (p21;p14) indicating possible susceptibility loci for schizophreniform disorder and mental retardation." J Neuropsychiatry Clin Neurosci **19**(3): 339.

Cottingham, R. W., Jr., R. M. Idury, et al. (1993). "Faster sequential genetic linkage computations." Am J Hum Genet **53**(1): 252-63.

Craddock, N., V. Khodel, et al. (1995). "Mathematical limits of multilocus models: the genetic transmission of bipolar disorder." Am J Hum Genet **57**(3): 690-702.

Craddock, N. and C. Lendon (1999). "Chromosome Workshop: chromosomes 11, 14, and 15." Am J Med Genet **88**(3): 244-54.

Curtis, D., G. Kalsi, et al. (2003). "Genome scan of pedigrees multiply affected with bipolar disorder provides further support for the presence of a susceptibility locus on chromosome 12q23-q24, and suggests the presence of additional loci on 1p and 1q." Psychiatr Genet **13**(2): 77-84.

Davis, S., M. Schroeder, et al. (1996). "Nonparametric simulation-based statistics for detecting linkage in general pedigrees." Am J Hum Genet **58**(4): 867-80.

Dean, F. B., S. Hosono, et al. (2002). "Comprehensive human genome amplification using multiple displacement amplification." Proc Natl Acad Sci U S A **99**(8): 5261-6.

Detera-Wadleigh, S. D., J. A. Badner, et al. (1999). "A high-density genome scan detects evidence for a bipolar-disorder susceptibility locus on 13q32 and other potential loci on 1q32 and 18p11.2." Proc Natl Acad Sci U S A **96**(10): 5604-9.

Devlin, B. and K. Roeder (1999). "Genomic control for association studies." Biometrics **55**(4): 997-1004.

Di Paolo, G., H. S. Moskowitz, et al. (2004). "Impaired PtdIns(4,5)P2 synthesis in nerve terminals produces defects in synaptic vesicle trafficking." Nature **431**(7007): 415-22.

Dimitrova, A., V. Milanova, et al. (2005). "Association study of myo-inositol monophosphatase 2 (IMPA2) polymorphisms with bipolar affective disorder and response to lithium treatment." Pharmacogenomics J **5**(1): 35-41.

Douglas, J. A., M. Boehnke, et al. (2000). "A multipoint method for detecting genotyping errors and mutations in sibling-pair linkage data." Am J Hum Genet **66**(4): 1287-97.

Drummond, A., B. Ashton, et al. (2007). "Geneious v3.5." from http://www.geneious.com/.

Duan, J., M. S. Wainwright, et al. (2003). "Synonymous mutations in the human dopamine receptor D2 (DRD2) affect mRNA stability and synthesis of the receptor." Hum Mol Genet **12**(3): 205-16.

Duan, S., R. Gao, et al. (2005). "A family-based association study of schizophrenia with polymorphisms at three candidate genes." Neurosci Lett **379**(1): 32-6.

Dudbridge, F. (2003). "Pedigree disequilibrium tests for multilocus haplotypes." Genet Epidemiol **25**(2): 115-21.

Duren W.L., E. M., Li M., and Boehnke M. (June 2004). "RELPAIR: A Program that Infers the Relationships of Pairs of Individuals Based on Marker Data. Version 2.0.1,."

Dyer, T. D., J. Blangero, et al. (2001). "The effect of pedigree complexity on quantitative trait linkage analysis." Genet Epidemiol **21 Suppl 1**: S236-43.

Easton, D. F., K. A. Pooley, et al. (2007). "Genome-wide association study identifies novel breast cancer susceptibility loci." Nature **447**(7148): 1087-93.

Ehm, M. and M. Wagner (1998). "A test statistic to detect errors in sib-pair relationships." Am J Hum Genet **62**(1): 181-8.

Elston, R. C. and J. Stewart (1971). "A general model for the genetic analysis of pedigree data." Hum Hered **21**(6): 523-42.

Endicott, J. and R. L. Spitzer (1978). "A diagnostic interview: the schedule for affective disorders and schizophrenia." Arch Gen Psychiatry **35**(7): 837-44.

Epstein, M. P., W. L. Duren, et al. (2000). "Improved inference of relationship for pairs of individuals." Am J Hum Genet **67**(5): 1219-31.

Evans, D. M. and L. R. Cardon (2004). "Guidelines for genotyping in genomewide linkage studies: single-nucleotide-polymorphism maps versus microsatellite maps." Am J Hum Genet **75**(4): 687-92.

Ewald, H., B. Degn, et al. (1998). "Support for the possible locus on chromosome 4p16 for bipolar affective disorder." Mol Psychiatry **3**(5): 442-8.

Falchi, M., P. Forabosco, et al. (2004). "A genomewide search using an original pairwise sampling approach for large genealogies identifies a new locus for total and low-density lipoprotein cholesterol in two genetically differentiated isolates of Sardinia." Am J Hum Genet **75**(6): 1015-31.

Fallin, D., A. Cohen, et al. (2001). "Genetic analysis of case/control data using estimated haplotype frequencies: application to APOE locus variation and Alzheimer's disease." Genome Res **11**(1): 143-51.

Farmer, A., A. Elkin, et al. (2007). "The genetics of bipolar affective disorder." Curr Opin Psychiatry **20**(1): 8-12.

Fingerlin, T. E., G. R. Abecasis, et al. (2006). "Using sex-averaged genetic maps in multipoint linkage analysis when identity-by-descent status is incompletely known." Genet Epidemiol **30**(5): 384-96.

Fishelson, M. and D. Geiger (2002). "Exact genetic linkage computations for general pedigrees." Bioinformatics **18 Suppl 1**: S189-98.

Frazer, K. A., D. G. Ballinger, et al. (2007). "A second generation human haplotype map of over 3.1 million SNPs." Nature **449**(7164): 851-61.

Gabriel, S. B., R. Salomon, et al. (2002). "Segregation at three loci explains familial and population risk in Hirschsprung disease." Nat Genet **31**(1): 89-93.

Ge, B., S. Gurd, et al. (2005). "Survey of allelic expression using EST mining." Genome Res **15**(11): 1584-91.

Gerber, D. J., D. Hall, et al. (2003). "Evidence for association of schizophrenia with genetic variation in the 8p21.3 gene, PPP3CC, encoding the calcineurin gamma subunit." Proc Natl Acad Sci U S A **100**(15): 8993-8.

Gilbody, S., S. Lewis, et al. (2007). "Methylenetetrahydrofolate reductase (MTHFR) genetic polymorphisms and psychiatric disorders: a HuGE review." Am J Epidemiol **165**(1): 1-13.

Gimelbrant, A., J. N. Hutchinson, et al. (2007). "Widespread monoallelic expression on human autosomes." Science **318**(5853): 1136-40.

Gladkevich, A., H. F. Kauffman, et al. (2004). "Lymphocytes as a neural probe: potential for studying psychiatric disorders." Prog Neuropsychopharmacol Biol Psychiatry **28**(3): 559-76.

Glaser, B., G. Kirov, et al. (2005). "Identification of a potential bipolar risk haplotype in the gene encoding the winged-helix transcription factor RFX4." Mol Psychiatry **10**(10): 920-7.

Gordon, D., C. Abajian, et al. (1998). "Consed: a graphical tool for sequence finishing." Genome Res **8**(3): 195-202.

Goring, H. H. and J. Ott (1997). "Relationship estimation in affected sib pair analysis of late-onset diseases." Eur J Hum Genet **5**(2): 69-77.

Gottesman, I. I. (1991). Schizophrenia genesis : the origins of madness. New York, Freeman.

Gould, T. D., J. A. Quiroz, et al. (2004). "Emerging experimental therapeutics for bipolar disorder: insights from the molecular and cellular actions of current mood stabilizers." Mol Psychiatry **9**(8): 734-55.

GraphPad_Software (version 4.03). "GraphPad Prism version 4.03 for Windows, GraphPad Software, San Diego California USA, www.graphpad.com."

Guo, J., M. R. Wenk, et al. (2003). "Phosphatidylinositol 4-kinase type IIalpha is responsible for the phosphatidylinositol 4-kinase activity associated with synaptic vesicles." Proc Natl Acad Sci U S A **100**(7): 3995-4000.

Guo, J., Z. Yang, et al. (2006). "Nudel contributes to microtubule anchoring at the mother centriole and is involved in both dynein-dependent and -independent centrosomal protein assembly." Mol Biol Cell **17**(2): 680-9.

Gurling, H. M., H. Critchley, et al. (2006). "Genetic association and brain morphology studies and the chromosome 8p22 pericentriolar material 1 (PCM1) gene in susceptibility to schizophrenia." Arch Gen Psychiatry **63**(8): 844-54.

Hackett, C. A. and L. B. Broadfoot (2003). "Effects of genotyping errors, missing values and segregation distortion in molecular marker data on the construction of linkage maps." Heredity **90**(1): 33-8.

Haines, J. L. and M. A. Pericak-Vance (1998). Approaches to Gene Mapping in Complex Human Diseases, Wiley-Liss.

Hallcher, L. M. and W. R. Sherman (1980). "The effects of lithium ion and other agents on the activity of myo-inositol-1-phosphatase from bovine brain." J Biol Chem **255**(22): 10896-901.

Hannula-Jouppi, K., N. Kaminen-Ahola, et al. (2005). "The axon guidance receptor gene ROBO1 is a candidate gene for developmental dyslexia." PLoS Genet **1**(4): e50.

HapMap (2003). "The International HapMap Project." Nature **426**(6968): 789-96.

Hattersley, A. T. and M. I. McCarthy (2005). "What makes a good genetic association study?" Lancet **366**(9493): 1315-23.

Hedrick, P. W. (1987). "Gametic disequilibrium measures: proceed with caution." Genetics **117**(2): 331-41.

Hennah, W., L. Tomppo, et al. (2007). "Families with the risk allele of DISC1 reveal a link between schizophrenia and another component of the same molecular pathway, NDE1." Hum Mol Genet **16**(5): 453-62.

Hodge, S. E. (1993). "Linkage analysis versus association analysis: distinguishing between two models that explain disease-marker associations." Am J Hum Genet **53**(2): 367-84.

Hodge, S. E., M. Boehnke, et al. (1999). "Loss of information due to ambiguous haplotyping of SNPs." Nat Genet **21**(4): 360-1.

Hodge, S. E., M. Durner, et al. (1993). "Better data analysis through data exploration." Am J Hum Genet **53**(3): 775-7.

Hoheisel, J. D. (2006). "Microarray technology: beyond transcript profiling and genotype analysis." Nat Rev Genet **7**(3): 200-10.

Holden, C. (2005). "Sex and the suffering brain." Science **308**(5728): 1574.

Hoogendoorn, B., N. Norton, et al. (2000). "Cheap, accurate and rapid allele frequency estimation of single nucleotide polymorphisms by primer extension and DHPLC in DNA pools." Hum Genet **107**(5): 488-93.

Hosono, S., A. F. Faruqi, et al. (2003). "Unbiased whole-genome amplification directly from clinical samples." Genome Res **13**(5): 954-64.

Hyman, S. E. (2007). "Can neuroscience be integrated into the DSM-V?" Nat Rev Neurosci **8**(9): 725-32.

Iafrate, A. J., L. Feuk, et al. (2004). "Detection of large-scale variation in the human genome." Nat Genet **36**(9): 949-51.

ISBD (2006). Bipolar disorders; from pathophysiology to treatment in the 21st century. 2nd Biennial Conference of the International Society for Bipolar Disorders, Edinburgh, UK.

Itokawa, M., T. Kasuga, et al. (2004). "Identification of a male schizophrenic patient carrying a de novo balanced translocation, t(4; 13)(p16.1; q21.31)." Psychiatry Clin Neurosci **58**(3): 333-7.

Iwamoto, K., M. Bundo, et al. (2004). "Expression of HSPF1 and LIM in the lymphoblastoid cells derived from patients with bipolar disorder and schizophrenia." J Hum Genet **49**(5): 227-31.

Iwamoto, K., C. Kakiuchi, et al. (2004). "Molecular characterization of bipolar disorder by comparing gene expression profiles of postmortem brains of major mental disorders." Mol Psychiatry **9**(4): 406-16.

Jamra, R. A., K. Klein, et al. (2006). "Association study between genetic variants at the PIP5K2A gene locus and schizophrenia and bipolar affective disorder." Am J Med Genet B Neuropsychiatr Genet **141**(6): 663-5.

Jean W. MacCluer, J. L. V. B. R. O. A. R. (1986). "Pedigree analysis by computer simulation." Zoo Biology **5**(2): 147-160.

John, S., N. Shephard, et al. (2004). "Whole-genome scan, in a complex disease, using 11,245 single-nucleotide polymorphisms: comparison with microsatellites." Am J Hum Genet **75**(1): 54-64.

Josée Dupuis, o. b. o. G. (2007). "Effect of linkage disequilibrium between markers in linkage and association analyses." Genetic Epidemiology **31**(S1): S139-S148.

Jungerius, B. J., M. L. Hoogendoorn, et al. (2007). "An association screen of myelin-related genes implicates the chromosome 22q11 PIK4CA gene in schizophrenia." Mol Psychiatry Online (1-9) doi: 10.1038.

Kasperaviciute, D., M. E. Weale, et al. (2007). "Large-scale pathways-based association study in amyotrophic lateral sclerosis." Brain **130**(Pt 9): 2292-301.

Kato, T. (2007). "Molecular genetics of bipolar disorder and depression." Psychiatry Clin Neurosci **61**(1): 3-19.

Kato, T., M. Ishiwata, et al. (2003). "Mechanisms of altered Ca2+ signalling in transformed lymphoblastoid cells from patients with bipolar disorder." Int J Neuropsychopharmacol **6**(4): 379-89.

Kent, W. J., C. W. Sugnet, et al. (2002). "The human genome browser at UCSC." Genome Res **12**(6): 996-1006.

Kimchi-Sarfaty, C., J. M. Oh, et al. (2007). "A "silent" polymorphism in the MDR1 gene changes substrate specificity." Science **315**(5811): 525-8.

Kimura, R., T. Nishioka, et al. (2005). "Allele-specific transcript quantification detects haplotypic variation in the levels of the SDF-1 transcripts." Hum Mol Genet **14**(12): 1579-85.

Kinoshita, Y., T. Suzuki, et al. (2005). "No association with the calcineurin A gamma subunit gene (PPP3CC) haplotype to Japanese schizophrenia." J Neural Transm **112**(9): 1255-62.

Kirk, K. M. and L. R. Cardon (2002). "The impact of genotyping error on haplotype reconstruction and frequency estimation." Eur J Hum Genet **10**(10): 616-22.

Kirkwood, B. R. and J. A. C. Sterne (2003). Essential medical statistics. Malden, Mass., Blackwell Science.

Knuuttila, J., Metzidis, A., Sammalisto, S., Peltonen, L., Perola, M., Saharinen, J. (2007). "CARTOGRAPHER: A tool to generate genetic marker maps for linkage analysis based on markers' physical location and genetic distances ".

Kohn, Y., E. Danilovich, et al. (2004). "Linkage disequlibrium in the DTNBP1 (dysbindin) gene region and on chromosome 1p36 among psychotic patients

from a genetic isolate in Israel: findings from identity by descent haplotype sharing analysis." Am J Med Genet B Neuropsychiatr Genet **128**(1): 65-70.

Kong, A. and N. J. Cox (1997). "Allele-sharing models: LOD scores and accurate linkage tests." Am J Hum Genet **61**(5): 1179-88.

Kong, A., D. F. Gudbjartsson, et al. (2002). "A high-resolution recombination map of the human genome." Nat Genet **31**(3): 241-7.

Kruglyak, L. and M. J. Daly (1998). "Linkage Thresholds for Two-stage Genome Scans." American journal of human genetics **62**(4): 994-995.

Lachman, H. M., J. R. Kelsoe, et al. (1997). "Linkage studies suggest a possible locus for bipolar disorder near the velo-cardio-facial syndrome region on chromosome 22." Am J Med Genet **74**(2): 121-8.

Lachman, H. M., E. Pedrosa, et al. (2007). "Increase in GSK3beta gene copy number variation in bipolar disorder." Am J Med Genet B Neuropsychiatr Genet **144**(3): 259-65.

Lachman, H. M., P. Stopkova, et al. (2005). "Association of schizophrenia in African Americans to polymorphism in synapsin III gene." Psychiatr Genet **15**(2): 127-32.

Lander, E. and L. Kruglyak (1995). "Genetic dissection of complex traits: guidelines for interpreting and reporting linkage results." Nat Genet **11**(3): 241-7.

Lander, E. S. and P. Green (1987). "Construction of multilocus genetic linkage maps in humans." Proc Natl Acad Sci U S A **84**(8): 2363-7.

Law, A. J., J. E. Kleinman, et al. (2007). "Disease-associated intronic variants in the ErbB4 gene are related to altered ErbB4 splice-variant expression in the brain in schizophrenia." Hum Mol Genet **16**(2): 129-41.

Le Hellard, S., A. J. Lee, et al. (2007). "Haplotype analysis and a novel allele-sharing method refines a chromosome 4p locus linked to bipolar affective disorder." Biol Psychiatry **61**(6): 797-805.

Lein, E. S., M. J. Hawrylycz, et al. (2007). "Genome-wide atlas of gene expression in the adult mouse brain." Nature **445**(7124): 168-76.

Lencz, T., C. Lambert, et al. (2007). "Runs of homozygosity reveal highly penetrant recessive loci in schizophrenia." Proc Natl Acad Sci U S A **104** (50):19942-7..

Lerer, B., R. H. Segman, et al. (2003). "Genome scan of Arab Israeli families maps a schizophrenia susceptibility gene to chromosome 6q23 and supports a locus at chromosome 10q24." Mol Psychiatry **8**(5): 488-98.

Lewis, C. (2006). Association Studies; Introduction**:** Powerpoint Presentation at Genetics of Complex Diseases course in CSHL, USA.

Li, Q., D. Hansen, et al. (2001). "Kendrin/pericentrin-B, a centrosome protein with homology to pericentrin that complexes with PCM-1." J Cell Sci **114**(Pt 4): 797-809.

Li, X., A. B. Friedman, et al. (2007). "Lithium regulates glycogen synthase kinase-3beta in human peripheral blood mononuclear cells: implication in the treatment of bipolar disorder." Biol Psychiatry **61**(2): 216-22.

Li, Y., A. Grupe, et al. (2006). "DAPK1 variants are associated with Alzheimer's disease and allele-specific expression." Hum Mol Genet **15**(17): 2560-8.

Liu, Y. L., C. S. Fann, et al. (2007). "More evidence supports the association of PPP3CC with schizophrenia." Mol Psychiatry **12**(10): 966-74.

Lo, H. S., Z. Wang, et al. (2003). "Allelic variation in gene expression is common in the human genome." Genome Res **13**(8): 1855-62.

Macgregor, S., P. M. Visscher, et al. (2002). "Is schizophrenia linked to chromosome 1q?" Science **298**(5602): 2277.

Majerus, P. W. (1992). "Inositol phosphate biochemistry." Annu Rev Biochem **61**: 225-50.

Marazziti, D., B. Dell'Osso, et al. (2006). "Common alterations in the serotonin transporter in platelets and lymphocytes of psychotic patients." Pharmacopsychiatry **39**(1): 35-8.

Mathieu, F., S. Miot, et al. (2008). "Association between the PPP3CC gene, coding for the calcineurin gamma catalytic subunit, and bipolar disorder." Behav Brain Funct **4**(1): 2.

McCarthy, C. (1998). Chromas.

McGuffin, P., J. Knight, et al. (2005). "Whole genome linkage scan of recurrent depressive disorder from the depression network study." Hum Mol Genet **14**(22): 3337-45.

McGuffin, P., M. J. Owen, et al. (2002). Psychiatric genetics and genomics. Oxford ; New York, Oxford University Press.

McPeek, M. S. and L. Sun (2000). "Statistical tests for detection of misspecified relationships by use of genome-screen data." Am J Hum Genet **66**(3): 1076-94.

Mendlewicz, J. and J. D. Rainer (1977). "Adoption study supporting genetic transmission in manic--depressive illness." Nature **268**(5618): 327-9.

Millar, J. K., J. C. Wilson-Annan, et al. (2000). "Disruption of two novel genes by a translocation co-segregating with schizophrenia." Hum Mol Genet **9**(9): 1415-23.

Minogue, S., J. S. Anderson, et al. (2001). "Cloning of a human type II phosphatidylinositol 4-kinase reveals a novel lipid kinase family." J Biol Chem **276**(20): 16635-40.

Moller, H. J. and H. A. Nasrallah (2003). "Treatment of bipolar disorder." J Clin Psychiatry **64 Suppl 6**: 9-17; discussion 28.

Mortensen, P. B., C. B. Pedersen, et al. (2003). "Individual and familial risk factors for bipolar affective disorders in Denmark." Arch Gen Psychiatry **60**(12): 1209-15.

Mowry, B. J., P. A. Holmans, et al. (2004). "Multicenter linkage study of schizophrenia loci on chromosome 22q." Mol Psychiatry **9**(8): 784-95.

Mukhopadhyay N, A. L., Schroeder M, Mulvihill WP, Weeks DE (2006). "Mega2 (Version 3.0 R9)."

Mukhopadhyay, N., L. Almasy, et al. (2005). "Mega2: data-handling for facilitating genetic linkage and association analyses." Bioinformatics **21**(10): 2556-7.

Murray, S. S., A. Oliphant, et al. (2004). "A highly informative SNP linkage panel for human genetic studies." Nat Methods **1**(2): 113-7.

Nackley, A. G., S. A. Shabalina, et al. (2006). "Human catechol-O-methyltransferase haplotypes modulate protein expression by altering mRNA secondary structure." Science **314**(5807): 1930-3.

Nemanov, L., R. P. Ebstein, et al. (1999). "Effect of bipolar disorder on lymphocyte inositol monophosphatase mRNA levels." Int J Neuropsychopharmcol **2**(1): 25-29.

Nevins, J. R. and A. Potti (2007). "Mining gene expression profiles: expression signatures as cancer phenotypes." Nat Rev Genet **8**(8): 601-9.

Ng, P. C. and S. Henikoff (2003). "SIFT: Predicting amino acid changes that affect protein function." Nucleic Acids Res **31**(13): 3812-4.

Nicolae, D. L. and N. J. Cox (2002). "MERLIN...and the geneticist's stone?" Nat Genet **30**(1): 3-4.

Nyholt, D. R. (2004). "A simple correction for multiple testing for single-nucleotide polymorphisms in linkage disequilibrium with each other." Am J Hum Genet **74**(4): 765-9.

O'Connell, J. R. and D. E. Weeks (1998). "PedCheck: a program for identification of genotype incompatibilities in linkage analysis." Am J Hum Genet **63**(1): 259-66.

O'Donnell, T., S. Rotzinger, et al. (2000). "Chronic lithium and sodium valproate both decrease the concentration of myo-inositol and increase the concentration of inositol monophosphates in rat brain." Brain Research **880**: 84-91.

Ohnishi, T., K. Yamada, et al. (2007). "A promoter haplotype of the inositol monophosphatase 2 gene (IMPA2) at 18p11.2 confers a possible risk for bipolar disorder by enhancing transcription." Neuropsychopharmacology **32**(8): 1727-37.

Ott, J. (1999). Analysis of human genetic linkage. Baltimore, Md. ; London, Johns Hopkins University Press.

Ott, J. and D. Rabinowitz (1997). "The effect of marker heterozygosity on the power to detect linkage disequilibrium." Genetics **147**(2): 927-30.

Palmour, R. M., S. Miller, et al. (1994). "A contribution to the differential diagnosis of the "group of schizophrenias": structural abnormality of chromosome 4." I Psychiatry Neurosci **19**(4): 270-7.

Park, N., S. H. Juo, et al. (2004). "Linkage analysis of psychosis in bipolar pedigrees suggests novel putative loci for bipolar disorder and shared susceptibility with schizophrenia." Mol Psychiatry **9**(12): 1091-9.

Pastinen, T., B. Ge, et al. (2005). "Mapping common regulatory variants to human haplotypes." Hum Mol Genet **14**(24): 3963-71.

Pastinen, T. and T. J. Hudson (2004). "Cis-acting regulatory variation in the human genome." Science **306**(5696): 647-50.

Pastinen, T., R. Sladek, et al. (2004). "A survey of genetic and epigenetic variation affecting human gene expression." Physiol Genomics **16**(2): 184-93.

Perlis, R. H., S. Purcell, et al. (2008). "Family-based association study of lithium-related and other candidate genes in bipolar disorder." Arch Gen Psychiatry **65**(1): 53-61.

Perneger, T. V. (1998). "What's wrong with Bonferroni adjustments." Bmj **316**(7139): 1236-8.

Pickard, B. J., A. Christofororou, et al. (2008). "Interacting haplotypes at the NPAS3 locus alter risk of schizophrenia and bipolar disorder." Mol Psychiatry Advanced Online Publication: doi 10.1038..

Pickard, B. S., M. P. Malloy, et al. (2006). "Cytogenetic and genetic evidence supports a role for the kainate-type glutamate receptor gene, GRIK4, in schizophrenia and bipolar disorder." Mol Psychiatry **11**(9): 847-57.

Pickard, B. S., M. P. Malloy, et al. (2005). "Disruption of a brain transcription factor, NPAS3, is associated with schizophrenia and learning disability." Am J Med Genet B Neuropsychiatr Genet **136**(1): 26-32.

Pompanon, F., A. Bonin, et al. (2005). "Genotyping errors: causes, consequences and solutions." Nat Rev Genet **6**(11): 847-59.

Pritchard, J. K. and M. Przeworski (2001). "Linkage disequilibrium in humans: models and data." Am J Hum Genet **69**(1): 1-14.

Pritchard, J. K., M. Stephens, et al. (2000). "Inference of population structure using multilocus genotype data." Genetics **155**(2): 945-59.

Purcell, S., S. S. Cherny, et al. (2003). "Genetic Power Calculator: design of linkage and association genetic mapping studies of complex traits." Bioinformatics **19**(1): 149-50.

R_Development_Core_Team (2006). "R: A Language and Environment for Statistical Computing."

Redon, R., S. Ishikawa, et al. (2006). "Global variation in copy number in the human genome." Nature **444**(7118): 444-54.

Rice, J., T. Reich, et al. (1987). "The familial transmission of bipolar illness." Arch Gen Psychiatry **44**(5): 441-7.

Risch, N. and K. Merikangas (1996). "The future of genetic studies of complex human diseases." Science **273**(5281): 1516-7.

Rowe, M. K., C. Wiest, et al. (2007). "GSK-3 is a viable potential target for therapeutic intervention in bipolar disorder." Neurosci Biobehav Rev **31**(6): 920-31.

Rozen, S. and H. Skaletsky (2000). Primer3 on the WWW for general users and for biologist programmer. Bioinformatics Methods and Protocols: Methods in Molecular Biology. . S. Misener and S. A. Krawetz. Totowa, NJ, Humana Press.

Saito, T., F. Guan, et al. (2001). "Mutation analysis of SYNJ1: a possible candidate gene for chromosome 21q22-linked bipolar disorder." Mol Psychiatry **6**(4): 387-95.

Saito, T., P. Stopkova, et al. (2003). "Polymorphism screening of PIK4CA: possible candidate gene for chromosome 22q11-linked psychiatric disorders." Am J Med Genet B Neuropsychiatr Genet **116**(1): 77-83.

Salyakina, D., S. R. Seaman, et al. (2005). "Evaluation of Nyholt's procedure for multiple testing correction." Hum Hered **60**(1): 19-25; discussion 61-2.

Sambrook, J., E. F. Fritsch, et al. (1989). Molecular cloning : a laboratory manual. Cold Spring Harbor, Cold Sprng Harbor Laboratory Press.

Sanders, A. R., J. Duan, et al. (2008). "No Significant Association of 14 Candidate Genes With Schizophrenia in a Large European Ancestry Sample: Implications for Psychiatric Genetics." Am J Psychiatry.

Savitsky, K., A. Bar-Shira, et al. (1995). "A single ataxia telangiectasia gene with a product similar to PI-3 kinase." Science **268**(5218): 1749-53.

Sawcer, S., H. B. Jones, et al. (1997). "Empirical genomewide significance levels established by whole genome simulations." Genet Epidemiol **14**(3): 223-9.

Sawcer, S. J., M. Maranian, et al. (2004). "Enhancing linkage analysis of complex disorders: an evaluation of high-density genotyping." Hum Mol Genet **13**(17): 1943-9.

Schaid, D. J., S. K. McDonnell, et al. (2007). "Affected relative pairs and simultaneous search for two-locus linkage in the presence of epistasis." Genet Epidemiol **31**(5): 431-49.

Schaid, D. J., S. K. McDonnell, et al. (2002). "Caution on pedigree haplotype inference with software that assumes linkage equilibrium." Am J Hum Genet **71**(4): 992-5.

Schreiner, A., M. Ruonala, et al. (2007). "Junction protein shrew-1 influences cell invasion and interacts with invasion-promoting protein CD147." Mol Biol Cell **18**(4): 1272-81.

Schumacher, J., R. Kaneva, et al. (2005). "Genomewide scan and fine-mapping linkage studies in four European samples with bipolar affective disorder suggest a new susceptibility locus on chromosome 1p35-p36 and provides further evidence of loci on chromosome 4q31 and 6q24." Am J Hum Genet **77**(6): 1102-11.

Schwab, S. G., M. Knapp, et al. (2006). "Evidence for association of DNA sequence variants in the phosphatidylinositol-4-phosphate 5-kinase IIalpha gene (PIP5K2A) with schizophrenia." Mol Psychiatry **11**(9): 837-46.

Seelan, R. S., A. Khalyfa, et al. (2007). "Deciphering the lithium transcriptome: Microarray profiling of lithium-modulated gene expression in human neuronal cells." Neuroscience **151**(4):1184-97.

Segurado, R., S. D. Detera-Wadleigh, et al. (2003). "Genome scan meta-analysis of schizophrenia and bipolar disorder, part III: Bipolar disorder." Am J Hum Genet **73**(1): 49-62.

Sengul, H., D. E. Weeks, et al. (2001). "A survey of affected-sibship statistics for nonparametric linkage analysis." Am J Hum Genet **69**(1): 179-90.

Sham, P. (1998). Statistics in human genetics. London, Arnold.

Shamir, A., R. P. Ebstein, et al. (1998). "Inositol monophosphatase in immortalized lymphoblastoid cell lines indicates susceptibility to bipolar disorder and response to lithium therapy." Mol Psychiatry **3**(6): 481-2.

Shih, M. C. and A. S. Whittemore (2001). "Allele-sharing among affected relatives: non-parametric methods for identifying genes." Stat Methods Med Res **10**(1): 27-55.

Shugart, Y. Y., J. Samuels, et al. (2006). "Genomewide linkage scan for obsessive-compulsive disorder: evidence for susceptibility loci on chromosomes 3q, 7p, 1q, 15q, and 6q." Mol Psychiatry **11**(8): 763-70.

Simon-Sanchez, J., S. Scholz, et al. (2007). "Genome-wide SNP assay reveals structural genomic variation, extended homozygosity and cell-line induced alterations in normal individuals." Hum Mol Genet **16**(1): 1-14.

Sjoholt, G., R. P. Ebstein, et al. (2004). "Examination of IMPA1 and IMPA2 genes in manic-depressive patients: association between IMPA2 promoter polymorphisms and bipolar disorder." Mol Psychiatry **9**(6): 621-9.

Sjoholt, G., A. Molven, et al. (1997). "Genomic structure and chromosomal localization of a human myo-inositol monophosphatase gene (IMPA)." Genomics **45**(1): 113-22.

Sklar, P. S., J. ; Fan, J.; Ferreira, M.; Perlis, R.; Chambert, K. ; Nimgaonkar, V; McQueen, M.B.; Faraone, S.V. ; Kirby, A.; de Bakker, P.I.K.; Ogdie, M.N.; Thase, M.E.; Sachs, G.S.; Todd-Brown, K. ; Gabriel, S.B. ; Sougnez, C ; Gates, C.; Blumenstiel, B.; Defelice, M.; Ardlie, K. ; Franklin, J. ; Muir, W.J.; McGhee, K.A. ; MacIntyre, D.A.; McLean, A.; VanBeck, M.; McQuillin, A.; Bass, N.J.; Robinson, M. ; Lawrence, J. ; Anjorin, A.; Curtis, D. ; Scolnick, E.M.; Daly,

M.J. ; Blackwood, D.H.; Gurling, H.M.D.; Purcell, S.H. (2008). "Whole-genome association study of bipolar disorder." Mol Psychiatry **13**(6): 558-69.

Smith, A. L. and M. M. Weissmann (1992). Epidemiology. Handbook of affective disorders. E. S. Paykel. Edinburgh, Churchill Livingstone**:** xii, 699 p.

Soares, J. C., C. S. Dippold, et al. (2001). "Increased platelet membrane phosphatidylinositol-4,5-bisphosphate in drug-free depressed bipolar patients." Neurosci Lett **299**(1-2): 150-2.

Sobel, E. and K. Lange (1996). "Descent graphs in pedigree analysis: applications to haplotyping, location scores, and marker-sharing statistics." Am J Hum Genet **58**(6): 1323-37.

Speed, T. and M. S. Waterman (1996). Genetic mapping and DNA sequencing. New York ; London, Springer.

SPSS_Inc. (2005). "SPSS for Windows, Rel. 14.0 2005. Chicago: SPSS Inc." SPSS for Windows, Rel. 14.0 2005. Chicago: SPSS Inc.

St Clair, D., D. Blackwood, et al. (1990). "Association within a family of a balanced autosomal translocation with major mental illness." Lancet **336**(8706): 13-6.

Stopkova, P., T. Saito, et al. (2003). "Polymorphism screening of PIP5K2A: a candidate gene for chromosome 10p-linked psychiatric disorders." Am J Med Genet B Neuropsychiatr Genet **123**(1): 50-8.

Stopkova, P., T. Saito, et al. (2004). "Identification of PIK3C3 promoter variant associated with bipolar disorder and schizophrenia." Biol Psychiatry **55**(10): 981-8.

Stopkova, P., J. Vevera, et al. (2004). "Analysis of SYNJ1, a candidate gene for 21q22 linked bipolar disorder: a replication study." Psychiatry Res **127**(1-2): 157-61.

Strachan, T. and A. P. Read (1999). Human molecular genetics. Oxford, BIOS Scientific.

Strassburg, C. P. and M. P. Manns (2002). "Autoantibodies and autoantigens in autoimmune hepatitis." Semin Liver Dis **22**(4): 339-52.

Stratton, M. (2008). "Genome resequencing and genetic variation." Nat Biotechnol **26**(1): 65-6.

Sullivan, P. F., C. Fan, et al. (2006). "Evaluating the comparability of gene expression in blood and brain." Am J Med Genet B Neuropsychiatr Genet **141**(3): 261-8.

Sullivan, P. F., M. C. Neale, et al. (2000). "Genetic epidemiology of major depression: review and meta-analysis." Am J Psychiatry **157**(10): 1552-62.

Taberlet, P., S. Griffin, et al. (1996). "Reliable genotyping of samples with very low DNA quantities using PCR." Nucleic Acids Res **24**(16): 3189-94.

Terwilliger, J. D. and J. Ott (1994). Handbook of human genetic linkage. Baltimore ; London, Johns Hopkins University Press.

Thalamuthu, A., I. Mukhopadhyay, et al. (2005). "A comparison between microsatellite and single-nucleotide polymorphism markers with respect to two measures of information content." BMC Genetics **6(Suppl 1)** (S27).

Thiele, H. and P. Nurnberg (2005). "HaploPainter: a tool for drawing pedigrees with complex haplotypes." Bioinformatics **21**(8): 1730-2.

Thompson, E. A. (2000). Statistical Inference from Genetic Data on Pedigrees.

Thomson, P. A., N. R. Wray, et al. (2005). "Sex-specific association between bipolar affective disorder in women and GPR50, an X-linked orphan G protein-coupled receptor." Mol Psychiatry **10**(5): 470-8.

Thomson, R., S. Quinn, et al. (2007). "The advantages of dense marker sets for linkage analysis with very large families." Hum Genet **121**(3-4): 459-68.

Tong, L. and E. Thompson (2008). "Multilocus lod scores in large pedigrees: combination of exact and approximate calculations." Hum Hered **65**(3): 142-53.

Tsuang, M. T., N. Nossova, et al. (2005). "Assessing the validity of blood-based gene expression profiles for the classification of schizophrenia and bipolar disorder: a preliminary report." Am J Med Genet B Neuropsychiatr Genet **133**(1): 1-5.

Turner, D. J., J. Shendure, et al. (2006). "Assaying chromosomal inversions by single-molecule haplotyping." Nat Methods **3**(6): 439-45.

Underwood, S. L., A. Christoforou, et al. (2006). "Association analysis of the chromosome 4p-located G protein-coupled receptor 78 (GPR78) gene in bipolar affective disorder and schizophrenia." Mol Psychiatry **11**(4): 384-94.

Vawter, M. P., E. Ferran, et al. (2004). "Microarray screening of lymphocyte gene expression differences in a multiplex schizophrenia pedigree." Schizophr Res **67**(1): 41-52.

Vazza, G., C. Bertolin, et al. (2007). "Genome-wide scan supports the existence of a susceptibility locus for schizophrenia and bipolar disorder on chromosome 15q26." Mol Psychiatry **12**(1): 87-93.

Venken, T. and J. Del-Favero (2007). "Chasing genes for mood disorders and schizophrenia in genetically isolated populations." Hum Mutat **28**(12): 1156-70.

Visscher, P. M., C. S. Haley, et al. (1999). "Detecting QTLs for uni- and bipolar disorder using a variance component method." Psychiatr Genet **9**(2): 75-84.

Wang, G. S. and T. A. Cooper (2007). "Splicing in disease: disruption of the splicing code and the decoding machinery." Nat Rev Genet **8**(10): 749-61.

Washizuka, S., C. Kakiuchi, et al. (2003). "Association of mitochondrial complex I subunit gene NDUFV2 at 18p11 with bipolar disorder." Am J Med Genet B Neuropsychiatr Genet **120**(1): 72-8.

Washizuka, S., C. Kakiuchi, et al. (2005). "Expression of mitochondria-related genes in lymphoblastoid cells from patients with bipolar disorder." Bipolar Disord **7**(2): 146-52.

Weeks, D. E., M. Lathrop, et al. (2006). Wellcome Trust Human Genome Analysis Course, Hinxton, Cambridge.

Wei, Y. J., H. Q. Sun, et al. (2002). "Type II phosphatidylinositol 4-kinase beta is a cytosolic and peripheral membrane protein that is recruited to the plasma membrane and activated by Rac-GTP." J Biol Chem **277**(48): 46586-93.

Weir, B. S., A. D. Anderson, et al. (2006). "Genetic relatedness analysis: modern data and new challenges." Nat Rev Genet **7**(10): 771-80.

Weissman, M. M., P. J. Leaf, et al. (1988). "Affective disorders in five United States communities." Psychol Med **18**(1): 141-53.

Wigginton, J. E. and G. R. Abecasis (2005). "PEDSTATS: descriptive statistics, graphics and quality assessment for gene mapping data." Bioinformatics **21**(16): 3445-7.

Wigginton, J. E., D. J. Cutler, et al. (2005). "A note on exact tests of Hardy-Weinberg equilibrium." Am J Hum Genet **76**(5): 887-93.

Wilkins, J. M., L. Southam, et al. (2007). "Extreme context specificity in differential allelic expression." Hum Mol Genet **16**(5): 537-46.

Williams, N. M., B. Glaser, et al. (2008). "Strong evidence that GNB1L is associated with schizophrenia." Hum Mol Genet **17**(4): 555-66.

Williams, N. M., M. I. Rees, et al. (1999). "A two-stage genome scan for schizophrenia susceptibility genes in 196 affected sibling pairs." Hum Mol Genet **8**(9): 1729-39.

Williams, R. S., L. Cheng, et al. (2002). "A common mechanism of action for three mood-stabilizing drugs." Nature **417**(6886): 292-5.

World Health Organization (2005). ICD-10, International statistical classification of diseases and related health problems. Geneva, World Health Organization**:** 1 CD-ROM.

World Health Organization. (1994). ICD-10 : international statistical classification of diseases and related health problems. Vol.3, Alphabetical index. Geneva, World Health Organization.

Wray, G. A. (2007). "The evolutionary significance of cis-regulatory mutations." Nat Rev Genet **8**(3): 206-16.

WTCCC (2007). "Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls." Nature **447**(7145): 661-78.

Yoon, I. S., P. P. Li, et al. (2001). "Altered IMPA2 gene expression and calcium homeostasis in bipolar disorder." Mol Psychiatry **6**(6): 678-83.

Yoonhee Kim, P. D. E. M. G. H. K. J. E. B.-W. (2008). "Examining the effect of linkage disequilibrium between markers on the Type I error rate and power of

nonparametric multipoint linkage analysis of two-generation and multigenerational pedigrees in the presence of missing genotype data." Genetic Epidemiology **32**(1): 41-51.

Yoshikawa, T., M. Kikuchi, et al. (2001). "Evidence for association of the myo-inositol monophosphatase 2 (IMPA2) gene with schizophrenia in Japanese samples." Mol Psychiatry **6**(2): 202-10.

Zhang, Y., Y. Zhou, et al. (2007). "Comparative analysis of selenocysteine machinery and selenoproteome gene expression in mouse brain identifies neurons as key functional sites of selenium in mammals." J. Biol. Chem. **283**(4): 2427-38.

Zhang, Y., S. N. Zolov, et al. (2007). "Loss of Vac14, a regulator of the signaling lipid phosphatidylinositol 3,5-bisphosphate, results in neurodegeneration in mice." Proc Natl Acad Sci U S A **104**(44): 17518-23.

Zondervan, K. T. and L. R. Cardon (2007). "Designing candidate gene and genome-wide case-control association studies." Nat Protoc **2**(10): 2492-501.

Zubenko, G. S., B. Maher, et al. (2003). "Genome-wide linkage survey for genetic loci that influence the development of depressive disorders in families with recurrent, early-onset, major depression." Am J Med Genet B Neuropsychiatr Genet **123**(1): 1-18.

9. Appendix A

Statistical Equations

TEST	DEFINITION	EQUATION
Mean	Average	X bar` = ∑x/n
Range	Distance between largest & smallest value	Highest value-lowest value
Variance	Measure of variation, aka second moment (m_2)	$S^2 = \sum(x-x\ bar)^2/n-1$
Standard deviation	Measures amount of variability in population	$S = \sqrt{\sum(x-x\ bar)^2/n-1}$
Coefficient of variation	Compares size of variation to size of observation	Cv =s/x bar
Standard Error	Measures amount of variability in the sample mean (depends on sample size)	s.e. =s/√n
Confidence Interval	Range of plausible values for population mean, given sample mean	95%CI=x bar -1.96x s.e. & x bar +1.96 x s.e.
T Test	Compares two means	$X\ bar_1 - x\ bar_0$/s.e., $d.f.=n_1+n_0-2$
One way ANOVA	Compares subgroups defined by one exposure	F=Between-groups MS (mean square)/Within groups MS, $d.f.=d.f._{between-groups}$, $d.f._{Within-groups}=k-1$, $n-k$
Two way ANOVA	Compares subgroups by more than one exposure	F=MS effect/MS residual
Correlation coefficient	Strength of the linear association	$r=\sum(x-x\ bar)(y-y\ bar)/\sqrt{[\sum(x-x\ bar)^2\sum(y-y\ bar)^2]}$
Linear interpolation	Constructs new data points from a discrete set of known data points	$y=y_a + ((x-x_a)(y_b-y_a)/(x_b-x_a))$ at the point (x,y) between (x_a, y_a) and (x_b, y_b)

Mean error rate per locus	Ratio between m_l, the number of single-locus genotypes including at least one allelic mismatch, and *nt* the number of replicated single-locus genotypes	$E_l=m_l/nt$
Empirical threshold	Calculate a desired significance level from simulations	(#>T)/(Total #) = 0.05 or 0.01
N^{th} moment	Calculated from a distribution, based on the powers of variance	$m_x=\sum(x-xbar)^N/n$
Coefficient of Skewness	A measure of the asymmetry of a distribution. For a symmetrical distribution, the coefficient is zero. Significant positive values have a long right tail. Significant negative values have a long left tail. A skewness value more than twice it's standard error indicates a departure from symmetry (Kirkwood and Sterne 2003; SPSS_Inc. 2005).	$m_3m_2^{-3/2}$
Coefficient of Kurtosis	A measure of the spread of the distribution. For a normal distribution the coefficient of kurtosis is zero. Positive kurtosis indicates the observations cluster more and have longer tails than those in normal distribution. Negative kurtosis indicates the observations cluster less and have shorter tails (Kirkwood and Sterne 2003; SPSS_Inc. 2005).	$m_4m_2^{-2}-3$

10. Appendix B

Parameters for Parametric Linkage Analysis

Parametric linkage analysis required disease locus parameters to be specified. Below are the files for each of the three disease models evaluated. The first line for each model notes the name "Trait" and the trait/disease frequency in the population. The following nine lines specify the penetrance function for each liability class, one to eight. Liability class one to four applies to individual in good mental health in age classes: <20 years, 20-30 years, 31-40 years and >40 years. Liability class five pertains to founder individuals who are married into the family. Liability class six relates to individuals with a bipolar disorder diagnosis and liability class seven to individuals with a recurrent major depression diagnosis. Liability class eight applies to individuals diagnosed with other psychiatric illness including anxiety states, alcoholism, single episode depression or minor depression. Please refer to chapter 1 and 6 for further explanation of the disease models.

a) Dominant model under the broad phenotypic model

```
TRAIT 0.03
LIABILITY = 1     0.016,0.250,0.250
LIABILITY = 2     0.024,0.390,0.390
LIABILITY = 3     0.038,0.610,0.610
LIABILITY = 4     0.047,0.750,0.750
LIABILITY = 5     0.047,0.750,0.750
LIABILITY = 6     0.005,0.290,0.290
LIABILITY = 7     0.047,0.750,0.750
LIABILITY = 8     0.500,0.500,0.500
OTHERWISE   0.500,0.500,0.500
```

b) Dominant model under the narrow phenotypic model

```
TRAIT 0.007
LIABILITY = 1     0.001,0.046,0.046
LIABILITY = 2     0.003,0.195,0.195
LIABILITY = 3     0.004,0.230,0.230
LIABILITY = 4     0.005,0.290,0.290
LIABILITY = 5     0.005,0.290,0.290
LIABILITY = 6     0.005,0.290,0.290
LIABILITY = 7     0.500,0.500,0.500
LIABILITY = 8     0.500,0.500,0.500
```

```
OTHERWISE   0.500,0.500,0.500
```

c) Recessive model under the broad phenotypic model

```
TRAIT 0.12  *     DB_Recessive
LIABILITY = 1     0.001,0.001,0.150
LIABILITY = 2     0.001,0.001,0.620
LIABILITY = 3     0.001,0.001,0.700
LIABILITY = 4     0.001,0.001,0.700
LIABILITY = 5     0.001,0.001,0.700
LIABILITY = 6     0.001,0.001,0.700
LIABILITY = 7     0.001,0.001,0.700
LIABILITY = 8     0.500,0.500,0.500
OTHERWISE   0.500,0.500,0.500
```

11. Appendix C

Published Paper